Spatial Statistics

Theory, modelling, and inference for spatial point patterns

Dominic Schuhmacher
Universität Bern

January 2011

Preface

These lecture notes accompanied a course by the same title held during the autumn semester 2010 at the University of Bern. They consist of a large probabilistic part on point process theory (Chapter 2), which may be read independently of the rest, and of a statistical part (everything else), which builds to some degree on Chapter 2, but does not require very detailed knowledge of it.

I would like to thank Andrew Atkinson, Benjamin Baumgartner, Chris Kopp, Kaspar Stucki, and Niki Zumbrunnen (in no particular order) for the correction of several typos in the original version and many interesting questions and discussions. Many thanks to Fabian Kück for pointing out some mistakes in later versions.

Further corrections, comments, or ideas for improvement are always very welcome!

Dominic Schuhmacher
Bern, January 2011
Göttingen, December 2014

Contents

Chapter 1

Introduction and data examples

In spatial statistics we study models for data that have a non-negligible spatial (geographic, geometric or topological) component. The subject can be roughly divided into three different domains.

Geostatistics

Math. objects: random fields, stochastic processes indexed over $\mathbb{R}^d$.
Typical application: predicting a spatial map of the occurence of some mineral resource based on a net of sample points.

Lattice Models

Math. objects: stochastic processes indexed over a regular or irregular lattice/graph.
Typical applications: denoising pixel images from a satellite or determining the prevalence of a certain disease in the various Swiss cantons.

Point patterns

Math. objects: random counting measures, random finite (or sometimes countably infinite) sets.
Typical applications: understanding the growing patterns of trees in a forest; many more later on in this chapter.

This course is solely concerned with the third domain, point patterns. Spatial point pattern models are fundamentally different from models encountered in multivariate statistics inasmuch as typically: a) the total number of points as well the positions of the points are random; and more importantly b) the positions of the points are usually not independent, but have a spatial dependence structure, most commonly in the sense that the presence of points in a certain region excites or inhibits the presence of other points nearby.

In the large second chapter we develop the fundamental theory of point processes on locally compact spaces (with special attention to $\mathbb{R}^d$), before we come to the more

statistical topics of descriptive quantities, point pattern models, and their inference. In the rest of the current chapter we present an overview of some rather advanced applications for spatial point pattern techniques. Analysing these data and studying the corresponding models in detail is beyond the scope of this course. However, the presented applications give us the motivation to investigate various point pattern tools and concepts later on, and will be rather easy to work out for the interested reader by herself, once the simpler models studied in this course are understood.

1.1 Galaxy redshift surveys

In the last few decades redshift surveys have become the main tool for creating 3d-maps of the universe. Redshift describes the phenomenon that under certain circumstances light from an object is shifted to a lower wavelenght and therefore appears "redder". The simplest cause for redshift is that an object moves away from an observer, which creates an analogue phenomenon as the Doppler effect for sound waves. However, most redshift observed on cosmological scales is not so much based on the relative velocities of objects directly, but rather on the expansion of the universe and some feature of space-time topology. Hubble's law states that the redshift of a galaxy (and correspondingly the speed with which it is receding from our galaxy) is proportional to its distance from us, which is exploited in constructing 3d-maps of the universe. Due to local disturbances, redshift surveys give only an approximate picture of the universe. Strictly speaking the data lives in a slightly distorted version of space referred to as "redshift space". Figure 1.1 shows an example of two slices through data from the recent 2dF Galaxy Redshift Survey. The whole dataset contains the 3d-positions of 221,414 galaxies.

Two things become apparent from looking at the picture. First the irregular web-like structure that consists of clusters, filaments, and (in 3d) membranes of galaxies around larger empty cells. This intricate structure is "real" and has become known as the *cosmic web*. Second there appear to be fine lines pointing towards the center (our galaxy) that have been dubbed *"fingers of god"*, but are pure artifacts due to redshift space distortions.

The overall structure of the universe has been studied intensively, and it is generally agreed that (apart from redshift distortions) the universe is on reasonably large scales homogeneous (the result of a stationary and isotropic point process, concepts introduced in Chapter 2), and that its pair correlation function[1] g follows a power law of the form

$$g(r) = 1 + \left(\frac{r_o}{r}\right)^{\gamma},$$

[1]Essentially the pair correlation function $g(r)$ gives the probability of having two points in infinitesimal balls at distance r apart, normalized with respect to the expected number of points in these balls.

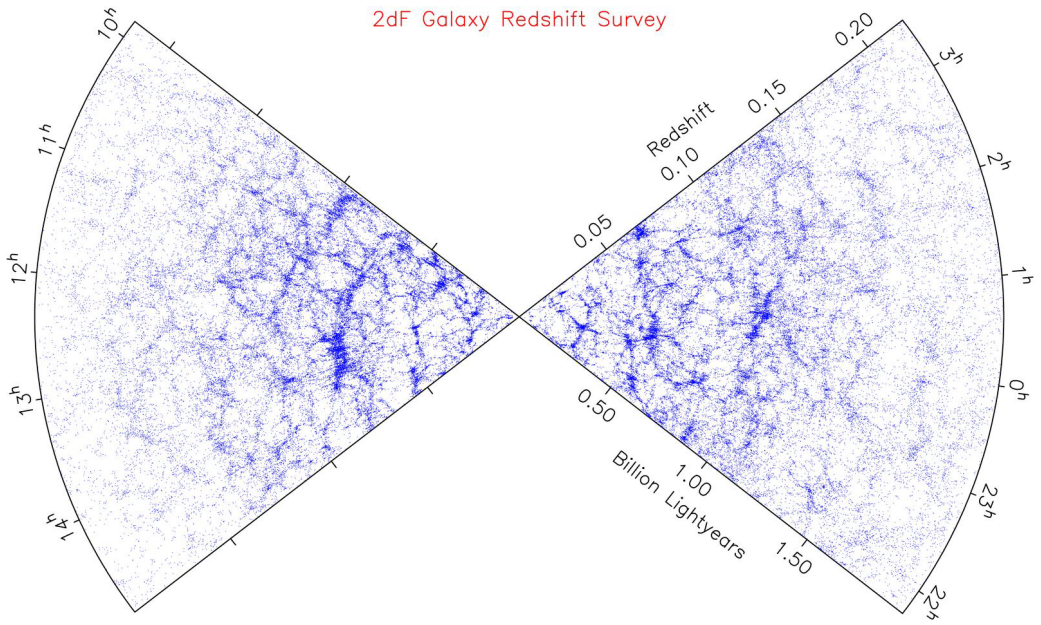

Figure 1.1: Two slices through the 2dF Galaxy Redshift Survey data.

where modern estimates of γ and r_o are around $\gamma = 1.8$ and $r_o = 5.77h^{-1}$ Megaparsecs[2]. This means that the galaxies are distributed more or less independently at large distances, but show a strong clustering behaviour at cosmologically close distances (note also the pole at 0). Furthermore the smooth decline of the pair correlation function over several orders of magnitude means that there are no characteristic scales. This is interpreted as one aspect of a "fractal" nature of the cosmic web.

Open questions include finding realistic and preferably simple point process models which have a pair correlation function of the above form. It is rather easy to find unrealistic models, which for example do not show the fractal behaviour of the cosmic web. Also studying higher-order correlation functions is an ongoing research topic.

More information can be found in the survey article by Jones et al. (2004)

1.2 Earthquake catalogues

A traditional application domain of point process modelling is statistical seismology. The data consists most frequently of the four-dimensional space-time process of epicentres equipped with so called marks (additional information attached to the points) in form of the magnitude of the seismic events. Such a process may be interpreted as a four-dimensional point process with points (t_i, x_i, y_i, M_i), where t_i is the time of the event,

[2]The "small" distance of approx. 265,000 light years (if my computations are correct). 1 Megaparsec (Mpc) is about $3.26 \cdot 10^6$ light years; h is Hubble's proportionality constant, the ratio between galaxy velocity and distance, which is estimated at a bit more than 70 (km/sec)/Mpc.

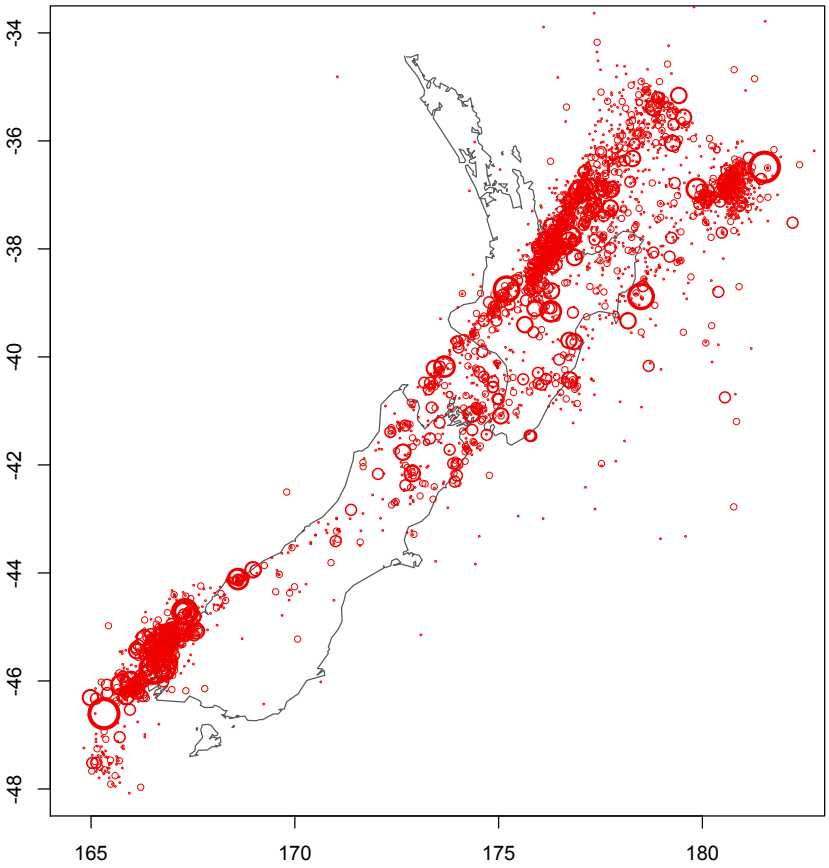

Figure 1.2: Epicentres of earthquakes of magnitude ≥ 4 that occured in the 10 years from 1/6/2000 to 31/5/2010 around New Zealand. 4148 events in total.

(x_i, y_i) are the spatial coordinates and M_i is the magnitude. Figure 1.2 gives an example in two dimensions, where the time component is omitted and the magnitude is shown via the circle sizes. For modelling, time plays a special role. From a theoretical point of view, this is because its natural ordering enables us to study point process characteristics that do not exist or are not as intuitive for processes on purely spatial coordinates. From a practical point of view, this is because we can have information available only from the past, but not from the future. Also the mark space plays a special role as it is usually set aside from any direct relation with the other two coordinates.

Statistical seimology has developed quite independently from the main stream of point pattern statistics, mainly due to specific physical realities in the development of earthquakes that have to be accounted for in modelling. One main feature is that earthquakes typically occur in spatio-temporal clusters of a main shock and several smaller aftershocks (and some "foreshocks") in the same region.

To give a flavour of the complexity and the main ideas of such models, we briefly look at Ogata's (1998) space-time ETAS (Epidemic Type Aftershock Sequence) model for

earthquakes of magnitude above some threshold M_c. One of the main ideas is that there is some *background seismicity* $\mu(x, y)$ that is constant in time on top of which the *clustering seismicity*, which is described by a branching process (each event, whether a main event or triggered by another event independently triggers further events) where offspring of magnitude $< M_o$ is killed. Like many models it is formulated in terms of the space-time conditional intensity $\lambda(t, x, y, M)$, which basically describes the probability of having a point of the four-dimensional process in an infinitesimal interval around (t, x, y, M) (normalized by the volume of the interval) given the whole history of the point process up time t. Conditional intensities are very important and we will study them for spatial point processes in detail in later parts of the course.

In the case of the ETAS model the space-time conditional intensity at (t, x, y, M) for earthquake data (t_i, x_i, y_i, M_i), $i = 1, \ldots, n$, is given as

$$\lambda(t, x, y, M) = J(M)\lambda(t, x, y)$$

with

$$\lambda(t, x, y) = \mu(x, y) + \sum_{i:t_i<t} \kappa(M_i)g(t - t_i)f(x - x_i, y - y_i \,|\, M_i),$$

where

- $J(M)$ is the probability density for the magnitude (independent of position!), modeled according to the so-called Gutenberg-Richter law as the shifted exponential density
$$J(M) = \beta e^{-\beta(M-M_o)} \quad \text{for } M \geq M_o.$$
- $\kappa(M)$ (cluster size factor for magnitude M) is the expected number of events triggered from an event of magnitude M, given by
$$\kappa(M) = Ae^{\alpha(M-M_o)} \quad \text{for } M \geq M_o.$$
- $g(t)$ is the probability density for the occurence times of triggered events, modeled according to the so-called Omori law (independent of position!), modeled as
$$g(t) = \frac{p-1}{c}\left(1 + \frac{t}{c}\right)^{-p} \quad \text{for } t > 0.$$
- $f(x, y \,|\, M)$ is the probability density for the locations of the triggered events, modelled either as
$$f(x, y \,|\, M) = \frac{1}{2\pi De^{\alpha(M-M_o)}} \exp\left(-\frac{x^2 + y^2}{2De^{\alpha(M-M_o)}}\right)$$
(exponential decay; modelling short range spatial dependence) or as
$$f(x, y \,|\, M) = \frac{q-1}{\pi De^{\alpha(M-M_o)}}\left(1 + \frac{x^2 + y^2}{De^{\alpha(M-M_o)}}\right)^{-q}$$
(power decay; modelling long range spatial dependence).

Given the background seismicity μ up to a constant ν the model is fitted by maximizing the log-likelihood

$$\log L(\theta) = \sum_{i=1}^{n} \log \lambda(t_i, x_i, y_i) - \iiint \lambda(t, x, y)\, dt\, dx\, dy$$

with respect to the parameter vector $\theta = (\nu, A, \alpha, c, p, D)$ (including also q in case of the power decay model), where the triple integral is taken over the whole space/time observation window. Maximum likelihood estimation in simpler models will be extensively treated in Chapter 5. If the spatial dependence in the background seismicity is unknown and there is no reliable estimate based on other data, an alternating algorithm can be used that goes back and forth between estimating μ at the data points (x_i, y_i), $i = 1, \ldots, n$, and maximizing the likelihood given these values, until convergence.

One central question for the space-time ETAS model is to test the null model of only background seismicity against the alternative that there are abnormal seismic events. Such events are unnatural activity or quiescence, which can be a precursor of a large earthquake.

1.3 Spatial epidemiology data

The data shown in Figure 1.3 is a classical example from spatial epidemiology. Given are the precise domicile locations of new cases of cancer of the larynx (red bullets, 58 cases) and of the lung (black pluses, 978 cases) recorded from 1974–1983 in the Chorley/South Ribble area of Lancashire, UK. The blue crosshairs mark the position of an industrial incinerator (no longer in use now). The main question is an assessment of whether the incinerator leads to an increased incidence of laryngeal cancer in its vicinity (note the cluster of four cases close to the incinerator). An influence of the incinerator on the incidence of lung cancer is precluded, so that the lunger cancer patients can be used as a control group, which allows to account for the spatially varying density of the susceptible population.

Modelling data of the above form has become increasingly popular from the 1980ies on, mainly in the context of concerns about increased incidence of certain types of cancer in the vicinty of nuclear installations. Several modelling attempts have been made that "destroy information", e.g. by modelling count information in a neighbourhood of the incinerator and in certain other neighbourhoods. However, it is not clear what a good (representative) neighbourhood of the incinerator or of other points should be. Nor is it clear what a good criterion is for defining clusters of points in order to analyse the data "cluster-wise".

Diggle (1990) proposed the following model for the intensity λ of cases as a function

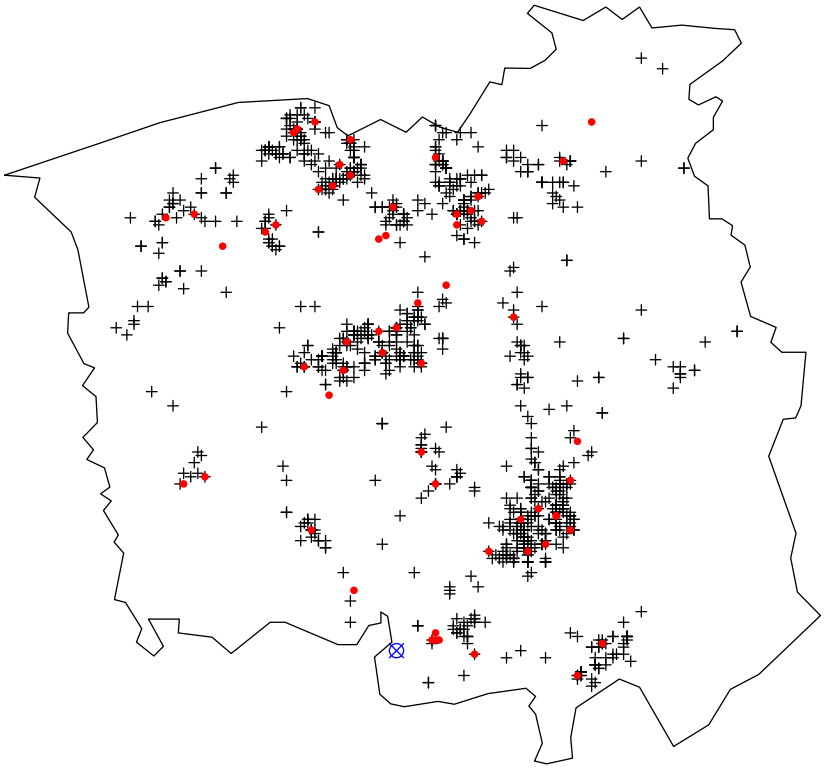

Figure 1.3: Cases of cancers (domiciles of patients) of the larynx (filled circles) and lung (pluses) 1974–1983 in the south of Lancashire, UK; and the location of an industrial incinerator.

of space[3]. He assumed that cases occur independently of one another, but with a non-uniform distribution, which yields a so-called inhomogeneous Poisson process. We will encounter Poisson processes and intensity functions again in Chapter 2. Suppose now that the incinerator is located at position x_0; then we model

$$\lambda(x) = \varrho\lambda_0(x)f(x - x_o; \theta).$$

Here ϱ is a location-independent scaling parameter that describes the prevalence of laryngeal cancer, $\lambda_0(x)$ is the intensity function of the population at risk and takes care of the spatial variation in the intensity of occurrences of laryngeal cancer in absence of an incinerator effect; finally f describes a potential incinerator effect, e.g.

$$f(u; \theta) = 1 + \theta_1 \exp(-\theta_2\|u\|),$$

where $\theta_1 \geq 0$ is the magnitude and $\theta_2 \geq 0$ describes the spatial scale of the incinerator effect. Note that $\theta_1 = 0$ corresponds to the absence of an incinerator effect.

Diggle (1990) and Diggle and Rowlingson (1994) fitted this model with different methods, which both revealed a significant incinerator effect. Model diagnostics along the lines that we will consider in Chapter 5 reveal (only) slight problems with the fit.

1.4 Positions of biological cells

Our last example is concerned with the amacrine cells in the retina of a rabbit (Figure 1.4). The retina is a neural structure at the back of the eye, which consists of various layers that convert light into electrical impulses. The amacrine cells reside in the so-called inner plexiform layer in the lower (back) part of the retina. The 152 white discs are "on" cells, which are excited by an increase of illumination, and the 142 black discs are "off" cells, which are excited by a decrease of illumination. The figure shows a 1060 by 662 μm section of a larger similar looking part of the retina, and is more or less to scale in the sense that the circle diameters roughly correspond to the usual cell sizes of 10 μm. Thus the regular pattern observed within the on cells and within the off cells cannot be (only) explained by the physical extensions of the cells. Between the on and off cells overlaps are possible, which is due to the fact that they reside in different layers that were projected on a common plane.

The main biological interest in this dataset is to decide between two developmental hypotheses. The *separate layer hypothesis* states that the on and off cells are initially formed in two separate layers which later fuse to form the mature retina. The *single layer*

[3] $\lambda(x)$ essentially denotes the probability that there is a point in an infinitesimal interval around x, normalized by the volume of this interval; compare the conditional intensity in the last example.

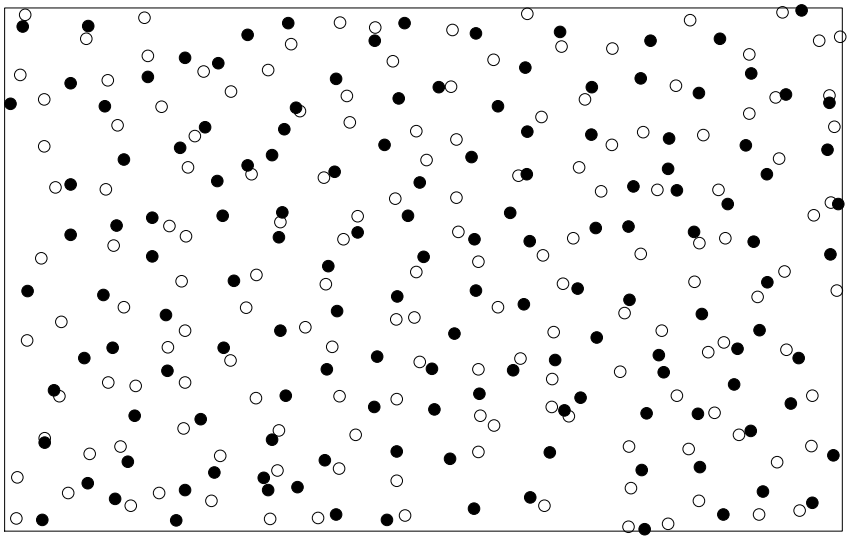

Figure 1.4: Amacrine cells of types on (transmits information when light goes on; empty circle) and off (transmits information when light goes off; filled circle) in the retina of a rabbit.

hypothesis states that the two types are initially undifferentiated in a single layer and only obtain their different functions at a later stage.

Tools for deciding which hypothesis is true include the bivariate (or cross-)K- and L-functions and parametric model fitting. At least for non-marked point patterns, we will study these tools in Chapters 3–5. Although some point pattern characteristics seem to indicate that the single layer hypothesis might hold, a more profound analysis gives strong evidence that in fact the separate layer hypothesis is true, i.e. that the white and black point patterns in Figure 1.4 are based on independent point process. It is uncontested, that each of them has a high degree of regularity within itself of course.

Chapter 2

Fundamental Point Process Theory

Point processes, which can be roughly described as random point patterns on some state space, are the mathematical objects that are at the centre of our attention.

2.1 Basics

As state space of our point processes, we always consider a topological space $(\mathcal{X}, \mathcal{T})$ that is Hausdorff, second countable, and locally compact. For the rest of these lecture notes we tacitly require these three properties whenever we consider a general state space. It can be shown that $\mathcal{X}$ is Polish, which means that it is separable and that there exists a metric d that induces $\mathcal{T}$ and makes $\mathcal{X}$ a complete metric space.

We call a set B (topologically) bounded if it is relatively compact, i.e. if its closure $\overline{B}$ is compact. It is easy to see that subsets of bounded sets are again bounded. It can be shown that any topologically bounded set is bounded in the metric d. The converse, however is not true in general, as can be seen from replacing d by the equivalent metric $\min(1, d)$, in which *every* set is metrically bounded. In $\mathbb{R}^d$ equipped with the usual Euclidean metric the two concepts coincide by the Heine–Borel theorem. In what follows, if used without further explanations, the word "bounded" will always mean "topologically bounded".

It can be shown that $\mathcal{X}$ contains always a sequence of open bounded sets $(B_i)_i$ such that $\bigcup_{i=1}^{\infty} B_i = \mathcal{X}$. By taking set differences of the B_i it can be seen that $\mathcal{X}$ can be *partitioned* into countably many bounded sets. By taking the closure of the B_i it follows that also $\mathcal{X} = \bigcup_{i=1}^{\infty} C_i$ for a sequence of *compact* sets C_i; hence $\mathcal{X}$ is σ-compact. By taking unions of the first i sets we obtain *increasing* sequences $(\widetilde{B}_i)$ and $(\widetilde{C}_i)$ with the above properties. By a simple compactness argument it follows that for every bounded set B there exists an $i \in \mathbb{N}$ such that $B \subset \widetilde{B}_i$.

We always equip $\mathcal{X}$ with its Borel-σ-algebra, which we denote by $\mathcal{B}$.

Denote by $\mathfrak{M}$ the set of all locally finite measures on $\mathcal{X}$, where a measure μ is called

locally finite if $\mu(B) < \infty$ for every bounded $B \in \mathcal{B}$. Denote by $\mathfrak{N}$ the set of all point measures on $\mathcal{X}$, where a *point measure* is a locally finite measure that takes only values in $\overline{\mathbb{Z}}_+ := \{0, 1, 2, \ldots\} \cup \{\infty\}$. The sets $\mathfrak{M}$ and $\mathfrak{N}$ are equiped with the σ-algebras $\mathcal{M}$ and $\mathcal{N}$, respectively, that are generated by the evaluation maps $\Phi_A : \mathfrak{M} \to \overline{\mathbb{R}}_+, \lambda \mapsto \lambda(A)$, for $A \in \mathcal{B}$, and $\Psi_A : \mathfrak{N} \to \overline{\mathbb{Z}}_+, \xi \mapsto \xi(A)$, for $A \in \mathcal{B}$, respectively. Hence for example $\mathcal{N} = \sigma\big(\{\xi \in \mathfrak{N} : \xi(A) \in R\} \,:\, A \in \mathcal{B}, R \subset \overline{\mathbb{Z}}_+\big)$.

Definition.

(a) An $\mathfrak{M}$-valued random element[1] Λ is called a *random measure* on $\mathcal{X}$.

(b) An $\mathfrak{N}$-valued random element Ξ is called a *point process* on $\mathcal{X}$.

Definition. Let $\mathfrak{N}^* := \{\xi \in \mathfrak{N} : \xi(\{x\}) \leq 1 \text{ for every } x \in \mathcal{X}\}$. A point process Ξ is called *simple* if $\mathbb{P}(\Xi \in \mathfrak{N}^*) = 1$. (Note that $\mathfrak{N}^* \in \mathcal{N}$; exercise!)

We denote the distribution of any random element Z either by $\mathscr{L}(Z)$ or by $\mathbb{P}Z^{-1}$ (since it is the image measure of the underlying probability measure $\mathbb{P}$ under Z, i.e. $\mathscr{L}(Z)(A) = \mathbb{P}(Z \in A) = \mathbb{P}(Z^{-1}(A)) =: (\mathbb{P}Z^{-1})(A)$ for every measurable A). In the remainder of this section, we formulate results for point measures and point processes only. The corresponding statements for general locally finite measures and random measures also hold with only a few obvious adaptations. We will need only the most basic random measure theory in the remainder of these lecture notes.

The σ-algebras $\mathcal{M}$ and $\mathcal{N}$ are chosen in such a way that all the "evaluations" $\Xi(B)$ and $\Lambda(B)$ for $B \in \mathcal{B}$ are random variables (with ∞ as a potential value if B is not bounded). The "evaluations" $\Xi(B)$ play in many respects a similar role as the "evaluations" of a stochastic process $(X(t))_t$ indexed by time. For instance the distribution of Ξ is completely determined by its finite dimensional distributions (short: fidi-distributions). The following results are formulated for point processes only, but analogous results hold for random measures.

Proposition 2.A. *Let Ξ and H ("capital Eta") be two point processes on $\mathcal{X}$ which have the same fidi-distributions for bounded and pairwise disjoint sets; that is,*

$$\mathbb{P}\big(\Xi(B_1) = k_1, \ldots, \Xi(B_r) = k_r\big) = \mathbb{P}\big(\mathrm{H}(B_1) = k_1, \ldots, \mathrm{H}(B_r) = k_r\big) \tag{2.1}$$

for all $r \in \mathbb{N}$, all bounded and pairwise disjoint $B_1, \ldots, B_r \in \mathcal{B}$, and all $k_1, \ldots, k_r \in \mathbb{Z}_+$.

Then Ξ and H have the same distribution.

Proof. We use a standard result about the uniqueness of measures: whenever we have two measures μ and ν on a common measurable space $(\mathcal{Y}, \mathcal{A})$ which coincide on a subsystem

[1] In this course, a *random element* is a measurable map from a probability space into a measurable space. The term *random variable* is used only for $\mathbb{R}$-valued random elements; sometimes the term *random vector* is used to denote an $\mathbb{R}^d$-valued random element.

$\mathcal{D} \subset \mathcal{A}$ that is closed under intersections and satisfies $\mathcal{Y} \in \mathcal{D}$ and $\sigma(\mathcal{D}) = \mathcal{A}$, then $\mu = \nu$ (see Kallenberg (2002), Lemma 1.17).

Define

$$\mathcal{D} := \Big\{ \{\xi \in \mathfrak{N}; \xi(B_1) = k_1, \ldots, \xi(B_r) = k_r\} : \\ r \in \mathbb{N},\, B_1, \ldots, B_r \in \mathcal{B} \text{ bounded},\, k_1, \ldots, k_r \in \mathbb{Z}_+ \Big\}.$$

It is clear that intersections of sets in $\mathcal{D}$ are again in $\mathcal{D}$. Furthermore $\mathfrak{N}$ is the element in $\mathcal{D}$ where $r = 1$, $B_1 = \emptyset$ and $k_1 = 0$.

Next we show that $\sigma(\mathcal{D}) = \mathcal{N}$. From the definition of $\mathcal{N}$, it can be verified that $\mathcal{N}$ is generated by the system

$$\mathcal{E} := \big\{ \{\xi \in \mathfrak{N} : \xi(B) = k\} \,:\, B \in \mathcal{B} \text{ bounded},\, k \in \mathbb{Z}_+ \big\},$$

because every set of the form $\Psi_A^{-1}(R) = \{\xi(A) \in R\}$ for general $A \in \mathcal{B}$ and $R \subset \overline{\mathbb{Z}}_+$ can be written as a countable union of sets in $\mathcal{E}$ (note by the σ-compactness of $\mathcal{X}$ that every $A \in \mathcal{B}$ can be written as a countable union of pairwise disjoint, bounded sets $B_i \in \mathcal{B}$). But then clearly $\mathcal{E} \subset \mathcal{D} \subset \sigma(\mathcal{E})$, hence $\sigma(\mathcal{D}) = \sigma(\mathcal{E}) = \mathcal{N}$.

It remains to show that $\mathbb{P}\Xi^{-1}(D) = \mathbb{P}\mathrm{H}^{-1}(D)$ for every $D \in \mathcal{D}$. Writing $B^1 := B$ and $B^0 := B^c$, we can see that for any $B_1, \ldots, B_r \in \mathcal{B}$ bounded

$$\begin{aligned} &\{\xi : \xi(B_1) = k_1, \ldots, \xi(B_r) = k_r\} \\ &\quad = \bigcup_{(l_{(e_1,\ldots,e_r)})} \big\{\xi : \xi\big(B_1^{e_1} \cap \ldots \cap B_r^{e_r}\big) = l_{(e_1,\ldots,e_r)} \text{ for all } (e_1, \ldots, e_r) \in \{0,1\}^r\big\}, \end{aligned} \tag{2.2}$$

where the union is taken over all finite "sequences" $(l_{(e_1,\ldots,e_r)})_{(e_1,\ldots,e_r)\in\{0,1\}^r}$ in $\mathbb{Z}_+$ that satisfy

$$\sum_{\substack{(e_1,\ldots,e_r)\in\{0,1\}^r \\ e_i = 1}} l_{(e_1,\ldots,e_r)} = k_i$$

for every $i \in \{1, \ldots, r\}$. Since in (2.2) we take the union of *disjoint* events, for which we know that $\mathbb{P}\Xi^{-1}$ and $\mathbb{P}\mathrm{H}^{-1}$ take the same value by Equation (2.1), it follows that $\mathbb{P}\Xi^{-1}(D) = \mathbb{P}\mathrm{H}^{-1}(D)$ for every $D \in \mathcal{D}$.

Altogether, by the result about uniqueness of measures quoted in the beginning, we have $\mathbb{P}\Xi^{-1} = \mathbb{P}\mathrm{H}^{-1}$. □

Rather surprisingly at first glance, the distribution of a simple point process is completely determined by the so-called void probabilities alone.

Proposition 2.B (Rényi–Mönch)**.** *Let* Ξ *and* H *be two simple point processes on* $\mathcal{X}$ *which have the same void probabilities for bounded sets; that is,*

$$\mathbb{P}\big(\Xi(B) = 0\big) = \mathbb{P}\big(\mathrm{H}(B) = 0\big) \tag{2.3}$$

for all bounded $B \in \mathcal{B}$.

Then Ξ *and* H *have the same distribution.*

An important concept for the proof that allows us to express statements about positions of points in terms of statements about point counts is the so-called dissecting system.

Definition. Let $A \in \mathcal{B}$. A sequence $\mathcal{S} = (\mathcal{S}_n)_n$ of finite partitions $\mathcal{S}_n = \{A_{ni} : 1 \leq i \leq m_n\}$ of A into Borel sets A_{ni} is called a *dissecting system for* A if

(a) $A_{ni} \cap A_{n+1,j} \in \{A_{n+1,j}, \emptyset\}$ for all $n \in \mathbb{N}$ and all i, j. *(Nesting property)*

(b) For all $x, y \in \mathcal{X}$ with $x \neq y$, there exists an $n \in \mathbb{N}$ and $i, j \in \{1, \ldots, m_n\}$ with $i \neq j$ such that $x \in A_{ni}$ and $y \in A_{nj}$. *(Point separation property)*

It can be shown that a dissecting system exists for any separable metric space (see Daley and Vere-Jones (2003), Proposition A2.1.IV), hence for our space $\mathcal{X}$, and then also for every set $A \in \mathcal{B}$.

Remark 2.C. *Based on the proof below it can be shown that in Proposition 2.B it is enough to require Equation* (2.3) *for all bounded* $B \in \mathcal{R}$*, where* $\mathcal{R}$ *is a dissecting ring*[2]*, i.e. the system of sets that are finite unions of elements of a dissecting system of* $\mathcal{X}$.

Proof of Proposition 2.B. We show that the void probabilities determine the fidi-distributions completely, i.e. we fix bounded and pairwise disjoint $B_1, \ldots, B_r \in \mathcal{B}$ and show that the left hand side of Equation (2.1) is expressible in terms of void probabilities $\mathbb{P}(\Xi(\widetilde{B}) = 0)$. The statement follows then from Proposition 2.A.

Let $\mathcal{S} = (\mathcal{S}_{nj})_n = \left((A_{n,j,i})_{1 \leq i \leq m_{nj}}\right)_n$ be a dissecting system for B_j. Define the random variable

$$Z_n(B_j) := \sum_{i=1}^{m_{nj}} 1\{\Xi(A_{n,j,i}) \geq 1\},$$

which gives the number of sets in $\mathcal{S}_{nj}$ that "contain points of Ξ".

We first show that $Z_n(B_j) \nearrow \Xi(B_j)$ almost surely as $n \to \infty$ for arbitrary j. For better readability, we suppress j in all notation for the duration of this argument. Since every A_{ni} is the union of sets from $\mathcal{S}_{n+1}$ and Ξ is $\overline{\mathbb{Z}}_+$-valued, we obtain that $Z_n(B)$ is increasing in n. Furthermore

$$Z_n(B) = \sum_{i=1}^{m_n} 1\{\Xi(A_{ni}) \geq 1\} \leq \sum_{i=1}^{m_n} \Xi(A_{ni}) = \Xi(B)$$

[2]The term ring refers to a ring of sets, i.e. any system of sets that is closed under finite set operations (intersections, unions, set differences). The classical example in $\mathbb{R}^d$ is the system of finite unions of bounded rectangles.

for every n, where strict inequality holds if and only if $\Xi(A_{ni}) \geq 2$ for some i. If it were the case that such an i existed for every $n \in \mathbb{N}$, there would be a decreasing (i.e. "nested") sequence $(A_{n,i_n})_n$ of sets with $\Xi(A_{n,i_n}) \geq 2$ for every n (decreasing, because otherwise there would be infinitely many disjoint sets in B with Ξ-mass ≥ 2, which contradicts the local finiteness of Ξ). By the point separation property it would follow that $A := \bigcap_{i\in\mathbb{N}} A_{n,i_n}$ contains at most one point. But then continuity of the measure from above would imply that

$$\Xi(A) = \lim_{n\to\infty} \Xi(A_{n,i_n}) \geq 2,$$

which either contradicts the fact that Ξ is a measure or the simplicity of Ξ. Therefore there exists an $n_0 \in \mathbb{N}$ such that $Z_n(B) = \Xi(B)$ for every $n \geq n_0$.

Re-introducing j into the notation, we have shown that $Z_n(B_j) \nearrow \Xi(B_j)$ for every j, almost surely. Since almost sure convergence implies convergence in distribution, it follows that

$$\lim_{n\to\infty} \mathbb{P}\big(Z_n(B_1) = k_1, \ldots, Z_n(B_r) = k_r\big) = \mathbb{P}\big(\Xi(B_1) = k_1, \ldots, \Xi(B_r) = k_r\big)$$

for any $k_1, \ldots, k_r \in \mathbb{Z}_+$.

We can now express the joint distribution of the $Z_n(B_j)$ (and hence of the $\Xi(B_j)$) in terms of void probabilities. Writing $C_{n,j;I_j} := B_j \setminus \bigcup_{i\in I_j} A_{n,j,i}$ and $\widetilde{C}_{n;I_1,\ldots,I_r} := \bigcup_{j=1}^r C_{n,j;I_j}$ for sets $I_j \subset \{1, \ldots, m_{nj}\}$, $1 \leq j \leq r$, we have

$$\begin{aligned}
&\mathbb{P}\big(Z_n(B_1) = k_1, \ldots, Z_n(B_r) = k_r\big) \\
&\quad = \sum_{I_1,\ldots,I_r} \mathbb{P}\Big(\forall j : \big(\Xi(C_{n,j;I_j}) = 0 \text{ and } \Xi(A_{n,j,i}) > 0 \text{ for every } i \in I_j\big)\Big) \\
&\quad = \sum_{I_1,\ldots,I_r} \Big[\mathbb{P}\big(\Xi(\widetilde{C}_{n;I_1,\ldots,I_r}) = 0\big) - \mathbb{P}\big(\textstyle\bigcup_{j=1}^r \bigcup_{i\in I_j} \big\{\Xi(A_{n,j,i} \cup \widetilde{C}_{n;I_1,\ldots,I_r}) = 0\big\}\big)\Big].
\end{aligned}$$

The sums in this formula are taken over all sets $I_1, \ldots, I_r$, where $I_j \subset \{1, \ldots, m_{nj}\}$ has exactly k_j elements, $j = 1, \ldots, r$. It follows from the Inclusion/Exclusion Formula that the probability in the very last line of the above equation is expressible in terms of void probabilites $\mathbb{P}(\Xi(A') = 0)$, where A' takes the form of unions of sets $A_{n,j,i} \cup \widetilde{C}_{n;I_1,\ldots,I_r}$ for $1 \leq j \leq r$ and $i \in I_j$. □

For a point process Ξ it is very natural to think of $\Xi(A)$ as the number of points of Ξ that lie in the set $A \in \mathcal{B}$. We have used this already in intuitive considerations. The following theorem makes this notion of a point process consisting of random points rigorous.

Proposition 2.D. *Let Ξ be a point process on $\mathcal{X}$. Then there are a $\overline{\mathbb{Z}}_+$-valued random variable M, $\mathbb{N}$-valued random variables $A_1, A_2, \ldots$, and $\mathcal{X}$-valued random elements*

$S_1, S_2, \ldots$ such that

$$\Xi = \sum_{i=1}^{M} A_i \delta_{S_i}.$$

Remark 2.E. *Conversely, it can be easily seen that, for corresponding M, A_i, and S_i, a random sum $\widetilde{\Xi} = \sum_{i=1}^{M} A_i \delta_{S_i}$ of random Dirac measures that is locally finite (for every ω) will always define a point process and the joint distribution of M, A_i, and S_i for $i \in \mathbb{N}$ determines the distribution of $\widetilde{\Xi}$ completely. Note on the other hand that the representation of a point process in Proposition 2.D as sum of Dirac measures is* by no means *unique.*

Proof of Proposition 2.D. We first show the existence of a deterministic representation if $\Xi = \xi \in \mathfrak{N}$ is non-random. The set $S := \{x \in \mathcal{X}; \xi(\{x\}) \geq 1\}$ is countable, because $\mathcal{X}$ can be partitioned into countably many bounded sets $\mathcal{X}_i$ and $S \cap \mathcal{X}_i$ is finite by the local finiteness of Ξ. Hence we may write $S = \{s_i : i \in I\}$ with the index set $I = \{1, \ldots, m\}$ for some $m \in \mathbb{N}$ or with $I = \mathbb{N}$, in which case we set $m := \infty$. Set furthermore $a_i := \xi(\{s_i\})$.

