AALBORG UNIVERSITY

Spatial Point Processes

Models, Simulation, and Statistical Inference

by

Tina Dam
Jan Brønnum Sørensen
Michael Hedegaard Thomsen

IR–99–2002 June 1999

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7E • 9220 Aalborg Øst • Denmark
Phone: +45 96 35 80 80 • Fax: +45 98 15 81 29

Spatial Point Processes

Models, Simulation, and Statistical Inference

A Master of Science Thesis in Statistics

submitted by

Tina Dam, Jan Brønnum Sørensen
and Michael Hedegaard Thomsen

Period

February 1st – June 4th 1999

Supervisor

Associate Professor Jesper Møller

Numbers printed: 40

Pages: 199

Abstract

This thesis presents the theory of spatial point processes. It introduces the general setting and characterization results for point processes, followed by definitions of the binomial and the Poisson point processes. The Poisson point process is important for the theory of point processes as it models complete spatial randomness. Various summary statistics are presented, and estimation and examples of the use of such are given. Emphasis is placed on Ripley's K-function and statistics related to it. Different classes of models are defined; Cox, cluster, hard-core, Markov, nearest neighbour Markov, and Gibbs processes and simulation of each of these models is explained. Markov chain Monte Carlo simulation is covered and here the Metropolis-Hastings algorithm is an important tool. We then present spatial birth–and–death processes that provide an alternative to the Metropolis-Hastings algorithm. Perfect simulation is treated in order to make it possible to simulate from an exact equilibrium distribution. Parametric inference is the final topic covered. Here pseudo-likelihood estimation, contrast-methods, Markov chain Monte Carlo maximum likelihood estimation, and stochastic gradient estimation is treated along with parametric bootstrapping, likelihood ratio and Fisher-Wald tests. Finally, six different point patterns are analysed mainly to exemplify the use of the presented theory.

Preface

The topic of this thesis is spatial point processes and the aim is to present a comprehensive mathematical and statistical coverage of the theory for this area of statistics.

Most of the theory is discussed thoroughly, but since very broad aspects of spatial statistics are considered, there are in some cases only a short description of the subject with references to further studies.

The theory presented is partly based on a series of lectures given in the fall semester of 1998 by Associate Professor Jesper Møller at Aalborg University. Further, the thesis builds on preliminary work made by the three authors in cooperation with Master students Ole Ravn Christensen and Bjarne Hansen also in the fall semester of 1998.

Main references can be found in the bibliography on page 189.

Aalborg University, June 4th 1999

Tina Dam Jan B. Sørensen Michael H. Thomsen

Contents

Notation

Below we list most of the symbols used in this thesis. Certain symbols attain more than one interpretation but this should not confuse the reader, as it is clear from the context, how the symbol is used.

Symbol	Usage	Used first on page
$\emptyset$	empty set	5
$\mathbb{1}[\dots]$	indicator-function	13
$\angle\kappa$	angle between κ and the abscissa	42
$\sim$	symmetric, reflexive relation	74
$\underset{x}{\sim}$	reflexive, symmetric neighbour relation	64
$\underset{x}{\sim}^2$	iterated Dirichlet relation	77
$\mathrm{Exp}(\cdot)$	exponential distribution	116
$\leq_E$	partial ordering relation	124
$\cdot$	inner product	131

Script letters	Usage	Used first on page
$\mathcal{B}$	Borel sets	5
$\mathcal{B}_0$	Bounded Borel sets	5
$\mathcal{B}_{\mathcal{L}}$	Borel sets of $\mathcal{L}$	9
$\mathcal{B}_{\mathcal{M}}$	Borel sets of $\mathcal{M}$	9
$\mathcal{C}$	σ-algebra for use with MCMC, $\mathcal{C} \subseteq \mathcal{N}$	90
$\mathscr{D}$	distribution of Markov chain	91
$\mathcal{F}$	element of $\mathcal{N}$	11
$\mathcal{L}$	space for locations	8
$\mathcal{M}$	space for marks	8
$\mathscr{M}$	common distribution of the marks	20
$\mathcal{N}$	σ-algebra on N	5
$\mathcal{R}$	measure space	121
$\mathscr{R}$	distribution on $\mathcal{R}$	121
$\mathcal{V}$	countable dense set in $\mathcal{X}$	5
$\mathcal{X}$	separable metric space	5

$\mathbb{B}\mathbb{B}$	Usage	Used first on page
$\mathbb{D}$	probability w.r.t. $\mathscr{D}$	91
$\mathbb{E}_0^!$	expectation of $\mathbb{P}_0^!$	35
$\mathbb{E}$	mean value	27
$\mathbb{N}_0$	natural numbers incl. zero	5
$\mathbb{N}$	natural numbers	7
$\mathbb{P}_\xi(\cdot)$	Palm distribution at the point ξ	29
$\mathbb{P}_0$	Palm distribution at the typical point	32
$\mathbb{P}_\xi^!$	reduced Palm distribution	30
$\mathbb{P}_0^!$	reduced Palm distribution at the typical point	32
$\mathbb{P}_t$	distribution of Markov chain	91
$\mathbb{P}$	probability distribution	91
$\mathbb{P}$	probability measure/distribution	6
$\mathbb{R}_+$	positive haf-line $(0, \infty)$	7
$\mathbb{R}^n$	n-dimensional Euclidean space	8

Greek letters	Usage	Used first on page
α, β, γ	positive constants	13
$\alpha^{(n)}$	nth order factorial moment measure	28
$\alpha_h^{(2)}$	second-order marked factorial moment measure	47
α	acceptance probability	101
α	total transition rate	116
$\chi(y \mid x)$	clique indicator function	75
$\partial(\cdot)$	neighbourhood	65
ϕ	measure	93
ϕ	transition function	121
γ_t	lag t auto-covariance	97
ξ, η, κ	Events in $\mathcal{X}$	8
λ	intensity (function)	15
$\lambda(\cdot;\cdot)$	Papangelou conditional intensity	65
$\lambda_\theta(x;\xi)$	parameterized Papangelou conditional intensity	134
$\lambda_\theta(\cdot)$	parameterized intensity function	134
$\lambda^{(n)}$	The nth order product density	31
μ	(intensity-)measure	13
π	target distribution	90
θ	element of Θ	129
$\sigma(\cdot)$	sigma-algebra	5
σ_g^2	asymptotic variance	97
σ	parameter of the soft-core process	72
τ	packing density	63
τ	waiting time	116
ν	(non-atomic-)measure	13
ζ	probability measure	96
$\zeta, \tilde{\zeta}$	kernels	96
φ	interaction function	68
$\{\Delta_t\}$	dominating process	117
$\Phi(y \mid x)$	interaction function	75
Λ	driving measure of a Cox process	55
Π	the distribution of a Poisson point process	23
Θ	parameter space	129
Ψ	probability space	6
Υ	σ-algebra	6

roman	Usage	Used first on page
$b(x,r)$	open ball with center x and radius r	13
$\bar{b}(\cdot,\cdot)$	closure of $b(\cdot,\cdot)$	20
$b(\cdot,\cdot)$	density for births	103
c_i	normalising constant	115
c_θ	unknown normalising constant	129
$c(\cdot,\cdot,\cdot)$	covariance function	149
$c_W(\xi)$	edge correcting factor	39
$d(\cdot,\cdot)$	death rate	116
d	metric on $\mathcal{X}$	5
$d_{\mathcal{L}}$	metric on $\mathcal{L}$	9
$d_{\mathcal{M}}$	metric on $\mathcal{M}$	9
$d(\cdot,\cdot)$	probability of choosing a point to delete	103
e_b	Epanecnikov kernel	44
$g(\cdot)$	pair-correlation function	36
g_h	mark-correlation function	47
$h(\cdot,\cdot,\cdot)$	Metropolis–Hastings ratio	101
k_b	kernel function with bandwidth b	38
$l(\theta)$	log likelihood function	134
l_i	element of $\mathcal{L}$	8
m_i	element of $\mathcal{M}$	8
p	density w.r.t. Π	23
p_n	equilibrium density	99
$pl_A(\theta;x)$	log pseudo likelihood function	135
$q(\cdot,\cdot,\cdot)$	proposal density	100
$r(\cdot)$	correlation function	57
$r(\cdot)$	probability of remaining in the current state	100
$r_\beta(\cdot)$	correlation function	149
$t(\cdot)$	n-dimensional measurable function	131
$t(\cdot;\cdot)$	statistic	148
t	a summary statistic	49
$t^*(r)$	upper envelope of t	49
$t_*(r)$	lower envelope of t	49
u	unit vector	142
w_t	weights	98
$w_{\xi,r}$	Ripley's edge correcting factor	41
x,y	point configuration	5
$x(B)$	number of points in $x \cap B$	5

ROMAN	Usage	Used first on page
$B(\cdot,\cdot)$	birth area rate	116
$B(\cdot)$	total birth rate	116
B	element of $\mathcal{B}$ or $\mathcal{B}_0$	5
$C(\xi \mid x)$	Dirichlet or Voronoi cell	76
$C(\cdot \times \cdot)$	first-order Campell measure	28
$C^{(n)}(\cdot)$	nth order Campell measure	28
$C^{!}(\cdot)$	first-order reduced Campell measure	28
$D(\cdot)$	total death rate	116
D	deviance	138
$D(\theta)$	discrepancy	148
$D(\cdot)$	dispersion of variable	28
E	locally finite subsets of $\mathcal{X}$ with positive density	90
E_n	fixed number subsets of $\mathcal{X}$ with positive density	90
$F(\cdot)$	empty space function	45
$G(\cdot)$	nearest neighbour distribution function	46
G_3	third-order nearest neighbour distribution function	48
$I(\theta)$	Fisher information	151
$J(\cdot)$	scaled version of the F and G-function	46
$K(\cdot)$	Ripley's K-function	34
K_{ij}	extended K-function	47
$L(\cdot)$	rescaled K-function	35
$L(\theta)$	likelihood function	134
L_a^a	lower process	124
M	mark-function for births and deaths	117
$M(\cdot)$	first-order moment measure	27
$M^{(n)}(\cdot)$	nth order moment measure	28
N	locally finite subsets of $\mathcal{X}$	5
$P(\cdot,\cdot)$	Markov transition kernel	91
$P^t(\cdot,\cdot)$	t-step transition probability	91
$\mathrm{PL}_A(\theta;x)$	pseudo-likelihood function	**??**
$Q(\cdot)$	proposal probability of adding a point	103

ROMAN	Usage	Used first on page
R	random field	57
R	simulation region	85
R^*	extended region containing R	89
$R(\cdot,\cdot)$	sub-stochastic kernel	100
R	radius	143
$S(x)$	pairs of neighbours	13
S^{n-1}	unit sphere in $\mathbb{R}^n$	8
T	partition	9
U_a^a	upper process	124
Z_ξ	point process associated to the point ξ	59
X, Y	point process	5
X_{TH}	thinned process	57
X_I	Matérn hard-core process of type I	62
X_{II}	Matérn hard-core process of type II	62
$\{X_t\}$	Markov chain	95
$\{X_t\}$	spatial birth-and-death process	118
W	observation window	23

Abbreviations	Usage	Used first on page
CFTP	Coupling From The Past	120
CLT	Central Limit Theorem	95
CSR	Complete Spatial Randomness	15
DBC	Detailed Balance Condition	100
GQP	General Quadratic Position	76
LGCP	Log Gaussian Cox Process	57
MCMC	Markov Chain Monte Carlo	90
MLE	Maximum Likelihood Estimate	129
MPLE	Maximum Pseudo-Likelihood Estimate	134
PLE	Pseudo-Likelihood Estimate	133
SBD	Spatial Birth-and-Death process	115
SLLN	Strong Law of Large Numbers	95
SSI	Simple Sequential Inhibition process	64

Chapter 1

Introduction

Spatial point processes are stochastic processes that govern the distribution and number of points in some region, which often is a subset of $\mathbb{R}^n$. Each realisation of a point process is called a point pattern where each point, called an event, represents some physical phenomenon. Figure 1.1 shows three such patterns, each observed on a square in $\mathbb{R}^2$.

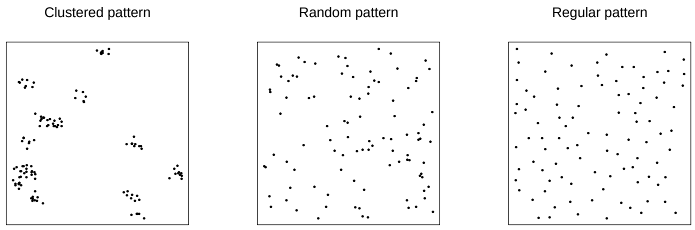

Figure 1.1: From left to right: Clustered, random and regular point patterns.

The primary aim of point process theory is to provide a structured method describing how to find a statistical model which offers a satisfactory explanation to the point pattern that has been observed.

Returning to the three point patterns we see that the first plot is rather clustered. This type of point pattern is often observed when for instance the seeds from a plant is spread in the proximity of the plant. The direct opposite of a clustered point pattern is a regular pattern which is seen in the right hand plot of the figure. Here we have no two events within a certain distance of each other. This could for example be used to describes trees in a forest or the centers of cells in some biological setup. Finally, in the middle plot we see an example of complete spatial randomness, i.e. the events are scattered completely randomly in the observation window.

Outline

Chapter 2 is dedicated to a formal definition of a point process, as well as a characterisation result for this type of stochastic process.

The Poisson point process is a cornerstone in the theory of point processes, and it is the main topic of Chapter 3. We explain how it can be derived naturally from a binomial process and we study different properties of this process. Especially the fundamental and widely utilized property of the Poisson point process that is given by the Slivnyak-Mecke formula is presented and proved in this chapter.

Summary statistics are fundamental elements in model selection and model validation procedures, and they are described in Chapter 4. Prior to introducing two commonly used summary statistics — Ripley's K-function and the pair correlation function g — the concepts of Palm distributions and intensity measures are described. Estimation of Ripley's K-function and the g-function is described for both homogeneous and inhomogeneous point processes, and an estimator of Ripley's K-function is shown to be unbiased for two different edge correcting factors. A short extension of the summary statistics for ordinary processes to marked and multivariate processes is given, and also a brief discussion of model selection and validation is included. Rounding off the chapter are examples of the use of the summary statistics introduced.

In Chapter 5 different classes of models for spatial point patterns are investigated. The log Gaussian Cox process is described as a model based on the ordinary Cox model. Cluster processes includes the Neyman–Scott models which in some cases are identical to the Cox process, and of which the Matérn and Thomas models are special cases. Following a description of hard-core models that prohibit events within a hard-core distance of each other, focus is placed on Markov processes describing how events interact more generally depending on their relative positions. The Strauss process is used as a classic example, and multi-scale and area interaction processes are also described. Further we treat Ord's process, the Soft-core process, and the Widom-Rowlinson point process. Nearest neighbour Markov processes extends the class of Markov models, before the chapter is ended with a description of the Gibbs process.

The research field of spatial point processes has turned out to produce theoretical difficulties, as many of the test statistics are impossible to derive on closed form. This is exemplified by the fact that Ripley's K-function as well as the g-function only have been found explicitly for a very limited number of point process models. The answer to these problems is simulations which is the topic of Chapter 6. Rejection sampling underpins many simulation techniques and thus has pole position in this chapter. Next Markov chain Monte Carlo methods are considered. Basic properties of Markov chains such as convergence criteria and results are described before the Metropolis–Hastings algorithm is introduced as a means of generating a Markov chain with a particular invariant distribution. We describe how to use the Metropolis–Hastings algorithm both in the conditional and unconditional case, i.e. the two situations where the number of points in the process is fixed, and where the number of points is allowed to vary. Then we give a discussion on how Markov chain Monte Carlo is used in practice. Spatial birth-and-death processes are continuous time Markov chains and provides an alternative way of performing simulation. The results given for spatial birth-and-death processes are further used in the final subject treated in this chapter; perfect simulation. In principle Markov chains has to be run infinitely in order to ensure that samples follows the equilibrium distribution exactly. Perfect simulation is a simulation technique that essentially provides a way of letting the Markov chain flag itself when it has reached equilibrium.

Chapter 7 deals with parametric inference, which concerns the estimation of the parameters in a model for a point process. Pseudo-likelihood and contrast-methods are presented, and Markov chain Monte Carlo maximum likelihood estimation including the case of missing data is also studied here. This chapter also includes a short introduction to the gradient algorithm, and a comparison of the effectiveness of the different methods. Finally, we treat parametric bootstrapping, likelihood ratio tests, and Fisher-Wald tests.

Finally, Chapter 8 contains analyses of six different point patterns using the presented theory.

Chapter 2

Point Processes

In this chapter we define, characterize and exemplify the concept of a point process. We consider a general setting which covers the examples of point processes discussed in Chapter 1, as well as those considered later in the report. Further details can be found in Carter & Prenter (1972), Daley & Vere-Jones (1988), and Stoyan, Kendall & Mecke (1995).

2.1 Preliminaries

Let $\mathcal{X}$ be a **separable metric space** with metric d, Borel sets $\mathcal{B}$ (the σ-algebra generated by the open balls), and bounded Borel sets $\mathcal{B}_0$. Separable means that $\mathcal{X}$ contains a countable dense set $\mathcal{V}$, i.e. the closure of $\mathcal{V}$ is $\mathcal{X}$. Throughout this report we assume separability as a means to obtain a convenient level of simplification. Specifically, separability of $\mathcal{X}$ ensures that a central theorem (Theorem 2.13) holds.

Let N be the set of **locally finite** subsets of $\mathcal{X}$; that is $N = \{x \subseteq \mathcal{X} : x(B) < \infty, \forall\, B \in \mathcal{B}_0\}$, where $x(B)$ denotes the number of points in $x \cap B$. Each $x \in N$ is called a **point configuration**, $x = \emptyset$ is the empty point configuration and a point ξ in a point configuration x is called an **event**.

Now equip N with the σ-algebra

$$\mathcal{N} \quad = \quad \sigma\left(\left\{\{x \in N : x(B) = m\} : m \in \mathbb{N}_0, B \in \mathcal{B}_0\right\}\right). \tag{2.1}$$

Thus $f_B : x \to x(B)$ becomes measurable for all $B \in \mathcal{B}_0$, since for all $M \subseteq \mathbb{N}_0$, $f_B^{-1}(M) \in \mathcal{N}$.

Definition 2.1 (Point Process)
*Let $\mathcal{X}$ be a separable metric space with $(N, \mathcal{N})$ as above. If X is a measurable function from some probability space $(\Psi, \Upsilon, \mathbb{P})$ to the measurable space $(N, \mathcal{N})$ then X is called a **point process** on $\mathcal{X}$.*

We shall identify X with its distribution $\mathbb{P}$, and typically $(\Psi, \Upsilon) = (N, \mathcal{N})$ so X is the identity mapping.

Note that we for ease of exposition assume that X is both locally finite and **simple** in the sense that any realisation of X consists of strictly different events. It is possible to relax these conditions but this is not something we will pursue in this report as most examples of point processes for applications will be locally finite and simple. If $X(\mathcal{X}) < \infty$ almost surely then X is called a **finite point process** (see for instance Carter & Prenter (1972)). For $\mathcal{X}$ bounded this will trivially be the case.

To help determine if a function X is measurable and hence a point process, we state the following lemma and proposition.

Lemma 2.2
Let $X : (\Psi, \Upsilon) \to (N, \mathcal{N})$ and $\mathcal{N} = \sigma(\mathcal{A})$. If $X^{-1}(F) \in \Upsilon$ for all $F \in \mathcal{A}$ then X is measurable.

Proof:
We need to verify that $X^{-1}(F) \in \Upsilon$ for all $F \in \mathcal{N}$. Since $\mathcal{N}$ is the σ-algebra generated by $\mathcal{A}$, it suffices to verify that $X^{-1}(F) \in \Upsilon$ if $F = \emptyset$, $F \in \mathcal{A}$, $F = A_1^c$, or $F = \cup_{i=1}^{\infty} A_i$ for any choice of A_i in $\mathcal{N}$. Because Υ is also a σ-algebra we see that

$$\begin{aligned} X^{-1}(\emptyset) &= \emptyset \\ X^{-1}(A_1^c) &= (X^{-1}(A_1))^c \\ X^{-1}(\cup_{i=1}^{\infty} A_i) &= \cup_{i=1}^{\infty} X^{-1}(A_i) \end{aligned}$$

are all members of Υ, hence proving the result. $\circ$

Proposition 2.3
A function $X : (\Psi, \Upsilon) \to (N, \mathcal{N})$ is measurable if and only if for all $B \in \mathcal{B}_0$, $X(B)$ is a random variable (i.e. $X(B)^{-1}(n) \in \Upsilon$ for any $n \in \mathbb{N}_0$).

Proof:
If the mapping $X(B) = (X(\psi))(B)$ from (Ψ, Υ) to $\mathbb{N}_0$ is a random variable for all $B \in \mathcal{B}_0$ then by definition it is measurable so in particular

$$\{\psi \in \Psi : (X(\psi))(B) = m\} \in \Upsilon$$

for any $m \in \mathbb{N}_0, B \in \mathcal{B}_0$.

Further, observing that

$$X^{-1}(\{x \in N : x(B) = m\}) \quad = \quad \{\psi \in \Psi : (X(\psi))(B) = m\} \in \Upsilon$$

we get by Lemma 2.2 that X is measurable.

Assume now that X is measurable, then it follows immediately that $X(B)$ is a random variable. ◦

The following example shows that it is possible to construct a process in accordance with Definition 2.1.

Example 2.4 (Homogeneous Poisson Point Process on $\mathbb{R}_+$)
Let $X_1, X_2, \dots$ be independent and exponentially distributed random variables with parameter $\lambda > 0$. Define a point process X on $\mathcal{X} = \mathbb{R}_+$ equipped with the Euclidean metric as

$$X = \{X_1, X_1 + X_2, \dots\}.$$

Then the underlying measurable space (Ψ, Υ) and X are given by

$$\begin{aligned}
\Psi &= \{(x_1, x_2, \dots) : x_i \in (0, \infty)\,, i = 1, 2, \dots\} \\
\Upsilon &= \sigma\left(\{\{\psi \in \Psi : \psi_1 \in B_1, \dots, \psi_n \in B_n\} : B_1, \dots, B_n \in \mathcal{B}\,, n \in \mathbb{N}\}\right) \\
X &: (x_1, x_2, \dots) \to \{x_1, x_1 + x_2, \dots\}.
\end{aligned}$$

The probability measure $\mathbb{P}$ will be calculated in Example 3.7.

To confirm that X is a point process we need to show that it is simple, locally finite and measurable. It is simple since all $x_i > 0$, so we turn to showing that $X(B) < \infty$ almost surely for all $B \in \mathcal{B}_0$.

For each $B \in \mathcal{B}_0$ there exists a $t > 0$, such that $B \subset (0, t]$. Then $X(B) \leq X((0, t])$, and using the fact that a sum $Y = X_1 + \cdots + X_k$ of independent exponentially distributed variables is gamma distributed we get,

$$\begin{aligned}
\mathbb{P}(X((0,t]) = k) &= \mathbb{P}(X_1 + \cdots + X_k \leq t; X_1 + \cdots + X_{k+1} > t) \\
&= \mathbb{P}(X_1 + \cdots + X_k \leq t; X_{k+1} > t - (X_1 + \cdots + X_k)) \\
&= \int_0^t \frac{\lambda^k}{(k-1)!} y^{k-1} e^{-\lambda y} e^{-\lambda(t-y)} \mathrm{d}y \\
&= \frac{(\lambda t)^k}{k!} e^{-\lambda t}
\end{aligned}$$

for $k \in \mathbb{N}$. Hence $X((0,t]) \sim \mathrm{Poisson}(\lambda t)$ is almost surely finite, so X is locally finite with probability one.

Next we show that X is measurable. According to Lemma 2.2 and Theorem 2.10 below it suffices to show that $\{X(B)=0\} \in \Upsilon$ for $B \in \mathcal{B}_0$. We have that

$$\begin{aligned}\{X(B)=0\} &= \left\{(x_1,x_2,\dots) \;\middle|\; \forall n \in \mathbb{N} : \sum_{i=1}^{n} x_i \notin B\right\} \\ &= \bigcap_{n=1}^{\infty} \left\{(x_1,x_2,\dots) \;\middle|\; \sum_{i=1}^{n} x_i \in B^c\right\} \\ &\in \Upsilon,\end{aligned}$$

because the mapping $(X_1,\dots X_n) \to \sum_{i=1}^{n} X_i$ is measurable and thus X is measurable.

Below we introduce four examples of commonly used point processes: point processes on $\mathbb{R}^n$, S^{n-1}, marked point processes, and multivariate point processes.

Example 2.5 (Point Processes on $\mathbb{R}^n$)
Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a Borel set equipped with the Euclidean metric $d(\xi,\eta)=\|\xi-\eta\|$, where $\xi,\eta \in \mathcal{X}$. In most applications $n=2$ and the events are observed in a rectangle or a disc. For theoretical considerations it is often convenient to take $\mathcal{X}=\mathbb{R}^n$.

For point processes in $\mathbb{R}^n$, we shall often assume stationarity and isotropy. A point process X on $\mathbb{R}^n$ is **stationary**, if its distribution is invariant under translation. That is, the distribution of X is identical to that of $X+v=\{x+v : x \in X\}$, for all $v \in \mathbb{R}^n$. As an analogue to stationarity, the distribution of an **isotropic** point process is invariant under rotation around the origin.

Example 2.6 (Point Processes on a Sphere)
Let $\mathcal{X}=S^{n-1}$, the unit sphere in $\mathbb{R}^n$. The natural metric d is the geodesic distance.

Example 2.7 (Marked Point Processes)
In this example we introduce the setting for marked point processes. Suppose X is a point process on a product space $\mathcal{X}=\mathcal{L}\times\mathcal{M}$, where $\mathcal{L}$ is a space for the locations of events while the 'mark space' $\mathcal{M}$ contains further information about the events. We call X a marked point process and view an event (l_i,m_i) in X as consisting of a location $l_i \in \mathcal{L}$ and a mark $m_i \in \mathcal{M}$. For instance, l_i may indicate the location of a tree observed within a region of a forest while m_i denotes the diameter of the tree in a certain height, i.e. the mark space is $\mathcal{M}=\mathbb{R}_+$. This may correspond to a ***disc process*** *where the location represents the center of a disc and the mark represents the radius. Another example of a marked point process for geometrical objects is a* ***line segment***

process. Here each location is the center of a line segment and each mark $m_i = (r_i, \theta_i) \in \mathbb{R}_+ \times [0, \pi) = \mathcal{M}$ consists of the length r_i of the line segment and the angle θ_i between the line segment and the abscissa.*

To ensure that $\mathcal{X}$ becomes a separable metric space, let $(\mathcal{L}, d_\mathcal{L})$ and $(\mathcal{M}, d_\mathcal{M})$ be separable metric spaces with Borel sets $\mathcal{B}_\mathcal{L}, \mathcal{B}_\mathcal{M}$ respectively. The metric on $\mathcal{X}$ is given by $d = \max(d_\mathcal{L}, d_\mathcal{M})$, i.e.

$$d((l_1, m_1), (l_2, m_2)) = \max(d_\mathcal{L}(l_1, l_2), d_\mathcal{M}(m_1, m_2)).$$

The Borel sets are

$$\mathcal{B} = \mathcal{B}_\mathcal{L} \otimes \mathcal{B}_\mathcal{M} = \sigma(\mathcal{B}_\mathcal{L} \times \mathcal{B}_\mathcal{M}) \tag{2.2}$$

*because of the following. Let $b((l,m), \varepsilon)$ denote the open ball in $\mathcal{X}$ with center (l, m) and radius ε. Thus, $b((l,m), \varepsilon) = b_\mathcal{L}(l, \varepsilon) \times b_\mathcal{M}(m, \varepsilon)$. It can be shown that (**??**) still holds if d is redefined as $d((l_1, m_1), (l_2, m_2)) = \{d_\mathcal{L}(l_1, l_2)^p + d_\mathcal{M}(m_1, m_2)^p\}^{1/p}$ with $p \in \mathbb{N}$.*

Example 2.8 (Multivariate Point Processes)
Let $X = (X_1, \ldots, X_n)$ be a multivariate point process, i.e. the X_i are point processes defined on the same probability space, taking values in $\mathcal{X}$. Observe that a multivariate point process can be considered as a marked point process Y with mark space $\mathcal{M} = \{1, \ldots, n\}$ referring to the n point processes in X, and location space $\mathcal{L} = \mathcal{X}$. Thus the ***superposition*** *$\cup_{i=1}^n X_i$ form the locations of Y. The metric $d_\mathcal{M}$ is the Euclidean distance whereby $\mathcal{B}_\mathcal{M}$ consists of all subsets of $\mathcal{M}$.*

2.2 Characterization of a Point Process

In the following we will establish a way to uniquely characterize a point process by means of its void probabilities.

Definition 2.9 (Void Probabilities)
The ***void probabilities*** *of a point process X are given by*

$$\mathbb{P}(X(B) = 0), \quad B \in \mathcal{B}_0.$$

In order to obtain the main result of this section, Theorem 2.13, we need the following theorem, where we use the concept of partitions. A **partition** of a set A is given by $T = \{T_1, \ldots, T_r\}$ if $A = \cup_{i=1}^r T_i$ and the T_i are disjoint. If S and T are partitions of A then S is **nested** within T if for all $S_i \in S$ there exists $T_j \in T$ so that $S_i \subseteq T_j$. This is written as $S < T$. A partition T of A is said to **separate** two points $a, b \in A$ if $a \in T_i$ and $b \notin T_i$.

Theorem 2.10
The σ-algebra $\mathcal{N}$ in (2.1) is generated by the following set

$$\mathcal{A} \quad = \big\{\{x \in N : x(B) = 0\} : B \in \mathcal{B}_0\big\}.$$

Proof:
Let $\mathcal{N}_0 = \sigma(\mathcal{A})$. By (2.1) it is obvious that $\mathcal{N}_0 \subseteq \mathcal{N}$. It thus remains to be shown that for a given $A \in \mathcal{B}_0$ and $m \in \mathbb{N}_0$,

$$\{x \in N : x(A) = m\} \in \mathcal{N}_0. \tag{2.3}$$

Since $\mathcal{X}$ is separable A contains a dense set $\mathcal{V} = \{v_1, v_2, \dots\}$. Now recursively construct nested partitions T^{nm} of A for $n, m \in \mathbb{N},\ m \le n$ so that they are nested $T^{11} > T^{21} > T^{22} > T^{31} > \cdots$ by

$$\begin{aligned}
T^{11} &= \{A \cap b(v_1, 1), A \backslash b(v_1, 1)\} \\
T^{n1} &= \bigcup_{T_j \in T^{n-1,n-1}} \{T_j \cap (A \cap b(v_1, 2^{1-n})),\ T_j \cap (A \backslash b(v_1, 2^{1-n}))\} \\
T^{nm} &= \bigcup_{T_j \in T^{n,m-1}} \{T_j \cap (A \cap b(v_m, 2^{m-n})),\ T_j \cap (A \backslash b(v_m, 2^{m-n}))\}
\end{aligned}$$

for $n > 1, m = 2, \dots, n$.

Now given any two different events $\xi, \eta \in A$, we show that for n sufficiently large $T^n = T^{nn}$ separates ξ and η. Let $d = d(\xi, \eta)$ where d is an appropriate metric on A. Since $\mathcal{V}$ is dense there exists a point $v_k \in \mathcal{V}$ so that $d(\xi, v_k) < \min\{\frac{1}{3}d, 1\}$. Thus $\xi \in b(v_k, 1)$ and if $\eta \notin b(v_k, 1)$ then T^k separates the two events. Otherwise there exists $l \in \mathbb{N}$ so that $\frac{1}{3}d \le 2^{-l} < \frac{2}{3}d$. Thus $\xi \in b(v_k, 2^{-l})$ and by the triangle inequality $\eta \notin b(v_k, 2^{-l})$ giving that $T^{k+l,k}$, and hence also T^{k+l} separates the two events.

Thus with $T^n = \{T_1^n, \dots, T_{r_n}^n\}$, $x(A) = m$ is equivalent to x having an event in exactly m of the sets T_j^n for n sufficiently large. That is, there exists a number $M \in \mathbb{N}$ such that for all $n \ge M$ there exists m different sets $T_{j_1}^n, \dots, T_{j_m}^n$ in T^n so that x has no events in $A \backslash \cup T_{j_i}^n$, while x has at least one event in each $T_{j_i}^n$. Hence

$$\begin{aligned}
&\{x \in N : x(A) = m\} = \\
&\bigcup_{M=1}^{\infty} \bigcap_{n=M}^{\infty} \bigcup_{1 \le j_1 < \cdots < j_m \le r_n} \left\{x \in N : x(A \backslash \bigcup_{i=1}^{m} T_{j_i}^n) = 0,\ x(T_{j_i}^n) > 0,\ i = 1, \dots, m\right\} \\
&\qquad \in \mathcal{N}_0
\end{aligned}$$

since all T_j^n are bounded Borel sets. $\circ$

In fact, using result in Daley & Vere-Jones (1988) we may even verify the following

Corollary 2.11
Suppose that $\mathcal{X}$ is a complete separable metric space. Then the σ-algebra $\mathcal{N}$ is separable, i.e. countable generated as

$$\mathcal{N} = \sigma\left(\{\{x \in N : x\left(b(v_i, 1/n)\right) = 0\} : i, n \in \mathbb{N}\}\right),$$

where $\mathcal{V} = \{v_1, v_2, \dots\}$ is a dense set in $\mathcal{X}$.

The next result is a subtle extension of Theorem 3.4 in Billingsley (1995), but the proof given there still applies.

Lemma 2.12
Let $\mathcal{F} = \sigma(\mathcal{A})$ be a σ-algebra generated by a set $\mathcal{A}$ which is closed under intersection. If P, Q are measures on $(\Omega, \mathcal{F})$ then

$$P(\Omega) = Q(\Omega) < \infty \quad \textit{and} \quad P = Q \textit{ on } \mathcal{A} \quad \Rightarrow \quad P = Q \textit{ on } \mathcal{F}.$$

Finally we are ready to give the main result of this chapter, which is part of what is known as Choquet's theorem, see Matheron (1975).

Theorem 2.13
The distribution of a point process X is uniquely determined by its void probabilities.

Proof:
By Theorem 2.10, $\mathcal{N}$ is generated by the set $\mathcal{A} = \{\{x \in N : x(B) = 0\} : B \in \mathcal{B}_0\}$ which is clearly closed under intersection. By Lemma 2.12 the distribution $\mathbb{P}$ of X is thus uniquely determined by $\mathbb{P}$ on $\mathcal{A}$ which is exactly the void probabilities. $\circ$

Chapter 3

Poisson Point Processes

This chapter concerns the simplest type of a non-trivial point process; the Poisson point process. The Poisson point process plays a very important role in the construction of most of the other models discussed in this report. First we consider the binomial point process which leads to the Poisson point process. Secondly we study properties of the Poisson point process including a very powerful result known as the Slivnyak-Mecke theorem. Finally we consider finite point processes that are specified by a density with respect to the Poisson point process. A detailed study of Poisson point processes is given in Kingman (1993); see also Daley & Vere-Jones (1988) and Stoyan et al. (1995).

3.1 Binomial Point Processes

To define binomial and later Poisson point processes we need to consider measures on $\mathcal{X}$. A measure μ on $\mathcal{X}$ is called **atomic**, if there exists a point $\xi \in \mathcal{X}$ so that $\mu(\xi) > 0$. Otherwise, the measure is called **non-atomic** or **diffuse**. Further, a measure μ is **locally finite** if $\mu(B) < \infty$ for all $B \in \mathcal{B}_0$. Finally, an event $\xi \in \mathcal{X}$ is μ-**uniformly** distributed on $B \in \mathcal{B}$, with $\mu(B) \in \mathbb{R}_+$, if $\mathbb{P}(\xi \in A) = \frac{\mu(A)}{\mu(B)}$ for all $A \in \mathcal{B}$, $A \subseteq B$.

Suppose now that ν is a non-atomic measure on $\mathcal{X}$, and let $\xi_1, \ldots, \xi_n$ be i.i.d. ν-uniformly distributed on a Borel set $B \in \mathcal{B}$, with $\nu(B) \in \mathbb{R}_+$. Let $X^{(n)}$ denote the union of these n events. We see that $X^{(n)}$ is finite and almost surely simple, and by Proposition 2.3 it is a point process because

$$X^{(n)}(A) \quad = \quad \sum_{i=1}^{n} \mathbb{1}[\xi_i \in A]$$

is a random variable for all $A \in \mathcal{B}$. This process is called a **binomial point process** and it has void probabilities

$$\mathbb{P}(X(A) = 0) \;=\; \left(\frac{\nu(B\backslash A)}{\nu(B)}\right)^n, \quad A \in \mathcal{B}.$$

For example, consider the Euclidean space with ν being the Lebesgue measure. In this case B could be any bounded Borel set.

The binomial point process is only defined on Borel sets B where $\nu(B) < \infty$, but in the following we will consider the asymptotic properties as B is enlarged. Let B_n be a sequence of Borel sets so that $B_i \subseteq B_{i+1}$ and $\nu(B_i) \in \mathbb{R}_+$ for $i \in \mathbb{N}$. Further, assume that $\lambda = \lim_{n\to\infty} \frac{n}{\nu(B_n)} \in \mathbb{R}_+$. For a trivial example, consider $\mathbb{R}^2$ with the Lebesgue measure and let $B_n = b(0, \sqrt{n})$ in which case $\lambda = \pi^{-1}$.

Given $k, l \in \mathbb{N}_0$ and two disjoint bounded Borel sets A and B, consider any $n \geq k + l$ such that $A \cup B \subseteq B_n$. Then

$$\begin{aligned}
&\mathbb{P}\left(X^{(n)}(A) = k, X^{(n)}(B) = l\right) \\
&= \mathbb{P}\left(X^{(n)}(A) = k, X^{(n)}(B) = l, X^{(n)}(B_n\backslash(A\cup B)) = n-(k+l)\right) \\
&= \binom{n}{k}\binom{n-k}{l}\left(\frac{\nu(A)}{\nu(B_n)}\right)^k\left(\frac{\nu(B)}{\nu(B_n)}\right)^l\left(\frac{\nu(B_n\backslash(A\cup B))}{\nu(B_n)}\right)^{n-(k+l)} \\
&= \frac{\nu(A)^k}{k!}\frac{\nu(B)^l}{l!}\frac{n}{\nu(B_n)}\frac{n-1}{\nu(B_n)}\cdots\frac{n-(k+l)+1}{\nu(B_n)} \\
&\quad \times \left(1 - \frac{\frac{\nu(A\cup B)}{\nu(B_n)}(n-(k+l))}{n-(k+l)}\right)^{n-(k+l)}.
\end{aligned}$$

Since

$$\left(1 - \frac{\frac{\nu(A\cup B)}{\nu(B_n)}(n-(k+l))}{n-(k+l)}\right)^{n-(k+l)} \to e^{-\lambda\nu(A\cup B)} \quad \text{when } n \to \infty$$

and $A \cap B = \emptyset$, we see that

$$\lim_{n\to\infty} \mathbb{P}\left(X^{(n)}(A) = k, X^{(n)}(B) = l\right) \;=\; \frac{(\lambda\nu(A))^k}{k!}e^{-\lambda\nu(A)}\frac{(\lambda\nu(B))^l}{l!}e^{-\lambda\nu(B)},$$

i.e. the number of events in two disjoint sets are independent when n tends to infinity. Further in the case where $B = \emptyset$ and $l = 0$ we see that the distribution of events in a bounded Borel set A tends to be Poisson distributed with mean $\lambda\nu(A)$.

3.2 Poisson Point Processes

The result above motivates the following definition where one may think of the measure μ as $\lambda\nu$.

Definition 3.1 (Poisson Point Processes)
*Let μ be a locally finite, non-atomic measure on a separable metric space $\mathcal{X}$. A **Poisson point process** on $\mathcal{X}$ is a point process satisfying the following conditions for all $B \in \mathcal{B}_0$,*

(P1) Poisson distribution: $X(B) \sim \mathit{Poisson}(\mu(B))$.

(P2) Independent scattering: *For all disjoint sets $A_1, \ldots, A_k \in \mathcal{B}_0$ $X(A_1), \ldots, X(A_k)$ are mutually independent.*

*We call μ the **intensity measure** of X since $\mu(B) = \mathbb{E}(X(B))$.*

If $\mathcal{X} = \mathbb{R}^n$ and μ has a density λ with respect to the Lebesgue measure, that is $\mu(B) = \int_B \lambda(\xi)\mathrm{d}\xi$, then λ is called the **intensity function** of X. If λ is constant on $\mathcal{X}$ then it is called the **intensity** or **rate** and X is a **homogeneous** Poisson point process, which is often referred to as **complete spatial randomness** (CSR). A homogeneous Poisson point process is thus both stationary and isotropic, since the Lebesgue measure is invariant under translation and rotation. If X is not homogeneous, it is called **inhomogeneous**.

In Theorem 3.2 it will be shown that there exists Poisson point processes. By (P1) a Poisson point process X has void probabilities $\mathbb{P}(X(B) = 0) = \exp(-\mu(B))$, and since Theorem 2.13 shows that the void probabilities uniquely determine the point process, we have that (P1) implies (P2). However, (P2) is included in the definition since it gives a good characterization of a Poisson point process, and also since it is needed for point processes where multiple points are allowed. It can be verified that for $\mathcal{X} \subseteq \mathbb{R}^n$, (P2) together with stationarity imply (P1), cf. Stoyan et al. (1995, p. 35).

Theorem 3.2 (Existence of a Poisson Point Process)
There exists a unique point process which satisfies Definition 3.1 and thus is a Poisson point process.

Proof:
First notice, that the uniqueness follows from the existence by Theorem 2.13.

Given a locally finite measure μ on a separable metric space $\mathcal{X}$, let $o \in \mathcal{X}$ be a fixed point. Then set $T_1 = b(o, 1)$ and $T_i = b(o, i) \backslash b(o, i-1)$ for $i = 2, 3, \ldots$ This gives a partition of $\mathcal{X}$ in disjoint sets $\{T_i\} \subset \mathcal{B}_0$.

Let $N_i \sim \text{Poisson}(\mu(T_i))$ for $i \in \mathbb{N}$ be mutually independent, with $N_i \equiv 0$ if $\mu(T_i) = 0$. For each $i \in \mathbb{N}$, conditional on $N_i = n_i$, generate a binomial process $X_i = \{X_{i1}, \ldots, X_{in_i}\}$ with the n_i points i.i.d. and μ-uniform on T_i. We thus have that $\mathbb{P}(X_{ij} \in A \mid N_i = n_i) = \frac{\mu(A)}{\mu(T_i)}$ for any Borel set $A \subseteq T_i$.

Conditional on all N_i, these binomial processes are assumed to be mutually independent. Now define X and thus $X(B)$ as

$$X = \bigcup_{i=1}^{\infty} \{X_{i1}, \ldots, X_{iN_i}\} \quad \text{and} \quad X(B) \quad = \quad \sum_{i=1}^{\infty} \sum_{j=1}^{N_i} \mathbb{1}[X_{ij} \in B].$$

To prove that X is a point process it suffices by Proposition 2.3 to show that $X(B)$ is a random variable for all $B \in \mathcal{B}_0$. First $\mathbb{1}[X_{ij} \in B]$ is a random variable, and since N_i can take only countably many values, so is $\sum_{j=1}^{N_i} \mathbb{1}[X_{ij} \in B]$. As B is bounded, only finitely many T_i contribute to $X(B)$, and thus $X(B)$ is a random variable.

To show that X satisfies property (P1) let $k \in \mathbb{N}_0$ and $B \in \mathcal{B}_0$. Then (taking $\frac{0}{0} = 0$ and $0^0 = 1$)

$$\begin{aligned}
&\mathbb{P}(X(T_i \cap B) = k) \\
&= \sum_{n=k}^{\infty} \mathbb{P}(N_i = n)\ \mathbb{P}(X(T_i \cap B) = k \mid N_i = n) \\
&= \sum_{n=k}^{\infty} \frac{\mu(T_i)^n}{n!}\ e^{-\mu(T_i)} \binom{n}{k} \left(\frac{\mu(T_i \cap B)}{\mu(T_i)}\right)^k \left(\frac{\mu(T_i \backslash B)}{\mu(T_i)}\right)^{n-k} \\
&= \frac{\mu(T_i \cap B)^k}{k!}\ e^{-\mu(T_i)} \sum_{n=k}^{\infty} \frac{\mu(T_i \backslash B)^{n-k}}{(n-k)!} \\
&= \frac{\mu(T_i \cap B)^k}{k!}\ e^{-\mu(T_i)}\ e^{\mu(T_i \backslash B)} \\
&= \frac{\mu(T_i \cap B)^k}{k!}\ e^{-\mu(T_i \cap B)}.
\end{aligned}$$

So $X(T_i \cap B) \sim \text{Poisson}(\mu(T_i \cap B))$ for all $i \in \mathbb{N}$ and they are mutually independent by the construction. Thus, since only finite many T_i intersect B,

$$\begin{aligned}
X(B) \quad &= \sum_{i: T_i \cap B \neq \emptyset} X(T_i \cap B) \ \sim\ \text{Poisson}\left(\sum_{i: T_i \cap B \neq \emptyset} \mu(T_i \cap B)\right) \\
&\sim\ \text{Poisson}(\mu(B))
\end{aligned}$$

which proves that X satisfies (P1).

Regarding property (P2) we for ease of exposition only treat the case of two sets but the proof generalizes easily to any number of sets.

Let $A, B \in \mathcal{B}_0$ be disjoint and let $k, l \in \mathbb{N}_0$. Observe that $X(T_i \cap A)$ and $X(T_j \cap B)$ by construction are independent for $i \neq j$. For $i = j$ we get

$$\begin{aligned}
&\mathbb{P}\big(X(T_i \cap A) = k; X(T_i \cap B) = l\big) \\
&\quad = \sum_{n=k+l}^{\infty} \mathbb{P}\big(X(T_i \cap A) = k; X(T_i \cap B) = l \mid N_i = n\big)\ \mathbb{P}(N_i = n) \\
&\quad = \sum_{n=k+l}^{\infty} \binom{n}{k}\binom{n-k}{l}\left(\frac{\mu(T_i \cap A)}{\mu(T_i)}\right)^k \left(\frac{\mu(T_i \cap B)}{\mu(T_i)}\right)^l \\
&\qquad\qquad \times \left(\frac{\mu(T_i \backslash (A \cup B))}{\mu(T_i)}\right)^{n-k-l} \frac{\mu(T_i)^n}{n!} e^{-\mu(T_i)} \\
&\quad = \frac{\mu(T_i \cap A)^k}{k!}\frac{\mu(T_i \cap B)^l}{l!} e^{-\mu(T_i)} \sum_{n=k+l}^{\infty} \frac{\mu(T_i \backslash (A \cup B))^{n-k-l}}{(n-k-l)!} \\
&\quad = \frac{\mu(T_i \cap A)^k}{k!}\frac{\mu(T_i \cap B)^l}{l!} e^{-\mu(T_i)} e^{\mu(T_i \backslash (A \cup B))} \\
&\quad = \frac{\mu(T_i \cap A)^k}{k!} e^{-\mu(T_i \cap A)} \frac{\mu(T_i \cap B)^l}{l!} e^{-\mu(T_i \cap B)} \\
&\quad = \mathbb{P}(X(T_i \cap A) = k)\ \mathbb{P}(X(T_i \cap B) = l).
\end{aligned}$$

This shows that $X(T_i \cap A)$ is independent of $X(T_i \cap B)$ for all i and thus $X(A)$ is independent of $X(B)$ which concludes (P2). $\circ$

The following propositions give four basic results for Poisson point processes, which are easy consequences of Theorem 2.13.

Proposition 3.3
Let X be a Poisson point process on $\mathcal{X}$ with intensity measure μ and let $A \in \mathcal{B}$. Then $X_A = X \cap A$ is a Poisson point process on A with intensity measure $\mu_A(\cdot) = \mu(\cdot \cap A)$.

Proof:
By Proposition 2.3, X_A is clearly a point process. Let $B \in \mathcal{B}_0(A)$ (the bounded Borel sets on A). Then

$$X_A(B) = X(B \cap A) = X(B) \sim \text{Poisson}(\mu(B)),$$

and since $\mu(B \cap A) = \mu(B)$, X_A satisfies (P1) and so X_A is a Poisson point process with intensity measure μ_A. $\circ$

Proposition 3.4
Let Y and Z be independent Poisson point processes on $\mathcal{X}$ with intensity measures μ_Y and μ_Z respectively. Then the superposition

$$X = Y \cup Z$$

is a Poisson point process with intensity measure $\mu = \mu_Y + \mu_Z$.

Proof:
Obviously, X is a point process, so in order to prove the result we calculate the void probabilities. Let $B \in \mathcal{B}_0$, then

$$\begin{aligned}\mathbb{P}(X(B)=0) &= \mathbb{P}(Y(B)=0; Z(B)=0)\\ &= \mathbb{P}(Y(B)=0)\mathbb{P}(Z(B)=0)\\ &= e^{-\mu_Y(B)}e^{-\mu_Z(B)} = e^{-(\mu_Y(B)+\mu_Z(B))}\\ &= e^{-\mu(B)},\end{aligned}$$

i.e. the void probability for a Poisson point process with intensity measure $\mu = \mu_Y + \mu_Z$. ∘

By induction the proposition naturally generalizes to the superposition of a finite number of independent Poisson point processes.

Proposition 3.5
Suppose X *is a Poisson point process on* $\mathcal{X}$ *with intensity measure* μ. *Let* $B \in \mathcal{B}_0$ *with* $\mu(B) > 0$ *and* $n \in \mathbb{N}_0$. *Conditional on* $X(B) = n$, X_B *is a binomial point process with* n *points.*

Proof:
Let $A \subseteq B$, $A \in \mathcal{B}_0$. Then

$$\begin{aligned}\mathbb{P}(X(A)=0 \mid X(B)=n) &= \frac{\mathbb{P}(X(A)=0, X(B\backslash A)=n)}{\mathbb{P}(X(B)=n)}\\ &= \frac{e^{-\mu(A)}\frac{\mu(B\backslash A)^n}{n!}e^{-\mu(B\backslash A)}}{\frac{\mu(B)^n}{n!}e^{-\mu(B)}}\\ &= \left(\frac{\mu(B\backslash A)}{\mu(B)}\right)^n,\end{aligned}$$

i.e. the void probabilities of a binomial process with n points. ∘

Proposition 3.6
Let X *be a Poisson point process on* $\mathcal{X}$ *with intensity measure* μ *and let* $A \in \mathcal{B}$. *Let* $r : \mathcal{X} \to [0,1]$ *be measurable and set*

$$Y = \{x_i \in X \cap A : Z_i < r(x_i)\}$$

where the Z_i *are i.i.d. uniformly on* $(0,1)$ *and independent of* X. *Then* Y *is a Poisson point process with intensity measure given by* $\tilde{\mu}(B) = \int_B r(w)\mathrm{d}\mu(w)$ *for* $B \in \mathcal{B}_0(A)$.

Proof:
It can be shown that Y is a point process – but here we restrict attention to proving that Y is a Poisson point process assuming that it is a point process. This is done by examining the void probabilities.

Let $B \in \mathcal{B}_0(A)$ and $X = \{X_1, \dots X_n\}$. If $\mu(B) > 0$ then the void probabilities for Y are

$$\begin{aligned}
\mathbb{P}(Y(B) = 0) &= \sum_{n=0}^{\infty} \mathbb{P}(X(B) = n)\mathbb{P}(Y(B) = 0 \mid X(B) = n) \\
&= \sum_{n=0}^{\infty} \frac{\mu(B)^n}{n!} e^{-\mu(B)} \prod_{i=1}^{n} \mathbb{P}(Z_i > r(X_i)) \\
&= \sum_{n=0}^{\infty} \frac{\mu(B)^n}{n!} e^{-\mu(B)} \left(\int_0^1 \int_B \mathbb{1}[z > r(w)] \frac{1}{\mu(B)} \mathrm{d}\mu(w)\mathrm{d}z \right)^n \\
&= \sum_{n=0}^{\infty} \frac{\mu(B)^n}{n!} e^{-\mu(B)} \left(\frac{1}{\mu(B)} \int_B 1 - r(w)\mathrm{d}\mu(w) \right)^n \\
&= e^{-\mu(B)} e^{\int_B (1-r(w))\mathrm{d}\mu(w)} = e^{-\int_B r(w)\mathrm{d}\mu(w)} \\
&= e^{-\tilde{\mu}(B)},
\end{aligned}$$

i.e. the void probabilities of a Poisson point process with intensity measure $\tilde{\mu}$. Note that if $\mu(B) = 0$, then $\tilde{\mu}(B) = 0$ and we obviously get

$$\mathbb{P}(Y(B) = 0) = e^{-\tilde{\mu}(B)} = 1.$$

$\circ$

Now we consider two examples of Poisson point processes.

Example 3.7 (Homogeneous Poisson Point Process on $\mathbb{R}_+$)
Using the setup in Example 2.4 we now show that the point process X treated there is a Poisson point process on $\mathbb{R}_+$ as might be expected. We have seen that $\mathbb{P}(X((0,t]) = 0) = \exp(-\lambda t)$ and using similar arguments it can be shown that $\mathbb{P}(X((s,t]) = 0) = \exp(-\lambda(t-s))$ which implies that $\mathbb{P}(X(B) = 0) = \exp(-\lambda|B|)$ for any bounded Borel set B of $\mathbb{R}_+$. Therefore X is a homogeneous Poisson point process with rate λ on $\mathbb{R}_+$.

Example 3.8 (Marked Poisson Point Processes)
*Let $X = \{(l_i, m_i)\}$ be a marked point process on $\mathcal{L} \times \mathcal{M}$, as in Example 2.7. Assume $L = \{(l_i)\}$ is a Poisson point process on $\mathcal{L}$ and that conditional on L, the marks are i.i.d. random variables taking values in $\mathcal{M}$. Then X is said to be a **marked Poisson point process**.*

To see that X is actually a Poisson point process, we examine its void probabilities. Let ν be the intensity measure of L and let $\mathcal{M}$ be the common distribution of the marks. Now let $B \in \mathcal{B}_0$ and $B_L \in \mathcal{B}_0(\mathcal{L})$ so that $\{l \mid \exists\, m : (l,m) \in B\} \subseteq B_L$. If $\nu(B_L) = 0$ we trivially have that

$$\mathbb{P}(X(B) = 0) \;=\; e^{-\mu(B)},$$

where $\mu = \nu \times \mathcal{M}$. Assuming that $\nu(B_L) > 0$, the void probabilities are given by

$$\mathbb{P}(X(B) = 0) \;=\; \sum_{n=0}^{\infty} \mathbb{P}(L(B_L) = n)\,\mathbb{P}(X(B) = 0 \mid L(B_L) = n).$$

The right-most probability reduces to $\mathbb{P}((l,m) \notin B)^n$, since by Proposition 3.5, conditioning on $L(B_L) = n$, L_{B_L} is a binomial point process and hence $l \sim \nu$-uniform(B_L) are i.i.d. for all $l \in L_{B_L}$. Further, $m \sim \mathcal{M}(\mathcal{M})$ are i.i.d. by assumption. Hence

$$\begin{aligned}
\mathbb{P}(X(B) = 0) &= \sum_{n=0}^{\infty}\left[\frac{e^{-\nu(B_L)}\nu(B_L)^n}{n!}\mathbb{P}\big((l,m) \notin B\big)^n\right] \\
&= \sum_{n=0}^{\infty}\left[\frac{e^{-\nu(B_L)}\nu(B_L)^n}{n!}\,\big(1 - \mathbb{P}\big((l,m) \in B\big)\big)^n\right] \\
&= e^{-\nu(B_L)}\sum_{n=0}^{\infty}\frac{\nu(B_L)^n}{n!}\left(1 - \int_{B_L}\int_{\mathcal{M}} \mathbb{1}[(l,m) \in B]\,\mathrm{d}\mathcal{M}(m)\frac{\mathrm{d}\nu(l)}{\nu(B_L)}\right)^n \\
&= e^{-\nu(B_L)}\sum_{n=0}^{\infty}\frac{\nu(B_L)^n}{n!}\left(1 - \frac{\mu(B)}{\nu(B_L)}\right)^n \\
&= e^{-\mu(B)}.
\end{aligned}$$

Hence X is a Poisson point process with intensity measure μ given by the product measure $\mu = \nu \times \mathcal{M}$.

For instance for a disc process, $\mathcal{M} = \mathbb{R}_+$, see Example 2.7. If we let $Z = \cup_{(l_i,m_i)\in X}\overline{b}(l_i,m_i)$, where $\overline{b}(\cdot,\cdot)$ denotes the closure of $b(\cdot,\cdot)$, then Z is an example of a Boolean model, c.f. Stoyan et al. (1995).

Next we show a theorem due to Slivnyak and Mecke which gives a convenient property of a Poisson point process. This property will be used several times later in this report.

Theorem 3.9 (Slivnyak-Mecke Formula)
Let $h : \mathcal{X}^n \times N \to [0,\infty)$ be a measurable function. For a Poisson point process

X with intensity measure μ,

$$\begin{aligned}
&\mathbb{E} \sum_{\xi_1,\dots,\xi_n \in X}^{\neq} h(\xi_1,\dots,\xi_n, X\backslash\{\xi_1,\dots,\xi_n\}) \\
&= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{E} h(\xi_1,\dots,\xi_n, X)\mathrm{d}\mu(\xi_1)\cdots\mathrm{d}\mu(\xi_n).
\end{aligned}$$

Proof:
We note that the theorem is equivalent to

$$\begin{aligned}
&\mathbb{E} \sum_{\xi_1,\dots,\xi_n \in X}^{\neq} h(\xi_1,\dots,\xi_n, X) \\
&= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{E} h(\xi_1,\dots,\xi_n, X \cup \{\xi_1,\dots,\xi_n\})\mathrm{d}\mu(\xi_1)\cdots\mathrm{d}\mu(\xi_n)
\end{aligned}$$

and proceed to prove this by induction. For $n = 1$ we first verify that

$$\mathbb{E} \sum_{\xi \in X \cap C} h(\xi, X) \quad = \quad \int_C \mathbb{E} h(\xi, X \cup \xi)\mathrm{d}\mu(\xi), \quad C \in \mathcal{B}_0, \tag{3.1}$$

where $X \cup \xi$ is the notation used for $X \cup \{\xi\}$. Since h is a non-negative measurable function it suffices to consider indicator functions $\mathbb{1}_F$, for $F \in \mathcal{B} \otimes \mathcal{N}$, c.f. Billingsley (1995, p. 184-185). This will henceforth be referred to as *standard measure theoretical arguments*. For such indicator functions which are of the special form

$$h(\xi, X) \quad = \quad \mathbb{1}[\xi \in A, X(B) = 0], \quad A, B \in \mathcal{B}_0 \tag{3.2}$$

both the left-hand side and right-hand side of (3.1) are measures on the σ-algebra $\mathcal{B} \otimes \mathcal{N}$. Naming these measures μ_1 and μ_2 respectively, we get that

$$\begin{aligned}
\mu_1(\mathcal{X} \times N) &= \mathbb{E} \sum_{\xi \in X \cap C} 1 = \mathbb{E} X(C) = \mu(C) < \infty \\
\mu_2(\mathcal{X} \times N) &= \int_C \mathbb{E}\, 1 \,\mathrm{d}\mu(\xi) = \mu(C),
\end{aligned}$$

that is μ_1 and μ_2 have the same upper bound.

Since $\{\{\xi \in A, X(B) = 0\} : A, B \in \mathcal{B}_0\}$ is closed under intersection and generates $\mathcal{B} \otimes \mathcal{N}$ then by Lemma 2.12 it suffices to show (3.1) for h on the form given in (3.2). We have that

$$
\begin{aligned}
\mathbb{E}\sum_{\xi\in X\cap C} h(\xi,X) &= \mathbb{E}(X(A\cap C)\mathbb{1}[X(B)=0]) \\
&= \mathbb{E}(X((A\cap C)\backslash B)\ \mathbb{1}[X(B)=0]) \\
&= \mathbb{E}(X((A\cap C)\backslash B))\ \mathbb{P}(X(B)=0) \quad \text{using (P2)} \\
&= \mu((A\cap C)\backslash B)\ \mathbb{P}(X(B)=0) \\
&= \int_{(A\cap C)\backslash B} \mathbb{P}(X(B)=0)\mathrm{d}\mu(\xi) \\
&= \int_{A\cap C} \mathbb{P}((X\cup\xi)(B)=0)\mathrm{d}\mu(\xi) \\
&= \int_{C} \mathbb{E}h(\xi,X\cup\xi)\mathrm{d}\mu(\xi).
\end{aligned}
$$

Let $\{C_i\}$ be a partition of $\mathcal{X}$ similarly to $\{T_i\}$ from the proof of Theorem 3.2. By the monotone convergence theorem we now obtain that

$$
\begin{aligned}
\mathbb{E}\sum_{\xi\in X} h(\xi,X) &= \mathbb{E}\sum_{i\in\mathbb{N}}\sum_{\xi\in X\cap C_i} h(\xi,X) \\
&= \sum_{i\in\mathbb{N}}\int_{C_i} \mathbb{E}h(\xi,X\cup\xi)\mathrm{d}\mu(\xi) \\
&= \int_{\mathcal{X}} \mathbb{E}h(\xi,X\cup\xi)\mathrm{d}\mu(\xi).
\end{aligned}
$$

Finally assume that the formula holds for $n-1\geq 0$. Then

$$
\begin{aligned}
&\mathbb{E}\sum_{\xi_1,\dots,\xi_n\in X}^{\neq} h(\xi_1,\dots,\xi_n,X) \\
&= \mathbb{E}\sum_{\xi_1,\dots,\xi_{n-1}\in X}^{\neq} \underbrace{\left(\sum_{\xi_n\in X\backslash\{\xi_1,\dots,\xi_{n-1}\}} h(\xi_1,\dots,\xi_n,X)\right)}_{f(\xi_1,\dots,\xi_{n-1},X)} \\
&= \int_{\mathcal{X}}\cdots\int_{\mathcal{X}} \mathbb{E}f(\xi_1,\dots,\xi_{n-1},X\cup\{\xi_1,\dots,\xi_{n-1}\})\mathrm{d}\mu(\xi_1)\cdots\mathrm{d}\mu(\xi_{n-1}),
\end{aligned}
$$

by the induction hypothesis. Further

$$\begin{aligned}
&\mathbb{E}f(\xi_1,\dots,\xi_{n-1},X\cup\{\xi_1,\dots,\xi_{n-1}\})\\
&\quad= \mathbb{E}\sum_{\xi_n\in X} h(\xi_1,\dots,\xi_n,X\cup\{\xi_1,\dots,\xi_{n-1}\})\\
&\quad= \int_{\mathcal{X}} \mathbb{E}h(\xi_1,\dots,\xi_n,X\cup\{\xi_1,\dots,\xi_n\})\mathrm{d}\mu(\xi_n)
\end{aligned}$$

and by Fubini's theorem we obtain the result. ∘

3.3 Finite Point Processes Specified by a Density

Later in this report we will consider finite point processes that have a density with respect to the distribution of a Poisson point process with finite intensity measure μ. We restrict attention to finite intensity measures without any great loss of generality because in applications we usually observe a point process within a window $W \in \mathcal{B}_0$, and restricting the intensity measure to W, it becomes finite. For ease of exposition, we therefore assume that $0 < \mu(\mathcal{X}) < \infty$.

Let Π denote the distribution of a Poisson point process X with intensity measure μ on $\mathcal{X}$. Then Π is a measure on $(N, \mathcal{N})$ given by (where we replace the integrals with $\mathbb{1}[\emptyset \in F]$ for $n = 0$)

$$\begin{aligned}
\Pi(F) &= \mathbb{P}(X\in F)\\
&= \sum_{n=0}^{\infty}\mathbb{P}(X(\mathcal{X})=n)\ \mathbb{P}(X\in F \mid X(\mathcal{X})=n)\\
&= \sum_{n=0}^{\infty}\frac{\mu(\mathcal{X})^n}{n!}e^{-\mu(\mathcal{X})}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}\mathbb{1}[\{x_1,\dots,x_n\}\in F]\frac{\mathrm{d}\mu(x_1)}{\mu(\mathcal{X})}\cdots\frac{\mathrm{d}\mu(x_n)}{\mu(\mathcal{X})}\\
&= \sum_{n=0}^{\infty}\frac{e^{-\mu(\mathcal{X})}}{n!}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}\mathbb{1}[\{x_1,\dots,x_n\}\in F]\mathrm{d}\mu(x_1)\cdots\mathrm{d}\mu(x_n),
\end{aligned}$$

for $F \in \mathcal{N}$ using Proposition 3.5 to obtain the third equality.

The processes we will consider have a density p w.r.t. Π, i.e.

$$\mathbb{P}(X\in F) = \int_F p(x)\mathrm{d}\Pi(x).$$

This density will often only be given up to proportionality. It is therefore natural to ask when a non-negative measurable function h is Π-integrable. First we notice by standard measure theoretical arguments that

$$\int_F h(x)\mathrm{d}\Pi(x) = \sum_{n=0}^{\infty} \frac{e^{-\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{1}[x \in F] h(x) \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n), \tag{3.3}$$

where $x = \{x_1, \dots, x_n\}$.

We now give two conditions that each is sufficient to ensure that h is Π-integrable.

Definition 3.10 (Stability Conditions)
*A non-negative measurable function h on N is **locally stable** if*

$$\exists K > 0, \forall x \in N, \xi \in \mathcal{X} : h(x \cup \xi) \leq K h(x)$$

*and **Ruelle stable** if*

$$\exists K > 0, c > 0, \forall x \in N : h(x) \leq c K^{x(\mathcal{X})}.$$

Proposition 3.11
Local stability implies Ruelle stability which in turn implies Π-integrability.

Proof:
Assume that h is locally stable and let $c = h(\emptyset)$. Then it easily follows by induction in $x(\mathcal{X})$ that h is Ruelle stable. Now assume that h is Ruelle stable. Then by (3.3),

$$\begin{aligned}
\int_N h(x)\mathrm{d}\Pi(x) &= \sum_{n=0}^{\infty} \frac{e^{-\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{1}[x \in N]\ h(x)\ \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n) \\
&\leq \sum_{n=0}^{\infty} \frac{e^{-\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} cK^n\ \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n) \\
&= c e^{-\mu(\mathcal{X})} \sum_{n=0}^{\infty} \frac{(K\mu(\mathcal{X}))^n}{n!} \\
&= c e^{(K-1)\mu(\mathcal{X})} \\
&< \infty,
\end{aligned}$$

and hence h is Π-integrable. ◦

Example 3.12
Let $h(x) = \beta^{x(\mathcal{X})}, \beta > 0$. *Clearly* h *is locally stable and thus* Π*-integrable, and by* (3.3)

$$\begin{aligned}
\int_N h(x) \mathrm{d}\Pi(x) &= \sum_{n=0}^{\infty} \frac{e^{-\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{1}[x \in N] \; \beta^n \; \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n) \\
&= \sum_{n=0}^{\infty} \frac{e^{-\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \beta^n \; \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n) \\
&= e^{-\mu(\mathcal{X})} \sum_{n=0}^{\infty} \frac{(\beta\mu(\mathcal{X}))^n}{n!} \\
&= e^{(\beta-1)\mu(\mathcal{X})},
\end{aligned}$$

where the second equality is due to the fact that μ *is non-atomic. Hence*

$$p(x) \quad = \quad e^{(1-\beta)\mu(\mathcal{X})} \beta^{x(\mathcal{X})}$$

is a density w.r.t. Π.

Finally the distribution of a point process X *with density* p *with respect to* Π *is given by*

$$\begin{aligned}
&\mathbb{P}(X \in F) \\
&= \sum_{n=0}^{\infty} \frac{e^{-\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{1}[x \in F] \; e^{(1-\beta)\mu(\mathcal{X})} \beta^n \; \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n) \\
&= \sum_{n=0}^{\infty} \frac{e^{-\beta\mu(\mathcal{X})}}{n!} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mathbb{1}[x \in F] \; \mathrm{d}(\beta\mu)(x_1) \cdots \mathrm{d}(\beta\mu)(x_n),
\end{aligned}$$

so X *is a Poisson point process with intensity measure* $\beta\mu$.

Chapter 4

Summary Statistics

In the process of finding a suitable model to describe a spatial point pattern, summary statistics are an important tool. They provide different kinds of information about the point pattern in question; information that can be related to characteristics of existing models in order to help identify a suitable model for the point pattern studied. First we introduce some theoretical concepts and afterwards we turn to specific summary statistics including the K- and g-functions among others. During the presentation of the functions we also consider how to estimate them. We then consider how to use summary statistics for model selection and validation and finally give some examples of the use of the summary statistics introduced. Although we introduce the terms in a very general setting, we lay the main emphasis on point processes in $\mathbb{R}^n$, as this theory is very applicable when confronted with a spatial point pattern in practice. Further coverage of summary statistics may be found in Ripley (1976), Ripley (1977), Diggle (1983), Stoyan & Stoyan (1994), and Stoyan et al. (1995).

4.1 Intensity Measures and Palm Distributions

We first consider the general setup where $\mathcal{X}$ is any separable metric space, and then turn attention to the case $\mathcal{X} = \mathbb{R}^n$.

The **first-order moment measure** of a point process X is given by

$$M(A) \quad = \quad \mathbb{E}X(A), \quad \text{for any } A \in \mathcal{B}_0.$$

Often, the first order moment measure is called the **intensity measure**, extending the definition given in Chapter 3. The **first-order Campbell mea-**

sure is defined as

$$\begin{aligned} C(A \times F) &= \mathbb{E}(X(A)\mathbb{1}[X \in F]) \\ &= \mathbb{E}\sum_{\xi \in X} \mathbb{1}[\xi \in A, X \in F], \end{aligned}$$

where $A \in \mathcal{B}_0$ and $F \in \mathcal{N}$.

The order of the moment and Campbell measures can be extended to an arbitrary order, where the **nth order moment measure** is given by

$$M^{(n)}(A_1 \times \cdots \times A_n) = \mathbb{E}(X(A_1)\cdots X(A_n)), \qquad A_i \in \mathcal{B}_0 \text{ for } i = 1,\ldots,n,$$

and the **nth-order Campbell measure** is

$$C^{(n)}(A_1 \times \cdots \times A_n \times F) = \mathbb{E}(X(A_1)\cdots X(A_n)\mathbb{1}[X \in F])$$

where $A_i \in \mathcal{B}_0$ for $i = 1,\ldots,n$, and $F \in \mathcal{N}$.

The moment and the Campbell measure of order n are related since $C^{(n)}(A_1 \times \cdots \times A_n \times N) = \mathbb{E}(X(A_1)\cdots X(A_n)) = M^{(n)}(A_1 \times \cdots \times A_n)$, and thus $C^{(n)}$ is finite if and only if $M^{(n)}$ is finite.

Further, the **first-order reduced Campbell measure** for a point process X on $A \times F$ is defined by

$$C^{!}(A \times F) = \mathbb{E}\sum_{\xi \in X} \mathbb{1}[\xi \in A,\ X\backslash\xi \in F], \quad A \in \mathcal{B}_0,\ F \in \mathcal{N}.$$

Having defined the moment measures, we see that the variance of $X(A)$ for $A \in \mathcal{B}_0$ is given by

$$\mathrm{Var}X(A) = \mathbb{E}(X(A)^2) - (\mathbb{E}X(A))^2 = M^{(2)}(A \times A) - M(A)^2.$$

Further, the **dispersion** of the variable $X(A)$ is given by

$$D(X(A)) = \frac{\mathrm{Var}(X(A))}{\mathbb{E}(X(A))} = \frac{\mathrm{Var}(X(A))}{M(A)}.$$

The variable $X(A)$ is **over dispersed** if $D(X(A)) > 1$ and **under dispersed** if $D(X(A)) < 1$. For a homogeneous Poisson point process $D(X(A)) = 1$ since $X(A)$ has a Poisson distribution.

Further, for $n \in \mathbb{N}$ the **nth-order factorial moment measure** $\alpha^{(n)}$ is given by

$$\alpha^{(n)}(A_1 \times \cdots \times A_n) = \mathbb{E}\left(\sum_{\xi_1,\ldots,\xi_n \in X}^{\neq} \mathbb{1}[\xi_1 \in A_1,\ldots,\xi_n \in A_n]\right), \qquad (4.1)$$

for $A_1, \ldots, A_n \in \mathcal{B}_0$. Note that $\alpha^{(n)}$ resembles $M^{(n)}$ closely, and in fact $\alpha^{(1)} = M^{(1)}$. If $h \geq 0$ is measurable, standard measure theoretical arguments gives that

$$\mathbb{E}\left(\sum_{\xi_1,\ldots,\xi_n\in X}^{\neq} h(\xi_1,\ldots,\xi_n)\right) = \int_{\mathcal{X}}\cdots\int_{\mathcal{X}} h(\xi_1,\ldots,\xi_n)\mathrm{d}\alpha^{(n)}(\xi_1,\ldots,\xi_n). \tag{4.2}$$

Conditional Distributions

Next we introduce the Palm distribution of a point process, which can be interpreted as the conditional distribution of X on $\mathcal{X}$ given that there is an event at ξ. One has to be careful when defining this conditional distribution as the probability of there being an event at ξ equals 0 almost surely. It is defined as follows.

Assume that M is a **σ-finite measure** on $\mathcal{X}$, that is $\mathcal{X} = \cup_{i=1}^{\infty} A_i$ with $M(A_i) < \infty$. For each $F \in \mathcal{N}$, $C(\cdot \times F)$ is **absolutely continuous** with respect to M, i.e. $M(A) = 0$ implies $C(A \times F) = 0$, for each $A \in \mathcal{B}_0$. So letting

$$\mathbb{P}_{\xi}(F) = \frac{\mathrm{d}C(\cdot \times F)}{\mathrm{d}M}(\xi) \tag{4.3}$$

we find that

$$C(A \times F) = \int_A \mathbb{P}_{\xi}(F)\mathrm{d}M(\xi).$$

It is possible to choose $\mathbb{P}_{\xi}(F)$ so that it becomes a **Markov kernel**, i.e. $\mathbb{P}_{\xi}(F)$ is measurable for all $F \in \mathcal{N}$ and $\mathbb{P}_{\xi}(F)$ is a probability measure on $\mathcal{N}$ for all $\xi \in \mathcal{X}$. Then the measure $\mathbb{P}_{\xi}$ is called the **Palm distribution at the point** ξ.

As mentioned, we can interpret the Palm distribution heuristically as a conditional probability. For a point process X the probability that $X \in F$ conditioned on that there is an event in $b(\xi,\varepsilon)$ can be approximated by $\mathbb{P}_{\xi}(F)$ as follows

$$\begin{aligned}
\mathbb{P}(X \in F \mid X(b(\xi,\varepsilon)) > 0) &= \frac{\mathbb{P}(X \in F, X(b(\xi,\varepsilon)) > 0)}{\mathbb{P}(X(b(\xi,\varepsilon)) > 0)} \\
&\approx \frac{\mathbb{E}(X(b(\xi,\varepsilon))\mathbb{1}[X \in F])}{\mathbb{E}X(b(\xi,\varepsilon))} \\
&\approx \frac{\mathrm{d}C(\cdot \times F)}{\mathrm{d}M}(\xi) \\
&= \mathbb{P}_{\xi}(F),
\end{aligned}$$

where $\varepsilon > 0$ is arbitrary small.

Assuming again that M is σ-finite, then there exists a Markov kernel $\mathbb{P}^!_\xi(F)$ such that

$$C^!(A \times F) = \int_A \mathbb{P}^!_\xi(F) \mathrm{d}M(\xi),$$

where $\mathbb{P}^!_\xi$ is called the **reduced Palm distribution**. This distribution can be interpreted as the conditional distribution of X on the reduced set $\mathcal{X}\backslash\xi$ given that there is an event at ξ, as shown for the Palm distribution.

The Campbell-Mecke Theorem

Next we introduce the Campbell-Mecke theorem which is an important result we will use later. The theorem was first given in a less general form by Campbell (1909) and has been extended by Mecke (1967) to its present form.

Theorem 4.1 (Campbell-Mecke Theorem)
Let $h : \mathcal{X} \times N \to [0, \infty)$ be a measurable function, and let X be a point process on $\mathcal{X}$. Then

$$\begin{aligned} \mathbb{E} \sum_{\xi \in X} h(\xi, X) &= \int_{\mathcal{X}} \int_N h(\xi, x) \mathrm{d}C(\xi, x) \\ &= \int_{\mathcal{X}} \int_N h(\xi, x) \mathrm{d}\mathbb{P}_\xi(x) \mathrm{d}M(\xi). \end{aligned} \tag{4.4}$$

Further,

$$\begin{aligned} \mathbb{E} \sum_{\xi \in X} h(\xi, X\backslash\xi) &= \int_{\mathcal{X}} \int_N h(\xi, x) \mathrm{d}C^!(\xi, x) \\ &= \int_{\mathcal{X}} \int_N h(\xi, x) \mathrm{d}\mathbb{P}^!_\xi(x) \mathrm{d}M(\xi). \end{aligned}$$

Proof:
By standard measure theoretical arguments it suffices to verify the result for $h(\xi, X) = \mathbb{1}[\xi \in A, X \in F]$, where $A \in \mathcal{B}_0$ and $F \in \mathcal{N}$. Thus by replacing h with the indicator function it follows immediately that

$$\begin{aligned} \mathbb{E} \sum_{\xi \in X} \mathbb{1}[\xi \in A, X \in F] &= C(A \times F) \\ &= \int_{\mathcal{X}} \int_N \mathbb{1}[\xi \in A, x \in F] \mathrm{d}\mathbb{P}_\xi(x) \mathrm{d}M(\xi). \end{aligned}$$

The second statement follows by similar arguments. ∘

Combining the Campbell-Mecke Theorem with the Slivnyak-Mecke formula (Theorem 3.9) we get the following theorem.

Theorem 4.2
Let X be a Poisson point process with intensity measure μ, and let $h : \mathcal{X} \times N \to [0, \infty)$ be measurable. Then $\mathbb{P} = \mathbb{P}^!_\xi$ for μ almost all ξ.

Proof:
Given a Poisson point process X we have, first by Theorem 4.1 and then by Theorem 3.9 that

$$\begin{aligned}
\mathbb{E} \sum_{\xi \in X} h(\xi, X \backslash \xi) &= \int_{\mathcal{X}} \int_{N} h(\xi, x) \mathrm{d}\mathbb{P}^!_\xi(x) \mathrm{d}M(\xi) \\
&= \int_{\mathcal{X}} \mathbb{E} h(\xi, X) \mathrm{d}\mu(\xi) \qquad (4.5) \\
&= \int_{\mathcal{X}} \int_{N} h(\xi, x) \mathrm{d}\mathbb{P}(x) \mathrm{d}\mu(\xi).
\end{aligned}$$

Since $\mu = M$ and (4.5) holds for all h, we conclude, that for μ almost all ξ, $\mathbb{P} = \mathbb{P}^!_\xi$. ∘

In fact, the property $\mathbb{P} = \mathbb{P}^!_\xi$ gives a unique characterization of the Poisson point process, see Jagers (1973).

The Case $\mathcal{X} = \mathbb{R}^n$

Assume now, that $\mathcal{X} = \mathbb{R}^n$. If $\alpha^{(n)}$ has a density $\lambda^{(n)}$ w.r.t. to the Lebesgue measure then $\lambda^{(n)}$ is called the **nth-order product density**, and is given by

$$\alpha^{(n)}(A_1 \times \cdots \times A_n) = \int_{A_1} \cdots \int_{A_n} \lambda^{(n)}(\xi_1, \ldots, \xi_n) \mathrm{d}\xi_1 \cdots \mathrm{d}\xi_n. \qquad (4.6)$$

Particularly, $\lambda^{(1)}(\cdot) = \lambda(\cdot)$ is called the **intensity function** again extending the definition given in Chapter 3 on page 15 and finally if λ is constant except on a Lebesgue null-set it is called the **intensity**. We notice, that for a stationary point process X this will be the case, since

$$\begin{aligned}
\alpha(A) &= \mathbb{E}X(A) = \mathbb{E}X(A + \eta) \qquad (4.7) \\
&= \int_A \lambda(\xi) \mathrm{d}\xi = \int_A \lambda(\xi + \eta) \mathrm{d}\xi
\end{aligned}$$

for all $\eta \in \mathcal{X}$ and $A \in \mathcal{B}_0$. Similar arguments give that if X is stationary then $\lambda^{(2)}(\xi, \eta)$ only depends on $\xi - \eta$ except on a Lebesgue null-set and if X is also isotropic then $\lambda^{(2)}(\xi, \eta)$ only depends on $\|\xi - \eta\|$ except on a Lebesgue null-set. Without loss of generality we redefine the density on the null-set in both cases, so that it has the mentioned property everywhere.

Applying Theorem 3.9 to (4.1) and (4.6), we see that for a Poisson point process with intensity function λ

$$\lambda^{(n)}(\xi_1, \dots, \xi_n) = \prod_{i=1}^{n} \lambda(\xi_i). \tag{4.8}$$

Now let X be a point process defined on $\mathcal{X}$ with intensity function λ. Then the first-order moment measure is given as

$$M(A) = \int_A \lambda(\xi) \mathrm{d}\xi,$$

and if X is stationary, then

$$M(A) = \lambda |A|.$$

The **Palm distribution at the typical point** denoted by $\mathbb{P}_0$, is given by $\mathbb{P}_0(F) = \mathbb{P}_\xi(F + \xi)$ for all $F \in \mathcal{N}$ and $\xi \in \mathcal{X}$, where

$$F + \xi = \{x + \xi = \{x_1 + \xi, x_2 + \xi, \dots\} \in N : x = \{x_1, x_2, \dots\} \in F\}.$$

Further, we call $\mathbb{P}_0^!$ the **reduced Palm distribution at the typical point**, and obtain similarly that $\mathbb{P}_0^!(F) = \mathbb{P}_\xi^!(F + \xi)$. The Palm distribution at ξ cannot be stationary whereas $\mathbb{P}_\xi^!$ can be stationary, see Stoyan et al. (1995, Example 4.3, p. 123).

Having defined the Palm distribution at the typical point, we now give a corollary to Theorem 4.1 which will be useful henceforth.

Corollary 4.3 (The Campbell-Mecke Theorem in $\mathbb{R}^n$)
Let $h : \mathcal{X} \times N \to [0, \infty)$ be a measurable function, and let X be a point process on $\mathcal{X} = \mathbb{R}^n$. If X is stationary with intensity λ then

$$\begin{aligned} \mathbb{E} \sum_{\xi \in X} h(\xi, X) &= \lambda \int_{\mathcal{X}} \int_N h(\xi, x) \mathrm{d}\mathbb{P}_\xi(x) \mathrm{d}\xi \\ &= \lambda \int_{\mathcal{X}} \int_N h(\xi, x + \xi) \mathrm{d}\mathbb{P}_0(x) \mathrm{d}\xi. \end{aligned} \tag{4.9}$$

Similarly for the reduced Palm distribution,

$$\begin{aligned}\mathbb{E}\sum_{\xi\in X} h(\xi, X\backslash\xi) &= \lambda\int_{\mathcal{X}}\int_{N} h(\xi,x)\mathrm{d}\mathbb{P}^!_\xi(x)\mathrm{d}\xi \\ &= \lambda\int_{\mathcal{X}}\int_{N} h(\xi,x+\xi)\mathrm{d}\mathbb{P}^!_0(x)\mathrm{d}\xi.\end{aligned}$$

Proof:
Let X be a stationary point process on $\mathcal{X}$ with intensity λ and review (4.4). Since

$$\begin{aligned}\int_{\mathcal{X}} \mathbb{1}[\xi\in A]\mathrm{d}M(\xi) &= \lambda|A| \\ &= \lambda\int_{\mathcal{X}} \mathbb{1}[\xi\in A]\mathrm{d}\xi,\end{aligned}$$

standard measure theoretical arguments and (4.4) gives the first equality in the corollary.

Using the fact that $\mathbb{P}_\xi(F) = \mathbb{P}_0(F-\xi)$ yields

$$\begin{aligned}\int_{N} \mathbb{1}[x\in F]\mathrm{d}\mathbb{P}_\xi(x) &= \int_{N} \mathbb{1}[x\in F-\xi]\mathrm{d}\mathbb{P}_0(x) \\ &= \int_{N} \mathbb{1}[x+\xi\in F]\mathrm{d}\mathbb{P}_0(x),\end{aligned}$$

which, again by standard measure theoretical arguments, concludes the proof. ∘

Now we can obtain an expression for $\mathbb{P}_\xi$ which is more constructive than (4.3). Using (4.9) we see that

$$\begin{aligned}\mathbb{E}\sum_{\eta\in X} \mathbb{1}[\eta\in A, X-\eta\in F] &= \lambda\int_{\mathcal{X}}\int_{N} \mathbb{1}[\eta\in A, x\in F]\mathrm{d}\mathbb{P}_0(x)\mathrm{d}\xi \\ &= \lambda|A|\int_{N} \mathbb{1}[x\in F]\mathrm{d}\mathbb{P}_0(x) \\ &= \lambda|A|\mathbb{P}_0(F) \\ &= \lambda|A|\mathbb{P}_\xi(F+\xi)\end{aligned}$$

and hence

$$\mathbb{P}_\xi(F) = \frac{\mathbb{E}\sum_{\eta\in X}\mathbb{1}[\eta\in A, X-\eta+\xi\in F]}{\lambda|A|}.$$

The following proposition states that if X is isotropic, then M, C, and the (reduced) Palm distribution are isotropic.

Proposition 4.4
If a point process X on $\mathcal{X} = \mathbb{R}^n$ is isotropic, then $M, C, \mathbb{P}_\xi$, and $\mathbb{P}^!_\xi$ are isotropic.

Proof:
Let r be any rotation around the origin. Since X is isotropic we have that $\mathbb{P}(X \in F) = \mathbb{P}(r(X) \in F)$. Hence

$$
\begin{aligned}
M(A) &= \mathbb{E}X(A) = \int_N x(A)\mathrm{d}\mathbb{P}(x) \\
&= \int_N r(x)(A)\mathrm{d}\mathbb{P}(x) = \mathbb{E}(r(X)(A)) \\
&= M_r(A) \\
C(A \times F) &= \mathbb{E}(X(A)\mathbb{1}[X \in F]) = \int_N x(A)\mathbb{1}[x \in F]\mathrm{d}\mathbb{P}(x) \\
&= \int_N r(x)(A)\mathbb{1}[r(x) \in F]\mathrm{d}\mathbb{P}(x) = \mathbb{E}(r(X)(A)\mathbb{1}[r(X) \in F]) \\
&= C_r(A \times F),
\end{aligned}
$$

where M_r and C_r are the measures for the rotation of X. Thus

$$
\begin{aligned}
\mathbb{P}_\xi(X \in F) &= \frac{\mathrm{d}C(\cdot \times F)}{\mathrm{d}M}(\xi) = \frac{\mathrm{d}C_r(\cdot \times F)}{\mathrm{d}M_r}(\xi) \\
&= \mathbb{P}_\xi(r(X) \in F)
\end{aligned}
$$

and hence $\mathbb{P}_\xi$ is isotropic. It follows directly by similar arguments that $\mathbb{P}^!_\xi$ is isotropic. ∘

4.2 Second Order Summary Statistics in $\mathbb{R}^n$

Second-order properties of spatial point processes describe how the interaction or spatial dependence amongst pairs of events varies through a particular space. These properties can be described by the second-order intensity of the spatial point pattern, by Ripley's K-function, or by the pair-correlation function g. The two latter functions will be described below.

4.2.1 Ripley's K-function

Ripley (1976) introduced the **K-function** for a stationary point process X as the mean number of further events within distance r from the typical point, divided by the intensity λ. Since X is stationary it suffices to look at arbitrary events within a bounded Borel set B. Formally this becomes

$$\lambda K(r) \quad = \quad \mathbb{E} \sum_{\xi \in B \cap X} \frac{X(b(\xi, r) \backslash \xi)}{\lambda |B|} \tag{4.10}$$

for $r \in \mathbb{R}_+$ and $B \in \mathcal{B}_0$ with $0 < |B|$. Using Corollary 4.3 we see that

$$\begin{aligned} \mathbb{E} \sum_{\xi \in B \cap X} X(b(\xi, r) \backslash \xi) &= \lambda \int_B \int_N x(b(\xi, r)) \mathrm{d}\mathbb{P}^!_\xi \mathrm{d}\xi \\ &= \lambda \int_B \mathbb{E}^!_\xi X(b(\xi, r)) \mathrm{d}\xi \\ &= \lambda |B| \mathbb{E}^!_0 X(b(0, r)), \end{aligned}$$

so that

$$\lambda K(r) \quad = \quad \mathbb{E}^!_0 X(b(0, r)),$$

which indeed is independent of the choice of B. Here $\mathbb{E}^!_0$ denotes the expectation w.r.t. $\mathbb{P}^!_0$.

Considering the case where X is not stationary, this definition is generalized by Baddeley, Møller & Waagepetersen (1999) in the following way

$$K_{inhom}(r) \quad = \quad \mathbb{E} \left[\sum_{\xi, \eta \in X}^{\neq} \frac{\mathbb{1}[\xi \in B, \|\eta - \xi\| \le r]}{\lambda(\xi) \lambda(\eta) |B|} \right],$$

if this expression is independent of B. They also investigate when K_{inhom} exists.

For practical purposes it is often more convenient to consider a rescaled version of the K-function called the **L-function** which is given by

$$L(r) = \left(\frac{K(r)}{|S^{n-1}|} \right)^{1/n},$$

where S^{n-1} is the unit sphere in $\mathbb{R}^n$. The L-function is particularly useful for diagnostic plotting purposes when $n = 2$, in which case $L(r) = \sqrt{K(r)/\pi}$. Further, the L-function is **variance-stabilizing** at least in the case of the Poisson point process, i.e. the variance is approximately constant for all r when considering the L-function, as a contrast to the K-function, where the variance increases as the radius increases (Besag 1977), see for instance Figure 4.4 ($K(r)$ and $L(r)$) in Section 4.8.

Example 4.5
For a homogeneous Poisson point process in $\mathbb{R}^n$ with rate λ,

$$\begin{aligned} K(r) &= \lambda^{-1}\mathbb{E}_0^! X(b(0,r)) \\ &= \lambda^{-1}\mathbb{E} X(b(0,r)) \\ &= |b(0,r)| \\ &= |S^{n-1}| r^n, \end{aligned}$$

since $\mathbb{P} = \mathbb{P}_0^!$. Hence, for $\mathbb{R}^2$, $K(r) = \pi r^2$, $L(r) = r$, and a plot of $L(r)$ versus r thus yields a straight line with slope 1. Otherwise, a plot of $L(r)$ versus r yielding a concave curve indicates a clustered point pattern, whereas a convex curve indicates a regular point pattern. See Section 4.8 for examples of this.

Finally, we stress the fact that the K-function does not uniquely characterize a point process. Baddeley & Silverman (1984) gives an example of a point process which is not Poisson but has the same K-function as the Poisson point process.

4.2.2 The g-function

Roughly speaking, the **pair-correlation function** g measures how often event pairs a certain distance apart occurs. The definition is based on the first– and second-order intensity functions defined in Section 4.1. Taking $a/0 = 0$ and $X = (x_1, x_2, \dots)$, the pair-correlation function can be expressed as

$$g(x_i, x_j) = \frac{\lambda^{(2)}(x_i, x_j)}{\lambda(x_i)\lambda(x_j)}, \quad \text{for } i, j = 1, 2, \dots \text{ and } i \neq j,$$

assuming that $\lambda(\cdot)$ and $\lambda^{(2)}(\cdot,\cdot)$ exists. Suppose that X is stationary and $\lambda \in \mathbb{R}_+$. If the second order intensity function $\lambda^{(2)}$ exists then by the arguments just below (4.7) we can choose $g(x_i, x_j) = g(x_i - x_j)$, and if we additionally assume isotropy of X, then $g(x_i, x_j) = g(\|x_i - x_j\|)$. In the latter case with $\|x_i - x_j\| = r$, large values of $g(r)$ indicate that event pairs of distance r appear frequently, while small values of $g(r)$ indicate that event pairs of distance r are rare. No event pairs of distance r occur if $g(r) = 0$. Finally, notice by (4.8) that for a homogeneous Poisson point process $g \equiv 1$.

If g exists then there is a one-to-one correspondence between K and g. Assuming stationarity, isotropy and $n = 2$ for simplicity, then the one-to-one correspondence is derived below. Using (4.2) and (4.6) we see that

$$\begin{aligned}
K(t) &= \frac{1}{\lambda^2|B|}\mathbb{E}\sum_{\xi,\eta\in X}^{\neq} \mathbb{1}[\xi \in B, \|\eta-\xi\| \leq t] \\
&= \frac{1}{\lambda^2|B|}\int_{\mathcal{X}}\int_{\mathcal{X}} \mathbb{1}[\xi \in B, \|\eta-\xi\| \leq t]\,\mathrm{d}\alpha^{(2)}(\xi,\eta) \\
&= \frac{1}{\lambda^2|B|}\int_{B}\int_{B_{\oplus t}} \mathbb{1}[\|\eta-\xi\| \leq t]\,\mathrm{d}\alpha^{(2)}(\xi,\eta) \\
&= \frac{1}{\lambda^2|B|}\int_{B}\int_{B_{\oplus t}} \mathbb{1}[\|\eta-\xi\| \leq t]\,\lambda^{(2)}(\xi,\eta)\,\mathrm{d}\eta\mathrm{d}\xi \\
&= \frac{1}{|B|}\int_{B}\int_{B_{\oplus t}} \mathbb{1}[\|\eta-\xi\| \leq t]\,g(\xi,\eta)\,\mathrm{d}\eta\mathrm{d}\xi,
\end{aligned}$$

where $B_{\oplus t} = \{\eta \in \mathcal{X} \mid \exists\ \xi \in B : \eta \in b(\xi,t)\}$. Since the inner integral only depends on ξ and η through $\|\eta-\xi\|$ we get that

$$\begin{aligned}
K(t) &= \frac{1}{|B|}\int_{B}\int_{b(0,t)} \mathbb{1}[\|\kappa\| \leq t]\,g(\|\kappa\|)\,\mathrm{d}\kappa\mathrm{d}\xi \\
&= \int_0^{2\pi}\int_0^t rg(r)\mathrm{d}r\mathrm{d}\theta \qquad (4.11) \\
&= 2\pi\int_0^t rg(r)\mathrm{d}r,
\end{aligned}$$

hence making it possible to calculate the K-function by knowledge of the pair-correlation function and *vice versa*.