It is easily seen that $\tilde{\xi} := \xi - \sum_{i=1}^{m} a_i \delta_{s_i} \in \mathfrak{N}$. We would like to show that $\tilde{\xi}(\mathcal{X}) = 0$.

Denote by $\mathfrak{U}(x) := \{U \in \mathcal{T} : x \in U\}$ the so-called neighbourhood filter of x. It can be seen that every $x \in \mathcal{X}$ has a neighbourhood $U(x) \in \mathfrak{U}(x)$ such that $\tilde{\xi}(U(x)) = 0$. For suppose there was a $x \in \mathcal{X}$ such that $\tilde{\xi}(U) \geq 1$ for every $U \in \mathfrak{U}(x)$. It can be shown that in our state space $\mathcal{X}$ there must be a sequence of bounded sets $U_n \in \mathfrak{U}(x)$ such that $U_n \searrow \{x\}$ as $n \to \infty$ (this follows most easily by using our metric d and setting $U_n := \mathbb{B}(x, 1/n) \cap V$, where $V \in \mathfrak{U}(x)$ is a fixed bounded set and $\mathbb{B}(x, 1/n)$ denotes the open $1/n$-ball at x, but it can also be shown based on the topological properties of $\mathcal{X}$ directly). Therefore

$$\tilde{\xi}(\{x\}) = \tilde{\xi}\big(\textstyle\bigcap_{n\in\mathbb{N}} U_n\big) = \lim_{n\to\infty} \tilde{\xi}(U_n) \geq 1,$$

which contradicts the fact that we have removed all the integer atoms from our point measure ξ.

So every point x does have a neighbourhood $U(x)$ with $\tilde{\xi}(U(x)) = 0$. Since $\mathcal{X}$ is σ-compact, the open cover $\big(U(x)\big)_{x\in\mathcal{X}}$ must have a countable subcover $\big(U(x_i)\big)_{i\in\mathbb{N}}$. Therefore

$$\tilde{\xi}(\mathcal{X}) = \tilde{\xi}\big(\textstyle\bigcup_{i\in\mathbb{N}} U(x_i)\big) \leq \displaystyle\sum_{i\in\mathbb{N}} \tilde{\xi}\big(U(x_i)\big) = 0.$$

In total, we have shown that $\xi = \sum_{i=1}^{m} a_i \delta_{s_i}$. It can be easily seen that, assuming the s_i are pairwise distinct, this representation is unique up to the enumeration of the pairs (s_1, a_1), (s_2, a_2), ...

The existence of a random representation is more tricky to show. We only give a sketch and refer to Kallenberg (1986), Lemma 2.3, for the full proof.

By the deterministic argument above, we can define maps m, s_1, a_1, s_2, a_2, ... in ξ, but this involves choosing for every ξ an enumeration of its points, and if we do not choose the different enumerations carefully, they will not fit together to give measurable maps.

The key is to find an "enumeration rule" that gives a unique result for every ξ and can be reduced to counting points in measurable subsets of $\mathcal{X}$ to prove the measurability. Such a rule is given as follows. Let $D = \{z_1, z_2, \ldots\}$ be a countable dense subset of $\mathcal{X}$ and d our usual metric on $\mathcal{X}$. We obtain a total order on $\mathcal{X}$ by defining $x \prec y$ if there exists a $k \in \mathbb{N}$ such that $d(x, z_i) = d(y, z_i)$ for every $i < k$ and $d(x, z_k) < d(y, z_k)$. Then enumerate the points of any $\xi \in \mathfrak{N}$ according to this order, i.e. such that $s_1(\xi) \prec s_2(\xi) \prec \ldots$.

It can then be shown that for every $i \in \mathbb{N}$ and $B \in \mathcal{B}$, the sets $s_i^{-1}(B) = \{\xi : s_i(\xi) \in B\}$ can be written as countable unions and intersections of point count events.

The total point count $m = \xi(\mathcal{X})$ is obviously a measurable function of ξ. For the point masses a_i an additional argument is needed that involves lifting $\xi = \sum_{i=1}^m a_i \delta_{s_i}$ to the larger space $\mathcal{X} \times \mathbb{N}$ by transforming it into the simple point measure $\xi = \sum_{i=1}^m \delta_{(s_i, a_i)}$. □

We finish this section by giving two important examples of point processes.

Example 2.F (Binomial process)**.** Let $m \in \mathbb{N}$ and let ν be a probability distribution on $\mathcal{X}$. Let furthermore $S_1, \ldots, S_m$ be i.i.d. random elements with distribution ν. Then the point process

$$\Xi = \sum_{i=1}^m \delta_{S_i}$$

is called a *binomial process* on $\mathcal{X}$. We denote its distribution by $\mathrm{Bin}(m, \nu)$.

The name stems from the defininig sum that is in a sense the point process analog of a sum of Bernoulli random variables and also from the fact that the one-dimensional distributions $\mathscr{L}(\Xi(B))$ are binomial distributions. More generally, the fidi-distributions are given by

$$\mathbb{P}\big(\Xi(B_1) = k_1, \ldots, \Xi(B_r) = k_r\big) = \begin{cases} \frac{m!}{k_1! \cdots k_r!} \nu(B_1)^{k_1} \cdots \nu(B_r)^{k_r} & \text{if } \sum_{i=1}^r k_i = m, \\ 0 & \text{otherwise,} \end{cases}$$

where $r \in \mathbb{N}$, $(B_i)_{1 \le i \le r}$ is a finite partition of $\mathcal{X}$ into measurable sets, and $k_1, \ldots, k_r \in \mathbb{Z}_+$. Hence $\big(\Xi(B_1), \ldots, \Xi(B_r)\big)$ is multinomially distributed with size parameter m and probability vector $(\nu(B_1), \ldots, \nu(B_r))$. This can be seen by analogy with the classical model for the multinomial distribution, where n balls are distributed independently with some fixed probabilities on r urns (here: n points are distributed independently on r disjoint sets). ◇

Although the binomial process has one of the simplest representations of all point processes as a sum of Dirac measures, it is not suited as a model for a "completely random" point process, because its point counts $\Xi(B_1)$ and $\Xi(B_2)$ are negatively correlated,

whenever B_1 and B_2 are disjoint and have positive ν-mass. Elementary computations for the multinomial distribution show that

$$\operatorname{Cov}\big(\Xi(B_1), \Xi(B_2)\big) = -m\,\nu(B_1)\,\nu(B_2).$$

The desired independence property is obtained by replacing m by a Poisson-distributed random variable M.

Example 2.G (Finite Poisson process)**.** Let λ be a finite measure on $\mathcal{X}$. Let furthermore M be a Poisson-distributed random variable with parameter $|\lambda| := \lambda(\mathcal{X}) < \infty$ (we write $M \sim \mathrm{Po}(|\lambda|)$ for this) and let $S_1, S_2, \dots$ be i.i.d. random elements with distribution $\lambda/|\lambda|$ that are independent also of M. Then the point process

$$\Xi = \sum_{i=1}^{M} \delta_{S_i}$$

is called a *Poisson process* with parameter measure λ. We denote its distribution by $\mathrm{Po}(\lambda)$.

As for the binomial process the name corresponds to the one-dimensional distributions. More generally, the fidi-distributions are given by

$$\mathbb{P}\big(\Xi(B_1) = k_1, \dots, \Xi(B_r) = k_r\big) = \frac{\lambda(B_1)^{k_1}}{k_1} e^{-\lambda(B_1)} \cdots \frac{\lambda(B_r)^{k_r}}{k_r} e^{-\lambda(B_r)},$$

where $r \in \mathbb{N}$, $B_1, \dots, B_r \in \mathcal{B}$ are disjoint sets, and $k_1, \dots, k_r \in \mathbb{Z}_+$. Hence $\Xi(B_1), \dots, \Xi(B_r)$ are independent with $\Xi(B_i) \sim \mathrm{Po}(\lambda(B_i))$.

This can be seen by the following computation. Writing $B := \bigcup_{i=1}^{r} B_i$ and $k := \sum_{i=1}^{r} k_i$, we have

$$\begin{aligned}
&\mathbb{P}\big(\Xi(B_1) = k_1, \dots, \Xi(B_r) = k_r\big) \\
&\quad= \sum_{m=0}^{\infty} \mathbb{P}\big(\Xi(B_1) = k_1, \dots, \Xi(B_r) = k_r \,\big|\, \Xi(\mathcal{X}) = m\big)\, \mathbb{P}\big(\Xi(\mathcal{X}) = m\big) \\
&\quad= \sum_{m=k}^{\infty} \frac{m!}{k_1! \cdots k_r! (m-k)!} \left(\frac{\lambda(B_1)}{|\lambda|}\right)^{k_1} \cdots \left(\frac{\lambda(B_r)}{|\lambda|}\right)^{k_r} \left(\frac{\lambda(B^c)}{|\lambda|}\right)^{m-k} \frac{|\lambda|^m}{m!} e^{-|\lambda|} \\
&\quad= \frac{\lambda(B_1)^{k_1}}{k_1!} e^{-\lambda(B_1)} \cdots \frac{\lambda(B_r)^{k_r}}{k_r!} e^{-\lambda(B_r)} \sum_{m=k}^{\infty} \frac{\lambda(B^c)^{m-k}}{(m-k)!} e^{-\lambda(B^c)} \\
&\quad= \frac{\lambda(B_1)^{k_1}}{k_1!} e^{-\lambda(B_1)} \cdots \frac{\lambda(B_r)^{k_r}}{k_r!} e^{-\lambda(B_r)},
\end{aligned}$$

where the second equality is obtained by the fact that $\mathscr{L}(\Xi \,|\, M = m) = \mathscr{L}\big(\sum_{i=1}^{m} S_i\big)$ is the distribution of a $\mathrm{Bin}(m, \lambda/|\lambda|)$ process, as treated in the previous example. ◇

The fact that point counts on disjoint sets are independent makes the Poisson process nice to deal with from a theoretical point of view and makes it well suited as a null model of "complete spatial randomness". Poisson processes will be studied in more detail in the next section.

2.2 The Poisson process

In the last section we have already defined the finite Poisson process, which is usually only of importance if $\mathcal{X}$ is a compact space. However, we would sometimes like to consider Poisson processes with infinitely many points on non-compact spaces such as all of $\mathbb{R}^d$. By the independence property and the fact that our state space $\mathcal{X}$ can always be partitioned into countably many bounded sets B_i, we can construct a general Poisson process on $\mathcal{X}$ as a patchwork of finite Poisson processes on B_i. While this construction is given below, it is nicer to define a Poisson process via its fidi-distributions.

Definition (General Poisson process)**.** Let $\lambda \in \mathfrak{M}$. Then a point process H on $\mathcal{X}$ is called a *Poisson process* with parameter measure λ if

(a) $\mathrm{H}(B) \sim \mathrm{Po}(\lambda(B))$ for every bounded $B \in \mathcal{B}$;[3]

(b) $\eta(B_1), \dots, \eta(B_r)$ are independent for every $r \in \mathbb{N}$ and every selection of bounded and pairwise disjoint sets $B_1, \dots, B_r \in \mathcal{B}$.

The existence of such a process follows from the construction in the following proposition. The distribution of a Poisson process is uniquely determined by the above properties because of Proposition 2.A. In what follows denote by $\lambda|_B$ the measure on $\mathcal{X}$ that is given by $\lambda|_B(A) := \lambda(A \cap B)$ for every $A \in \mathcal{B}$. We will sometimes tacitly interprete $\lambda|_B$ as a measure on B (or even on an arbitrary superset of B).

Proposition 2.H. *Let $\lambda \in \mathfrak{M}$, and let $(\mathcal{X}_i)_{i\in\mathbb{N}}$ be a partition of $\mathcal{X}$ into bounded measurable sets so that every bounded B has non-empty intersection only with finitely many of the $\mathcal{X}_i$.*[4] *For each $i \in \mathbb{N}$, construct a finite $\mathrm{Po}(\lambda|_{\mathcal{X}_i})$-process H_i on $\mathcal{X}_i$ in such a way that H_i, $i \in \mathbb{N}$, are independent. Then H, defined by $\mathrm{H}(A) := \sum_{i\in\mathbb{N}} \mathrm{H}_i(A \cap \mathcal{X}_i)$ for every $A \in \mathcal{B}$, is a $\mathrm{Po}(\lambda)$-process on $\mathcal{X}$.*

Proof. It is easily checked that H is a point process. In particular all of its realizations are locally finite, because for every bounded set $B \in \mathcal{B}$ there are an $l \in \mathbb{N}$ and $i_1, \dots, i_l \in \mathbb{N}$ such that $B \cap \mathcal{X}_i = \emptyset$ for $i \notin \{i_1, \dots, i_l\}$ and hence $\mathrm{H}(B) = \sum_{j=1}^{l} \mathrm{H}_{i_j}(B \cap \mathcal{X}_{i_j}) < \infty$.

It remains to show that the two properties from the Poisson process definition are satisfied.

(a) Let $B \in \mathcal{B}$ bounded. For $i_1, \dots, i_l \in \mathbb{N}$ such that $B \cap \mathcal{X}_i = \emptyset$ if $i \notin \{i_1, \dots, i_l\}$, we have $\mathrm{H}(B) = \sum_{j=1}^{l} \mathrm{H}_{i_j}(B \cap \mathcal{X}_{i_j})$, where the $\mathrm{H}_{i_j}(B \cap \mathcal{X}_i)$ are independent and $\mathrm{Po}(\lambda(B \cap \mathcal{X}_{i_j}))$-distributed. Therefore $\mathrm{H}(B)$ is Poisson-distributed with parameter $\sum_{j=1}^{l} \lambda(B \cap \mathcal{X}_{i_j}) =$

[3] We set $\mathrm{Po}(0) := \delta_0$.

[4] Such a partition exists in our space $\mathcal{X}$: Compare the beginning of Chapter 2, where it was mentioned that there is an increasing sequence of bounded sets $\widetilde{B}_i$ with $\bigcup_i \widetilde{B}_i = \mathcal{X}$ such that every bounded set $B \in \mathcal{B}$ is contained in one of the $\widetilde{B}_i$s. The required partition is then given by $\mathcal{X}_i = \widetilde{B}_i \setminus \bigcup_{j=1}^{i-1} \widetilde{B}_j$.

$\lambda(B)$.

(b) Let $r \in \mathbb{N}$, and let $B_1, \dots, B_r \in \mathcal{B}$ be bounded disjoint sets. Then $\mathrm{H}(B_k) = \sum_{i=1}^{\infty} \mathrm{H}_i(B_k \cap \mathcal{X}_i)$. Since H_i are Poisson processes that are independent of one another, it follows that $\mathrm{H}_i(B_k \cap \mathcal{X}_i)$, $1 \le k \le r$, $i \in \mathbb{N}$, are all independent. Thus $\mathrm{H}(B_k)$, $1 \le k \le r$, are independent, because they are functions of disjoint selections from $\mathrm{H}_i(B_k \cap \mathcal{X}_i)$, $1 \le k \le r$, $i \in \mathbb{N}$. □

By mixing of Poisson processes with respect to the parameter measure, a much larger class of point processes can be obtained.

Definition. Let Λ be a random measure on $\mathcal{X}$. We call H a *Cox process* on $\mathcal{X}$ with *directing measure* Λ and write $\mathrm{H} \sim \mathrm{Cox}(\Lambda)$ if $\mathscr{L}(\mathrm{H} \,|\, \Lambda = \lambda) = \mathrm{Po}(\lambda)$ for $\mathbb{P}\Lambda^{-1}$-a.e. λ. The distribution of this process is completely determined by

$$\mathbb{P}(\mathrm{H} \in D) = \int_{\mathfrak{M}} \mathrm{Po}(\lambda)(D)\, \mathbb{P}\Lambda^{-1}(d\lambda)$$

for every $D \in \mathcal{N}$.

We were explicitly asking for the independence property (b) in the Poisson process definition, but not for the Poisson distribution of the point counts. It would therefore be interesting to know if the Poisson distribution is necessary. Conversely, we may ask ourselves if Poisson-distributed point counts are sufficient for the independence property. These questions are much easier to decide if we restrict ourselves to simple point processes.

We first give a lemma that shows that simplicity is easily determined if we know that a point process is Poisson.

Lemma 2.I. *A* $\mathrm{Po}(\lambda)$*-Process is simple if and only if its parameter measure* λ *does not contain any atoms, i.e.* $\lambda(\{x\}) = 0$ *for all* $x \in \mathcal{X}$.

Proof. Suppose that $\lambda(\{x\}) = \varepsilon > 0$ for some $x \in \mathcal{X}$. Then

$$\mathbb{P}(\mathrm{H} \notin \mathfrak{N}^*) \ge \mathbb{P}\big(\mathrm{H}(\{x\}) \ge 2\big) > 0,$$

because $\mathrm{H}(\{x\}) \sim \mathrm{Po}(\varepsilon)$. Thus H is not simple.

Conversely, suppose that $\lambda(\{x\}) = 0$ for every $x \in \mathcal{X}$. Choose a partition $(\mathcal{X}_i)$ of $\mathcal{X}$ into bounded measurable sets and note that

$$\mathbb{P}(\mathrm{H} \notin \mathfrak{N}^*) = \mathbb{P}\big(\textstyle\bigcup_{i\in\mathbb{N}} \{\mathrm{H}|_{\mathcal{X}_i} \notin \mathfrak{N}^*\}\big) \le \displaystyle\sum_{i\in\mathbb{N}} \mathbb{P}(\mathrm{H}|_{\mathcal{X}_i} \notin \mathfrak{N}^*).$$

Since $\mathrm{H}|_{\mathcal{X}_i}$ is a finite Poisson process on $\mathcal{X}$ whose expectation measure $\lambda|_{\mathcal{X}_i}$ has no atoms, it is enough to show that $\mathbb{P}(\mathrm{H} \notin \mathfrak{N}^*) = 0$ for processes H of the later kind.

Represent such an H as $\sum_{i=1}^{M} \delta_{S_i}$, where $M \sim \mathrm{Po}(|\lambda|)$ and $S_1, S_2, \ldots \sim \lambda/|\lambda|$ i.i.d. (independent also of M). We then have

$$\mathbb{P}(\mathrm{H} \notin \mathfrak{N}^*) \le \mathbb{P}\big(\exists i, j \in \mathbb{N}, i \ne j \colon S_i = S_j\big) \le \sum_{i \ne j} \mathbb{P}(S_i = S_j) = 0,$$

because by conditioning on S_j, using the independence of S_i and S_j and the fact that λ has no atoms, each of the summands $\mathbb{P}(S_i = S_j)$ is zero. □

We now decide on the questions whether for simple point processes the Poisson-distributed point counts are necessary and/or sufficient. The answer to the second question is an unqualified "yes".

Proposition 2.J. *Let Ξ be a simple point process on $\mathcal{X}$, and let $\lambda \in \mathfrak{M}$ be without atoms. If*

$$\mathbb{P}(\Xi(B) = 0) = e^{-\lambda(B)}$$

for every $B \in \mathcal{B}$ bounded (or for all bounded elements of a dissecting ring), then Ξ is a $\mathrm{Po}(\lambda)$-process. Note that we can leave away neither the condition that λ is a measure, nor that it is locally finite, nor that it is without atoms.

Proof. This is a direct consequence of the Rényi–Mönch theorem 2.B. □

There is clearly one restriction to the necessity of Poisson-distributed point counts. Namely, if H is a simple Poisson process, one can always add a point at a fixed location $x \in \mathcal{X}$ with a certain probability $p > 0$, independently of H. The result will *not* be a Poisson process, because its total number of points in a bounded set that contains x will be the sum of a Poisson and an independet Bernoulli random variable, which is never Poisson distributed.

Definition. We say a point process Ξ has a fixed atom at $x \in \mathcal{X}$, if $\mathbb{P}(\Xi(\{x\}) \ge 1) > 0$.

The following proposition says that no other restrictions are necessary.

Proposition 2.K. *Let Ξ be a simple point process on $\mathcal{X}$ without fixed atoms that satisfies the independence property (b) from the Poisson process definition. Then Ξ is a Poisson process.*

For the proof we make a certain detour, which allows us to plunge into the wonderful world of Poisson approximation and to pick up a general fact about point processes without fixed atoms in connection with dissecting systems.

Lemma 2.L (Generalized law of small numbers)**.** *For every $n \in \mathbb{N}$, let $I_{n1}, \dots, I_{n,m_n}$ be independent indicators with $p_{ni} := \mathbb{E}I_{ni} \in [0,1]$ such that*

$$\max_{1 \le i \le m_n} p_{ni} \longrightarrow 0 \quad \text{as } n \to \infty.$$

If

$$\lambda_n := \sum_{i=1}^{m_n} p_{ni} \longrightarrow \lambda \quad \text{as } n \to \infty$$

for some $\lambda \in [0, \infty]$, then

$$W_n := \sum_{i=1}^{m_n} I_{ni} \xrightarrow{\mathscr{D}} \mathrm{Po}(\lambda),$$

where $\mathrm{Po}(0) = \delta_0$ and $\mathrm{Po}(\infty) = \delta_\infty$.

Proof. We give a simple proof for the case $\lambda < \infty$. The case $\lambda = \infty$ is treated in Problem 3 (Exercise Sheet 2).

Let $Z_{ni} \sim \mathrm{Po}(p_{ni})$ and $J_{ni} \sim \mathrm{Be}\big((1-p_{ni})e^{p_{ni}}\big)$ all be independent for $1 \le i \le m_n$. Since the above statement is about convergence in distribution of the random variable W_n, we may construct this random variable in a particular way (as long as the distribution comes out correctly) and work with properties of the special construction. Let therefore $I_{ni} := 1 - J_{ni}\, 1\{Z_{ni} = 0\}$, $1 \le i \le m_n$, which obviously gives independent indicators with $\mathbb{P}(I_{ni} = 0) = \mathbb{P}(J_{ni} = 1)\mathbb{P}(Z_{ni} = 0) = 1 - p_{ni}$, and set $W_n := \sum_{i=1}^{m_n} I_{ni}$, which accordingly has the right distribution. Set furthermore $Z_n := \sum_{i=1}^{m_n} Z_{ni} \sim \mathrm{Po}(\lambda_n)$.

Now

$$\begin{aligned}
d_{\mathrm{TV}}\big(\mathscr{L}(W_n), \mathscr{L}(Z_n)\big) &:= \frac{1}{2} \sum_{k=0}^{\infty} \big|\mathbb{P}(W_n = k) - \mathbb{P}(Z_n = k)\big| \\
&\le \mathbb{P}(W_n \neq Z_n) \\
&\le \sum_{i=1}^{m_n} \mathbb{P}(I_{ni} \neq Z_{ni}) \\
&= \sum_{i=1}^{m_n} p_{ni}(1 - e^{-p_{ni}}) \\
&\le \sum_{i=1}^{m_n} p_{ni}^2 \\
&\le \lambda_n \max_{1 \le i \le m_n} p_{ni} \longrightarrow 0,
\end{aligned}$$

where the first inequality follows by a short computation (see Problem 2, Ex. Sheet 2). The total variation distance d_{TV} defined on the left hand side of the above inequality is in fact an important metric between probability distributions, but this need not concern us here and is mentioned simply for information.

Let then $Z \sim \mathrm{Po}(\lambda)$. We obtain the required statement, since for any $k \in \mathbb{Z}_+$

$$\big|\mathbb{P}(W_n = k) - \mathbb{P}(Z = k)\big| \le \big|\mathbb{P}(W_n = k) - \mathbb{P}(Z_n = k)\big| + \big|\mathbb{P}(Z_n = k) - \mathbb{P}(Z = k)\big| \longrightarrow 0,$$

where the second summand goes to zero, because $\lambda_n \to \lambda$. □

Remark. We have shown in the proof above that

$$d_{\mathrm{TV}}\big(\mathscr{L}(W_n), \mathscr{L}(Z_n)\big) \le \sum_{i=1}^{m_n} p_{ni}^2,$$

which from the practical point of view of Poisson approximation is much more useful than a mere convergence proof. However, the above bound is by no means the best one can do. Le Cam (1960) showed among other things that

$$d_{\mathrm{TV}}\big(\mathscr{L}(W_n), \mathscr{L}(Z_n)\big) \le 4.5 \max_{1\le i\le m_n} p_{ni},$$

which in particular means that Poisson approximation is good if the p_{ni} get uniformly small, even if λ_n gets very large. Modern estimates by the Stein-Chen method (initiated by Stein (1972); Chen (1975)) give

$$d_{\mathrm{TV}}\big(\mathscr{L}(W_n), \mathscr{L}(Z_n)\big) \le \min\Bigg(\sum_{i=1}^{m_n} p_{ni}^2, \max_{1\le i\le m_n} p_{ni}\Bigg).$$

The second lemma needed for the proof of Proposition 2.K says that for a point process without fixed atoms the probabilities that individual partition sets of a dissecting system contain any points go uniformly to zero.

Lemma 2.M. *Let Ξ be a point process on $\mathcal{X}$ without fixed atoms. Let furthermore $B \in \mathcal{B}$ be bounded and $\mathcal{S} = (\mathcal{S}_n) = \big((A_{ni})_{1\le i\le m_n}\big)$ be a dissecting system for B.*
Then $\max_{1\le i\le m_n} \mathbb{P}\big(\Xi(A_{ni}) \ge 1\big) \to 0$ as $n \to \infty$.

Proof. Suppose that the statement is not true. Let $\varepsilon > 0$ be such that for infinitely many $n \in \mathbb{N}$ we have $\max_i \mathbb{P}\big(\Xi(A_{ni}) \ge 1\big) \ge \varepsilon$. Passing over to a suitable subsequence of $(\mathcal{S}_n)$ we may assume without loss of generality that for *every* $n \in \mathbb{N}$ there exists an $i_n \in \{1, \dots, m_n\}$ with $\mathbb{P}\big(\Xi(A_{ni}) \ge 1\big) \ge \varepsilon$. We now argue in a similar way as for part of the Rényi–Mönch proof.

Since the partition $\mathcal{S}_1$ is finite, there is a $j_1 \in \{1, \dots, m_1\}$ so that A_{1,j_1} contains infinitely many of the sets A_{n,i_n}, $n \in \mathbb{N}$; and then inductively, if A_{k,j_k} contains infinitely many of the sets A_{n,i_n}, there must be a $j_{k+1} \in \{1, \dots, m_{k+1}\}$ so that $A_{k+1,j_{k+1}} \subset A_{k,j_k}$ also contains infinitely many of these sets (because $\mathcal{S}_{k+1}$ is finite). This yields a sequence $(j_n)_{n\in\mathbb{N}}$, where $1 \le j_n \le m_n$, $A_{n,j_n} \supset A_{n+1,j_{n+1}}$ and

$$\mathbb{P}\big(\Xi(A_{n,j_n}) \ge 1\big) \ge \varepsilon$$

for every $n \in \mathbb{N}$. By the point separation property of the dissecting system $A := \bigcap_{n\in\mathbb{N}} A_{n,j_n}$ contains at most one point. But since $\Xi(A_{n,j_n}) \searrow \Xi(A)$ by continuity of measures from above, we have $1\{\Xi(A_{n,j_n}) \geq 1\} \searrow 1\{\Xi(A) \geq 1\}$ and thus by the monotone convergence theorem convergence of the expections such that

$$\mathbb{P}\big(\Xi(A) \geq 1\big) \geq \varepsilon > 0.$$

This is impossible if A is empty, so $A = \{x\}$ for some $x \in \mathcal{X}$. But then the inequality states that Ξ has a fixed atom at x. □

Proof of Proposition 2.K. Let $\mathcal{S} = (\mathcal{S}_n)_n = \big((A_{ni})_{1\leq i\leq m_n}\big)_n$ be a dissecting system of $\mathcal{X}$. For every $A \in \mathcal{B}$, we obtain a dissecting system $\mathcal{S}^{(A)} = (\mathcal{S}_n^{(A)})_n = \big((A_{ni}^{(A)})_{1\leq i\leq m_n}\big)_n$ of A by setting $A_{ni}^{(A)} := A_{ni} \cap A$.

Let $Z_n(A) := \sum_{i=1}^{m_n} 1\{\Xi(A_{ni}^{(A)}) \geq 1\}$. In the proof of the Rényi–Mönch theorem, we have seen that $Z_n(A) \nearrow \Xi(A)$ if A is bounded, which can be easily extended to general $A \in \mathcal{B}$ by monotone convergence.

Therefore

$$\begin{aligned}
\lambda(A) &:= \lim_{n\to\infty} \sum_{i=1}^{m_n} \mathbb{P}\big(\Xi(A_{ni}^{(A)}) \geq 1\big) \\
&= \lim_{n\to\infty} \mathbb{E} Z_n(A) \\
&= \mathbb{E}\big(\lim_{n\to\infty} Z_n(A)\big) \\
&= \mathbb{E}\,\Xi(A) \in [0,\infty]
\end{aligned}$$

gives a well-defined measure λ on $\mathcal{X}$. The σ-additivity follows immediately from the σ-additivity of Ξ and Tonelli's theorem.

Now we have for bounded $B \in \mathcal{B}$ that $Z_n(B)$ converges to $\Xi(B)$ almost surely. But also $Z_n(B) = \sum_{i=1}^{m_n} I_{ni}$, where $I_{ni} := 1\{\Xi(A_{ni}^{(B)}) \geq 1\}$ with $\max_{1\leq i\leq m_n} \mathbb{E} I_{ni} \to 0$ by Lemma 2.M and $\sum_{i=1}^{m_n} \mathbb{E} I_{ni} \to \lambda(B)$. Lemma 2.L implies then that $Z_n(B) \xrightarrow{\mathscr{D}} \mathrm{Po}(\lambda(B))$.

Thus $\Xi(B) \sim \mathrm{Po}(\lambda(B))$ by the uniqueness of weak limits. Furthermore $\lambda(B) < \infty$, because otherwise $\Xi(B) = \infty$ a.s. would contradict the fact that Ξ is point process. This completes the proof our statement. □

A particularly easy (and "particularly random") Poisson process is obtained if $\mathcal{X} = \mathbb{R}^d$ and $\lambda = \ell\, Leb^d$. We call such a Poisson process *homogeneous* (or *stationary*, see Section 2.4) with intensity $\ell \geq 0$, and write $\mathrm{Po}_\ell := \mathrm{Po}(\ell Leb^d)$. This concept could be generalized to topological groups[5] $\mathcal{X}$, where the role of Lebesgue measure is taken over by the corresponding Haar measure.

[5]Our space $\mathcal{X}$ is a topological group if it is a group where the group operation $[(g,h) \mapsto gh]$ and the inversion map $[g \mapsto g^{-1}]$ are continuous, the former with respect to the product topology on $\mathcal{X} \times \mathcal{X}$.

In what follows we give various alternative constructions of Poisson processes on Euclidean spaces. The main difference compared to the standard construction given in Example 2.G (and extended in Proposition 2.H) is that the new constructions are point-based rather than region-based. By this we mean that instead of simulating the point process in a fixed bounded region B they generate points of the process sequentially (according to some canonical ordering) until a certain condition is met, e.g. the first m points are found or the set of locations in $\mathbb{R}^d$ that are closer to the origin than to any points of the process is determined. In the point-based case the shape and size of the "observation window" is typically random.

The homogeneous Poisson process on the real half-line

Let $\mathcal{X} = \mathbb{R}_+$. In this context one often refers to $\mathbb{R}_+$ as the time axis and to the points of a process as (time of) events. The homogeneous Poisson process $\mathbb{R}_+$ with intensity ℓ has been analyzed extensively in classical probability theory. It can be constructed by generating i.i.d. exponentially distributed random variables $T_1, T_2, \ldots$ with parameter ℓ (i.e. mean $1/\ell$). Setting $S_i := \sum_{j=1}^{i} T_j$ we obtain $\Xi := \sum_{i=1}^{\infty} \delta_{S_i}$ as the desired Poisson process on $\mathbb{R}_+$ with intensity ℓ.

The proof may be sketched as follows. Obviously the point process defined above is simple, because $\mathbb{P}(T_j = 0) = 0$, so by Proposition 2.J it is enough to show that

$$\mathbb{P}\big(\Xi(B) = 0\big) = e^{-\ell|B|} \tag{2.4}$$

for any $B \subset \mathbb{R}_+$ that is a finite union of pairwise disjoint intervals of the form $(a_1, a_1 + t_1]$, $\ldots$, $(a_r, a_r + t_r]$. For a single such interval with left endpoint in 0 this is easily seen, since

$$\mathbb{P}\big(\Xi((0,t]) = 0\big) = \mathbb{P}(T_1 > t) = e^{-\ell t}.$$

For the more general $B = \bigcup_{k=1}^{r}(a_k, a_k + t_k]$, we can use a famous property of the exponential distribution known as "memorylessness": if $T \sim \mathrm{Exp}(\ell)$, then it is easily seen that

$$\mathbb{P}(T > t + s \,|\, T > t) = \mathbb{P}(T > s)$$

for all $t, s \geq 0$. Using this and the independence of the T_j, it can be shown for any $a, t > 0$, $m \in \mathbb{N}$, $s_1, \ldots, s_m \in (0, a]$ with $s_1 < s_2 < \ldots < s_m$ that

$$\begin{aligned}
&\mathbb{P}\big(\Xi((a,a+t]) = 0 \;\big|\; \Xi|_{[0,a]} = \textstyle\sum_{i=1}^{m}\delta_{s_i}\big) \\
&\qquad = \mathbb{P}\big(T_{m+1} > (a - s_m) + t \;\big|\; T_1 = t_1, \ldots, T_m = t_m, T_{m+1} > a - s_m\big) \\
&\qquad = \mathbb{P}\big(T_{m+1} > (a - s_m) + t \;\big|\; T_{m+1} > a - s_m\big) \\
&\qquad = \mathbb{P}(T_{m+1} > t) = e^{-\ell t},
\end{aligned}$$

where $t_1 := s_1$ and $t_i := s_i - s_{i-1}$ for $i \geq 2$. From this it follows that Equation (2.4) holds for arbitrary intervals $B = (a, a+t]$ in $\mathbb{R}_+$ and then also for finite unions of such intervals.

The homogeneous Poisson process on the real line

Let $\mathcal{X} = \mathbb{R}$. Construct i.i.d. $\mathrm{Exp}(\ell)$-distributed random variables $T_1^+, T_1^-, T_2^+, T_2^-, \dots$ and set

$$S_i := \begin{cases} \sum_{j=1}^{i} T_i^+ & \text{if } i \geq 1; \\ -\sum_{j=1}^{-i} T_i^- & \text{if } i \leq -1. \end{cases}$$

Then $\Xi := \sum_{\mathbb{Z}\setminus\{0\}} \delta_{S_i}$ is a Po_ℓ-process on $\mathbb{R}$. In other words, the Poisson process on $\mathbb{R}$ is obtained by patching a Poisson process on $\mathbb{R}_+$ and on $\mathbb{R}_-$ together in the same way as we did for Proposition 2.H. The only difference is that we are using (finitely many) *general* Poisson processes as components, but the proof of conditions (a) and (b) from the definition is exactly the same.

This construction of a Poisson process on $\mathbb{R}$ leads to a paradoxical implication. If we look at the inter-point (or "inter-event") interval that contains zero, the length of this interval is the sum of two $\mathrm{Exp}(\ell)$ random variables and hence has a $\Gamma(2, \ell)$-distribution. However, "all other" inter-point intervals have a length that is $\mathrm{Exp}(\ell)$-distributed. Note in particular that we could do the same construction and come to the same conclusion for any fixed number τ instead of zero.

This phenomenon has been dubbed *inspection paradoxon*, essentially because the mere fact that we are *inspecting* the Poisson process at a certain time τ seems to change the distribution of its inter-event times. This is nonsense of course. What is true, however, is that if we are inspecting at a fixed time τ, then the distribution of the time we have to wait until we see the next event occur is $\mathrm{Exp}(\ell)$, and hence the same as if we had just seen an event occurring (for this reason the phenomenon is also known as *waiting time paradoxon*). In particular the expected time is $1/\ell$.

Many people would find this counterintuitive, arguing that the situations should be the same as if you picked an inspection time uniformly at random from an interval of (independent) $\mathrm{Exp}(\ell)$-distributed length, which would give an expected waiting time of $1/(2\ell)$. Intuition can be helped by considering a similar situation from the Poisson process' point of view: suppose the events are fixed and we pick an inspection time τ uniformly at random from a very large interval. Then it is clearly more likely that we land in a long inter-event interval than in a short one. So the expected wating time until the next event is (considerably) longer than if we just take a "typical" $\mathrm{Exp}(\ell)$-distributed interval.

The homogeneous Poisson process on $\mathbb{R}^d$

Let $\mathcal{X} = \mathbb{R}^d$. We construct a Poisson process radially from the origin as proposed in Quine and Watson (1984). Let $U_1, T_1, U_2, T_2, \dots$ be independent random elements, where the U_i are uniformly distributed on the unit sphere $\{x \in \mathbb{R}^d \colon \|x\| = 1\}$ and the T_i are $\mathrm{Exp}(\ell\alpha_d)$-valued, where $\alpha_d = \pi^{d/2}/\Gamma(d/2+1)$ denotes the volume of the unit ball in $\mathbb{R}^d$. Setting

further more $R_i := \big(\sum_{j=1}^{i} T_j\big)^{1/d}$, the process $\Xi := \sum_{i=1}^{\infty} \delta_{R_i U_i}$ is a Poisson process on $\mathbb{R}^d$ with intensity ℓ.

We only give a sketch of the proof. It seems a bit awkward that we have to check properties about point counts in general bounded sets $B \in \mathcal{B}$, where (closed) balls $\mathbb{B}(0,r)$ with centre 0 and radius $R > 0$ would be much more convenient to handle. In fact it can be shown quite easily that in order to satisfy the condition on the void probabilities from Proposition 2.J it suffices to show for every $r > 0$ that

(i) $\Xi\big(\mathbb{B}(0,r)\big) \sim \mathrm{Po}(\ell\alpha_d r^d)$;
(ii) Given $\Xi\big(\mathbb{B}(0,r)\big) = m$ the random vectors $R_1U_1, \ldots, R_mU_m$ are i.i.d., uniformly distributed on $\mathbb{B}(0,r)$.

Statement (i) is easily established via

$$\begin{aligned}\mathbb{P}\big(\Xi(\mathbb{B}(0,r)) = m\big) &= \mathbb{P}\big(\big(\textstyle\sum_{i=1}^{\infty}\delta_{R_i}\big)((0,r]) = m\big)\\ &= \mathbb{P}\big(\big(\textstyle\sum_{i=1}^{\infty}\delta_{R_i^d}\big)((0,r^d]) = m\big)\\ &= \mathrm{Po}(\ell\alpha_d r^d)(\{m\}),\end{aligned}$$

because by the construction of the homogeneous Poisson process on $\mathbb{R}_+$ from above $\sum_{i=1}^{\infty}\delta_{R_i^d}$ is a $\mathrm{Po}_{\ell\alpha_d}$-process.

For essentially the same reason the joint distribution of $(U_1, \ldots, U_m, R_1^d, \ldots, R_m^d)$ given $\Xi\big(\mathbb{B}(0,r)\big) = m$ is that of independent random variables $(\tilde{U}_1, \ldots, \tilde{U}_m, \tilde{R}_1^d, \ldots, \tilde{R}_m^d)$, where the $\tilde{U}_i$ are uniformly distributed on the unit sphere and the $\tilde{R}_i$ are uniformly distributed on $[0, r^d]$. Applying the transformation theorem to integrals over the corresponding density, we obtain that $\tilde{R}_1\tilde{U}_1$, $\tilde{R}_2\tilde{U}_2$, ... are i.i.d. uniform on $\mathbb{B}(0,r)$ as required for statement (ii).

2.3 Moment measures

Moments are useful quantities to summarize distributions of random variables. In the same way, moment measures are useful quantities to summarize distributions of point processes (and in fact also of random measures).

Definition. Let Ξ be a point process on $\mathcal{X}$ and $k \in \mathbb{N}$. Suppose that

$$\mu_k(B_1 \times \ldots \times B_k) := \mathbb{E}\big(\Xi(B_1)\cdots\Xi(B_k)\big) < \infty \qquad (*)$$

for all bounded $B_1, \ldots, B_k \in \mathcal{B}$. In this case we say that *the k-th moment measure exists.* The unique measure μ_k on $(\mathcal{X}^k, \mathcal{B}^k)$ defined via (*) is called the *"k-th moment measure"*. In particular we call $\mathbb{E}\Xi := \mu_1$ the expectation measure of Ξ.