4.3 Estimation

In this section, estimation of the intensity function, Ripley's K-function, and the g-function is described separately for homogeneous and inhomogeneous point processes.

4.3.1 Estimating the Intensity Function

If we are to estimate either the K- or the g-function we need an estimate of the intensity function λ.

Homogeneous Case

A natural estimator of λ (also called the **basic estimator**) is the number of events $X(W)$ observed in a bounded region W divided by the area of that region, i.e.

$$\hat{\lambda} = X(W)/|W|.$$

For a homogeneous Poisson point process, this is the maximum likelihood estimator of λ (see Chapter 7).

Inhomogeneous Case

For an inhomogeneous point process it may be more appropriate to estimate the intensity locally over the observation window W instead of using the basic estimate from the homogeneous case.

One method of doing so is called **binning**. A partition of W is made with elements called bins and the basic estimator is found for each bin. If we assume λ to be constant on each bin, we are done. More often it may be appropriate to apply some smoothing algorithm to smoothen out the estimate of λ from the basic estimates on the bins.

Another method, treated in Silverman (1986), is called **kernel estimation**. To explain this concept we give an example where the kernel is Gaussian and the point process is defined on $\mathbb{R}$. However, the principle easily generalizes to other kernels and spaces. Assume that X is a point process on $\mathcal{X} = \mathbb{R}$. Having observed X only on $W \subset \mathcal{X}$ we wish to estimate the intensity function λ on W. For all $s \in W$ we may try with

$$\hat{\lambda}_b(s) = \sum_{\xi \in X \cap W} k_b(s - \xi),$$

where $k_b(z)$ is the density of a normal distribution with mean 0 and variance b at the point z. The left hand part of Figure 4.1 illustrates this.

Generally, k_b is the kernel with **bandwidth** b and the choice of bandwidth has a huge impact on the estimate. A small bandwidth (variance in this example) will concentrate the intensity function narrowly around the observed events while a large bandwidth will make the intensity function close to constant.

A point worth noticing is that the estimate given above is biased downwards because contributions from unobserved events in X but outside W are missing. The right hand part of Figure 4.1 clearly illustrates this. This error is corrected by estimating λ by

$$\hat{\lambda}_b(s) = \sum_{\xi \in X \cap W} k_b(s - \xi)/c_W(\xi), \qquad s \in W,$$

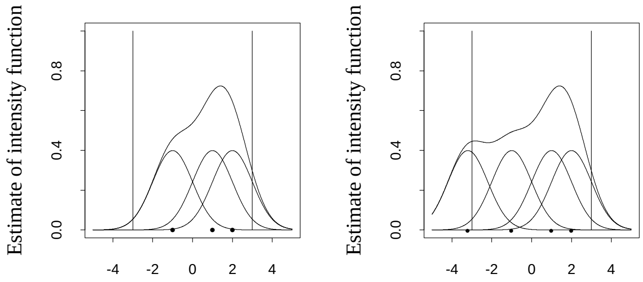

Figure 4.1: Use of a Gaussian kernel to estimate the intensity function.

where $1/c_W(\xi)$ is an **edge correcting factor** given by

$$c_W(\xi) = \int_W k_b(s-\xi)\mathrm{d}s.$$

With this correction we clearly have that

$$\int_W \hat{\lambda}_b(s)\mathrm{d}s = x(W),$$

thus making our estimate agree with the observed data.

However, Baddeley et al. (1999) argue that using $\hat{\lambda}_b$ in the estimate of K_{inhom} (see Section 4.3.2) makes it severely biased downwards, and suggest the modified kernel estimator

$$\tilde{\lambda}_b(s) = \sum_{\xi \in X \cap W \setminus s} k_b(s-\xi)/c_W, \qquad s \in W,$$

where c_W is as above. Note that $\hat{\lambda}_b(s) = \tilde{\lambda}_b(s)$ for $s \in W \setminus X$.

4.3.2 Estimating the K-function

We again assume that a point process X on $\mathcal{X}$ has been observed only on $W \subset \mathcal{X}$, where $W \in \mathcal{B}_0$.

Homogeneous Case

If we try estimating K by letting $B = W$ in

$$\lambda\hat{K}(r) = \sum_{\xi \in B \cap X} \frac{X(b(\xi, r)\backslash\xi)}{\lambda|B|}$$

we will need information about X outside W which is not available. Instead we may use **minus sampling** where we let $B = W_{\ominus r} = \{w \in W : b(w, r) \subseteq W\}$ which gives an unbiased estimate if $|W_{\ominus r}| > 0$ and λ is known. However, λ is not normally known, but Ripley (1981) argues that the use of $\hat{\lambda}$ instead of λ does not affect the unbiasedness of $\hat{K}$ seriously.

The use of minus sampling does have a serious drawback. As r increases more of the observed data is not utilized for estimation, so the method is not suitable unless r is small compared to the size of W. For a better method we again have to look at edge effects. An obvious way to do this is by observing that

$$X(b(\xi, r)) \approx X(W \cap b(\xi, r))\frac{|b(\xi, r)|}{|W \cap b(\xi, r)|}$$

leading to the estimate of λK given by

$$\widehat{\lambda K}(r) = \sum_{\xi \in W \cap X} \frac{X(W \cap b(\xi, r)\backslash\xi)}{X(W)} \frac{|b(\xi, r)|}{|W \cap b(\xi, r)|}.$$

However, according to Stoyan & Stoyan (1994) this estimator is biased, so instead they propose a less intuitive edge correction which is unbiased under certain restrictions.

In the following theorem let $W_\xi = \{s + \xi : s \in W\}$ for any $\xi \in \mathcal{X}$.

Theorem 4.6
Consider a stationary point process X on $\mathcal{X} = \mathbb{R}^n$ with intensity λ. Then

$$\widehat{\lambda^2 K}(r) = \sum_{\xi, \eta \in X \cap W}^{\neq} \frac{\mathbb{1}[\|\eta - \xi\| \leq r]}{|W_\xi \cap W_\eta|}$$

is an unbiased estimator of $\lambda^2 K(r)$ if $|W_\xi \cap W_\eta| > 0$ almost surely for any pair $\xi, \eta \in X \cap W$.

Proof:

Consider (4.10) which may be rewritten as

$$\begin{aligned}
\lambda^2 K(r) &= \mathbb{E}\left[\sum_{\xi,\eta\in X}^{\neq} \frac{\mathbb{1}[\xi\in B, \|\eta-\xi\|\leq r]}{|B|}\right] \\
&= \lambda \int_{\mathcal{X}}\int_{N} \sum_{\eta\in x+\xi} \frac{\mathbb{1}[\xi\in B, \|\eta-\xi\|\leq r]}{|B|} \mathrm{d}\mathbb{P}_0^!(x)\mathrm{d}\xi \\
&= \lambda \int_{N} \sum_{\kappa\in x} \int_{\mathcal{X}} \frac{\mathbb{1}[\xi\in B, \|\kappa\|\leq r]}{|B|} \mathrm{d}\xi\mathrm{d}\mathbb{P}_0^!(x) \\
&= \lambda \int_{N} \sum_{\kappa\in x} \mathbb{1}[\|\kappa\|\leq r]\mathrm{d}\mathbb{P}_0^!(x) \\
&= \lambda \int_{N} \sum_{\kappa\in x} \int_{\mathcal{X}} \frac{\mathbb{1}[\xi\in W\cap W_{-\kappa}, \|\kappa\|\leq r]}{|W\cap W_{-\kappa}|} \mathrm{d}\xi\mathrm{d}\mathbb{P}_0^!(x) \\
&= \lambda \int_{N} \sum_{\kappa\in x} \int_{\mathcal{X}} \frac{\mathbb{1}[\xi\in W, \xi+\kappa\in W, \|\kappa\|\leq r]}{|W\cap W_{-\kappa}|} \mathrm{d}\xi\mathrm{d}\mathbb{P}_0^!(x) \\
&= \lambda \int_{\mathcal{X}}\int_{N} \sum_{\eta\in x+\xi} \frac{\mathbb{1}[\xi\in W, \eta\in W, \|\eta-\xi\|\leq r]}{|W_\xi\cap W_\eta|} \mathrm{d}\mathbb{P}_0^!(x)\mathrm{d}\xi \\
&= \mathbb{E}\left[\sum_{\xi,\eta\in X\cap W}^{\neq} \frac{\mathbb{1}[\|\eta-\xi\|\leq r]}{|W_\xi\cap W_\eta|}\right],
\end{aligned}$$

provided that $|W_\xi\cap W_\eta|>0$ for $\{\xi,\eta\}\subseteq X\cap W$, $\xi\neq\eta$, almost surely. ∘

The factor $|W|/|W_\xi\cap W_\eta|$ is the edge correcting factor. Note that this method incorporates all known points and thus no information is lost as opposed to minus sampling.

Another edge correction factor was given by Ripley (1976). He considered the stationary and isotropic case, and used the heuristic approximation,

$$X(b(\xi,r+\varepsilon)\backslash b(\xi,r)) \approx X((b(\xi,r+\varepsilon)\backslash b(\xi,r))\cap W)\frac{|\partial b(\xi,r)|}{|\partial b(\xi,r)\cap W|},$$

for ε small where $w_{\xi,r}=\frac{|\partial b(\xi,r)|}{|\partial b(\xi,r)\cap W|}$ is called **Ripley's edge correcting factor**.

Next we show that $w_{\xi,r}$ gives an unbiased estimate for $\lambda^2K(r)$ in $\mathbb{R}^2$ if r satisfies a natural restriction and X is isotropic. To do this we need the following lemma.

Lemma 4.7
Let X be a stationary and isotropic point process in $\mathbb{R}^2$. Then

$$\int_{\mathcal{X}}\int_{\mathcal{X}} h(\xi,\eta)\mathrm{d}\alpha^{(2)}(\xi,\eta) = \frac{\lambda^2}{2\pi}\int_{\mathbb{R}_+}\int_{\mathcal{X}}\int_0^{2\pi} h(\xi,\xi+r(\cos\theta,\sin\theta))\,\mathrm{d}\theta\,\mathrm{d}\xi\,\mathrm{d}K(r).$$

Proof:
By standard measure theoretical arguments and Lemma 2.12 it suffices to consider the case where $h(\xi,\eta) = \mathbb{1}[\xi \in A, \eta-\xi \in B]$ for $A \in \mathcal{B}_0$ and $B = \{\kappa \in \mathbb{R}^2 : \|\kappa\| \in I,\ \angle\kappa \in J\}$, where $I = (s,t]$ and $J \subset [0,2\pi)$ are bounded intervals. Here $\angle\kappa$ denotes the counterclockwise angle between the abscissa and κ. By (4.2) and Corollary 4.3 we have that

$$\begin{aligned}
\int_{\mathcal{X}}\int_{\mathcal{X}} h(\xi,\eta)\mathrm{d}\alpha^{(2)}(\xi,\eta) &= \mathbb{E}\sum_{\xi\in X}\sum_{\eta\in X\setminus\xi} h(\xi,\eta)\\
&= \mathbb{E}\sum_{\xi\in X}\mathbb{1}[\xi\in A]\sum_{\eta\in X\setminus\xi}\mathbb{1}[\eta-\xi\in B]\\
&= \lambda\int_{\mathcal{X}}\int_{N}\mathbb{1}[\xi\in A]\sum_{\eta\in x+\xi}\mathbb{1}[\eta-\xi\in B]\mathrm{d}\mathbb{P}_0^!(x)\,\mathrm{d}\xi\\
&= \lambda\int_{\mathcal{X}}\int_{N}\mathbb{1}[\xi\in A]\sum_{\kappa\in x}\mathbb{1}[\|\kappa\|\in I,\angle\kappa\in J]\mathrm{d}\mathbb{P}_0^!(x)\,\mathrm{d}\xi.
\end{aligned}$$

Since X is isotropic, we have by Proposition 4.4 that $\mathbb{P}_0^!$ is isotropic and thus

$$\int_N\sum_{\kappa\in x}\mathbb{1}[\|\kappa\|\in I,\angle\kappa\in J]\mathrm{d}\mathbb{P}_0^!(x) = \frac{|J|}{2\pi}\int_N\sum_{\kappa\in x}\mathbb{1}[\|\kappa\|\in I]\mathrm{d}\mathbb{P}_0^!(x).$$

Using this and Corollary 4.3 we obtain that

$$\begin{aligned}
\int_{\mathcal{X}}\int_{\mathcal{X}} h(\xi,\eta)\mathrm{d}\alpha^{(2)}(\xi,\eta) &= \frac{|J|}{2\pi}\lambda\int_{\mathcal{X}}\int_{N}\mathbb{1}[\xi\in A]\sum_{\kappa\in x}\mathbb{1}[\|\kappa\|\in I]\mathrm{d}\mathbb{P}_0^!(x)\,\mathrm{d}\xi\\
&= \frac{|J|}{2\pi}\mathbb{E}\sum_{\xi\in X}\sum_{\eta\in X\setminus\xi}\mathbb{1}[\xi\in A,\|\xi-\eta\|\in(s,t]]\\
&= \frac{|J|}{2\pi}\lambda^2|A|(K(t)-K(s))\\
&= \frac{\lambda^2}{2\pi}\int_{\mathbb{R}_+}\int_{\mathcal{X}}\int_0^{2\pi}\mathbb{1}[\xi\in A,r\in I,\theta\in J]\,\mathrm{d}\theta\,\mathrm{d}\xi\,\mathrm{d}K(r),
\end{aligned}$$

which concludes the proof. ◦

Assuming isotropy we are now ready to present and show unbiasedness of the estimator for K using Ripley's edge correction factor.

Theorem 4.8
Consider a stationary and isotropic point process X on $\mathbb{R}^2$. Let $W \in \mathcal{B}_0, |W| > 0$ and

$$r^* = \sup \left\{ r > 0 : \int\limits_W \mathbb{1}\left[W \cap \partial b(\xi, r) \neq \emptyset\right] \mathrm{d}\xi > 0 \right\}.$$

Then

$$\widehat{\lambda^2 K}(r) = \frac{1}{|W|} \sum_{\xi,\eta \in X \cap W}^{\neq} \mathbb{1}[\|\xi - \eta\| \leq r] w_{\xi,\|\xi-\eta\|}$$

is an unbiased estimator whenever $r < r^*$.

For example, if $W = [0,1]^2$ then $r^* = \sqrt{2}$.

Proof:
Let $t < r^*$. The edge correcting factor $w_{\xi,r}$ may be written as

$$\begin{aligned} w_{\xi,r} &= \frac{|\partial b(\xi, r)|}{|\partial b(\xi, r) \cap W|} \\ &= \frac{2\pi r}{r \int_0^{2\pi} \mathbb{1}[\xi + r(\cos\theta, \sin\theta) \in W] \mathrm{d}\theta} \end{aligned}$$

Using this, (4.2), and Lemma 4.7, we see that

$$\begin{aligned} &\mathbb{E} \sum_{\xi,\eta \in X}^{\neq} \mathbb{1}[\xi, \eta \in W, \|\xi - \eta\| \leq t] w_{\xi,\|\xi-\eta\|} \\ &= \int\limits_{\mathcal{X}} \int\limits_{\mathcal{X}} \mathbb{1}[\xi, \eta \in W, \|\xi - \eta\| \leq t] w_{\xi,\|\xi-\eta\|} \mathrm{d}\alpha^{(2)}(\xi, \eta) \\ &= \frac{\lambda^2}{2\pi} \int\limits_{\mathbb{R}_+} \int\limits_N \int_0^{2\pi} \mathbb{1}[\xi \in W,\ \xi + r(\cos\theta, \sin\theta) \in W,\ r \leq t] w_{\xi,r} \,\mathrm{d}\theta \,\mathrm{d}\xi \,\mathrm{d}K(r) \\ &= \lambda^2 \int_0^t \int\limits_W \mathrm{d}\xi \,\mathrm{d}K(r) \\ &= \lambda^2 |W| K(t), \end{aligned}$$

which completes the proof. ∘

We thus have two competing estimators for the stationary and isotropic case and it would be natural to investigate which is the most efficient, i.e. which has the smallest variance. It would also be interesting to see how well the estimator that uses Ripley's edge correction factor behaves under deviation from the assumption of isotropy. This is not investigated in this report, but a discussion is found in Stoyan & Stoyan (1994).

Inhomogeneous Case

An unbiased estimator of $K_{inhom}(t)$ is given by

$$\hat{K}_{inhom}(t) = \frac{1}{|W|} \sum_{\xi,\eta \in X \cap W}^{\neq} w_{\xi,\|\xi-\eta\|} \frac{\mathbb{1}[\|\xi-\eta\| \leq t]}{\lambda(\xi)\lambda(\eta)}, \qquad \text{for } t < r^*,$$

where $w_{\xi,\|\xi-\eta\|}$ is Ripley's edge correcting factor. If $\lambda(\cdot)$ is unknown then it is substituted by the estimate $\tilde{\lambda}_b(\cdot)$ as given in Section 4.3.1. For further details see Baddeley et al. (1999).

Estimating the L-function

An obvious estimator for the L-function in both the homogeneous and inhomogeneous case is

$$\hat{L} = \left(\frac{\hat{K}}{|S^{n-1}|}\right)^{1/n},$$

where $\hat{K}$ is one of the estimators given above. For $n = 2$ this reduces to $\hat{L} = \sqrt{\frac{\hat{K}}{\pi}}$.

4.3.3 Estimating the g-function

Homogeneous Case

For the isotropic case we use (4.11) to see that the following kernel estimator is appropriate.

$$\lambda^2 \hat{g}(r) = \sum_{\xi,\eta \in X \cap W}^{\neq} \frac{w_{\eta,\|\xi-\eta\|} k_b(r - \|\xi-\eta\|)}{2\pi r |W|},$$

where k_b is a kernel function, for example the Epanecnikov kernel e_b with bandwidth $b > 0$ given by

$$e_b(s) = \mathbb{1}[|s| \leq b] \frac{3}{4b} \left(1 - \frac{s^2}{b^2}\right).$$

As before we substitute λ with the estimate $\hat{\lambda}$. Instead of the edge correcting factor $w_{\eta,\|\xi-\eta\|}$ we could use $|W|/|W_\xi \cap W_\eta|$ as presented earlier. As usual the choice of bandwidth is far more important than the choice of kernel, and the value of the resulting non-parametric estimate is highly dependent on the chosen bandwidth. Stoyan & Stoyan (1994) discuss how to choose b and settle on the value $b = c\lambda^{-1/2}$, $c \in [0.1, 0.2]$.

Inhomogeneous Case

For simplicity assume that $g(\xi, \eta)$ exists and is both stationary and isotropic. Then an estimate for the pair correlation function g is given by

$$\hat{g}(r) = \sum_{\xi,\eta \in X \cap W}^{\neq} \frac{w_{\xi,\|\xi-\eta\|} k_b(r - \|\xi - \eta\|)}{2\pi r |W| \hat{\lambda}(\xi)\hat{\lambda}(\eta)},$$

where k_b is any kernel.

4.4 Other Summary Statistics in $\mathbb{R}^n$

Other useful summary statistics besides Ripley's K-function and the g-function exist. Three commonly used ones are described briefly below. For further details about the summary statistics treated in this section, the reader should consult Ripley (1981), Diggle (1983), Stoyan & Stoyan (1994), and Stoyan et al. (1995).

4.4.1 The F-function

The empty space function $F(r)$ is the cumulative distribution function for random-point-nearest-event distances, meaning that it denotes the probability that the distance from a random point to the nearest event is less than or equal to r. For a stationary point process X and an arbitrary point $\xi \in \mathbb{R}^n$, the empty space function is given by

$$F(r) = \mathbb{P}(X(b(\xi, r)) > 0).$$

For simplicity consider $\mathbb{R}^2$ and a regular grid on W and let n denote the number of grid points in $W_{\ominus r}$ and d_i the distance from the ith grid point to its nearest event in $X \cap W$. Then we can estimate F by

$$\hat{F}(r) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[d_i \leq r], \qquad \text{for } r > 0.$$

4.4.2 The G-function

Suppose X is stationary. The nearest neighbour distribution function G, cf. Cowie (1967), is the cumulative distribution function for random-event-nearest-event distances, and can be defined by

$$G(r) = \mathbb{P}^!_\xi(X(b(\xi, r)) > 0),$$

where ξ is an arbitrary event in $\mathbb{R}^n$. For simplicity consider $\mathbb{R}^2$ and let n denote the number of observed events in $W_{\ominus r}$ and e_i the distance from the ith such event to its nearest neighbour in $X \cap W_{\ominus r}$. Then we can estimate G by

$$\hat{G}(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[e_i \leq r], \qquad \text{for } r > 0.$$

4.4.3 The J-function

The J-function, based on the F and G functions described above, was introduced by Baddeley & van Lieshout (1995) as

$$J(r) = \frac{1 - G(r)}{1 - F(r)} = \frac{\mathbb{P}^!_0(X(b(0, r)) = 0)}{\mathbb{P}(X(b(0, r)) = 0)}, \qquad \text{for } r \geq 0 \text{ and } F(r) < 1.$$

In the Poisson case $J = 1$, $J < 1$ indicates clustering, and $J > 1$ indicates regularity. An obvious estimator for $J(r)$ is

$$\hat{J}(r) = \frac{1 - \hat{G}(r)}{1 - \hat{F}(r)}, \qquad \text{for } r \geq 0 \text{ and } \hat{F}(r) < 1.$$

Example 4.9
For a homogeneous Poisson process X with intensity λ on $\mathbb{R}^n$ we have that

$$F(r) = 1 - \exp(-\lambda|S^{n-1}|r^n) = G(r),$$

for all $r > 0$, since $\mathbb{P} = \mathbb{P}^!_0$. Hence $J \equiv 1$.

4.5 Summary Statistics for Marked Point Processes

In this section we briefly consider summary statistics for marked point processes. There exists several methods to extend the summary statistics to marked point processes. Cox & Isham (1980) contains a general introduction to this theory, while Diggle (1983) focus on multi-variate point processes, extending the definition of the K and G functions. For example the extension of the K function denoted by K_{ij}, is defined as the number of events of type j that are found within a certain distance of a typical event of type i.

Here we restrict attention to the marked point processes as defined in Example 2.7 (see also Example 3.8). More specifically we let X be a point process on $\mathcal{X} = \mathbb{R}^n \times \mathcal{M}$. Now the summary statistics already described in Sections 4.2 and 4.4 may be used on the locations (L) of the events in X, which is the reason that we here consider how to summarize the inter-mark correlation. This is done through the mark correlation function, which is described below. For further reading, see Stoyan & Stoyan (1994).

First, let $h : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ be a measurable function. For instance, if each mark is an angle then $h(m_1, m_2)$ could be the absolute difference between the angles m_1 and m_2.

As an analogue to $\alpha^{(2)}$, we now define the **second-order marked factorial moment measure** w.r.t. h by

$$\alpha_h^{(2)}(B_{L_1} \times B_{L_2}) \;=\; \mathbb{E}\Big(\sum_{(l_1,m_1),(l_2,m_2)\in X}^{\neq} h(m_1, m_2)\, \mathbb{1}[l_1 \in B_{L_1}, l_2 \in B_{L_2}]\Big), \tag{4.12}$$

where B_{L_1} and B_{L_2} are bounded Borel sets in $\mathbb{R}^n$.

If $\alpha_h^{(2)}$ has a density $\lambda_h^{(2)}$ w.r.t. the Lebesgue measure then it is called the **h-product density** and we can write

$$\alpha_h^{(2)}(B_{L_1}, B_{L_2}) = \int_{B_{L_1}} \int_{B_{L_2}} \lambda_h^{(2)}(l_1, l_2)\mathrm{d}l_2\mathrm{d}l_1. \tag{4.13}$$

Let

$$\gamma_h(l_1, l_2) = \frac{\lambda_h^{(2)}(l_1, l_2)}{\lambda^{(2)}(l_1, l_2)}, \quad \lambda^{(2)}(l_1, l_2) \neq 0$$

where $\lambda^{(2)}$ is the second-order product density of L, assuming that it exist. We now define the **mark-correlation function** w.r.t. h as

$$g_h(l_1, l_2) = \frac{\gamma_h(l_1, l_2)}{\overline{h}},$$

where $\overline{h} = \int_{\mathcal{M}} \int_{\mathcal{M}} h(m_1, m_2)\mathrm{d}\mathfrak{M}\mathrm{d}\mathfrak{M}$, i.e. the mean value of h, and $\mathfrak{M}$ is the distribution of the marks.

Assuming that g_h exists we see that if the marks are independent of the locations then we find by rewriting (4.12) that

$$\begin{aligned}\alpha_h^{(2)}(B_{L_1}, B_{L_2}) &= \overline{h}\,\mathbb{E} \sum_{(l_1,m_1),(l_2,m_2)\in X} \mathbb{1}[l_1 \in B_{L_1}, l_2 \in B_{L_2}] \\ &= \overline{h}\,\alpha^{(2)}(B_{L_1}, B_{L_2}).\end{aligned}$$

Since this holds for all measurable sets B_{L_1}, B_{L_2} we have by (4.13) that $\lambda_h^{(2)} = \overline{h}\,\lambda^{(2)}$ whereby $g_h = 1$. Further, $g_h > 1$ for pairs of locations that result in marks with higher than average values of h, and $g_h < 1$ in the opposite case. For instance, we might have that g_h is proportional to the distance between locations so that $g_h(r) < 1$ for small distances r and $g_h(r) > 1$ for large r.

Assuming that g_h exists and that it is both stationary and isotropic we may estimate it in the same manner we estimated g in Section 4.3.3. Hence,

$$\hat{g}_h(r) = \sum_{(l_1,m_1),(l_2,m_2)\in X\cap W}^{\neq} \frac{h(m_1,m_2)}{\overline{h}} \frac{w_{l_1,\|l_1-l_2\|} k_b(r - \|l_1 - l_2\|)}{2\pi r \lambda(l_1)\lambda(l_2)|W|},$$

where k_b is a kernel function, for example the Epanecnikov kernel. In Section 4.8 we show plots of such estimates.

As a final remark, we note that the choice of h depends entirely on the application at hand. Thus it may be appropriate to use several h functions that each target a different aspect.

4.6 Higher-Order Summary Statistics

It is possible to consider third- and higher-order summary statistics for point processes in $\mathbb{R}^n$. For instance consider the G-function from Section 4.4. A natural extension to third-order could be

$$G_3(r) = \mathbb{P}_\xi^!\big(\exists \eta, \kappa \in X \cap b(\xi, r) : \|\eta - \kappa\| < r\big).$$

For further coverage we refer to Stoyan & Stoyan (1994).

4.7 Model Selection and Validation

When confronted with a point configuration x, the first thing to do is to decide on a suitable model that appears to fit the particular configuration well (see Chapter 5 for a description of different models). Since it is difficult visually to distinguish between point configurations from different types of models, the use of summary statistics may be a guideline. Plotting the summary statistics and comparing these plots to how they typically would be for the different models should help deciding on a suitable type of model. Of course, any knowledge about the 'physical' objects that are represented by the events should not be disregarded either, but rather be used to complement the plots, in order to choose an appropriate model.

After deciding on a type of model and estimating any involved parameter in the model, see Chapter 7, it is prudent to try to validate the model. One way of doing this is through the use of summary statistics. Make $n-1$ independent samples $x_1, \ldots, x_{n-1}$ from the model, using the techniques described in Chapter 6. Assuming x follows the model, we then have n independent samples. Thus for any summary statistic t at range r we get that

$$\mathbb{P}(t_x(r) > \max_{i=1,\ldots,n-1} t_i(r)) = \mathbb{P}(t_x(r) < \min_{i=1,\ldots,n-1} t_i(r)) = \frac{1}{n}.$$

We now call

$$t^*(r) = \max_{i=1,\ldots,n-1} t_i(r)$$

the **upper envelope** of t and

$$t_*(r) = \min_{i=1,\ldots,n-1} t_i(r)$$

the **lower envelope** of t and make a plot of t_x, t^* and t_* for an r-range.

Making plots for all the described summary statistics results in a kind of model validation. It should be noted that if we let the model selection and the validation be based on the same criteria, this may lead to a greater degree of acceptance. However, it is hoped that using several summary statistics, this is significantly reduced.

In the following section we consider how to use these plots in practice and in Section 4.8 we show some examples.

4.7.1 Summary Statistics in Practice

When calculating summary statistics we need to decide on how many values of r to use. On one hand we want many values to get a fine resolution but on the other hand we do not want the calculations to be too slow. As a compromise we have chosen to use 100 values of r evenly spaced. Further, we will use 99 samples for making the envelopes getting a probability of data being outside the envelopes of 2% for each value of r.

This still leaves two problems

- What range shall be used for the values of r?

- How do we determine when to reject a hypothesis?

Assume that our observation window is the unit square. Then it is possible to consider values of r from 0 to $\sqrt{2}$. However, if we consider values of r near $\sqrt{2}$ we have very little data supporting the analysis and we will be considering interaction on the scale of the entire observation window which does not seem right. Instead, we propose to use the range from 0 to 0.2 because in this range we have more data. However, prior knowledge about the dataset should also be used to settle on the range of value of r in each case.

For the second problem, we would expect approximately 2% of the data being outside the envelopes. If we for the sake of the argument assume that the probability of data being outside the envelopes for any two different values of r is independent and let Z be the number of values of r where data is outside the envelopes, then Z follows a binomial distribution with parameters $n = 100$ and $p = 0.02$. Hence, $\mathbb{P}(Z \leq 4) = 0.95$ and $\mathbb{P}(Z \leq 5) = 0.99$. However, because the assumption of independence is clearly false, we instead propose to look at all five summary statistics ($K/L, g, F, G$, and J) together. Using the same 100 values of r for all five summary statistics we are likely to get that the J function is only defined for around 60 of these values. We thus have 460 different measurements of data that each are outside the envelopes with probability 2%. If we assume independence then we get with probability 95% that there are 14 or fewer of the 460 outside the envelopes. We hence propose a rough rule of accepting the hypothesis if there is less than 3% total outside and no single summary statistic has more than 7% outside, naturally also taking the actual plots into consideration in each case.

4.8 Examples

As previously mentioned, summary statistics can be an important tool for model selection. However, for this to be effective, it is important to know how the summary statistics behave for different kinds of point processes. For the three point pattern shown in Figure 4.2 we have calculated the various summary statistics from Section 4.2 and 4.4, and these are shown in Figures 4.3, 4.4, and 4.5. There is an obviously clustered pattern, a CSR pattern and a very regular pattern. When later faced with a point pattern with less obvious characteristics, see Chapter 8, we will estimate and plot its summary statistics and compare these plots to the plots shown in this section.

To indicate the use of envelopes we have made upper and lower envelopes for each of the three patterns using 99 samples from a homogeneous Poisson point process with $\lambda = 100$. Not surprisingly, the hypothesis of CSR is rejected for two of the patterns, see Figures 4.3, 4.4, and 4.5.

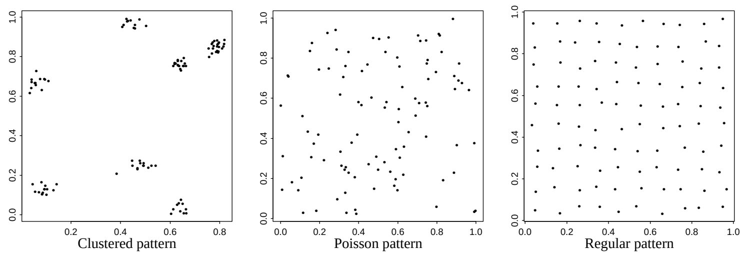

Figure 4.2: Three point patterns. A clustered (left), a Poisson (middle) and a very regular pattern (right).

Finally, we illustrate the use of the marked pair-correlation function in Figure 4.6. Here we have three different marked point patterns, all with the same locations. The marks are angles between 0 and 2π and we choose h to be the angle difference, i.e. $h(m_1, m_2) = \min(|m_1 - m_2|, 2\pi - |m_1 - m_2|)$. We further assume, that $\mathcal{M}$ is the uniform distribution on $[0, 2\pi)$ so $\overline{h} = \frac{\pi}{2}$. The three patterns are made in an ad-hoc manner to get one pattern where the marks are positively correlated for locations close together, one pattern where the marks are independent of the locations, and finally a pattern where the correlation is negative. This is also clearly reflected in the plots of the mark pair-correlation function g_h.

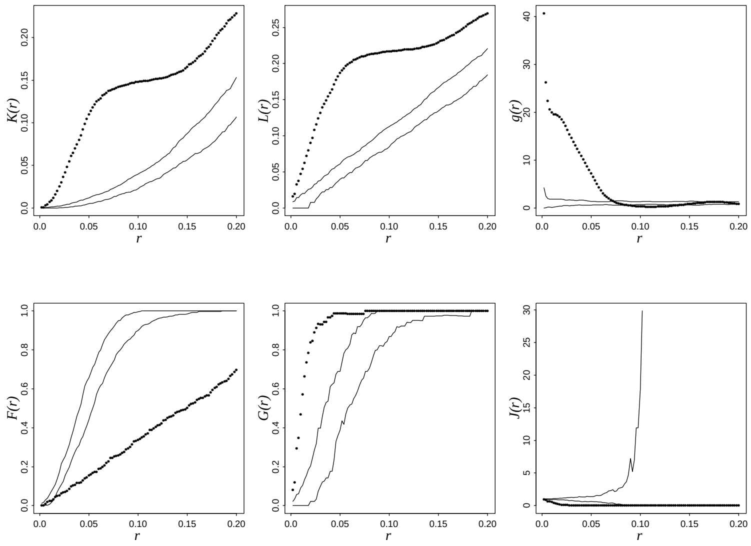

Figure 4.3: Summary statistics for the clustered pattern.

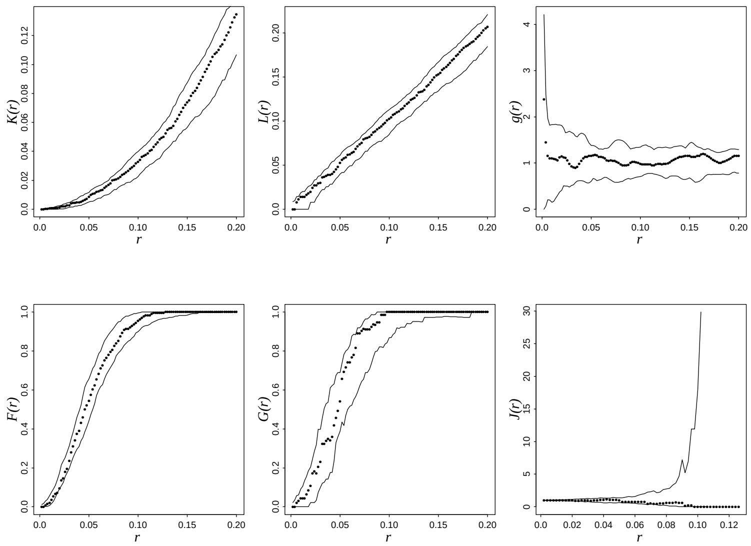

Figure 4.4: Summary statistics for the CSR pattern.

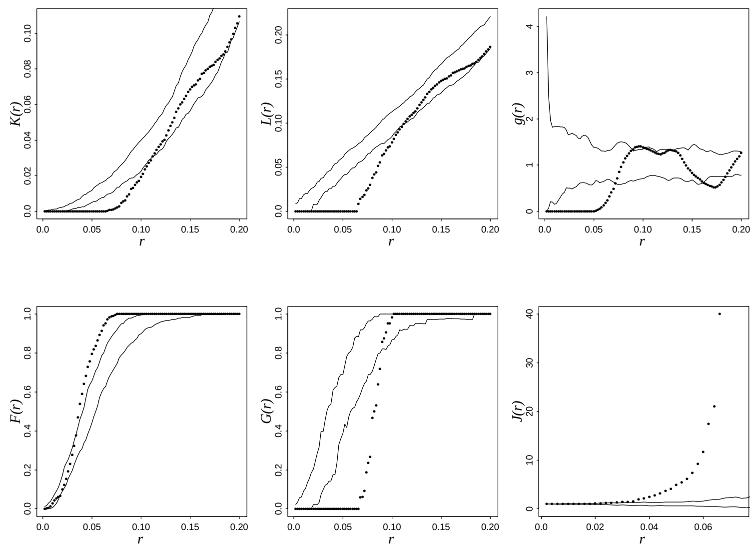

Figure 4.5: Summary statistics for the regular pattern.

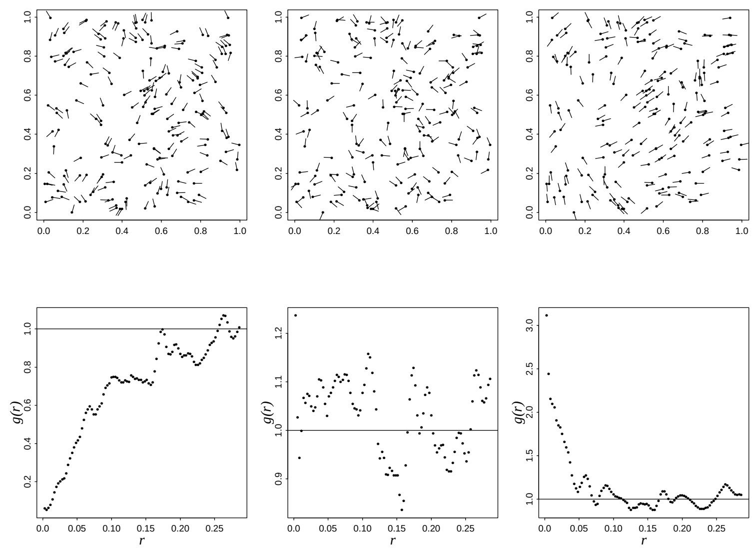

Figure 4.6: Three marked point patterns with same locations. Marks are angles between 0 and 2π and h is the angle difference (above). Estimates of the mark pair-correlation function (below).

Chapter 5

Models

In this chapter we introduce six types of point process models, namely Cox, cluster, hard-core, Markov, nearest neighbour Markov, and Gibbs point process models. Many of these are, as we shall see, closely related to each other and to the Poisson point process which was introduced in Chapter 3. We study some basic properties of these processes, while simulation procedures are described in Chapter 6. Throughout this chapter we only consider point processes defined on $\mathcal{X} = \mathbb{R}^n$, except where specifically noted. However, some of the treated models naturally generalizes to other spaces.

5.1 Cox Processes

In this section we concentrate on Cox processes which were introduced by Cox (1955). Further reading may be found in Grandell (1976) and Kingman (1993). As a special case, we shall present the log Gaussian Cox processes introduced by Møller, Syversveen & Waagepetersen (1998).

A natural way to generalize the Poisson point process is to let the intensity measure be a **random measure**, i.e. a measurable function taking values that are measures.

Definition 5.1 (Cox Processes)
Let Λ be a random diffuse σ-finite measure on $\mathcal{B}$. Suppose that X is a point process so that the conditional distribution of X given $\Lambda = \mu$ is a Poisson point process with intensity measure μ. Then X is called a Cox process with ***driving measure*** *Λ.*

Often the Cox process is referred to as the **doubly stochastic Poisson point**

process since both X and the intensity measure Λ are random.

By definition of the factorial moment (Section 4.1) and conditional expectation we have for $B_1, \ldots, B_n \in \mathcal{B}_0$ that

$$\begin{aligned}
\alpha^{(n)}(B_1 \times \cdots \times B_n) &= \mathbb{E}\left[\mathbb{E} \sum_{\xi_1,\ldots,\xi_n \in X}^{\neq} \mathbb{1}[\xi_1 \in B_1, \ldots, \xi_n \in B_n] \mid \Lambda\right] \\
&= \mathbb{E}\left(\Lambda(B_1)\cdots\Lambda(B_n)\right),
\end{aligned}$$

where the second equality follows since X conditional on $\Lambda = \mu$ is a Poisson point process with intensity measure μ. Using this, we obtain the dispersion of a Cox process,

$$\begin{aligned}
\frac{\mathrm{Var}X(B)}{\mathbb{E}X(B)} &= \frac{M^{(2)}(B \times B) - M(B)^2}{M(B)} \\
&= \frac{\alpha^{(2)}(B \times B) + M(B) - M(B)^2}{M(B)} \\
&= \mathrm{Var}\Lambda(B) + 1 \\
&\geq 1.
\end{aligned}$$

Hence, a Cox process may be appropriate in cases where a Poisson point process is expected but the dispersion is too large to support this.

If Λ has density $\tilde{\lambda}$ w.r.t. the Lebesgue measure we obtain

$$\mathbb{E}\left(\Lambda(B_1)\cdots\Lambda(B_n)\right) = \int_{B_1} \cdots \int_{B_n} \mathbb{E}[\tilde{\lambda}(\xi_1)\cdots\tilde{\lambda}(\xi_n)]\mathrm{d}\xi_1 \cdots \mathrm{d}\xi_n$$

and hence by (4.6)

$$\lambda^{(n)}(\xi_1, \ldots, \xi_n) = \mathbb{E}(\tilde{\lambda}(\xi_1)\cdots\tilde{\lambda}(\xi_n)). \tag{5.1}$$

The void probabilities of a Cox process are given by

$$\mathbb{P}(X(B) = 0) = \mathbb{E}(\exp(-\Lambda(B))).$$

A very simple example of a Cox process is the Poisson point process with intensity measure μ which is obtained by letting $\Lambda \equiv \mu$. Other, more interesting examples are considered below.

Example 5.2 (Mixed Poisson Processes)
A mixed Poisson process is constructed by letting Λ be the Lebesgue measure multiplied by a non-negative random variable Z. The Cox process driven by Λ is then called a ***mixed Poisson process****. If for instance Z is gamma-distributed, then X is called a* ***negative binomial point process****, since the number of events follow a negative binomial distribution.*

Example 5.3 (Random Independent Thinning of a Poisson Process) *Let X be a Poisson point process on $\mathcal{X}$ with intensity measure μ, and let R be a random field where any realisation r of R is a measurable function from $\mathcal{X}$ to $(0,1)$. Observe the thinned process*

$$X_{TH} = \{X_i \in X : Z_i < R(X_i)\}$$

where the Z_i are i.i.d. uniformly distributed on $(0,1)$ and independent of R and X. By Proposition 3.6 X_{TH} conditional on $R = r$ is a Poisson point process with intensity measure $\tilde{\mu}(B) = \int_B r(\xi)\mathrm{d}\mu(\xi)$, and hence X_{TH} is a Cox process driven by

$$\Lambda(B) = \int_B R(\xi)\mathrm{d}\mu(\xi).$$

5.1.1 Log Gaussian Cox Processes

Log Gaussian Cox processes (in the following abbreviated LGCPs) as introduced by Møller et al. (1998) are Cox processes where the logarithm of the intensity function is a Gaussian field. In the following we only consider the stationary case; see Brix & Møller (1998) for a study of the non-stationary case.

Let $Y = \{Y(s) : s \in \mathbb{R}^n\}$ be a stationary **Gaussian field** with mean μ, variance σ^2, and correlation function $r(\cdot)$, that is

$$(Y(s_1), \dots, Y(s_n)) \sim N_n\left((\mu, \dots, \mu), \sigma^2 r(s_i - s_j)_{i,j=1,\dots,n}\right).$$

A **log Gaussian Cox process** is a Cox process, where the driving measure is given by

$$\Lambda(B) = \int_B e^{Y(s)}\mathrm{d}s.$$

This is well defined if $s \to Y(s)$ is continuous almost everywhere, the specific conditions are described in Møller et al. (1998). We thus assume this to be the case and observe that by (5.1)

$$\begin{aligned}\lambda^{(n)}(\xi_1, \dots, \xi_n) &= \mathbb{E}\exp(Y(\xi_1) + \dots + Y(\xi_n)) \\ &= \exp\left(n\mu + \frac{n\sigma^2}{2} + \sigma^2 \sum_{1 \le i < j \le n}^{n} r(\xi_i - \xi_j)\right) \qquad (5.2)\end{aligned}$$

since $Y(\xi_1) + \dots + Y(\xi_n)$ is normally distributed with mean $n\mu$ and variance $\sum_{i,j=1}^n \sigma^2 r(\xi_i - \xi_j)$. Further, we see that

$$\lambda^{(n)}(\xi_1, \dots, \xi_n) = \lambda^n \prod_{1 \le i < j \le n} g(\xi_i - \xi_j),$$

since

$$g(\xi_i - \xi_j) = \frac{\lambda^{(2)}(\xi_i, \xi_j)}{\lambda^2} = \exp(\sigma^2 r(\xi_i - \xi_j)). \tag{5.3}$$

Theorem 5.4
A stationary LGCP is determined by its intensity and pair-correlation function.

Proof:
Let X be a LGCP with underlying Gaussian field Y. Then the void probabilities of X are

$$\mathbb{P}(X(B) = 0) = \mathbb{E} \exp\left(- \int_B e^{Y(s)} \mathrm{d}s \right),$$

which are determined by (μ, σ^2, r).

Now by (5.2)

$$\lambda = \lambda^{(1)}(\cdot) = e^{\mu + \sigma^2/2}$$

and by (5.3),

$$g(\xi_i - \xi_j) = \exp(\sigma^2 r(\xi_i - \xi_j)).$$

Taking $g(0) = \exp(\sigma^2)$ we see that (λ, g) is in one-to-one correspondence with (μ, σ^2, r) and hence X is determined by λ and g. ○

Cox processes have advantages in being a flexible class for modelling clustering as the intensity measure Λ may model a stochastic heterogeneity. They can be naturally extended to multivariate Cox processes where each variable has a driving measure (Cox & Isham 1980).

For the LGCPs we further have simple methods for model validation and parameter estimation, and there are no problems with edge effects since the distribution of a log Gaussian Cox process on a bounded subset is known. In particular, if $\hat{g}$ is a non-parametric estimate of g, $\hat{c} = \log \hat{g}$ is an estimate of the covariance function which may be used for further parametric inference, see Section 7.5. In Møller et al. (1998) another appealing property of LGCPs is described; the unobserved intensity surface can be predicted from a realization of a LGCP observed within a bounded window using a so-called Metropolis-adjusted Langvin algorithm for conditional simulation.