Remark. Standard extension and uniqueness theorems from measure theory (see e.g. Kallenberg, 2002, Theorem 2.5 and Lemma 1.17 [6]) show that the above definition results indeed

[6] Formulated for finite measures; so we apply 1.17 to the measures restricted to the sets $\mathcal{X}_{i1} \times \ldots \times \mathcal{X}_{ik}$, $i \in \mathbb{N}$, of a partition of $\mathcal{X}^k$ into bounded rectangles.

in a unique measure μ_k on $\mathcal{X}^k$.

For $k \geq 2$ it is helpful to consider the k-fold product measure Ξ^k on $\mathcal{X}^k$, which can be represented for $\Xi = \sum_{i=1}^M \delta_{S_i}$ as

$$\Xi^k = \sum_{i_1,\ldots,i_k=1}^{M} \delta_{(S_{i_1},\ldots,S_{i_k})}.$$

It is clear by Tonelli's theorem that $\mathbb{E}(\Xi^k)$ is a measure that satisfies (*), and hence

$$\mu_k = \mathbb{E}(\Xi^k)$$

for every $k \in \mathbb{N}$.

The "raw" moment measures are usually not quite what one wants to consider. They have the aesthetic flaw that, for $k \geq 2$ and Ξ not almost surely zero, μ_k always assigns positive mass to any of the diagonal spaces $\Delta_{i_1,\ldots,i_l} := \{(x_1,\ldots,x_k) \in \mathcal{X}^k;\, x_{i_1} = \ldots = x_{i_l}\}$, where $l \geq 2$ and $i_1,\ldots,i_l \in \{1,2,\ldots,k\}$ are pairwise different. This rules out the possibility of having a density with respect to Lebesgue measure, which is often intuitively easier to understand than some general measure. We therefore define an alternative moment measure that has any artificial diagonal mass removed.

Definition. Let Ξ be a point process on $\mathcal{X}$ with existing k-th moment measure. For $m \in \mathbb{Z}_+$, write $m^{[k]} := m(m-1)\ldots(m-k+1)$. The unique measure $\mu_{[k]}$ on $(\mathcal{X}^k, \mathcal{B}^k)$ that is defined via

$$\mu_{[k]}(B_1^{k_1} \times \ldots \times B_r^{k_r}) := \mathbb{E}\big(\Xi(B_1)^{[k_1]} \cdots \Xi(B_r)^{[k_r]}\big) \qquad (**)$$

for $r \in \mathbb{N}$, $k_1,\ldots,k_r \in \mathbb{N}$ with $\sum_{i=1}^r k_i = k$, and $B_1,\ldots,B_r \in \mathcal{B}$ bounded and pairwise disjoint, is called the *k-th factorial moment measure* of Ξ.

Remark. It is easily seen that every bounded rectangle can be written as a finite union of pairwise disjoint rectangles $B_{j1}^{k_{j1}} \times \ldots \times B_{jr_j}^{k_{jr_j}}$ of the above form, so that by additivity Equation (**) determines μ_k for all bounded rectangles (without destroying σ-additivity). Existence and uniqueness follow then in the same way as for the "raw" moment measure.

Again it is helpful (already for σ-additivity) to consider the corresponding product measure of the point process. Let $\Xi^{[k]}$ be the factorial product measure of Ξ, which is most easily defined, for $\Xi = \sum_{i=1}^M \delta_{S_i}$, as

$$\Xi^{[k]} := \sum_{i_1,\ldots,i_k=1}^{M,\neq} \delta_{(S_{i_1},\ldots,S_{i_k})},$$

where the symbol $\neq$ on top of the summation sign indicates that the sum is taken over all $i_1,\ldots,i_k \in \{1,\ldots,M\}$ that are pairwise different. Then $\mathbb{E}(\Xi^{[k]})$ is a measure, and it is easily seen that it satisfies (**), so that

$$\mu_{[k]} = \mathbb{E}(\Xi^{[k]})$$

for every $k \in \mathbb{N}$.

Note that $\mu_{[k]}$, $k \geq 2$, still has positive mass on certain diagonal spaces if Ξ is non-simple.

Example 2.N (Moment measures of Poisson processes)**.** Let H be a Po(λ)-process on $\mathcal{X}$ with parameter measure λ. Furthermore let $\Delta := \{(x_1, x_2) \in \mathcal{X}^2 \colon x_1 = x_2\}$ and $\varphi_\Delta \colon \mathcal{X} \to \mathcal{X}^2, x \mapsto (x, x)$, and denote by $\lambda_\Delta := \lambda\varphi_\Delta^{-1}$ the "diagonal measure" on $\mathcal{X}^2$. Note that this means that $\lambda_\Delta(A_1 \times A_2) = \lambda(A_1 \cap A_2)$ for all $A_1, A_2 \in \mathcal{B}$.

We then obtain

(i) $\mu_1 = \lambda$;

(ii) $\mu_2 = \lambda^2 + \lambda_\Delta$;

(iii) $\mu_{[k]} = \lambda^k$ for every $k \in \mathbb{N}$.

(i) is immediately clear because $\mathrm{H}(B) \sim \mathrm{Po}(\lambda(B))$ for every bounded $B \in \mathcal{B}$.

We next show statement (iii). Let $Z \sim \mathrm{Po}(\ell)$. Then, using that $m^{[k]} = 0$ for $m < k$,

$$\mathbb{E}\big(Z^{[k]}\big) = \sum_{m=0}^{\infty} m^{[k]} \frac{\ell^m}{m!} e^{-\ell} = \ell^k \sum_{m=k}^{\infty} \frac{\ell^{m-k}}{(m-k)!} e^{-\ell} = \ell^k.$$

By this and the independence property of the Poisson process we have for all $k_1, \ldots, k_r \in \mathbb{N}$ with $\sum_{i=1}^r k_i = k$ and $B_1, \ldots, B_r \in \mathcal{B}$ bounded and pairwise disjoint that

$$\mathbb{E}\big(\mathrm{H}(B_1)^{[k_1]} \cdots \mathrm{H}(B_r)^{[k_r]}\big) = \mathbb{E}\big(\mathrm{H}(B_1)^{[k_1]}\big) \cdots \mathbb{E}\big(\mathrm{H}(B_r)^{[k_r]}\big) = \lambda(B_1)^{k_1} \cdots \lambda(B_r)^{k_r}.$$

Hence $\mu_k = \lambda^k$.

Statement (ii) follows in the same way by noting that

$$\mathbb{E}\big(\mathrm{H}(B_i)^2\big) = \mathbb{E}\big(\mathrm{H}(B_i)(\mathrm{H}(B_i) - 1)\big) + \mathbb{E}\mathrm{H}(B_i) = \lambda(B_i)^2 + \lambda(B_i) = (\lambda^2 + \lambda_\Delta)(B_i \times B_i).$$

We could also derive $\mu_{[k]}$ for general k in a similar way and would obtain that we have to add to λ_k for each subspace $\Delta_{i_1,\ldots,i_l}$ a "diagonal measure" $\lambda_{\Delta_{i_1,\ldots,i_l}}$ on $\mathcal{X}^k$ that is defined in a similar way as λ_Δ above (loosely said, it is λ^{k-l+1} "projected" onto $\Delta_{i_1,\ldots,i_l}$). Comparing μ_k to $\mu_{[k]}$ in the Poisson process case, it becomes very understandable that we usually do not want to trouble ourselves with raw moment measures. $\diamond$

Many useful functionals of a point process $\Xi = \sum_{i=1}^M \delta_{S_i}$ in spatial statistics are of the form $F(\Xi) = \int f(x)\,\Xi(dx) = \sum_{i=1}^M f(S_i)$. The following two results allow us to compute moments of such functionals based on moment measures of Ξ.

Proposition 2.O (Campbell's Formula)**.** *Let Ξ be a point process on $\mathcal{X}$ with existing expectation measure and $f \colon \mathcal{X} \to \mathbb{R}_+$ measurable. Then*

$$\mathbb{E}\bigg(\int f(x)\,\Xi(dx)\bigg) = \int f(x)\,(\mathbb{E}\Xi)(dx) \;\in \overline{\mathbb{R}}_+.$$

Proof. The proof follows from a standard extension argument. First, let $f := 1_A$ be an indicator function with $A \in \mathcal{B}$. Then the statement is just the definition of the expectation measure. The statement holds for simple functions, that is functions of the form $\sum_{i=1}^n a_i 1_{A_i}$ with $a_i \geq 0$ and $A_i \in \mathcal{B}$ by the linearity of expectation and integral. We then use the fact that if f is a non-negative measurable function, it can be approximated by an increasing sequence (f_n) of non-negative simple functions (see e.g. Kallenberg, 2002, Lemma 1.11). Using this approximation we have

$$\mathbb{E}\left(\int f(x)\,\Xi(dx)\right) = \lim_{n\to\infty} \mathbb{E}\left(\int f_n(x)\,\Xi(dx)\right) = \lim_{n\to\infty} \int f_n(x)\,\mu_1(dx) = \int f(x)\,\mu_1(dx).$$

by several applications of Levi's monotone convergence theorem. □

Applying Campbell's Formula to Ξ^k, we obtain more generally:

Corollary 2.P. *Let Ξ be a point process on $\mathcal{X}$ with existing k-th moment measure and $f\colon \mathcal{X} \to \mathbb{R}_+$ measurable. Then*

$$\mathbb{E}\left(\int f(x)\,\Xi(dx)\right)^k = \int f(x_1)\cdots f(x_k)\,\mu_k(dx_1\,\ldots\,dx_k) \;\in\; \overline{\mathbb{R}}_+.$$

□

2.4 Stationarity and Isotropy

Stationarity and isotropy are important properties that simplify estimation and inference for point processes. In their strongest forms these properties simply say that the distribution of a point process does not change if all of its realizations are subjected to a common translation (stationarity) or rotation (isotropy). In general we may consider an arbitrary group action on $\mathcal{X}$.

Definition (Invariance)**.** Let $\star$ be an action[7] of a group $\mathcal{G}$ on $\mathcal{X}$ such that $\theta_g\colon \mathcal{X} \to \mathcal{X}$, $\theta_g(x) := g \star x$, is a measurable map for every $g \in \mathcal{G}$. It is known from group theory (and very easy to see) that $\Theta := \{\theta_g\colon g \in \mathcal{G}\}$ with composition of functions as operation is again a group, where $\theta_g^{-1} = \theta_{g^{-1}}$. Define then the maps $T_g\colon \mathfrak{N} \to \mathfrak{N}, T_g(\xi) := \xi\theta_g^{-1} = \sum_{i=1}^m \delta_{\theta_g(s_i)}$ for $\xi = \sum_{i=1}^m \delta_{s_i} \in \mathfrak{N}$.
Committing a minor naming crime we say that (the distribution of) a point process Ξ is $(G,\star)$-*invariant* or Θ-*invariant* if $\mathscr{L}(T_g(\Xi)) = \mathscr{L}(\Xi)$ for every $g \in \mathcal{G}$.

Invariances can be studied in various contexts and many general statements do not depend on the concrete group $\mathcal{G}$ that is considered. However, for spatial statistics the most

[7] i.e. $\star\colon \mathcal{G}\times\mathcal{X} \to \mathcal{X}, (g,x) \mapsto g\star x$ such that for all $g,h \in \mathcal{G}$ and $x \in \mathcal{X}$ the properties $(gh)\star x = g\star(h\star x)$ and $e \star x = x$ hold, where e is the neutral element of $\mathcal{G}$.

important types of invariances are the ones with respect to translations and rotations on $\mathbb{R}^d$.

Definition. A point process Ξ on $\mathcal{X} = \mathbb{R}^d$ is called

(a) *stationary* if it is invariant with respect to the group of translations on $\mathbb{R}^d$ (induced by $\mathbb{R}^d$ operating on $\mathcal{X}$ by addition);

(b) *isotropic* if it is invariant with respect to the group of rotations on $\mathbb{R}^d$ (induced by $\mathrm{SO}(d)$ operating on $\mathcal{X}$ by matrix multiplication).

From here on we use Θ only for the group of translations $\theta_a : \mathbb{R}^d \to \mathbb{R}^d, \theta_a(x) := x + a$ and write Ψ for the group of rotations $\psi_A : \mathbb{R}^d \to \mathbb{R}^d, \psi_A(x) := Ax$. For the time being we concentrate on stationarity. For many applied purposes a weaker concept suffices.

Definition. A point process Ξ on $\mathcal{X} = \mathbb{R}^d$ with existing k-th moment measure μ_k is called *k-th order stationary* if μ_j for $j = 1, \ldots, k$ is invariant under diagonal shifts, i.e.

$$\mu_j\big(\theta_a^{-1}(B_1) \times \cdots \times \theta_a^{-1}(B_j)\big) = \mu_j\big(B_1 \times \cdots \times B_j\big)$$

for all bounded $B_1, \ldots, B_j \in \mathcal{B}$ and all $a \in \mathbb{R}^d$.

Remark. The following are straightforward consequences from the definition, the proof of which is left to the reader.

(i) A stationary point process Ξ whose k-th moment measure exists is k-th order stationary for any k. For this reason stationary point processes are sometimes called *strongly stationary* for better distinction.

(ii) The k-th order stationarity property implies that

$$\mu_{[k]}\big(\theta_a^{-1}(B_1) \times \cdots \times \theta_a^{-1}(B_k)\big) = \mu_{[k]}\big(B_1 \times \cdots \times B_k\big)$$

for all bounded $B_1, \ldots, B_k \in \mathcal{B}$ and all $a \in \mathbb{R}^d$.

Intuitively k-th order stationarity means that μ_k and $\mu_{[k]}$ do not contain any information "in the direction of the main diagonal $\Delta := \{(x_1, \ldots, x_k) \colon x_1 = x_2 = \ldots = x_k\}$". This can be used to define "reduced" moment measures $\check{\mu}_k$ and $\check{\mu}_{[k]}$ with one component $\in \mathbb{R}^d$ removed. In the proposition below we only consider the very important special case of the reduced second factorial moment measure $\mathcal{K} := \check{\mu}_{[2]}$, called the *$\mathcal{K}$-measure*. For a point process on $\mathbb{R}^2$ this measure lives on $\mathbb{R}^2$ as well, so that we can draw a contour plot or a heat map of an estimate of its density (if available), which usually gives a very good idea of the second order properties of the underlying point process.

Proposition 2.Q. *Let Ξ be a point process on $\mathbb{R}^d$.*

(i) If Ξ is first order stationary, then its expectation measure is of the form $\mu_1 = m_1 \mathrm{Leb}^d$ for some $m_1 \in \mathbb{R}_+$.

(ii) Suppose that $\mathbb{P}(\Xi(\mathbb{R}^d) = 0) < 1$. *If* Ξ *is second order stationary, there exists a unique measure* $\mathcal{K}$ *on* $\mathbb{R}^d$ *such that*

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} h(x,y)\,\mu_{[2]}\big(d(x,y)\big) = m_1^2 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h(x,x+y)\,\mathcal{K}(dy)\,dx$$

for every measurable function $h : \mathcal{X}^2 \to \mathbb{R}_+$.

Proof. (i) This follows from a well-known theorem in measure theory, which says that Lebesgue measure is up to a factor the only translation invariant, locally finite measure on $\mathbb{R}^d$ (see e.g. Amann and Escher, 2001, Theorem 5.19).

(ii) Define $\varphi\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d, (x,y) \mapsto (x, y-x)$, which is continuous (hence measurable) and bijective. The transformation theorem for integrals yields that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} h(x,y)\,\mu_{[2]}\big(d(x,y)\big) = \int_{\mathbb{R}^d\times\mathbb{R}^d} h(x,x+y)\,\mu_{[2]}\varphi^{-1}\big(d(x,y)\big). \tag{2.5}$$

We first show that

$$\mu_{[2]}\varphi^{-1}\big(\theta_a^{-1}(A_1)\times A_2\big) = \mu_{[2]}\varphi^{-1}\big(A_1\times A_2\big)$$

for every $a \in \mathbb{R}^d$ and all $A_1, A_2 \in \mathcal{B}^d$. Writing $D_a\colon \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}^d\times\mathbb{R}^d, (x,y) \mapsto (\theta_a(x), \theta_a(y))$ for the diagonal a-shift, we obtain

$$\begin{aligned}\varphi^{-1}\big(\theta_a^{-1}(A_1)\times A_2\big) &= \{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d\colon x\in\theta_a^{-1}(A_1), y-x\in A_2\}\\ &= D_a^{-1}\big(\{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d\colon x\in A_1, y-x\in A_2\}\big)\\ &= D_a^{-1}\big(\varphi^{-1}(A_1\times A_2)\big).\end{aligned}$$

Hence by the invariance of $\mu_{[2]}$ under diagonal shifts, which extends from rectangles to arbitrary measurable sets,

$$\mu_{[2]}\varphi^{-1}\big(\theta_a^{-1}(A_1)\times A_2\big) = \mu_{[2]}\big(D_a^{-1}\big(\varphi^{-1}(A_1\times A_2)\big)\big) = \mu_{[2]}\varphi^{-1}\big(A_1\times A_2\big).$$

We next show that every locally finite measure ν on $\mathbb{R}^d\times\mathbb{R}^d$ with the property $\nu\big(\theta_a^{-1}(A_1)\times A_2\big) = \nu\big(A_1\times A_2\big)$ is of the form $Leb^d \otimes \check{\nu}$ for some measure $\check{\nu}$ on $\mathbb{R}^d$. For fixed bounded $B\in\mathcal{B}^d$ it is easily seen that

$$\nu_B(A) := \nu(A\times B)$$

defines a locally finite measure on $(\mathbb{R}^d, \mathcal{B}^d)$ with

$$\nu_B(\theta_a^{-1}A) = \nu_B(A)$$

for every $A\in\mathcal{B}^d$. By the theorem quoted in part (i) there is exists an $m_B \geq 0$ such that

$$\nu(A\times B) = \nu_B(A) = m_B Leb^d(A)$$

for every $A, B \in \mathcal{B}^d$ where B is bounded. Since $\nu([0,1]^d \times \cdot)$ is a measure, we can define the measure $\check{\nu}$ on $\mathbb{R}^d$ by $\check{\nu}(B) := m_B$ for bounded $B \in \mathcal{B}^d$. It follows that $\nu = Leb^d \otimes \check{\nu}$.

Setting $\nu := \mu_{[2]}\varphi^{-1}$ and $\mathcal{K} := \frac{1}{m_1^2}\check{\nu}$, the required integral equation follows from Equation (2.5). Choosing in the integral equation $h := 1_{\varphi^{-1}([0,1]^d \times A)}$ for arbitrary $A \in \mathcal{B}^d$, it follows that

$$\mathcal{K}(A) = (Leb^d \otimes \mathcal{K})\big([0,1]^d \times A\big) = \mu_{[2]}\big(\varphi^{-1}\big([0,1]^d \times A\big)\big)/m_1^2. \tag{2.6}$$

Hence the measure $\mathcal{K}$ is uniquely determined. □

Remark. (i) With the help of Proposition 2.Q(i) it is clear that a Poisson process is first order stationary if and only if it is homogeneous. For any such process H it is easily seen that $\mathrm{H}\theta_a^{-1}$ satisfies again the defining properties of a Po_ℓ-process. Thus the concepts of homogeneity, k-th order stationarity, and strong stationarity are equivalent for Poisson processes.

(ii) By the fact that φ^{-1} is a linear transform on $\mathbb{R}^d \times \mathbb{R}^d$ with determinant 1, we obtain that $(Leb^d \otimes Leb^d)\varphi^{-1} = Leb^d \otimes Leb^d$. Thus it follows from Example 2.N(iii) and Equation (2.6) that the $\mathcal{K}$-measure of the Po_ℓ-process is Leb^d.

(iii) For certain authors the $\mathcal{K}$-measure is defined as m_1 times our $\mathcal{K}$-measure.

We finish this section by giving an example of useful point processes that live on a more complicated space than $\mathbb{R}^d$. This leads us into the realms of stochastic geometry. Since a detailed account of stochastic geometry is beyond the scope of these lecture notes, many of the statements below are without proof and the interested reader is referred to Schneider and Weil (2008).

Example 2.R (Point processes of closed sets)**.** Let $\mathcal{F}$ be the system of closed subsets of $\mathbb{R}^d$ equipped with the topology $\mathcal{T}_\mathcal{F}$ generated by

$$\{\mathcal{F}_G \colon G \subset \mathbb{R}^d \text{ open}\} \cup \{\mathcal{F}^C \colon C \subset \mathbb{R}^d \text{ compact}\}$$

where $\mathcal{F}_G := \{F \in \mathcal{F} \colon F \cap G \neq \emptyset\}$ is the system of sets "hitting" a given set G and $\mathcal{F}^C := \{F \in \mathcal{F} \colon F \cap C = \emptyset\}$ is the system of sets missing a given set C. Topologies constructed like this are known as "hit-and-miss" topologies. The particular topology $\mathcal{T}_\mathcal{F}$ is known as the *Fell topology* (also: topology of closed convergence).

The topological subspace $\mathcal{X} := \mathcal{F}' := \mathcal{F} \setminus \{\emptyset\}$ can be shown to be a locally compact second countable Hausdorff space, so that we may consider point processes on this space.[8] A point process Ξ on $\mathcal{F}'$ is called *stationary* if it is invariant with respect to the group of set translations $\{\theta_a \colon a \in \mathbb{R}^d\}$, where $\theta_a \colon \mathcal{F} \to \mathcal{F}$, $F \mapsto a + F := \{a + x \colon x \in F\}$.

[8] The space $\mathcal{F}$ has the same properties and is even compact. However, this is rather a disadvantage as it restricts our point processes to having only finitely many points (by local finiteness).

A rather nice class of such point processes is formed by the so-called particle processes.

Definition. A point process Ξ on $\mathcal{F}'$ is called a *particle process* if it is concentrated on the subset $\mathcal{C}' := \{C \in \mathcal{F}' \colon C \text{ is compact}\}$ and its expectation measure μ_1 exists.

It can be shown (this is essentially Lemma 2.3.1 in Schneider and Weil, 2008) that the existence of μ_1 is equivalent to the fact that the expected number of "points" Z_i of Ξ that have non-empty intersection with a fixed set C is finite for every compact $C \subset \mathbb{R}^d$.

Particle processes are helpful for understanding random closed sets (RACS), i.e. random elements of $(\mathcal{F}, \sigma(\mathcal{T}_{\mathcal{F}}))$. It can be shown that every RACS can be written as the union set $Z_\Xi := \bigcup_{i=1}^M Z_i$ of a simple particle process $\Xi = \sum_{i=1}^M \delta_{Z_i}$ (Schneider and Weil, 2008, Theorem 4.3.1).

It may be shown further that the class of union sets of stationary Poisson particle processes coincides with the class of so-called stationary Boolean models (Schneider and Weil, 2008, Theorem 4.3.1). The latter are RACS of the form $Z = \bigcup_{i=1}^M (S_i + Z_i)$, where $\mathrm{H} = \sum_{i=1}^M \delta_{S_i}$ is a Po_ℓ-process on $\mathbb{R}^d$ and $Z_1, Z_2, \ldots$ are i.i.d. random compact sets in $\mathcal{C}'$ that are independent also of H and satisfy $\mathbb{E}\big(Leb^d(Z_1 + B(0,r))\big) < \infty$ for some $r > 0$.

Due to this characterization result it is easy to compute the capacity functional T of a stationary Boolean model Z, which gives the hitting probability for any compact set C and is an important distributional descriptor for RACS; namely

$$T(C) := \mathbb{P}(Z \cap C \neq \emptyset) = \mathbb{P}(\Xi(\mathcal{F}_C) \geq 1) = 1 - \exp(-\mu_1(\mathcal{F}_C)).$$

Using more detailed knowledge about the duality between stationary Boolean models and Poisson particle processes, this may be further evaluated to yield

$$T(C) = 1 - \exp\big(-\ell\, \mathbb{E}(Leb^d(Z_1^* + C))\big)$$

for every compact $C \subset \mathbb{R}^d$, where $Z_1^* := \{-x \colon x \in Z_1\}$. ◇

2.5 Conditioned Point Processes

In this section we study conditional distributions of a point process Ξ. Loosely speaking, we mean by this distributions of the form $\mathscr{L}\big(\Xi|_A \,\big|\, \Xi|_B\big)$ for disjoint sets A and B. Intuitively the most interesting situations occur if either A or B degenerate to a one-point set (or more generally a finite set). However, it is then without further theory mathematically no longer clear how the corresponding conditional distributions can be defined. We consider

- **Interior conditioning:** Conditioning on the presence of a point of the process at a fixed location $x \in \mathcal{X}$, leading to a precise definition for "$\mathscr{L}\big(\Xi \,\big|\, \Xi(\{x\}) = 1\big)$". The

key object in this context is known as a Palm kernel and the necessary methods are known as Palm theory.

- **Exterior conditioning:** Formalization of the "pressure" the process Ξ generates around a fixed location $x \in \mathcal{X}$ with regard to the existence of a point at this location, leading to a precise definition of "$\mathscr{L}\big(\Xi(\{x\}) \;\big|\; \Xi\,|_{\mathcal{X}\setminus\{x\}} = \xi\big)$". The key objects are known as Papangelou kernel and conditional intensity.

In fact the above intuitions are usually only valid if our point processes are simple. The non-simple case causes various complications. In what follows we sometimes restrict ourselves to simple point processes for this reason.

2.5.1 Palm theory

Recall the general definition of conditional probabilities of the form $\mathbb{P}(Y \in B \,|\, X = x)$ for a $(\mathcal{X}, \mathcal{A}_\mathcal{X})$-valued random variable X and a $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$-valued random variable Y. $\mathbb{P}(Y \in B \,|\, X = \cdot)$ is defined as the $\mathbb{P}X^{-1}$-a.s. unique measurable map $\psi_B \colon \mathcal{X} \to \mathbb{R}_+$ that satisfies

$$\int_A \psi_B(x)\, \mathbb{P}X^{-1} = \mathbb{P}(X \in A, Y \in B) \quad \text{for all } A \in \mathcal{A}_\mathcal{X},\ B \in \mathcal{A}_\mathcal{Y}.$$

If the space $\mathcal{Y}$ is nice (e.g. a complete separable metric space with Borel σ-algebra), then there is a so-called regular conditional probability, i.e. a version[9] of $\psi_B(x) = P(Y \in B \,|\, X = x)$ that is a probability measure in B.

For a point processes Ξ on $\mathcal{X}$ (now used for our state space again) without fixed atom at $x \in \mathcal{X}$, we have $\Xi(\{x\}) = 0$ almost surely, so that $\mathbb{P}(\Xi \in D \,|\, \Xi(\{x\}) = 1)$ is not determined by the usual definition of conditional probabilities. However, we may use a similar approach as above by integrating over x with respect to $\mathbb{E}\Xi$.

General assumption: For the remainder of this section let Ξ be a point process on $\mathcal{X}$ with existing expectation measure $\mathbb{E}\Xi = \mu_1$.

Definition. (a) The *Campbell measure* C on $\mathcal{X} \times \mathfrak{N}$ (equipped with the product σ-algebra $(\mathcal{B} \otimes \mathcal{N})$) is given by

$$C(A \times D) := \mathbb{E}\big(\Xi(A) 1\{\Xi \in D\}\big) \quad \text{for all } A \in \mathcal{B}, D \in \mathfrak{N}.$$

(b) The *reduced Campbell measure* $C^!$ on $\mathcal{X} \times \mathfrak{N}$ is given by

$$C^!(A \times D) := \mathbb{E}\bigg(\int_A 1\{\Xi - \delta_x \in D\}\, \Xi(dx)\bigg) \quad \text{for all } A \in \mathcal{B}, D \in \mathfrak{N}.$$

[9] $\tilde{\psi}$ is a version of ψ if $\tilde{\psi}_B = \psi_B$ almost surely, for every $B \in \mathcal{A}_\mathcal{Y}$

That the above equations define unique measures on $\mathcal{X} \times \mathfrak{N}$ follows by our usual measure existence and uniqueness theorems. Note that the term $\Xi - \delta_x$ in the above integral is only needed where $\Xi(\{x\}) \geq 1$, in which case $\Xi - \delta_x \in \mathfrak{N}$. If, more pedantically, we would like to have a measurable integrand that is defined on all of $\Omega \times \mathcal{X}$, we should write $\Xi - \delta_x 1\{\Xi(\{x\}) \geq 1\}$.

The proposition below imitates the construction of (regular) conditional probabilities, where the role of $\mathbb{P}(X \in A, Y \in B)$ is taken by $C(A \times D)$ and $C^!(A \times D)$. Recall that a *probability kernel* K from a measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ to another measurable space $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ is a map $\mathcal{X} \times \mathcal{B}_\mathcal{Y} \to \mathbb{R}_+$ such that

(a) $K(\cdot, A)$ is a measurable function for every $A \in \mathcal{B}_\mathcal{Y}$;

(b) $K(x, \cdot)$ is a probability measure for every $x \in \mathcal{X}$.

Proposition 2.S. *There are a.s. unique*[10] *probability kernels Q and $Q^!$ from $\mathcal{X}$ to $\mathfrak{N}$ such that*

(i) $\int_A Q_x(D)\ (\mathbb{E}\Xi)(dx) = C(A \times D)$ *for all $A \in \mathcal{B}, D \in \mathfrak{N}$;*

(ii) $\int_A Q^!_x(D)\ (\mathbb{E}\Xi)(dx) = C^!(A \times D)$ *for all $A \in \mathcal{B}, D \in \mathfrak{N}$.*

Proof. We consider only the measure C; the argument for $C^!$ is virtually identical.

Fix $D \in \mathcal{N}$ and choose a partition $(\mathcal{X}_i)_{i \in \mathbb{N}}$ of $\mathcal{X}$ into bounded measurable sets. Since

$$C(\mathcal{X}_i \times D) \leq C(\mathcal{X}_i \times \mathfrak{N}) = \mu_1(\mathcal{X}_i) < \infty,$$

we obtain that $C(\cdot \times D)$ is σ-finite.

Furthermore, any μ_1-null set $A \in \mathcal{B}$ is also a $C(\cdot \times D)$-null set, since

$$C(A \times D) \leq \mu_1(A) = 0.$$

Hence $C(\cdot \times D)$ is absolutely continuous with respect to μ_1.

By the Radon–Nikodym theorem there is a μ_1-a.s. unique map $Q(\cdot, D)\colon \mathcal{X} \to \mathbb{R}_+$ such that

$$C(A \times D) = \int_A Q(x, D)\ \mu_1(dx) \quad \text{for every } A \in \mathcal{B}.$$

The more laborious part of the proof consists of showing that there is a version of Q that is a probability kernel. It can be shown (see also Exercise Sheet 2) that there is a metric d on $\mathfrak{N}$ such that $(\mathfrak{N}, d)$ is a complete separable metric space whose Borel σ-algebra is $\mathcal{N}$. We then may use the same construction of the probability kernel (and the same argument for uniqueness) as in the case of regular conditional probabilities. See Kallenberg (2002), Theorem 6.3 for details. □

[10] i.e. $Q(x, \cdot)$ and $Q^!(x, \cdot)$ may be altered only on a $(\mathbb{E}\Xi)$-null set in x. We write $Q_x(\cdot)$ instead of $Q(x, \cdot)$, and $Q^!_x(\cdot)$ instead of $Q^!(x, \cdot)$.

Definition. We call Q and $Q^!$ from Proposition 2.S the *Palm kernel*[11] and the *reduced Palm kernel* of Ξ, respectively.

If $(\Xi_x)_{x\in\mathcal{X}}$ is a collection of point processes where Ξ_x has distribution Q_x each individual Ξ_x is (somewhat inconsistently) called a *Palm process given a point at* x. If $(\Xi_x^!)_{x\in\mathcal{X}}$ is a collection of point processes where every $\Xi_x^!$ has distribution $Q_x^!$ each individual $\Xi_x^!$ is called a *reduced Palm process given a point at* x.

We can always define the Ξ_x, $x \in \mathcal{X}$, on a common probability space, at least as long as no condition about the joint distribution is required.[12]

Theorem 2.T (Campbell–Mecke for "raw" Palm processes)**.** *For any $\mathcal{B}\otimes\mathcal{N}$-measurable function $h : \mathcal{X} \times \mathfrak{N} \to \mathbb{R}_+$, we have*

$$\mathbb{E}\left(\int_{\mathcal{X}} h(x,\Xi)\ \Xi(dx)\right) = \int_{\mathcal{X}} \mathbb{E}h(x,\Xi_x)\ (\mathbb{E}\Xi)(dx). \tag{2.7}$$

Proof. The proof follows essentially from the same extension argument as the Campbell formula, but is a bit more difficult now due to the presence of the product space. We start with h of the form $h(x,\xi) = 1_A(x)1_D(\xi)$ for $A \in \mathcal{B}$ and $D \in \mathcal{N}$. Then

$$\begin{aligned}\mathbb{E}\left(\int_{\mathcal{X}} h(x,\Xi)\ \Xi(dx)\right) &= \mathbb{E}\big(\Xi(A)1_D(\Xi)\big)\\ &= \int_A \mathbb{P}(\Xi_x \in D)\ \mu_1(dx)\\ &= \int_{\mathcal{X}} \mathbb{E}h(x,\Xi_x)\ \mu_1(dx)\end{aligned}$$

by definition of the Palm process.

Next we consider general indicators $h = 1_G$ for $G \in \mathcal{B}\otimes\mathcal{N}$ and use a monotone class argument (see Kallenberg (2002), Theorem 1.1). Let

$$\mathcal{D} := \{G \in \mathcal{B}\otimes\mathcal{N}\colon \text{Equation (2.7) holds for } h = 1_D\}$$

and

$$\mathcal{C} := \{A \times D\colon A \in \mathcal{B}, D \in \mathcal{N}\}.$$

Then $\mathcal{C} \subset \mathcal{D}$ by the first part of the proof and $\mathcal{C}$ is stable under intersections. On the other hand $\mathcal{D}$ is (in the terminology of Kallenberg (2002)) a λ-system, i.e. $\mathcal{X}\times\mathfrak{N} \in \mathcal{D}$ and $\mathcal{D}$ is closed under proper differences ($G_1, G_2 \in \mathcal{D}$, $G_1 \subset G_2$ implies $G_2 \setminus G_1 \in \mathcal{D}$; by linearity

[11] This has of course nothing to do with the fruit of a palm tree ;-), but the name was chosen in honour of the Swedish electrical engineer and statistician Conrad Palm (1907–1951) who introduced in his Ph.D. thesis "Intensitätsschwankungen im Fernsprechverkehr" (1943) among many other topics "Palm probabilities" for stationary point processes on $\mathbb{R}$.

[12] By the Łomnicki–Ulam theorem (Kallenberg, 2002, Corollary 6.18) or, exploiting the fact that $\mathfrak{N}$ is Polish, also by the Kolmogorov extension theorem.

of integrals) and the taking of increasing limits ($(G_n) \in \mathcal{D}^{\mathbb{N}}$ where $G_n \subset G_{n+1}$ for all n implies $\bigcup_n G_n \in \mathcal{D}$; by monotone convergence). Therefore $\mathcal{B} \otimes \mathcal{N} = \sigma(\mathcal{C}) \subset \mathcal{D} \subset \mathcal{B} \otimes \mathcal{N}$, and thus the statement holds for general indicators.

It is then straightforward to see that the statement holds for simple functions (by linearity of integrals) and for general measurable functions (by monotone convergence). □

Corollary 2.U. *We have always $\mathbb{P}\big(\Xi_x(\{x\}) \geq 1\big) = 1$ for $\mathbb{E}\Xi$-almost every $x \in \mathcal{X}$. Thus $\Xi_x - \delta_x$ is a well-defined point process*[13]*. We have $\Xi_x - \delta_x \sim Q^!_x$, i.e. the reduced Palm process given a point at x is obtained by removing x from the corresponding "raw" Palm process.*

Corollary 2.V (Campbell–Mecke)**.** *For any $\mathcal{B}\otimes\mathcal{N}$-measurable function $h : \mathcal{X}\times\mathfrak{N} \to \mathbb{R}_+$, we have*

$$\mathbb{E}\bigg(\int_{\mathcal{X}} h(x, \Xi - \delta_x)\, \Xi(dx)\bigg) = \int_{\mathcal{X}} \mathbb{E}h(x, \Xi^!_x)\, (\mathbb{E}\Xi)(dx). \tag{2.8}$$

Proof of Corollaries 2.U and 2.V. As usual, let $(\mathcal{X}_i)_{i\in\mathbb{N}}$ be a partition of $\mathcal{X}$ into bounded measurable sets. Let $h(x,\xi) := 1_{\mathcal{X}_i}(x)1\{\xi(\{x\}) \geq 1\}$ for all $x \in \mathcal{X}$ and $\xi \in \mathfrak{N}$. The function h is measurable because it can be shown that for an arbitrary dissecting system $\big((A_{ni})_{1\leq i\leq m_n}\big)_n$

$$\{(x,\xi)\colon \xi(\{x\}) \geq 1\} = \bigcap_{n\in\mathbb{N}} \bigcup_{i=1}^{m_n} \big(A_{ni} \times \{\xi\colon \xi(A_{ni}) \geq 1\}\big) \in \mathcal{B} \otimes \mathcal{N}.$$

Theorem 2.T implies that

$$\begin{aligned}\mu_1(\mathcal{X}_i) &= \mathbb{E}\bigg(\int_{\mathcal{X}_i} 1\{\Xi(\{x\}) \geq 1\}\, \Xi(dx)\bigg)\\ &= \int_{\mathcal{X}_i} \mathbb{P}\big(\Xi_x(\{x\}) \geq 1\big)\, \mu_1(dx).\end{aligned}$$

Hence

$$\int_{\mathcal{X}_i} \big(1 - \mathbb{P}\big(\Xi_x(\{x\}) \geq 1\big)\big)\, \mu_1(dx) = 0,$$

so that $\mathbb{P}\big(\Xi_x(\{x\}) \geq 1) = 1$ for $\mathbb{E}\Xi$-almost every $x \in \mathcal{X}_i$. Since a countable union of null sets is a null set, we obtain that $\mathbb{P}\big(\Xi_x(\{x\}) \geq 1) = 1$ for $\mathbb{E}\Xi$-almost every $x \in \mathcal{X}$. By this we can identify $\Xi_x - \delta_x$ with $\Xi_x - \delta_x 1\{\Xi_x(\{x\}) \geq 1\} \in \mathfrak{N}$.

For measurable $h\colon \mathcal{X} \times \mathfrak{N} \to \mathbb{R}_+$ we define

$$\tilde{h}(x,\xi) := h\big(x, \xi - \delta_x 1\{\xi(\{x\}) \geq 1\}\big).$$

[13]To be precise, $\Xi_x - \delta_x 1\{\Xi_x(\{x\}) \geq 1\}$, which is obtained from $\Xi_x - \delta_x$ by modification on a $\mathbb{P}$-null set, is a well-defined point process for $\mathbb{E}\Xi$-almost every x (note that Ξ_x is only defined for $\mathbb{E}\Xi$-almost every x). We usually suppress the factor $1\{\Xi_x(\{x\}) \geq 1\}$ for visual pleasure.

Being a composition of measurable functions, $\tilde{h}$ is measurable. Another application of Theorem 2.T yields

$$\begin{aligned}\mathbb{E}\biggl(\int_{\mathcal{X}} h(x, \Xi - \delta_x)\, \Xi(dx)\biggr) &= \mathbb{E}\biggl(\int_{\mathcal{X}} \tilde{h}(x, \Xi)\, \Xi(dx)\biggr) \\ &= \int_{\mathcal{X}} \mathbb{E}\tilde{h}(x, \Xi_x)\, \mu_1(dx) \\ &= \int_{\mathcal{X}} \mathbb{E}h(x, \Xi_x - \delta_x)\, \mu_1(dx).\end{aligned}$$

Choosing $h = 1_{A\times D}$ for $A \in \mathcal{B}$, $D \in \mathcal{N}$, we obtain that $\Xi_x - \delta_x \sim Q_x^!$. Consequently, the right hand side of the above equation is equal to

$$\int_{\mathcal{X}} \mathbb{E}h(x, \Xi_x^!)\, \mu_1(dx),$$

which proves Corollary 2.V. □

Since the "raw" Palm distribution is by Corollary 2.U only a slightly degenerate variant of the reduced Palm distribution, we will often make statements only about the latter. However, it is usually a bit easier to use the defining equation for the raw Palm distribution (see Proposition 2.S(i)) to prove these statements.

The next theorem says that a Poisson process is its own reduced Palm process. At least in the world of simple point processes it is also true that Poisson processes are the only processes with this property. In what follows the addition "for $\mathbb{E}\Xi$-almost every $x \in \mathcal{X}$" is sometimes tacitly omitted when Palm processes are involved.

Theorem 2.W (Slivnyak–Mecke)**.** *If Ξ is a Poisson process, then $\mathscr{L}(\Xi_x^!) = \mathscr{L}(\Xi)$.*

If conversely Ξ is a simple point process with $\mathscr{L}(\Xi_x^!) = \mathscr{L}(\Xi)$ for $\mathbb{E}\Xi$-almost every x, then Ξ is a Poisson process.

For the proof we need the following crucial characterization of the Poisson distribution

Lemma 2.X. *A random variable Z on $\mathbb{Z}_+$ is $\mathrm{Po}(\ell)$-distributed if and only if for every bounded function $g\colon \mathbb{Z}_+ \to \mathbb{R}$ the following equality holds:*

$$\mathbb{E}\bigl(Z\, g(Z)\bigr) = \ell\, \mathbb{E}g(Z+1). \tag{2.9}$$

Proof. If $Z \sim \mathrm{Po}_\ell$, Equation (2.9) follows by straightforward computation. If on the other hand Equation (2.9) holds for every g as specified above, choose for $k \in \mathbb{Z}_+$ the particular functions g_k defined as $g_k(j) = 1\{j = k\}$ for any $j \in \mathbb{Z}_+$. This yields $\mathbb{P}(Z = k) = \frac{\ell}{k}\mathbb{P}(Z = k-1)$ for every $k \geq 1$ and therefore the existence of a constant $c > 0$ such that $\mathbb{P}(Z = k) = c\frac{\ell^k}{k!}$. Since probabilites have to sum up to one, $c = e^{-\ell}$. □

Proof of Theorem 2.W. Let Ξ be a Po(λ)-process on $\mathcal{X}$. By Corollary 2.U we may alternatively show that

$$\int_A \mathbb{P}(\Xi + \delta_x \in D)\ \lambda(dx) = C(A \times D) \tag{2.10}$$

for all $A \in \mathcal{B}$ and $D \in \mathcal{N}$. The proof of Proposition 2.A shows that any finite measure on $\mathfrak{N}$ is completely determined by its values on the sets

$$D_{B_1,\dots,B_r;k_1,\dots,k_r} := \{\xi \in \mathfrak{N}\colon \xi(B_1) = k_1, \dots, \xi(B_r) = k_r\},$$

where $B_1, \dots, B_r \in \mathcal{B}$ are bounded and pairwise disjoint and $k_1, \dots, k_r \in \mathbb{Z}_+$. It is therefore sufficient to prove that

$$\int_A \mathbb{P}(\Xi + \delta_x \in D_{B_1,\dots,B_r;k_1,\dots,k_r})\ \lambda(dx) = C(A \times D_{B_1,\dots,B_r;k_1,\dots,k_r}) \tag{2.11}$$

for all such sets and for bounded $A \in \mathcal{B}$, whence Equation (2.10) is obtained, first for bounded A, and then by σ-additivity for any $A \in \mathcal{B}$ and $D \in \mathcal{N}$.

Writing $B := \bigcup_{i=1}^r B_i$, we obtain (2.11) via

$$\begin{aligned}
&\int_A \mathbb{P}(\Xi + \delta_x \in D_{B_1,\dots,B_r;k_1,\dots,k_r})\ \lambda(dx)\\
&\quad= \int_{A\setminus B} \mathbb{P}\big(\forall j\colon \Xi(B_j) = k_j\big)\ \lambda(dx)\\
&\qquad+ \sum_{i=1}^r \int_{A\cap B_i} \mathbb{P}\big(\Xi(B_i) + 1 = k_i \text{ and } \forall j \neq i\colon \Xi(B_j) = k_j\big)\ \lambda(dx)\\
&\quad= \lambda(A \setminus B)\, \mathbb{P}\big(\forall j\colon \Xi(B_j) = k_j\big)\\
&\qquad+ \sum_{i=1}^r \lambda(A \cap B_i)\, \mathbb{P}\big(\Xi(B_i) = k_i - 1\big)\, \mathbb{P}\big(\forall j \neq i\colon \Xi(B_j) = k_j\big)\\
&\quad= \mathbb{E}\big(\Xi(A \setminus B) 1\{\forall j\colon \Xi(B_j) = k_j\}\big) + \sum_{i=1}^r \mathbb{E}\big(\Xi(A \cap B_i) 1\{\forall j\colon \Xi(B_j) = k_j\}\big)\\
&\quad= \mathbb{E}\big(\Xi(A) 1\{\Xi \in D_{B_1,\dots,B_r;k_1,\dots,k_r}\}\big).
\end{aligned}$$

The penultimate equality follows from the fact that Ξ-counts of disjoint sets are independent and by

$$\begin{aligned}
\lambda(A \cap B_i)\, \mathbb{P}\big(\Xi(B_i) = k_i - 1\big) &= \sum_{l=1}^{k_i} \lambda(A \cap B_i)\, \mathbb{P}\big(\Xi(A \cap B_i) = l - 1\big)\, \mathbb{P}\big(\Xi(B_i \setminus A) = k_i - l\big)\\
&= \sum_{l=1}^{k_i} \mathbb{E}\big(\Xi(A \cap B_i)\, 1\{\Xi(A \cap B_i) = l\}\big)\, \mathbb{P}\big(\Xi(B_i \setminus A) = k_i - l\big)\\
&= \mathbb{E}\big(\Xi(A \cap B_i) 1\{\Xi(B_i) = k_i\}\big),
\end{aligned}$$

where the second equality is a consequence of Lemma 2.X.

To show the converse direction in the theorem, let $B \in \mathcal{B}$ be a bounded set and $g : \mathbb{Z}_+ \to \mathbb{R}$ an arbitrary bounded function. Set $h(x, \xi) := 1_B(x)\, g(\xi(B))$, which defines a measurable function $\mathcal{X} \times \mathfrak{N} \to \mathbb{R}_+$, so that by Theorem 2.T and $\mathcal{L}(\Xi_x) = \mathcal{L}(\Xi + \delta_x)$

$$\begin{aligned}\mathbb{E}\big(\Xi(B)\, g(\Xi(B))\big) &= \mathbb{E}\bigg(\int h(x, \Xi)\; \Xi(dx)\bigg) \\ &= \int \mathbb{E}h(x, \Xi + \delta_x)\; \mu_1(dx) \\ &= \mu_1(B)\, \mathbb{E}\big(g(\Xi(B) + 1)\big).\end{aligned}$$

Thus Lemma 2.X implies that $\Xi(B) \sim \mathrm{Po}(\mu_1(B))$ for every bounded $B \in \mathcal{B}$. By Proposition 2.J we obtain that Ξ is a $\mathrm{Po}(\mu_1)$-process. □

Without proof we mention the following result, which further justifies our interpretation of the Palm distribution Q_x as the distribution of the point process given a point at x.

Proposition 2.Y (Special case of Kallenberg (1986), Theorem 12.8.)**.** *Let Ξ be simple and let $\big((A_{ni})_{1 \le i \le m_n}\big)_n$ be a dissecting system for $\mathcal{X}$. For arbitrary $x \in \mathcal{X}$, denote by $A_n(x)$ the unique set among A_{ni}, $1 \le i \le m_n$, that contains x. Then*

$$\mathbb{P}(\Xi_x \in D) = \lim_{n\to\infty} \mathbb{P}\big(\Xi \in D \;\big|\; \Xi(A_n(x)) \ge 1\big) = \lim_{n\to\infty} \mathbb{P}\big(\Xi \in D \;\big|\; \Xi(A_n(x)) = 1\big).$$

For non-simple point processes one has to be careful, because the intuition of the Palm distribution as the distribution given a point at some location x usually fails. This is illustrated by a Poisson process H whose expectation measure λ has an atom at some location x, i.e. $\alpha := \lambda(\{x\}) > 0$. By Theorem 2.W its reduced Palm process is again H, so

$$\mathbb{P}\big(\mathrm{H}_x(\{x\}) = k + 1\big) = \mathbb{P}\big(\mathrm{H}(\{x\}) = k\big) = \frac{\alpha^k}{k!} e^{-\alpha}$$

for any $k \in \mathbb{Z}_+$. On the other hand, writing $Z := \mathrm{H}(\{x\})$, we have

$$\mathbb{P}(Z = k + 1 \,|\, Z \ge 1) = \frac{\mathbb{P}(Z = k + 1)}{\mathbb{P}(Z \ge 1)} = \frac{\alpha^{k+1} e^{-\alpha}}{(k + 1)!\,(1 - e^{-\alpha})}$$

and

$$\mathbb{P}(Z = k + 1 \,|\, Z = 1) = 1\{k = 0\}$$

for any $k \in \mathbb{Z}_+$. Thus neither $\mathscr{L}(\mathrm{H} \,|\, \mathrm{H}(\{x\}) \ge 1)$ nor $\mathscr{L}(\mathrm{H} \,|\, \mathrm{H}(\{x\}) = 1)$ corresponds to the Palm distribution of H although $\mathbb{E}\mathrm{H}(\{x\}) > 0$.

Palm processes for stationary point processes

A number of important results can be obtained for stationary point processes on $\mathbb{R}^d$. The following theorem provides an explicit representation of the Palm distribution that has

an interesting interpretation (see Remark 2.Ä). We write $T_x : \mathfrak{N} \to \mathfrak{N}, \xi \mapsto \xi\theta_x^{-1}$ for the shift of point measures by the vector x (cf. Section 2.4).

Theorem 2.Z. *Let Ξ be a stationary point process on $\mathbb{R}^d$ with expectation measure $m_1 Leb^d$, where $m_1 \in (0,\infty)$. For any $A \in \mathcal{B}$ with $0 < |A| < \infty$ we have*

$$Q_0^!(D) = \frac{1}{m_1 |A|} \mathbb{E}\bigg(\int_A 1\{\Xi - \delta_x \in T_x(D)\}\, \Xi(dx) \bigg) \tag{2.12}$$

for all $D \in \mathcal{N}$. Furthermore

$$Q_x^! = Q_0^! T_x^{-1};$$

that is,

$$\mathscr{L}(\Xi_x^!) = \mathscr{L}(T_x(\Xi_0^!))$$

for Leb^d-almost every x.

Proof. Define $\mu(A) := \mathbb{E}\big(\int_A 1\{\Xi - \delta_x \in T_x(D)\}\, \Xi(dx)\big)$ for $A \in \mathcal{B}$. Obviously μ is a locally finite measure, and since

$$\begin{aligned}
\int_{\theta_y^{-1}(A)} 1\{\Xi - \delta_x \in T_x(D)\}\, \Xi(dx) &= \int_{\theta_y^{-1}(A)} 1\{T_x^{-1}(\Xi) - \delta_0 \in D\}\, \Xi(dx) \\
&= \int_A 1\{T_{x-y}^{-1}(\Xi) - \delta_0 \in D\}\, \Xi\theta_y^{-1}(dx) \\
&= \int_A 1\{T_x^{-1}(\Xi\theta_y^{-1}) - \delta_0 \in D\}\, \Xi\theta_y^{-1}(dx) \\
&= \int_A 1\{\Xi\theta_y^{-1} - \delta_x \in T_x(D)\}\, \Xi\theta_y^{-1}(dx)
\end{aligned}$$

and Ξ is stationary, we can see that μ is translation invariant, hence a multiple of Lebesgue measure. Therefore, writing $K_0(D)$ for the right hand side of (2.12), it is clear that the definition of $K_0(D)$ does not depend on A, and that K_0 is a probability measure. Defining $K(x,\cdot) := K_0 T_x^{-1}$, it is immediately clear that K is a probability measure in the second component, and it can be shown that K is measurable in the first component.

We now use a monotone class argument as in the proof of Theorem 2.T to show that $K(x,\cdot) = Q_x^!$. It can be seen that

$$\mathcal{D} := \bigg\{ G \in \mathcal{B} \otimes \mathcal{N} \colon \mathbb{E}\bigg(\int_{\mathcal{X}} 1_G(x, \Xi - \delta_x)\, \Xi(dx) \bigg) = m_1 \int_{\mathcal{X}} \int_{\mathfrak{N}} 1_G(x,\xi)\, K(x, d\xi)\, dx \bigg\}$$

is a λ-system and, writing $A \diamond D := \{(x,\xi) \in \mathcal{X} \times \mathfrak{N} \colon x \in A, \xi \in T_x(D)\}$, that

$$\mathcal{C} := \big\{ A \diamond D \colon A \in \mathcal{B}, D \in \mathcal{N} \big\}$$

is stable under intersections and generates $\mathcal{B} \otimes \mathcal{N}$. By the definition of K_0 and $K(x,\cdot)$ we have

$$\mathbb{E}\bigg(\int_A 1\{\Xi - \delta_x \in T_x(D)\}\, \Xi(dx) \bigg) = m_1 |A| K_0(D) = m_1 \int_A K(x, T_x(D))\, dx$$

and hence $\mathcal{C} \subset \mathcal{D}$. The monotone class theorem yields then that $\mathcal{D} = \mathcal{B} \otimes \mathcal{N}$. Thus we may in particular choose for the set G any rectangle $A \times D$, yielding

$$\mathbb{E}\left(\int_A 1\{\Xi - \delta_x \in D\} \, \Xi(dx) \right) = m_1 \int_A K(x, D) \, dx$$

for all $A \in \mathcal{B}$ and $D \in \mathcal{N}$. Hence $Q_x^! = K(x, \cdot)$ for Leb^d-almost every x. □

Remark 2.Ä. *Theorem 2.Z allows another interpretation of Palm probabilities in the stationary case which is very important for point pattern statistics. The indicator in Equation* (2.12) *counts a success if the event "D shifted to x" occurs for the point pattern* $\Xi - \delta_x$*. The corresponding Palm probability is obtained as an "expected success rate" of all points in any given set A of finite positive volume. We may thus interprete* $Q_0^!(D)$ *intuitively as the probability that, seen from a "typical" point of the process* Ξ*, the rest of the process satisfies D (where any reference to the origin is replace by a reference to the point).*

For the sake of illustration assume that $\Xi = \sum_{i=1}^M \delta_{S_i}$ *is simple and think of the event* $D = \{\xi \in \mathfrak{N} : \xi(B(0,r)) = 0\}$ *that there is no point within distance* $r > 0$ *of the origin. Then* $\{\Xi - \delta_{S_i} \in T_{S_i}(D)\} = \{\Xi(\dot{B}(S_i, r)) = 0\}$*, where* $\dot{B}(x,r) = B(x,r) \setminus \{x\}$ *denotes the punctured closed ball. We may then interprete* $Q_0^!(D)$ *as the probabililty that there are no further points within distance* $r > 0$ *of a "typical" point.*

From here on, when we refer to a typical point *of a stationary point process, it will always have a rigorous meaning in terms of the Palm distribution, even if we do not mention this explicitly.*

Another interesting result for stationary point processes is that the expectation measure of the reduced Palm process given a point at the origin coincides up to a normalizing constant with the $\mathcal{K}$-measure. Analogous results would exist for higher moment measures of the reduced Palm process and higher reduced factorial moment measures of the original process (which were not introduced in this course).

Proposition 2.B̈. *Let* Ξ *be a stationary point process on* $\mathcal{X} = \mathbb{R}^d$ *with expectation measure* $m_1 Leb^d$*, where* $m_1 \in (0, \infty)$*. Then the expectation measure* $\mu_1^!$ *of the Palm process* $\Xi_0^!$ *exists if and only if the second moment measure of* Ξ *exists. In this case we have*

$$\mu_1^! = m_1 \mathcal{K}.$$

Proof. Let $A \in \mathcal{B}$, and define $[0,1]^d \diamond A := \{(x,y) \in \mathcal{X}^2 \colon x \in [0,1]^d, y \in \theta_x(A)\}$. By the Campbell–Mecke formula (Corollary 2.V) for the function $h(x, \xi) = 1_{[0,1]^d}(x) \, \xi(\theta_x(A))$, we

obtain that

$$\begin{aligned}\mathbb{E}\big(\Xi^{[2]}([0,1]^d \diamond A)\big) &= \mathbb{E}\bigg(\int_{[0,1]^d} (\Xi - \delta_x)(\theta_x(A))\ \Xi(dx)\bigg) \\ &= m_1 \int_{[0,1]^d} \mathbb{E}\Xi^!_x(\theta_x(A))\ dx \\ &= m_1 \int_{[0,1]^d} \mathbb{E}\Xi^!_0(A)\ dx \\ &= m_1 \mathbb{E}\Xi^!_0(A),\end{aligned} \tag{2.13}$$

where the first equality becomes clear if we represent the point processes in the usual way as $\Xi = \sum_{i=1}^{M} \delta_{S_i}$ and correspondingly $\Xi^{[2]} = \sum_{i,j=1}^{M,\neq} \delta_{(S_i,S_j)}$.

Equation (2.13) implies that $\mu_1^!$ exists if and only if $\mathbb{E}\big(\Xi^{[2]}([0,1]^d \diamond B)\big)$ is finite for all bounded $B \in \mathcal{B}$. By the stationarity in the set $[0,1]^d$ this can be seen to be equivalent to the fact that $\mathbb{E}\big(\Xi^{[2]}(B_1 \times B_2)\big) < \infty$ for all bounded $B_1, B_2 \in \mathcal{B}$, which by $m_1 < \infty$ is equivalent to the existence of μ_2.

Under these existence conditions, Equation (2.6) identifies the left hand side of Equation (2.13) as $m_1^2 \mathcal{K}(A)$, which completes the proof. □

Higher order Palm processes

We have only treated distributions of point processes given one point at a certain location. One could easily generalize the definitions and most theorems to incorporate distributions of point processes given $k \in \mathbb{N}$ points at k different locations. This leads under the assumption that the k-th moment measure μ_k of the point process Ξ exists to higher order Palm and reduced Palm processes $\Xi_{x_1,\ldots,x_k}$ and $\Xi^!_{x_1,\ldots,x_k}$, respectively. Concentrating once more on the reduced versions we can obtain in particular the following generalizations of results in this section.

Theorem 2.C̈ (k-th order Campbell–Mecke)**.** *For any $\mathcal{B}^k \otimes \mathcal{N}$-measurable function $h : \mathcal{X}^k \times \mathfrak{N} \to \mathbb{R}_+$, we have*

$$\begin{aligned}&\mathbb{E}\bigg(\int_{\mathcal{X}^k} h\big(x_1,\ldots,x_k; \Xi - \textstyle\sum_{i=1}^k \delta_{x_i}\big)\ \Xi^{[k]}(dx_1 \ldots dx_k)\bigg) \\ &\qquad = \int_{\mathcal{X}^k} \mathbb{E}h\big(x_1 \ldots, x_k; \Xi^!_{x_1,\ldots,x_k}\big)\ \mu_{[k]}(dx_1 \ldots dx_k).\end{aligned} \tag{2.14}$$

Theorem 2.D̈ (k-th order Slivnyak–Mecke)**.** *If Ξ is a simple Poisson process, then $\mathscr{L}(\Xi^!_{x_1,\ldots,x_k}) = \mathscr{L}(\Xi)$ for $\mu_{[k]}$-almost every $(x_1,\ldots,x_k) \in \mathcal{X}^k$.*

2.5.2 Papangelou kernels and conditional intensities

Due to time constraints we only give a brief glimpse at what is essentially the "opposite" form of conditioning. We would like to formalize the probabilities "$\mathbb{P}\big(\Xi(\{x\}) = 1 \;\big|\; \Xi\,|\,_{\mathcal{X}\setminus\{x\}} = \xi\big)$", which in a sense is done in the analogous way as in Palm theory, but the construction gets a little more difficult due to technical problems.

We consider a simple point process Ξ,[14] and require what is known as "Condition (Σ)", which says that for every bounded $B \in \mathcal{B}$,

$$\mathbb{P}(\Xi(B) = 0 \,|\, \Xi|_{B^c}) > 0 \quad \text{almost surely.}$$

While this condition is satisfied in many contexts, it is for example violated if Ξ is a binomial process (with more than one point).

One can show then that the reduced Campbell measure may be disintegrated with respect to the distribution of the point process, i.e. there is a a.s. unique locally finite kernel R[15] from $\mathfrak{N}^*$ to $\mathcal{X}$ such that

$$C^!(B \times D) = \int_D R(B \,|\, \xi)\, \mathbb{P}\Xi^{-1}(d\xi)$$

for all bounded $B \in \mathcal{B}$ and all $D \in \mathcal{N}^*$. We call this kernel the *Papangelou kernel* of Ξ.

Under more restrictive conditions the Papangelou kernel has a density $\lambda(x \,|\, \xi)$ with respect to Lebesgue measure, which is known as the *Papangelou* or *conditional intensity*, such that for $\mathbb{P}\Xi^{-1}$-almost every ξ

$$R(A \,|\, \xi) = \int_A \lambda(x \,|\, \xi)\; Leb^d(dx).$$

for all $A \in \mathcal{B}$. This density can be characterized via Ξ directly, as the (for $\mathbb{P}\Xi^{-1}$-almost every $\xi \in \mathfrak{N}^*$) unique function $\lambda(\cdot \,|\, \xi)\colon \mathcal{X} \to \mathbb{R}_+$ that satisfies the *Georgii–Nguyen–Zessin-Formula*:

$$\mathbb{E}\bigg(\int_{\mathcal{X}} h(x, \Xi - \delta_x)\; \Xi(dx)\bigg) = \int_{\mathcal{X}} \mathbb{E}\big(h(x,\Xi)\lambda(x \,|\, \Xi)\big)\; dx$$

for "nice" measurable functions $h : \mathcal{X} \times \mathfrak{N}^* \to \mathbb{R}_+$.

In the point process models considered from Chapter 3 onward, the conditional intensity $\lambda(\cdot \,|\, \xi)$ can be defined by a simple explicit formula, so that the precise theoretical construction will not be needed. Nevertheless, it is helpful to understand how $\lambda(\cdot \,|\, \xi)$ relates to the general theory of conditioning.

[14] Assume without loss of generality that Ξ is $(\mathfrak{N}^*, \mathcal{N}^*)$-valued.
[15] We write $R(A \,|\, \xi)$ instead of $R(\xi, A)$.

Chapter 3

Point pattern models and descriptive statistics

In this chapter we tailor various objects and results from general point process theory to our needs in spatial statistics and introduce point estimators of various characteristic quantities introduced earlier. From now on we have the following simplifications:

- $\mathcal{X} \subset \mathbb{R}^d$.
- All point processes considered are simple; without loss of generality we assume that they take only values in $(\mathfrak{N}^*, \mathcal{N}^*)$ (sometimes written as $(\mathfrak{N}^*(\mathcal{X}), \mathcal{N}^*(\mathcal{X}))$ for clarification), where $\mathcal{N}^* = \{D \cap \mathfrak{N}^* \colon D \in \mathcal{N}\}$. We usually identify $\xi = \sum_{i=1}^m \delta_{s_i} \in \mathfrak{N}^*$ with $\{s_1, \ldots, s_m\}$, so that we may write for example $\sum_{x \in \xi} f(x)$ instead of $\sum_{i=1}^m f(s_i)$ or $\int f(x)\, \xi(dx)$.
- Our point processes usually have an expectation measure with a density with respect to Leb^d. We denote this density by λ and call it the *intensity* of the point process. So in contrast with the previous chapter λ is from now on a function $\mathcal{X} \to \mathbb{R}_+$. We use μ_1 or λ^* to refer to the parameter/expectation measure of a Poisson process.
- $\mathcal{W}$ (for "window") denotes a compact subset of $\mathbb{R}^d$ on which we observe our point process. Our data will always be a single(!) point pattern on $\mathcal{W}$. The underlying point process model lives on the superset $\mathcal{X} \supset \mathcal{W}$, where typically $\mathcal{X} = \mathcal{W}$ or $\mathcal{X} = \mathbb{R}^d$. The latter case most commonly occurs when we (assume that we) observe a part of a stationary point process.

3.1 Point process densities

A very convenient approach to dealing with many common distributions on $\mathbb{R}$ is working with their densities with respect to Lebesgue measure. When we replace $\mathbb{R}$ with the space

$\mathfrak{N}^*$, we would like to have another simple reference measure with the help of which we can define a large class of useful point process models. This measure is the "standard" Poisson distribution $\mathrm{Po}_1 = \mathrm{Po}(Leb^d)$.

Definition. Let Ξ be a point process Ξ on $\mathcal{X} = \mathcal{W}$. We call a measurable function $f : \mathfrak{N} \to \mathbb{R}_+$ with

$$\mathbb{P}(\Xi \in D) = \int_D f(\xi)\, P_1(d\xi) \quad \text{for any } D \in \mathcal{N}^*$$

a *density* of Ξ.

As usual the Radon–Nikodym theorem guarantees that a density exists if and only if $\mathscr{L}(\Xi) \ll \mathrm{Po}_1$, i.e. every Po_1-null set D satisfies $\mathbb{P}(\Xi \in D) = 0$; then the density is Po_1-almost surely unique.

We begin with the rather simple example of the density of a general Poisson process.

Proposition 3.A. *Let* H *be a Poisson process on* $\mathcal{X} = \mathcal{W}$ *whose expectation measure* λ^* *has a density* λ *with respect to* Leb^d*-measure. Then* H *has a density given by*

$$f(\xi) = e^{-\lambda^*(\mathcal{W})+|\mathcal{W}|} \prod_{x\in\xi} \lambda(x)$$

for every $\xi \in \mathfrak{N}^*$.

Proof. Let $\mathrm{H}_1 \sim \mathrm{Po}_1$. For arbitrary $D \in \mathcal{N}^*$ we obtain for the function f specified above that

$$\begin{aligned}
&\mathbb{P}(\mathrm{H} \in D)\\
&= \sum_{m=0}^{\infty} \mathbb{P}(\mathrm{H} \in D \,|\, \mathrm{H}(\mathcal{W}) = m)\, \frac{\lambda^*(\mathcal{W})^m}{m!} e^{-\lambda^*(\mathcal{W})}\\
&= \sum_{m=0}^{\infty} \frac{\lambda^*(\mathcal{W})^m}{m!} e^{-\lambda^*(\mathcal{W})} \int_{\mathcal{W}} \cdots \int_{\mathcal{W}} 1\big\{\{x_1,\dots,x_m\} \in D\big\} \frac{\lambda(x_1)}{\lambda^*(\mathcal{W})} \cdots \frac{\lambda(x_m)}{\lambda^*(\mathcal{W})}\, dx_1 \cdots dx_m\\
&= \sum_{m=0}^{\infty} \frac{|\mathcal{W}|^m}{m!} e^{-|\mathcal{W}|} \int_{\mathcal{W}} \cdots \int_{\mathcal{W}} 1\big\{\{x_1,\dots,x_m\} \in D\big\} f(\{x_1,\dots,x_m\}) \frac{1}{|W|} \cdots \frac{1}{|W|}\, dx_1 \cdots dx_m\\
&= \sum_{m=0}^{\infty} \mathbb{E}\big(1\{\mathrm{H}_1 \in D\} f(\mathrm{H}_1) \;\big|\; \mathrm{H}_1(\mathcal{W}) = m\big)\, \mathbb{P}\big(\mathrm{H}_1(\mathcal{W}) = m\big)\\
&= \int_D f(\xi)\ \mathrm{Po}_1(d\xi).
\end{aligned}$$

Hence f is a density of H. □

Remark 3.B. *We only consider densities for point processes on compact subsets of* $\mathbb{R}^d$*. For point processes on all of* $\mathbb{R}^d$ *an appropriate description in terms of densities becomes far more technical and complicated. If we just extend the above definition naively to point*

processes on $\mathcal{X} = \mathbb{R}^d$, many natural and interesting point processes are excluded. For example a Po_ℓ-process on $\mathbb{R}^d$ does not have a density (with respect to Po_1) in this sense, unless $\ell = 1$.

This can be seen by considering a sequence of disjoint sets $A_i \in \mathcal{B}^d$ with $|A_i| = 1$ and the event $D = \{\xi \in \mathfrak{N}^*(\mathbb{R}^d)\colon \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \xi(A_i) = \ell\} \in \mathcal{N}^*$. By the strong law of large numbers we obtain for $\mathrm{H} \sim \mathrm{Po}_\ell$ and $\mathrm{H}_1 \sim \mathrm{Po}_1$ that $\mathbb{P}(\mathrm{H}_1 \in D) = 0$ (since $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathrm{H}_1(A_i) = 1$ almost surely), but $\mathbb{P}(\mathrm{H} \in D) = 1$. Hence $\mathrm{Po}_\ell \not\ll \mathrm{Po}_1$.

When doing parametric inference, we concentrate on so-called Gibbs processes.

Definition. A function $f : \mathfrak{N}^* \to \mathbb{R}_+$ is called *hereditary* if $f(\xi) = 0$ implies $f(\eta) = 0$ for any point patterns $\xi, \eta \in \mathfrak{N}^*$ with $\xi \subset \eta$.

Definition. A point process Ξ on $\mathcal{W}$ with hereditary density is called a *Gibbs process*.

We now give a simple definition of the conditional intensity, which was mentioned at the end of Chapter 2, for the special case of Gibbs processes. The intuition is that $\lambda(x \,|\, \xi)$ describes the intensity of the point process at x conditional on the point process everywhere else being ξ, i.e. in some loose sense $\lambda(x \,|\, \xi)\, dx = \mathbb{E}(\Xi(dx) \,|\, \Xi \setminus dx = \xi)$.

Definition. Let Ξ be a Gibbs process on $\mathcal{W}$ with density f. The function $\lambda(\cdot \,|\, \cdot)\colon \mathcal{W} \times \mathfrak{N}^*(\mathcal{W}) \to \mathbb{R}_+$ that is given by

$$\lambda(x \,|\, \xi) = \begin{cases} \frac{f(\xi \cup \{x\})}{f(\xi)} & \text{if } x \notin \xi, \\ \frac{f(\xi)}{f(\xi \setminus \{x\})} & \text{if } x \in \xi, \end{cases}^1$$

where we set $\frac{0}{0} := 0$, is called *conditional intensity* of Ξ.

Remark 3.C. *(i) It can be seen that $\lambda(\cdot \,|\, \cdot)$ is $\mathcal{B} \otimes \mathcal{N}^*(\mathcal{W})$-measurable.*

(ii) It follows immediately from Proposition 3.A that the conditional intensity of a Poisson process coincides with its intensity, i.e. for any $\mathrm{Po}(\lambda(x)dx)$-process $\lambda(\cdot \,|\, \xi) = \lambda(\cdot)$ for every $\xi \in \mathfrak{N}^*$.

(iii) The conditional intensity determines the distribution of a Gibbs process completely. Clearly,

$$f(\{x_1, \ldots, x_m\}) = f(\emptyset) \prod_{i=1}^m \lambda(x_i \,|\, \{x_1, \ldots, x_{i-1}\}),$$

where $f(\emptyset)$ is determined as

$$f(\emptyset) = \left(\int_{\mathfrak{N}^*} \textstyle\prod_{i=1}^m \lambda(x_i \,|\, \{x_1, \ldots, x_{i-1}\})\ \mathrm{Po}_1(d\{x_1, \ldots, x_m\}) \right)^{-1}$$

if the integral is finite.

[1] From the point of view of the theory in Chapter 2 and for most considerations later on, the second case is not relevant since $\mathbb{P}(x \in \Xi) = 0$ for every $x \in \mathcal{W}$ by the fact that $\mathscr{L}(\Xi) \ll \mathrm{Po}_1$.

The following theorem is important for computations and shows that the conditional intensity is just a special case of the (Papangelou) conditional intensity mentioned at the end of Chapter 2.

Theorem 3.D (Georgii–Nguyen–Zessin Formula)**.** *For a Gibbs process Ξ on $\mathcal{W}$ with conditional intensity $\lambda(\cdot \,|\, \cdot)$, the equation*

$$\mathbb{E}\biggl(\int_{\mathcal{W}} h(x, \Xi \setminus \{x\})\, \Xi(dx)\biggr) = \int_{\mathcal{W}} \mathbb{E}\bigl(h(x,\Xi)\lambda(x \,|\, \Xi)\bigr)\, dx$$

holds for every measurable function $h : \mathcal{X} \times \mathfrak{N}^* \to \mathbb{R}_+$.

Proof. Letting $\mathrm{H}_1 \sim \mathrm{Po}_1$, we have

$$\begin{aligned}
\mathbb{E}\biggl(\int_{\mathcal{W}} h(x, \Xi \setminus \{x\})\, \Xi(dx)\biggr) &= \mathbb{E}\biggl(\int_{\mathcal{W}} h(x, \mathrm{H}_1 \setminus \{x\})\, f(\mathrm{H}_1)\, \mathrm{H}_1(dx)\biggr) \\
&= \int_{\mathcal{W}} \mathbb{E}\bigl(h(x, \mathrm{H}_1)\, f(\mathrm{H}_1 \cup \{x\})\bigr)\, dx \\
&= \int_{\mathcal{W}} \mathbb{E}\bigl(h(x, \mathrm{H}_1)\, \lambda(x \,|\, \mathrm{H}_1)\, f(\mathrm{H}_1)\bigr)\, dx \\
&= \int_{\mathcal{W}} \mathbb{E}\bigl(h(x, \Xi)\, \lambda(x \,|\, \Xi)\bigr)\, dx,
\end{aligned}$$

where the second equality follows by the Slivnyak–Mecke theorem (via the Campbell–Mecke theorem for the function $[(x,\xi) \mapsto h(x,\xi)\, f(\xi \cup \{x\})]$) and the third equality makes use of the fact that the density f is hereditary. □

Proposition 3.E. *Let Ξ be a Gibbs process on $\mathcal{W}$ with existing expectation measure. Then*

(i) $\mathbb{E}\Xi \ll \mathrm{Leb}^d$ with density $\lambda(x) = \mathbb{E}\bigl(\lambda(x \,|\, \Xi)\bigr)$ for $x \in \mathcal{W}$.

(ii) The reduced Palm process $\Xi_x^!$ is a Gibbs process for $\mathbb{E}\Xi$-almost every $x \in \mathcal{W}$ with density

$$f_x^!(\xi) := \frac{f(\xi \cup \{x\})}{\lambda(x)}, \qquad \xi \in \mathfrak{N}^*,$$

for any $x \in \mathcal{W}$ with $\lambda(x) > 0$.

Proof. Statement (i) follows from the GNZ-formula with $h(x,\xi) = 1_A(x)$ for all $x \in \mathcal{W}$ and $\xi \in \mathfrak{N}^*$, where $A \in \mathcal{B}$ is arbitrary.

Statement (ii) has a similar proof as Theorem 3.D. By the Slivnyak–Mecke theorem (again via Campbell–Mecke) for the function $h(x,\xi) := 1_A(x) 1_D(\xi) f(\xi)$, where $A \in \mathcal{B}_{\mathcal{W}}$

and $D \in \mathcal{N}^*(\mathcal{W})$ are arbitrary, we have

$$\begin{aligned}\mathbb{E}\Big(\int_A 1\{\Xi \setminus \{x\} \in D\}\, \Xi(dx)\Big) &= \mathbb{E}\Big(\int_A 1\{\mathrm{H}_1 \setminus \{x\} \in D\} f(\mathrm{H}_1)\, \mathrm{H}_1(dx)\Big) \\ &= \int_A \mathbb{E}\big(1\{\mathrm{H}_1 \in D\}\, f(\mathrm{H}_1 \cup \{x\})\big)\, dx \\ &= \int_A \mathbb{E}\Big(\frac{1\{\mathrm{H}_1 \in D\}\, f(\mathrm{H}_1 \cup \{x\})}{\lambda(x)}\Big)\lambda(x)\, dx,\end{aligned}$$

where we set the ratio to zero (or some other value) if $\lambda(x) = 0$. The last equality holds because by part (i) and hereditarity

$$\lambda(x) = \mathbb{E}\big(\lambda(x \,|\, \Xi)\big) = \mathbb{E}\big(f(\mathrm{H}_1 \cup \{x\})\big),$$

which means that $\lambda(x) = 0$ implies that $f(\mathrm{H}_1 \cup \{x\}) = 0$ almost surely.

Hence by the definition of the reduced Palm process

$$\begin{aligned}\mathbb{P}(\Xi_x^! \in D) &= \mathbb{E}\Big(\frac{1\{\mathrm{H}_1 \in D\}\, f(\mathrm{H}_1 \cup \{x\})}{\lambda(x)}\Big) \\ &= \int_D \frac{f(\xi \cup \{x\})}{\lambda(x)}\ \mathrm{Po}_1(d\xi)\end{aligned}$$

for $\mathbb{E}\Xi$-almost every $x \in \mathcal{W}$, whence we obtain that $f_x^!$ is a density of $\Xi_x^!$ (no matter how we define $f_x^!$ for $x \in \mathcal{W}$ with $\lambda(x) = 0$).

Hereditarity of $f_x^!$ where $\lambda(x) > 0$ follows immediately from the hereditarity of f since $f(\xi \cup \{x\}) = 0$ implies $f(\eta \cup \{x\}) = 0$ for $\xi \subset \eta$. □

With the help of the conditional intensity we can now define formally what we understand by attractiveness and repulsiveness of point processes.

Definition. A Gibbs process Ξ is called *attractive* if $\lambda(x \,|\, \xi) \le \lambda(x \,|\, \eta)$ for all $\xi, \eta \in \mathfrak{N}^*$ with $\xi \subset \eta$. It is called *repulsive* if $\lambda(x \,|\, \xi) \ge \lambda(x \,|\, \eta)$ for all $\xi, \eta \in \mathfrak{N}^*$ with $\xi \subset \eta$.

3.2 Parametric families of point pattern models

As we have seen in Chapter 1 real point pattern data typically seem to have a far more complex structure than that generated by a Poisson process. In this section we present some commonly used parametric models of Gibbs point processes that have very intuitive densities.

Unfortunately it is only possible to specify these densities up to a normalizing constant c which cannot be computed analytically. This is a major nuissance for the computation of maximum likelihood estimators in Chapter 5. It is also one of the main reasons for the highly important role of the conditional intensity: since $\lambda(x \,|\, \xi)$ is defined as the ratio of two values of the same density, the multiplicative constant c cancels out and the conditional intensity can be calculated explicitly.

3.2.1 The Strauss process

For parameters $R > 0$ (interaction range), $\beta > 0$ (intensity control), $\gamma \in [0, 1]$ (interaction control), the density of the *(homogeneous) Strauss process on* $\mathcal{W}$ is given as

$$f(\xi) := f_{\text{Strauss}(R;\beta,\gamma)}(\xi) := c\beta^{|\xi|}\gamma^{s_R(\xi)}, \quad \xi \in \mathfrak{N}^*, \tag{3.1}$$

where $|\xi| = \xi(\mathcal{W})$ denotes the total number of points of ξ and

$$s_R(\xi) = \frac{1}{2} \sum_{x,y\in\xi,\, x\neq y} 1\{\|x - y\| \leq R\}$$

is the number of pairs of points in ξ that are "R-close", i.e. no more than distance R apart. Consequently the conditional intensity of the Strauss process may be computed as

$$\lambda(x \,|\, \xi) = \beta\gamma^{s_R(x;\xi)}, \tag{3.2}$$

where

$$s_R(x;\xi) = \sum_{y\in\xi\setminus\{x\}} 1\{\|x - y\| \leq R\}.$$

Note how the unwelcome factor c has disappeared. Figure 3.1 shows simulations for various values of γ.