5.2 Cluster Processes

In the following we consider cluster models which for example may be used for modelling plants that spread grain. Further reading may be found in Daley & Vere-Jones (1988), Cressie (1993), and Stoyan et al. (1995).

Definition 5.5 (Cluster Processes)
Let Y be a point process. For each $\xi \in Y$, let Z_ξ be a finite point process associated to the point ξ. Then the ***superposition***

$$X = \bigcup_{\xi \in Y} Z_\xi \tag{5.4}$$

is called a ***cluster process****.*

We name Z_ξ a **cluster**, and its members are called **daughters**. The elements of Y are referred to as **mothers**, and Y is called the **mother process** of X.

A conceptually fundamental type of cluster process is the **Poisson cluster process**, and all cluster models described subsequently are based thereupon.

Definition 5.6 (Poisson Cluster Processes)
Let X be a cluster process, where the mother process Y is a Poisson point process and the clusters Z_ξ are conditionally independent given Y. Then X is called a Poisson cluster process.

Theorem 5.7
Suppose X is a Poisson cluster process with mother process Y and conditional on Y each cluster Z_ξ is a Poisson point process with finite intensity measure μ_ξ such that for any $B \in \mathcal{B}_0$

$$\Lambda(B) = \sum_{\xi \in Y} \mu_\xi(B) < \infty \quad \textit{almost surely.}$$

Then X is a Cox process driven by Λ.

Proof:
The void probabilities of X are given by

$$\begin{aligned}\mathbb{P}(X(B) = 0) &= \mathbb{E} \prod_{\xi \in Y} \mathbb{P}(Z_\xi(B) = 0) \\ &= \mathbb{E} \prod_{\xi \in Y} \exp(-\mu_\xi(B)) \;=\; \mathbb{E} \exp(-\Lambda(B)),\end{aligned}$$

which are the void probabilities of a Cox process driven by Λ. ∘

Next we define the **Neyman-Scott process** which is due to work by Neymann & Scott (1958).

Definition 5.8 (Neyman–Scott Processes)
Let X be a Poisson cluster process on $\mathcal{X}$. If the counts $Z_\xi(\mathcal{X})$ are i.i.d. and the events $\eta \in \cup_{\xi \in Y}(Z_\xi - \xi)$ are i.i.d. with density p w.r.t. the Lebesgue measure on

$\mathbb{R}^n$, then X is called a Neyman–Scott process. In particular, if we in addition assume that $Z_\xi(\mathcal{X})$ has a Poisson distribution, then X is called a Neyman–Scott Poisson process.

Giving the product densities for a Neyman–Scott process conditional on a mother point process we present the following corollary to Theorem 5.7.

Corollary 5.9
Let Y be a mother point process of a Neyman-Scott process X on $\mathcal{X}$, so that conditional on Y each cluster Z_ξ is a Poisson point process with intensity function $cp(\cdot - \xi)$, where c is a constant and p is a density w.r.t. the Lebesgue measure. Then $X \mid Y$ is a Poisson point process with intensity function $\lambda(\xi \mid Y) = c\sum_{\eta \in Y} p(\eta - \xi)$ w.r.t. the Lebesgue measure, and X has product densities

$$\lambda^{(n)}(\xi_1, \ldots, \xi_n) = \mathbb{E}(\lambda(\xi_1 \mid Y) \cdots \lambda(\xi_n \mid Y)).$$

Proof:
By Theorem 5.7, X is a Cox process driven by $\Lambda(\cdot) = \sum_{\xi \in Y} \mu_\xi(\cdot)$. Since μ_ξ has intensity function $cp(\cdot - \xi)$, we get that $X \mid Y = y$ has intensity function $\lambda(\cdot \mid Y = y) = c\sum_{\xi \in Y} p(\cdot - \xi)$. Finally, the product densities follow by (5.1). $\circ$

Theorem 5.10
Assume that X is a Neyman–Scott Poisson process on $\mathcal{X}$, where Y is stationary with intensity $\lambda_Y > 0$, and conditional on Y each cluster Z_ξ has intensity function $cp(\cdot - \xi)$. Then X is stationary with intensity $\lambda = c\lambda_Y$, and pair correlation function given by

$$g(\xi) = 1 + \frac{1}{\lambda_Y} h(\xi),$$

where

$$h(\xi) = \int_{\mathcal{X}} p(\eta - \xi) p(\eta) \mathrm{d}\eta$$

is the density function for $Z_1 - Z_2$, the vector difference between two daughters of the same mother, i.e. Z_1, Z_2 are i.i.d. with density p. Moreover, if $p(\cdot)$ is radially symmetric, that is $p(\xi) = p(\|\xi\|)$, then X is isotropic and $h(\xi)$ depends only on $\|\xi\|$.

Proof:
Since the centered clusters $Z_\xi - \xi$ are i.i.d. and independent of Y, stationarity of Y implies stationarity of X. Moreover, given that Y is isotropic and p is radially symmetric, X is isotropic and $h(\xi) = h(\|\xi\|)$. By Corollary 5.9 and

the Slivnyak–Mecke theorem (Theorem 3.9)

$$\begin{aligned}
\lambda &= \mathbb{E}\lambda(\xi \mid Y) = c\mathbb{E}\sum_{\eta\in Y} p(\eta-\xi) \\
&= c\int_{\mathcal{X}} p(\eta-\xi)\mathrm{d}\Lambda_Y(\xi) \\
&= c\lambda_Y \int_{\mathcal{X}} p(\eta-\xi)\mathrm{d}\xi = c\lambda_Y.
\end{aligned}$$

The second-order density can be written as

$$\begin{aligned}
\lambda^{(2)}(\xi,0) &= \mathbb{E}(\lambda(\xi \mid Y)\lambda(0 \mid Y)) = c^2\mathbb{E}\sum_{\eta\in Y}\sum_{\kappa\in Y} p(\eta-\xi)p(\kappa) \\
&= c^2\left(\mathbb{E}\sum_{\eta,\kappa\in Y}^{\neq} p(\eta-\xi)p(\kappa) + \mathbb{E}\sum_{\eta\in Y} p(\eta-\xi)p(\eta)\right) \\
&= c^2\left(\lambda_Y^2 \int_{\mathcal{X}}\int_{\mathcal{X}} p(\eta-\xi)p(\kappa)\mathrm{d}\eta\mathrm{d}\kappa + \lambda_Y\int_{\mathcal{X}} p(\eta-\xi)p(\eta)\mathrm{d}\eta\right).
\end{aligned}$$

The doubly integral equals 1 since it can be split into two integrals over densities, and the last integrand is simply $h(\xi)$. Thus by definition of the pair correlation function g (see Section 4.2.2)

$$g(\xi) = 1 + \frac{1}{\lambda_Y}h(\xi).$$

○

Special cases of the Neyman–Scott models can be obtained by varying the distribution of the number of daughters or of their spread. Two examples of Neyman-Scott models are Matérn cluster (Matérn 1960) and Thomas processes, which only differ with respect to the distribution of the daughters.

Example 5.11 (Matérn Cluster Processes)
Suppose Y is a Poisson point process with intensity λ. Let the number of daughters of each mother point $y_i \in Y$ have a Poisson distribution with mean ν, and let the daughters be uniformly distributed within a ball of radius R around the mother y_i. Then the superposition of the daughter events constitutes a Matérn cluster process.

Example 5.12 (Thomas Processes)
For the Matérn cluster processes the daughter events were assumed to be uniformly distributed in a ball with radius R around the mother point. Replacing

the uniform distribution by the symmetric normal distribution we obtain the Thomas process. An appealing property of the Thomas process is that the summary statistics $K(r)$ and $g(r)$ are known, and can be written on closed form. Let Z_i, $i = 1, 2$ be independent random variables with $Z_i \sim N_n(0, \sigma^2 I)$, then $Z_1 - Z_2 \sim N_n(0, 2\sigma^2 I)$ with density function

$$h(\xi) = \frac{1}{(4\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\xi\|^2}{4\sigma^2}\right).$$

Hence, by Theorem 5.10 the pair correlation function is given by

$$g(r) = 1 + \frac{1}{\lambda_Y (4\pi\sigma^2)^{n/2}} \exp\left(-\frac{r^2}{4\sigma^2}\right),$$

and in the two-dimensional case Ripley's K-function amounts to

$$K(r) = \pi r^2 + \frac{1 - \exp\left(-\frac{r^2}{4\sigma^2}\right)}{\lambda_Y},$$

as seen by (4.11).

Remarks

As we have seen, certain cluster processes and certain Cox processes can be shown to be identical. In particular, all Neyman–Scott Poisson processes are identical to a Cox process, but the reverse relation is generally unknown, i.e. when a Cox process is identical to a cluster process (Cressie 1993). Note that the former relationship includes the Matérn and Thomas processes.

5.3 Hard-core Processes

A **hard-core point process** is a point process in which the distance between events is larger than a specified minimum distance R. Such a model could be used when each event need to be in an area all by itself, such as for instance trees in a wood which need sunlight to do photo-synthesis.

5.3.1 Matérn Hard-core Processes

Here we define type I and II of the **Matérn hard-core process** as introduced by Matérn (1960). They both evolve from **dependent thinning** of a homogeneous Poisson point process.

Type I

Let $Y = \{y_i\}$ be a homogeneous Poisson point process of intensity λ. From this we define the Matérn hard-core process X_I of type I as

$$X_I = \{y_i \in Y : \|y_i - y_j\| > r \; \forall j \neq i\}.$$

Type II

For the type II model the thinning is less radical as an event within distance r of another is retained under the following circumstances: Let Y be defined as before. Then the Matérn hard-core process X_{II} of type II is given as

$$X_{II} = \{y_i \in Y : U(y_i) > U(y_j), \; \forall \, y_j \in Y \cap b(y_i, r) \backslash y_i\},$$

where $U(y_1), U(y_2), \ldots,$ are i.i.d. with some continuous distribution and independent of Y. Without loss of generality we assume that $U \sim \text{Uniform}(0,1)$.

Clearly, X_I and X_{II} are stationary since Y is stationary. Their intensities are given by the original intensity λ times the probability of retaining each event. For X_I this probability is equal to the probability that no events in Y fall within a ball of radius R, i.e. $\exp(-\lambda|b(0,r)|)$ since Y is a homogeneous Poisson point process. Regarding X_{II}, we observe that with $X_t = \{y \in Y : U(y) > t\}$ the retaining probability is

$$\begin{aligned}
\mathbb{P}(y \in X_{II} \mid y \in Y) &= \int_0^1 \mathbb{P}(X_t(b(0,r)) = 0) \mathrm{d}t \\
&= \int_0^1 \exp(-(1-t)\lambda|b(0,r)|) \mathrm{d}t \\
&= \frac{1 - \exp(-\lambda|b(0,r)|)}{\lambda|b(0,r)|},
\end{aligned}$$

and thus the respective intensities λ_I and λ_{II} are given by

$$\lambda_I = \lambda \exp(-\lambda|b(0,r)|) \tag{5.5}$$

$$\lambda_{II} = \frac{1 - \exp(-\lambda|b(0,r)|)}{|b(0,r)|}. \tag{5.6}$$

This means that (as expected) $\lambda_I < \lambda_{II}$ if X_I and X_{II} stem from the same Y. To specify this difference in intensity we introduce the notion of a **packing density** τ, which is the expected proportion of $\mathbb{R}^n$ covered by balls with centre in points $\xi \in X$ and radius $r/2$. The packing density under Matérn hard-core processes of type I or II is given by the volume of a ball multiplied with the intensity, that is

$$\begin{aligned}\tau_I &= |b(0,r/2)|\lambda_I \\ \tau_{II} &= |b(0,r/2)|\lambda_{II}\end{aligned}$$

Combining these expressions with (5.5) and (5.6) the maximal packing densities obtained are

$$\begin{aligned}\max_{\lambda\in\mathbb{R}_+} \tau_I &= \frac{1}{2^n e} \overset{n=2}{\approx} 0,092 \\ \sup_{\lambda\in\mathbb{R}_+} \tau_{II} &= \frac{1}{2^n} \overset{n=2}{=} 0,25.\end{aligned}$$

These values are very small compared to the respective maximal packing density $\tau_{\max}$, which in two dimensions is $\tau_{\max} = \frac{\pi\sqrt{3}}{6} \approx 0,907$ corresponding to an equilateral triangulation. Currently, $\tau_{\max}$ is not known for $d \geq 3$.

5.3.2 Simple Sequential Inhibition Process (SSI)

The SSI model by Diggle, Besag & Gleaves (1976) is defined on some bounded region $B \in \mathcal{B}_0$ through the inductive procedure:

1. Generate the first point x_1 as $x_1 \sim \text{Uniform}(B)$.
2. Given $x_1, \ldots, x_{j-1}$ then generate $x_j \sim \text{Uniform}(B \setminus \bigcup_{i=1}^{j-1} b(x_i, r))$.
3. The number of events n is determined by that $B \setminus \bigcup_{i=1}^{n} b(x_i, r) = \emptyset$.

Tanemura (1988) has found that the packing density for this process is $\tau_{SSI} \approx 0,547$ in $\mathbb{R}^2$ and thus it models very regular patterns well.

5.4 Markov Point Processes

In this section we treat Markov point processes which were first introduced by Ripley & Kelly (1977). These are finite point processes with a density w.r.t. the distribution of a Poisson point process with finite intensity measure, see Section 3.3.

Let $\sim$ be a reflexive, symmetric relation on $\mathcal{X}$. The **neighbourhood** $\partial(A)$ of any $A \subseteq \mathcal{X}$ is defined by

$$\partial(A) = \{\xi \in \mathcal{X} : \xi \sim \eta, \text{ for some } \eta \in A\}.$$

If $\xi \sim \eta$ we call ξ and η **neighbours**. This definition is often referred to as the **Ripley-Kelly Markov relation**.

A simple example of two points being related is $\xi \sim \eta \Leftrightarrow d(\xi, \eta) < R$ for some $R \in \mathbb{R}_+$.

Definition 5.13 (Markov Point Processes)
A point process X with density p w.r.t. the distribution Π of a Poisson point process with finite intensity measure μ is a Markov point process with respect to $\sim$ if the following properties are satisfied for all $x \in N$:

1. *$p(x) > 0$ implies $p(y) > 0$, for all $y \subseteq x$.*
2. *If $p(x) > 0$ and $\xi \in \mathcal{X} \backslash x$, then*

$$\lambda(x; \xi) = \frac{p(x \cup \xi)}{p(x)}$$

depends only on ξ and $x \cap \partial(\xi) = \{\eta \in x : \eta \sim \xi\}$.

The first property of Definition 5.13 is called the **hereditary** property, $\lambda(x; \xi)$ is called the **Papangelou conditional intensity**, and the second property is called the **local Markov** property, because the Papangelou conditional intensity only depends on x through the neighbourhood of ξ. We note that the local stability condition (Definition 3.10) implies the hereditary property. Further, if $\sim$ is the total relation, i.e. all pairs of points are neighbours, then Markov reduces to hereditary.

Let $N_f \subseteq N$ denote the space of finite point configurations. A function $f : N_f \to [0, \infty)$ satisfying property 1 and 2 in Definition 5.13 is called a **Markov function** with respect to $\sim$. The following proposition gives a result about certain Markov functions. It was first given by Kelly & Ripley (1976) for the case where multiple points are allowed.

Proposition 5.14
Let $\sim$ be an arbitrary reflexive and symmetric relation on $\mathcal{X}$ so that there exist different points $\xi, \kappa, \eta_1, \eta_2, \ldots \in \mathcal{X}$ such that $\xi \sim \eta_i, \xi \sim \kappa$, and $\eta_i \not\sim \kappa$, for $i = 1, 2, \ldots$ Then $f(\cdot)$ is a Markov function with respect to $\sim$ where $\lambda(x; \xi)$ only depends on x through $x(\partial(\xi))$ if and only if

$$f(x) = \alpha \beta^{x(\mathcal{X})} \gamma^{S(x)}$$

for some constants $\alpha, \beta, \gamma \geq 0$ and $S(x) = \sum_{\{\zeta, \nu\} \subseteq x} \mathbb{1}[\zeta \sim \nu]$, taking $0^0 = 1$.

Proof:
Assume that $f(x) = \alpha \beta^{x(\mathcal{X})} \gamma^{S(x)}$. Hence we show that f is a Markov function.

1. Now $f(y) = \alpha\beta^{y(\mathcal{X})}\gamma^{S(y)}$, where $y \subseteq x$. Assume contradictory that $f(x) \neq 0$ and $f(y) = 0$. If

   - $\alpha = 0$ then $f(x) = 0$ which is a contradiction.
   - $\beta = 0$ then $y(\mathcal{X}) > 0$ and hence $x(\mathcal{X}) > 0$ and $f(x) = 0$.
   - $\gamma = 0$ then $S(y) > 0$ and hence $S(x) > 0$ so $f(x) = 0$.

   Thus $f(\cdot)$ is hereditary.

2. Assume $f(x) > 0$. Then, for $\zeta \notin x$

$$\begin{aligned}\lambda(x;\zeta) &= \frac{f(x \cup \zeta)}{f(x)} = \frac{\alpha\beta^{(x\cup\zeta)(\mathcal{X})}\gamma^{S(x\cup\zeta)}}{\alpha\beta^{x(\mathcal{X})}\gamma^{S(x)}} \\ &= \beta\gamma^{x(\partial(\zeta))}.\end{aligned}$$

   Hence $\lambda(x;\zeta)$ only depends on x through $x(\partial(\zeta))$ which verifies the local Markov property of Definition 5.13.

We now assume $f(\cdot)$ to be Markov so that $\lambda(x;\zeta) = g(x(\partial(\zeta)) \geq 0$. Then λ and f are one-to-one up to proportionality using the convention $\frac{0}{0} = 0$. It therefore suffices to verify that $g(x(\partial(\zeta)) = \beta\gamma^{x(\partial(\zeta))}$ for all $\zeta \in \mathcal{X}\backslash x$, where $x \in N_f$.

Set $\alpha = f(\emptyset)$, and since f is hereditary,

$$g(0) = 0 \quad \Rightarrow \quad g(n) = 0, \ \ \forall\, n \in \mathbb{N}_0. \tag{5.7}$$

Set $\beta = g(0)$ and $\gamma = g(1)/\beta$ if $\beta > 0$ and set $\gamma = 0$ if $\beta = 0$. We now claim that

$$g(n) = \beta\gamma^n, \ \ n \in \mathbb{N}_0. \tag{5.8}$$

To prove this, we proceed by induction. By (5.7), if $\beta = 0$ then (5.8) clearly holds, so suppose $\beta > 0$. By definition of β, (5.8) is satisfied when $n = 0$. Assume that (5.8) holds for $n < k$, and let $x = \{\eta_1, \ldots, \eta_{k-1}\}$. Then, still using the convention $\frac{0}{0} = 0$ together with the hereditary property of $f(\cdot)$, we get

$$\begin{aligned}
g(k) &= g((x \cup \kappa)(\partial(\xi))) \\
&= \lambda(x \cup \kappa; \xi) \\
&= \frac{f(x \cup \{\kappa, \xi\})}{f(x \cup \kappa)} \\
&= \frac{f(x \cup \{\kappa, \xi\})}{f(x \cup \xi)} \frac{f(x \cup \xi)}{f(x)} \frac{f(x)}{f(x \cup \kappa)} \\
&= \lambda(x \cup \xi; \kappa) \frac{\lambda(x; \xi)}{\lambda(x; \kappa)} \\
&= g((x \cup \xi)(\partial(\kappa))) \frac{g(x(\partial(\xi)))}{g(x(\partial(\kappa)))} \\
&= g(1) \frac{g(k-1)}{g(0)} \\
&= \beta\gamma \frac{\beta\gamma^{k-1}}{\beta} \\
&= \beta\gamma^k,
\end{aligned}$$

which shows that the claim (5.8) is satisfied. ◦

We notice that if $\gamma \leq 1$ then $\beta^{x(\mathcal{X})}\gamma^{S(x)}$ is locally stable and hence by Proposition 3.11 it is Π-integrable. We may thus choose α so that $p(x) = \alpha\beta^{x(\mathcal{X})}\gamma^{S(x)}$ is a density w.r.t. Π. This is called a **Strauss point process** (Strauss 1975) and will be treated further in Example 5.18.

The following example considers the relation between Markov and Poisson point processes.

Example 5.15 (Poisson Point Processes)
Let X be a Poisson point process with finite intensity measure $\lambda\mu$, where $\lambda \in \mathbb{R}_+$ is a parameter. By Example 3.12, X has a density w.r.t. Π given by

$$p(x) = \exp((1-\lambda)\mu(\mathcal{X}))\lambda^{x(\mathcal{X})}.$$

This density is Markov w.r.t. any symmetric, reflexive relation $\sim$ since $p(x) > 0$ for all $x \in \mathcal{X}$ and $\lambda(x; \xi) = \lambda$ which is a constant.

In Example 3.8 we observed that it was possible to consider a marked Poisson point process as a Poisson point process. In the following example we verify a similar result for marked Markov point processes.

Example 5.16 (Marked Markov Point Processes)
Consider $\mathcal{X} = \mathcal{L} \times \mathcal{M}$ and let Π be the distribution of a Poisson point process on $\mathcal{X}$ with finite intensity measure $\mu = \nu \times \mathcal{M}$, where ν is the intensity measure

of a Poisson point process $\tilde{\Pi}$ on $\mathcal{L}$ and $\mathcal{M}$ is the common distribution of the marks, as in Example 3.8.

Let $X = \{(l_i, m_i)\}$ be a finite marked point process on $\mathcal{X}$ where the locations $L = \{l_i\}$ have a density $\tilde{p}$ w.r.t. $\tilde{\Pi}$ and the marks M are i.i.d. with density m w.r.t. $\mathcal{M}$.

Now consider a mark-independent relation given by $(l_1, m_1) \sim (l_2, m_2) \Leftrightarrow l_1 \approx l_2$, where $\approx$ is a symmetric, reflexive relation on $\mathcal{L}$.

Since the marks are i.i.d. we may write the density of X w.r.t. Π as

$$p(x) = \tilde{p}(l) \prod_{i=1}^{x(\mathcal{X})} m(m_i).$$

If λ_X and λ_L denote the Papangelou intensities for X and L respectively, the Papangelou intensity for X becomes

$$\begin{aligned} \lambda_X(x; (\xi, m_\xi)) &= \frac{p((\xi, m_\xi) \cup x)}{p(x)} \\ &= \frac{\tilde{p}(l \cup \xi) m(m_\xi)}{\tilde{p}(l)} \\ &= \lambda_L(l; \xi) m(m_\xi). \end{aligned}$$

From this we conclude that if L is Markov with respect to $\approx$ then X is Markov with respect to $\sim$. Further, if $m(\cdot) > 0$ and X is Markov w.r.t. $\sim$ then so is L w.r.t. $\approx$.

We are now about to state the Hammersley-Clifford theorem for Markov point processes which is a characterization result first given by Hammersley & Clifford (1971).

In order to give this result we define a point configuration x to be a **clique** w.r.t. a reflexive, symmetric relation $\sim$, if

$$\forall \xi, \eta \in x : \xi \sim \eta.$$

Note that the empty point configuration as well as singletons, are cliques, since $\sim$ is reflexive. Further, we say that $\varphi : N_f \to [0, \infty)$ is an **interaction function** if $\varphi(x) = 1$ whenever x is not a clique.

Theorem 5.17 (Hammersley-Clifford)
Let $f : N_f \to [0; \infty)$. Then f is a Markov function w.r.t. $\sim$ if and only if there exists an interaction function φ such that for all $x \in N_f$

$$f(x) = \prod_{y \subseteq x} \varphi(y). \tag{5.9}$$

Proof:
This proof is due to Ripley & Kelly (1977). First assume that (5.9) holds. If $f(x) > 0$ then by (5.9) $\varphi(y) > 0$ for all $y \subseteq x$, and hence $f(z) = \prod_{y \subseteq z} \varphi(y) > 0$ for $z \subseteq x$. Thus f is hereditary.

In order to show that $\lambda(x;\xi)$ only depends on ξ and $x \cap \partial(\xi)$, consider

$$\begin{aligned} \lambda(x;\xi) &= \frac{f(x \cup \xi)}{f(x)} \\ &= \frac{\prod_{y \subseteq x} \varphi(y) \prod_{y \subseteq x} \varphi(y \cup \xi)}{\prod_{y \subseteq x} \varphi(y)} \\ &= \prod_{y \subseteq x} \varphi(y \cup \xi). \end{aligned}$$

If there exist $\eta \in y$ such that $\eta \not\sim \xi$ then $\varphi(y \cup \xi) = 1$ since φ is an interaction function. Hence it suffices to take the product over all $y \subseteq x : y \subseteq \partial(\xi)$, in which case $\lambda(x;\xi)$ only depends on $x \cap \partial(\xi)$ and ξ, proving the local Markov property.

It now remains to show the existence of an interaction function φ observing (5.9) given that f is Markov with respect to $\sim$. Define φ inductively by

$$\begin{aligned} \varphi(\emptyset) &= f(\emptyset) \\ \varphi(x) &= \begin{cases} \dfrac{f(x)}{\prod_{y \subset x} \varphi(y)} & \text{if } x \text{ is a clique,} \\ 1 & \text{otherwise,} \end{cases} \end{aligned}$$

where we let $\frac{0}{0} = 0$. Note, that φ is well-defined as $\prod_{y \subset x} \varphi(y) = 0$ implies that there exist $y \subset x$ so that $f(y) = 0$ and hence $f(x) = 0$ because of the hereditary property of a Markov function. Three situations can occur:

- x is not a clique and $f(x) = 0$. Then there exist events $\xi, \eta \in x : \xi \not\sim \eta$ and $\varphi(x) = 1$. If

  $$\prod_{y \subset x} \varphi(y) = 0 \quad \Rightarrow \quad f(x) = 0 = \prod_{y \subseteq x} \varphi(y), \tag{5.10}$$

  (5.9) holds. Suppose that $\prod_{y \subset x} \varphi(y) > 0$. Then $f(y) > 0$ for all $y \subset x$, and

  $$0 = \frac{f(x)}{f(x \backslash \eta)} f(x \backslash \eta) = \frac{f(x)}{f(x \backslash \eta)}. \tag{5.11}$$

  We know that f is local Markov and hence (5.11) depends only on η and $x \cap \partial(\eta)$. So removing ξ from the ratio yields the same value and hence

  $$\frac{f(x)}{f(x \backslash \eta)} = \frac{f(x \backslash \xi)}{f(x \backslash \{\xi, \eta\})} > 0$$

which is clearly a contradiction, i.e. $\prod_{y\subset x}\varphi(y)=0$.

- x is a clique and $f(x)=0$. Again we have (5.10) and hence (5.9) holds for $\prod_{y\subset x}\varphi(y)=0$. If $\prod_{y\subset x}\varphi(y)>0$, then $\varphi(x)=0$ by definition and we thus have that

$$f(x)=0=\prod_{y\subseteq x}\varphi(y),$$

by which (5.9) holds.

- $f(x)>0$; we then use induction on $x(\mathcal{X})=n$. For $n=0$, (5.9) holds. Now assume that $n>0$ and x being a clique. Then

$$f(x)=\varphi(x)\prod_{y\subset x}\varphi(y)=\prod_{y\subseteq x}\varphi(y)$$

and the theorem holds.

Assuming $n>0$ and that x is not a clique, we have that $x=z\cup\{\xi,\eta\},\xi\not\sim\eta$, and since f is Markov, $f(z)>0$ and $f(z\cup\xi)>0$. Thus

$$\begin{aligned}
f(x) &= \frac{f(z\cup\{\xi,\eta\})}{f(z\cup\xi)}f(z\cup\xi)\\
&= \frac{f(z\cup\eta)}{f(z)}f(z\cup\xi)\\
&= \frac{\prod_{y\subseteq z\cup\eta}\varphi(y)}{\prod_{y\subseteq z}\varphi(y)}\prod_{y\subseteq z\cup\xi}\varphi(y)\\
&= \left(\prod_{y\subseteq z}\varphi(y\cup\eta)\right)\prod_{y\subseteq z\cup\xi}\varphi(y)\\
&= \prod_{y\subseteq x}\varphi(y)
\end{aligned}$$

and hence (5.9) holds.

This completes the proof. ○

Example 5.18 (Strauss Processes)
A specific Markov point process is the previously mentioned Strauss point process (Strauss 1975) which has a density w.r.t. Π *given by*

$$p(x)=\alpha\beta^{x(\mathcal{X})}\gamma^{S(x)},$$

where $S(x) = \sum_{i<j} \mathbb{1}[\|x_i - x_j\| < R]$, *and* $\beta, R > 0$, *and* $0 \leq \gamma \leq 1$ *are parameters. Here we let* $0^0 = 1$.

The corresponding interaction function φ *w.r.t.* $\xi \sim \eta \Leftrightarrow \|\xi - \eta\| < R$ *is given by*

$$
\begin{aligned}
\varphi(\emptyset) &= \alpha \\
\varphi(\xi) &= \beta \\
\varphi(\{\xi, \eta\}) &= \gamma\, \mathbb{1}[\xi \sim \eta] \\
\varphi(y) &= 1 \qquad \text{for } y(\mathcal{X}) > 2.
\end{aligned}
$$

If $\gamma = 1$ *we have a Poisson point process with intensity measure* $\beta\mu(\cdot)$, *where* μ *is the intensity measure of* Π. *If* $\gamma = 0$ *we have a hard-core point process obtained by considering this Poisson point process conditional on that no pair of events can be closer than* R *units. For* $0 < \gamma < 1$ *there is repulsion between the events which is increasing as* γ *decreases. The case* $\gamma > 1$ *would imply clustering but in this case the density can generally be shown not to be integrable (Kelly & Ripley 1976).*

Both the Strauss and the following multi-scale point process are **pairwise interaction** point processes because the interaction function is 1 for all sets of more than two events.

Example 5.19 (Multi-scale Process)
The Strauss process can be extended to a multi-scale process (Penttinen 1984) where the interaction function is the same except that with $l = \|\xi, \eta\|$

$$
\varphi(\{\xi, \eta\}) = \begin{cases} \gamma_1 & \text{if } l < R_1 \\ \gamma_2 & \text{if } R_1 \leq l < R_2 \\ \vdots & \vdots \\ \gamma_m & \text{if } R_{m-1} \leq l < R_m \\ 1 & \text{if } l \geq R_m, \end{cases}
$$

where $\gamma_1, \ldots, \gamma_m \in [0, 1]$ *and* $R_1 < R_2 < \cdots < R_m$. *Hence, it is Markov w.r.t. the relation* $\xi \sim \eta \Leftrightarrow \|\xi - \eta\| < R_m$.

In principle, any pairwise interaction function where the interaction depends solely on the distance between events, may be approximated by such a multi-scale process.

One could naturally generalize to n-**wise interaction** for any fixed n, but even this does not cover all Markov models as shown in the next example.

Example 5.20 (Area Interaction Point Processes)
An area interaction point process (Baddeley & van Lieshout 1995) has the following locally stable density

$$p(x) = \alpha\beta^{x(\mathcal{X})}\gamma^{-|U(x)|} \tag{5.12}$$

where $U(x) = \cup_{x_i \in x} b(x_i, R)$, *and parameters* $\beta, \gamma, R \geq 0$. *Again we let* $0^0 = 1$. *This is a Markov process since the density may be factorized by the interaction function* φ *given by*

$$\begin{aligned}
\varphi(\emptyset) &= \alpha \\
\varphi(\xi) &= \beta\gamma^{-|b(\xi,R)|} \\
\varphi(y) &= \gamma^{(-1)^{y(\mathcal{X})}|\bigcap_{y_i \in y} b(y_i,R)|} \quad \text{for } y(\mathcal{X}) > 1.
\end{aligned}$$

Since the Papangelou conditional intensity is

$$\lambda(x,\xi) = \beta\gamma^{-|b(\xi,R)\setminus U(x)|}$$

this is an interaction function w.r.t. the relation $\sim$ *which is given by* $\xi \sim \eta \Leftrightarrow \|\xi - \eta\| \leq 2R$. *In the following section we shall return to the area interaction point process in Example 5.31.*

If $\gamma \in (0,1)$ *there is repulsion,* $\gamma = 1$ *yields the Poisson process with intensity measure* $\beta\mu(\cdot)$, *and* $\gamma > 1$ *results in clustering.*

In the next example we briefly look at a pairwise interaction point process where all pairs of events are neighbours, i.e. the interaction is not of finite range as opposed to the multi-scale model.

Example 5.21 (Soft-core Point Processes)
The soft-core point process (Ogata & Tanemura 1984) is a point process with density p *given by*

$$p(x) = \alpha\beta^{x(\mathcal{X})} \prod_{\{\xi,\eta\}\subseteq x} \left[1 - \exp\left(-\frac{\|\xi-\eta\|^2}{\sigma^2}\right)\right],$$

where $\beta, \sigma > 0$ *are the parameters. It hence models repulsion that is very strong for events close together and decreases with distance, but never reaches zero.*

Example 5.22 (Hybrid Hard/Soft–core model)
The ***hybrid hard/soft–core point process*** *has density*

$$p(x) = \alpha\beta^{x(\mathcal{X})} \prod_{\substack{\xi,\eta \in x \\ d=d(\xi,\eta)<R_1}} \mathbb{1}[d > R_0]\exp(\rho(d - R_1)),$$

where R_0 *is the hard-core,* R_1 *is the range of the soft-core, and* ρ *is the power of the soft-core. Hence the hybrid hard/soft–core process is a process with a hard core and a soft edge.*

In the final example we present the **Widom-Rowlinson bivariate model** by Widom & Rowlinson (1970) which embraces the area interaction point process.

Example 5.23 (Widom-Rowlinson's Bivariate Point Process)
Let X_1 and X_2 be i.i.d. Poisson point processes on $\mathcal{X}$ with finite intensity measure $\tilde{\mu}$. Then the bivariate point process $X = (X_1, X_2)$ (see Example 2.8) is a point process on $\mathcal{X} \times \mathcal{X}$ with distribution $\Pi \times \Pi$, where Π denotes the marginal distribution of X_1 and X_2. It can also be considered as a marked Poisson point process (see Example 3.8) with intensity measure $\mu = 2\tilde{\mu} \times \mathcal{M}$, where $\mathcal{M}$ is uniform on $\{1, 2\}$.

*A **Widom-Rowlinson bivariate point process** $Y = (Y_1, Y_2)$ on $\mathcal{X} \times \mathcal{X}$ has a density p on the form*

$$p(y) \quad = \quad \alpha\beta_1^{y_1(\mathcal{X})}\beta_2^{y_2(\mathcal{X})}\mathbb{1}[d(\xi, \eta) > R, \ \forall \xi \in y_1, \eta \in y_2]$$

w.r.t. the distribution of X. The parameters are $\beta_1, \beta_2, R \geq 0$ and since it is locally stable it is a density. Existence of the Widom-Rowlinson bivariate point process can also be seen in another way, since Y is obtained from two independent Poisson point processes conditional on the hard-core condition between events of different types.

The Papangelou conditional intensity is given by

$$\lambda(y; (\xi, 1)) \quad = \quad \beta_1 \mathbb{1}[y_2 \cap b(\xi, R) = \emptyset]$$

assuming without loss of generality that ξ is of type 1, and hence it is Markov w.r.t. the relation $(\xi, i) \sim (\eta, j) \Leftrightarrow (i \neq j \wedge d(\xi, \eta) < R) \vee (\xi = \eta)$.

It is thus a pairwise interaction point process with interaction function

$$\begin{aligned}
\varphi(\emptyset) &= \alpha \\
\varphi((\xi, i)) &= \beta_i \\
\varphi(\{(\xi, i), (\eta, j)\}) &= \mathbb{1}[(\xi, i) \not\sim (\eta, j)] \\
\varphi(y) &= 1 \qquad \textit{otherwise.}
\end{aligned}$$

Observe, that conditionally on $Y_2 = y_2$ we hence get that Y_1 has a density $p_1(y_1) = \alpha_1 \beta_1^{y_i}$ w.r.t. $\Pi_{\mathcal{X} \setminus \bigcup_{\xi \in y_2} b(\xi, R)}$ and by Example 3.12 it is thus a Poisson point process with intensity measure $\beta_1 \tilde{\mu}$ restricted to $\mathcal{X} \setminus \cup_{\xi \in y_2} b(\xi, R)$.

We finally turn to showing how this models relates to the area interaction point *process. Consider the marginal density $\tilde{p}$ of Y_1*

$$\begin{aligned}\tilde{p}(y_1) &= \int_N p(y_1, y_2)\mathrm{d}\Pi(y_2) \\ &= \alpha\beta_1^{y_1(\mathcal{X})}e^{-\tilde{\mu}(\mathcal{X})}\int_N \frac{\beta_2^{y_2(\mathcal{X})}}{y_2(\mathcal{X})!}\mathbb{1}[y_2 \subseteq \mathcal{X}\backslash U(y_1)]\mathrm{d}\Pi(y_2) \\ &= \alpha\beta_1^{y_1(\mathcal{X})}e^{-\tilde{\mu}(\mathcal{X})}e^{\beta_2\tilde{\mu}(\mathcal{X}\backslash U(y_1))} \\ &= \alpha\; e^{(\beta_2-1)\tilde{\mu}(\mathcal{X})}\;\beta_1^{y_1(\mathcal{X})}\;e^{-\beta_2\tilde{\mu}(U(y_1))}\end{aligned}$$

where $U(y_1) = \cup_{\xi\in y_1} b(\xi, R)$. Letting $\gamma = e^{\beta_2}$ and $\tilde{\mu}$ be the Lebesgue measure we notice that this is on the form (5.12), but only incorporates the attractive case as $\gamma \geq 1$.

5.5 Nearest-Neighbour Markov Point Processes

In this section we consider nearest-neighbour Markov point processes; a class of models that generalizes Markov point processes. We only give an introduction to the subject and refer to Baddeley & Møller (1989) for details and results.

For Markov point processes, two points are neighbours regardless of which point configuration they are events in. In nearest-neighbour Markov point processes we change this so that the neighbourhood relation depends on the actual point configuration. For a specific example of this difference between Markov point processes in the Ripley-Kelly sense and the nearest-neighbour Markov point processes, see Example 5.28.

Definition 5.24 (Nearest-Neighbour Markov Point Processes)
Let X be a point process with density p w.r.t. a Poisson point process Π. Let $H \in \mathcal{N}$ be a hereditary set, i.e. $x \in H \Rightarrow y \in H$ for $y \subseteq x$, so that $\Pi(H) = 1$. For each $x \in H$ let $\underset{x}{\sim}$ be a symmetric, reflexive relation on x. The point process X is then a nearest-neighbour Markov point process with respect to $\underset{x}{\sim}$ if the following properties are satisfied for all $x \in N$:

(1) $p(x) > 0$ implies $p(y) > 0$ for all $y \subseteq x$.

(2) If $p(x) > 0$ and $\xi \in \mathcal{X}\backslash x$ so that $x \cup \xi \in H$, then $\lambda(x;\xi)$ depends only on ξ, $\partial(\xi \mid x)$, and the restrictions of $\underset{x}{\sim}$ and $\underset{x\cup\xi}{\sim}$ to $\partial(\xi \mid x)$,

*where $\partial(\xi \mid x) = \{\eta \in x : \eta \underset{x\cup\xi}{\sim} \xi\}$ is the x-**neighbourhood** of ξ.*

We notice that if $H = \mathcal{X}$ and the relations $\underset{x}{\sim}$ are independent of x in the sense that there exists a symmetric, reflexive relation $\sim$ on $\mathcal{X}$ so that for all $x \in N$ the restriction of $\sim$ to x is equal to $\underset{x}{\sim}$, then the definition above reduces to that of a Markov point process (Definition 5.13). According to Baddeley & Møller (1989) the reason why H is used instead of the entire space $\mathcal{X}$ is that it may be impossible to define or guarantee certain properties of $\underset{x}{\sim}$ on the whole of $\mathcal{X}$.

Before giving examples we present an extension of Hammersley-Clifford's theorem to nearest-neighbour Markov point processes. To do so we need a few definitions, most of which resembles those used for Markov point processes.

If $f : H \to [0, \infty)$ satisfies both (1) and (2) in Definition 5.24 then f is a **nearest-neighbour Markov function** w.r.t. $\underset{x}{\sim}$. For $y \subseteq x \in H$ define that y is a $\underset{x}{\sim}$ **clique** if all pairs of events of y are $\underset{x}{\sim}$ related and further define the **clique indicator function** as $\chi(y \mid x) = \mathbb{1}[y \text{ is a } \underset{x}{\sim} \text{ clique}]$.

Definition 5.25 (Interaction Function)
A function on the form $\Phi(y \mid x) = \varphi(y)^{\chi(y|x)}$ *(taking* $0^0 = 0$*) defined on all* $y \subseteq x \in H$ *is called an interaction function w.r.t.* $\underset{x}{\sim}$ *if* $\varphi : H \to [0, \infty)$ *satisfies*

(I1) φ *is hereditary,*

(I2) $\varphi(x) > 0,\ \varphi(\xi \cup \partial(\xi \mid x)) > 0 \quad \Rightarrow \quad \varphi(x \cup \xi) > 0$

for all $\xi \in \mathcal{X}, x \cup \xi \in H$.

Finally we need two consistency conditions on the system of relations $\underset{x}{\sim}$ that allows various results to be shown.

Definition 5.26 (Consistency Conditions)
Let H *be a hereditary subset of* $\mathcal{N}$, *let* $y \subset z \in H$, *and let* $\xi, \eta \in \mathcal{X} \backslash z$ *so that* $x = z \cup \{\xi, \eta\} \in H$. *The two consistency conditions on the system of relations* $\underset{x}{\sim}$ *are then given by :*

(C1) $\chi(y \mid z) \neq \chi(y \mid z \cup \xi) \quad \Rightarrow \quad y \subset \partial(\xi \mid y)$

(C2) $\xi \underset{x}{\not\sim} \eta \quad \Rightarrow \quad \chi(y \mid z \cup \xi) + \chi(y \mid z \cup \eta) = \chi(y \mid z) + \chi(y \mid x)$.

We are now ready to present Hammersley-Clifford's theorem, referring to Baddeley & Møller (1989) for a proof.

Theorem 5.27 (Hammersley-Clifford)
Let $f : H \to [0, \infty)$ where H is a hereditary subset of $\mathcal{X}$. If $\underset{x}{\sim}$ satisfies (C1) and Φ is an interaction function w.r.t. $\underset{x}{\sim}$ so that

$$f(x) \quad = \quad \prod_{y \subseteq x} \Phi(y \mid x) \quad \forall x \in H \tag{5.13}$$

then f is a nearest-neighbour Markov function w.r.t. $\underset{x}{\sim}$.

If $\underset{x}{\sim}$ satisfies (C1) and (C2) then the reverse is true. That is, if f is a nearest-neighbour Markov function w.r.t. $\underset{x}{\sim}$ then there exists an interaction function Φ w.r.t. $\underset{x}{\sim}$ so that (5.13) is true.

With this result in mind, we present some examples of nearest-neighbour Markov point processes.

Example 5.28 (Sequential Neighbours on $\mathbb{R}$)
Let $\mathcal{X} \subseteq \mathbb{R}$ and let $x = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$ where $x_i < x_j$ if $i < j$. If we define $x_i \underset{x}{\sim} x_j \Leftrightarrow |i - j| \leq 1$, we see that the cliques are sets of zero or one event and the sets $\{x_i, x_{i+1}\}$ for $i = 1, \dots, n-1$. Clearly $\underset{x}{\sim}$ satisfies (C1) and thus, by Hammersley-Clifford (Theorem 5.27), f is a nearest-neighbour Markov function if it is on the form

$$f(x) \quad = \quad \varphi(\emptyset) \prod_{i=1}^{n} \varphi(x_i) \prod_{i=1}^{n-1} \varphi(\{x_i, x_{i+1}\}).$$

This is not a Markov function in the Ripley-Kelly sense since $x = \{2, 4\}$ has $2 \underset{x}{\sim} 4$, but $x = \{2, 3, 4\}$ has $2 \underset{x}{\not\sim} 4$, so no symmetric, reflexive relation $\sim$ on $\mathcal{X}$ exists so that its restriction on x is equal to $\underset{x}{\sim}$ for all $x \in N$.

By extending $\mathcal{X}$ from $\mathbb{R}$ to $\mathbb{R}^2$ we obtain another example due to Baddeley & Møller (1989) which in practice may be used when describing nuclei cells.

Example 5.29 (Dirichlet Tessellation)
Let $\mathcal{X}$ be a bounded Borel set in $\mathbb{R}^2$ and $x = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$. We say that x is in ***General Quadratic Position*** *(GQP) if*

- *x has no multiple points,*
- *no three points in x lie on the same line i $\mathbb{R}^2$,*
- *no four points in x lie on (the boundary of) the same sphere in $\mathbb{R}^2$,*

and let $H = \{x \in \mathcal{X} : x \text{ is in } GQP\}$ which is clearly a hereditary set.

*We then define the **Dirichlet** or **Voronoi cell** $C(\xi \mid x)$ of $\xi \in x$ as*

$$C(\xi \mid x) = \{\eta \in \mathcal{X} : \|\xi - \eta\| \leq \|\kappa - \eta\| \quad \forall \kappa \in x\},$$

i.e. the set of points in $\mathcal{X}$ that are closer or as close to ξ as to any other event in x.

*The concept of GQP ensures that edges of cells $C(\xi \mid x), \xi \in x$ cannot be parallel, and two cells cannot intersect in only one point. We define the **Dirichlet relation** by $\xi \underset{x}{\sim} \eta \Leftrightarrow C(\xi \mid x) \cap C(\eta \mid x) \neq \emptyset$.*

In Figure 5.1 a Dirichlet tessellation of biological cells is shown. An analysis of this dataset can be seen in Chapter 8.

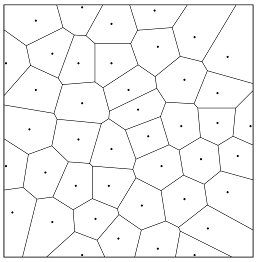

Figure 5.1: Dirichlet tessellation of biological cell data.

In Baddeley & Møller (1989) it is shown that $\Pi(H) = 1$, that $\underset{x}{\sim}$ satisfies the consistency conditions (C1) and (C2), and that no $\underset{x}{\sim}$ clique consists of more than 3 points.

Applying the Hammersley-Clifford Theorem we hence obtain that f is a nearest-neighbour Markov function w.r.t. the Dirichlet relation if and only if

$$f(x) = \varphi(\emptyset) \prod_{\xi \in x} \varphi(\xi) \prod_{\substack{\{\xi,\eta\} \in x \\ \chi(\{\xi,\eta\} \mid x) = 1}} \varphi(\{\xi,\eta\}) \prod_{\substack{\{\xi,\eta,\kappa\} \in x \\ \chi(\{\xi,\eta,\kappa\} \mid x) = 1}} \varphi(\{\xi,\eta,\kappa\})$$

for some φ satisfying (I1) and (I2) in Definition 5.25.

The example above can be generalized to n dimensions, see Baddeley & Møller (1989) and Møller (1994).