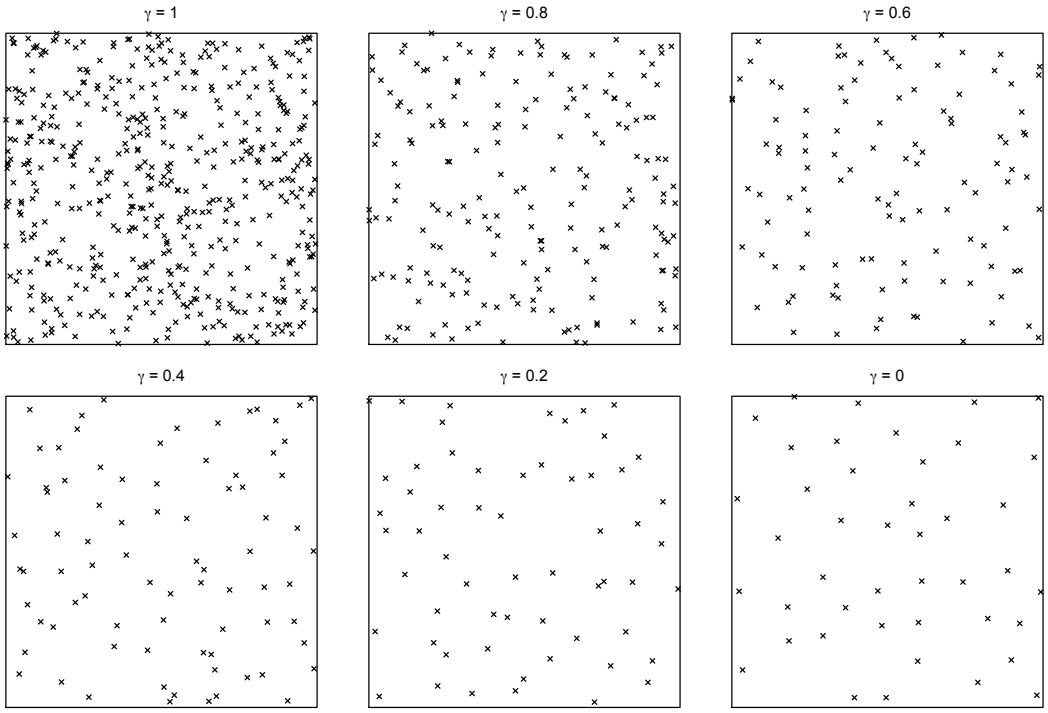

Figure 3.1: Simulations of Strauss processes for various values of γ, where $R = 0.1$ and $\beta = 500$. The realized numbers of points were 496, 188, 113, 81, 59, and 43.

From the density formula (3.1) it is clear that the Strauss$(R; \beta, 1)$-process coincides with the Po_β process. For general $\gamma \in [0, 1]$ the process can be seen from Equation (3.2)

to be repulsive. For $\gamma = 0$ no R-close point pairs are permitted anymore (with probability 1). In this case we refer to Ξ as the standard hard core process. The idea behind this name is that we may imagine that there are hard spheres of radius $R/2$ around the points that do not allow them to be closer than distance R. There are other types of hard core processes like the Matérn type I hard core process encountered in the exercises (Exercise Sheet 3, Problem 8).

It is tempting to allow $\gamma > 1$ in order to encourage clustered point patterns. This, however, is not possible because the corresponding "density" function is not integrable (see Exercise Sheet 5). A useful alternative of a "Strauss-like" process that does allow for clustering is *Geyer's saturation process* (more economically referred to as *Geyer process* in this lecture, although there are other Geyer processes). Its density is given for arbitrary $\gamma \in \mathbb{R}_+$ as

$$f(\xi) := f_{\mathrm{Geyer}(R,s_0;\beta,\gamma)}(\xi) := c\beta^{|\xi|} \prod_{x\in\xi} \gamma^{\min(s_0, s_R(x;\xi))}. \tag{3.3}$$

Thus the number of additional points in the R-neighbourhood of any point that the density gets "rewarded for" (in the case $\gamma > 1$) is limited to s_0. This leads to an integrable density. The Geyer process can be seen to be attractive for $\gamma \geq 1$ and repulsive for $\gamma \leq 1$. In the latter case, note that $\mathrm{Geyer}(R, \infty; \beta, \gamma) = \mathrm{Strauss}(R; \beta, \gamma^2)$.

Some simulations for varying values of s_0 and $\gamma > 1$ can be seen in Figure 3.2.

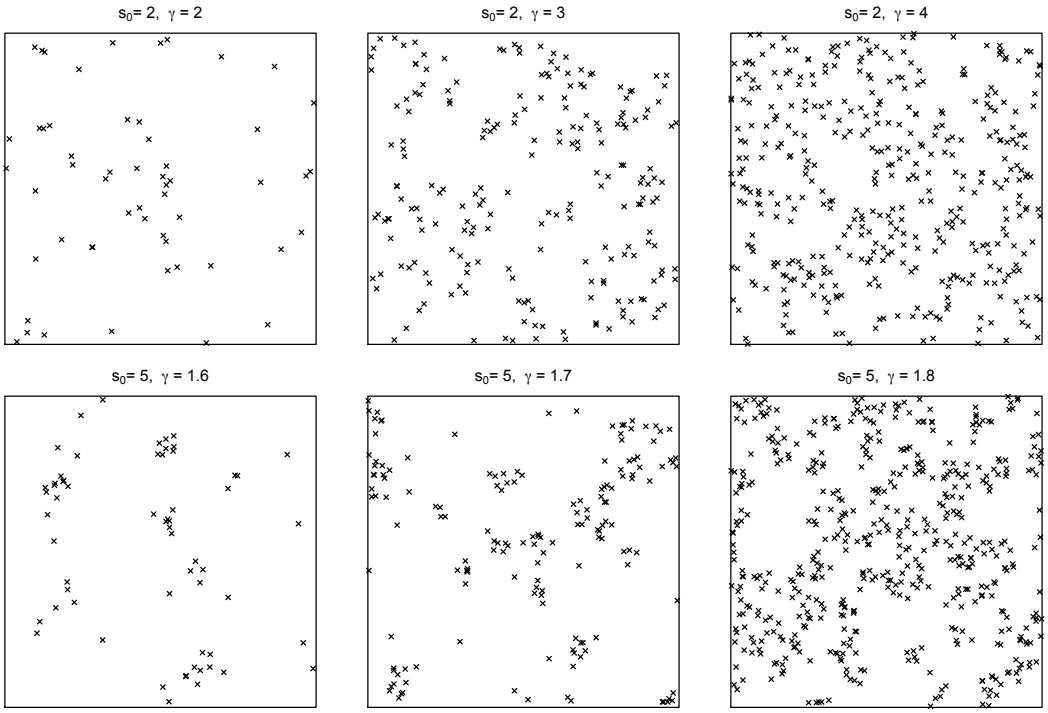

Figure 3.2: Simulations of Geyer's saturation processes for various values of s_0 and γ, where $R = 0.05$ and $\beta = 20$. The realized numbers of points were 55, 209, 401, 61, 158, and 508.

3.2.2 The area-interaction process

Many point pattern models commonly used are so-called pairwise interaction processes, which means essentially that their densities can be described completely as a product of functions of single points and of pairs of points. A very versatile model that considers interactions of arbitrarily high order is the area-interaction process. It can handle spatial aggregation as well as inhibition. For parameters $R > 0$ (interaction range), $\beta > 0$ (intensity control), $\eta > 0$ (interaction control), the density of the *(homogeneous) area-interaction process (AIP)* is given as

$$f(\xi) := f_{\mathrm{AIP}(R;\beta,\eta)}(\xi) := c\,\beta^{|\xi|}\,\eta^{-|U_R(\xi)|}, \tag{3.4}$$

where

$$U_R(\xi) = \bigcup_{x\in\xi} B(x,R).$$

The conditional intensity may be computed as

$$\lambda(x\,|\,\xi) = \beta\,\eta^{-|B(x,R)\setminus\bigcup_{y\in\xi\setminus\{x\}} B(y,R)|}. \tag{3.5}$$

Therefore the area-interaction process is repulsive for $\eta \le 1$ and attractive for $\eta \ge 1$. For $\eta = 1$ we obtain the Po_β-process. Some simulations for varying values of η can be seen in Figure 3.3; note that this time the value of β was adjusted to roughly stabilize the expected number of points.

3.2.3 Inhomogeneous models

Inhomogeneous versions of the models presented so far can be easily obtained by letting β vary with the position of the points of ξ. Just replace in Equations (3.1), (3.3), and (3.4) the factor $\beta^{|\xi|}$ by $\prod_{x\in\xi}\beta(x)$ and in Equations (3.2) and (3.5) the factor β by $\beta(x)$, where β is now an integrable function $\mathcal{W}\to\mathbb{R}_+$.

It is readily checked that the statements about repulsiveness and attractiveness of the processes remain the same, and that with $\gamma = 1$ for the inhomogeneous Strauss and Geyer processes and with $\eta = 1$ for the inhomogeneous area-interaction process, the Poisson processes with intensity function β is obtained.

3.3 Descriptive Statistics

In this section we study the most important tools for exploratory data analysis. Most of the statistics considered are (more or less natural) non-parametric estimators of point process characteristics introduced in Chapter 2. Without making any distributional assumptions on the underlying point process, there is usually not much that can be said

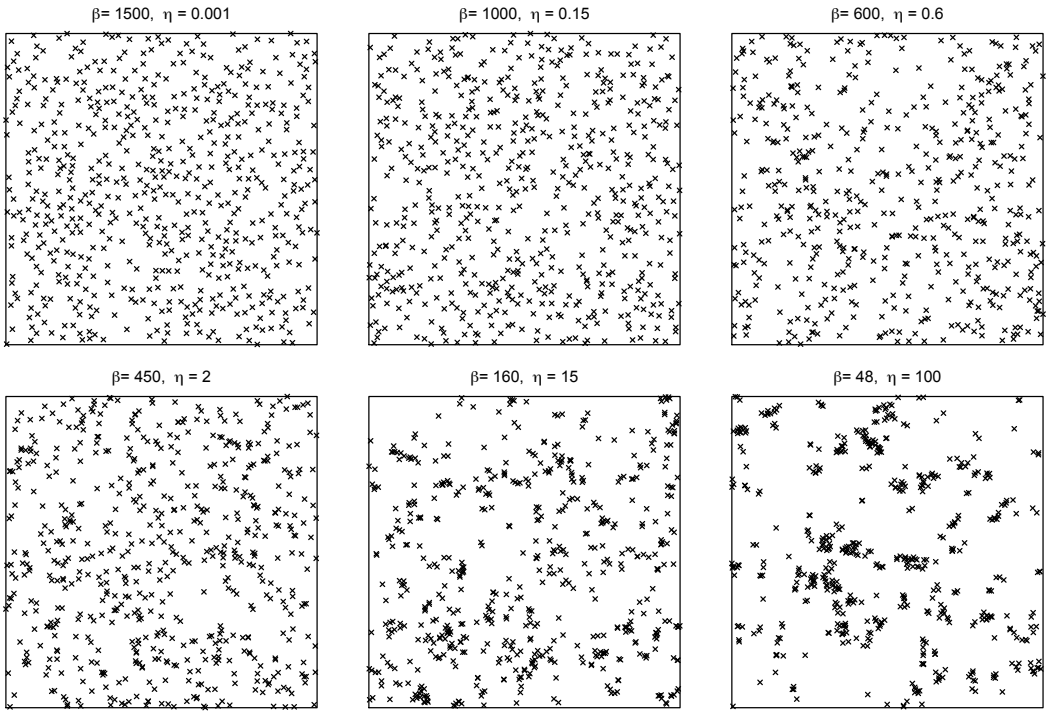

Figure 3.3: Simulations of area interaction processes for various values of η, where $R = 0.02$ and β was manually adjusted in such a way that the expected number of points remains roughly the same. The realized numbers of points were 598, 599, 524, 596, 570, and 536.

about the theoretical properties of these estimators. With distributional assumptions such statements are often limited to Poisson processes. One common goal is to look for natural estimators that are unbiased (or "ratio-unbiased" in some cases) and keep an eye on the variance in common situations.

Let always ξ denote our data point pattern on $\mathcal{W}$ and Ξ the underlying point process on $\mathcal{X} \supset \mathcal{W}$. We always require that $|\mathcal{W}| > 0$.

3.3.1 First order characteristics

For stationary point processes Ξ, the constant intensity λ is most naturally estimated by

$$\widehat{\lambda} := \frac{|\xi|}{|\mathcal{W}|}.$$

The estimator is unbiased since $\mathbb{E}|\Xi| = \lambda|\mathcal{W}|$. If Ξ is a Po_λ-process, then

$$\mathrm{Var}\big(\widehat{\lambda}\big) = \frac{1}{|\mathcal{W}|^2}\,\mathrm{Var}(\Xi(\mathcal{W})) = \frac{\lambda}{|\mathcal{W}|},$$

whence the Bienaymé–Chebyshev inequality yields that

$$\widehat{\lambda} = \lambda + O_p\big(|\mathcal{W}|^{-1/2}\big)$$

as $|\mathcal{W}| \to \infty$. Note that this statement uses $\mathrm{Var}(\Xi(\mathcal{W})) = O(|\mathcal{W}|)$ and is therefore not true for general stationary point processes. Note further that in point pattern statistics based on one point pattern there is usually no sample size n that we can let go to infinity. For stationary point processes the role of n is played by the volume of the observation window.

Turning now to general (i.e. not necessarily stationary) point processes, the common technique to obtain intensity estimates is kernel density estimation, although in principle most techniques from probability density estimation could be adapted to the present situation. Let $\kappa : \mathbb{R}^d \to \mathbb{R}_+$ be a probability density function, called in this context a *kernel*, which is usually assumed to be symmetric with respect to the origin, and often even rotationally symmetric. Let $\kappa_h(z) := \frac{1}{h^d}\kappa(z/h)$, be a scaled version of the kernel, where h is known as its *bandwidth*. We then define

$$\hat{\lambda}(z) := c_h(z) \sum_{x \in \xi} \kappa_h(z - x),$$

where

$$c_h(z) := \left(\int_{\mathcal{W}} \kappa_h(z - x)\, dx \right)^{-1}. \tag{3.6}$$

The term $c_h(z)$ takes care of "edge correction" under the assumption that $\mathcal{X} = \mathbb{R}^d$ or at least that $\mathcal{X}$ is "considerably larger" than $\mathcal{W}$.[2] For reasonable kernels $c_h(z)$ is (almost) 1 if z is well in the interior of $\mathcal{W}$, but it is larger close to the edge of $\mathcal{W}$, thus compensating for the fact that we do not get the contributions to the density estimate from points that lie outside the window $\mathcal{W}$. Note if the true density $\lambda(z) = \lambda$ is constant, that the above estimator is unbiased by Campbell's formula:

$$\mathbb{E}\widehat{\lambda}(z) = \mathbb{E}\frac{\int_{\mathcal{W}} \kappa_h(z - x)\, \Xi(dx)}{\int_{\mathcal{W}} \kappa_h(z - x)\, dx} = \lambda.$$

In general, $\hat{\lambda}(z)$ is biased due to the redistribution of mass by the kernels κ_h, but the above correction still removes "unnecessary" additional bias close to the edge.

It is known from probability density estimation that the precise shape of the kernel is not so important (as long as it is reasonable, e.g. centered at the origin and unimodal). What is important is the choice of the bandwidth h, which governs a trade-off between bias and variance of the estimator (as a rule: small h results in small bias and large variance, and large h results in large bias and small variance).

Certain criteria have been developed for the choice of the bandwidth in intensity estimation that are based on mean square error computations for a stationary isotropic

[2]Consequently, if $\mathcal{X} = \mathcal{W}$, there is no need for this kind of edge correction, and we would simply replace $c_h(z)$ by 1. However, there is another problem then: since we know that our point process is concentrated on $\mathcal{W}$, we are not happy about the fact that intensity mass of points close to the edge of $\mathcal{W}$ may spill to $\mathcal{W}^c$. The common approach to this is the reflection of kernels at the boundary of $\mathcal{W}$.

Cox process (see Diggle (1985), Berman and Diggle (1989), and Diggle (2003), Section 8.2). Note that bandwidth rules from probability density estimation are usually not appropriate for intensity estimation.

It is often a good idea to inspect the estimated density for several different bandwidths. A good choice of the bandwidth often also depends on the purpose for which the intensity estimate is needed. For example, in order to detect departures from stationarity, where typically a simple trend is all we can hope to detect, a larger bandwidth might be more appropriate. The default choice in the `spatstat` function `density` (where the kernel is always Gaussian, and the bandwidth is consequently given via the parameter `sigma` with a default for rectangular windows being 1/8 of the shorter window side) serves this purpose rather well.

For illustration we consider a simulated point pattern, displayed in Figure 3.4.

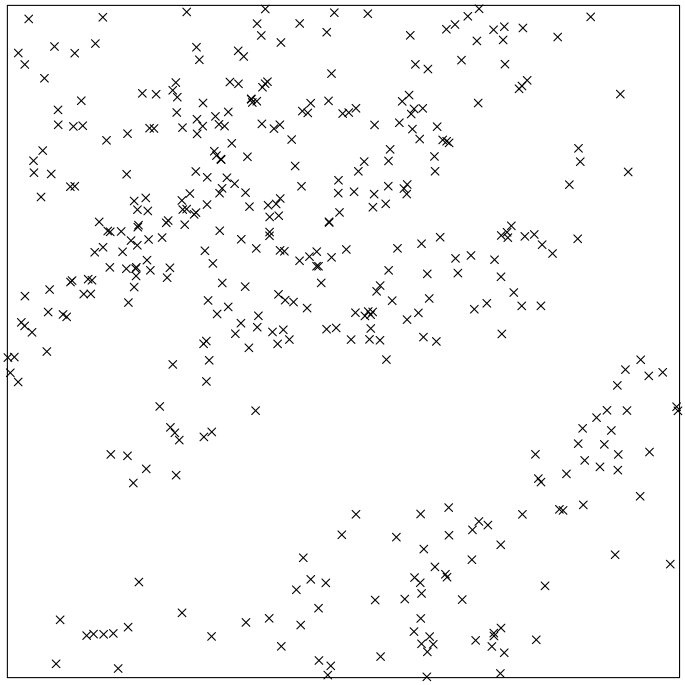

Figure 3.4: Simulated point pattern with 397 points.

Figure 3.5 shows the intensity estimate for two different choices of h. On the left the default choice of $h = 0.125$ (1/8th of the window length) was used, on the right the choice of $h = 0.025$, which might be more appropriate if (for some reason) we decide that the point process is non-stationary and we are interested in the true density per se.

From the point pattern itself and the two intensity pictures in Figure 3.5, we would usually judge that stationarity is still a reasonable assumption. There is a long "channel" from the south-west to north-east without any points, and if possible we would certainly try to find out the reason why there are no points there (if this were a real data set). In

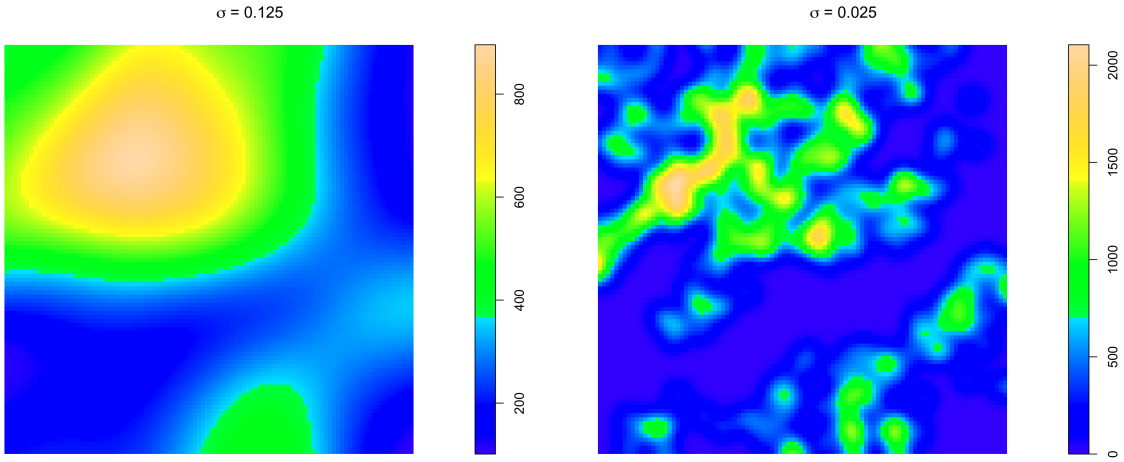

Figure 3.5: Intensity estimates for $h = 0.125$ and $h = 0.025$.

any case there is no simple large-scale feature that speaks against stationarity. By the way, the true distribution from which the data pattern was generated is in fact stationary.

3.3.2 Second order characteristics

We focus here on the case of stationary point processes or, from a more applied point of view, on data ξ on $\mathcal{W}$ that is judged "to may have well come from a stationary point process" Ξ on $\mathbb{R}^d$. Let $\lambda > 0$ be the (constant) intensity and assume that the second moment measure exists. At least for the K-function defined farther below there is an "inhomogeneous version" which can be found in the literature (see also the spatstat function `Kinhom`).

$\mathcal{K}$-measure

We start by the most general descriptor in the stationary case, which is the $\mathcal{K}$-measure introduced in Section 2.4. Writing $\mathcal{W}_x := x + \mathcal{W}$ for the translated observation window, we consider the following estimator for the quantity $\mathcal{K}_* := \lambda^2 \mathcal{K}$,

$$\widehat{\mathcal{K}}_*(B) := \sum_{x,y \in \xi,\, x \neq y} \frac{1_B(y-x)}{|\mathcal{W}_x \cap \mathcal{W}_y|} \tag{3.7}$$

for any bounded $B \in \mathcal{B}^d$ such that $|\mathcal{W} \cap \mathcal{W}_z| > 0$ for all $z \in B$ (setting $\frac{0}{0} := 0$). Proposition 2.B̈ implies that $\mathcal{K}_* = \lambda\, \mathbb{E}\Xi_0^!$. Therefore

$$\sum_{x,y \in \xi,\, x \neq y} \frac{1_B(y-x)}{|\mathcal{W}|} = \frac{1}{|\mathcal{W}|} \sum_{x \in \xi} \sum_{y \in \xi \setminus \{x\}} 1_{x+B}(y) = \frac{1}{|\mathcal{W}|} \int_{\mathcal{W}} (\xi \setminus \{x\})(x+B)\, \xi(dx) \tag{3.8}$$

seems like a reasonable estimator of $\mathcal{K}_*(B)$ since it is a spatial average over $x \in \xi$ of the number of points in "B as seen from the point x" (excluding x itself). More formally, it

can be seen from Theorem 2.Z that

$$\mathcal{K}_*(B) = \lambda\,\mathbb{E}\Xi_0^!(B) = \frac{1}{|\mathcal{W}|}\mathbb{E}\bigg(\int_{\mathcal{W}} (\Xi \setminus \{x\})(x+B)\;\Xi(dx)\bigg).$$

However, since ξ is only a realization of $\Xi|_{\mathcal{W}}$ not of Ξ (i.e. points outside $\mathcal{W}$ are not observed) the estimator in (3.8) is in fact biased, as it systematically underestimates $\mathcal{K}_*(B)$. We therefore weigh each summand on the left hand side of (3.8) with the edge correction factor $\frac{|\mathcal{W}|}{|\mathcal{W}_x \cap \mathcal{W}_y|}$ to obtain the estimator (3.7). Note that $|\mathcal{W}_x \cap \mathcal{W}_y| = |\mathcal{W} \cap \mathcal{W}_{y-x}|$, so that this factor is one over the fraction of the volume of all $z \in \mathcal{W}$ from which the difference $y - x$ can be observed, i.e. lies in $\mathcal{W}$, as compared to the total volume of $\mathcal{W}$.

This form of edge correction is known as *translational edge correction*. The following proposition shows that it works as it should.

Proposition 3.F. *For any bounded $B \in \mathcal{B}^d$ as specified above $\widehat{\mathcal{K}}_*(B)$ is an unbiased estimator of $\mathcal{K}_*(B)$.*

Proof. Write

$$h(x,y) := 1_{\mathcal{W}}(x) 1_{\mathcal{W}}(y) \frac{1_B(y-x)}{|\mathcal{W} \cap \mathcal{W}_{y-x}|}$$

for all $x, y \in \mathbb{R}^d$. By Campbell's formula and Proposition 2.Q we obtain that

$$\begin{aligned}
\mathbb{E}\widehat{\mathcal{K}}_*(B) &= \mathbb{E}\bigg(\int_{\mathbb{R}^d \times \mathbb{R}^d} h(x,y)\;\Xi^{[2]}\big(d(x,y)\big)\bigg) \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} h(x,y)\;\mu_{[2]}\big(d(x,y)\big)\bigg) \\
&= \lambda^2 \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(x, x+y)\;\mathcal{K}(dy)\;dx \\
&= \lambda^2 \int_{\mathbb{R}^d} \frac{1_B(y)}{|\mathcal{W} \cap \mathcal{W}_y|} \int_{\mathbb{R}^d} 1_{\mathcal{W}}(x) 1_{\mathcal{W}}(x+y)\;dx\;\mathcal{K}(dy) \\
&= \lambda^2 \int_{\mathbb{R}^d} 1_B(y)\,\frac{|\mathcal{W} \cap \mathcal{W}_{-y}|}{|\mathcal{W} \cap \mathcal{W}_y|}\;K(dy) \\
&= \lambda^2 \mathcal{K}(B) \\
&= \mathcal{K}_*(B).
\end{aligned}$$

□

Note that the variance of the above estimator becomes large on sets B that are relatively far away from the origin (with regard to the window size), because then $|\mathcal{W}_x \cap \mathcal{W}_y|$ for the relevant x and y is small.

We now estimate the actual $\mathcal{K}$-measure as

$$\widehat{\mathcal{K}}(B) := \frac{\widehat{\mathcal{K}}_*(B)}{\widehat{\lambda^2}},$$

where

$$\widehat{\lambda^2} = \frac{|\xi|(|\xi|-1)}{|\mathcal{W}|^2},$$

which is a slightly corrected version of the natural plug-in estimator $\left(\widehat{\lambda}\right)^2$. We use $\widehat{\lambda^2}$ because it is an unbiased estimator of λ^2 under Poisson assumption (not in general!), so that $\widehat{\mathcal{K}}$ is so-called *ratio-unbiased*, meaning for a general parameter θ of the form θ_1/θ_2 that the estimator is the ratio $\hat{\theta} = \hat{\theta}_1/\hat{\theta}_2$ of two unbiased estimators $\hat{\theta}_1$ for θ_1 and $\hat{\theta}_2$ for θ_2.

As a descriptive statistic for visual inspection it is again most instructive to look at the corresponding kernel estimate of the density of $\mathcal{K}$ (assuming it exists), given by

$$\widehat{\widetilde{g}}(z) = \frac{1}{\widehat{\lambda^2}} \sum_{x,y \in \xi,\, x \neq y} \frac{\kappa_h(z-(y-x))}{|\mathcal{W}_x \cap \mathcal{W}_y|}. \tag{3.9}$$

We concentrate on values $z \in \mathbb{R}^d$ that are "well inside" the domain of $\widehat{\widetilde{g}}$ due to the large variance of the estimator close to the boundary of the domain. Consequently, we do not have to trouble ourselves with an additional edge correction due to kernels that lie outside the domain of $\widehat{\widetilde{g}}$ (as we did for the kernel estimator of the intensity function).

Figure 3.6 depicts the output of the `spatstat` function `Kmeasure`, which computes essentially (and presumably even exactly) the estimator $\widehat{\widetilde{g}}(z)$, for the point pattern in Figure 3.4. The "+" denotes the origin. There are three striking features: first there is a

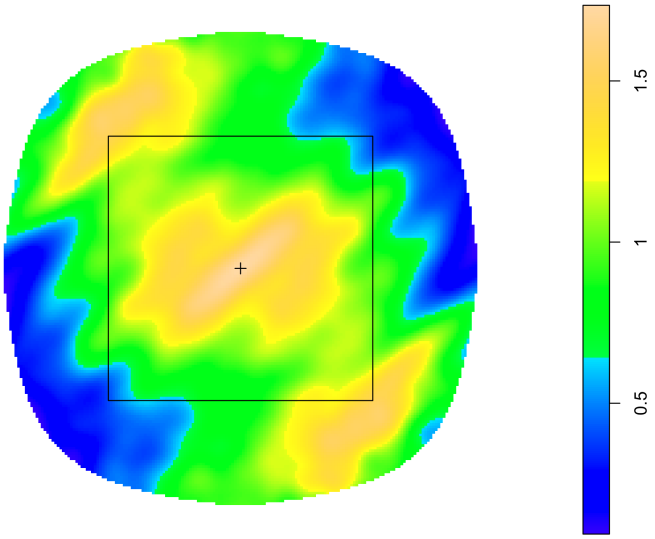

Figure 3.6: $\mathcal{K}$-measure density estimate for $\sigma = 0.04$ with the square $[-0.5, 0.5]^2$ for easy reference of distances.

high density ellipse (hard to see on black-and-white printouts) around the origin in south-westerly to north-easterly direction, which indicates clustering with a preferred tendency in these directions. Secondly, this ellipse together with the fact that there is a general

tendency in the same directions, rather than just a roughly circular picture, indicates that there is pretty strong anisotropy in the point process distribution. Thirdly, there seems to be some periodicity in the north-westerly to south-easterly directions with period of about $1/\sqrt{2}$, as seen from the high-density blotches that are at that distance from the origin. This can be largely attributed to the "channel" in the point pattern. Since the periodicity blotches are already rather far away from the origin it is not too clear if this is a real feature of the point process distribution or an artifact due to sampling.

In fact, the point pattern in Figure 3.4 was generated as a Poisson cluster process. First a homogeneous Poisson process of "parent points" was generated with intensity 30 on a window that was far larger than the unit square. Then each realized point x was replaced by a Poisson process with intensity $12\,\varphi_{x,\Sigma}$, where $\varphi_{x,\Sigma}$ denotes the Gaussian p.d.f. centered at x and having covariance matrix Σ, where in our example

$$\Sigma = 0.004 \cdot \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}.$$

The isotropic case

Suppose now that the point process Ξ on $\mathbb{R}^d$ is stationary and isotropic with intensity $\lambda > 0$ and existing second moment measure. This has two important consequences.

First, the distribution of Ξ is invariant under rotations around *any* centre in $\mathbb{R}^d$.[3] This follows from the fact that a rotation around $x \in \mathbb{R}^d$ may be decomposed into a translation by the vector $-x$, then a rotation around 0 and then a translation by x, all operations under which $\mathscr{L}(\Xi)$ is invariant.

Second, the Palm process $\Xi_0^!$ is isotropic as well, which implies by Proposition 2.B̈ that the $\mathcal{K}$-measure is rotationally invariant. Isotropy of the Palm process can be obtained by Theorem 2.Z as follows. Recall that for any $x \in \mathbb{R}^d$ we write $\theta_x(y) := y + x$ for the translation by the vector $x \in \mathbb{R}^d$ and ψ_Φ for the rotation by the matrix $\Phi \in \mathrm{SO}(d)$. Furthermore $T_x \colon \mathfrak{N}^* \to \mathfrak{N}^*, \xi \mapsto \xi\theta_x^{-1}$ and $R_\Phi \colon \mathfrak{N}^* \to \mathfrak{N}^*, \xi \mapsto \xi\psi_\Phi^{-1}$ for the corresponding transformations of point patterns. Let $D \in \mathcal{N}^*$ and write $\widetilde{D} := R_\Phi(D)$. Furthermore we

[3] For this statement we do not need any requirements about the moment measures of course.

choose an $A \in \mathcal{B}^d$ with $0 < |A| < \infty$ and write $\widetilde{A} := \psi_\Phi(A)$. Then we have

$$\begin{aligned}
\mathbb{P}\big(\Xi_0^! \in \widetilde{D}\big) &= \frac{1}{\lambda|\widetilde{A}|}\mathbb{E}\bigg(\int_{\widetilde{A}} 1\{\Xi - \delta_x \in T_x(\widetilde{D})\}\, \Xi(dx)\bigg) \\
&= \frac{1}{\lambda|\widetilde{A}|}\mathbb{E}\bigg(\int_{\widetilde{A}} 1\{R_\Phi(\Xi) - \delta_x \in T_x(\widetilde{D})\}\, R_\Phi(\Xi)(dx)\bigg) \\
&= \frac{1}{\lambda|A|}\mathbb{E}\bigg(\int_{A} 1\{R_\Phi(\Xi) - \delta_{\psi_\Phi(x)} \in T_{\psi_\Phi(x)}(\widetilde{D})\}\, \Xi(dx)\bigg) \\
&= \frac{1}{\lambda|A|}\mathbb{E}\bigg(\int_{A} 1\{\Xi - \delta_x \in R_\Phi^{-1}\big(T_{\psi_\Phi(x)}(\widetilde{D})\big)\big)\}\, \Xi(dx)\bigg) \\
&= \frac{1}{\lambda|A|}\mathbb{E}\bigg(\int_{A} 1\{\Xi - \delta_x \in T_x\big(R_\Phi^{-1}(\widetilde{D})\big)\}\, \Xi(dx)\bigg) \\
&= \mathbb{P}\big(\Xi_0^! \in D\big),
\end{aligned}$$

where the second equality follows by isotropy and the third equality follows by the transformation theorem and the fact that $|A| = |\widetilde{A}|$ by the rotational invariance of Lebesgue measure.

Since $\mathcal{K}$ is rotation-invariant around zero, we may consider simpler statistics that still contain the full information of the $\mathcal{K}$-measure.[4]

Definition. Let Ξ be a stationary and isotropic point process on $\mathbb{R}^d$ whose second moment measure exists.

(a) The *K-function* (widely known as *Ripley's K-function*) $\mathbb{R}_+ \to \mathbb{R}_+$ is given by

$$K(r) := \mathcal{K}\big(B(0,r)\big) \quad \text{for every } r \in \mathbb{R}_+.$$

(b) The *L-function* (less widely known as *Besag's L-function*) $\mathbb{R}_+ \to \mathbb{R}_+$ is given by

$$L(r) := \big(K(r)/\alpha_d\big)^{1/d} \quad \text{for every } r \in \mathbb{R}_+,$$

where $\alpha_d = |B(0,1)| = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$.

(c) The *pair correlation function* $g\colon \mathbb{R}_+ \to \mathbb{R}$ is given by

$$g(r) := \frac{K'(r)}{\omega_d r^{d-1}} \quad \text{for almost every } r \in \mathbb{R}_+,$$

where $\omega_d = \nu_{d-1}(\partial B(0,1)) = d\alpha_d$ denotes the surface measure of the unit ball.[5] Note that K is by definition an increasing function on $\mathbb{R}_+$, so it is almost everywhere differentiable (see for example Elstrodt, 2007, Chapter VII, Theorem 4.5).

[4] In the case of the pair correlation function: under reasonable additional assumptions.

[5] One way to define ν_{d-1} as a measure is by setting it to $(d-1)$-dimensional Hausdorff measure.

All of the above functions contain the same information, but represent it in different ways. The K-function was historically the first and has the simplest definition. However, the other two functions are nowadays often preferred as they are easier to interpret. The L-function has the advantage that it makes deviations from a Poisson model better visible for small r and that it is more objectively interpretable since sample fluctuations are stabilized over the whole domain of the function (see the example below and further considerations in Chapter 4).

Under reasonable additional assumptions the pair correlation function has a very intuitive interpretation as the following derivation shows. Assume that $\mathcal{K}$ has a rotationally invariant density $\tilde{g}\colon \mathbb{R}^d \to \mathbb{R}_+$, which means that there is a function $\tilde{g}_0\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\tilde{g}(z) = \tilde{g}_0(\|z\|) \quad \text{for every } z \in \mathbb{R}^d.$$

Then by the transformation theorem

$$\begin{aligned} K(r) &= \int_{B(0,r)} \tilde{g}(z)\, dz \\ &= \int_0^r \omega_d s^{d-1} \tilde{g}_0(s)\, ds, \end{aligned}$$

and thus by the fundamental theorem of calculus for the Lebesgue integral (Elstrodt, 2007, Chapter VII, Theorem 4.14)

$$g(r) = \frac{K'(r)}{\omega_d r^{d-1}} = \tilde{g}_0(r) \quad \text{for almost every } r \geq 0. \tag{3.10}$$

This allows the following intuition: Since Ξ is simple the expected number of points in an "infitesimal subset" dx of $\mathbb{R}^d$ corresponds to the probability that there is a point in dx. Hence the pair correlation $g(r)$ may be understood as the "normalized infinitesimal probability" that two points of the process occur at distance r of one another. Somewhat more precisely, the probability that two points occur in infinitesimal sets dx and dy at distance r apart is given by

$$g(r) \cdot \lambda\, dx \cdot \lambda\, dx.$$

We would usually like to compare the graphs of the three functions introduced above to those of a homogeneous Poisson process.

Proposition 3.G. *Let* H *be a* Po_λ*-process. Then*

(i) $K(r) = \alpha_d r^d$ *for all* $r \in \mathbb{R}_+$*;*

(ii) $L(r) = r$ *for all* $r \in \mathbb{R}_+$*;*

(iii) $g(r) = 1$ *for all* $r > 0$*.*

Proof. We have seen in Section 2.4 that the $\mathcal{K}$-measure of a Po_λ-process is Leb^d. The statements follow then immediately from the definitions of the functions. □

Estimators for the three functions are given as follows. For the K-function we may obviously just use the estimator (3.7) for the $\mathcal{K}$-measure applied to balls $B = B(0,r)$. For the L-function we can use the corresponding plug-in estimator. Hence

$$\widehat{K}_{\mathrm{trans}}(r) := \widehat{\mathcal{K}}\big(B(0,r)\big) \quad \text{and} \quad \widehat{L}_{\mathrm{trans}}(r) := \big(\widehat{K}_{\mathrm{trans}}(r)/\alpha_d\big)^{1/d},$$

for $0 \le r < r_{\mathrm{trans}} := \sup\{\tilde{r} \ge 0 \colon |\mathcal{W} \cap \mathcal{W}_z| > 0 \text{ for all } z \in B(0,r)\}$. For example, if $\mathcal{W}$ is a rectangle/cuboid in $\mathbb{R}^d$, then r_{trans} is equal to the shortest side length. We know by Proposition 3.F that $\widehat{K}_{*,\mathrm{trans}}(r) = \widehat{\lambda}^2 \widehat{K}_{\mathrm{trans}}(r)$ is an unbiased estimator of the uncorrected K-function value $K_*(r) := \lambda^2 K(r)$ for every $r \in [0, r_{\mathrm{trans}})$.

It is more efficient to make use of the isotropy assumption when doing edge correction. We set $\widehat{K}_{\mathrm{iso}}(r) := \widehat{K}_{*,\mathrm{iso}}(r)/\widehat{\lambda}^2$ with

$$\widehat{K}_{*,\mathrm{iso}}(r) := \sum_{x,y \in \xi,\, x \neq y} \frac{1\{\|y-x\| \le r\}}{\tilde{\omega}_d(x,y)\, |\mathcal{W}^{(\|y-x\|)}|}$$

for $0 \le r < r_{\mathrm{iso}}$, where

$$\tilde{\omega}_d(x,y) = \frac{\nu_{d-1}\big(B(x, \|y-x\|) \cap \mathcal{W}\big)}{\nu_{d-1}\big(B(x, \|y-x\|)\big)}$$

is the fraction of the surface of the ball centered at x and touching y which lies in $\mathcal{W}$ and

$$\mathcal{W}^{(\tilde{r})} := \{x \in \mathcal{W} \colon \partial B(x, \tilde{r}) \cap \mathcal{W} \neq \emptyset\}$$

is the set of locations in $\mathcal{W}$ from which locations at distance $\tilde{r}$ are observable. The upper bound r_{iso} is just $\sup\{\tilde{r} \ge 0 \colon |\mathcal{W}^{(\tilde{r})}| > 0\}$, which in the case of a rectangular/cuboidal window $\mathcal{W}$ is the length of its body diagonal.

The idea of the edge correction is roughly the following. The factor $1/\tilde{\omega}_d(x,y)$ corrects for the fact that, seen from a point x of ξ, instead of each single point $y \neq x$ that we observe at distance $\tilde{r}$, we would expect by isotropy to have a true number of points at distance "about $\tilde{r}$" that is equal to the total surface of the $\tilde{r}$-ball divided by the observed surface. The factor $1/|\mathcal{W}^{(\|y-x\|)}|$ corrects for the fact that the total number of points we expect to see in this way at distance $\tilde{r}$ is based on those locations x in $\mathcal{W}$ only from which it is at all possible to see the distance $\tilde{r}$.

Formally, it can be shown in a similar way as in Proposition 3.F that $\widehat{K}_{*,\mathrm{iso}}(r)$ is unbiased for $\lambda^2 K(r)$ (see Exercise Sheet 5). The estimator $\widehat{K}_{\mathrm{iso}}(r)$ has usually a smaller variance than $\widehat{K}_{\mathrm{trans}}(r)$ if Ξ is really isotropic. Furthermore $\widehat{K}_{\mathrm{iso}}(r)$ is defined on a larger interval (compare the statements about rectangular windows). However, it is known to be rather sensitive to departures from isotropy, which is why a comparison with $\widehat{K}_{\mathrm{trans}}(r)$ is usually advisable.

We now turn to the estimation of the pair correlation function, which in a sense is less nice, because we have to use kernels again. Since it is difficult to estimate K' directly, the

usual way is via the relation (3.10). Note that in view of the estimator of the density of $\mathcal{K}$ in (3.9) the estimators below are very natural. Let $\kappa : \mathbb{R} \to \mathbb{R}_+$ be a (usually symmetric) kernel function (i.e. a p.d.f.) and $\kappa_h(t) := \frac{1}{h}\kappa(t/h)$ the scaled version with bandwidth $h > 0$. Then

$$\hat{g}(r) := \frac{1}{\widehat{\lambda^2}} \sum_{x,y \in \xi,\, x \neq y} \frac{\kappa_h(r - \|y - x\|)}{\omega_d\, r^{d-1}\, |\mathcal{W}_x \cap \mathcal{W}_y|}$$

and

$$\hat{g}_{\mathrm{iso}}(r) := \frac{1}{\widehat{\lambda^2}} \sum_{x,y \in \xi,\, x \neq y} \frac{\kappa_h(r - \|y - x\|)}{\tilde{\omega}_d(x,y)\, |\mathcal{W}^{(\|y-x\|)}|}$$

are our estimators of the pair correlation function with translational and isotropic edge correction respectively.

Modern recommendations go towards replacing $\widehat{\lambda^2}$ here and in the estimators for the K-function by more complex variants that tend to even out fluctuations of the corresponding estimators of the denominators. The reader is referred to Section 4.3.3 in Illian et al. (2008) for details.

Since the example from Figure 3.4 was clearly anisotropic we consider a new one. The point pattern in Figure 3.7 was simulated from a Strauss$(0.05; 100, 0.3)$-process on the unit square.

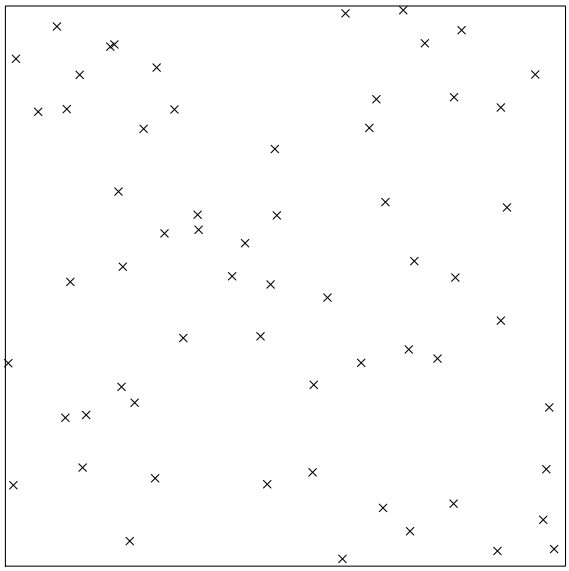

Figure 3.7: Point pattern simulated from the Strauss$(0.05; 100, 0.3)$-distribution. The realized number of points was 62.

Figures 3.8 and 3.9 give the corresponding estimates of the K-, L- and pair correlation functions using the two types edge correction. The theoretical functions for a homogeneous Poisson process are also given. The repulsiveness in the point pattern is clearly seen in all of the functions: the expected number of points (K-,L-functions) and the probability

of points within closer distances are clearly too small. Note that the "critical distance" of (roughly) 0.05 stands out somewhat in all of the graphs: there is a slight hint of a "cusp" in the K- and also the L-function (which will be used in the so-called *cusp-point method* presented in Chapter 5) and there is a substantial increase of the pair correlation function. The K- and L-functions catch up with the theoretical functions at around 0.1 and are then roughly the same.

Also note the effect of the various bandwidth choices in the estimation of the pair correlation function.

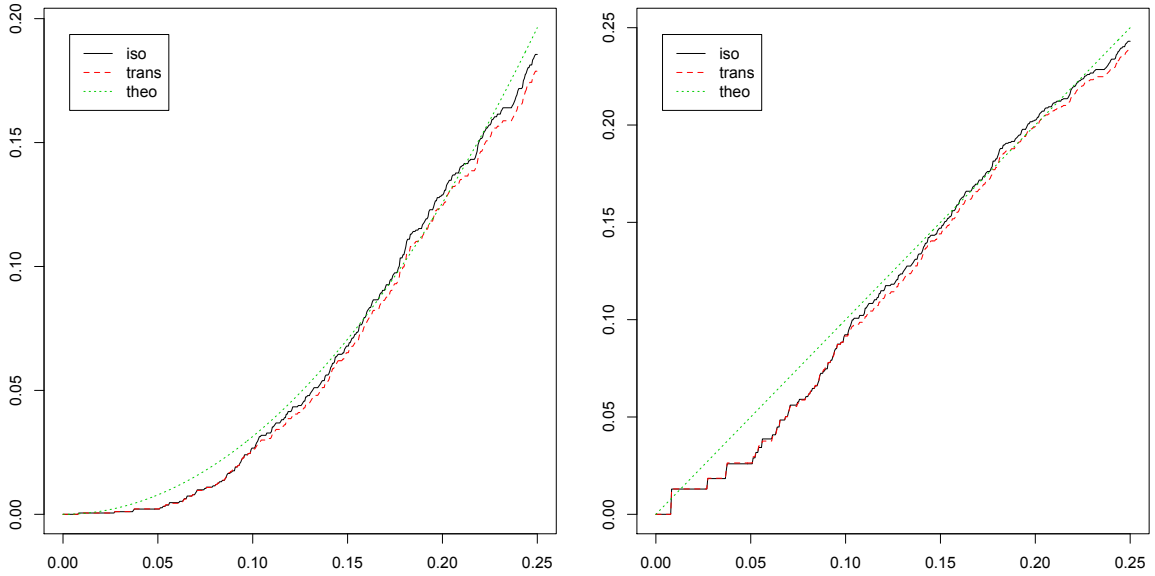

Figure 3.8: Estimated K- and L-functions for the Strauss point pattern in Figure 3.7 using translational and isotropic edge correction.

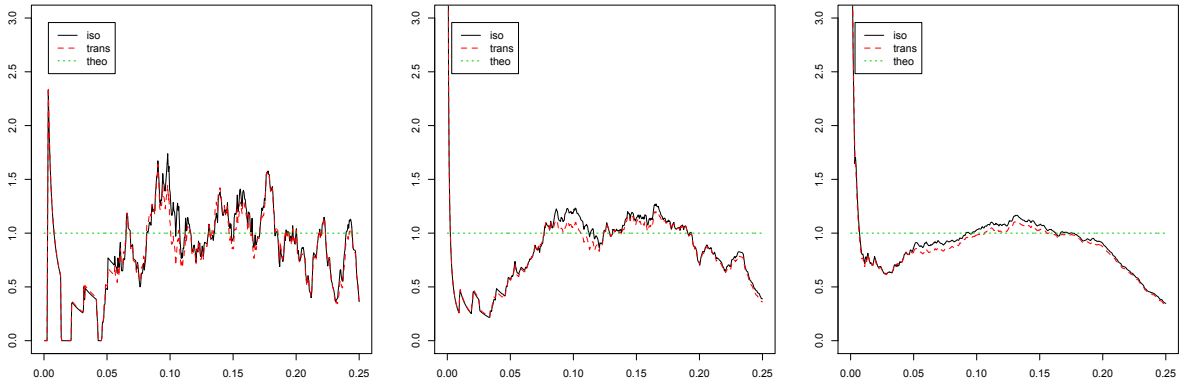

Figure 3.9: Estimated pair correlation functions for the Strauss point pattern using a rectangular kernel with bandwidths 0.003, 0.01, and 0.03 and applying translational and isotropic edge correction.

Of course it would be more informative to take sampling error into account, e.g. for investigating departure from the Poisson assumption to compare the estimated curve

against the usual variation in estimated curves for Poisson samples. While exact theoretical calculations are not feasible, good approximations exist in some cases and it is always possible to use a Monte Carlo based approach. These ideas will be made more explicit in Chapter 4.

3.3.3 Distance based characteristics: the J-function

We stay in the stationary context, but do without isotropy. In this subsection another descriptive statistic of a point process, the so-called J-function, is presented, which has very nice theoretical properties and may complement information gained from K-, L- and pair correlation functions. The objects introduced in what follows are all based on distributions of certain distances in a stationary point process. We always assume that Ξ has finite intensity $\lambda > 0$.

Definition. (a) The *F-function* (also known as *empty space function* or *spherical contact distribution function*) $\mathbb{R}_+ \to [0,1]$ is given by

$$F(r) := \mathbb{P}\big(\Xi(B(0,r)) \geq 1\big) \quad \text{for every } r \in \mathbb{R}_+.$$

(b) The *G-function* (also known as *nearest neighbour distance distribution function*) $\mathbb{R}_+ \to [0,1]$ is given by

$$G(r) := \mathbb{P}\big(\Xi_0^!(B(0,r)) \geq 1\big) \quad \text{for every } r \in \mathbb{R}_+.$$

(c) The *J-function* (van Lieshout and Baddeley, 1996) $A_J \to \mathbb{R}_+$ is given by

$$J(r) := \frac{1-G(r)}{1-F(r)} \quad \text{for every } r \in A_J := \{\tilde{r} \in \mathbb{R}_+ \colon F(\tilde{r}) < 1\}.$$

Remark 3.H. *For arbitrary $x \in \mathbb{R}^d$ and $\xi \in \mathfrak{N}^*$, let $D_x(\xi) := \inf\{\|y-x\| \colon y \in \xi\}$ (where $\inf \emptyset := \infty$ as usual) be the distance from x to the closest point of ξ. Then $\xi(B(x,r)) \geq 1$ if and only if $D_x(\xi) \leq r$.*

By this and the fact that Ξ is stationary we can see that F is the distribution function of the distance from an arbitrary fixed location in $\mathbb{R}^d$ to the nearest point of Ξ.

In a similar vein, using the typical point interpretation of Remark 2.Ä, we may interpret G as the distribution function of the distance from a typical point of Ξ to its nearest neighbour.

The J-function combines F and G in a clever way, where $J(r)$ is the ratio of the probability that the r-ball around a typical point is empty (i.e. contains no points of Ξ) to the probability that the r-ball around an arbitrary location is empty. Consequently, we may interpret values $J(r) \geq 1$ as indicating regularity (repulsiveness) and $J(r) \leq 1$ as indicating clustering (attractiveness) in the point process.

From an applied point of view, the F-function is a rather poor statistic for detecting departures from the Poisson model. The G-function is reasonably useful, but typically does not reveal more than the other descriptive functions (and sometimes clearly less). We focus on the J-function for this reason.

An important theoretical property of the J-function is that it can detect the "range of interactions" of Ξ.

Proposition 3.I (Van Lieshout–Baddeley)**.** *Let Ξ be a stationary point process on $\mathbb{R}^d$ with intensity $\lambda > 0$ whose conditional intensity $\lambda(\cdot \,|\, \cdot)$ exists.*[6]

(i) $F(r) = 1$ implies $G(r) = 1$ and

$$J(r) = \mathbb{E}\biggl(\frac{\lambda(0 \,|\, \Xi)}{\lambda} \;\Big|\; \Xi\bigl(B(0,r)\bigr) = 0 \biggr) \quad \text{for all } r \in A_J.$$

(ii) Suppose that $\lambda(0 \,|\, \xi) = \lambda(0 \,|\, \emptyset)$ for every $\xi \in \mathfrak{N}^*$ with $\xi(B(0,R)) = 0$. We refer to this property as "Ξ has (finite) interaction range $R > 0$". Then

$$J(r) = \frac{\lambda(0 \,|\, \emptyset)}{\lambda} \quad \text{for all } r \in A_J \text{ with } r \geq R.$$

Proof. (i) Note that the Campbell–Mecke theorem in connection with the GNZ-formula yields that for any $A \in \mathcal{N}^*$ and any bounded measurable $h\colon \mathbb{R}^d \times \mathfrak{N}^* \to \mathbb{R}_+$

$$\lambda \int_A \mathbb{E}h(x, \Xi_x^!)\, dx = \mathbb{E}\biggl(\int_A h(x, \Xi \setminus \{x\})\, \Xi(dx) \biggr) = \int_A \mathbb{E}\bigl(h(x,\Xi) \lambda(x \,|\, \Xi) \bigr)\, dx.$$

Since the density of a σ-finite measure is almost everywhere unique, we have

$$\lambda\, \mathbb{E}h(x, \Xi_x^!) = \mathbb{E}\bigl(h(x,\Xi) \lambda(x \,|\, \Xi) \bigr)$$

for Leb^d-a.e. $x \in \mathbb{R}^d$. We choose $h(x,\xi) := 1\{\xi(B(x,r)) = 0\}$. Using $\Xi_0^!$ as prototypal notation for a point process with distribution $\mathscr{L}\bigl(T_x^{-1}(\Xi_x^!)\bigr)$, which is the same for almost every $x \in \mathbb{R}^d$,[7] and using in a similar way $\lambda(0 \,|\, \xi)$ as prototypal notation for $\lambda(x \,|\, T_x(\xi))$ (we have not studied conditional intensities under the aspect of stationarity), we obtain that

$$\lambda\, \mathbb{P}\bigl(\Xi_0^!(B(0,r)) = 0\bigr) = \mathbb{E}\bigl(1\{\Xi(B(0,r)) = 0\} \lambda(0 \,|\, \Xi) \bigr).$$

Therefore

$$1 - G(r) = \mathbb{P}\bigl(\Xi_0^!(B(0,r)) = 0\bigr) = \mathbb{E}\biggl(\frac{\lambda(0 \,|\, \Xi)}{\lambda} 1\{\Xi(B(0,r)) = 0\} \biggr).$$

[6] This uses the general definition of the conditional intensity as the function $\lambda(\cdot \,|\, \cdot)$ that satisfies the GNZ-formula, mentioned in Subsection 2.5.2. Note that we have a simple explicit formula available only for Gibbs processes on compact subsets $\mathcal{W}$ of $\mathbb{R}^d$ (cf. Remark 3.B), which can never be a stationary process due to the limited window and its edge effects.

[7] Note that we have used $\Xi_0^!$ in this way many times before for stationary Ξ, last time when defining $G(r) := \mathbb{P}\bigl(\Xi_0^!(B(0,r)) \geq 1\bigr)$.

On the other hand

$$1 - F(r) = \mathbb{P}\big(\Xi(B(0,r)) = 0\big),$$

so that $F(r) = 1$ implies $G(r) = 1$, and if $F(r) < 1$, then

$$J(r) = \frac{1 - G(r)}{1 - F(r)} = \mathbb{E}\bigg(\frac{\lambda(0 \,|\, \Xi)}{\lambda} \;\bigg|\; \Xi\big(B(0,r)\big) = 0\bigg)$$

(ii) For $r \geq R$ the prerequisite yields that $\xi(B(0,r)) = 0$ implies $\lambda(0 \,|\, \xi) = \lambda(0 \,|\, \emptyset)$ for every $\xi \in \mathfrak{N}^*$. Therefore by part (i)

$$\begin{aligned} 1 - G(r) &= \mathbb{E}\bigg(\frac{\lambda(0 \,|\, \Xi)}{\lambda} 1\{\Xi(B(0,r)) = 0\}\bigg) \\ &= \mathbb{E}\bigg(\frac{\lambda(0 \,|\, \emptyset)}{\lambda} 1\{\Xi(B(0,r)) = 0\}\bigg) \\ &= \frac{\lambda(0 \,|\, \emptyset)}{\lambda}\big(1 - F(r)\big), \end{aligned}$$

which yields the statement. □

A further advantage of the theoretical J-function is that it can be (partially) computed for more models than many other descriptive functions. We give just two simple examples.

Example 3.J. (i) If Ξ is a Po_λ-process, then $\mathscr{L}(\Xi_0^!) = \mathscr{L}(\Xi)$. Therefore $G(r) = F(r) = 1 - \exp(-\lambda \alpha_d r^d) > 0$, whence

$$J(r) = 1 \quad \text{for every } r \in \mathbb{R}.$$

(ii) If Ξ is a Strauss$(R; \beta, \gamma)$-process,[8] then

$$J(r) = \frac{\beta}{\lambda} \geq 1 \quad \text{for every } r \in \mathbb{R} \text{ with } r \geq R.$$

This can be seen as follows. For any point pattern ξ with $\xi\big(B(0,R)\big) = 0$ we have

$$\lambda(0 \,|\, \xi) = \beta \gamma^{s_R(0;\xi)} = \beta;$$

see (3.2). Thus Ξ has interaction range R and Proposition 3.I(ii) implies that $J(r) = \beta/\lambda$. Proposition 3.E(i) implies further that

$$\lambda = \mathbb{E}\big(\lambda(0 \,|\, \Xi)\big) = \beta \mathbb{E}(\gamma^{s_R(0;\Xi)}) \leq \beta,$$

because $\gamma \leq 1$.

[8] This should strictly speaking be a Strauss process on all of $\mathbb{R}^d$. We have not introduced such an object, but the reader may think of Ξ as defined on an arbitrarily large compact window $\mathcal{W}$ by restricting a finite Strauss$(R; \beta, \gamma)$-process $\widetilde{\Xi}$ on a "considerably larger" window $\widetilde{\mathcal{W}} \supset \mathcal{W}$ (e.g. $\widetilde{\mathcal{W}} = \mathcal{W} + B(0, cR)$, where $c >> 1$). The important theoretical property of the Strauss$(R; \beta, \gamma)$-process on $\mathbb{R}^d$ used here is that its conditional intensity corresponds to that of a finite process, i.e. $\lambda(x \,|\, \xi) = \beta \gamma^{s_R(x;\xi)}$ for $x \in \mathbb{R}^d$ and $\xi \in \mathfrak{N}^*(\mathbb{R}^d)$.

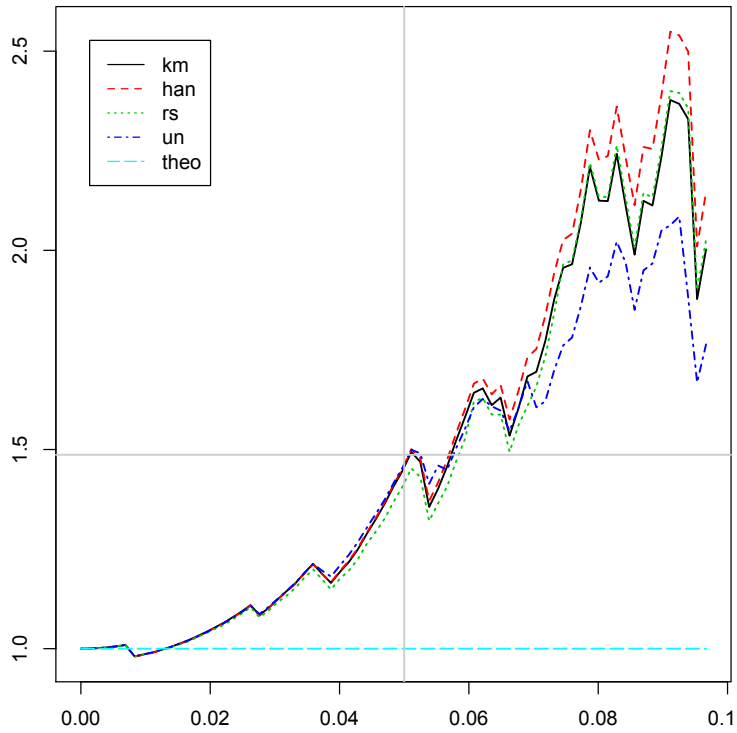

Figure 3.10: Estimated J-function for the Strauss point pattern in Figure 3.7 using various types of edge corrections, and using no edge correction (type "un").

Estimation of J is a somewhat unsatisfying topic. The best estimators known are simply plug-in estimators of the form

$$\widehat{J}(r) := \frac{1 - \widehat{G}(r)}{1 - \widehat{F}(r)},$$

where $\widehat{F}(r)$ and $\widehat{G}(r)$ are estimators of $F(r)$ and $G(r)$, respectively, that are based on similar kinds of edge corrections. Instead of going into details we describe a further possibility that was presented in Baddeley et al. (2000): it is actually possible to estimate the J-function without edge correction and to obtain results that are often a bit better than those for edge corrected estimators. Intuitively, the reason is that systematic errors in the uncorrected F- and G-estimators tend to cancel each other out.

Thus

$$\widehat{J}_{\mathrm{un}}(r) := \frac{1 - \widehat{G}_{\mathrm{un}}(r)}{1 - \widehat{F}_{\mathrm{un}}(r)},$$

where the natural uncorrected estimators for $F(r)$ and $G(r)$ are

$$\widehat{F}_{\mathrm{un}}(r) := \widehat{F}_{\mathrm{un},I}(r) := \frac{1}{|I|} \sum_{x \in I} 1\{D_x(\xi) \leq r\}$$

for a fine regular grid $I \subset \mathcal{W}$, and

$$\widehat{G}_{\mathrm{un}}(r) := \frac{1}{|\xi|} \sum_{x \in \xi} 1\{D_x(\xi \setminus \{x\}) \leq r\}.$$

Figure 3.10 shows estimates of the J-function for the Strauss point pattern in Figure 3.7 using various types of edge corrections (each time similar kinds for $\widehat{F}$ and $\widehat{G}$ as described above) and also using no edge correction (yielding $\widehat{J}_{\text{un}}$). The faint vertical line marks the interaction range of $R = 0.05$, the faint horizontal line marks $\beta/\lambda \approx 1.487$, which is the constant function value of the theoretical J-function of a Strauss$(0.05; 100, 0.3)$-process for $r \geq R$. The estimates seem quite reasonable up to around $r = 0.07$, and then quickly deteriorate. To obtain an idea about bias and standard deviation of the estimators, we have computed the mean and standard deviation of the estimated functions from 1000 independent simulations of a Strauss$(0.05; 100, 0.3)$-process. It can be seen that the uncorrected estimator does reasonably well. It is slightly

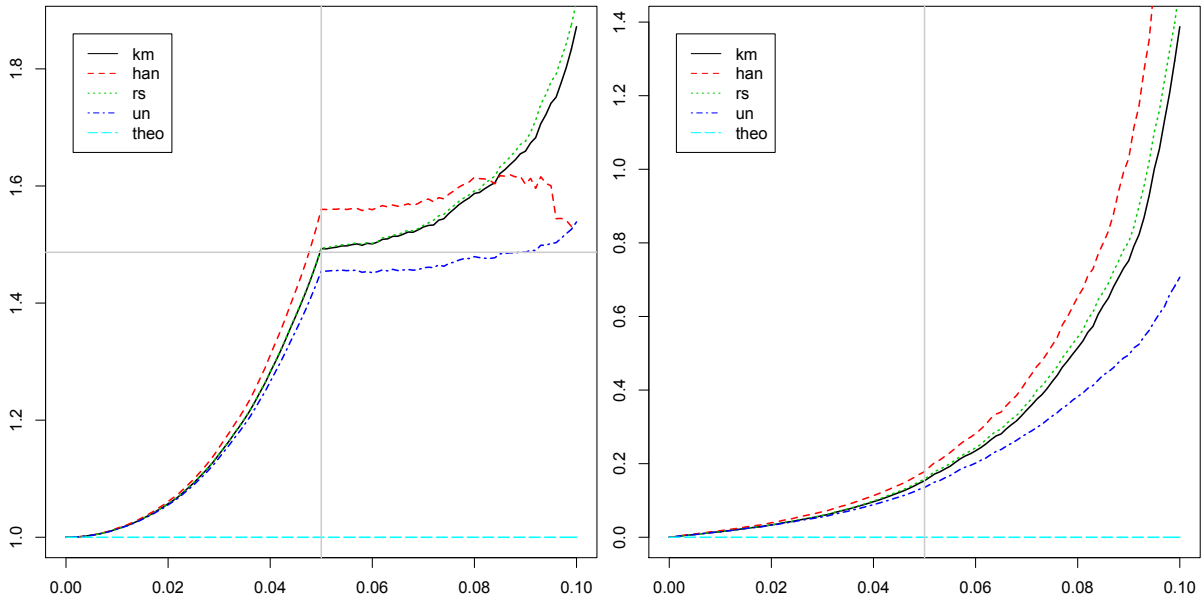

Figure 3.11: Monte Carlo mean (left) and standard deviation (right) of the estimated J-function using various types of edge corrections; based on 1000 simulations from Strauss$(0.05; 100, 0.3)$.

biased towards 1, which can be theoretically explained (see Baddeley et al., 2000), but it detects the interaction range as well as the other estimators, and remains rather more faithfully at a constant value for $r \geq R$. Furthermore it has clearly the smallest standard deviation of all the estimators.

We end the analysis with a word of warning. Although superficially the J-function and the pair correlation function are depicted in similar ways — and in particular both are compared with the constant value 1 of a Poisson process —, their interpretations are *very* different. Most prominently, for the J-function departures above 1 indicate regularity and departures below 1 indicate clustering, whereas for the pair correlation function it is the other way round. Another important point is that the J-function describes a cumulative quantity (like the K- and L-functions do), the ratio of probabilities of having points at *up*

to distance r, whereas the pair correlation function is non-cumulative and the meaning of its function values is more intuitive.

Chapter 4

CSR and other goodness-of-fit tests

We stay in the framework of Chapter 3. Recall in particular that our data consists of a single point pattern ξ observed in a compact window $\mathcal{W} \subset \mathbb{R}^d$.

A test for complete spatial randomness (CSR) is often the first step before any more serious inferential analysis starts. The term CSR, which is widely used in the various application fields of point pattern statistics, means nothing else but that the underlying point process Ξ is a homogeneous Poisson process with some intensity $\lambda > 0$.

If we fail to reject the null hypothesis of CSR in one or maybe several tests, then it is usually not worth doing further inferential analysis. So while the homogeneous Poisson process is a desirable object for the theoretical statistician, it is usuallly rather a disappointment for the data analyst.

4.1 CSR tests

Note that when testing for CSR, we are actually testing for two very different properties of Ξ at once:

(a) *Spatial homogeneity:* Ξ is (at least first-order) stationary, i.e. there is some $\lambda > 0$ such that $\mathbb{E}\Xi(A) = \lambda|A|$ for any $A \subset \mathcal{W}$. Assuming that given $\Xi(\mathcal{W}) = m$ the points of Ξ are i.i.d. (no point interactions), then spatial homogeneity means that they are uniformly distributed on $\mathcal{W}$.

(b) *No point interactions:* Given $\Xi(\mathcal{W}) = m$ the points of Ξ can be defined in such a way that they are i.i.d. The assumption is then usually that Ξ is a (possibly inhomogeneous) Poisson process.

Note that we cannot make any statements about the distribution of the total number of points of Ξ based on a single point pattern alone. In particular we cannot really distinguish between a Poisson and a Binomial process (say) although it is often clear from meta-information on the data whether the total number of points should be modeled as

random or not, the second case being far more rare in spatial settings.

In what follows we distinguish between CSR tests that rather target detecting spatial inhomogeneity and CSR tests that rather target detecting point interactions. *The distinction is based on intuitive considerations with regard to the construction of the tests and should not be taken too seriously!* In either category CSR may very well be rejected due the property of the other category (or typically a mixture of both properties).

4.1.1 Tests (rather) for spatial inhomogeneity

A classical approach is to subdivide $\mathcal{W}$ into k subregions of equal area, count the number of points in each subregion, and do a χ^2 goodness-of-fit test for uniformity of the k-valued distribution of subregion membership. This is known as *quadrat count method* and the subregions are known as *quadrats* even if they have a more complex shape. Using this procedure for spatial data is nowadays somewhat frowned upon by many (mathematical) statisticians, because the subdivision into quadrats is usually a very arbitrary decision, involving at the very least a choice of quadrat size (too small quadrats and too large quadrats give very small power; quadrats should reflect somehow the scale of inhomogeneity of the distribution) and reasonably also a choice of quadrat shape.

As a more modern alternative we consider *spatial Kolmogorov–Smirnov tests*. The name is maybe somewhat misleading as these are just ordinary univariate Kolmogorov–Smirnov tests, but with respect to the distribution of a spatial covariate. The procedure runs as follows.

Spatial Kolmogorov–Smirnov test

1. Choose a one-dimensional spatial covariate (if inferential statements are to be made: without looking at the data!). It should be a (more or less) continuous quantity that is defined on all of $\mathcal{W}$. Examples include some coordinate (e.g. x- or y-coordinate in 2-d), the distance from a prespecified point, or some kind of additional information that has been mapped for the whole window (e.g. nitric oxide content of soil).
2. Evaluate the covariate at the data points ξ and form the empricial cumulative distribution function (e.c.d.f.) $\widehat{F}_n$ of these values.
3. Choose a fine grid I in $\mathcal{W}$, evaluate the covariate at each location of I, and form the e.c.d.f. of these values. This is an approximation of the theoretical or "predicted" c.d.f. F.
4. Perform a Kolmogorov–Smirnov test with $\widehat{F}_n$ and F. Under the assumption that the points $S_1, \ldots, S_m$ of $\Xi|_{\mathcal{W}}$ are i.i.d. given $\Xi(\mathcal{W}) = m$, this test examines the null hypothesis whether the covariate evaluated at S_i follows the theoretical distribution of the covariate evaluated at a random vector that is uniformly distributed on $\mathcal{W}$.

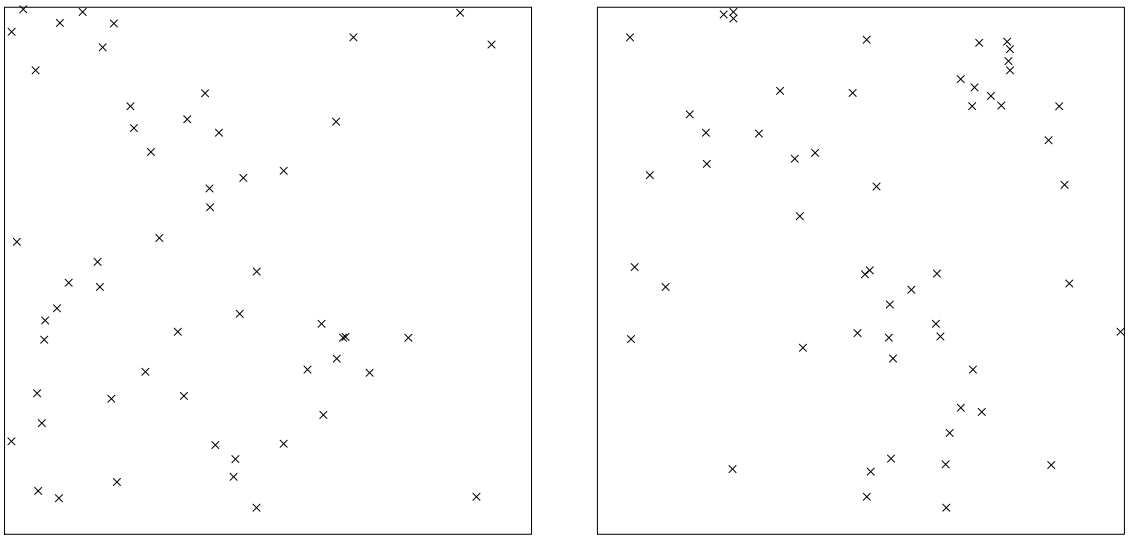

Figure 4.1: Point patterns simulated from inhomogeneous Poisson process distributions with an exponential trend in the x-coordinate (left) and with a Gaussian mixture intensity function consisting of three "hills" (right); compare Figure 4.2. The realized numbers of points were 55 and 56, respectively.

We recall here the gist of Kolmogorov–Smirnov test theory, omitting certain details and all the proofs. The latter may be found for example in Dümbgen (2010), Section 3.2 and Chapter 8. Let $X_1, \ldots, X_n$ be i.i.d. random variables following some distribution function F. Write

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq t\}$$

for the e.c.d.f., and consider the (two-sided) Kolmogorov–Smirnov statistic given by

$$D_n := \sup_{t \in \mathbb{R}} \big| \widehat{F}_n(t) - F(t) \big| .$$

The key theorem is the following.

Theorem 4.A. *Among all continuous distribution functions F, the distribution of D_n is the same.*

The distribution function G_n of D_n and its α-quantiles $q_n(\alpha)$ can be computed numerically and the values can be found in tables. We then obtain

$$C_{n,\alpha} := \big\{ (t,y) \in \mathbb{R} \times [0,1] \colon y \in \big[\widehat{F}_n(t) - q_n(1-\alpha) \,,\, \widehat{F}_n(t) + q_n(1-\alpha) \big] \big\}$$

as a (simultaneous) confidence band for $\widehat{F}_n$ and

$$1 - G_n(D_n)$$

as a p-value for the null hypothesis that the true distribution function is F.

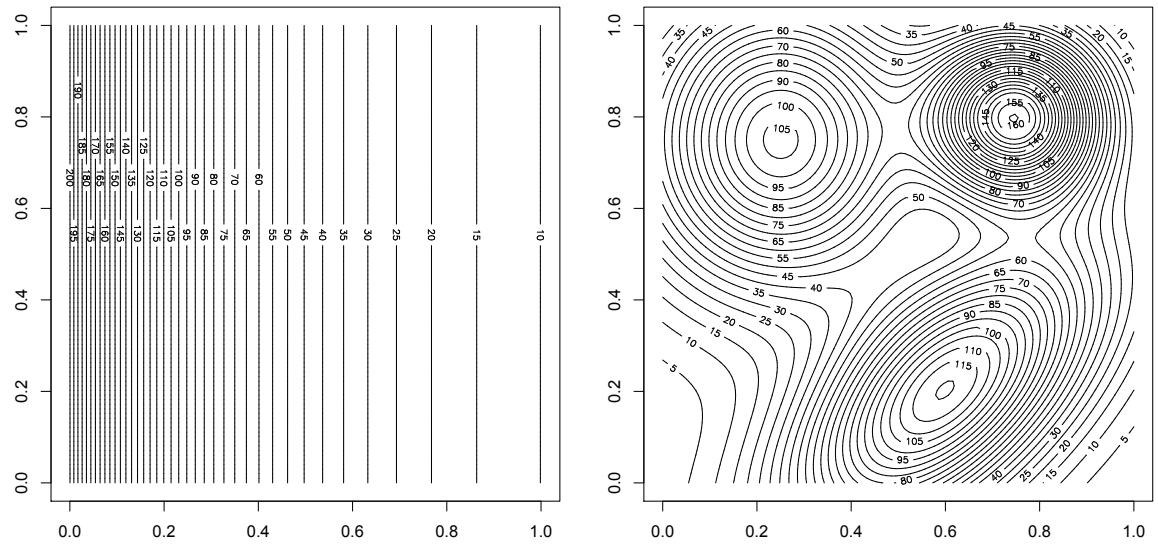

Figure 4.2: Contour plots of the intensity functions used for Figure 4.1.

If n is large, then instead of the distribution of D_n the limiting distribution in the following theorem is used.

Theorem 4.B. *For any continuous distribution function F we have*

$$\sqrt{n} D_n \xrightarrow{\mathscr{D}} \sup_{t\in[0,1]} |B(t)| \quad \text{as } n \to \infty,$$

where $(B(t))_{t\in[0,1]}$ is a Brownian bridge.

The distribution of $\sup_{t\in[0,1]} |B(t)|$, which again can be computed numerically,[1] is known as the Kolmogorov distribution.

We apply the spatial Kolmogorov–Smirnov test to the two simulated data sets in Figure 4.1. They have been generated from inhomogeneous Poisson process distributions with intensities shown in Figure 4.2. The results are summarized in Figures 4.3 and 4.4. Unfortunately the Kolmogorov–Smirnov test functions in R and spatstat do not provide us with the abovementioned confidence band.

As covariates we examine for the first point pattern the x- and y-coordinates. As we might have expected, the p-value for the x-coordinate is very small. Since in real data examples it is unrealistic to assume that we get the direction of a supposed trend exactly right, we also consider the function $f(x, y) := x + y$ as a spatial covariate, which corresponds to the (scaled) displacement along the bisecting line. Again the p-value was very small. That it was this small the second time was probably a bit due to a lucky realization of our point process. It can be seen from the third plot in Figure 4.3 that the

[1] We have in fact that $\mathbb{P}(\sup_t |B(t)| > z) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2z^2)$ for every $z > 0$. See Dümbgen (2010), Satz 8.14, for a proof.

e.c.d.f. fits the theoretical distribution function rather well up to $t = 1$, but then quickly deteriorates due to the fact that the upper right triangle of the observation window hardly contains any points: in the end the D_n statistic is almost as large as in the case of the x-coordinate.

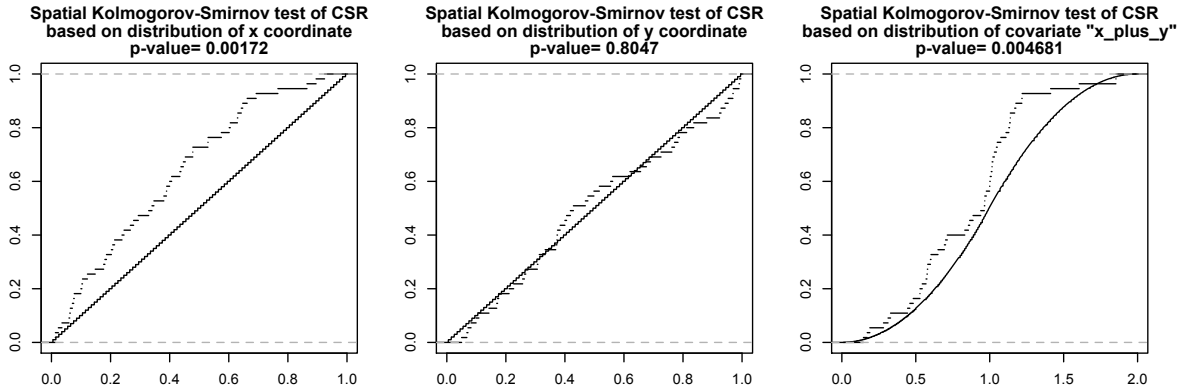

Figure 4.3: Empirical and theoretical distribution functions for the first point pattern in Figure 4.1 using x-coordinate, y-coordinate, and the function $f(x, y) := x + y$ as spatial covariates.

For the second point pattern the x- and y-coordinates are maybe somewhat suspicious, but not significantly so. In the third panel, we imitate the case of additional map information being available on $\mathcal{W}$: Suppose our point pattern describes a plant species that prefers higher altitudes, and suppose further that the altitude profile was given by the right panel in Figure 4.2. By using the altitude as a spatial covariate we would arrive at the rightmost plot in Figure 4.4. Since we have *simulated* the point pattern using the very same intensity, it is of course no surprise that the Kolmogorov–Smirnov test is highly significant. Nevertheless this little thought experiment illustrates the use of a map information covariate rather well.

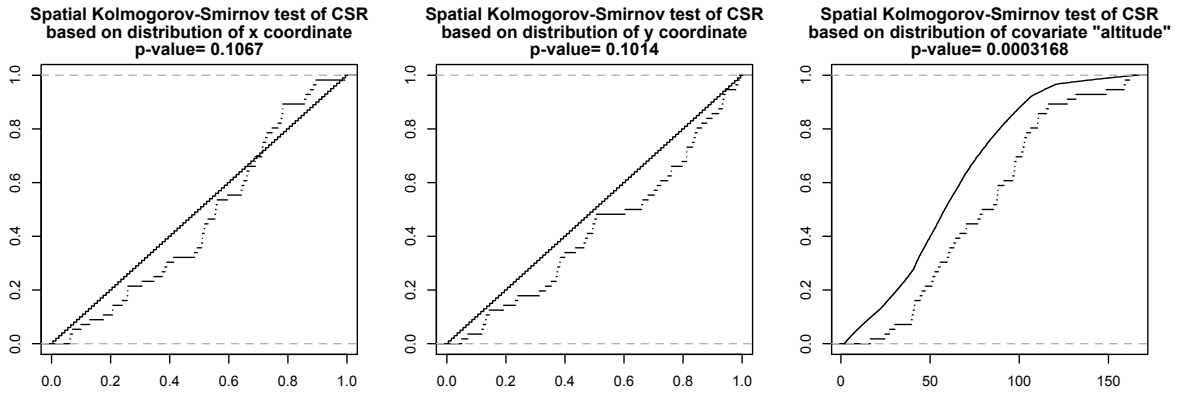

Figure 4.4: Empirical and theoretical distribution functions for the second point pattern in Figure 4.1 using x-coordinate, y-coordinate, and the (artificial) altitude as covariate.

4.1.2 Tests (rather) for point interactions

In Chapter 3 we have searched for spatial point interactions exploratively by looking at descriptive functions like the L-, J- or pair correlation function. It seems therefore natural to base inference about point interactions on the same functions. Unfortunately, as briefly mentioned in Chapter 3, very little can be said about the distribution of the proposed estimators for these functions. We usually can make statements about unbiasedness or ratio-unbiasedness, and for some descriptive functions, there are approximation formulae for the variance, mostly under Poisson process assumptions; see for example the approximations of the variance of $\widehat{K}_{\mathrm{iso}}$ under Poisson assumptions by Ripley and by Lotwick and Silverman, implemented in the `spatstat` function `Kest` via the option `var.approx=TRUE`. In the case of K-function estimators there are also statements about asymptotic normality, so that an asymptotic test using a theoretical distribution is possible.