Example 5.30 (Ord's Process)
Let $\underset{x}{\sim}$ be the Dirichlet relation and consider the ***iterated Dirichlet relation*** $\underset{x}{\sim}^2$ *given by*

$$\xi \underset{x}{\sim}^2 \eta \Leftrightarrow \exists \kappa \in x : \xi \underset{x}{\sim} \kappa \text{ and } \eta \underset{x}{\sim} \kappa.$$

According to Baddeley & Møller (1989) $\underset{x}{\sim}^2$ can be shown to satisfy (C1). Hence, with H as in Example 5.29, the density $p : H \to \mathbb{R}_+$ is given by

$$p(x) = \alpha \beta^{x(\mathcal{X})} \prod_{\xi \in x} g(C(\xi \mid x))$$

is (by Hammersley-Clifford) seen to be a nearest-neighbour Markov function w.r.t. $\underset{x}{\sim}^2$ for $\beta, g(\cdot) > 0$, since $C(\xi \mid x)$ is a function of $\partial(\xi \mid x)$. The function g could for instance be the Lebesgue measure.

The point process with density p w.r.t. Π was first introduced by Ord in the discussion of Ripley (1977).

The last two examples consider the **connected component relation** $\underset{x}{\sim}$ which is defined from any symmetric, reflexive relation $\sim$ by $\xi \underset{x}{\sim} \eta$ if there exists $\{x_1, \ldots, x_n\} \subseteq x$ so that $\xi = x_1 \sim x_2 \sim \cdots \sim x_n = \eta$.

Example 5.31 (Area Interaction Process)
The area interaction point process with radius R from Example 5.20 may naturally be seen as a nearest-neighbour Markov point process with respect to a connected component relation $\underset{x}{\sim}$ where $\sim$ is given by $\xi \sim \eta \Leftrightarrow \|\xi - \eta\| \leq 2R$. An area interaction point process has the following density

$$p(x) \quad = \quad \alpha \beta^{x(\mathcal{X})} \gamma^{-|U(x)|}$$

where $U(x) = \cup_{x_i \in x} b(x_i, R)$.

Define $\Phi(y \mid x) = \varphi(y)^{\chi(y|x)}$ by

$$\begin{aligned}
\varphi(\emptyset) &= \alpha \\
\varphi(y) &= \beta^{y(\mathcal{X})} \gamma^{-|U(y)|} \quad \text{if } y \neq \emptyset \text{ is a maximal clique w.r.t. } \underset{x}{\sim} \\
\varphi(y) &= 1 \quad \text{otherwise.}
\end{aligned}$$

Thus

$$p(x) \quad = \quad \prod_{y \subseteq x} \Phi(y \mid x)$$

and since Φ clearly is an interaction function and $\underset{x}{\sim}$ satisfies (C1), then, by Hammersley-Clifford, X is a nearest-neighbour Markov point process with respect to $\underset{x}{\sim}$.

The final example is due to Baddeley, van Lieshout & Møller (1996) and shows that certain cluster processes are nearest-neighbour Markov point processes. We refer to this article for proofs.

Example 5.32 (Cluster Process)
Consider a cluster process X as defined in Section 5.2 on page 59. Assume Y is finite and for all $\xi \in Y$, $\mathbb{P}(Z_\xi \subseteq b(\xi, r)) = 1$. If Y is a Poisson point process then X is a nearest-neighbour Markov point process with respect to the connected component relation with $\sim$ of range $2R$. If Y is Ripley-Kelly Markov (e.g. the Strauss process) of range r and the clusters are non-empty, then X is a nearest-neighbour Markov point process w.r.t. the connected component relation with $\sim$ of range $2R + r$. However, X is in general not *Ripley-Kelly Markov in this situation.*

5.6 Gibbs Point Processes

Gibbs point processes originate in statistical physics, where they concern the existence and uniqueness of equilibrium states for dynamic particle systems given certain local dynamic specifications (Ruelle 1969). In this section we first define Gibbs point processes for finite range Markov processes before giving the definition for the general case. The material presented here builds partly on Preston (1976), Georgii (1988), and Stoyan et al. (1995).

Let φ be an interaction function of finite range, i.e. $\varphi : N \to [0, \infty)$ is measurable with $\varphi(x) = 1$ if there exists $\xi, \eta \in x$ so that $d(\xi, \eta) > R$, where R is a constant. Let Π be the distribution of a Poisson point process with locally finite intensity measure μ on $\mathcal{X}$ and for $A \in \mathcal{B}_0$ let Π_A denote Π restricted to A, i.e. $X_A = X \cap A \sim \Pi_A$ if $X \sim \Pi$. Further, for $B \in \mathcal{B}$ let N_B be the set of locally finite point configurations in B and let $\mathcal{N}_B = \{F_B : F \in \mathcal{N}\}$, where $F_B = \{x \cap B : x \in F\}$.

Assume for the moment that μ is finite and that X is a finite Markov point process with interaction function φ. Thus the density of X w.r.t. Π is hereditary and given by

$$
\begin{aligned}
p(x) &= \prod_{y \subseteq x} \varphi(y) = \prod_{z \subseteq x_{\partial(A)}} \varphi(z) \prod_{z \subseteq x_{A^c}} \varphi(z) \left(\prod_{z \subseteq x_{\partial(A) \setminus A}} \varphi(z) \right)^{-1} \\
&= \frac{p(x_{\partial(A) \setminus A} \cup x_A)}{p(x_{\partial(A) \setminus A})} \, p(x_{A^c}) \\
&= \lambda(x_{\partial(A) \setminus A}; x_A) \, p(x_{A^c}),
\end{aligned}
$$

where $\lambda(x_{\partial(A) \setminus A}; x_A)$ is a natural extension of the Papangelou conditional intensity.

For $p(x_{A^c}) > 0$, the conditional distribution of X given $X_{A^c} = x_{A^c}$ has density $p_A(\cdot \mid x_{A^c})$ w.r.t. Π_A given by

$$\begin{aligned} p_A(x_A \mid x_{A^c}) &= \frac{p(x_A \cup x_{A^c})}{\int\limits_{N_A} p(y_A \cup x_{A^c}) \mathrm{d}\Pi_A(y_A)} \\ &= \frac{\lambda(x_{\partial(A)\setminus A}; x_A) p(x_{A^c})}{p(x_{A^c}) \int\limits_{N_A} \lambda(x_{\partial(A)\setminus A}; y_A) \mathrm{d}\Pi_A(y_A)} \\ &= c_A(x_{\partial(A)\setminus A})\ \lambda(x_{\partial(A)\setminus A}; x_A), \end{aligned}$$

which does not depend on $x_{A^c \setminus \partial(A)}$.

Motivated by this we generalize to the case where μ is locally finite, defining a **local specification** (in the sense of Preston (1976)) as a set of densities

$$p_A(x_A \mid x_{\partial(A)\setminus A}) = c_A(x_{\partial(A)\setminus A})\ \lambda(x_{\partial(A)\setminus A}; x_A) \tag{5.14}$$

w.r.t. Π_A defined for all $A \in \mathcal{B}_0$ and all $x_{\partial(A)\setminus A} \in N_{\partial(A)\setminus A}$ with $p(x_{\partial(A)\setminus A}) > 0$. This ensures that a local specification is consistent in the manner considered by Preston (1976) and Georgii (1988).

Definition 5.33 (Gibbs Point Process: Finite Range of Interaction)
A point process X on $\mathcal{X}$ is a Gibbs point process with finite range R of interaction w.r.t. the local specification (5.14) *if the distribution of X agrees with all the conditional densities in the local specification. That is,*

$$\mathbb{P}(X_A \in F_A \mid X_{A^c} = x_{A^c}) = \int\limits_{F_A} p_A(x_A \mid x_{\partial(A)\setminus A}) \mathrm{d}\Pi_A(x_A) \tag{5.15}$$

for all $A \in \mathcal{B}_0$, $F_A \in \mathcal{N}_A$ and $x_{A^c} \in N_{A^c}$ where $p(x_{\partial(A)\setminus A}) > 0$.

Given a local specification it is by no means clear that a Gibbs point process exists nor that the specification is restrictive enough to define a unique process. We refer to Preston (1976) and Georgii (1988) for a detailed treatment of existence and uniqueness of Gibbs point processes.

Example 5.34 (Poisson Point Processes)
Let a local specification be given by letting all the conditional densities be equal to 1*. Then* (5.15) *reduces to*

$$\mathbb{P}(X_A \in F_A \mid X_{A^c} = x_{A^c}) = \Pi_A(F),$$

which is obtained by letting $\mathbb{P} = \Pi$ be the distribution of a Poisson point process with intensity measure μ on $\mathcal{X}$. Thus the Poisson point process may be regarded as a Gibbs point process.

For an alternative approach, consider a Poisson point process X with intensity measure $\lambda\mu$. By Example 3.12 it has a density $p(x) = \exp((1-\lambda)\mu(\mathcal{X}))\lambda^{x(\mathcal{X})}$ w.r.t. Π, and hence $\lambda(x_{\partial(A)\setminus A}; x_A)$ is given by $\lambda^{x(A)}$ so again by Example 3.12

$$p_A(x_A \mid x_{\partial(A)\setminus A}) = \lambda^{x(A)} \exp((1-\lambda)\mu(A)).$$

By Definition 5.33 we see that X is a Gibbs process since

$$\begin{aligned} \mathbb{P}(X_A \in F_A \mid X_{A^c} = x_{A^c}) &= \mathbb{P}(X_A \in F_A) \\ &= \int_{F_A} \lambda^{x(A)} \exp((1-\lambda)\mu(A)) \mathrm{d}\Pi_A(x), \end{aligned}$$

i.e. the conditional densities in the local specification coincide with the distribution of X.

We are now ready to extend the definition of Gibbs point processes with Definition 5.35 below, see Nguyen & Zessin (1979). It may be shown, that this general definition reduces to Definition 5.33 in the case of finite range of interaction, see also Example 5.37.

Definition 5.35 (Gibbs Point Process: General Case)
Let μ be a locally finite measure on a separable metric space $\mathcal{X}$ and let X be a point process on $\mathcal{X}$. If there exists a non-negative measurable function m on $\mathcal{X} \times N$ so that

$$\mathbb{E} \int_{\mathcal{X}} h(\xi, X) m(\xi, X) \mathrm{d}\mu(\xi) = \mathbb{E} \sum_{\xi \in X} h(\xi, X\setminus\xi) \tag{5.16}$$

for all non-negative measurable functions h on $\mathcal{X} \times N$, then X is a Gibbs point process.

If X is a finite range Markov point process then m is the Papangelou conditional intensity as will be shown in Example 5.37.

By (5.16) and the definition of the reduced Campbell measure we observe that

$$\begin{aligned} \mathbb{E} \int_{\mathcal{X}} \mathbb{1}[\xi \in A, X \in F] m(\xi, X) \mathrm{d}\mu(\xi) &= \mathbb{E} \sum_{\xi \in X} \mathbb{1}[\xi \in A, X\setminus\xi \in F] \\ &= C^!(A \times F). \end{aligned}$$

Hence, (5.16) implies that

$$C^!(A \times F) = \int_F \int_A m(\xi, x)\ \mathrm{d}\mu(\xi)\mathrm{d}\mathbb{P}(x), \tag{5.17}$$

and by standard measure theoretical arguments (5.17) also implies (5.16).

Example 5.36 (Poisson Point Process Revisited)
If X is a Poisson point process with intensity measure μ we see that letting $m \equiv 1$, (5.16) reduces to Theorem 3.9 (Slivnyak-Mecke), so the Poisson point process with intensity measure μ is a Gibbs process.

Example 5.37 (Finite Range Markov Point Processes)
Let X be a finite range Markov point process with interaction function φ and let Π denote the Poisson point process with locally finite intensity measure μ. Now take $m(\xi, x) = \lambda(x; \xi)$ and examine (5.16). Observe that

$$\begin{aligned} \mathbb{E} \sum_{\xi \in X} h(\xi, X \backslash \xi) &= \int_N \sum_{\xi \in x} h(\xi, x \backslash \xi)\; p(x) \mathrm{d}\Pi(x) \\ &= \mathbb{E} \sum_{\xi \in Y} h'(\xi, Y \backslash \xi), \end{aligned} \tag{5.18}$$

where $Y \sim \Pi$ and $h'(\xi, Y\backslash\xi) = h(\xi, Y\backslash\xi)p(Y)$. Further,

$$\begin{aligned} \mathbb{E} \int_{\mathcal{X}} h(\xi, X)\lambda(X;\xi)\mathrm{d}\mu(\xi) &= \int_N \int_{\mathcal{X}} h(\xi, x)\frac{p(x \cup \xi)}{p(x)}\mathrm{d}\mu(\xi)p(x)\mathrm{d}\Pi(x) \\ &= \mathbb{E} \int_{\mathcal{X}} h'(\xi, Y)\mathrm{d}\mu(\xi), \end{aligned} \tag{5.19}$$

since p is hereditary.

Finally we see by Example 5.36 that (5.18) is equal to (5.19), and thus (5.16) is fulfilled for a finite range Markov point process.

Chapter 6

Simulation

Due to their analytical intractability there is obviously a need for simulating spatial point processes. One might wish to see what a "typical" realization of a certain point process looks like, and there is a need to simulate point patterns for summary statistics used for model selection and validation, in order to do Monte Carlo tests (see e.g. Diggle (1983)), to do approximate maximum likelihood estimation (Geyer & Møller 1994, Geyer 1999) or to investigate the performance of certain inference procedures. Moreover, simulation may be used to do prediction or more generally, to do Bayesian inference, see e.g. Møller et al. (1998) and Heikkinen & Arjas (1998).

In this chapter we consider different methods to simulate the point processes described in Chapters 3 and 5. First we examine a general method for simulating point patterns, namely rejection sampling. Next we consider simulation of binomial and Poisson processes, as they are fundamental in simulating other point processes. When the density we wish to sample from only is known up to proportionality, a different sampling scheme may be used. This is done using Markov chain Monte Carlo theory, and here the Metropolis-Hastings algorithm becomes useful. We conclude the chapter with an introduction to spatial birth-and-death processes and perfect simulation.

6.1 Rejection Sampling

Rejection sampling can be used for simple models, such as simulation of the binomial process and for the Gibbs sampler.

Let the situation be as in Section 3.3. Suppose we wish to sample from a distribution with density p (referred to as the target density) w.r.t. some measure (which for example could be Π). Assume further that p is known up to proportionality, i.e. $p(\cdot) = c_1 h(\cdot)$. If it is difficult to sample directly from this distribution, we may instead sample from another distribution with density $f(\cdot) = c_2 g(\cdot)$ w.r.t. Π. This other distribution has to be easy to sample from, and it is used as a way to sample from $p(\cdot)$ as described in the following algorithm.

The procedure is summarized in the following algorithm.

Algorithm [Rejection Sampler]

Assume that there exists a $c > 0$ such that for all $x \in N : h(x) \leq cg(x)$.

1. Generate X from f. Independently of X generate U from Uniform$([0,1])$.
2. If $U \leq \frac{h(X)}{cg(X)}$ (taking $\frac{0}{0} = 0$), return $Z = X$ and stop; otherwise return to step 1.

The following Proposition provides useful information about the output from the rejection sampler.

Proposition 6.1
The output Z of the rejection sampler follows the target density p, and the number of loops before the acceptance state is reached is geometrically distributed with

$$\mathbb{P}(\text{Acceptance}) = \mathbb{P}\left(U \leq \frac{h(X)}{cg(X)}\right) = \frac{c_2}{c_1 c}.$$

Proof:
Let $F \in \mathcal{N}$, then

$$\begin{aligned}
\mathbb{P}\left(X \in F, U \leq \frac{h(X)}{cg(X)}\right) &= \int_F \frac{h(x)}{cg(x)} f(x) \mathrm{d}\Pi(x) = \int_F \frac{p(x)/c_1}{cf(x)/c_2} f(x) \mathrm{d}\Pi(x) \\
&= \frac{c_2}{c_1 c} \int_F p(x) \mathrm{d}\Pi(x),
\end{aligned}$$

whereby $\mathbb{P}(\text{Acceptance}) = \frac{c_2}{c_1 c}$ and $\mathbb{P}(Z \in F) = \mathbb{P}(X \in F \mid \text{Acceptance}) = \int_F p(x) \mathrm{d}\Pi(x)$. ∘

It is worth pointing out that we need to find g such that both h/g is bounded and g is integrable w.r.t. Π. Also it is of importance that g is a 'good' enveloping function for the rejection sampler, such that the acceptance probabilities will be as large as possible. In order to fulfill this criterion, $cg(\cdot)$ is supposed to be close to $h(\cdot)$.

Ripley (1977) discusses rejection sampling for a hard-core model fitted to a particular dataset, where the number of events is kept fixed. The acceptance probability for that particular sample is

$$\mathbb{P}(\text{Acceptance}) = 10^{-24},$$

i.e. the rejection sampler is very infeasible for this example.

Because of the often low acceptance probability, it seems that rejection sampling is impractical for simulation of some point processes.

6.2 Simulation of Various Point Processes

In this section we investigate how to make simulations of the point processes described in Chapters 3 and 5. As we shall see, the simulation procedures for some point processes follows directly from their definition.

We introduce different ways of simulating the Poisson point process, but first we naturally turn to simulation of the binomial point process. For simplicity we restrict attention to simulation on bounded regions in $\mathbb{R}^n$, but the procedures can (at least in theory) easily be extended to cover other spaces.

Simulation of the Binomial Point Process

When simulating the binomial point process on a square $R = [0, 1]^2$ we simply need to distribute independent events uniformly on R. Here a practical problem arises, as a random number generator is needed. Implementing a pseudo random number generator solves this problem to some extent, but since this is only pseudo random we can by no means expect to get a sequence of events that are completely independent and uniformly distributed (see e.g. Ripley (1987) for further details). Disregarding this problem, we let $y_1, y_2, \dots$ be a sequence of independent uniformly distributed variables on $[0, 1]$. Then the events $x_i = (y_{2i-1}, y_{2i})$ for $i = 1, 2, \dots$ are uniformly distributed on R and we only have to choose the number of events to be simulated. It is straightforward to generalize this technique to n-dimensional hyper-cubes $[0, 1]^n$ and through translations and scaling, a binomial point process can be simulated on any n-dimensional rectangle or hyper-cube. In practice this method will cover most applications. However, if the region R is not rectangular or a hyper-cube, we have to apply different methods.

Rejection Sampling: A hyper-cube $W \subseteq \mathbb{R}^n$ is chosen so that the region R, where the events are to be distributed, is enclosed in W. Now use the method of sampling explained above on W. When an event falls in R it will be uniformly distributed in R. Events that fall in W but outside R are rejected. When the required number of events are found (all situated in R) they constitute the binomial point process on R. Effectiveness of this method is of course maximized when W is chosen as small as possible. To see that this actually corresponds to rejection sampling, let the un-normalized density h on R be given by

$$h(\xi) = \begin{cases} g(\xi) & \text{if } \xi \in R \\ 0 & \text{otherwise} \end{cases}$$

where $g \equiv 1$ is the un-normalized density of the binomial process.

Approximation: k disjoint hyper-cubes are chosen so that they cover the region R in question. A hyper-cube is selected with probability proportional to the volume of the hyper-cube divided by the volume of the k hyper-cubes combined. In the selected hyper-cube we generate an event uniformly. Repeatedly select a hyper-cube and generate an event until the required number of events are obtained. This method can be used together with rejection sampling to ensure satisfactory simulations on complicated regions.

For specific regions there exists more efficient methods. For example, if R is a disc, the radial simulation procedure in Quine & Watson (1984) can be used or if R has some sort of symmetry it is possible to use the method with transformation of coordinates as in Stoyan et al. (1995).

Simulation of the Homogeneous Poisson Process

As was shown in Section 3.3, conditioning on the number of events in a Poisson point process, results in a binomial point process. This makes the simulation methods described in relation to the binomial point process applicable to simulations in the Poissonian case. Hence the simulation of a Poisson point process can be completed by following the two steps:

1. Generate a number n from a Poisson distribution with mean $\lambda|R|$.
2. Simulate a binomial point process with n events on R.

Figure 6.1 shows the result of such a simulation. Note that the plots may seem to exhibit clustering, but they are in fact examples of complete spatial randomness.

Since simulation of a binomial point process has been addressed we now consider the simulation of a Poisson distribution. Two standard methods (a) and (b) are described below:

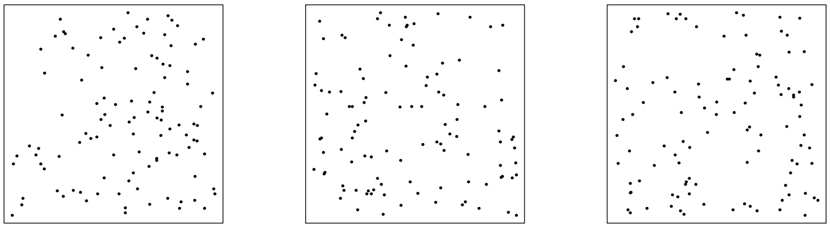

Figure 6.1: Three simulations of a homogeneous Poisson point process with intensity $\lambda = 100$ on the unit square.

(a) Generate $Y \sim \text{Uniform}([0,1])$. If q is the cumulative distribution function for the Poisson distribution with mean $\lambda|R|$, then the smallest n, so that $q(n) > Y$, will be Poisson distributed with the desired mean. Thus iteratively calculate $q(n) = q(n-1) + p(n)$ until $q(n) > Y$, where $q(0) = p(0)$ and p is the density of the Poisson distribution. Note that

$$p(n) \quad = \quad \frac{(\lambda|R|)^n}{n!} \exp(-\lambda|R|)$$

can be calculated recursively as $p(n) = \frac{\lambda|R|}{n} p(n-1)$ for $n \in \mathbb{N}$.

(b) Another way of simulating the Poisson distribution takes advantage of the construction made in Example 2.4. There we discovered that for the point process $X = \{X_1, X_1 + X_2, \dots\}$, where X_i are i.i.d. exponentially distributed with parameter λ, we have that

$$m = X((0, |R|]) \quad \sim \quad \text{Poisson}(\lambda|R|),$$

where $m \geq 0$ is the largest integer so that

$$\sum_{i=1}^{m} X_i \leq |R|.$$

Thus iteratively simulating X_i until the sum gets larger than the size of R provides a way to simulate the Poisson distribution. Note that X_i may be generated as $-\frac{1}{\lambda} \log U_i$, where $U_i \sim \text{Uniform}([0,1])$.

Simulation of Inhomogeneous Poisson Point Processes

Simulating an inhomogeneous Poisson point process Y with intensity function λ on a region $R \subset \mathbb{R}^n$ may be done using **independent thinning**.

Suppose $\sup_{r \in R} \lambda(r) \leq K < \infty$. Simulate a homogeneous Poisson point process X with rate K on R and let

$$Y = \left\{ x_i \in X : Z_i < \frac{\lambda(x_i)}{K} \right\},$$

where the Z_i are uniformly i.i.d. on $(0, 1)$ and independent of X.

Hence, by Proposition 3.6, Y is a Poisson point process with intensity measure $\mu(B) = \int_B \lambda(\xi) \mathrm{d}\xi$, where $B \in \mathcal{B}(R)$.

Figure 6.2 shows the result of such a simulation. Note that events are sparse in the upper half of the plots.

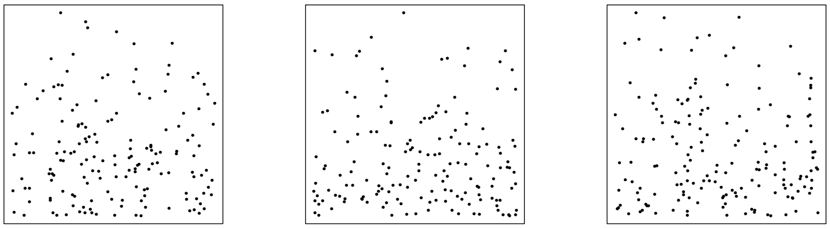

Figure 6.2: Three simulations of an inhomogeneous Poisson point process with intensity function $\lambda(x, y) = 500(1 - \sqrt{y})$ on the unit square.

Simulation of Cox and Log Gaussian Cox Processes

As the Cox process arises from a two-step procedure, so does simulation of this process. Hence following the definition of a Cox process in Section 5.1 a method for simulating this type of process becomes apparent. In the first step a method to find a realization μ of the random measure Λ on $\mathcal{X} = \mathbb{R}^n$ has to be found. Note that the existence of such a method by no means is obvious, however suggestions are given in Cressie (1993). Now the Cox process can be simulated by simulating a Poisson process with intensity measure μ.

Simulation of the log Gaussian Cox process also involves two steps. First the Gaussian field is simulated, and secondly, conditional on the Gaussian field, an inhomogeneous Poisson process is simulated. Simulation of the Gaussian field is described in Wood & Chan (1994) by means of a diagonalization by the two-dimensional discrete Fourier transform, while Lantuéjoul (1994) describes several other methods. In practice, the region of interest is represented by a grid and we approximate the Gaussian process by the Gaussian distribution on the grid. Assume we restrict the LGCP to the unit square on $\mathcal{X} \subseteq \mathbb{R}^2$, where we consider the Gaussian field $Y = (Y_{ij})_{(i,j)\in I}$; here I is a square lattice approximating the unit square. The inhomogeneous Poisson process can then be simulated in two steps:

1. Within each cell C_{ij} of the grid simulate a homogeneous Poisson Process with intensity $\lambda_{ij} = \exp(y_{ij})$.
2. Thinning the homogeneous Poisson process with intensity given by $\lambda_{max} = \max\limits_{(i,j)\in I} \lambda_{ij}$, so that the probability of retaining a point in the ijth cell is $\lambda_{ij}/\lambda_{max}$.

Simulation of Matérn Cluster and Thomas Processes

The definitions of the Matérn cluster and Thomas processes entail descriptions of how they are to be simulated on bounded regions; see Section 5.2. However, we need to consider the situation where a mother process outside the bounded region R yields daughters that lie inside the region. These boundary effects can be reduced heavily by simulating on an extended region R^* containing R such that a mother point outside R^* yield daughter points inside R with a negligible probability.

Further it is possible to use another procedure. This procedure is taking advantage of the result in Section 5.2, i.e. the fact that the Matérn and Thomas processes can be described as Cox processes. Hence the simulation procedure for Cox processes described above is applicable here.

Simulation of Hard-core Processes

Ways of simulating the hard-core processes described in Section 5.3 follow immediately from their definitions. For example, consider the Matérn hard-core process of type II, recalling that this process essentially is a dependent thinning applied to a homogeneous Poisson process with intensity λ.

Let Y denote a homogeneous Poisson process with intensity λ, and let W be the window of observation. Simulate Y on $W_{\oplus r}$ and attach independently to each point a uniform mark U distributed on $(0,1)$. Then perform the thinning described in Section 5.3. The desired sample then consists of the retained points in W.

Simulation of Markov Point Processes

Simulation of Markov point processes can be done in several ways. We shall treat the Metropolis–Hastings algorithm, Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953) and Hastings (1970), in Section 6.4, and further treat spatial birth-and-death processes in Section 6.6. We refer the reader to Møller (1999) for a historical introduction.

6.3 Markov Chains

In this section we describe a general theory called Markov chain Monte Carlo (MCMC) which we will use to simulate certain finite point processes. A main reference for MCMC is Meyn & Tweedie (1993). See Møller (1999) for MCMC for point processes and Gilks, Richardson & Spiegelhalter (1996) for more practical aspects of MCMC.

Let X be a finite point process with distribution π which has a density p w.r.t. Π. We call π the **target distribution** and consider how to generate a chain of point processes $X_0 \sim \mathbb{P}_0, X_1 \sim \mathbb{P}_1, \ldots$ all on the same space so that $\lim_{t\to\infty} \mathbb{P}_t = \pi$. In Section 6.3.1 we formalize this and consider properties of the chain. This allows us to obtain sufficient conditions for the desired convergence as shown in Section 6.3.2. Finally, we consider a strong law of large numbers and a central limit theorem for such chains in Section 6.3.3.

However, before doing so we need to establish the spaces that we will be considering. Because we use many results from Meyn & Tweedie (1993), these spaces must have separable σ-algebras. In order to use these results in Section 6.4 we consider two cases; the fixed number case, where we condition on $X(\mathcal{X}) = n$ and the random number case with no conditioning.

For the fixed number case we will consider the space $(E, \mathcal{C})$ given by

$$\begin{aligned} E &= E_n = \{x \in N : x(\mathcal{X}) = n, p(x) > 0\} \\ \mathcal{C} &= \{F \in \mathcal{B}^n : p(x) > 0 \ \forall x \in F\}. \end{aligned} \tag{6.1}$$

Since $\mathcal{X}$ is separable it contains a dense set $\{v_1, v_2, \ldots\}$ and $\mathcal{B} = \sigma(b(v_i, \frac{1}{j}) : (i,j) \in \mathbb{N}^2)$. Therefore $\mathcal{B}$ is separable making $\mathcal{B}^n$ and also $\mathcal{C}$ separable.

For the random number case $(E, \mathcal{C})$ is given by

$$\begin{aligned} E &= \bigcup_{n=0}^{\infty} E_n = \{x \in N : p(x) > 0\} \\ \mathcal{C} &= \{F \in \mathcal{N} : p(x) > 0 \ \forall x \in F\} \end{aligned} \tag{6.2}$$

where $\mathcal{C}$ is separable by Corollary 2.11.

6.3.1 Properties of Markov Chains

Define a **Markov kernel** $P(\cdot,\cdot)$ as a function satisfying

$$\begin{aligned} P(x,\cdot) &: \text{ Is a probability measure on } \mathcal{C} \text{ for all } x \in E \\ P(\cdot,F) &: \ E \to [0,1] \text{ is measurable for all } F \in \mathcal{C}. \end{aligned}$$

Definition 6.2 (Time-homogeneous Markov Chain)
For a given Markov kernel P and a distribution ρ on $(E,\mathcal{C})$ define a time-homogeneous Markov chain $\{X_0, X_1, \dots\}$ by its distribution $\mathcal{D}$ on $(\prod_0^\infty E, \prod_0^\infty \mathcal{C})$ given by the following conditional finite dimensional distributions

$$\begin{aligned} \mathcal{D}(X_0) &= \rho \\ \mathcal{D}(X_t \mid X_0 = x_0, \dots, X_{t-1} = x_{t-1}) &= P(x_{t-1}, \cdot) \end{aligned}$$

for all $t \in \mathbb{N}$.

We notice, that by Kolmogorov's existence theorem, cf. Billingsley (1995), the distribution $\mathcal{D}$ exists. See Meyn & Tweedie (1993) for details.

The chain is Markov because it only depends on the past through the immediate past, and time-homogeneous because P does not depend on t. Letting $\mathbb{P}_t = \mathcal{D}(X_t)$ we get that

$$\mathbb{P}_t(F) = \int_E P(x,F)\mathrm{d}\mathbb{P}_{t-1}(x) \tag{6.3}$$

for $F \in \mathcal{C}$.

Further, we let the **t-step transition probability** be given by

$$\begin{aligned} P^t(x,F) &= \mathbb{P}_t(F \mid X_0 = x) \\ &= \int_E P(y,F)P^{t-1}(x,\mathrm{d}y), \end{aligned} \tag{6.4}$$

taking $P^1 = P$.

Definition 6.3 (Equilibrium Distribution)
Let a Markov chain be generated by a Markov transition kernel P. Then a probability distribution $\mathbb{P}$ on $(E,\mathcal{C})$ is called the ***equilibrium distribution*** *if*

$$\lim_{t\to\infty} P^t(x,F) = \mathbb{P}(F) \quad \text{for } \mathbb{P}\text{-a.a. } x \in E,\ \forall\, F \in \mathcal{C}.$$

Our aim now is to obtain sufficient conditions on P to ensure that our target distribution π is the equilibrium distribution of the Markov chain generated by

P. It is important that these conditions are fairly easy to verify and that they suggest how to construct P.

First we introduce a necessary condition.

Definition 6.4 (Invariant Distribution)
*A probability distribution $\mathbb{P}$ is called an **invariant distribution** of a Markov chain generated by P if*

$$\mathbb{P}(F) \quad = \quad \int_E P(x,F)\mathrm{d}\mathbb{P}(x) \quad \forall\, F \in \mathcal{C}. \tag{6.5}$$

Proposition 6.5
If $\mathbb{P}$ is the equilibrium distribution of a Markov chain then $\mathbb{P}$ is an invariant distribution of the chain.

Proof:
By Definition 6.3 and (6.4) we see that

$$\begin{aligned}
\mathbb{P}(F) &= \lim_{t\to\infty} P^t(x,F) \\
&= \lim_{t\to\infty} \int_E P(y,F)P^{t-1}(x,\mathrm{d}y) \\
&= \lim_{t\to\infty} \int_E P(y,F)P^{t}(x,\mathrm{d}y) \\
&= \int_E P(y,F)\mathrm{d}\mathbb{P}(y)
\end{aligned}$$

hence yielding (6.5). ∘

By (6.3) $\mathbb{P}_t = \pi \Rightarrow \mathbb{P}_{t+1} = \pi$, so if the chain has reached equilibrium ($\mathbb{P}_t = \pi$) then the chain stays in equilibrium.

The following definition considers a sufficient condition for being invariant. It has the advantage that it is often easier to construct a transition kernel with this condition in mind, hence making the target distribution an invariant distribution of the chain.

Definition 6.6 (Reversible Markov Chain)
A Markov chain is reversible w.r.t. a probability measure $\mathbb{P}$ on $(E,\mathcal{C})$ if

$$\int_B P(x,A)\mathrm{d}\mathbb{P}(x) \quad = \quad \int_A P(x,B)\mathrm{d}\mathbb{P}(x) \tag{6.6}$$

for all $A, B \in \mathcal{C}$.

We observe that by taking $B = E$, (6.6) reduces to (6.5). Hence, if a Markov chain is reversible w.r.t. $\mathbb{P}$ then $\mathbb{P}$ is an invariant distribution of the chain and thus by Proposition 6.5 it may be the equilibrium distribution of the chain.

So, by constructing a Markov chain which is reversible w.r.t. the target distribution, the chain might converge as desired. In the next section we consider what further conditions to impose in order to get convergence of the chain.

6.3.2 Convergence Results for Markov Chains

Two concepts, irreducibility and aperiodicity, play a key role in investigating the convergence of a Markov chain. Irreducibility ensures that all parts of the state space can be reached with the Markov chain, no matter which state we started the chain in, and aperiodicity ensures that the Markov chain has no cyclic behavior. We refer to Meyn & Tweedie (1993) for proofs of the results given here.

Definition 6.7 (ϕ-irreducible)
A Markov chain is ϕ-irreducible if ϕ is a measure on $\mathcal{C}$ with $\phi(E) > 0$ and

$$\exists t \in \mathbb{N} : P^t(x, F) > 0,$$

for all $x \in E$ and all $F \in \mathcal{C}$ with $\phi(F) > 0$.

Proposition 6.8
Let π be an invariant distribution of a Markov chain. If the Markov chain is ϕ-irreducible for some measure ϕ then it is also π-irreducible and π is the only invariant distribution of the chain.

This proposition is a useful tool when checking whether a Markov chain is π-irreducible or not, since the proposition expresses that it suffices to check if a very simple measure is irreducible on the state space, for example $\phi(A) = \mathbb{1}[\emptyset \in A]$. Further, if the chain has π as invariant distribution and is irreducible then π is the only possible equilibrium distribution of the chain by Proposition 6.5.

Definition 6.9 (Periodicity)
If there exists an integer $q > 1$ and non-empty disjoint sets $F_0, \ldots, F_{q-1} \in \mathcal{C}$ such that for $i = 0, \ldots, q-1$ and for all $x \in F_i$

$$P(x, F_j) \quad = \quad 1, \qquad \text{for } j = i + 1 \mod q,$$

then the Markov chain with Markov transition kernel P is called periodic.

If a Markov chain is not periodic, it is called **aperiodic**.

In order to present the first convergence result we have to choose an appropriate norm for the convergence.

Definition 6.10 (Total Variation Norm)
Let ν be a signed measure on the measure space $(E, \mathcal{C})$. Then the norm defined as

$$\|\nu\|_v = \sup_{F \in \mathcal{C}} \{\nu(F)\} - \inf_{F \in \mathcal{C}} \{\nu(F)\}$$

is called the total variation norm.

Proposition 6.11
A Markov chain with invariant distribution π is π-irreducible and aperiodic if and only if

$$\|P^t(x, \cdot) - \pi(\cdot)\|_v \to 0$$

for $t \to \infty$ and π-a.a. $x \in E$.

So, if we construct a Markov chain which is reversible w.r.t. π, irreducible and aperiodic, then the chain converges for π-almost all x. To ensure convergence for all x we need the following concept which is closely related to irreducibility.

Definition 6.12 (Harris Recurrence)
A Markov chain with invariant distribution π is Harris recurrent if for all $x \in E$ and for all $F \in \mathcal{C}$ with $\pi(F) > 0$

$$\mathbb{D}(\exists t : X_t \in F \mid X_0 = x) = 1, \tag{6.7}$$

where $\mathbb{D}$ denotes the probability under $\mathcal{D}$.

Definition 6.13 (Ergodic Markov Chain)
A Harris recurrent and aperiodic Markov chain is called ergodic.

Further, any element $y \in E$ is called an **ergodic atom** if for all $x \in E$ we have that $\mathbb{D}(\exists t : X_t = y \mid X_0 = x) = 1$.

Proposition 6.14
A Markov chain with invariant distribution π is ergodic if and only if for all $x \in E$

$$\|P^t(x, \cdot) - \pi(\cdot)\|_v \to 0$$

for $t \to \infty$.

Hence, an ergodic Markov chain converges for all x.

6.3.3 Limit Theorems for Markov Chains

The **strong law of large numbers** (SLLN) and **central limit theorem** (CLT) are well-known limit theorems from probability theory and statistics, and in this section these limit theorems are adapted to the framework of Markov chains. In order to formulate the CLT, issues relating to the rate of convergence of the Markov chain are treated. For the SLLN no assumption of rate of convergence (or aperiodicity for that matter) is required. Unless otherwise noted proofs of the theorems given here can be found in Meyn & Tweedie (1993).

Assuming that $\{X_t\}$ is a Markov chain with invariant distribution π, the expected value of a real valued function g w.r.t. π, $\mathbb{E}_\pi g$, is naturally approximated by the sample geometric average $\bar{g}_n$ which is given by

$$\bar{g}_n = \frac{1}{n}\sum_{t=s}^{n+s} g(X_t), \tag{6.8}$$

where s typically (but not necessarily) is chosen large enough so that the Markov chain can be deemed to be in equilibrium. The strong law of large numbers then expresses under which conditions $\bar{g}_n$ converges to the true value of $\mathbb{E}_\pi g$.

Theorem 6.15 (Strong Law of Large Numbers)
Let $\{X_t\}$ be a Harris recurrent Markov chain with invariant distribution π. Then for $g : E \to \mathbb{R}$ measurable where $\mathbb{E}_\pi g$ exists we have that

$$\mathbb{D}\left(\lim_{n\to\infty} \bar{g}_n = \mathbb{E}_\pi g\right) = 1$$

for any initial distribution ρ.

Even if the chain is known to converge, the rate of convergence may be too slow for practical purposes. Geometric ergodicity is a stronger form of ergodicity, and essentially targets the problem of the rate of convergence.

Definition 6.16 (Geometric Ergodicity)
A Markov chain $\{X_t\}$ with invariant distribution π is geometrically ergodic if it is Harris recurrent and there exists $r \in [0,1)$ such that for all $x \in E$ and $t \in \mathbb{N}$

$$\|P^t(x,\cdot) - \pi(\cdot)\|_v \leq M(x)r^t$$

for some constant r and some function $M(x) < \infty$.

If we take $M(x) \equiv M$ to be constant, the Markov chain is **uniformly ergodic**. Møller (1999) considers the rate of convergence of a specific Strauss point process but finds that the bounds obtained are too weak for practical purposes.

So, despite the fact that geometrical ergodicity seems appealing, it may be of little practical use concerning the rate of convergence for point processes. It is, however, a prerequisite for the CLT presented later.

The notion of a small set is convenient for the verification of geometric or uniform ergodicity.

Definition 6.17 (Small Set)
A set $C \in \mathcal{C}$ is small if there exists constants $\varepsilon > 0, t \in \mathbb{N}$ and a probability measure ζ on $(E, \mathcal{C})$ so that

$$P^t(x, \cdot) \quad \geq \quad \varepsilon\zeta(\cdot)$$

for all $x \in C$.

Without loss of generality we may assume that $\varepsilon < 1$ in which case

$$P^t(x, \cdot) \quad = \quad \varepsilon\zeta(\cdot) + (1-\varepsilon)\frac{P^t(x, \cdot) - \varepsilon\zeta(\cdot)}{1-\varepsilon}.$$

Thus $P^t(x, \cdot)$ for $x \in C$ may be seen as a mixture of $\zeta(\cdot)$ and $\tilde{\zeta}(x, \cdot) = \frac{P^t(x,\cdot)-\varepsilon\zeta(\cdot)}{1-c}$, and we may construct P^t by first choosing between the kernels ζ and $\tilde{\zeta}$ with probability ε and $1-\varepsilon$ respectively and then making the transition using the chosen kernel. So, every time the chain is in C there is a probability ε that the future will be completely independent of the past, because the transition is made using ζ making it independent of x. Hence, parts of the chain will be i.i.d.

Geometric ergodicity may be difficult to establish directly, in which case it may be more appropriate to verify an alternative criterion that, if satisfied, implies geometric ergodicity.

Definition 6.18 (Geometric Drift Condition)
A Markov chain satisfies the geometric drift condition if there exists a function $V : E \to [1, \infty)$ and a small set $C \in \mathcal{C}$ and constants $\beta < 1$ and $b < \infty$ such that for all $x \in E$

$$\int_E V(y)P(x, \mathrm{d}y) \quad \leq \quad \beta V(x) + b\,\mathbb{1}[x \in C].$$

Theorem 6.19
If a Markov chain is aperiodic and satisfies the geometric drift condition then it is geometrically ergodic.

Theorem 6.20
A Markov chain is uniformly ergodic if and only if E is small.

The central limit theorem presented below gives the distribution of the estimator in (6.8) for geometrically ergodic Markov chains. A similar version of the CLT exists for uniformly ergodic Markov chains, but since many Markov chains does not satisfy this strict condition we only assume geometric ergodicity.

Theorem 6.21 (Central Limit Theorem)
Let $\{X_t\}$ be a geometrically ergodic Markov chain with invariant distribution π, and let $g : E \to \mathbb{R}$. If $\mathbb{E}_\pi |g|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$ or the chain is reversible w.r.t. π and $\mathbb{E}_\pi g^2 < \infty$ then

$$\sqrt{n}(\bar{g}_n - \mathbb{E}_\pi g) \to N(0, \sigma_g^2) \quad \text{as } n \to \infty,$$

where the asymptotic variance σ_g^2 is given by

$$\sigma_g^2 \quad = \quad \mathrm{Var}_\pi g(X_0) + 2 \sum_{t=1}^{\infty} \mathrm{Cov}_\pi(g(X_0), g(X_t)).$$

For a proof see Chan & Geyer (1994) for the case $\varepsilon > 0$ and Roberts & Rosenthal (1996) for the case of reversibility.

Estimating the Variance

It is naturally interesting to estimate the asymptotic variance σ_g^2 which can be rewritten as

$$\sigma_g^2 \quad = \quad \gamma_0 + 2 \sum_{t=1}^{\infty} \gamma_t,$$

where γ_t is the **lag t auto-covariance** of the stationary time series $g(X_0)$, $g(X_1), \cdots$;

$$\gamma_t \quad = \quad \mathrm{Cov}_\pi(g(X_i), g(X_{t+i})), \quad i \in \mathbb{N}.$$

From standard time series analysis methods (see e.g. Chatfield (1996)) an estimator for γ_t is given by

$$\hat{\gamma}_t \quad = \quad \frac{1}{n} \sum_{i=1}^{n-t} \left(g(X_i) - \bar{g}_n\right)\left(g(X_{t+i}) - \bar{g}_n\right), \quad t = 0, 1, 2, \ldots, n-1, \qquad (6.9)$$

where $\frac{1}{n}$ often is replaced by the more intuitive $\frac{1}{n-t}$ which ensures that $\{\hat{\gamma}_{r,s}\}_{r,s=0,1,\ldots,n-2}$ is positive semi-definite. Note that stationarity implies that $\hat{\gamma}_{r,s} = \hat{\gamma}_{|r-s|}$.

A natural estimator for σ_g^2 is

$$\hat{\sigma}_g^2 = \hat{\gamma}_0 + 2\sum_{t=1}^{n} \hat{\gamma}_t. \tag{6.10}$$

However, this estimator is inconsistent which leads to the consideration of two alternative estimation procedures: **window estimation** and **batching**. We refer to Ripley (1988) regarding the latter method.

In window estimation the idea is to down-weigh the large-lag terms in the estimate for the empirical auto-covariance in (6.9), that is

$$\hat{\sigma}_g^2 = w_0\hat{\gamma}_0 + 2\sum_{t=1}^{n-1} w_t\hat{\gamma}_t,$$

where $w_t \in [0,1]$ are the weights, which are said to comprise a **lag window**, hence the name 'window estimation'. If, for instance, we want to cut-off (ultimate down-weighting) auto-covariance at a certain lag K, the weights simply become

$$w_t = \mathbb{1}[t \leq K], \quad 0 < K < n-1.$$

Geyer (1992) shows that if the Markov chain is ϕ-irreducible, reversible and in equilibrium, then the series

$$\Gamma_s = \gamma_{2s} + \gamma_{2s+1}, \quad s = 0, 1, 2, \ldots$$

is a strictly positive, strictly decreasing and strictly convex function of s. Let the estimate of Γ_s be $\hat{\Gamma}_s = \hat{\gamma}_{2s} + \hat{\gamma}_{2s+1}$ and set

$$\begin{aligned} K_{pos} &= \max\{t \in \mathbb{N} : \hat{\Gamma}_s > 0 \ , \forall\, t\}, \\ K_{mon} &= \max\{t \leq K_{pos} : \hat{\Gamma}_s < \hat{\Gamma}_{s-1}, \forall s \leq t\}, \\ K_{con} &= \max\{t \leq K_{mon} : \hat{\Gamma}_0, \ldots, \hat{\Gamma}_t \text{ is convex}\}. \end{aligned}$$

Using $w_t = \mathbb{1}[t \leq 2K+1]$ with K being K_{pos}, K_{mon} or K_{con} gives the initial positive, monotone and convex sequence estimators $\hat{\sigma}^2_{pos}$, $\hat{\sigma}^2_{mon}$ and $\hat{\sigma}^2_{con}$ respectively. According to Geyer (1992, Theorem 3.2) we have that for any of the three initial sequential estimators mentioned above ($\hat{\sigma}^2_{seq}$)

$$\liminf_{n\to\infty} \hat{\sigma}^2_{seq} \geq \hat{\sigma}^2_g.$$

Further, we have by definition that $\hat{\sigma}^2_{con} \leq \hat{\sigma}^2_{mon} \leq \hat{\sigma}^2_{pos}$. Hence, of the three estimators, $\hat{\sigma}^2_{con}$ is the smallest consistent overestimate of σ_g^2.