However, by far the most popular and also reliable method for making inference about point interactions is performing Monte Carlo tests. While computer simulation may seem a little unsatisfactory to some from the point of view of theoretical statistics, it has the advantage that it can provide us with tests that exactly observe a given significance level $\alpha > 0$, which is something the asymptotic tests simply cannot. The procedure runs as follows.

Pointwise $(1-\alpha)$ *Monte Carlo envelopes*

1. Choose a descriptive function $\Phi\colon A \to \mathbb{R}_+$, $A \subset \mathbb{R}$, among those presented in Chapter 3 (e.g. L-function), and choose an estimation method for this function (e.g. plug-in, with isotropic edge-correction). Denote by $A' \subset A$ the subset of the domain that is of interest.
2. Given a data point pattern ξ, compute the estimate $\widehat{\Phi}_\xi$ of Φ on A'.
3. Choose $n \in \mathbb{N}$ and $k \in \{1, 2, \ldots, \lfloor n/2 \rfloor\}$ such that $\frac{2k}{n+1} = \alpha$ (assuming $\alpha \in (0,1)$ is rational). Simulate n point patterns $\eta_1, \ldots, \eta_n$ from the $\mathrm{Po}_{\xi(\mathcal{W})/|\mathcal{W}|}$-distribution on $\mathcal{W}$ and compute their estimates $\widehat{\Phi}_{\eta_i}$, $1 \le i \le n$, on A'.
4. Denoting for any r by $\widehat{\Phi}_{(j)}(r)$ the value that has rank j among $\widehat{\Phi}_{\eta_1}(r), \ldots, \widehat{\Phi}_{\eta_n}(r)$, the set
$$\mathrm{Env}_{n,\alpha} := \big\{(r,y) \in A' \times \mathbb{R}_+ \colon y \in \big[\widehat{\Phi}_{(k)}(r), \widehat{\Phi}_{(n-k+1)}(r)\big]\big\} \tag{4.1}$$
is a pointwise non-rejection region for CSR at level α. This means that for any $r_0 \in A'$, we have $\mathbb{P}\big((r_0, \widehat{\Phi}_{\mathrm{H}}(r_0)) \notin \mathrm{Env}_{n,\alpha}\big) = \alpha$ if H is a $\mathrm{Po}_{\xi(\mathcal{W})/|\mathcal{W}|}$-process on $\mathcal{W}$.

We refer to $\mathrm{Env}_{n,\alpha}$ briefly as *simulation envelope* at level α. *Care should be taken not to confound a non-rejection region with a confidence interval!* The set $\mathrm{Env}_{n,\alpha}$ is often used as a descriptive statistic in order to obtain an impression of the typical variation in the estimates of Φ based on a Poisson process. If we can prespecify a single value

of r_0 at which to check if $(r_0, \widehat{\Phi}_\xi(r_0)) \in \mathrm{Env}_{n,\alpha}$, then we obtain an exact CSR test at level α. However, it is often in applications not clear what a good value of r_0 should be before observing the data. Multiple testing with Holm correction might be performed but it should be noted that there is typically a strong positive correlation between the test results for values of r that are not too far away from each other, and therefore such a correction is quite likely to be very conservative.

A better way is to perform the Monte Carlo test based on a statistic that takes a whole range of values of r_0 into account. A very convenient choice is the maximal absolute difference from the theoretical descriptive function, among other things because it allows again for a graphical representation. Write

$$T(\eta) := T_\Phi(\eta) := \sup_{r \in A'} \big| \widehat{\Phi}_\eta(r) - \Phi_{\mathrm{CSR}}(r) \big|$$

for $\eta \in \mathfrak{N}^*(\mathcal{W})$, where Φ_{CSR} denotes the theoretical Φ-function under CSR. This statistic is only sensible for descriptive functions that have more or less stable sample variance over the considered domain; our favourite function in this respect is the L-function.

Simultaneous $(1-\alpha)$ *Monte Carlo envelopes*

1. Choose a descriptive function $\Phi\colon A \to \mathbb{R}_+$, $A \subset \mathbb{R}$, among those presented in Chapter 3 (e.g. L-function), and choose an estimation method for this function (e.g. plug-in, with isotropic edge-correction). Denote by $A' \subset A$ the subset of the domain that is of interest.
2. Given a data point pattern ξ, compute the statistic $T(\xi)$.
3. Choose $n \in \mathbb{N}$ and $k \in \{1, 2, \dots, n\}$ such that $\frac{k}{n+1} = \alpha$ (assuming $\alpha \in (0,1)$ is rational). Simulate n point patterns $\eta_1, \dots, \eta_n$ from the $\mathrm{Po}_{\xi(\mathcal{W})/|\mathcal{W}|}$-distribution on $\mathcal{W}$ and compute their statistic values $T(\eta_i)$, $1 \le i \le n$.
4. Denoting by $T_{(1)}, \dots, T_{(n)}$ the ordered $T(\eta_i)$-values, we obtain

   $$\mathrm{Env}^*_{n,\alpha} := \big\{ (r,y) \in A' \times \mathbb{R}_+ \colon y \in \big[\Phi_{\mathrm{CSR}}(r) - T_{(n-k+1)}, \Phi_{\mathrm{CSR}}(r) + T_{(n-k+1)} \big] \big\}$$

   as a simultaneous non-rejection region for CSR at level α. This means that we have $\mathbb{P}\big(\exists r \in A' \colon (r, \Phi_{\mathrm{H}}(r)) \notin \mathrm{Env}^*_{n,\alpha}\big) = \alpha$ if H is a $\mathrm{Po}_{\xi(\mathcal{W})/|\mathcal{W}|}$-process on $\mathcal{W}$.

Note the different role of k in this procedure: while the significance level for the pointwise envelope was $2k/(n+1)$ it is now $k/(n+1)$.

For more general descriptive functions that do not necessarily have stable sampling variance a statistic of the form

$$T_q(\eta) := T_{q,\Phi}(\eta) := \int_{A'} w(r) \big| \widehat{\Phi}_\eta(r) - \Phi_{\mathrm{CSR}}(r) \big|^q \, dr,$$

where $q \ge 1$ and $w\colon A' \to \mathbb{R}_+$ is some (typically decreasing) weight function, may be

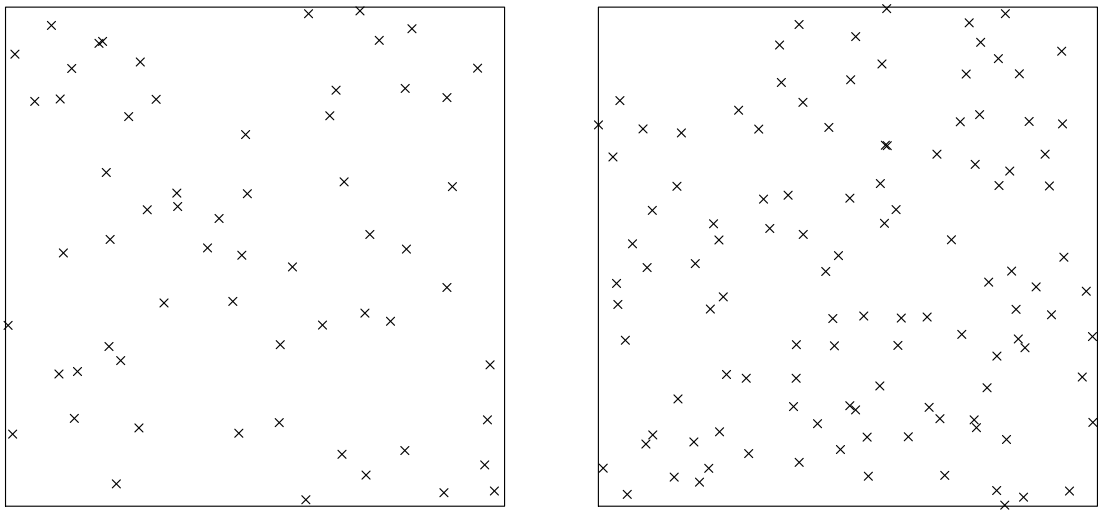

Figure 4.5: Two point patterns simulated from the Strauss$(0.05; 100, 0.3)$- and the Strauss$(0.05; 200, 0.3)$-distributions. The first pattern is the same as in Figure 3.7. The realized numbers of points were 62 and 114, respectively.

appropriate. Of course the weight function may also be introduced in the supremum statistic T.

A further modification is that we may want to allow A' to depend on the point pattern η. This is particularly useful (in theory even necessary) for the J-function, where the upper limit of the interval where J can be estimated depends on the largest ball of empty space that we can find in $\mathcal{W}$, and hence on η. From our construction it is clear that a point pattern dependent A' in the definition of T and T_q changes nothing about the validity of the Monte Carlo procedure described.

We illustrate pointwise and simultaneous simulation envelopes once more for the Strauss point pattern in Figure 3.7 (shown again in the left panel of Figure 4.5). Figure 4.6 shows the pointwise 95% envelopes for L-, J- and pair correlation functions, and the left panel of Figure 4.7 shows the simultaneous 95% envelope as described above for the L-function. We can see that the deviation from the theoretical function in Figure 4.6 is clearly still suspicious if we take the sample variation into account, and also significant if we have fixed a value of t_0 around 0.05 in advance. When doing a strict simultaneous test the result is a very close call, not actually visible from the left plot in Figure 4.7. Looking at the exact numbers, it is seen that the kink at about $r = 0.0508$ lies actually outside the simulation envelope. Therefore we can reject CSR. However, it should be noted that due to the randomized nature of the Monte Carlo test, it is not at all certain that a second run of the test will lead to the same decision. This may be seen as a certain disadvantage of Monte Carlo testing, mainly because it provides some room for manipulating test results: if a certain non-rejection result is a close call, a (shady) statistician can rerun the test a

few times until the Monte Carlo method rejects the null hypothesis and present only the last result.

For comparison we have also simulated from the Strauss$(0.05; 200, 0.3)$-distribution; see the right panel of Figure 4.5. That is, we have just replaced the parameter $\beta = 100$ by $\beta = 200$. As we would have expected the rejection of CSR is much clearer here, and it is very unlikely that a second run of the test would give a different result.

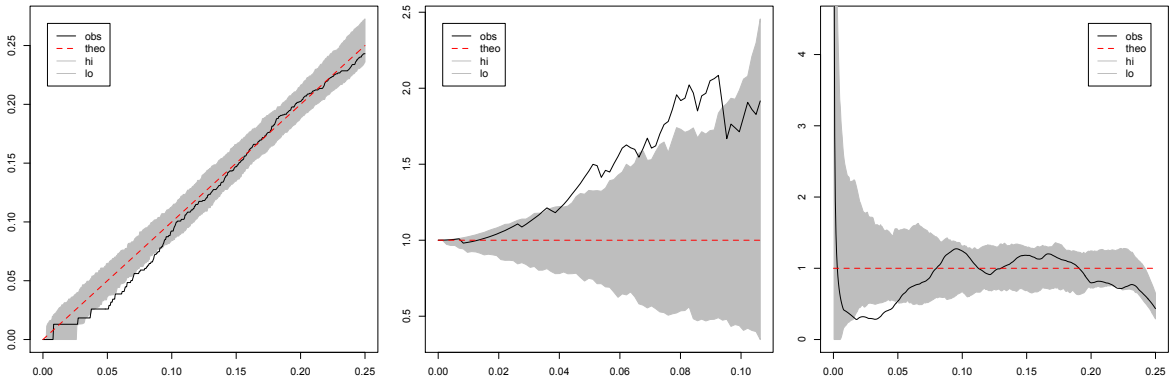

Figure 4.6: Estimates of L-, J-, and pair correlation functions for the Strauss point pattern in the left panel of Figure 4.5, together with pointwise 95% envelopes for CSR based on 199 simulations.

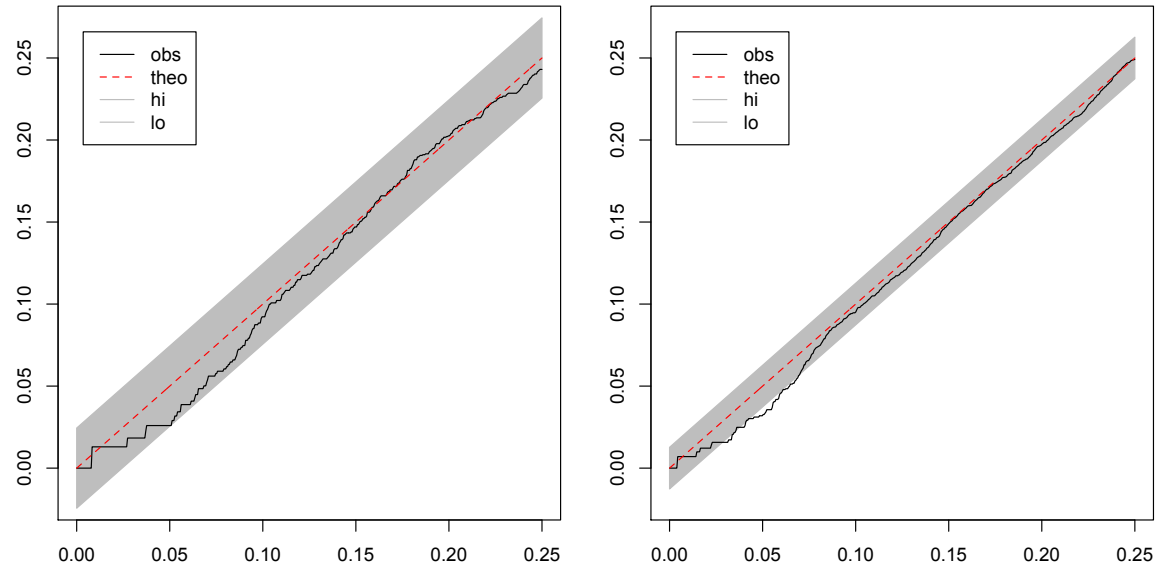

Figure 4.7: Estimates of the L-functions for the Strauss point patterns in Figure 4.5, together with simultaneous 95% envelopes for CSR based on 199 simulations. Although not clearly visible, the estimated function on the left hand side leaves the simulation envelope at around 0.05.

4.2 General goodness-of-fit tests

In principle the methods from Subsection 4.1.2 may be used for testing not only for a homogeneous Poisson process, but for *any* point process distribution from which we can simulate. A minor problem for the simultaneous simulation envelopes is that typically for non-Poisson process distributions, we do not have a formula for the theoretical Φ-function available (as a replacement of Φ_{CSR} in the simultaneous Monte Carlo procedure). However, we may always estimate it as the average Φ-estimate from further $\tilde{n}$ Monte Carlo samples (different from the ones used to construct the envelope), essentially by the fact that our estimators are approximately unbiased for $\tilde{n}$ not too small.

A bigger problem is often the source of the null hypothesis distribution. The most typical application of general goodness-of-fit tests is when we have fitted a certain point process model and would like to check the goodness-of-fit of this model. If this test is based on the same data as the model fitting, it will be highly conservative, indicating a reasonable fit, when in fact the null hypothesis "should be rejected". We postpone the discussion of this problem to Section 5.4, where we consider model diagnostics from a more general perspective.

Chapter 5

Parametric model fitting

In this section we fit exponential parametric families $(P_\theta)_{\theta\in\Theta}$ of Gibbsian point process distributions to a given data point pattern ξ. The parameter vector θ has components that control point interactions and/or the influence of spatial covariates. The latter are functions of the location $x \in \mathcal{W}$ that are known (or at least in principle obtainable) on the whole observation window $\mathcal{W}$. We do not treat here the case of "point-specific" covariates, which informally may be described as additional information attached to the points of ξ that does not make sense at non-point locations. Typical examples for a point pattern that describes tree positions are the nitrogen oxide concentration in the soil (which is a spatial covariate, in principle measurable whether there is a tree or not), and the diameter at breast height (which clearly does not make sense at non-tree locations).

The following methods are commonly used for model fitting.

- ◇ **Maximum likelihood:** All-purpose method in statistics, which in general gives estimators with good properties; in the present context for most models tedious, because the likelihood can only be evaluated numerically and accuracy comes at a high computational cost.
- ◇ **Maximum pseudolikelihood:** Replacement for maximum likelihood which is easier to deal with theoretically and computationally, but is often worse than maximum likelihood.
- ◇ **Takacs–Fiksel method:** Estimation based on equality of the densities in the GNZ- and the Campbell–Mecke equations, leading in the stationary case to

$$\frac{1}{|\mathcal{W}|}\mathbb{E}_\theta\Bigg(\sum_{x\in\Xi\cap\mathcal{W}} h(x,\Xi\setminus\{x\})\Bigg) = \mathbb{E}_\theta\big(h(y,\Xi)\lambda_\theta(y\,|\,\Xi)\big) \tag{5.1}$$

  for arbitrary fixed $y \in \mathcal{W}$. For some appropriate choice of functions h_i, the estimate of θ is obtained by bringing estimates of the left and right hand side of (5.1) as close as possible together.
- ◇ **Minimum contrast method:** Based on a somewhat similar idea. Choose a de-

scriptive function $\Phi\colon I \to \mathbb{R}$ (e.g. among those introduced in Chapter 3), where $I \subset \mathbb{R}$ is an interval. Estimate θ by minimizing

$$\Delta(\theta) = \int_I \bigl|\widehat{\Phi}(r) - \Phi_\theta(r)\bigr|^\beta \, dr$$

for some $\beta > 0$, where $\widehat{\Phi}$ is the estimated Φ-function for the data point pattern and Φ_θ is the theoretical function under distribution P_θ (often approximated by simulation).

After this brief overview we focus on the maximum likelihood and pseudolikelihood methods. We restric ourselves to exponential families of Gibbs processes, which are introduced in the next section.

5.1 Exponential families

In what follows we consider point processes defined on a compact set $\mathcal{X}$.

Definition. A probability model of the form $(P_\theta)_{\theta\in\Theta}$, where[1] $\Theta \subset \mathbb{R}^k$ and P_θ are distributions of Gibbs point processes on $\mathcal{X}$, is called *exponential family* if the density f_θ of P_θ takes the form

$$f_\theta(\xi) = c(\theta)\, b(\xi) \exp(\theta^\top T(\xi)) \quad \text{for every } \xi \in \mathfrak{N}^*. \tag{5.2}$$

We refer to θ as *canonical parameter* and to T as *canonical sufficient statistic* (compare the usual theory of exponential families on $\mathbb{R}^d$).

For an exponential family the conditional intensities are given by

$$\lambda_\theta(x \,|\, \xi) = B(x;\xi) \exp\bigl(\theta^\top S(x;\xi)\bigr),$$

where $B(x;\xi) = b(\xi \cup \{x\})/b(\xi)$ and $S(x,\xi) = T(\xi \cup \{x\}) - T(\xi)$. Note that hereditarity implies that $b(\xi \cup \{x\}) = 0$ if $b(\xi) = 0$ (since the other factors in Equation (5.2) must be positive as long as $\theta \in \mathbb{R}^k$).

The factor $b(\xi)$ describes our null model under the assumption that there is no influence of the statistic $T(\xi)$. We convene to normalize b so that it is a density. Hence for $\theta = 0 \in \mathbb{R}^k$, we have $f_0(\xi) = b(\xi)$ (and $c(0) = 1$). Typically our null model will be the homogeneous Poisson process, which means that $b(\xi) = 1$ for every ξ.

Example 5.A.
(i) Inhomogeneous Poisson process with log-affine intensity: $P_\theta := \mathrm{Po}(\lambda_\theta)$ with

[1] We sometimes allow that certain components of θ take the value $-\infty$ for nonnegative T; these cases typically require some extra attention.

$\lambda_\theta(x) = B(x)\exp(\theta^\top S(x))$ on $\mathcal{X}$, where B and S are chosen in such a way that λ_θ is integrable. Then

$$f_\theta(\xi) = \exp\Big(-\int \big(\lambda_\theta(x) - 1\big)\, dx\Big)\Big(\prod_{x\in\xi} B(x)\Big)\exp\Big(\theta^\top\big(\textstyle\sum_{x\in\xi} S(x)\big)\Big)$$

is in exponential family form with

$$b(\xi) = \prod_{x\in\xi} B(x) \quad \text{and} \quad T(\xi) = \sum_{x\in\xi} S(x);$$

note that also $c(\theta)$ is known here explicitly.

(ii) Homogeneous Strauss process: We try to cast Strauss$(R;\beta,\gamma)$ in exponential family form, where θ is a transformation of at least some of the parameters R, β and γ. We do not succeed for the parameter R, but obtain for fixed $R > 0$

$$f_\theta(\xi) = c(\theta)\exp\big(\log(\beta)|\xi| + \log(\gamma)s_R(\xi)\big),$$

where $\theta = (\log(\beta), \log(\gamma))^\top$ is the canonical parameter, and $T(\xi) = (|\xi|, s_R(\xi))^\top$ is the canonical sufficient statistic. Here we may want to allow $\theta_2 := \log(\gamma) = -\infty$. Any parameter such as R that does not fit the exponential form is called an *irregular parameter*. It is typical for interaction ranges to be irregular parameters. In what follows we treat only regular parameters in the general estimation theory and deal with irregular parameters separately.

(iii) Inhomogeneous Strauss process: Bringing together examples (i) and (ii), we can see in a similar way as above that for fixed $R > 0$ the distributions Strauss$(R;\beta_\psi(\cdot),\gamma)$, where $\beta_\psi(x) = B(x)\exp(\psi^\top S(x))$ is integrable for $\psi \in \Psi \subset \mathbb{R}^{k-1}$, form an exponential family with

$$\theta = \begin{pmatrix} \psi \\ \log(\gamma) \end{pmatrix} \in \Theta = \Psi \times (-\infty, 0] \quad \text{and} \quad T(\xi) = \begin{pmatrix} \sum_{x\in\xi} S(x) \\ s_R(\xi) \end{pmatrix} \in \mathbb{R}^k.$$

As before we may want to allow $\log(\gamma) = -\infty$.

(iv) Area-interaction process: In a similar vein, we can understand AIP$(R;\beta_\psi(\cdot),\gamma)$, where β_ψ is as in example (iii), as an exponential family with irregular parameter R and

$$\theta = \begin{pmatrix} \psi \\ \log(\eta) \end{pmatrix} \in \Theta = \Psi \times \mathbb{R} \quad \text{and} \quad T(\xi) = \begin{pmatrix} \sum_{x\in\xi} S(x) \\ -|U_R(\xi)| \end{pmatrix} \in \mathbb{R}^k.$$

◇

We conclude this section by taking a brief look at the log-likelihood function of an exponential family, which for a data point pattern $\xi \in \mathfrak{N}^*$ has the form

$$\log f_\theta(\xi) = \theta^\top T(\xi) + \log(b(\xi)) + \log(c(\theta)) \tag{5.3}$$

for $\theta \in \Theta$. This is the simplest form of the log-likelihood, for the situation where $\mathcal{X} = \mathcal{W}$, i.e. there are no edge effects to consider.

Our goal is maximization of (5.3) in θ. We can see that the first summand is unproblematic and the second summand is unnecessary since it does not depend on θ. The last summand, however, can be very annoying. It involves the integral

$$\frac{1}{c(\theta)} = \int_{\mathfrak{N}^*} b(\xi) \exp(\theta^\top T(\xi)) \ \mathrm{Po}_1(d\xi), \tag{5.4}$$

which in most cases cannot be evaluated analytically and is difficult to evaluate efficiently by numerical approaches because of the large space $\mathfrak{N}^*$. This is why we investigate maximum pseudolikelihood as a replacement of the maximum likelihood method in the next section.

5.2 Maximum pseudolikelihood

This method was first devised for the case of Markov lattice process in Besag (1974). In a lattice process finitely many random variables $(X_i)_{i\in I}$ are attached to the nodes of a regular grid I. The random variables X_i (typically) depend on each other in a way that reflects their spatial arrangement on the grid. The idea of maximum pseudolikelihood in this situation is the following: instead of using the joint density of $(X_i)_{i\in I}$ at $(x_i)_{i\in I}$ we use the density of independent random variables $(\widetilde{X}_i)_{i\in I}$, where $\widetilde{X}_i \sim \mathscr{L}\big(X_i \,|\, X_j = x_j, j \neq i\big)$. When translating this idea into the world of point processes, we obtain a Poisson process $\widetilde{\Xi}$ instead of the lattice process $(\widetilde{X}_i)_{i\in I}$ and we obtain the conditional intensity $\lambda(x \,|\, \xi)$ instead of the density of the conditional distribution $\mathscr{L}\big(X_i \,|\, X_j = x_j, j \neq i\big)$.

Recalling that the density of a $\mathrm{Po}(\lambda(x)dx)$-process on $\mathcal{W}$ is given by

$$\exp\Big(-\int_{\mathcal{W}} \big(\lambda(x) - 1\big) \ dx\Big) \prod_{x\in\xi} \lambda(x), \tag{5.5}$$

we are lead to the following definition. Note that we tacitly leave out the summand $|\mathcal{W}|$, which does not change the maximization problem in θ.

Definition. For a probability model $(P_\theta)_{\theta\in\Theta}$ of Gibbs processes on $\mathcal{W}$ with conditional intensities $\lambda_\theta(\cdot \,|\, \xi)$, we define the *log-pseudolikelihood function* by

$$\mathrm{PL}(\theta) := \mathrm{PL}(\theta; \xi) := \sum_{x\in\xi} \log \lambda_\theta(x \,|\, \xi) - \int_{\mathcal{W}} \lambda_\theta(x \,|\, \xi) \ dx \tag{5.6}$$

for $\theta \in \Theta$.

Definition. We define the *maximum pseudolikelihood estimator (MPLE)* as

$$\hat{\theta} := \hat{\theta}_{\mathrm{MPL}} := \operatorname*{arg\,max}_{\theta\in\Theta} \mathrm{PL}(\theta).$$

(So far, we do not know anything about existence and uniqueness, so that $\hat{\theta}$ is a set which may be empty; if $\hat{\theta}$ is a one-point set, it is identified with its only element.)

Once again the above definitions implicitly assume that $\mathcal{X} = \mathcal{W}$. We will discuss suitable corrections in the presence of edge effects in Subsection 5.2.4.

In the exponential family case, where for simplicity we assume now that $b(\xi) = 1$, i.e. our null model is the standard Poisson process, we have

$$\mathrm{PL}(\theta) = \theta^\top \Big(\sum_{x \in \xi} S(x;\xi) \Big) - \int_{\mathcal{W}} \exp\big(\theta^\top S(x;\xi)\big)\, dx \tag{5.7}$$

for every $\theta \in \Theta$.

For an inhomogeneous Poisson process, intensity and conditional intensity coincide, so that the pseudolikelihood function and the likelihood function are the same. Hence we deal in what follows implicitly also with maximum likelihood estimation for inhomogeneous Poisson process families of the form given in Example 5.A(i). Regarding the properties of the estimators studied below stronger results can be obtained in the Poisson family case.

5.2.1 Existence and uniqueness of MPLEs

For the sake of simplicity we require the function $S(\cdot;\xi)$ to be bounded on $\mathcal{W}$ for our data point pattern ξ. This requirement is satisfied for every $\xi \in \mathfrak{N}^*$ in the families of Example 5.A, provided that $S(\cdot)$ is bounded in the inhomogeneous families.

Proposition 5.B. *Suppose that $S(\cdot;\xi)$ is bounded on $\mathcal{W}$. Then*

(i) PL *is $\mathbb{R}$-valued and concave.*

(ii) If

$$\int_{\mathcal{W}} S(x;\xi) S(x;\xi)^\top \, dx$$

is positive definite, then PL *is strictly concave. If $\Theta \subset \mathbb{R}^k$ is convex, it follows that there is at most one maximizer of* PL *on Θ.*

(iii) If for every $u \in \mathcal{S}^{k-1} := \{y \in \mathbb{R}^k \colon \|y\| = 1\}$ we have either

(α) $\big|\{x \in \mathcal{W} \colon u^\top S(x;\xi) > 0\}\big| > 0$ or

(β) $u^\top \big(\sum_{x \in \xi} S(x;\xi)\big) < 0$,

then PL *is coercive in the sense that*

$$\lim_{\substack{\|\theta\| \to \infty \\ \theta \in \Theta}} \mathrm{PL}(\theta) = -\infty.^2$$

If $\Theta \subset \mathbb{R}^k$ is closed, it follows that there is at least one maximizer of PL *on Θ.*

[2] If Θ is bounded, we make no statement here, and the next statement is still true because every $\mathbb{R}$-valued concave function is continuous.

Proof. Using the boundedness of $S(\cdot;\xi)$ to ensure that we may interchange differentiation and integration, we obtain as gradient and Hessian matrix

$$\nabla \operatorname{PL}(\theta) = \sum_{x\in\xi} S(x;\xi) - \int_{\mathcal{W}} \exp(\theta^\top S(x;\xi))\, S(x;\xi)\, dx \tag{5.8}$$

and

$$D^2 \operatorname{PL}(\theta) = -\int_{\mathcal{W}} \exp(\theta^\top S(x;\xi))\, S(x;\xi)\, S(x;\xi)^\top\, dx. \tag{5.9}$$

(i) The Hessian is negative semidefinite, because the integrand in (5.9) is always positive semidefinite. Thus PL is concave.

(ii) If $\int_{\mathcal{W}} S(x;\xi)S(x;\xi)^\top\, dx$ is positive definite, we have for any $v \in \mathbb{R}^k \setminus \{0\}$ that $\int_{\mathcal{W}} v^\top S(x;\xi)S(x;\xi)^\top v\, dx > 0$, and hence $\big|\{x \in \mathcal{W}\colon v^\top S(x;\xi)S(x;\xi)^\top v > 0\}\big| > 0$. Since $\exp(\theta^\top S(x;\xi)) > 0$ for all $\theta \in \mathbb{R}^k$ and $x \in \mathcal{W}$, we obtain

$$\int_{\mathcal{W}} \exp(\theta^\top S(x;\xi))\, v^\top S(x;\xi)S(x;\xi)^\top v\, dx > 0 \quad \text{for every } v \in \mathbb{R}^k \setminus \{0\}\ .$$

Thus PL is strictly concave.

(iii) Without loss of generality we may extend PL to all of $\mathbb{R}^k$ by just using the form (5.7) also for $\theta \in \Theta^c$. Since $S(\cdot;\xi)$ is bounded this is still a $\mathbb{R}$-valued function that is concave. It can be shown that coercivity follows if

$$\lim_{r\to\infty} u^\top \nabla \operatorname{PL}(ru) < 0 \tag{5.10}$$

for every $u \in \mathcal{S}^{k-1}$ (cf. Exercise Sheet 6). We have

$$u^\top \nabla \operatorname{PL}(ru) = u^\top \left(\sum_{x\in\xi} S(x;\xi)\right) - \int_{\mathcal{W}} \exp(ru^\top S(x;\xi))\, u^\top S(x;\xi)\, dx.$$

If $\big|\{x \in \mathcal{W}\colon u^\top S(x,\xi) > 0\}\big| > 0$, then by the theorem of monotone convergence the integral term goes to infinity, so that (5.10) holds. Otherwise the integral term goes to zero by the same theorem. If $u^\top\big(\sum_{x\in\xi} S(x;\xi)\big) < 0$ then (5.10) is still satisfied. □

The following corollary gives a single condition for existence of a unique MPLE that is easier to check in practice (see Exercise Sheet 6).

Corollary 5.C. *Suppose that $S(\cdot;\xi)$ is bounded on $\mathcal{W}$ and that $\Theta \subset \mathbb{R}^k$ is closed and convex. Denote by χ the image measure of Lebesgue measure on $\mathcal{W}$ under $S(\cdot;\xi)$, i.e. $\chi(B) := \big|\{x \in \mathcal{W}\colon S(x;\xi) \in B\}\big|$ for every measurable $B \subset \mathbb{R}^k$. Denote by C the convex*

cone generated by $\operatorname{supp}(\chi)$.[3] *Then a unique maximum pseudolikelihood estimate exists in* Θ *if*

$$\sum_{x\in\xi} S(x;\xi) \in \operatorname{int}(C). \tag{5.11}$$

Remark 5.D. *We do not take into account here maximizers "at infinity", i.e. generalized maximizers* θ *of* PL *for which one or several components are plus or minus infinity. These might correspond to reasonable models in some situations; see for example the stationary Strauss family with* $\log(\gamma) = -\infty$.

Proof of Corollary 5.C. Condition (5.11) implies that $\operatorname{supp}(\chi)$ is not contained in a linear subspace of $\mathbb{R}^k$ of dimension $k-1$ because otherwise $\operatorname{int}(C) = \emptyset$. Hence for any $v \in \mathbb{R}^k \setminus \{0\}$ we have $\big|\{x \in \mathcal{W}\colon v^\top S(x;\xi) \neq 0\}\big| = \chi(\{y \in \mathbb{R}^k\colon v^\top y \neq 0\}) > 0$ and therefore that $\int_{\mathcal{W}} S(x;\xi)S(x;\xi)^\top \, dx$ is positive definite. It is then enough to show that either Condition (α) or Condition (β) from Proposition 5.B holds. Denote by $C^* := \{y \in \mathbb{R}^k\colon y^\top z \leq 0 \text{ for all } z \in C\}$ the so-called *polar cone* of C.

Let $u \in \mathcal{S}^{k-1}$. If $u \in C^*$, then $u^\top \sum_{x\in\xi} S(x;\xi) < 0$ (i.e. Condition (β) is satisfied) because $\sum_{x\in\xi} S(x;\xi)$ is in the interior of C. If $u \in (C^*)^c$, there exists $v \in \operatorname{supp}(\chi)$ with $u^\top v > 0$ (if not, then for every $w \in C$ we would have $u^\top w \leq 0$ because w can be written as a nonnegative linear combination of elements of $\operatorname{supp}(\chi)$; but this contradicts $u \notin C^*$). By continuity of the scalar product, there must be an open neighbourhood U_v of v such that $u^\top z > 0$ for every $z \in U_v$. But then Condition (α) is satisfied, because $\chi(U_v) > 0$ by the fact that v is in the support of χ. □

5.2.2 Asymptotic properties of MPLEs

Deriving asymptotic properties of MPLEs in general is a rather involved topic requiring long lists of special conditions and rather intricate proofs. We restrict ourselves to the special class of pairwise interaction processes with finite range of interaction and give a rough sketch of special cases of poineering results of Jensen and Møller (1991) and Jensen and Künsch (1994) to convey an idea. Further developments can be found in Mase (1995), Mase (2000), Billiot et al. (2008), and Coeurjolly and Drouilhet (2010).

For the sake of completeness we briefly recall the definitions of the key asymptotic properties we are going to discuss, formulated directly in terms of parametric point pattern models. Suppose that $(P_\theta)_{\theta\in\Theta}$, where $\Theta \subset \mathbb{R}^k$, is a family of point process distributions on

[3] The support of a measure μ on the Borel σ-algebra $\mathcal{B}$ of some topological space $\mathcal{Y}$ is defined as

$$\operatorname{supp}(\mu) := \{y \in \mathcal{Y}\colon \mu(U) > 0 \text{ for every open neighbourhood } U \text{ of } y\}.$$

Under very general conditions it can be alternatively defined as the smallest closed set C such that $\mu(C^c) = 0$.

$\mathbb{R}^d$ and that $(\mathcal{W}_n)$ is an increasing sequence of compact subsets of $\mathbb{R}^d$ with $\mathcal{W}_n \nearrow \mathbb{R}^d$. For each $n \in \mathbb{N}$ let $\hat{\theta}_n := \hat{\theta}_n(\Xi)$ be an estimator for θ in the parametric family $(P_\theta^{(n)})_{\theta\in\Theta}$, where $P_\theta^{(n)}$ is the restriction of the distribution P_θ to $\mathcal{W}_n$. We then call the sequence $(\hat{\theta}_n)$

(a) *consistent* if for every $\theta \in \Theta$

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{under } P_\theta;$$

(b) *asymptotically normal* if for every $\theta \in \Theta$ there is a symmetric positiv definite matrix $\Sigma_\theta \in \mathbb{R}^{k\times k}$ such that

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \xrightarrow{\mathscr{D}} \mathcal{N}_k(0, \Sigma_\theta) \quad \text{under } P_\theta;$$

in this context, Σ_θ is known as the *asymptotic covariance matrix* of $(\hat{\theta}_n)$;

(c) *asymptotically efficient*[4] or *best asymptotically normal (BAN)* if it is asymptotically normal with asymptotic covariance matrix $\Sigma_\theta = I_\theta^{-1}$, where

$$I_\theta := \mathbb{E}_\theta\big((\nabla \log f_\theta(\Xi))(\nabla \log f_\theta(\Xi))^\top\big)$$

denotes the *Fisher information matrix*; it can be seen that under rather general technical conditions $I_\theta = -\mathbb{E}_\theta\big(D^2 \log f_\theta(\Xi)\big)$ and $\Sigma_\theta - I_\theta^{-1}$ is positive semidefinite, so that I_θ^{-1} may be interpreted as the best possible asymptotic covariance matrix.

In brief the message for the MPLE is that it is often consistent and asymptotically normal under suitable conditions, but in general not asymptotically efficient. Empirical results furthermore suggest that as far as finite sample results are concerned the MPLE can have considerable bias and be inefficient in distribution families where point interactions are strong.

In order to be a little more precise consider a family $(P_\theta)_{\theta\in\Theta}$ of stationary pairwise interaction processes with finite interaction range $R > 0$. In such a family the densities are of the form

$$f_\theta(\xi) = c(\theta) \exp\Bigg(\theta_1^\top |\xi| + \theta_2^\top \sum_{\{x,y\}\subset\xi} \varphi(x-y)\Bigg),$$

where $\theta = (\theta_1, \theta_2)^\top$ and $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ is point symmetric with respect to zero with $\varphi(z) = 0$ if $\|z\| > R$. It is implicitly understood that the sum is taken only over true two-point subsets of ξ. Note that for example the stationary Strauss process is of this form.

Under existence and uniqueness conditions and technical conditions on φ and the window sequence $(\mathcal{W}_n)$, Jensen and Møller (1991) and Jensen and Künsch (1994) obtain consistency and asymptotic normality in the inhibitory case $\varphi \le 0$. For more general

[4]There are more general definitions of asymptotic efficiency that do not need an asymptotic normality context.

φ consistency is obtained if φ is bounded (and further technical conditions apply) and asymptotic normality is obtained if φ is rotationally symmetric and has a hard core, i.e. $\varphi(z) = -\infty$ for $\|z\| \le r$ with some $r \in (0, R]$ (and further technical conditions apply).

5.2.3 Approximate computation of MPLEs: The Berman–Turner device

While the maximum pseudolikelihood method avoids the numerical approximation of the difficult integral (5.4), there is still an integral over $\mathcal{W}$ to evaluate in (5.6). We approximate this integral by a finite sum as

$$\int_{\mathcal{W}} \lambda_\theta(x \,|\, \xi) dx \;\approx\; \sum_{j=1}^{m} \lambda_\theta(u_j \,|\, \xi) w_j,$$

where u_j are points in $\mathbb{R}^d$ and w_j are appropriate weights compensating for the volume that each point "controls". Such an approximation is called a *quadrature rule*. We choose the points u_j as the union of ξ with a set of (typically many more) dummy points. In our context there are two common possibilities:

1) (Voronoi weights; computationally more expensive). Pick the dummy points either systematically (e.g. on a regular grid) or randomly (e.g. uniformly) in $\mathcal{W}$. Then choose the quadrature weight w_j to be the area/volume of the Voronoi cell of u_j, which is the set of all locations in $\mathbb{R}^d$ that are closer to u_j than to any u_i with $i \neq j$.

2) (Counting weights; computationally cheaper). Partition $\mathcal{W}$ into tiles of equal area/volume a (typically many more than data points). Pick exactly one dummy point per tile placed either systematically or at random. Let $w_j = a/n_j$ where n_j is the total number of points in the same tile as u_j.