Example 6.22
As an example we simulate a Strauss process with parameters $\beta = 150, \gamma = 0.5$, and $R = 0.1$ w.r.t. a unit rate Poisson point process on a 3 times 3 area. We

make 100.000 *iterations and throw away the first* 10.000 *as burn-in. Using the remaining* 90.000 *samples we find that the average number of events within the center unit square is* 56.4 *with an estimated variance of* 1.4 *giving a* 95% *confidence interval of* $54.0 - 58.8$. *Similarly we find that the average number of neighbours has a variance of* 1.05 *giving a* 95% *confidence interval of* $27.7 - 31.7$. *For these two estimates* K_{con} *is* 3 *and* 1 *respectively.*

6.4 Metropolis–Hastings Algorithms

Several algorithms exist for constructing Markov chains with a given equilibrium distribution. Here we describe a general class of algorithms known as **Metropolis–Hastings algorithms** which have their origin in Metropolis et al. (1953). The Metropolis–Hastings algorithm described here was introduced by Hastings (1970), and as usual we restrict attention to finite point processes.

The core of this section describes how to construct a Metropolis–Hastings algorithm that is reversible with respect to a target distribution π. Under further conditions the constructed Markov chain is geometrically ergodic, and hence we can apply the SLLN and CLT (see Section 6.3.3). We treat both the fixed- and the random number case, and start with the former since this will ease the introduction to the random number case. From a historical point of view, this was also how the Metropolis–Hastings algorithm was presented first, because it at that time was too computationally expensive to simulate from the random number case.

Let X be a finite point process with density

$$p(x) \quad \propto \quad k(x(\mathcal{X}))g(x), \quad x \in N, \tag{6.11}$$

w.r.t. Π, where $k(\cdot) \geq 0$ and $g(\cdot)$ are known measurable functions. Suppose that $g(\cdot)$ in some sense describes the interaction structure of the process, e.g. if X is a Strauss process (see Example 5.18) then $k(x(\mathcal{X})) = \beta^{x(\mathcal{X})}$ and $g(x) = \gamma^{S(x)}$.

6.4.1 Fixed Number Case

This section focus on the case where we condition on $X(\mathcal{X}) \equiv n$, i.e. the number of events in the process X is kept fixed. In this conditional case (6.11) reduces to

$$p_n(x) \propto g(x) \tag{6.12}$$

w.r.t. μ^n. Assume subsequently that we wish to construct a Markov chain $\{X_t\}$ with state space $(E_n, \mathcal{C})$ as in (6.1) and p_n as equilibrium density, i.e. we construct a Markov chain with equilibrium distribution π.

The Metropolis–Hastings algorithm essentially consists of two steps. First a proposal for a new state is made and secondly the proposal is either accepted or rejected. Here we consider the single point updating Metropolis–Hastings algorithms as surveyed in Møller (1999). If $X_t = x$ is the current state of the Markov chain, then it is proposed to make a transition to $X_{t+1} = (x \cup \xi)\backslash\eta$, where $\eta \in x \cup \xi$ and $\xi \in \mathcal{X}\backslash x$. The proposal is accepted with some probability α which equals zero if $(x \cup \xi)\backslash\eta \notin E$. If the proposal is rejected the next state remains x, that is $X_{t+1} = x$. There are two possible methods in which the above two steps can be undertaken.

(i) First select $\eta \in x$ with some probability (may depend on x), and then generate ξ according to some density (may depend on x and η) w.r.t. μ.

(ii) First an extra point ξ is generated according to some density w.r.t. μ (may depend on x), and then a point $\eta \in x$ is selected with some probability (may depend on x).

Let the proposal density $q(x, \cdot, \cdot)$ be the joint density for (ξ, η) such that

$$\int_{\mathcal{X}} \sum_{\eta \in x} q(x, \eta, \xi) \mathrm{d}\mu(\xi) = 1,$$

for instance it could be given by $q(x, \eta, \xi) \equiv \frac{1}{n\mu(\mathcal{X})}$.

We define a **sub-stochastic kernel** by

$$R(x, F) = \int_{\mathcal{X}} \sum_{\eta \in x} q(x, \eta, \xi)\, \alpha(x, \eta, \xi)\, \mathbb{1}[(x \cup \xi)\backslash\eta \in F]\, \mathrm{d}\mu(\xi),$$

which is used to describe accepted proposals, different from the current state x. Letting $r(x) = 1 - R(x, E)$ be the probability of remaining in the current state, we can now define the Markov transition kernel by

$$P(x, F) = R(x, F) + \mathbb{1}[x \in F] r(x), \quad \forall\, F \in \mathcal{C}. \tag{6.13}$$

Now we establish the following **detailed balance condition** (DBC)

$$\begin{gathered} q(x, \eta, \xi)\, \alpha(x, \eta, \xi)\, g(x) = \\ q((x \cup \xi)\backslash\eta, \xi, \eta)\, \alpha((x \cup \xi)\backslash\eta, \xi, \eta)\, g((x \cup \xi)\backslash\eta), \end{gathered} \tag{6.14}$$

which in fact ensures reversibility of the Markov chain.

Proposition 6.23
The DBC (6.14) *makes the Markov chain reversible.*

Proof:
We need to show that $\int_G P(x,F)\mathrm{d}\pi(x) = \int_F P(y,G)\mathrm{d}\pi(y)$, which by (6.13) is equivalent to showing that $\int_G R(x,F)\mathrm{d}\pi(x) = \int_F R(y,G)\mathrm{d}\pi(y)$. Here we let $y = (x\cup\xi)\backslash x_i$ with x_i being any point in x and $c = \exp(-\mu(\mathcal{X}))/n!$, by which we get

$$
\begin{aligned}
\int_G R(x,F)\,\mathrm{d}\pi(x) &= c\int_G R(x,F)\,p(x)\,\mathrm{d}\mu^n(x)\\
&= c\int_G\int_{\mathcal{X}}\sum_{\eta\in x} q(x,\eta,\xi)\,\alpha(x,\eta,\xi)\,\mathbb{1}[(x\cup\xi)\backslash\eta\in F]\,\mathrm{d}\mu(\xi)\,p(x)\,\mathrm{d}\mu^n(x)\\
&= c\int_{\mathcal{X}}\dots\int_{\mathcal{X}}\sum_{\eta\in x} q(x,\eta,\xi)\alpha(x,\eta,\xi)\mathbb{1}[(x\cup\xi)\backslash\eta\in F, x\in G]p(x)\mathrm{d}\mu(\xi)\mathrm{d}\mu^n(x)\\
&= cn\int_{\mathcal{X}}\dots\int_{\mathcal{X}} q(x,x_i,\xi)\alpha(x,x_i,\xi)\mathbb{1}[(x\cup\xi)\backslash x_i\in F, x\in G]p(x)\mathrm{d}\mu(\xi)\mathrm{d}\mu^n(x)\\
&= cn\int_{\mathcal{X}}\dots\int_{\mathcal{X}} q(y,\xi,x_i)\alpha(y,\xi,x_i)\mathbb{1}[(y\cup x_i)\backslash\xi\in G, y\in F]p(y)\mathrm{d}\mu(x_i)\mathrm{d}\mu^n(y)\\
&= c\int_{\mathcal{X}}\dots\int_{\mathcal{X}}\sum_{\kappa\in y} q(y,\kappa,x_i)\alpha(y,\kappa,x_i)\mathbb{1}[(y\cup x_i)\backslash\kappa\in G, y\in F]p(y)\mathrm{d}\mu(x_i)\mathrm{d}\mu^n(y)\\
&= c\int_{\mathcal{X}}\dots\int_{\mathcal{X}} R(y,G)\mathbb{1}[y\in F]p(y)\mathrm{d}\mu^n(y)\\
&= \int_F R(y,G)\mathrm{d}\pi(y),
\end{aligned}
$$

and hence

$$\int_G P(x,F)\mathrm{d}\pi(x) = \int_F P(y,G)\mathrm{d}\pi(y)$$

whereby the Markov chain is reversible. ∘

Rewriting (6.14) we see that

$$\alpha(x,\eta,\xi) = h(x,\eta,\xi)\alpha((x\cup\xi)\backslash\eta,\xi,\eta),$$

where $h(x,\cdot,\cdot)$ is called the **Metropolis–Hastings ratio** and is given by

$$h(x,\eta,\xi) = \frac{q((x\cup\xi)\backslash\eta,\xi,\eta)g((x\cup\xi)\backslash\eta)}{q(x,\eta,\xi)g(x)} \tag{6.15}$$

for $q(x, \eta, \xi) > 0, x \in E$ and $(x \cup \xi) \backslash \eta \in E$.

In order to make the acceptance probability maximal (see Peskun (1973)) we let

$$\alpha(x, \eta, \xi) = \min\{1, h(x, \eta, \xi)\}. \tag{6.16}$$

We can summarize the above to yield a rough description of the Metropolis–Hastings algorithm using method (ii):

Algorithm [Metropolis–Hastings — Fixed Number Case]

1. Set $t = 0$ and let x_0 be an arbitrary state in E_n.
2. Generate a proposal $(x_t \cup \xi) \backslash \eta$ from the proposal density $q(x_t, \eta, \xi)$.
3. Sample a value u from the uniform distribution on $(0, 1)$.
4. If $u \leq \alpha(x_t, \eta, \xi)$ then set $x_{t+1} = (x \cup \xi) \backslash \eta$ else set $x_{t+1} = x_t$.
5. Set $t = t + 1$ and repeat from step 2.

6.4.2 The Gibbs Sampler

The fixed number case of the Metropolis–Hastings algorithm is related to the widely used **Gibbs sampler** as named by Geman & Geman (1984), which originates in statistical physics where it is known as the **heat bath algorithm**.

Assume that N is a product space, say $N = \mathcal{X}_1 \times \cdots \times \mathcal{X}_r$. With the Gibbs sampler, the idea is to update one of these r components at a time, selecting the component to update either deterministically or randomly such that all components are eventually updated.

For the ith component, a proposal density q_i for going from state x to state y given by

$$q_i(x, y) = \begin{cases} p(y_i \mid x_{-i}) & \text{if } x_{-i} = y_{-i} \\ 0 & \text{otherwise} \end{cases} \tag{6.17}$$

is used, where x_{-i} denotes x except its ith component and p is the density of the target distribution π. As opposed to the Metropolis Hastings algorithm the proposal is always accepted.

Since (6.17) makes the following detailed balance condition

$$p(x)q_i(x, y) = p(y)q_i(y, x)$$

satisfied, updating a single component is reversible. Hence, updating the components in any order will result in π being an invariant distribution but not necessarily in the chain being reversible.

The Gibbs sampler relates to the fixed case Metropolis Hastings algorithm in the following way.

First, $N = \mathcal{X}^n$ is a product space so the Gibbs sampler may be used. Now select which component (event) i to update giving each of the n components a probability of $\frac{1}{n}$. We then use (6.17) to get the proposal density for the ith component

$$\begin{aligned} q_i(x, x_{-i} \cup \xi) &= p(\xi \mid x_{-i}) \\ &= \frac{g(x_{-i} \cup \xi)}{\int_{\mathcal{X}} g(x_{-i} \cup \kappa) \mathrm{d}\mu(\kappa)}. \end{aligned}$$

This gives a joint proposal density of

$$q(x, \eta, \xi) = \frac{g(x \backslash \eta \cup \xi)}{n \int_{\mathcal{X}} g(x \backslash \eta \cup \kappa) \mathrm{d}\mu(\kappa)}$$

resulting in an acceptance probability of

$$\alpha(x, \eta, \xi) \equiv 1.$$

So with this particular choice of q the Metropolis Hastings algorithm is a Gibbs sampler. We shall use the Gibbs sampler in Example 6.34 for sampling from the Widom-Rowlinson model.

6.4.3 The Random Number Case

We now consider the case where the number of events in the process X is allowed to be random, which causes the state space to be of varying dimension; $E = \bigcup_{n=0}^{\infty} E_n$ with E_n and $\mathcal{C}$ as in (6.2). Also let $\mathcal{C}_n$ denote $\mathcal{C}$ restricted to E_n. In order to accommodate this, the fixed number case Metropolis Hastings algorithm is modified as described by Geyer & Møller (1994).

The proposal density $q(x, \cdot, \cdot)$ of $P(x, \cdot)$ is now constructed such that with probability $Q(x)$ we propose to add a new point from a density $b(x, \xi)$ with respect to μ. Then with probability $1 - Q(x)$ we propose to delete a random point $\eta \in x$, which is chosen with probability $d(x, \eta)$; if there are no points in the current configuration, we do nothing. The proposal is accepted with acceptance probability $\alpha(x, y)$, where x is the current state, and y is the proposed state.

Assume subsequently for $x \in E_n$, $\xi \in \mathcal{X}$, and $x \cup \xi \in E_{n+1}$ that

$$Q(x) > 0, \quad Q(x \cup \xi) < 1, \quad d(x \cup \xi, \eta) > 0, \text{ and } b(x, \xi) > 0,$$

which are necessary conditions in order to make the Markov chain irreducible.

For $F_m \in \mathcal{C}_m$ and $x \in E_n$ we have $P(x, F_m) = 0$ if $|m - n| > 1$. Otherwise we let

$$P(x, G_{n+1}) \quad = \quad Q(x) \int\limits_{\mathcal{X}} \mathbb{1}[x \cup \xi \in G_{n+1}] b(x, \xi) \alpha(x, x \cup \xi) \mathrm{d}\mu(\xi), \tag{6.18}$$

and

$$\begin{aligned} P(x \cup \xi, F_n) \quad = \quad & (1 - Q(x \cup \xi))\Big\{ \sum_{\eta \in x \cup \xi} \mathbb{1}[(x \cup \xi) \backslash \eta \in F_n] \\ & \times d(x \cup \xi, \eta) \alpha(x \cup \xi, (x \cup \xi) \backslash \eta)\Big\}. \end{aligned} \tag{6.19}$$

The following **detailed balance condition** ensures reversibility: for any $x \in E$ and $\xi \in \mathcal{X}$,

$$\begin{gathered} Q(x)\, b(x, \xi)\, \alpha(x, x \cup \xi)\, p(x) = \\ (1 - Q(x \cup \xi))\, d(x \cup \xi, \xi)\, \alpha(x \cup \xi, x)\, p(x \cup \xi). \end{gathered} \tag{6.20}$$

Proposition 6.24
The DBC (6.20) *makes the Markov chain reversible.*

Proof:
Let $x \in F_n$ and $x \cup \xi \in G_{n+1}$, where $F_n \in \mathcal{C}_n$ and $G_{n+1} \in \mathcal{C}_{n+1}$.

Recall that reversibility w.r.t. π implies that

$$\int\limits_{G_{n+1}} P(x \cup \xi, F_n) \mathrm{d}\pi(x \cup \xi) \quad = \quad \int\limits_{F_n} P(x, G_{n+1}) \mathrm{d}\pi(x). \tag{6.21}$$

By (6.18) we have

$$\begin{aligned} & \int\limits_{F_n} P(x, G_{n+1}) \mathrm{d}\pi(x) \\ = \quad & \int\limits_{F_n} P(x, G_{n+1}) p(x) \mathrm{d}\Pi(x) \\ = \quad & c \int\limits_{F_n} \int\limits_{\mathcal{X}} Q(x)\, b(x, \xi)\, \alpha(x, x \cup \xi)\, p(x) \\ & \qquad \times \mathbb{1}[x \in F_n, x \cup \xi \in G_{n+1}] \mathrm{d}\mu(\xi) \mathrm{d}\mu^n(x) \\ = \quad & c \int\limits_{\mathcal{X}} \cdots \int\limits_{\mathcal{X}} (1 - Q(x \cup \xi))\, d(x \cup \xi, \xi)\, \alpha(x \cup \xi, x)\, p(x \cup \xi) \\ & \qquad \times \mathbb{1}[x \in F_n, x \cup \xi \in G_{n+1}] \mathrm{d}\mu(\xi) \mathrm{d}\mu^n(x) \end{aligned} \tag{6.22}$$

where $c = \frac{e^{-\mu(\mathcal{X})}}{n!}$. On the other hand we have

$$\begin{aligned}
&\int_{G_{n+1}} P(x\cup\xi, F_n)\mathrm{d}\pi(x\cup\xi)\\
&= \int_{G_{n+1}} P(x\cup\xi, F_n)p(x\cup\xi)\mathrm{d}\Pi(x\cup\xi)\\
&= \frac{c}{(n+1)}\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}(1-Q(x\cup\xi))\sum_{\eta\in x\cup\xi} d(x\cup\xi,\eta)\alpha(x\cup\xi,(x\cup\xi)\backslash\eta)\,p(x\cup\xi)\\
&\qquad\times\mathbb{1}[x\cup\xi\in G_{n+1},(x\cup\xi)\backslash\eta\in F_n]\mathrm{d}\mu(\xi)\mathrm{d}\mu^n(x), \qquad (6.23)
\end{aligned}$$

where for $n = 0$ we substitute $\mathrm{d}\mu^n(x)$ by $\mathbb{1}[x = \emptyset]$. Splitting (6.23) into $n + 1$ terms, one for each $\eta \in x \cup \xi$, and integrating these terms w.r.t. $\mu^{n+1}(x \cup \xi)$, we see that the sum without loss of generality can be replaced by $(n+1)\mathbb{1}[x \in F_n]d(x\cup\xi,\xi)\alpha(x\cup\xi, x)$, and thus (6.23) = (6.22), proving the proposition. ○

Since all terms in (6.20) by assumption are greater than 0, the **Metropolis–Hastings ratio** for adding a point is given by

$$h(x,\xi) = \frac{p(x\cup\xi)}{p(x)}\frac{(1-Q(x\cup\xi))}{Q(x)}\frac{d(x\cup\xi,\xi)}{b(x,\xi)}.$$

The detailed balance condition (6.20) is hence fulfilled if

$$\begin{aligned}
\alpha(x\cup\xi, x) &= \min\{1, 1/h(x,\xi)\}\\
\alpha(x, x\cup\xi) &= \min\{1, h(x,\xi)\}.
\end{aligned}$$

There are many ways to choose $b(\cdot,\cdot)$, $d(\cdot,\cdot)$ and $Q(\cdot)$, but here we consider the simplest situation, where

$$Q(\cdot)\equiv\frac{1}{2},\quad b(\cdot,\cdot)\equiv\frac{1}{\mu(\mathcal{X})},\quad d(x\cup\xi,\cdot)\equiv\frac{1}{n+1}\quad\text{if } x\cup\xi\in E_{n+1},$$

and hence

$$h(x,\xi) = \lambda(x;\xi)\frac{\mu(\mathcal{X})}{n+1}\quad\text{if } x\cup\xi\in E_{n+1},$$

where $\lambda(x;\xi)$ is the Papangelou conditional intensity. With these choices of $b(\cdot,\cdot), d(x \cup \xi,\cdot)$ and $Q(\cdot)$ we prove that the Metropolis-Hastings algorithm under certain restrains on the Markov transition kernel satisfies the conditions for convergence derived in Section 6.3 (see Geyer (1999), which extends Geyer & Møller (1994)). One of these is the local stability property (see Section 3.3), which implies the hereditary property since

$$p(x\cup\xi)\le Kp(x) \qquad (6.24)$$

implies that $p(x \cup \xi) = 0$ whenever $p(x) = 0$ for $x \in E$ and $\xi \in \mathcal{X}$.

Lemma 6.25
If the density p is locally stable then for any $m \geq 0$

$$C = \{x \in E : x(\mathcal{X}) \leq m\}$$

is a small set.

Proof:
Let $x(\mathcal{X}) = n \geq 1$. We can use the local stability property to write the probability of deleting a point $\eta \in x$ as

$$\begin{aligned}
P(x, \text{delete } \eta) &= (1 - Q(x)) \sum_{\eta \in x} d(x,\eta)\alpha(x, x\backslash\eta) \\
&= \frac{1}{2} \sum_{\eta \in x} \frac{1}{n} \min\left\{1, \frac{n}{\lambda(x\backslash\eta;\eta)\mu(\mathcal{X})}\right\} \\
&\geq \frac{1}{2} \frac{1}{K\mu(\mathcal{X})},
\end{aligned}$$

if $1 \leq n \leq m$ and if we choose K large enough. So $P(x, \text{ delete } \eta) \geq k > 0$ for $x \in C$ where k is some constant. Hence for $x \in C$ and $t > m$ we get that $P^t(x, \{\emptyset\}) \geq k^t$. This suggests that we take $\zeta(A) = \mathbb{1}[\emptyset \in A]$, so

$$P^t(x, A) \geq k^t \zeta(A), \quad \forall x \in C, \; A \in \mathcal{C}.$$

If $x(\mathcal{X}) = 0$, then $P^t(\emptyset, A) = (1 - Q(\emptyset))^t \mathbb{1}[\emptyset \in A]$. Hence C is small by Definition 6.17. ◦

Proposition 6.26
A Markov chain with a locally stable density p, is uniformly ergodic if and only if there exists an $m < \infty$ such that $E = \cup_{n=0}^{m} E_n$.

Proof:
Assume that $E = \cup_{n=0}^{m} E_n$, then by Lemma 6.25 and Theorem 6.20, the chain is uniform ergodic. Assume now that $\{X_t\}$ is a uniformly ergodic Markov chain. By Theorem 6.20 E is small and hence

$$\exists t : P^t(x, F) > 0 \quad \forall F \in \mathcal{C}.$$

Let $x \in E$ such that $s = x(\mathcal{X}) < \infty$ and hence $\{x\} \in \mathcal{C}$. Suppose that there does not exist an $m \in \mathbb{N}$ so that $E = \cup_{n=0}^{m} E_n$, hence there exists a $x' \in E : x'(\mathcal{X}) > t + s$. Then clearly $P^t(x', \{x\}) = 0$ which is a contradiction. ◦

Theorem 6.27
If the density p is locally stable, then the Metropolis-Hastings algorithm is aperiodic and geometrically ergodic.

Proof:
Since p is locally stable, $\emptyset \in E$, and since $P(\emptyset, \{\emptyset\}) \geq \frac{1}{2} > 0$, the chain is aperiodic. To prove the geometric ergodicity property it is sufficient to show the geometric drift condition by Theorem 6.19. The local stability property gives $\lambda(\cdot, \cdot) \leq K$. Let $V(x) = l^n$ where $n = x(\mathcal{X})$, $l > 1$ are constants. Then we can write,

$$\int_E V(y)P(x, \mathrm{d}y) = l^{n+1}P(x, \text{adding some } \xi) + l^{n-1}P(x, \text{deleting some } \eta) \\ + l^n P(x, \text{retaining } x). \quad (6.25)$$

Each of the three terms on the right hand side are now evaluated.

$$\begin{aligned} P(x, \text{adding } \xi) &= Q(x) \int_{\mathcal{X}} \mathbb{1}[x \cup \xi \in E_{n+1}]\, b(x, \xi)\, \alpha(x, x \cup \xi)\, \mathrm{d}\mu(\xi) \\ &= \frac{1}{2} \int_{\mathcal{X}} \mathbb{1}[x \cup \xi \in E_{n+1}] \min\left\{1, \lambda(x; \xi) \frac{\mu(\mathcal{X})}{n+1}\right\} \frac{\mathrm{d}\mu(\xi)}{\mu(\mathcal{X})} \\ &\leq \frac{1}{2} \min\left\{1, \frac{K\mu(\mathcal{X})}{n+1}\right\} \\ &\leq \frac{\varepsilon}{2}, \end{aligned}$$

if $\varepsilon \geq \frac{K\mu(\mathcal{X})}{n+1}$.

$$\begin{aligned} P(x, \text{deleting } \eta) &= (1 - Q(x)) \sum_{\eta \in x} d(x, \eta) \alpha(x, x \backslash \eta) \\ &= \frac{1}{2} \sum_{\eta \in x} \frac{1}{n+1} \min\left\{1, \frac{1}{\lambda(x \backslash \eta; \eta)} \frac{n+1}{\mu(\mathcal{X})}\right\} \\ &= \frac{1}{2} \sum_{\eta \in x} \min\left\{\frac{1}{n+1}, \frac{1}{K\mu(\mathcal{X})}\right\} \\ &= \frac{1}{2}, \end{aligned}$$

if $n + 1 \geq K\mu(\mathcal{X})$. So we get the retaining probability

$$\begin{aligned} P(x, \text{retaining } x) &= 1 - P(x, \text{adding } \xi) - P(x, \text{deleting } \eta) \\ &\leq 1 - P(x, \text{deleting } \eta) = \frac{1}{2} \end{aligned}$$

for $n+1 \geq K\mu(\mathcal{X})$. Hence if we define a real constant $N_{K,\varepsilon}$ satisfying $N_{K,\varepsilon} > \max\left\{\frac{1}{\varepsilon}K\mu(\mathcal{X}), K\mu(\mathcal{X})\right\}$, we obtain for $n+1 > N_{K,\varepsilon}$ that (6.25) can be rewritten

$$\begin{aligned}\int_E V(y)P(x,\mathrm{d}y) &\leq l^{n+1}\varepsilon\frac{1}{2} + l^{n-1}\frac{1}{2} + l^n\frac{1}{2} \\ &= l^n\left(l\varepsilon\frac{1}{2} + \frac{1}{2l} + \frac{1}{2}\right).\end{aligned}$$

Letting $\varepsilon \to 0$ this converges to $l^n(\frac{1}{2l} + \frac{1}{2})$ where $(\frac{1}{2l} + \frac{1}{2})$ is less than 1. This means that for ε small enough there exists a constant $\beta < 1$ so that

$$\begin{aligned}\int_E V(y)P(x,\mathrm{d}y) &\leq \beta V(x), \text{ if } x \notin C \\ \int_E V(y)P(x,\mathrm{d}y) &\leq l^{N_{K,\varepsilon}+1} = b \text{ if } x \in C,\end{aligned}$$

where $C = \{x \in E : x(\mathcal{X}) < N_{K,\varepsilon}\}$. Using Lemma 6.25 we see that C is a small set which completes the argument. $\circ$

Hence, the Metropolis-Hastings algorithm is geometric ergodic which makes it possible to apply the SLLN and CLT.

6.5 Using MCMC in Practice

In this section we consider some more practical aspects of how to use MCMC for spatial point processes. For further reading we recommend Gilks et al. (1996).

Three main applications of MCMC in the present context are

- Making samples from the target distribution for use for example in model selection and validation (see Section 4.7).
- Finding the mean value of some function g under the target distribution.
- Approximate maximum likelihood estimation for parametric inference.

In the following we treat the first two of these applications, deferring the third to Section 7.3.

6.5.1 MCMC and Model Validation

When describing how to use summary statistics for model validation in Section 4.7 we assumed for simplicity that independent samples were used for making upper and lower envelopes. In this section we first investigate how to make independent or at least nearly independent samples using MCMC. Further, we describe a clever method for making such envelopes without the demand for independence.

Throughout this section we use a Metropolis-Hastings algorithm to simulate a Strauss process with parameters $\beta = 150, \gamma = 0.5$ and $R = 0.1$ w.r.t. a unit rate Poisson point process on a 3 times 3 area to exemplify and compare the different methods.

Making Nearly Independent Samples

The first thing to consider is how long the chain has to be run before is can be considered to have reached equilibrium, called the **burn-in** of the chain. The method normally used to determine the burn-in is to plot various statistics versus the number of iterations — possibly with the Markov chain started in different extreme states. Figure 6.3 shows such plots.

After $5.000 - 10.000$ iterations all plots seem to have "forgotten" their starting state and, notably, to have reached the same distribution. So after a burn-in of 10.000 iterations we assume that the chain effectively has reached equilibrium and thus any configuration after this should follow the Strauss model. However, we are never sure that equilibrium has been reached. Perhaps we have considered too few statistics, too few starting states or too few iterations to determine if a chain has reached equilibrium. These problems may be overcome by the perfect sampling method, which we will discuss in Section 6.7, but usually this comes at a cost in computation time.

The next problem to ponder is how to get a batch of nearly independent samples. A straight forward approach is to run a batch of Markov chains, taking one sample from each. However, because each chain must first burn-in this can be rather time costly. Instead, we might just try using a longer chain, taking a sample for every kth iteration, where k is chosen so that these samples are deemed nearly independent.

For two samples to be nearly independent it seems appropriate that they do not have any events in common, or at least only a few. Again using the Strauss process as an example we discard the burn-in and take the first remaining sample as a reference. We then calculate what percentage of the events in this sample is retained in the following samples of the chain, plotting the result in Figure 6.4. We notice that in this case it takes 5.000 or more iterations

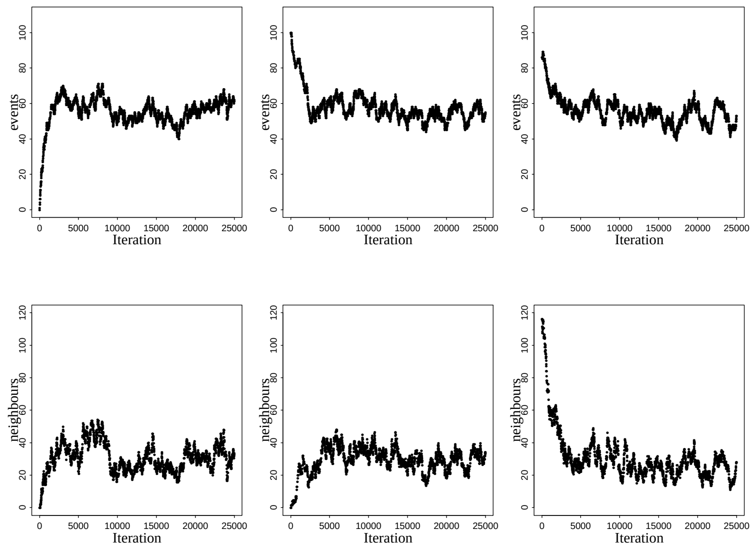

Figure 6.3: Output analysis for three Markov chains simulating a Strauss process. Plots for the three chains are shows from left to right. The chains have different starting states but are all generated by the same Metropolis-Hastings algorithm on a 3×3 area. The plots above show the number of events in the center unit square and the plots below show the number of neighbours in the center unit square.

before the percentage of retained events is near zero, so this approach hold no significant advantage compared to running a batch of Markov chains.

Since we will be using the samples to make summary statistics it may be more relevant to investigate how correlated such statistics are for the samples from the chain. Using (6.9) to estimate auto-correlations we investigate how soon these tend to zero. In Figure 6.5 we have done so for three summary statistics on a chain simulating the Strauss process finding that after $2.000-3.000$ iterations the auto-correlations have almost vanished, hence reducing the time cost by a factor of 3 to 5.

Using Exchangeable Statistics

We will now consider a technique developed by Besag & Clifford (1989) for making envelopes that does not demand the use of independent samples.

Let x_0 denote the given data point pattern and suppose we wish to investigate if x_0 follows a specific distribution by making envelopes for various summary statistics. Assume that we have a Markov chain with this distribution as equilibrium distribution.

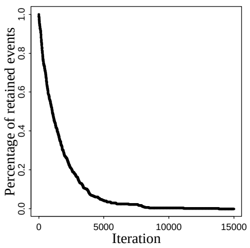

Figure 6.4: Percentage of events in first sample after burn-in that are retained in subsequent samples. Samples are from a Strauss process.

Now, start this Markov chain in x_0 and run it backwards for m iterations. If the chain is reversible this is the same as running the chain forwards. Let x_{start} denote the point configuration reached after the m iterations. Now run the chain forward n times each time for m iterations and each time starting in x_{start}. Let $x_1, \ldots, x_n$ be the point configurations reached with the n different runs, see Figure 6.6.

If x_0 follows the target distribution then $x_0, x_1, \ldots, x_n$ are exchangeable because they are conditional independent given x_{start}, as they all evolve independently in m steps from x_{start}. Hence, we get that for any statistic t that the distribution of $t(x_0), t(x_1), \ldots, t(x_n)$ is exchangeable, so for instance

$$\mathbb{P}\big(t(x_0) > \max_{i=1,\ldots,n} t(x_i)\big) \quad \leq \quad \frac{1}{n+1}$$

where equality is obtained if $t(x_0)$ almost surely is not equal to $t(x_i)$ for any $i \in \{1, \ldots, n\}$.

If x_0 does not follow the target distribution then we have to select m large enough so that the statistics become exchangeable. Again using the Markov chain for a specific Strauss process on a 3×3 area we let x_0 be a realisation of a binomial point process with 600 events on this area, roughly the right number of events for the Strauss process, and let $m = 1000$. Figure 6.7 shows the corresponding summary statistic plots for the center 1×1 area, clearly rejecting the hypothesis that this pattern is from the specific Strauss process.

So in this particular example, we only needed 1.000 iterations per sample compared to $2.000 - 10.000$ samples if we wanted nearly independent samples.

Bias Using the Two Methods for Envelopes

In the following we investigate some properties of the informal test we get by using summary statistics and envelopes either from nearly independent samples

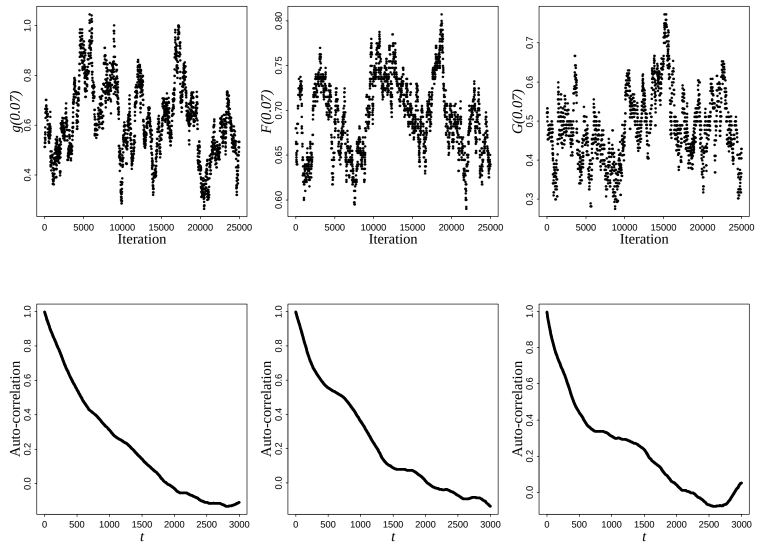

Figure 6.5: Auto-correlations for three summary statistics $g(0.07), F(0.07)$, and $G(0.07)$, based on a Markov chain of length 25.000 after burn-in.

or exchangeable statistics to test if a given point configuration comes from a specified model.

Consider the use of nearly independent samples taken from a chain before it has reached equilibrium, i.e. the burn-in used is too short. If we used a batch of Markov chains started in different states then this would probably result in wider envelopes than under the true distribution. If we used one long chain we might have that the first samples taken are not from the distribution while the last samples are, again giving wider envelopes. These envelopes hence results in a degree of acceptance that is larger than the significance level aimed for.

Now assume that the nearly independent samples used are all from the distribution, but are somewhat dependent. It then seems intuitively likely that this dependence will result in the summary statistics obtained having less wide envelopes than summary statistics from independent samples which naturally leads to a smaller degree of acceptance.

Finally, consider the use of exchangeable statistics. If the data x_0 does in fact follow the distribution then the test has the specified level of acceptance as soon as the number of iterations m is large enough that the values we compare are different, which is likely to occur for even very small values of m. However, if x_0 does not follow the model and we use a small m then the chain may not have moved far enough away from x_0 for us to detect this, leading to a greater

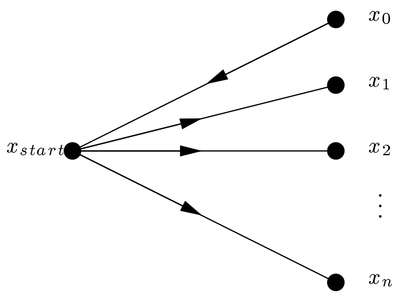

Figure 6.6: Markov chain is run backwards from x_0 to x_{start} and then forwards n times from x_{start} to $x_1, \ldots, x_n$.

degree of acceptance in this case.

The conclusion is that if we use nearly independent samples from one long chain and are confident in the burn-in used then we will be very confident in any acceptance but less confident in a rejecting and *vice versa* using exchangeable statistics. So, an obvious idea is to use both methods, accepting or rejecting if they agree and investigating further if they disagree.

6.5.2 Estimating the Mean of a Function Using MCMC

When estimating the mean of a function, the SLLN (Theorem 6.15) is the natural foundation. However, we need to decide on a finite number of samples that still gives a good approximation. Intuitively, we want the chain to have visited every part of the distribution and that it only contains samples from the distribution. Hence, we discard the burn-in, because we do not consider these first samples to follow the target distribution. Further, it seems advisable to run the chain for at least so long that the last samples are deemed independent of the first sample after the burn-in.

This approach may lead to very long chains which may produce storage problems. One way to reduce this problem is to incorporate the function into the chain and then only store the value of the function as opposed to the entire sample. However, we may want to estimate the mean of several functions using the same chain and these functions may not always be known beforehand. Depending on the implementation used it may also be impractical to incorporate the function(s) into the algorithm that makes the chain or it may even be too time consuming to calculate the function for all samples. A way to solve these problems is through **spacing** where only every kth sample is stored and thus used to estimate the mean.

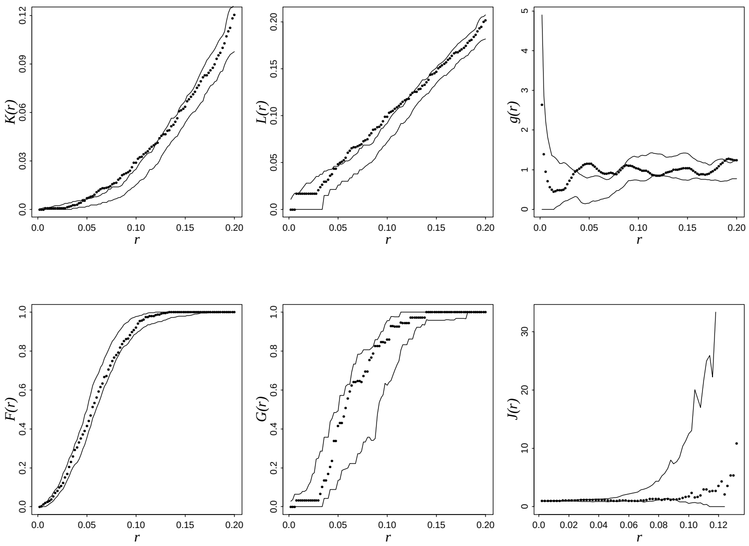

Figure 6.7: Summary statistics for a binomial point pattern using 39 Strauss samples for envelopes.

Importance Sampling

Another problem we may face, is that we wish to find the mean of a function under a distribution which is hard to simulate from. In this case we may sometimes use a technique called **importance sampling**, which not only provide easier sampling but also allows us to use a single chain to do inference about several values of the parameter, as long as these in the vicinity of the value with which the chain is generated. Further, importance sampling can be used as a means to reduce the asymptotic variance (as treated in Ripley (1987)).

Suppose we have obtained a Markov chain sample $x_1, \ldots, x_n$ from a target distribution π_1, but we want to estimate $\mathbb{E}_{\pi_2} g$ for some real function g under π_2. Further assume that π_1 and π_2 both have known densities p_1 and p_2 w.r.t. Π respectively, and that $p_1(x) = 0$ implies $p_2(x) = 0$. Then, taking $\frac{0}{0} = 0$,

$$\begin{aligned}
\mathbb{E}_{\pi_2} g(X) &= \int_N g(x) p_2(x) \mathrm{d}\Pi(x) \\
&= \int_N g(x) \frac{p_2(x)}{p_1(x)} p_1(x) \mathrm{d}\Pi(x) \\
&= \mathbb{E}_{\pi_1} \left[g(X) \frac{p_2(X)}{p_1(X)} \right],
\end{aligned}$$

can be approximated by $\frac{1}{n}\sum_{i=1}^{n} g(x_i)\frac{p_2(x_i)}{p_1(x_i)}$. Now the problem is that we normally do not know p_1 and p_2. We assume in the following that $p_1 = \frac{h_1}{c_1}$ and $p_2 = \frac{h_2}{c_2}$ where h_1 and h_2 are known functions but c_1 and c_2 are normalising constants which might be unknown. Then

$$\begin{aligned} \mathbb{E}_{\pi_2} g(X) &= \mathbb{E}_{\pi_1}\left[g(X)\frac{h_2(X)}{h_1(X)} \Big/ \frac{c_2}{c_1} \right] \\ &= \mathbb{E}_{\pi_1}\left[g(X)\frac{h_2(X)}{h_1(X)} \right] \Big/ \mathbb{E}_{\pi_1}\left[\frac{h_2(X)}{h_1(X)} \right] \end{aligned}$$

can be approximated by the ratio unbiased estimator

$$\left\{ \frac{1}{n}\sum_{i=1}^{n} g(x_i)\frac{h_2(x_i)}{h_1(x_i)} \right\} \Big/ \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{h_2(x_i)}{h_1(x_i)} \right\}. \tag{6.26}$$

Importance sampling only works in practice if p_2 is close enough to p_1, i.e. the importance weights $w(x_i) \propto \frac{h_2(x_i)}{h_1(x_i)}$ are all of the same magnitude. Otherwise, the generated chain which is considered to cover π_1 well may cover π_2 badly.

6.6 Spatial Birth-and-Death Processes

This section provides an alternative to the Metropolis–Hastings based simulation of Section 6.4, introducing the theory of spatial birth-and-death processes (SBD). Birth-and-death processes are used in different areas of statistics, but their use in spatial statistics is due to the close relationship to Gibbs processes (Preston 1977) and to simulation of spatial point patterns (Ripley 1977). In this section we briefly introduce what is meant by a spatial birth-and-death process and then present a construction that allows any locally stable finite point process to be simulated using a SBD, a result which will be used in Section 6.7 for perfect simulation of such point processes. This section is concluded with a short comparison of Metropolis–Hastings based simulation versus simulation using spatial birth-and-death processes. Main references are Preston (1977), Kelly & Ripley (1976), Ripley (1977), Stoyan et al. (1995), Møller (1989), Baddeley & Møller (1989), and Kendall & Møller (1999).

A spatial birth-and-death process on $\mathcal{X}$ is a continuous time Markov chain (as opposed to the discrete Markov chain described in Section 6.3) so that the transition probability of the process at a given time instance only depends on the current point configuration; the Markov property in time. In fact it is a jump process which changes at certain time instants either by the birth of a new point, which is added to the process, or by the death of an existing point, which is deleted.

6.6.1 Formal Treatment of Spatial Birth-and-Death Processes

A spatial birth-and-death process is a continuous time Markov chain $\{X_t : t \geq 0\}$ where a realisation of X_t for a fixed t is a finite point configuration on $\mathcal{X}$. If $X_t = x$ and τ denotes the **waiting time** from time t until a transition happens, i.e.

$$X_s = X_t \text{ for } t \leq s < t + \tau, \ X_{t+\tau} \neq X_t,$$

then there are only two possible types of transitions. One is the **birth** of an event $\xi \in \mathcal{X}$ so $X_{t+\tau} = x \cup \xi$ and the other is the **death** of an event $x_i \in x$ so $X_{t+\tau} = x \backslash x_i$.

Given a target distribution π, the aim is to specify how and when to make births and deaths, so that the distribution of X_t as t tends to infinity is π given $X_0 = x_0$ for all $x_0 \in E$. We assume that π has the density p w.r.t. the intensity measure μ on $\mathcal{X}$ of a Poisson point process Π.

The births and deaths are controlled by a **birth rate** and a **death rate** respectively. Let the birth and death rate be non-negative measurable functions b and d on $N_f \times \mathcal{X}$ and let the **birth area rate** be

$$B(x, A) = \int_A b(x, \xi) \mathrm{d}\mu(\xi) \tag{6.27}$$

for $A \in \mathcal{B}$ and $x \in \mathcal{X}$. Further, the total birth and death rates, $B(x)$ and $D(x)$, are

$$B(x) = B(x, \mathcal{X}) \quad \text{and} \quad D(x) = \mathbb{1}[x \neq \emptyset] \sum_{\xi \in x} d(x \backslash \xi, \xi),$$

and the total transition rate is $\alpha(x) = B(x) + D(x)$. These rates are then used to determined the waiting time τ and further what transition will occur at time $t + \tau$. Let $\mathrm{Exp}(a)$ denote the exponential distribution with mean $\frac{1}{a}$ taking $\mathrm{Exp}(0) \equiv \infty$. Conditional on $X_t = x$, let

$$\begin{aligned} \tau_0 &\sim \mathrm{Exp}(B(x)) \\ \tau_i &\sim \mathrm{Exp}(d(x \backslash \xi, x_i)) \quad \text{for } x_i \in x, \end{aligned}$$

all mutually independent and independent of everything before time t, and let

$$\tau = \min\{\tau_0, \tau_1, \ldots, \tau_{x(\mathcal{X})}\},$$

so that $\tau \sim \mathrm{Exp}(\alpha(x))$. If $\tau = \tau_0$ then a new event is born at time $t + \tau$ from the density $\frac{b(x,\xi)}{B(x)}$ w.r.t. μ and if $\tau = \tau_i$ then x_i dies at time $t + \tau$.

We refer to Preston (1977) for sufficient conditions on b and d that ensure the existence, uniqueness and convergence of such spatial birth-and-death processes, noting that if the detailed balance condition

$$p(x)b(x,\xi) \quad = \quad p(x \cup \xi)d(x,\xi) \; > \; 0 \qquad \text{if } p(x \cup \xi) > 0 \tag{6.28}$$

is fulfilled then the process is time-reversible and p is the unique invariant density. For further details we refer to Møller (1989) - which also states necessary conditions for obtaining a geometrically fast convergence to equilibrium.

With these conditions in mind we can choose the birth and death rates in different ways. If the birth rate is set to be constant, e.g. $b \equiv 1$, then the death rate will be $d(x,\xi) = \frac{p(x)}{p(x \cup \xi)}$. But this means that the process is very likely to delete the point that was just added, implying that we often have to deal with a very slow convergence, see Ripley (1977). Instead, one often choose to let the death rate be constant, e.g. $d(\cdot,\cdot) \equiv 1$, resulting in the birth rate being the Papangelou conditional intensity $\lambda(x;\xi)$. A critical element in this approach was until recently, that it required the total birth rate, i.e. that the integral for $B(x)$ had to be evaluated, making it difficult to implement an efficient algorithm. But, as we shall see in the following section there is a way around this problem.

Finally, if p is locally stable, b is the Papangelou conditional intensity and $d \equiv 1$ then existence, uniqueness, convergence and obviously detailed balance are ensured, see Møller (1999), so henceforth we work with this setup.

6.6.2 Thinning from a Dominating Process

In this section we will show how to construct a spatial birth-and-death process that converges towards any given locally stable density p. To do so we make a **dominating process** $\Delta = \{\Delta_t : t \geq 0\}$ and through thinning from Δ we obtain a process which converges towards the target distribution π.

The dominating process Δ is a spatial birth-and-death process with birth rate $b \equiv C$ and death rate $d \equiv 1$, where $C \geq \sup_{x \cup \xi \in N_f} \lambda(x;\xi)$. Further, Δ_0 is a Poisson point process with intensity measure $C\mu$, so by Example 3.12 it has a density $\tilde{p}(y) = e^{(1-C)\mu(\mathcal{X})}C^{y(\mathcal{X})}$ w.r.t. the Poisson point process with intensity measure μ and hence Δ is time-reversible and in equilibrium by (6.28). The waiting time τ from time t is realised by

$$\begin{aligned} \tau_0 &\sim \operatorname{Exp}(C\mu(\mathcal{X})) \\ \tau_i &\sim \operatorname{Exp}(1), \end{aligned}$$

so $\tau \sim \operatorname{Exp}(C\mu(\mathcal{X}) + y(\mathcal{X}))$ if $\Delta_t = y$.

In order to do thinning from Δ we define a mapping $M : \mathbb{R} \to [0,1]$ so that

$$\begin{array}{rcll} M(t) & \sim & \text{Uniform}((0,1)) & \text{if } \Delta \text{ has a transition at time } t \\ M(t) & = & 0 & \text{otherwise,} \end{array}$$

ensuring that (Δ, M) is also time-reversible.

Now we show how to obtain the target SBD process $\{X_t\}$ from (Δ, M). Let $\Delta_t = y$ and $X_t = x$ where $x \subseteq y$, and let τ be the waiting time before a transition takes place in Δ. Then X_t is determined by (Δ, M) via

$$\begin{aligned} \Delta_{t+\tau} = y \backslash \xi \quad &\Rightarrow \quad X_{t+\tau} = x \backslash \xi \\ \Delta_{t+\tau} = y \cup \xi \quad &\Rightarrow \quad X_{t+\tau} = \begin{cases} x \cup \xi & \text{if } M(t+\tau) \leq \frac{\lambda(x,\xi)}{C} \\ x & \text{otherwise} \end{cases} \\ & \qquad X_s = X_t \quad \text{if } t \leq s < t + \tau. \end{aligned}$$

In the following we will show that $\{X_t\}$ is a spatial birth-and-death process with birth and death rates given by

$$\begin{array}{rcl} b(x,\xi) & = & \lambda(x;\xi) \\ d(\cdot,\cdot) & = & 1. \end{array} \tag{6.29}$$

Proposition 6.28
Let $J-1$ be geometrically distributed with parameter q, i.e. $\mathbb{P}(J-1=n) = q(1-q)^n$ for $n = 0,1,\ldots,$ and let $M \mid J-1=n$ be gamma distributed with shape parameter $n+1$ and inverse scale parameter c. Then

$$M \sim \text{Exp}(cq).$$

Proof:
Since $M \mid J-1 = n \sim \Gamma(n+1,c)$ we have that the density of M is given by

$$\begin{aligned} f(m) &= \sum_{n=0}^{\infty} q(1-q)^n \frac{c^{n+1}}{n!} m^n \exp(-cm) \\ &= cq\exp(-cm) \sum_{n=0}^{\infty} \frac{1}{n!}(cm(1-q))^n \\ &= cq\exp(-cqm), \end{aligned}$$

and hence $M \sim \text{Exp}(cq)$. ◦

Proposition 6.29
Conditional on $X_t = x$, the waiting time for a birth in the target SBD is

$$\tau_0 \quad \sim \quad \text{Exp}(B(x))$$

where $B(x) = \int\limits_{\mathcal{X}} \lambda(x;\xi)\mathrm{d}\mu(\xi)$ *and the waiting times* τ_i *for deaths* $x_i \in x$ *are*

$$\tau_i \sim \mathrm{Exp}(1),$$

and τ_0 *and* τ_i *are mutually independent.*

Proof:
The waiting times for deaths follow directly from the corresponding waiting times for deaths in Δ.

Let $(\xi_1, \xi_2, \dots)$ be the sequence of events born in Δ after time t so that ξ_i is born at time t_i where $t_1 < t_2 < \dots$ and let $t_0 = t$. Since $(t_i - t_{i-1}) \sim \mathrm{Exp}(C\mu(\mathcal{X}))$ for $i = 1, 2, \dots$ are i.i.d. we have that $(t_i - t) \sim \Gamma(i, C\mu(\mathcal{X}))$.

Now, let $I = \inf\{i \in \mathbb{N} : M(t_i) \le \lambda(x;\xi_i)/C\}$ correspond to the first event born in X after time t. Since the ξ_i are i.i.d., the $M(t_i)$ are i.i.d., and the ξ and the $M(t_i)$ are independent, we obtain that

$$\begin{aligned} \mathbb{P}(M(t_i) \le \lambda(x;\xi_i)/C) &= \int\limits_{\mathcal{X}} \int_0^1 \mathbb{1}\left[m \le \frac{\lambda(x;\xi)}{C}\right] \mathrm{d}m \frac{\mathrm{d}\mu(\xi)}{\mu(\mathcal{X})} \qquad (6.30) \\ &= \frac{B(x)}{C\mu(\mathcal{X})}. \end{aligned}$$

So $I - 1$ is geometrically distributed with parameter $q = \frac{B(x)}{C\mu(\mathcal{X})}$. Further $\tau_0 = (t_I - t)|I) \sim \Gamma(I, C\mu(\mathcal{X}))$. By applying Proposition 6.28 the result $\tau_0 \sim \mathrm{Exp}(B(x))$ thus follows. Finally it follows immediately that τ_0 and τ_i are mutually independent. ∘

As a consequence of this result we see that $B(x)$ and d have the correct values in order to satisfy (6.29). Hence, it remains only to show that we can take $b(x,\xi) = \lambda(x;\xi)$, i.e. an event which is born comes from a density $\frac{\lambda(x;\xi)}{B(x)}$ w.r.t. μ.

Proposition 6.30
Let $X_t = x$, *let the next transition in* X *be a birth at time* $t + \tau$ *and let* ξ *denote the point that is born. Then*

$$\mathbb{P}(\xi \in A \mid X_{t+\tau} = x \cup \xi, X_s = x, t \le s < t + \tau) = \int\limits_A \frac{\lambda(x;\xi)}{B(x)} \mathrm{d}\mu(\xi)$$

for any $A \in \mathcal{B}$.

Proof:

$$
\begin{aligned}
&\mathbb{P}\left(\xi \in A \mid X_{t+\tau} = x \cup \xi, X_s = x, t \leq s < t+\tau\right) \\
&\quad = \mathbb{P}\left(\xi \in A \ \mid\ M(t+\tau) \leq \frac{\lambda(x;\xi)}{C}\right) \\
&\quad = \frac{\mathbb{P}(\xi \in A, M(t+\tau) \leq \frac{\lambda(x;\xi)}{C})}{\mathbb{P}(M(t+\tau) \leq \frac{\lambda(x;\xi)}{C})} \\
&\quad = \frac{\int_A \lambda(x;\xi)\mathrm{d}\mu(\xi)}{B(x)}
\end{aligned}
$$

as seen by (6.30). $\circ$

We have thus proven that the target process X behaves as a spatial birth-and death process with b and d as in (6.29) from time t to time $t+\tau$ and by induction this holds for the entire process. Finally, we know from Section 6.6.1 that X converges towards π in distribution.

Comparing SBD and MH–methods

Comparing the two methods of simulation it is obvious that the Metropolis–Hastings method is easier to implement. In the case of the Strauss process Clifford & Nicholls (1994) found that the Metropolis–Hastings algorithm is simplest and most efficient in most cases. Also, the theory for the Metropolis–Hastings method is to the best of our knowledge much more developed than simulation based on spatial birth-and-death processes.

On the other hand it seems so far that perfect simulation (treated in the next section) based on spatial birth-and-death processes is simpler than perfect simulation using the Metropolis–Hastings method (as studied in Kendall & Møller (1999)).

6.7 Perfect Simulation

Simulation using classical MCMC methods as considered so far fail to be exact because the algorithms only simulate a Markov chain for a finite period of time, i.e. for such MCMC methods to be exact, they would have to be run for infinite time. Here we present a method called **exact** or **perfect simulation** based on **coupling from the past** (CFTP). The concept of coupling from the past was introduced by Propp & Wilson (1996). Here we give an introduction based on papers by Propp & Wilson (1996), Häggström, van Lieshout & Møller (1999) and Kendall & Møller (1999). An alternative way to do perfect simulation is by a complicated rejection sampler called Fill's Algorithm (Fill 1998, Thönnes 1999, Møller & Schladitz 1999) but here we restrict attention to coupling from the past.

Assume that we have an ergodic Markov chain $\{X_t\}$ with transition kernel P and equilibrium distribution π. We now represent this Markov chain by a **transition function** $\phi : E \times \mathcal{R} \to E$, such that the stochastic recursive sequence

$$X_t = \phi(X_{t-1}, R_t),$$

where R_t are i.i.d. random variables defined on some measure space $\mathcal{R}$ with distribution $\mathfrak{R}$ used for determining the transition from state X_{t-1} to X_t. In implementation ϕ is given by the computer code and R_t is a vector (possible of random length) of uniform numbers. The connection between the transition function and the Markov transition kernel is given by

$$\mathfrak{R}(\phi(x, R) \in F) = P(x, F).$$

Let $\{R\}_a^b = (R_a, R_{a+1}, \ldots, R_b)$ and let

$$X_a^b(x) \quad = \quad \phi(\ldots(\phi(\phi(x, R_{a+1}), R_{a+2})\ldots), R_b),$$

i.e. the result of running the chain from state x at time a to time b using $\{R\}_{a+1}^b$. With this notational setup we are ready to present the algorithm that constitutes perfect simulation.

Algorithm [Perfect Simulation]

1. For $t = 1, 2, 4, \ldots$ repeat:
2. If $t = 1$ then generate a realisation of R_0, otherwise generate realisations of $\{R\}_{-t+1}^{-t/2}$ and reuse the previously generated realisations of $\{R\}_{-t/2+1}^0$.
3. until $X_{-t}^0(x) = X_{-t}^0(y) \quad \forall x, y \in E$.
4. Output X_{-t}^0.

When $X_{-t}^0(x) = X_{-t}^0(y)$ the two chains are said to **coalesce** or to be in **coalescence**. Let $T = \inf_{n \in \mathbb{N}} \{t = 2^{n-1} : X_{-t}^0(x) = X_{-t}^0(y), \forall x, y \in E\}$ be the (stochastic) time so that all chains run from time $-T$ to time 0 are in coalescence. For the algorithm to terminate a.s. T naturally has to be a.s. finite, so this has to be checked for the specific Markov chain. The algorithm uses coupling from the past as all chains are coupled by using the same realisations of the random variables for the past, i.e. time prior to 0.

The following theorem shows that the output of the algorithm is in fact a sample from π.

Theorem 6.31
If T is a.s. finite then the distribution of the output X^0_{-T} is π.

Proof:
Let $F \in \mathcal{C}$. Since $X^0_{-t} = X^0_{-T}$ for any $t \geq T$ we get that

$$\begin{aligned}
\mathbb{D}(X^0_{-T} \in F) &= \mathbb{D}(\lim_{t\to\infty} X^0_{-t} \in F) &&= \mathbb{E}\,\mathbb{1}[\lim_{t\to\infty} X^0_{-t} \in F] \\
&= \mathbb{E} \lim_{t\to\infty} \mathbb{1}[X^0_{-t} \in F] &&= \lim_{t\to\infty} \mathbb{E}\,\mathbb{1}[X^0_{-t} \in F] \\
&= \lim_{t\to\infty} \mathbb{D}(X^0_{-t} \in F) &&= \lim_{t\to\infty} \mathbb{D}(X^t_0 \in F) \\
&= \pi(F)
\end{aligned}$$

using the dominated convergence theorem in the second line and the fact that the Markov chain is time-homogeneous in the third line. ∘

Before we illustrate the perfect simulation algorithm with an example we will consider the effect of using the scheme $t = 1, 2, 4, \ldots$

It is easily seen that we could use any strictly increasing sequence of positive integers for the algorithm and still get samples from π. Hence, it might seem that letting t step through $1, 2, 4, \ldots$ is inefficient, but actually this is fairly efficient. Assume that it takes t time units to check coalescence from time $-t$, and let the realisations of all the random variables R be given. Let T^* be the minimum integer so that all the Markov chains are in coalescence if run from time $-T^*$ to time 0. Letting n be the integer, so that $2^{n-1} < T^* \leq 2^n$, the time-complexity of the algorithm is

$$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1 = 4 * 2^{n-1} - 1 < 4T^*,$$

so it is less then four times the minimum time needed. For comparison, the scheme $T = 1, 2, 3, \ldots$ gives a time-complexity that is $\frac{1}{2}(T^* + 1)T^*$.

Another important issue is that the algorithm reuses the realisations. If it instead generated new realisations the algorithm would not work properly as we shall see towards the end of the following simple example.

Example 6.32 (Perfect Sampling on a Finite State Space)
Consider a point process X with only three possible outcomes $\{0, 1, 2\}$ and distribution π given by that

$$\pi(0) = \pi(1) = \pi(2) = \frac{1}{3}.$$

Obviously, we could sample from π directly with ease, but in the following we use it to illustrate perfect simulation.

Consider a Markov chain given by the kernel P where

$$P(0,\{0\}) = P(0,\{1\}) = P(1,\{0\}) = P(1,\{2\}) = P(2,\{1\}) = P(2,\{2\}) = \frac{1}{2},$$

which clearly is ergodic and has π as equilibrium distribution. Figure 6.8 illustrates this chain.

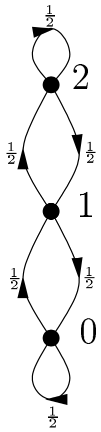

Figure 6.8: Transition probabilities for a random walk defined as $\{0, 1, 2\}$ with reflecting barriers.

Then each R_t could be represented by u for up or d for down each with probability $\frac{1}{2}$ and obvious effect. Now, running the algorithm, we first set $t = 1$. No matter if R_0 is u or d the algorithm will not stop. Now, at $t = 2$, there are 4 different possibilities for $\{R\}_{-1}^{0}$: (u, u), (u, d), (d, u), and (d, d), each happens with probability $\frac{1}{4}$. The (u, u) results in an output of 2, (d, d) result in 0 and the middle two (u, d) and (d, u) do not produce an output. If the algorithm continues to $t = 4$, there are 8 different possibilities: (u, u, u, d), (u, u, d, u), (u, d, u, d), (u, d, d, u), (d, u, u, d), (d, u, d, u), (d, d, u, d), (d, d, d, u). Note that there are no $(\cdot, \cdot, u, u)$ or $(\cdot, \cdot, d, d)$ because then the algorithm would already have stopped. Of these 8 possibilities one produces the result 2, four output 1, one output 0 and two give no output. At this time we have a preliminary distribution $\mathbb{P}_4$ where $\mathbb{P}_4(\text{no output}) = \frac{2}{16}$, $\mathbb{P}_4(1) = \frac{4}{16}$, and $\mathbb{P}_4(0) = \mathbb{P}_4(2) = \frac{5}{16}$. Continuing the algorithm using inductive arguments, we see that when $t = s$ we get that $\mathbb{P}_s(\text{no output}) = \frac{2}{s^2}$, $\mathbb{P}_s(1) = \frac{\lfloor s^2/3 \rfloor - 1}{s^2}$ and $\mathbb{P}_s(0) = \mathbb{P}_s(2) = \frac{\lfloor s^2/3 \rfloor}{s^2}$ which indeed converges to π.

Finally, consider what would happen if we did not reuse the realisations of the random variable. For $t = 4$ we would now also get $(\cdot, \cdot, u, u)$ and $(\cdot, \cdot, d, d)$ realisations for a total of 16 different possibilities. Of these, two give no output, five give 0, four give 1 and five give 2, naturally corresponding to $\mathbb{P}_4$ as described above. However, combined with the fact the algorithm could stop when $t = 2$

it is easily seen that the output will be severely biased towards 0 *and* 2*, already when* $t = 4$ *we have a preliminary distribution* $\mathbb{P}_4^*$ *where* $\mathbb{P}_4^*(\textit{no output}) = \frac{2}{32}$, $\mathbb{P}_4^*(1) = \frac{4}{32}$ *and* $\mathbb{P}_4^*(0) = \mathbb{P}_4^*(2) = \frac{13}{32}$.

6.7.1 Monotone Coupling From the Past

Naturally, not all examples are that simple. Often, the state space is infinite and hence we need a method for detecting when the infinite number of chains are all in coalescence.

Definition 6.33 ((Anti)-Monotone Transition Function)
A transition function ϕ *is monotone w.r.t. a partial order relation* $\leq_E$ *on* E *if* $x \leq_E y \Rightarrow \phi(x,r) \leq_E \phi(y,r)$ *for all* $x, y \in E$ *and all* $r \in \mathcal{R}$*. Conversely,* ϕ *is anti-monotone w.r.t.* $\leq_E$ *if* $x \leq_E y \Rightarrow \phi(y,r) \leq_E \phi(x,r)$.

For a monotone or anti-monotone transition function we get certain properties. Let $\leq_E$ be a partial order relation on E and let ϕ be a monotone transition function w.r.t. $\leq_E$. Assume that l is a unique minimum w.r.t. $\leq_E$ and that u is a unique maximum w.r.t. $\leq_E$, i.e. $l \leq_E x \leq_E u$ for all $x \in E$. Let $L_a^b = X_a^b(l)$ and $U_a^b = X_a^b(u)$ for any $b \geq a$, taking $X_a^a(x) = x$.

Then the following **sandwiching properties**

$$L_a^a \leq_E L_a^b \leq_E X_a^b(x) \leq_E U_a^b \leq_E U_a^a \quad \forall a, b \in \mathbb{Z}, a < b, \quad \forall x \in E$$

and the following **funneling properties**

$$L_b^c \leq_E L_a^c \leq_E U_a^c \leq_E U_b^c \quad \forall a, b, c \in \mathbb{Z}, a < b < c$$

are easily seen to be satisfied.

Conversely, if ϕ is anti-monotone w.r.t. $\leq_E$, define L_a^b and U_a^b inductively as

$$\begin{aligned} L_a^a &= l \\ U_a^a &= u \\ L_a^b &= \phi(U_a^{b-1}, R_b) \\ U_a^b &= \phi(L_a^{b-1}, R_b) \end{aligned}$$

again resulting in the sandwiching and funneling properties being satisfied, this trick is due to Kendall (1998).

In either case we call L and U lower and upper processes respectively.

This clearly solves the problem of deciding when the infinite number of chains have coalesced. If $L_{-t}^0 = U_{-t}^0$ we get the coalescence of all chains $X_{-t}^0(\cdot)$ by the sandwiching properties, so we only need to run these two chains.

Example 6.34 (Perfect Simulation of the Widom-Rowlinson Model) *Consider the bivariate Widom-Rowlinson model (see Häggström et al. (1999) and Example 5.23) $X = (Y, Z)$ on $\mathcal{X}$. Since the conditional distribution of one variate, say Y, given the other is a Poisson point process with intensity measure $\beta_Y \mu$ on $\mathcal{X} \setminus \bigcup_{\xi \in Z} b(\xi, r)$, it seems appropriate to apply a Gibbs sampler, see Section 6.4.2.*

Hence, we get a Markov chain $\{X_t\} = \{(Y_t, Z_t)\}$ with transition function ϕ given by

$$\begin{aligned}
(Y_t, Z_t) &= \phi((Y_{t-1}, Z_{t-1}), (V_t, W_t)) \\
Y_t &= V_t \setminus \bigcup_{\xi \in Z_{t-1}} b(\xi, r) \\
Z_t &= W_t \setminus \bigcup_{\xi \in Y_t} b(\xi, r)
\end{aligned}$$

where the V_t and the W_t are independent Poisson point processes with intensity measure $\beta_Y \mu$ and $\beta_Z \mu$ respectively.

Since this is a Gibbs sampler it has the correct invariant distribution and by Theorem 6.20 it is uniformly ergodic if E is small. But, for all $x \in E$ and all $F \in \mathcal{C}$ we have that

$$P(x, F) \geq \mathbb{P}(V_t = W_t = \emptyset) \mathbb{1}[(\emptyset, \emptyset) \in F]$$

so E is small by Definition 6.17 taking $t = 1$, $\varepsilon = \mathbb{P}(V_t = W_t = \emptyset) > 0$ and $\zeta(F) = \mathbb{1}[(\emptyset, \emptyset) \in F]$, and hence the chain is uniformly ergodic.

Clearly, ϕ is monotone w.r.t. the partial order relation

$$(y, z) \leq_E (y', z') \quad \Leftrightarrow \quad y \subseteq y' \ \wedge \ z' \subseteq z$$

and we need to find a unique minima and maxima. If we technically allow ϕ and $\leq_E$ to operate on the extended space $E \cup \{\mathcal{X}\}$ then $(\emptyset, \mathcal{X})$ is a unique minimum and $(\mathcal{X}, \emptyset)$ is a unique maximum. Alternatively, we could consider ***quasi-minima/maxima****. The state $(\cdot, V_t)$ is a quasi-minimum at time $t-1$, because*

$$\phi((\cdot, V_t)) \;=\; (\emptyset, W_t) \;\leq_E\; \phi((y, z)) \quad \forall (y, z) \in E,$$

and equivalently $(\cdot, \emptyset)$ is a quasi-maximum.

Hence, when running the perfect sampling algorithm, we run the chains $L^0_{-t} = X^0_{-t}$(quasi-minimum) and $U^0_{-t} = X^0_{-t}$(quasi-maximum) and output when these are in coalescence.

Alternatively, ϕ is anti-monotone w.r.t.

$$(y, z) \leq_E (y', z') \quad \Leftrightarrow \quad y' \subseteq y \ \wedge \ z \subseteq z',$$

leading to similar results.

6.7.2 Dominated Coupling from the Past

In this section we describe how to use spatial birth and death processes to make perfect sampling from any locally stable point process, based on the work by Kendall & Møller (1999).

Recall from Section 6.6.2 the dominating process Δ with marks M for deaths and births and the thinning process used to simulate a locally stable target point process X. We will now describe a method called **dominated CFTP** to detect when all chains run from time $-t$ are in coalescence at time 0 and equilibrium hence is reached.

Since (Δ, M) is time-reversible, it is possible to run backwards simply by running forwards and interchanging births and deaths. So, run (Δ, M) backwards in time until $\Delta = \emptyset$ and let $-T_\Delta$ denote the time when this happens. Since $\emptyset$ is an ergodic atom, T_Δ will almost surely be finite. Now run a target process X from time $-T_\Delta$ to time 0 starting in state $\emptyset$, using (Δ, M) to determine transitions. Then X at time 0 will be a sample from the target distribution by arguments similar to the proof of Theorem 6.31.

However theoretically simple and appealing this approach is, it will often be infeasible for applications because T_Δ will be too large and the algorithm hence too slow. Instead, an approach using lower and upper processes w.r.t. the partial ordering of set inclusion, i.e. $x \leq_E y \Leftrightarrow x \subseteq y$ may be used.

Again we run (Δ, M) backwards but this time only for a set amount of time, say until time $-t$. At time $-t$ we let the lower and upper processes be

$$\begin{aligned} L_{-t}(-t) &= \emptyset \\ U_{-t}(-t) &= \Delta(-t). \end{aligned}$$

Now we run these two processes forwards in time using (Δ, M). If Δ experiences a death of ξ at time s then both L and U experience the death of ξ at time s provided s is part of L and U respectively. If Δ experiences a birth of η at time s things get a bit more complicated. Let $x_{min} = L_{-t}(s-)$ be the state of L just prior to time s and similarly let $x_{max} = U_{-t}(s-)$. Consider a target process $X(s-) = x$ where $x_{min} \subseteq x \subseteq x_{max}$. If for all such x the event η would be added to X then η is added to L and if for any such x it would be added to X it is added to U. So, η is added to L if

$$M(s) \leq \min\{\lambda(x;\eta)/C : x_{min} \subseteq x \subseteq x_{max}\} \tag{6.31}$$

and η is added to U if

$$M(s) \leq \max\{\lambda(x;\eta)/C : x_{min} \subseteq x \subseteq x_{max}\}, \tag{6.32}$$

where C is as defined in Section 6.6.2.

Since any target process is dominated by Δ, i.e. $X(s) \subseteq \Delta(s)$, we have by this construction that both the sandwiching and funneling properties are satisfied. Hence, running the perfect sampling algorithm, extending (Δ, M) backwards in time with each step until the lower and upper processes coalesce at time 0, results in a sample from the target distribution.

In general, finding the extrema in (6.31) and (6.32) is no easy task. However, for attractive point processes the task is rather simple since

$$\begin{aligned} \min \lambda(x;\eta) &= \lambda(x_{\min};\eta) \\ \max \lambda(x;\eta) &= \lambda(x_{\max};\eta), \end{aligned}$$

and similarly for repulsive point processes such as the Strauss process where

$$\begin{aligned} \min \lambda(x;\eta) &= \lambda(x_{\max};\eta) \\ \max \lambda(x;\eta) &= \lambda(x_{\min};\eta). \end{aligned}$$

6.7.3 Remarks on Perfect Simulation

We now have a theoretical way of making exact samples from a target distribution. However, when we try to implement the theory on a computer we run into the problem of making a true random number generator. Since we have to settle for a pseudo random number generator and also only have finite precision in numbers and calculations, the implementation will not result in exact samples. However, this problem also applies to standard MCMC methods and to simulation procedures in general.

If we disregard this hopefully minor problem, we may generate i.i.d. exact samples from the distribution by running the algorithm repeatedly or if we do not need independence we may start a regular Markov chain from an obtained perfect sample, eliminating the need for burn-in.

Chapter 7

Parametric Inference

In this chapter we investigate several methods for parametric inference for point processes. The primary setup is that we have data x which we assume to be a realization of a point process X with a parameterized density p_θ w.r.t. a Poisson point process Π with finite diffuse intensity measure μ and then wish to estimate θ within some parameter space Θ. If p_θ is explicitly known we may use standard maximum likelihood estimation (MLE), but often we only know that

$$p_\theta(x) \quad = \quad \frac{g_\theta(x)}{c_\theta}$$

where g_θ is explicitly known but the normalizing constant c_θ is unknown and analytically intractable.

In fact it is only in the case of a Poisson point process that the maximum likelihood estimate is easily calculated. For the homogeneous Poisson point process with intensity λ on $\mathcal{X}$ the likelihood function is (see Example 3.12)

$$p(x;\theta) \quad = \quad p(x;\lambda) \;=\; \exp(-(\lambda-1)\mu(\mathcal{X}))\lambda^{x(\mathcal{X})},$$

and so the MLE is $\hat{\lambda} = \frac{x(\mathcal{X})}{\mu(\mathcal{X})}$. In the inhomogeneous case certain iterative optimization methods are available (Baddeley & Turner 1998).

In Section 7.2 we describe pseudo-likelihood estimation (PLE) which is a method for estimating θ without considering c_θ. Essentially, this method takes the likelihood function for the homogeneous Poisson point process

$$\begin{aligned} p(\theta; x) &\propto \exp\left(-\int_{\mathcal{X}} \lambda \mathrm{d}\mu(\xi)\right) \lambda^{x(\mathcal{X})} \\ &= \exp\left(-\int_{\mathcal{X}} \lambda(x;\xi) \mathrm{d}\mu(\xi)\right) \prod_{\xi \in x} \lambda(x \backslash \xi; \xi) \end{aligned} \tag{7.1}$$

which does not depend on c_θ and uses the expression (7.1), called the pseudo-likelihood function, to estimate θ for any point process. Intuitively, this approach should work well when we consider processes that have no or little interaction between events but it might perform less well with strong interaction.

In Section 7.3 we consider MCMC maximum likelihood estimation which uses MCMC and importance sampling to estimate θ. MCMC maximum likelihood estimation is considered to be better than pseudo-likelihood estimation, especially in case of strong interaction, but PLE is much faster and is hence often used in preliminary investigations or when several different parametric model classes are investigated.

In Section 7.4 we very briefly introduce a stochastic gradient algorithm for estimating the parameters and empirically compare this to PLE and MCMC maximum likelihood estimation on 500 samples of a particular Strauss process.

In some cases we may not even know the density up to proportionality and hence we can neither use PLE nor use MCMC maximum likelihood estimation. We may instead use minimum contrast methods as presented in Section 7.5.

In Section 7.6 we finally consider how to find the distribution of the estimate, to make a confidence region, how to make tests of hypotheses expressing that the data follows some type of model given that it is assumed to follow a less restricted type of model, i.e. that the data is a realisation of a Poisson point process given that it is assumed to be a realisation of a Strauss process.

To present these different estimation methods we first introduce the exponential family models which comprise a large class of models, including many of the models treated in Chapter 5.

7.1 MLE in Exponential Family Models

The density p_θ w.r.t. Π of an **exponential family model** is on the form

$$p_\theta(x) = \frac{1}{c_\theta} h(x) \exp(\theta \cdot t(x)),$$

where $\theta \in \Theta \subseteq \mathbb{R}^n$ parameterizes the density, $h(x)$ is a measurable function, $t(x)$ is a n-dimensional measurable function, and $\cdot$ denotes the inner product. For more material on exponential family models than presented in this section we refer to Barndorff-Nielsen (1978).

We are said to have a **minimal representation** if $t(X)$ with $X \sim p_\theta$ is not concentrated on a $(n-1)$-dimensional subspace (to avoid having over-parameterized the model) and Θ is n-dimensional, so that it makes sense to consider derivatives of p_θ along the axes of $\mathbb{R}^n$ everywhere in Θ. For example, Θ could be open, rectangular, etc.

In the following we investigate the MLE of exponential family models, assuming that we have a minimal representation and further that Θ is convex. As we shall see, this ensures some desirable results about the maximum likelihood function of exponential family models. Often the parameter space Θ is given by the maximal parameter space

$$\Theta = \left\{ \theta \in \mathbb{R}^n \ : \ \int_N h(x) \exp(\theta \cdot t(x)) \mathrm{d}\Pi(x) < \infty \right\} \tag{7.2}$$

which is seen to be convex by Hölder's inequality.

Proposition 7.1
The MLE $\hat{\theta}$ exists if and only if $\mathbb{E}_\theta t(X) = t(x)$ has a solution $\theta \in \Theta$.

Proof:
The likelihood function can be written as

$$L(\theta) = \frac{1}{c_\theta} \exp(\theta \cdot t(x)),$$

and hence the log-likelihood function is

$$l(\theta) = -\log c_\theta + \theta \cdot t(x).$$

Differentiation gives

$$\frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta} = -\frac{1}{c_\theta} \frac{\mathrm{d}c_\theta}{\mathrm{d}\theta} + t(x),$$

and observing that

$$\begin{aligned} \frac{\mathrm{d}c_\theta}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta} \int_N h(x) \exp(\theta \cdot t(x)) \mathrm{d}\Pi(x) \\ &= \int_N t(x) h(x) \exp(\theta \cdot t(x)) \mathrm{d}\Pi(x) \\ &= c_\theta \mathbb{E}_\theta t(X), \end{aligned}$$

we get the likelihood equation

$$\mathbb{E}_\theta t(X) = t(x). \tag{7.3}$$

Further we see that l is concave since

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\theta \mathrm{d}\theta^\top} l(\theta) &= \frac{1}{c_\theta^2} \frac{\mathrm{d}c_\theta}{\mathrm{d}\theta} \frac{\mathrm{d}c_\theta}{\mathrm{d}\theta^\top} - \frac{1}{c_\theta} \frac{\mathrm{d}^2}{\mathrm{d}\theta \mathrm{d}\theta^\top} c_\theta \\
&= \frac{c_\theta \mathbb{E}_\theta t(X) c_\theta \mathbb{E}_\theta t(X)^\top}{c_\theta^2} - \frac{1}{c_\theta} \frac{\mathrm{d}}{\mathrm{d}\theta} \left(c_\theta \mathbb{E}_\theta t(X)^\top \right) \\
&= \mathbb{E}_\theta t(X) \mathbb{E}_\theta t(X)^\top - \frac{1}{c_\theta} \frac{\mathrm{d}}{\mathrm{d}\theta} \int_N t(x)^\top h(x) \exp(\theta \cdot t(x)) \mathrm{d}\Pi(x) \\
&= \mathbb{E}_\theta t(X) \mathbb{E}_\theta t(X)^\top - \mathbb{E}_\theta (t(X) t(X)^\top) \\
&= -\mathrm{Var}_\theta t(X)
\end{aligned}$$

and also strictly concave as the minimal representation ensures that $\mathrm{Var}_\theta(\alpha \cdot t(X)) > 0$ for all $\alpha \in \mathbb{R}^n \setminus \{0\}$.

Hence the MLE $\hat{\theta}$ exists if and only if $\mathbb{E}_\theta t(X) = t(x)$ has a solution for $\theta \in \Theta$. ∘

Using results in Barndorff-Nielsen (1978) it is further possible to show the following.

Theorem 7.2
Assume that we have a minimal representation where Θ is given by (7.2). Let C be the interior of the closure of the smallest convex set containing a support of $t(X)$, where $X \sim p_\theta$. Finally, let $T = \{\mathbb{E}_\theta t(X) \mid \theta \in int(\Theta)\}$, where int denotes the interior of a set, and assume $X = x$ is observed.

Then the MLE $\hat{\theta}$ exists if and only if $t(x) \in C$. Further, $\hat{\theta}$ is found in the interior of Θ if $t(x) \in T$ and otherwise it lies on the boundary of Θ.

Example 7.3 (MLE for the Strauss Process)
Consider the Strauss process from Example 5.18. For a fixed R this is an exponential family model where $h(x) \equiv 1$, $t(x)^\top = (x(\mathcal{X}), S(x))$ and $\theta^\top = (\log \beta, \log \gamma)$. Hence, $\Theta = \mathbb{R} \times (-\infty, 0]$ omitting the case $\gamma = 0$ where $\log(\gamma) = -\infty$.

Since both $\beta > 0$ and $\gamma > 0$, the support of $t(X)$ will contain the set $\left\{ (n, i_n), \ldots, \left(n, \frac{n(n-1)}{2}\right) : n = 0, 1, \ldots \right\}$ for constants $i_n \in \mathbb{N}_0$ depending on R and $\mathcal{X}$ indicating the minimum possible number of neighbours given n events, so $C \subseteq \mathbb{R}_+ \times \mathbb{R}_+$. For small n we have that $i_n = 0$ but as n tends to infinity i_n also tends to infinity. By Theorem 7.2 we obtain that the MLE exists if and only if $S(x) > i_n$ where $n = x(\mathcal{X})$, i.e. there are at least one pair of neighbours above the minimum possible number.

Consider the boundary of Θ *where* $\gamma = 1$*. Here we have a Poisson point process so we get an upper bound for* $\mathbb{E}t(X)$ *given by*

$$\begin{aligned} \mathbb{E}_{(\beta,1)} X(\mathcal{X}) &= \beta\mu(\mathcal{X}) \\ \mathbb{E}_{(\beta,1)} S(X) &\leq \frac{\beta^2}{2}\mu(\mathcal{X})\mu(b(0,R)) \end{aligned}$$

because the average number of neighbours per event is bounded by $\beta\mu(b(0,R))$*. Hence, it appears that*

$$T = \left\{(n,s) \in \mathbb{N}\times\mathbb{N} : s < \frac{n^2\mu(b(0,R))}{2\mu(\mathcal{X})}\right\}.$$

and we thus have to estimate $\gamma < 1$ *if* $t(x) \in T$ *and* $\gamma = 1$ *otherwise, which is hardly surprising. We note that if* $s > \frac{n^2\mu(b(0,R))}{2\mu(\mathcal{X})}$ *then it might be more appropriate to fit some sort of a cluster model than the Strauss process.*

7.2 Pseudo-likelihood Estimation

Pseudo-likelihood estimation may be considered as an alternative to maximum likelihood estimation. Studies in pseudo-likelihood estimation were introduced by Besag (1974) for lattice processes, and extended to point processes, specifically the Strauss process, by Besag (1977).

Here we start with a brief look at the pseudo-likelihood estimate (PLE) for a parametric model of a random vector, which is used to obtain a similar definition of the PLE for a spatial point process. Essentially the idea is to approximate a spatial point process by a random vector, and then use the pseudo-likelihood function for the latter.

This section is based on Ripley (1988), Jensen & Møller (1991), and Baddeley & Turner (1998).

7.2.1 Original Definition of Pseudo-likelihood

Let $X = (X_1, \ldots, X_n)$ be a random vector which has a distribution given by some parameterized density p_θ. If Y is a maximal sub-vector of X so that the components of Y are mutually independent given the components in X that are not in Y, denoted by $X\backslash Y$, then Y is called a **coding set** of X and

$$p_\theta(y \mid x\backslash y) = \prod_{\xi\in y} p_\theta(\xi \mid x\backslash y) = \prod_{\xi \in y} p_\theta(\xi \mid x\backslash\xi)$$

is called a **coding function**. Now, we could select any coding set for X and maximize the corresponding coding function w.r.t. θ and use the resulting θ as an estimate. Instead, we combine all the coding functions into a single function

$$\mathrm{PL}\,(\theta; x) = \prod_{\xi \in x} p_\theta\,(\xi \mid x \backslash \xi)$$

called the **pseudo-likelihood function** and wish to maximize this function w.r.t. θ in order to obtain the **maximum pseudo-likelihood estimate** (MPLE).

7.2.2 Pseudo-likelihood for Spatial Point Processes

Assume that p_θ is hereditary and Ruelle stable and let $A \in \mathcal{B}_0$ with $\mu(A) > 0$. Then the conditional density of X_A given $X_{A^c} = x_{A^c}$ w.r.t. Π_A which was also considered in Section 5.6 is given by

$$p_\theta(x_A \mid x_{A^c}) = \begin{cases} \dfrac{p_\theta(x_A \cup x_{A^c})}{\int_A p_\theta(y \cup x_{A^c}) \mathrm{d}\Pi_A(y)} & \text{if } p_\theta(x_{A^c}) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7.4}$$

Consider nested partitions $T^i = \{T_1^i, \ldots, T_{r_i}^i\}$ of A, for instance those used in the proof of Theorem 2.10, and assume that $\lim_{i\to\infty} r_i = \infty$ and $\lim_{i\to\infty} r_i \delta_i^2 = 0$, where $\delta_i = \max_{1 \le j \le r_i} \mu(T_j^i)$. Eventually $x(T_j^i) \le 1$ and under suitable continuity conditions, see Besag, Milne & Zachary (1982), X_A can therefore be approximated by a random vector $(X_1, \ldots X_{r_i})$ where each X_j represents either no event or a uniformly distributed event within T_j^i. This motivates the following definition.

Definition 7.4 (Pseudo-likelihood Function)
The pseudo-likelihood function for a point process X on a set $A \in \mathcal{B}$ is given by

$$\mathrm{PL}_A(\theta; x) = \exp(-\mu(A)) \lim_{i\to\infty} \prod_{j=1}^{r_i} p_\theta\left(x_{T_j^i} \mid x_{\mathcal{X} \backslash T_j^i}\right).$$

The choice of A depends on the application. Typically we only observe x within some bounded window W in which case we let $A \subset W$ so that interaction between events within A and events outside W is negligible. In particular, we can often select A for a Markov point process of finite range so there is no interaction, i.e. $\partial(A) \subseteq W$.

To show that the PL_A is well-defined and does not depend on the choice of nested partitions, we consider the parameterized Papangelou conditional intensity

$$\lambda_\theta(x; \xi) \quad = \quad \frac{p_\theta(x \cup \xi)}{p_\theta(x)} \quad = \quad \frac{p_\theta(x_A \cup \xi \mid x_{A^c})}{p_\theta(x_A \mid x_{A^c})}$$

by (7.4) for $\xi \in A \backslash x$.

Theorem 7.5
$\mathrm{PL}_A(\theta; x)$ *is well defined for* Π-a.a. x *and is given by*

$$\mathrm{PL}_A(\theta; x) = \exp\left(- \int_A \lambda_\theta(x;\xi)\mathrm{d}\mu(\xi) \right) \prod_{\xi \in x_A} \lambda_\theta(x \backslash \xi; \xi). \tag{7.5}$$

For a proof of this theorem we refer to Jensen & Møller (1991), where also consistency of the maximum pseudo-likelihood estimate is proved for a Markov point process of finite range. Jensen & Künsch (1994) show asymptotic normality of the MPLE for certain pairwise interaction processes.

Note that for a Markov process the pseudo-likelihood function only depends on x through $x_{\partial(A)}$, so if $\partial(A) \subseteq W$ it may be calculated based on the observed data, x_W. Notice further, that in the case of a Poisson process with intensity function $\lambda_\theta(\cdot)$, we have that $\lambda_\theta(x;\xi) = \lambda_\theta(\xi)$, and hence the pseudo-likelihood function reduces to the ordinary likelihood function. So, intuitively the MPLE should be close to the MLE for processes with no or little interaction between events but it might perform less well with strong interaction.

By differentiating the logarithm of (7.5) under the assumption that we may interchange integration and differentiation, for instance the strong assumption that $\frac{\mathrm{d}}{\mathrm{d}\theta}\lambda_\theta(x;\xi)$ is uniformly bounded in $\xi \in A$ and $\theta \in \Theta$ for each fixed x, we obtain the **log pseudo-likelihood equations**

$$\sum_{\xi \in X_A} \frac{\mathrm{d}}{\mathrm{d}\theta} \log \lambda_\theta(x \backslash \xi; \xi) \quad = \quad \int_A \frac{\mathrm{d}}{\mathrm{d}\theta} \lambda_\theta(x;\xi)\mathrm{d}\mu(\xi). \tag{7.6}$$

Recall, that if X is a finite range Markov process then by Example 5.37 it is a Gibbs process and hence

$$\mathbb{E}_\theta \sum_{\xi \in X_A} \tilde{h}(\xi, X \backslash \xi) \quad = \quad \mathbb{E}_\theta \int_A \tilde{h}(\xi, X)\lambda_\theta(x;\xi)\mathrm{d}\mu(\xi) \tag{7.7}$$

for any measurable function $\tilde{h}$, provided the means exist. Letting

$$\tilde{h}(\xi, X) \quad = \quad \frac{\mathrm{d}}{\mathrm{d}\theta} \log \lambda_\theta(x;\xi)$$

we see that omitting the means from (7.7) reduces it to (7.6), so the MPLE may be considered as a form of moment estimation. Hence pseudo-likelihood estimation is also a special case of **Takacs-Fiksel estimation** (Takacs 1986, Fiksel 1984, Fiksel 1988) where the idea is to choose $m \geq \dim \Theta$ different test functions $h_1, \ldots, h_m$ and then use (7.7) to obtain m estimation equations by omitting the $\mathbb{E}$-signs.

Consider now the case where the density p_θ belongs to the exponential family models, see Section 7.1, and assume a minimal representation. Further, in order to ensure the hereditary property of p_θ we assume that h is hereditary. Defining, for $x \in N$ and $\xi \in \mathcal{X}\backslash x$,

$$t(x;\xi) = t(x \cup \xi) - t(x),$$

$$h(x;\xi) = \frac{h(x \cup \xi)}{h(x)},$$

where we let $\frac{0}{0} = 0$, we find that

$$\lambda_\theta(x;\xi) = h(x;\xi)\exp(\theta \cdot t(x;\xi)),$$

which by Theorem 7.5 gives the log pseudo-likelihood function

$$\begin{aligned} pl_A(\theta;x) &= \log \mathrm{PL}_A(\theta;x) \\ &= \theta \cdot \sum_{\xi \in X_A} (\log h(x;\xi) + t(x\backslash\xi;\xi)) - \int_A h(x;\xi)\exp(\theta \cdot t(x;\xi))\mathrm{d}\mu(\xi). \end{aligned}$$

Observe that for any $\alpha \in \mathbb{R}^n$ and $\theta \in \Theta$ we have

$$\begin{aligned} &\alpha^\top \left\{ \frac{\mathrm{d}^2}{\mathrm{d}\theta\mathrm{d}\theta^\top} pl_A(\theta;x) \right\} \alpha \\ &= \alpha^\top \left\{ -\int_A t(x;\xi)t(x;\xi)^\top h(x;\xi)\exp(\theta \cdot t(x;\xi))\mathrm{d}\mu(\xi) \right\} \alpha \\ &= -\int_A [\alpha \cdot t(x;\xi)]^2 h(x;\xi)\exp(\theta \cdot t(x;\xi))\mathrm{d}\mu(\xi) \\ &\leq 0. \end{aligned}$$

Since the second derivative of the log pseudo-likelihood equation is negative semi-definite we conclude that $pl_A(\cdot)$ is concave.

The following example shows the pseudo-likelihood equations for a Strauss process.

Example 7.6 (Strauss Process)
Continuing from Example 7.3 we see that for the Strauss process

$$\lambda_\theta(x;\xi) \quad = \quad \beta\gamma^{s(x;\xi)} = \exp\left(\log\beta + s(x;\xi)\log\gamma\right).$$

Hence, inserting in (7.6) *yields the equations*

$$\begin{aligned} x(A) &= \beta \int_A \gamma^{s(x;\xi)} \mathrm{d}\mu(\xi) \\ \sum_{\xi \in x_A} s(x \backslash \xi; \xi) &= \beta \int_A s(x;\xi) \gamma^{s(x;\xi)} \mathrm{d}\mu(\xi). \end{aligned}$$

7.2.3 Computing the MPLE

In this section we consider an approximation method which enables us to calculate the integral in (7.5), based on Baddeley & Turner (1998). We shall only evaluate the exponential family models where $h \equiv 1$, as this makes the use of generalized linear methods possible. We obtain a log-linear conditional intensity since

$$\log \lambda_\theta(x;\xi) = \theta \cdot t(x;\xi).$$

Now the integral in (7.5) is approximated by

$$\int_A \lambda_\theta(x;\xi)\mathrm{d}\mu(\xi) \approx \sum_{j=1}^m \lambda_\theta(x;u_j)w_j,$$

where $u_1, \dots, u_m \in A$ and $w_1 > 0, \dots, w_m > 0$ are weights so that $\sum_{i=1}^m w_i = \mu(A)$. The simplest way to choose the weights is to let $w_i = \frac{1}{m}$. A slightly more complicated scheme is to partition A into l cells of equal size and then let $w_i = \frac{1}{le}$ where e is the number of u_j in the same cell. One could also weigh each point by the size of its Dirichlet cell w.r.t. the other points or by some other scheme.

We let the data x be a subset of $\{u_1, \dots, u_m\}$ and define $\lambda_\theta(x;\xi)$ and $t(x;\xi)$ for $\xi \in x$ as $\lambda_\theta(x\backslash\xi;\xi)$ and $t(x\backslash\xi;\xi)$. Then the log pseudo-likelihood is given by

$$\begin{aligned} pl_A(\theta;x) &= \sum_{\xi \in x_A} \log \lambda_\theta(x\backslash\xi;\xi) - \int_A \lambda_\theta(x;\xi)\mathrm{d}\mu(\xi) \\ &\approx \sum_{j=1}^m (y_j \log \lambda_j - \lambda_j) w_j, \end{aligned} \tag{7.8}$$

where $\lambda_j = \lambda_\theta(x;u_j)$ and $y_j = \frac{z_j}{w_j}$ for which $z_j = 1$ if $u_j \in x$ and $z_j = 0$ otherwise.