We can then approximate the log-pseudolikelihood in (5.7), as

$$\mathrm{PL}(\theta) \approx \sum_{j=1}^{m} \big[y_j \, \theta^\top S(u_j; \xi) - \exp(\theta^\top S(u_j; \xi)) \big] w_j, \tag{5.12}$$

where

$$y_j = \begin{cases} 1/w_j & \text{if } u_j \text{ is a data point,} \\ 0 & \text{if } u_j \text{ is a dummy point.} \end{cases}$$

However, the right hand side of (5.12) is up to an additive constant the log-likelihood of a weighted Poisson regression model: Consider observations $y_1, \ldots, y_n \in \mathbb{Z}_+$ and covariate vectors $z_1, \ldots, z_n \in \mathbb{R}^k$ stemming from independent random pairs $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ with identical conditional distributions

$$\mathscr{L}(Y_i \,|\, Z_i = z) = \mathrm{Po}(\nu_\theta(z)),$$

where

$$\log\bigl(\nu_\theta(z)\bigr) = \theta^\top z.$$

Then the log-likelihood of this model (given $Z_1 = z_1, \ldots, Z_n = z_n$) is

$$\widetilde{L}(\theta) = \log\biggl(\prod_{j=1}^{n} \frac{\nu_\theta(z_j)^{y_j}}{y_j!} e^{-\nu_\theta(z_j)}\biggr) = \sum_{j=1}^{n} \bigl[y_j\, \theta^\top z_j - \exp(\theta^\top z_j)\bigr] - \sum_{j=1}^{n} \log(y_j!).$$

The second sum on the right hand side is not important for maximization in θ. Apart from the fact that here the y_j-values are assumed to be in $\mathbb{Z}_+$ and the weights w_j are equal to 1, this is exactly the approximating term from (5.12) for the log-pseudolikelihood, where $z_j = S(u_j; \xi)$. Software packages that fit Poisson regression models can usually handle weighted summands and non-integer values $y_j \geq 0$; in any case the `glm`-function in R can. In conclusion we have found ourselves a neat way to reduce (approximate) maximum pseudolikelihood estimation to the application of a proved and tested algorithm.

5.2.4 A short overview of further topics

Edge effects

So far we have assumed for simplicity that the point pattern ξ exists only on the window $\mathcal{W}$ on which we observe it (i.e. $\mathcal{X} = \mathcal{W}$). The more common situation is that our data are a clipping of a larger point pattern on $\mathcal{X} \supset \mathcal{W}$. Assuming then that there are no points outside $\mathcal{W}$ can introduce substantial bias if interactions and the corresponding interaction distances are sizeable.

The simplest correction method is the so-called *border correction*. Assume that the conditional intensity depends on ξ only within a certain distance, i.e.

$$\lambda(x \,|\, \xi) = \lambda(x \,|\, \xi|_{B(x,R)})$$

for some $R > 0$. This is the case for all the point pattern models introduced in Chapter 3. We then consider in the Berman–Turner approximation (5.12) only (data and dummy) points u_j inside the reduced window $\mathcal{W} \ominus B(0,R) := \{x \in \mathcal{W} \colon x + B(0,R) \subset \mathcal{W}\}$, but still use ξ on all of $\mathcal{W}$.

The disadvantage of the border correction method is that considerable amounts of data are disregarded, especially in higher dimensions and if R constitutes a substantial fraction of the window size. Alternative correction methods include *translational* and *isotropic correction*, which work in similar ways as the corresponding methods presented in Section 3 for the descriptive functions. In the case of a rectangular window there is also the *toroidal correction method*, which generates artificial neighbours of points close to the boundary by identifying the opposite edges of the rectangle.

Estimation of irregular parameters

As we have seen in Section 5.1 distribution families of interest sometimes contain parameters that cannot be modelled in exponential family form, but have to be plugged into the corresponding methods as known. Such irregular parameters are typically the interaction ranges. There are several ways of obtaining estimators for these parameters as well.

One applied method that is sometimes a bit dubious is the so-called *cusp point method*, which derives parameter estimates from suspicious points in descriptive functions. A typical application of this method is in the stationary Strauss family, where the estimated K- and L functions are known to form a cusp at (roughly) the right interaction range if the point pattern is large enough. With some good will such a cusp can be found at about $r = 0.0508$ (from inspecting the plot and then looking up the cusp point in the numerical data) in Figure 3.8. Although the point pattern and therefore the cusp is really too small to make this a convincing conclusion.

A more objective method is provided by maximum profile pseudolikelihood. We write now $\mathrm{PL}(R;\theta)$ to emphasize the dependence of the pseudolikelihood function on the irregular parameter R, which in general we assume to lie in a set $\mathcal{R} \subset \mathbb{R}^q$. Denote by $\mathrm{PPL}(R) := \max_{\theta\in\Theta} \mathrm{PL}(R;\theta)$ the profile log-pseudolikelihood. We compute

$$(\hat{R}, \hat{\theta}) := \operatorname*{arg\,max}_{R\in\mathcal{R},\, \theta\in\Theta} \mathrm{PL}(R;\theta)$$

by first computing

$$\hat{R} = \operatorname*{arg\,max}_{R\in\mathcal{R}} \mathrm{PPL}(R)$$

and then (assuming $\hat{R}$ is unique)

$$\hat{\theta} = \operatorname*{arg\,max}_{\theta\in\Theta} \mathrm{PL}(\hat{R};\theta).$$

The profile log-pseudolikelihood is in general not a concave (nor even unimodal) function. If the space $\mathcal{R}$ is one-dimensional and not too large it may be maximized by brute force, i.e. by evaluating PPL on a reasonably fine grid.

Confidence regions and P-values

In spite of the asymptotic results that are available, inference about (regular) model parameters is usually not done via asymptotic theory, but via parametric bootstrap. One main reason for this is that for a given sample ξ the resulting χ^2-approximation of the log-pseudolikelihood ratio may be very bad. For the considerations below assume that the log-pseudolikelihood is strictly concave and coercive such that a unique maximum pseudolikelihood estimator exists.

The (parametric) bootstrap paradigm says that the estimated distribution $P_{\hat{\theta}}$ may be used in place of the true distribution P_{θ_0}. To obtain confidence regions we generate N independent samples from $P_{\hat{\theta}}$, compute their parameter estimates $\hat{\theta}_1, \dots, \hat{\theta}_N$ and consider a set containing all estimates up to the k-th "extremest"[5] (in some sense) as a $(1 - \frac{k}{N+1})$-confidence region. Componentwise two-sided $(1 - \frac{2k}{N+1})$-confidence intervals are obtained as $[\hat{\theta}^i_{(k)}, \hat{\theta}^i_{(N+1-k)}]$, where $\hat{\theta}^i_{(1)}, \dots, \hat{\theta}^i_{(N)}$ denotes the ordered values of the i-th components of $\hat{\theta}_1, \dots, \hat{\theta}_N$.

P-values for the null hypothesis $\Psi^\top \theta = \eta$, where $\Psi \in \mathbb{R}^{k \times l}$ and $\eta \in \mathbb{R}^l$ are obtained as follows. Define a (in a sense more general version than before of the) profile log-pseudolikelihood as

$$\mathrm{PPL}(\eta) := \max_{\theta \in \Theta \colon \Psi^\top \theta = \eta} \mathrm{PL}(\theta)$$

and denote the corresponding maximizer by $\hat{\theta}^{(\eta)}_{\mathrm{MPL}}$. We consider the log-pseudolikelihood ratio statistic (or deviance statistic)

$$\tau(\tilde{\xi}) := 2 \log\left(\frac{\mathrm{PL}(\hat{\theta}_{\mathrm{MPL}})}{\mathrm{PPL}(\eta)}\right) = 2 \log\left(\frac{\mathrm{PL}(\hat{\theta}_{\mathrm{MPL}})}{\mathrm{PL}(\hat{\theta}^{(\eta)}_{\mathrm{MPL}})}\right), \quad \tilde{\xi} \in \mathfrak{N}^*.$$

Simulating N independent point patterns $\xi_1, \dots, \xi_N$ from $P_{\hat{\theta}^{(\eta)}_{\mathrm{MPL}}}$ and computing corresponding parameter estimates $\hat{\theta}_1, \dots, \hat{\theta}_N$, the bootstrap P-value for testing the null hypothesis $\Psi^\top \theta = \eta$ is obtained as

$$\frac{N + 2 - r(\xi)}{N + 1},$$

where $r(\xi)$ denotes the rank of $\tau(\xi)$ among $\tau(\xi_1), \dots, \tau(\xi_N)$.

5.3 (Numerical) maximum likelihood

We still consider an exponential family $(P_\theta)_{\theta \in \Theta}$ where we assume for simplicity that $P_0 = \mathrm{Po}_1$. That is, we have densities

$$f_\theta(\xi) = Z(\theta)^{-1} \exp(\theta^\top T(\xi)), \quad \xi \in \mathfrak{N}^*,$$

where

$$Z(\theta) = \frac{1}{c(\theta)} = \int_{\mathfrak{N}^*} \exp(\theta^\top T(\xi)) \ \mathrm{Po}_1(d\xi) \in (0, \infty). \tag{5.13}$$

Hence our log-likelihood takes the form

$$L(\theta) = \theta^\top T(\xi) - \log(Z(\theta)).$$

As stated before, the integral in (5.13) is problematic. It can almost never be computed analytically, but has to be approximated numerically. Such approximations have been

[5] i.e. we discard the $k - 1$ extremest estimates.

extensively investigated, but the topic remains a difficult one. The main impulse has come from physics, mainly statistical mechanics, where large systems of particles are studied. In this context the normalizing constant $Z(\theta)$ is known as the *partition function*.

There are nifty ways how the computational burden can be reduced and reasonably accurate approximations can be obtained in (sometimes) non-prohibitive time. We consider here just a rather naïve approach, which nevertheless can be quite successful when paired with maximum pseudolikelihood estimation.

We start by computing the gradient and the Hessian matrix of the log-likelihood. In order to be able to interchange differentiation and integration below, we require that $\mathbb{E}_\theta(\|T(\Xi)\|^2) < \infty$ for all $\theta \in \Theta$. We then have

$$\nabla Z(\theta) = \int_{\mathfrak{N}^*} \exp(\theta^\top T(\xi))\, T(\xi)\ \mathrm{Po}_1(d\xi) = Z(\theta)\, \mathbb{E}_\theta T(\Xi)$$

and

$$\begin{aligned} D^2 Z(\theta) &= D\biggl(\int_{\mathfrak{N}^*} \exp(\theta^\top T(\xi))\, T(\xi)\ \mathrm{Po}_1(d\xi)\biggr) \\ &= \int_{\mathfrak{N}^*} \exp(\theta^\top T(\xi))\, T(\xi) T(\xi)^\top\ \mathrm{Po}_1(d\xi) \\ &= Z(\theta)\, \mathbb{E}_\theta\bigl(T(\Xi)T(\Xi)^\top\bigr). \end{aligned}$$

Therefore

$$\nabla L(\theta) = T(\xi) - \frac{1}{Z(\theta)} \nabla Z(\theta) = T(\xi) - \mathbb{E}_\theta T(\Xi)$$

and

$$\begin{aligned} D^2 L(\theta) &= -D\biggl(\frac{1}{Z(\theta)} \nabla Z(\theta)\biggr) \\ &= \frac{1}{Z(\theta)^2} \nabla Z(\theta) \nabla Z(\theta)^\top - \frac{1}{Z(\theta)} D^2 Z(\theta) \\ &= \bigl(\mathbb{E}_\theta T(\Xi)\bigr)\bigl(\mathbb{E}_\theta T(\Xi)\bigr)^\top - \mathbb{E}_\theta\bigl(T(\Xi)T(\Xi)^\top\bigr) \\ &= -\operatorname{Var}_\theta\bigl(T(\Xi)\bigr), \end{aligned}$$

where the above form of the product rule is readily checked componentwise. A first result we have from this is that the log-likelihood is concave and, provided that $T(\Xi)$ is not concentrated on an affine subspace of $\mathbb{R}^k$, even strictly concave.

Naïve numerical approximations of gradient and Hessian are then given by the following Monte Carlo method. Generate N independent realizations $\xi_1, \dots, \xi_N$ from P_θ and approximate theoretical mean and variance by the corresponding sample quantities,

$$\begin{aligned} \mathbb{E}_\theta T(\Xi) &= \frac{1}{N} \sum_{i=1}^N T(\xi_i) =: \overline{T(\xi)}, \\ \operatorname{Var}_\theta T(\Xi) &= \frac{1}{N-1} \sum_{i=1}^N \bigl(T(\xi_i) - \overline{T(\xi)}\bigr)\bigl(T(\xi_i) - \overline{T(\xi)}\bigr)^\top. \end{aligned} \tag{5.14}$$

5.3.1 Newton Method

Since L is a (typically strictly) concave function that is twice continuously differentiable and we have explicit (albeit only approximative) expressions for the gradient and Hessian, we may use the Newton method in order to try to find its maximum (if it exisits). Up to a potential step size correction, which might be necessary if the "curvature" of the log-likelihood surface is "small" in an explored region, this means that beginning with some starting value θ_0, we use the iteration procedure

$$\theta_{n+1} = \theta_n - \left(D^2 L(\theta_n)\right)^{-1} \nabla L(\theta_n) \quad \text{for } n \geq 0. \tag{5.15}$$

This iteration step is obtained by approximating L in θ_n by a quadratic function and taking the maximum of this function as new value θ_{n+1}. Taylor approximation in θ_n yields the function

$$L(\theta \,|\, \theta_n) = L(\theta_n) + \nabla L(\theta_n)^\top (\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^\top D^2 L(\theta_n)(\theta - \theta_n).$$

Assuming that the Hessian is strictly negative definite the maximizing value θ_{n+1} must satisfy

$$0 = \nabla L(\theta_n) + D^2 L(\theta_n)(\theta_{n+1} - \theta_n),$$

which yields Equation (5.15).

Each Newton step requires the simulation of point patterns $\xi_1, \ldots, \xi_N$. One would expect that each time an order of magnitude of thousand point patterns is needed for the approximations in (5.14) to be good. What is more, the point patterns ξ_i typically have to be simulated by Markov Chain Monte Carlo, requiring thousands to tens of thousands of Markov Chain steps for each pattern. For this reason the naïve approximation approach paired with the Newton method is usually very, very slow.

5.3.2 Huang–Ogata one-step method

Huang and Ogata (1999) propose a simple but powerful method which often gives a surprisingly good estimator that approximates the MLE. The idea is to start with a reasonably good initial estimator, *namely the MPLE*, and then make a single Newton step, based on a (very) rough approximation in (5.14), using just $N = 100$. A step size correction may still be beneficial although the authors do not comment on this.

As we will see in Section 5.5 this method can do very well in practice. The idea seems to follow a more comprehensive principle. Inagaki (1973) showed in the case of n independent observations that the estimator obtained by a single Newton step on the log-likelihood surface starting from a consistent estimator is "asymptotically equivalent" to the MLE.

5.4 Model diagnostics

5.4.1 Residuals

In analogy with linear models we may want to consider residuals of fitted point pattern models. Although there is not one unique way to define such residuals, the concept introduced in Baddeley et al. (2005) is quite natural and has been used rather successfully. We introduce a *raw residual measure* (a *signed* measure actually) on $\mathcal{W}$, which is given by

$$R_{\hat{\theta}}(B) = \xi(B) - \int_B \lambda_{\hat{\theta}}(x \,|\, \xi)\, dx$$

for any $B \in \mathcal{B}_{\mathcal{W}}$. As mentioned before the information contained in a spatial point pattern consists of presence *and* of absence of points in space, which is reflected in the fact that we consider a residual *measure*, not just residuals at the points of the data pattern ξ.

Note that we have

$$\mathbb{E}_\theta R_\theta(B) = 0$$

for any $B \in \mathcal{B}_{\mathcal{W}}$ and any θ by the Georgii-Nguyen-Zessin-Formula. If the model is correct, we expect (the density of) $R_{\hat{\theta}}$ to fluctuate around zero. One way of producing a diagnostic plot is too look at the smoothed residual field $(s(x)\colon x \in \mathcal{W})$, i.e. the density of a kernel smoothed version of the measure ξ minus the measure $[B \mapsto \int_B \lambda_{\hat{\theta}}(x \,|\, \xi)\, dx]$. This should look as uninformative as possible in order to point to a good fit.

Variance computations show that the raw residuals in general do not fluctuate evenly over the whole of $\mathcal{W}$. It may therefore be better to look at a *Pearson residual measure*, which is given by

$$R^{\mathrm{Pears}}_{\hat{\theta}}(B) = \sum_{x \in \xi \cap B} \frac{1}{\sqrt{\lambda_{\hat{\theta}}(x \,|\, \xi)}} - \int_B \sqrt{\lambda_{\hat{\theta}}(x \,|\, \xi)}\, dx$$

for any $B \in \mathcal{B}_{\mathcal{W}}$, assuming that $\lambda_{\hat{\theta}}(x \,|\, \xi) > 0$ if $x \in \xi$. This type of residuals tends to be better suited for detecting unusual regions (mainly unusual data points where there should be none).

In general one may consider any GNZ-difference for a residual measure, i.e. for any function $h_{\hat{\theta}}(x; \xi)$ we may consider h-weighted residuals

$$R^{(h)}_{\hat{\theta}}(B) = \sum_{x \in \xi \cap B} h_{\hat{\theta}}(x; \xi \setminus \{x\}) - \int_B h_{\hat{\theta}}(x; \xi)\, \lambda_{\hat{\theta}}(x \,|\, \xi)\, dx$$

and the GNZ-formula guarantees that

$$\mathbb{E}_\theta R^{(h)}_\theta(B) = 0$$

for any $B \in \mathcal{B}_{\mathcal{W}}$ and any $\theta \in \Theta$.

Plotting the smoothed residual field is mainly a means for detecting problems with the fitted trend. Problems with interpoint interactions are better detected by looking at QQ-plots of the smoothed residual field against the expected residual field, both evaluated on a fine grid of locations in $\mathcal{W}$. The expected residual field is usually approximated by Monte Carlo averaging.

All of these diagnostic plots are considered for numerical examples in Section 5.5.

5.4.2 Goodness-of-fit tests

It is also a good idea to use one or several of the descriptive functions Φ introduced in Chapter 3, and compare the function $\hat{\Phi}$ estimated from the data point pattern ξ to the theoretical function for the fitted model (obtained by Monte Carlo sampling if necessary).

Plotting pointwise Monte Carlo envelopes is helpful to visually assess the goodness-of-fit. Note, however, that formal Monte Carlo testing is questionable since we have fitted the model parameters based on the same data. Some authors propose to use descriptive functions that "focus on different aspects of the point process distribution than those used in the estimating procedure" for formal tests.

5.5 Numerical examples

We first consider the point pattern in Figure 3.7, simulated from a Strauss$(0.05; 100, 0.3)$-distribution. Using our knowledge about the true interaction range $R = 0.05$, we obtained parameter estimates of $\hat{\beta} = 76.219$ and $\hat{\gamma} = 0.347$ with maximum pseudolikelihood and border correction, $\hat{\beta} = 85.812$ and $\hat{\gamma} = 0.283$ with maximum pseudolikelihood and translational correction, and of $\hat{\beta} = 85.890$ and $\hat{\gamma} = 0.331$ with the Huang–Ogata method and border correction. The maximum pseudolikelihood estimates do not involve any simulation because the dummy points in the quadrature rule were chosen as a fine regular grid. The Huang–Ogata estimate on the other hand may vary considerably in subsequent fits due to the sampling involved for the approximation (5.14) with $N = 100$.

Figure 5.1 shows the bootstrap sampling distribution in the β-γ-space. The points are scattered about rather wildly, but since our original point pattern was not very large we cannot expect wonders. Note the (curved) horizontal line structures in the points, which correspond to different numbers of R-close pairs in the simulated point patterns. It can be shown that if there are no R-close pairs in a point pattern then the maximum pseudolikelihood estimator for γ will be 0, which strictly speaking is a degenerate solution corresponding to $-\infty$ in the canonical parameter space.

Parameterwise 95% confidence intervals based on the above bootstrap sample are given

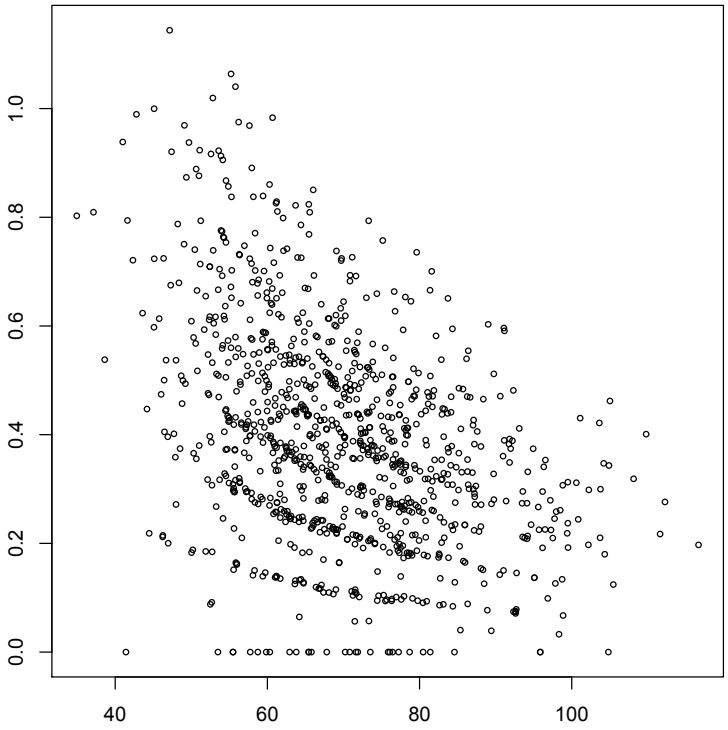

Figure 5.1: Parameter estimates from 999 parametric bootstrap samples for the point pattern in Figure 3.7 using the maximum pseudolikelihood method with border correction.

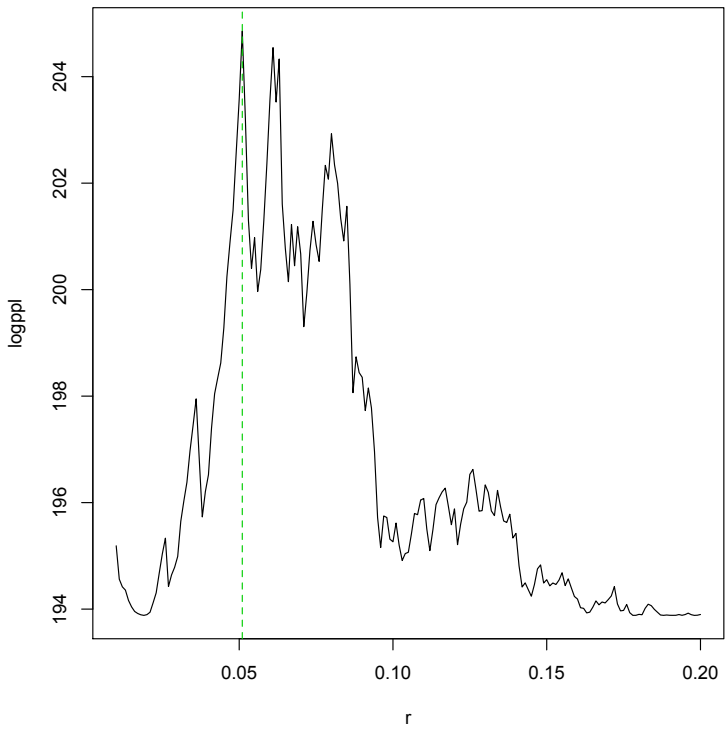

Figure 5.2: Profile log-pseudolikelihood plot for the Strauss example from Figure 3.7 using translational edge correction.

by $[47.157, 98.822]$ for β and $[0, 0.857]$ for γ.

If we do not use our knowledge of the true interaction range R, we may estimate it by maximizing the profile log-pseudolikelihood. A plot of this function is given in Figure 5.2. The maximum is (somewhat luckily) attained at 0.510, which then leads to a very similar fit for the regular parameters as above.

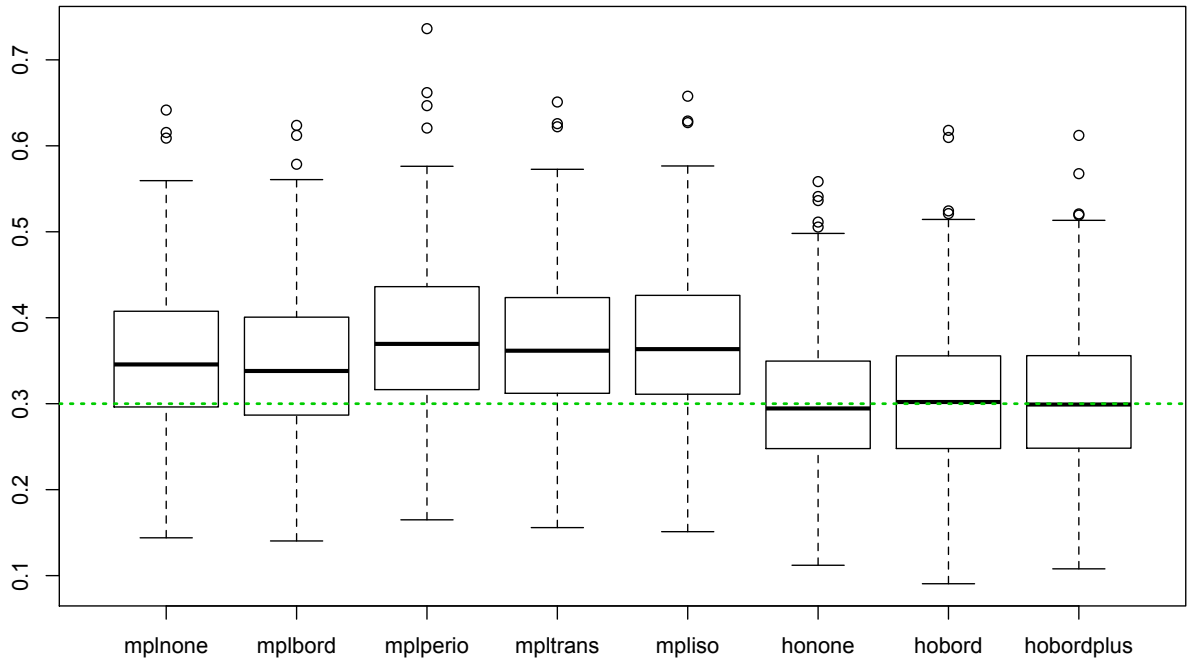

Figure 5.3: Summaries of Monte Carlo distributions of estimators for γ in the Strauss$(R; \beta, \gamma)$-model for maximum pseudolikelihood and Huang–Ogata technique under various edge correction methods. The boxplots are based on 500 simulations. The window was $[0, 1]^2$ and the true parameters were $R = 0.05$ (assumed to be known), $\beta = 250$, and $\gamma = 0.3$.

For a better comparison of the performance of the maximum pseudolikelihood versus the Huang–Ogata method under various kinds of edge corrections, we have simulated 500 point patterns from the Strauss$(0.05; 250, 0.3)$-distribution. The Monte Carlo distributions of the resulting estimators for γ are summarized in Figure 5.3. The rightmost boxplot is for the Huang–Ogata method with border correction, but for $N = 1000$ simulations in the approximation (5.14). We see that the maximum pseudolikelihood estimators exhibit considerable bias, but the Huang–Ogata estimators do not. Furthermore the standard deviations of the Huang–Ogata estimators tend to be a bit smaller than those of the MPLEs. Means and standard deviations are given in the following table.

mplnone	mplbord	mplperio	mpltrans	mpliso	honone	hobord	hobordplus
0.350	0.343	0.376	0.365	0.367	0.298	0.303	0.303
0.0823	0.0845	0.0862	0.0822	0.0837	0.0752	0.0810	0.0793

As a second example we consider a point pattern simulated as a clipping on $[0,1]^2$ of a larger Strauss$(0.02;\beta(\cdot),0.3)$-process with a somewhat complicated spatial trend, namely $\beta(x,y) = \exp(7.5 - 4x^2 + 2y)$, where x, y denote the coordinates in $\mathbb{R}^2$. The point pattern is shown in Figure 5.4.

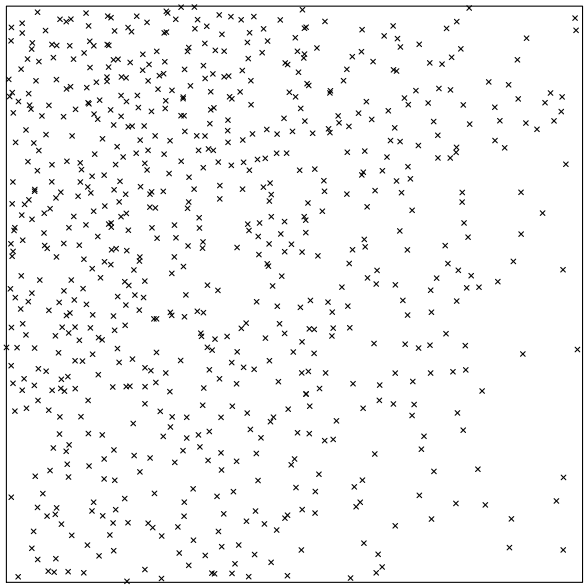

Figure 5.4: Large point pattern from an inhomogeneous Strauss$(0.02;\beta(\cdot),0.3)$-distribution with a quadratic trend. The realized number of points was 738.

Again we make use of our knowledge of the true $R = 0.02$. Since R is small compared to the side lengths of the window, we use in all analyses the border correction method. We fit an inhomogeneous Strauss family with general quadratic trend, i.e. the family from Example 5.A(iii) with $B(x,y) = 1$ and $S(x,y) = (1, x, y, x^2, xy, y^2)^\top$, so that $k = 7$. We obtain parameter estimates of

$$\hat{\psi} = (6.922, 0.668, 2.113, -3.387, -0.380, -0.382)^\top \quad \text{and} \quad \hat{\gamma} = 0.375$$

for the maximum pseudolikelihood and

$$\hat{\psi} = (7.205, 0.285, 2.369, -3.461, -0.278, -0.461)^\top \quad \text{and} \quad \hat{\gamma} = 0.292$$

for the Huang–Ogata method, which are both reasonably close to the true parameters of

$$\psi = (7.5, 0, 2, -4, 0, 0)^\top \quad \text{and} \quad \gamma = 0.3.$$

Figure 5.5 compares the true trend surface with the two estimates. Once more the Huang–Ogata method seems to do quite a bit better, most notably in estimating the interaction parameter γ.

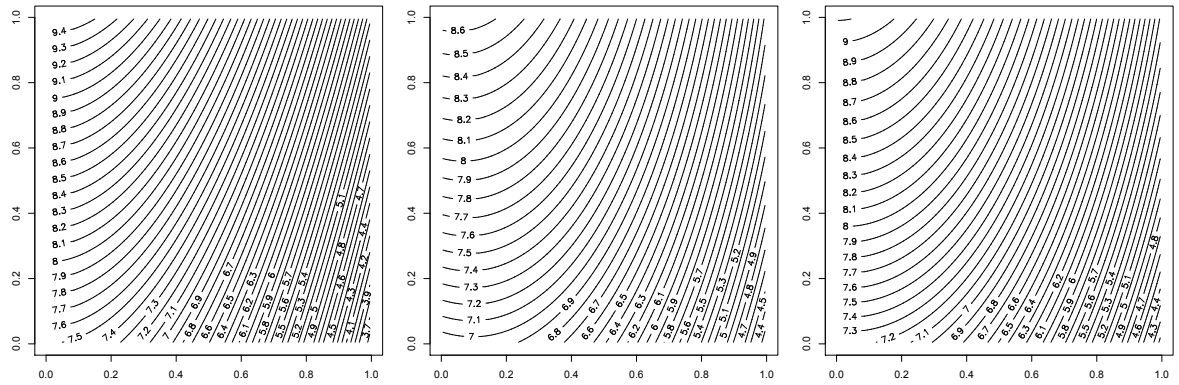

Figure 5.5: From left to right: true trend surface, MPL estimate, and Huang–Ogata estimate of trend surface.

Let us finally consider some model diagnostics. First we examine the residuals for problems with the spatial trend. We restrict ourselves to the raw residuals. The smoothed residual field in the lower right corner of the left panel in Figure 5.6 shows no suspicious structure. This confirms that our fitted trend is reasonable. On the other hand, if we had fitted an inhomogeneous Strauss process with only a linear trend to the data, the diagnostic plots would have looked as in the right panel of Figure 5.6. The fit is clearly inadequate. There is structure in the direction of the x-coordinate, with high residuals in the centre and low residuals towards the left and right boundary. This is saying that compared to the fitted model our data point pattern has too many points in the central region and to few points in the outer regions (in x-direction). The simplest way to amend this is by introducing a quadratic trend in the x-direction. Of course, knowing the true model, we can see that this is exactly what was missing.

Next we examine the resiudals for problems with point interactions. We draw QQ-plots as explained in Section 5.4. Note that the bandwidth of the smoothing kernels is crucial for these plots. Since our suspected (or in fact: known) interaction works on a scale of $R = 0.02$, we choose a bandwidth σ that is somewhat larger, say $\sigma = 0.05$. This is quite a bit smaller than the default σ of roughly 0.14 used by spatstat. The corresponding plots for the raw residuals are given in Figure 5.7. The left plot is for our fitted Strauss model and gives us nothing to worry about. For comparison, fitting a Poisson process with quadratic trend results in the QQ-plot to the right, which shows that the latter model is clearly inadequate.

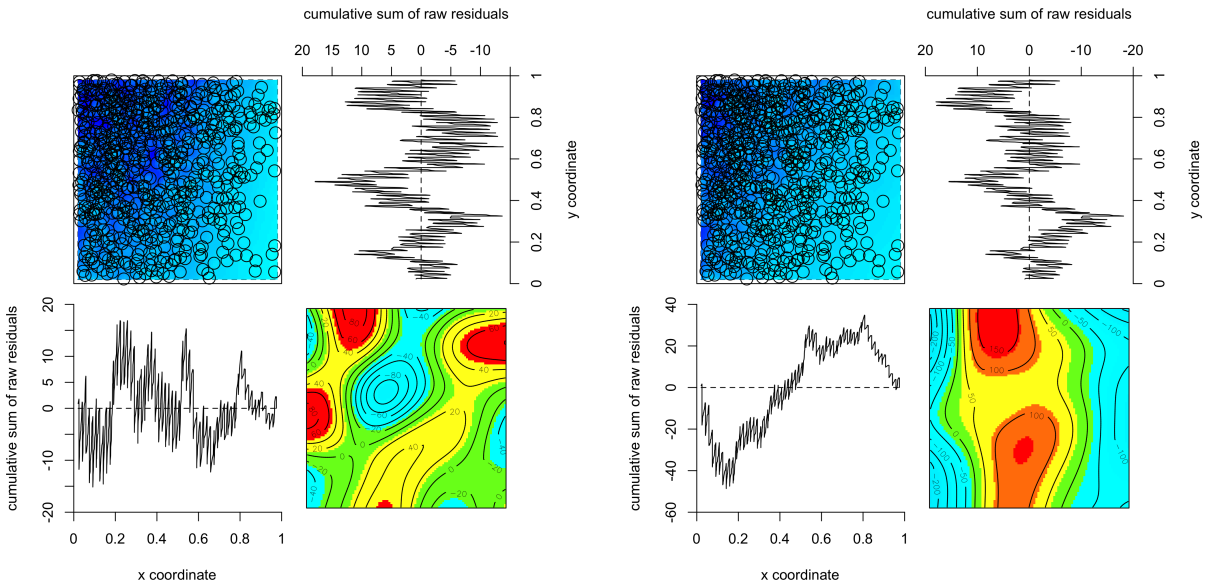

Figure 5.6: Trend diagnosis. Left for the (correct) Strauss model with quadratic trend, right for the Strauss model with only a linear trend.

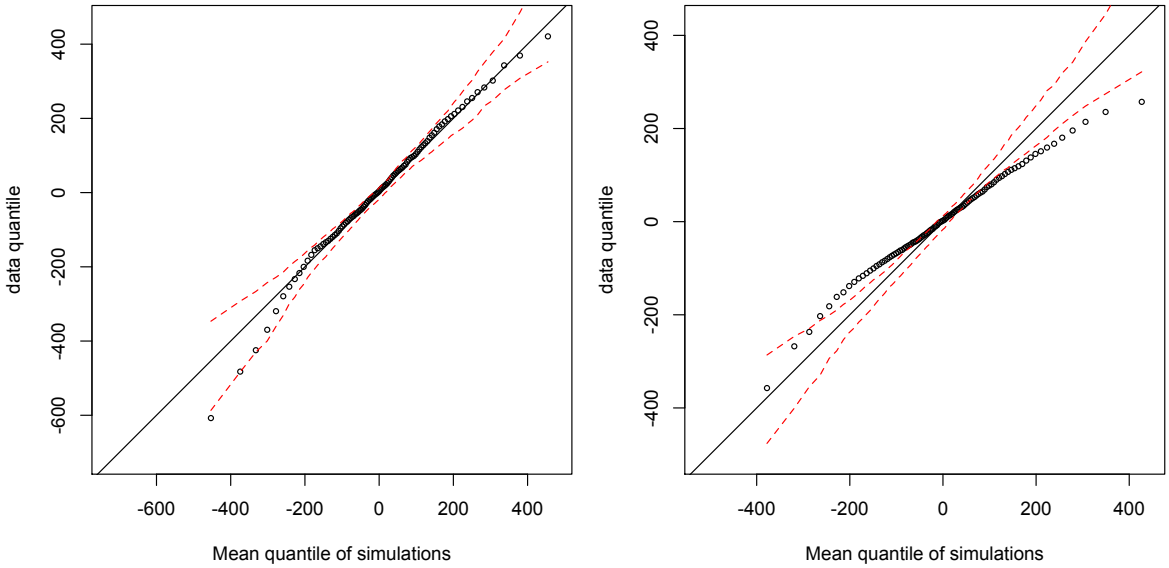

Figure 5.7: Interaction diagnosis. Left for the (correct) Strauss model with quadratic trend, right for the Poisson model with quadratic trend.

Bibliography

Amann, H. and Escher, J. (2001). *Analysis III*. Birkhäuser, Basel.

Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(5):617–666. With discussion and a reply by the authors.

Baddeley, A. J., Kerscher, M., Schladitz, K., and Scott, B. (2000). Estimating the J function without edge correction. *Statistica Neerlandica*, 54(3):315–328.

Berman, M. and Diggle, P. J. (1989). Estimating weighted integrals of the second order intensity of a spatial point process. *J. R. Statist. Soc. B*, 51:81–92.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, 36:192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

Billiot, J.-M., Coeurjolly, J.-F., and Drouilhet, R. (2008). Maximum pseudolikelihood estimator for exponential family models of marked Gibbs point processes. *Electron. J. Stat.*, 2:234–264.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probability*, 3(3):534–545.

Coeurjolly, J.-F. and Drouilhet, R. (2010). Asymptotic properties of the maximum pseudo-likelihood estimator for stationary Gibbs point processes including the Lennard-Jones model. *Electron. J. Stat.*, 4:677–706.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I*. Probability and its Applications (New York). Springer-Verlag, New York, second edition. Elementary theory and methods.

Diggle, P. J. (1985). A kernel method for smoothing point process data. *Appl. Statist.*, 34(2):138–147.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J. R. Statist. Soc. A*, 153(3):349–362.

Diggle, P. J. (2003). *Statistical analysis of spatial point patterns*. Arnold, London, second edition.

Diggle, P. J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of elevated risk. *J. R. Statist. Soc. A*, 157(3):433–440.

Dümbgen, L. (2010). *Empirische Prozesse (in German)*. Lecture Notes, University of Bern.

Elstrodt, J. (2007). *Maß- und Integrationstheorie*. Springer, Berlin, Heidelberg, fifth edition.

Huang, F. and Ogata, Y. (1999). Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics*, 8(3):510–530.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, Chichester.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum liklihood estimator. *Ann. Inst. Statist. Math.*, 25:1–26.

Jensen, J. L. and Künsch, H. R. (1994). On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Ann. Inst. Statist. Math.*, 46(3):475–486.

Jensen, J. L. and Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *Ann. Appl. Probab.*, 1(3):445–461.

Jones, B. J. T., Martínez, V. J., Saar, E., and Trimble, V. (2004). Scaling laws in the distribution of galaxies. *Rev. Modern Phys.*, 76(X):1211–1266.

Kallenberg, O. (1986). *Random measures*. Akademie-Verlag, Berlin, fourth edition.

Kallenberg, O. (2002). *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition.

Mase, S. (1995). Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes. *Ann. Appl. Probab.*, 5(3):603–612.

Mase, S. (2000). Marked Gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Math. Nachr.*, 209:151–169.

Ogata, Y. (1998). Space-time point process models for earthquake occurrences. *Ann. Inst. Statist. Math.*, 50:379–402.

Quine, M. P. and Watson, D. F. (1984). Radial generation of n-dimensional Poisson processes. *J. Appl. Probab.*, 21(3):548–557.

Schneider, R. and Weil, W. (2008). *Stochastic and integral geometry*. Springer, Berlin, Heidelberg.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, Berkeley, Calif. Univ. California Press.

van Lieshout, M. N. M. and Baddeley, A. J. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361.