The expression in (7.8) formally corresponds to the log-likelihood of a generalized linear model consisting of independent Poisson variables $Y_j \sim \text{Poisson}(\lambda_j)$, except that Y_j may be non-integer, taken with the weights w_j and log as the link-function, that is $\log \lambda_j = \theta \cdot t(x; u_j)$, where $t(x; u_j)$ is interpreted as an explanatory variable. It is then possible to compute the MPLE by fitting the link-function. The problem of finding the weighted functions is overcome by the use of a statistical program such as S-PLUS. In fact, it is easily seen that (7.8) is concave, so alternatively Newton-Raphson iteration may be used to find the approximate MPLE.

Finally, $pl_A(\theta; x)$ may be found by inserting the obtained estimates in (7.8), or if the software returns the deviance D, by the equation

$$-pl_A(\theta; x) \quad \approx \quad \frac{D}{2} + x(A) + \sum_{i=1}^{m} z_i \log w_i.$$

Estimating Parameters in General

Often we have that some parameters in a particular model may be transformed into canonical parameters for an exponential family form while other parameters may not. For instance this is the case for the Strauss process where we have an exponential family form for a fixed R, as seen in Example 5.18. The following example shows how R is estimated in this case, but the example easily generalizes to other models.

Example 7.7 (Estimating Parameters for the Strauss Model)
For the Strauss process we can estimate β and γ using the MPLE for a fixed value of R. To estimate R we calculate the value of the log pseudo-likelihood function for a range of R and select the one that results in the highest value.

Let $s(x; \xi; R)$ denote $s(x; \xi)$ at range R. Investigating the log pseudo-likelihood function for the Strauss process we observe that if

$$\sum_{\xi \in x_A} s(x; \xi; R_1) \quad = \quad \sum_{\xi \in x_A} s(x; \xi; R_2)$$

for $R_1 < R_2$ then pl_A based on R_2 will be strictly larger than pl_A based on R_1, so we only need to check a finite number of values of R, namely those corresponding to inter-event distances in the observed data.

In the leftmost plot in Figure 7.1, these values of R for a particular data set (see Section 8.1) are plotted against the corresponding maximal log pseudo-likelihood values, and in the rightmost plot, we have zoomed in to observe the fact that pl_A is strictly increasing between two inter-event distances. In this particular case, the value of $R = 3.4$ results in the highest value of pl_A.

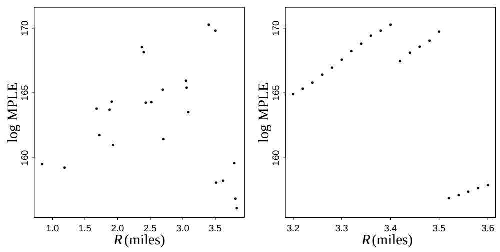

Figure 7.1: A plot of the log pseudo-likelihood values versus the interaction-radius (R), used to find the most appropriate R-value.

7.3 MCMC Maximum Likelihood Estimation

The likelihood equation $\mathbb{E}_\theta t(X) = t(x)$, see Equation (7.3), would suggest to approximate $\mathbb{E}_\theta t(X)$ by Monte Carlo methods. Instead of doing this we describe in this section a method which approximates the log likelihood function, following Geyer (1999).

Consider the density $p_\theta(\cdot)$ w.r.t. Π given as

$$p_\theta(x) = \frac{1}{c_\theta} g_\theta(x),$$

which could be on exponential family form. Further, suppose that $p(x) = \frac{1}{c} g(x)$ is another density w.r.t. Π, so that $g(x) = 0$ implies $g_\theta(x) = 0$ for all $\theta \in \Theta$. As we shall see, this facilitates the definition of a more convenient log-likelihood ratio, which can be approximated using importance sampling.

Usually $g(x)$ is chosen to be $g_\psi(x)$ for some $\psi \in \Theta$, but it could also be a mixture as for example $g(x) = \sum_{i=1}^k w_i g_{\psi_i}(x)$ with $w_i > 0$ and $\sum_{i=1}^k w_i = 1$. It is important how ψ is chosen, and an ideal choice would of course be the unknown value of the exact MLE $\hat{\theta}$ assuming this exists. Here we shall only pursue the method of importance sampling, see Section 6.5.2, where ψ should be close to θ. In the following we write $p(x) = p_\psi(x) = \frac{1}{c_\psi} g_\psi(x)$.

Now we can define a likelihood ratio

$$L(\theta) = \frac{p_\theta(x)}{p_\psi(x)} = \frac{c_\psi}{c_\theta} \frac{g_\theta(x)}{g_\psi(x)}$$

and hence the log likelihood ratio is

$$l(\theta) = \log \frac{g_\theta(x)}{g_\psi(x)} - \log \frac{c_\theta}{c_\psi}. \tag{7.9}$$

Since the first term in (7.9) consists of the un-normalised densities, it is completely known, but the second part involving the normalising constants is not known on closed form. In order to approximate the expression it is rewritten as

$$\begin{aligned}
\frac{c_\theta}{c_\psi} &= \frac{1}{c_\psi}\int_N g_\theta(x)\mathrm{d}\Pi(x) \\
&= \int_N \frac{1}{c_\psi} g_\psi(x)\frac{g_\theta(x)}{g_\psi(x)}\mathrm{d}\Pi(x) \\
&= \mathbb{E}_\psi\left[\frac{g_\theta(X)}{g_\psi(X)}\right],
\end{aligned}$$

which means that we can use MCMC methods to calculate the log-likelihood function. The MCMC approximation to (7.9) is then found by replacing the expectation by its approximation, i.e.

$$\begin{aligned}
l_k(\theta) &= \log\frac{g_\theta(x)}{g_\psi(x)} - \log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{g_\theta(x_i)}{g_\psi(x_i)}\right) \\
&\approx \log\frac{g_\theta(x)}{g_\psi(x)} - \log\mathbb{E}_\psi\left[\frac{g_\theta(X)}{g_\psi(X)}\right]
\end{aligned}$$

where $x_1, \ldots, x_k$ is a Markov chain sample from a Harris recurrent Markov chain $X_1, X_2, \ldots$ with invariant distribution specified by the density p_ψ. The MCMC maximum likelihood estimate is then found by maximizing $l_k(\theta)$.

In the following we assume that both $g_\theta(x)$ and $g_\psi(x)$ are given on exponential family form with a minimal representation, i.e. $g_\theta(x) = h(x)\exp(\theta \cdot t(x))$ and $g_\psi(x) = h(x)\exp(\psi \cdot t(x))$. Then the log-likelihood equation is

$$l_k(\theta) = (\theta - \psi)\cdot t(x) - \log\frac{1}{k}\sum_{i=1}^{k}\exp((\theta-\psi)\cdot t(x_i)),$$

and thus the derivative becomes

$$\frac{\mathrm{d}}{\mathrm{d}\theta}l_k(\theta) = t(x) - \frac{\frac{1}{k}\sum_{i=1}^{k} t(x_i)\exp((\theta-\psi)\cdot t(x_i))}{\frac{1}{k}\sum_{i=1}^{k}\exp((\theta-\psi)\cdot t(x_i))}. \tag{7.10}$$

Here we recognize the second term on the right-hand side of (7.10) as the ratio estimator given in (6.26). Thus this is a case of importance sampling with importance weights defined as

$$w_{k,\theta,\psi}(x_i) = \frac{\exp((\theta-\psi)\cdot t(x_i))}{\sum_{j=1}^{k}\exp((\theta-\psi)\cdot t(x_j))}.$$

This gives the Monte Carlo score function

$$\frac{\mathrm{d}}{\mathrm{d}\theta}l_k(\theta) = t(x) - \mathbb{E}_{k,\theta,\psi}t(X),$$

where $\mathbb{E}_{k,\theta,\psi}t(X)$ is the MCMC approximation of $\mathbb{E}_\theta t(X)$, i.e.

$$\mathbb{E}_{k,\theta,\psi}t(X) = \sum_{i=1}^{k} w_{k,\theta,\psi}(x_i)t(x_i).$$

Bear in mind that the approximation is local in the sense that it only holds in the case where ψ is close enough to $\hat{\theta}$.

The MCMC approximation of the Fisher information is the second derivative of the log-likelihood, yielding

$$-\frac{\mathrm{d}^2}{\mathrm{d}\theta\mathrm{d}\theta^\top} l_k(\theta) = \mathrm{Var}_{k,\theta,\psi}t(X), \tag{7.11}$$

by which $l_k(\theta)$ is seen to be concave.

Now the approximate MLE $\hat{\theta}_k$ is found by solving

$$t(x) = \mathbb{E}_{k,\theta,\psi}t(X),$$

using for instance Newton-Raphson exploring the fact the l_k is concave.

Finally we give a central result stating that if k is large enough and we have an exponential family model then the MCMC maximum likelihood estimate $\hat{\theta}_k$ converges to the maximum likelihood estimate $\hat{\theta}$. To do so we need the following lemma where a **relatively open** set Θ is a set which is open in the smallest affine subspace that contains Θ.

Lemma 7.8
Let Θ be a relatively open, convex subset of $\mathbb{R}^n$, and let $f_1, f_2, \ldots$ be a sequence of finite, convex functions on $\mathbb{R}^n$. If $f_1, f_2, \ldots$ converges pointwise towards a function f on a dense subset of Θ then $f_1, f_2, \ldots$ converges uniformly on any compact subset of Θ.

For a proof we refer to Rockafellar (1970).

Theorem 7.9
Suppose we have an exponential family representation where Θ is convex, let data be x, suppose the MLE $\hat{\theta}$ exists and is unique, and let

$$l_k(\theta) = \log \frac{g_\theta(x)}{g_\psi(x)} - \log\left(\frac{1}{k}\sum_{i=1}^{k}\frac{g_\theta(x_i)}{g_\psi(x_i)}\right),$$

be the Monte Carlo approximation of the log-likelihood

$$l(\theta) = \log\frac{g_\theta(x)}{g_\psi(x)} - \log\frac{c_\theta}{c_\psi}.$$

Then with probability one w.r.t. the distribution $\mathcal{D}$ of the Markov chain $X_1, X_2, \ldots$ we have that:

1. *There exists an integer $K > 0$ so that for $k > K$, $l_k(\theta)$ has a maximizer $\hat{\theta}_k \in \Theta$.*
2. $\lim_{k\to\infty} \hat{\theta}_k = \hat{\theta}$.

Proof:
We only prove the case where $\hat{\theta}$ lies in the interior of Θ but with a few modifications this proof also apply to the case where $\hat{\theta}$ lies on the boundary of Θ.

We will prove the result by showing that for any closed ball $\overline{b}(\hat{\theta}, \varepsilon) \subset \Theta$ there exists a constant K so that for any $k > K$ we have that the maximizer $\hat{\theta}_k$ of $l_k(\theta)$ exists and that $\hat{\theta}_k \in \overline{b}(\hat{\theta}, \varepsilon)$. Finally, letting $\varepsilon \to 0$ concludes the proof.

Let $\varepsilon > 0$ so that $\overline{b}(\hat{\theta}, \varepsilon) \subset \Theta$. From Section 7.1 we have that the likelihood function l is strictly concave, so

$$\exists \delta > 0 \ : \quad l(\hat{\theta} + \varepsilon u) + 2\delta \ < \ l(\hat{\theta}) \qquad \forall u \in \mathbb{R}^n \text{ with } ||u|| = 1. \tag{7.12}$$

By the SLLN (Theorem 6.15) l_k converges pointwise to l as $k \to \infty$ almost surely which by Lemma 7.8 implies that l_k converges uniformly to l on any compact subset almost surely. Hence, there exists a constant K so that for any $k > K$ and any unit vector $u \in \mathbb{R}^n$ we have that

$$\begin{aligned} ||l_k(\hat{\theta} + \varepsilon u) - l(\hat{\theta} + \varepsilon u)|| \ &< \ \delta \\ ||l_k(\hat{\theta}) - l(\hat{\theta})|| \ &< \ \delta \end{aligned}$$

which combined with (7.12) shows that $l_k(\hat{\theta}) > l_k(\hat{\theta} + \varepsilon u)$. Since l_k is concave we conclude that $\hat{\theta}_k \in \overline{b}(\hat{\theta}, \varepsilon)$ which was to be proved. ∘

7.3.1 Trust-regions

By Theorem 7.9 we have that the estimate $\hat{\theta}_k$ obtained will be close to $\hat{\theta}$ for any ψ used if k is large enough. However, we may need k to be very large making this impractical. A simple way to improve this is to let ψ be the MPLE which hopefully is not too far from $\hat{\theta}$. Complementary to this is the use of **trust-regions** as described by Geyer & Thompson (1992). The idea is to use an iterative procedure that moves the estimate in the direction of the true value of $\hat{\theta}$ restricting the length of each move to avoid "over-shooting" the target.

Trust Region Algorithm

Using an appropriate metric d on Θ, a length δ, and a constant j, the procedure is this

1. Let ψ_1 be any "appropriate" initial parameter value, for example the MPLE and let $i = 1$.

2. Repeat:

3. Generate a Markov chain sample from p_{ψ_i}.

4. Find a maximizer $\hat{\theta}_k \in \Theta_i = \Theta \cap \{\theta : d(\theta, \psi_i) < \delta\}$.

5. Let $i = i + 1$ and let $\psi_i = \hat{\theta}_k$.

6. Until ψ_i is indistinguishable from the previous j estimates $\psi_{i-j}, \ldots, \psi_{i-1}$.

7. Return ψ_i.

Certainly the notion of "indistinguishable" is vague and leaves some variation, but numerous other variations are also possible, for instance to return $\frac{1}{j+1} \sum_{k=i-j}^{i} \psi_k$ instead of ψ_i.

For applications, we will use the maximum metric

$$d((\theta_1, \ldots, \theta_m), (\psi_{i,1}, \ldots, \psi_{i,m})) \quad = \quad \max \left\{ \frac{\|\theta_l - \psi_{i,l}\|}{\delta_l} : l = 1, \ldots, m \right\}$$

and length $\delta = 1$, such that $\theta_l \in (\psi_{i,l} - \delta_l; \psi_{i,l} + \delta_l)$ for $l = 1, \ldots, m$ (see for example Section 7.4).

7.3.2 Estimating non-Canonical Parameters

The method considered so far deals mainly with estimating the canonical parameters for a model on exponential family form. In this section we will consider how to estimate any remaining parameters.

Consider again the Strauss process where $(\log \beta, \log \gamma)$ is the canonical parameter that may be estimated for a fixed value of R. Let $S(x; R)$ denote $S(x)$ at range R. From the density of the Strauss process it is easily seen that if $S(x; R_1) = S(x; R_2)$ and $R_1 < R_2$ then $p_{(\beta,\gamma,R_1)}(x) \leq p_{(\beta,\gamma,R_2)}(x)$ for any value of (β, γ) and x, so R_2 is the better estimate. This is similar to what was seen in Example 7.7 for the PLE and allows us to compare only a finite number of values of R, namely the inter-event distances $R_1, \ldots, R_q$.

The conceptually simplest way to do this is to find the approximate MLE $(\hat{\beta}_i, \hat{\gamma}_i)$ of (β, γ) for each value R_i and then estimate the normalizing constants

$c_{(\hat{\beta}_i,\hat{\gamma}_i,R_i)}$ using for instance MCMC, i.e.

$$\begin{aligned}\frac{1}{c_{(\hat{\beta}_i,\hat{\gamma}_i,R_i)}} &= \int_{\mathcal{X}} \hat{\beta}_i^{y(\mathcal{X})} \hat{\gamma}_i^{S(y;R_i)} \mathrm{d}\Pi(y) \\ &\approx \frac{1}{k} \sum_{j=1}^{k} \hat{\beta}_i^{x_{j_i}(\mathcal{X})} \hat{\gamma}^{S(x_{j_i};R_i)}\end{aligned}$$

where $x_{1_i}, \ldots, x_{k_i}$ are samples from the Strauss process with parameters $(\hat{\beta}_i, \hat{\gamma}_i, R_i)$ for $i = 1, \ldots, q$. Finally, these approximated normalizing constants are inserted in the density to find the value R_i that results in the largest likelihood.

Naturally, this method depends heavily on the error made approximating the constants and it is very time-consuming. So, often a faster method like pseudo-likelihood estimation is used to estimate the non-canonical parameters simply to save time, or to narrow it down to a few candidates that can then be checked.

7.3.3 Missing Data

In the previous sections we have described how to find the MCMC maximum likelihood estimate when full data is available. In this section we will show how to adapt these methods to the case where data is only partially observed.

Assume that X is a point process on $\mathcal{X} = W \cup V$ where $W \cap V = \emptyset$ and that we only have observed $w = x_W$ leaving $v = x_V$ as missing data. Recalling that $\Pi = \Pi_W \times \Pi_V$, we define a likelihood function

$$\begin{aligned}L(\theta) &= \frac{\int_N p_\theta(v \cup w)\Pi_V(v)}{\int_N p_\psi(v \cup w)\Pi_V(v)} \\ &= \frac{c_\theta^{-1} \int_N g_\theta(v \cup w)\Pi_V(v)}{c_\psi^{-1} \int_N g_\theta(v \cup w)\Pi_V(v)} \\ &= \frac{c_\theta^{-1} c(\theta|w)}{c_\psi^{-1} c(\psi|w)}\end{aligned}$$

where ψ is any element in Θ giving a log likelihood function of

$$l(\theta) = \log \frac{c(\theta|w)}{c(\psi|w)} - \log \frac{c_\theta}{c_\psi}.$$

This is then estimated by

$$l_k(\theta) = \log \left(\frac{1}{k} \sum_{i=1}^{k} \frac{g_\theta(v_i \cup w)}{g_\psi(v_i \cup w)} \right) - \log \left(\frac{1}{k} \sum_{i=1}^{k} \frac{g_\theta(x_i)}{g_\psi(x_i)} \right)$$

where $v_1, \dots, v_k$ are samples from the unnormalized conditional density $g_\psi(\cdot \cup w)$ w.r.t. Π_V and $x_1, \dots, x_k$ are samples from p_ψ.

If we again assume that p_θ is on exponential family form we get that the Monte Carlo score function and approximate Fisher information are given by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta} l_k(\theta) &= \mathbb{E}_{k,\theta,\psi,w} t(X_V \cup w \mid w) - \mathbb{E}_{k,\theta,\psi} t(X) \\ -\frac{\mathrm{d}^2}{\mathrm{d}\theta \mathrm{d}\theta^\top} l_k(\theta) &= -\mathrm{Var}_{k,\theta,\psi} t(X_V \cup w \mid w) + \mathrm{Var}_{k,\theta,\psi} t(X), \end{aligned}$$

where

$$\mathbb{E}_{k,\theta,\psi,w} t(X_V \cup w \mid w) = \frac{\frac{1}{k}\sum_{i=1}^k t(v_i \cup w) \exp((\theta - \psi) \cdot t(v_i \cup w))}{\frac{1}{k}\sum_{i=1}^k \exp((\theta - \psi) \cdot t(v_i \cup w))}.$$

Hence, we find the MCMCMLE $\hat{\theta}_k$ be solving

$$\mathbb{E}_{k,\theta,\psi} t(X_V \cup w \mid w) = \mathbb{E}_{k,\theta,\psi} t(X). \tag{7.13}$$

We notice in particular that l_k is not necessarily concave, so solving this equation corresponds to solving polynomials (of possibly high degree) where the powers may be real values.

Both due to this problem and due to the fact that this method is computationally expensive we propose the following modification which to the best of our knowledge has not been proposed before. Instead of solving (7.13) we propose solving

$$t(x_W) = \mathbb{E}_{k,\theta,\psi} t(X_W)$$

which only requires the unconditional samples and results in an increasing (or decreasing) polynomial. We further suggest to take edge effects into account. For example for the Strauss process we will do this by replacing $t(x_W)$ with $t_W(x)$ given by

$$t_W(x)^\top = \left(x(W), \frac{1}{2} \sum_{\xi \in W} s(x;\xi) \right)$$

which we may calculate exactly for the samples but have to estimate using edge correction for the data x_W, for instance by counting pairs of neighbours on a torus. In the following section we investigate how these estimates perform both compared to each other and compared to other types of estimation.

7.4 Comparing Inference Methods

In this section we first compare the three different ways of doing MCMC maximum likelihood estimation with missing data. We then briefly introduce a stochastic gradient algorithm for estimating parameters and then compare this to both MPLE and MCMC maximum likelihood estimation using a case of 500 samples from a particular Strauss process also considering this as a missing data problem.

To compare the three different ways of doing MCMC maximum likelihood estimation with missing data we consider a single sample from a Strauss process on a 3 times 3 area observed only on the unit square. We then run each of the three algorithms 200 times starting in the MPLE at each run and compare the results. In each run we use 2500 conditional samples and 2500 unconditional samples for solving (7.13) and 2500 unconditional samples for each of the other methods. All samples are made using a Metropolis-Hastings sampler with a burn-in of 20000 and a spacing of 20. Hence, the first method takes approximately twice the time, of the other two methods, to run. The results are summarized in Table 7.1.

	$\mathbb{E}_{k,\theta,\psi} t(X_V \cup w \mid w)$	$t(x_W)$	$t_W(x)$
$\mathbb{E} \log \hat{\beta}$	4.92	4.98	4.98
Var $\log \hat{\beta}$	0.016	0.001	0.001
$\mathbb{E} \log \hat{\gamma}$	-1.11	-1.09	-1.05
Var $\log \hat{\gamma}$	0.021	0.001	0.001

Table 7.1: Parameter estimates using three different methods of MCMCMLE with missing data.

Based on these results we conclude that in this particular case the two last methods perform equally well and somewhat better than the first method due to their much lower variance. We hence continue using the method with $t_W(x)$ in the sequel, but could also have used $t(x_W)$.

Stochastic Gradient Algorithm

The **stochastic gradient algorithm** as presented here is a slight modification of an algorithm presented in Younes (1988) and assumes that we have a density on exponential family form.

Let x be the observed data, m be some positive integer, and $c = (c_1, \ldots, c_n)^\top$ be a vector of positive (small) constants.

1. let $\theta_0 \in \Theta$ be an appropriate initial value, let $x_0 = x$, and let $i = 1$
2. let x_i be the output of some sampler with $p_{\theta_{i-1}}$ as equilibrium density after running the sampler for a (short) number of steps starting in state x_{i-1}

3. let $\theta_i = \theta_{i-1} + i^{-\frac{1}{2}} \operatorname{diag}\left(c \cdot (t(x) - t(x_i))^\top\right)$
4. let $i := i + 1$
5. if $i \leq m$ then repeat from step 2
6. return θ_m,

where diag is the diagonal of the matrix.

The idea behind this algorithm is quite intuitive. It simply tries to adjust the parameters in the "right" direction using finer and finer corrections. We refer to Younes (1988) and Moyeed & Baddeley (1991) for a more thorough investigation of the algorithm.

Comparing MCMCMLE, MPLE and the Gradient Estimate

To compare the three methods of estimation, pseudo-likelihood, MCMC maximum likelihood and the stochastic gradient algorithm, we make 500 samples from a Strauss process with $\beta = 200$, $\gamma = 0.2$ and $R = 0.1$ on a 3 times 3 area retaining only the data within the unit square and estimate $\log \beta$ and $\log \gamma$ for each of these samples leaving R fixed at 0.1. Both the MCMCMLE and gradient algorithms are started in the MPLE in order to make them more efficient. For the gradient algorithm $m = 250$, $c = (1, 0.01)^\top$ and a Metropolis-Hastings algorithm run for 500 iterations is used. For the MCMCMLE a trust region of $\delta_\beta = 20$ and $\delta_\gamma = 0.3$ is used letting the algorithm in Section 7.3.1 run for 10 iterations with $j = 1$ using 2500 samples from a Metropolis-Hastings algorithm with a burn-in of 20000 and a spacing of 20 in each iteration.

The result can be seen in Figure 7.2 and Table 7.2. MCMCMLE has a mean that is closer to the true value than MPLE and the gradient estimate is clearly the worst of the three. Further, the variance of MCMCMLE is smaller than the variance of the gradient estimate and here the MPLE performs the worst. The time used to run the algorithms was approximately 1 hour for the pseudo-likelihoods, 10 hours for the gradient algorithm and 60 hours for the MCMC maximum likelihoods, i.e. 12 minutes per sample for MCMCMLE. The main result obtained from this comparison is that the PLE performs remarkably well even compared to the MCMCMLE that is approximately 100 times slower, at least in this particular case.

7.5 Contrast Methods

In situations where the likelihood or pseudo-likelihood function cannot be determined, contrast methods can be used for parametric inference. Assume X is

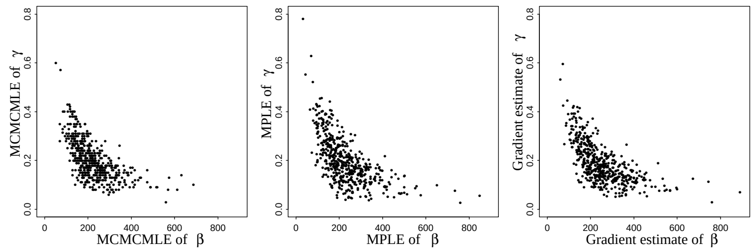

Figure 7.2: Parameter estimates for 500 samples from a Strauss process with $\beta = 200, \gamma = 0.2$ and $R = 0.1$. Only β and γ has been estimated with R fixed at 0.1.

	Parameter	MCMCMLE	MPLE	Gradient
$\mathbb{E} \log \hat{\beta}$	5.30	5.34	5.35	5.46
Var $\log \hat{\beta}$	-	0.141	0.186	0.159
Median $\log \hat{\beta}$	5.30	5.33	5.35	5.47
$\mathbb{E} \log \hat{\gamma}$	-1.61	-1.68	-1.72	-1.76
Var $\log \hat{\gamma}$	-	0.163	0.267	0.195
Median $\log \hat{\gamma}$	-1.61	-1.66	-1.69	-1.75

Table 7.2: Parameter estimates using three different methods of estimation on 500 samples from a Strauss process with $\log \beta = 5.30, \log \gamma = -1.6$ and $R = 0.1$. Only $\log \beta$ and $\log \gamma$ has been estimated with R fixed at 0.1.

a stationary and isotropic point process observed in some window $W \in \mathcal{B}_0$ with $|W| > 0$, and let the model for X depend on the parameter vector $\theta \in \Theta \subseteq \mathbb{R}^n$. Let $t(a;\theta) : \mathbb{R}_+ \times \Theta \to \mathbb{R}$ be a statistic for that model, and let $\hat{t}(a)$ be the estimate of $t(a;\theta)$. Define the discrepancy $D(\theta)$ between the model and data as

$$D(\theta) = \int_{\varepsilon}^{a_0} \left\{ \hat{t}(a)^{\alpha} - t(a;\theta)^{\alpha} \right\}^2 \mathrm{d}a, \tag{7.14}$$

where $0 \leq \varepsilon < a_0$ and $\alpha > 0$ are user specified constants. The estimate of θ is then determined by the value of θ that minimizes $D(\theta)$. The roles of ε, a_0 and α and how to choose them are described by Diggle (1983), Stoyan & Stoyan (1994), and Møller et al. (1998).

When dealing with a cluster process (see Chapter 5), we do not know the density on a closed form. In such situation we could use contrast methods to estimate the parameters. Consider for example the Thomas process (Example 5.12), where the theoretical K- and g-functions are known. Thus, minimizing

$$D(\lambda_Y, \sigma) = \int_{\varepsilon}^{a_0} \left\{ \hat{K}(r)^{\alpha} - K(r;\lambda_Y,\sigma)^{\alpha} \right\}^2 \mathrm{d}r, \tag{7.15}$$

we find the parameters λ_Y and σ that fits the data best. In praxis we replace the integral by a sum.

Sometimes we do not even know the statistic $t(a;\theta)$ on closed form. In these cases, we may approximate it by simulating from the model, using the obtained Monte Carlo mean instead. In Section 8.2 we apply this method to fit cluster processes.

We illustrate the use of this method by considering a log Gaussian Cox process as described in Møller et al. (1998) (see also Section 5.1.1).

Assume X is a LGCP observed in a window $W \in \mathcal{B}_0$, with $|W| > 0$. Let Y denote the underlying stationary and isotropic Gaussian field (see Section 5.1.1) with mean μ and covariance function given by

$$c(a, \sigma^2, \beta) \quad = \quad \mathrm{Cov}(Y(s_1), Y(s_2)) \quad = \quad \sigma^2 r_\beta(a), \tag{7.16}$$

where $a = \|s_1 - s_2\|$ is the distance between two points s_1 and s_2, σ^2 is the variance, and $r_\beta(\cdot)$ is a correlation-function parameterized by $\beta > 0$, for example $r_\beta(\alpha) = r(\alpha/\beta)$ where r is a known correlation function. The functional form of $r_\beta(\cdot)$ may be determined by plotting a non-parametric estimate for $c(\cdot)$ which we derive in the following.

Recall that $X \mid Y$ is a Poisson point process with intensity function $\exp(Y(\cdot))$, while the intensity and pair correlation function for X are

$$\lambda = \exp\left(\mu + \frac{\sigma^2}{2}\right) \quad \text{and} \quad g(a) = \exp(c(a)),$$

where μ and σ^2 are the mean and variance of the Gaussian surface respectively. Let $\hat{g}(a)$ be a non-parametric estimate of the pair correlation function $g(a)$ as described in Section 4.3. An obvious estimator $\hat{c}(\cdot)$ of $c(a)$ is hence $\hat{c}(a) = \log \hat{g}(a)$. Thus for the LGCP the parameters we need to estimate are $(\mu, \sigma^2, \beta) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. If we have observed $X \cap W = \{x_1, \ldots, x_n\}$, a natural estimator of λ is

$$\hat{\lambda} = \frac{n}{|W|}.$$

Suppose we have an estimate $\hat{\sigma}^2$ of σ^2. Using the method of moments, we find $\hat{\mu}$ by

$$\frac{n}{|W|} = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right),$$

which results in the estimate $\hat{\mu} = \log \hat{\lambda} - \frac{\hat{\sigma}^2}{2}$.

Consequently it only remains to estimate β and σ^2 which is done using the minimum contrast method described above. Letting $\theta = (\sigma^2, \beta)$ and $t(a;\theta) = c(a;\theta)$, then according to (7.14) we have to minimize

$$\int_\varepsilon^{a_0} \left\{\hat{c}(a)^\alpha - c(a;\theta)^\alpha\right\}^2 \mathrm{d}a. \tag{7.17}$$

The solution is found by inserting (7.16) in (7.17) giving a polynomial of degree 2 in $\sigma^{2\alpha}$. Hence, $\hat{\beta}$ is found by maximizing $\frac{B(\beta)^2}{A(\beta)}$ w.r.t. β, where

$$B(\beta) = \int_{\varepsilon}^{a_0} \{\hat{c}(a) r_\beta(a)\}^\alpha \, \mathrm{d}a$$

and

$$A(\beta) = \int_{\varepsilon}^{a_0} r_\beta(a)^{2\alpha} \mathrm{d}a,$$

and if we obtain a solution $\hat{\beta}$, then

$$\hat{\sigma}^2 = \left[\frac{B(\hat{\beta})}{A(\hat{\beta})}\right]^{\frac{1}{\alpha}}.$$

7.6 Parametric Bootstrap and Hypothesis Tests

In this section we consider how to find the distribution of a parameter estimate to make a confidence region, and how to make tests of hypotheses expressing that the data follows some type of model given that the data is assumed to follow a less restricted type of model, i.e. that the data is a realisation of a Poisson point process given that it is assumed to be a realisation of a Strauss process.

7.6.1 Parametric Bootstrap

Regardless of which method is used to find an estimate of the parameters it would also be of interest to find the distribution of this estimate in order to make a confidence interval or region. In some specific cases it may be appropriate to assume a particular distribution, but more often we will resort to simulations, using a method called **parametric bootstrap**. Let $\hat{\theta}$ be the parameter estimate, simulate a number of point processes $x_1, \ldots, x_n$ from $p_{\hat{\theta}}$ and find estimates $\hat{\theta}_1, \ldots, \hat{\theta}_n$ based on these realisations. These estimates, see Figure 7.2 for an example, are then used as an empirical distribution of the parameter estimate which can be used to make confidence intervals or regions.

7.6.2 Likelihood Ratio Test

Assuming we have accepted that data x follows some type of model, for example a Strauss process, it may be of interest to test a hypothesis concerning specific constrains on the parameters, for example that $\gamma = 1$ corresponding to a Poisson point process. A method very similar to the likelihood ratio test for maximum likelihood estimates may then be used. Let $\hat{\theta}$ be the estimate under the full model and $\hat{\theta}_0$ be the estimate under the hypothesis. Assuming that these estimates are MPLE's we then accept the hypothesis if

$$\frac{\mathrm{PL}_A(\hat{\theta}_0; x)}{\mathrm{PL}_A(\hat{\theta}; x)} > c$$

where c is some constant between 0 and 1. To determine the value of c associated with a specific significance level we typically simulate a number of samples from $p_{\hat{\theta}}$ to get an empirical distribution of the ratio under the hypothesis. Similarly for MCMCMLE where we accept the hypothesis if

$$\frac{L(\hat{\theta}_0; x)}{L(\hat{\theta}; x)} = \frac{c_{\hat{\theta}} g_{\hat{\theta}_0}}{c_{\hat{\theta}_0} g_{\hat{\theta}}} = \frac{c_{\hat{\theta}}}{c_{\psi_1}} \frac{c_{\psi_1}}{c_{\psi_2}} \cdots \frac{c_{\psi_{k-1}}}{c_{\psi_k}} \frac{c_{\psi_k}}{c_{\theta_0}} \frac{g_{\hat{\theta}_0}}{g_{\hat{\theta}}} > c$$

where ψ_1 is close to $\hat{\theta}$, ψ_{i-1} is close to ψ_i for $i = 2, \dots k$ and ψ_k is close to $\hat{\theta}_0$ so that importance sampling may be used to estimate the normalizing constants. It is then hoped that the error in each of these estimates cancel each other rather than accumulate.

7.6.3 Fisher-Wald Test

Another test is the **Fisher-Wald test** which is applicable for MCMC maximum likelihood estimates.

Let $\hat{\theta}_k$ be the MCMCMLE, let $I(\theta)$ denote the Fisher information, and let $I(\hat{\theta})$ denote its MCMC approximation. Suppose, that

$$\hat{\theta}_k \approx N_n(\theta, I(\theta)^{-1}).$$

According to Geyer (1999) it is still an open research question if this is the case but in the sequel we assume this.

Let $\theta^\top = (\theta_1^\top, \theta_2^\top)$ where θ_1 is p dimensional and θ_2 is $n-p$ dimensional. If we now wish to test if $\theta_1 = v$ where v is a given vector and $(I(\theta)^{-1})_{11}$ denotes the upper left p times p sub-matrix of the Fisher information then we may use the fact that

$$(\hat{\theta}_{k_1} - \theta_1)^\top ((I(\theta)^{-1})_{11})^{-1} (\hat{\theta}_{k_1} - \theta_1) \approx \chi_p^2$$

inserting $I(\hat{\theta})$ obtained by (7.11) instead of the Fisher information and v instead of θ_1 to obtain a value that may be compared to the chi-square distribution.

Chapter 8

Analyses of Spatial Point Patterns

In this chapter we present analyses of six point patterns. The main reason for carrying out these analyses is to examine how to apply the theory in practice, so the results are somewhat secondary. Hence, we have aimed at using as much of the theory presented as possible.

Most calculations are made using the statistical programming language S-Plus but to speed up computation, some of the algorithms have been implemented in C. Unless otherwise stated, these algorithms for analysing the data have been implemented by the authors, using the theory of the previous chapters. Certain algorithms are documented in Appendix A, and the rest may be required by sending an e-mail to hedegard@@bigfoot.com, but it should be noted that documentation is sparse or non-existing.

First a widely analysed spatial point pattern, the Spanish towns which stem from Glass & Tobler (1971), is analysed. It is a standard point pattern in the S-Plus library spatial. We apply a Strauss, a two-step multi-scale, and a soft-core model to this dataset. This analysis will be documented in a very detailed manner, while we present the later five more briefly.

Secondly we analyse a dataset containing positions of short-leaf pines kindly supplied by Dr. Jorge Mateu, Dept. of Mathematics, University of Jaume I, Spain. Here we use contrast methods for Matérn cluster and the Thomas process.

Next is a dataset containing biological cells (Crick & Lawrence 1975), where an attempt to fit an SSI process, a Matérn hard-core type II, a multi-scale, a hybrid soft/hard-core process, and Ord's process is made.

Swedish pines from Venables & Ripley (1994) is then analyzed and the Strauss, the multi-scale, and Ord's process are applied to fit the data.

Further analysed is a dataset where the positions of trees of the Norwegian spruce specie are given. To this we have tried to fit a Strauss, a two-step-, and a three-step multi-scale model.

The last analysis is of a marked dataset, containing the positions of blighted and living wild bamboos. Separately the two types can be described by CSR, but no model has been fitted to the entire data – only a suggestion to an extension of the bivariate Cox process is made.

These last two datasets are also supplied by Dr. Jorge Mateu.

We have used much of the theory described in Chapters 2 to 7 and in most cases it has been possible for us to fit a model, and hence these analyses can serve as a foundation for the practical aspects of statistical inference in spatial point processes.

8.1 Spanish Towns

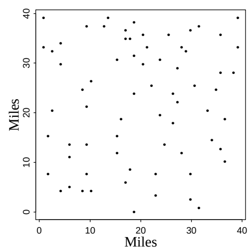

Figure 8.1: Map of the center of 69 Spanish towns within an area of 40×40 miles.

The Spanish towns data, which can be seen in Figure 8.1, comes from Glass & Tobler (1971) and shows the locations of the centres of 69 Spanish towns of unknown sizes, positioned at latitude $^o2\ '30\ W$ and longitude $^o39\ '47\ N$ and distributed within a 40×40 miles area. The area is assumed to be homogeneous in climate, physiography, transport, population density, and economy.

Analyses of this dataset can be seen in Glass & Tobler (1971), Ripley (1977), Ripley (1979), Ripley (1988), and Ogata & Tanemura (1984). Ripley (1977) found that the dataset can be described by a Strauss process. Using the K-function he finds the interaction radius to be 3.5 miles, and further that $\gamma = 0.5$.

Heuristically this model also seems like a good choice, because one would expect some sort of repulsion between the towns. But despite this and the fact that the model is well documented, we start our analysis of the Spanish towns by considering whether the point pattern can be described by complete spatial randomness or not.

8.1.1 Test for Poisson Distribution

Checking for CSR we use the summary statistics described in Chapter 4. To each summary statistic, we calculate envelopes of 99 simulated Poisson point processes. Figure 8.2 shows plots of the various summary statistics for the Spanish towns data (dots) together with the envelopes for the simulated Poisson point processes (full lines). Though it may seem irrelevant to plot both the K and the L function, we have done so for completeness sake.

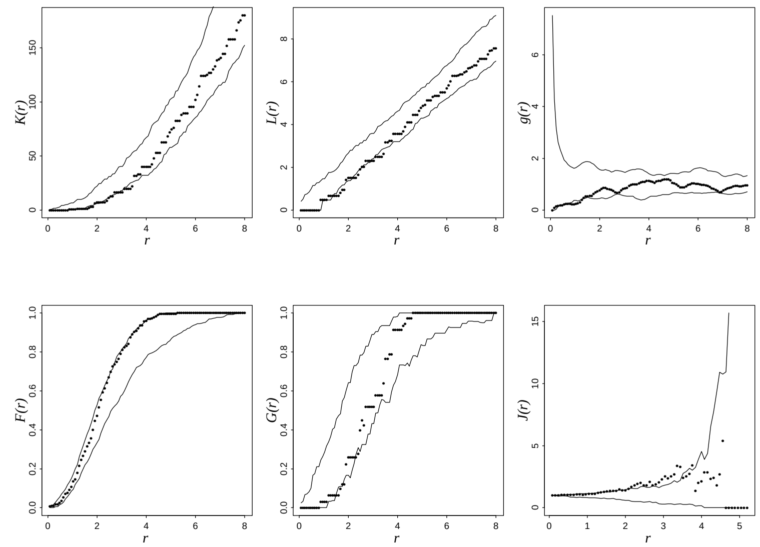

Figure 8.2: The summary statistics for the Spanish towns data, plotted with envelopes for 99 simulated Poisson processes.

Since 19% of the data is outside the envelopes for the K/L-function, and there are 9%, 6%, 6%, and 40% outside the envelopes for the g, F, G, and J-function respectively, it is clear that the data cannot be described by a Poisson point process. Expecting only 2% of the data to be outside the envelopes, we reject the hypothesis that the data comes from a Poisson point process.

Using Figure 4.5 in Section 4.8, we see that the data might come from a process which models regularity to some extend.

8.1.2 Fitting a Strauss Model

When fitting a model, we assume that the data not only follows the model on the observed area, but rather on a larger area. We thus have to deal with a missing data problem, where we assume data follows a Strauss distribution on a 120×120 area, from which we only observe the 40×40 square in the middle.

We first try to fit the data to the Strauss model described in Section 5.4, which has density on the form

$$p(x) \quad \propto \quad \beta^{x(\mathcal{X})}\gamma^{S(x)}.$$

In order to specify the model we need to estimate the parameters β and γ. Further, we have to take care of the interaction radius R which is a parameter in $S(x)$. In the following we use the pseudo-likelihood estimation approach, which is described in Section 7.3.1.

Pseudo-likelihood Estimation

Below we estimate the parameters β and γ by use of pseudo-likelihood estimation. We use the same approach as in Example 7.7 for estimating the interaction radius (in fact Example 7.7 is based on the Spanish towns data). Thus in the sequel we use the value of R that maximizes the log-pseudo-likelihood which in the example was found to be $\hat{R} \approx 3.4$ which is close to be in agreement with Ripley (1977).

However, to take care of edge-effects, there are two ways to estimate the parameters – either to use minus-sampling or to use the torus convention. The first approach yields the estimates $\hat{\beta}_m = 0.150$ and $\hat{\gamma}_m = 0.38$, where the m indicates that minus-sampling has been used, while the latter gives $\hat{\beta}_t = 0.112$ and $\hat{\gamma}_t = 0.42$

Hence, we have two sets of parameters, which should describe the data to some extend. We thus proceed with model-checking.

Model Checking

We simulate 99 Strauss processes with parameters given as above, and calculate envelopes for the summary statistics, which are to be compared with the summary statistics for the Spanish towns data.

Thus we simulate a Markov chain for a Strauss process with parameters $(\hat{\beta}_m, \hat{\gamma}_m)$ and $(\hat{\beta}_t, \hat{\gamma}_t)$ respectively and interaction radius $\hat{R}$. This yields the envelopes shown in Figures 8.3 and 8.4.

For the estimates found using minus-sampling, we find that 2%, 4%, 1%, 0%, and 2% of the data are outside the envelopes for the K/L, g, F, G, and J functions respectively. These values are acceptable, and hence we choose to accept the model with these parameters.

However, the model where the parameters has been estimated on a torus, seems to approximate the data even better, since for the different summary statistics, we here only have 1% of the data outside the envelopes for every summary statistic.

Even though we have just accepted the Strauss model with parameters $(\hat{\beta}_t, \hat{\gamma}_t, \hat{R})$ for the Spanish towns, we continue the analysis in the search for better models and estimates.

MCMC Maximum Likelihood Estimation

We now do MCMC maximum likelihood estimation in order to see if it is possible to make an even better fit of the parameters in the Strauss model.

Dealing with a missing data problem, we have to take care of edge-effects, so applying the method described in Section 7.3.3 we calculate the number of pairs of neighbours by wrapping the data around a torus and find that $S(x) = 30$, taking $R = \hat{R}$.

We use the approach described in Section 7.3.1 to estimate the parameters where we start by using $\psi_1 = (0.188, 0.3)$. To make one iteration of the procedure, a burn-in of 20.000 samples is made, and afterwards every 20th sample is used for the importance sampling, taking a total of 2.500 samples. In Figure 8.5 we see values of β and γ plotted against the number of iterations of the procedure. Note that the reason why the γ- (and also the β-) values lie on lines is due to rounding off. Though we could use the pseudo-likelihood estimates, we use $\beta_0 = 0.188$ and $\gamma_0 = 0.3$ to illustrate the effect of using trust-regions. The leftmost plots has trust-regions $\delta_\beta = 0.031$ and $\delta_\gamma = 0.3$ (which for this particular situation is equal to not using trust-regions, since the parameters are not changing that much), whereas the two rightmost plots has $\delta_\beta = 0.003$ and $\delta_\gamma = 0.03$.

As can be seen in all four plots, the procedure yields parameters $\hat{\beta}_{MLE} \approx 0.115$ and $\hat{\gamma}_{MLE} \approx 0.44$. Also, we see that it necessary to discard a number of iterations, since the initial parameter is far from the MCMCMLE. For the leftmost plots, the procedure yields values close to the MCMCMLE relatively quickly (6 iterations), in contrast to the rightmost plots, where it takes around 30 iterations before the parameters are stabilized. However, when 'equilibrium' is reached for the rightmost plots, there is very little variance, as opposed to the leftmost plots.

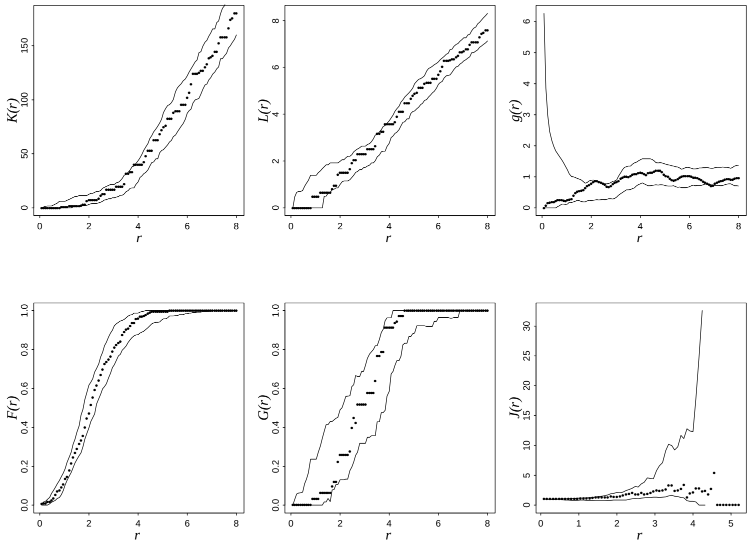

Figure 8.3: Envelopes for a Strauss process with parameters $(\hat{\beta}_m, \hat{\gamma}_m, \hat{R})$. Summary statistics for the Spanish towns data are shown as dots.

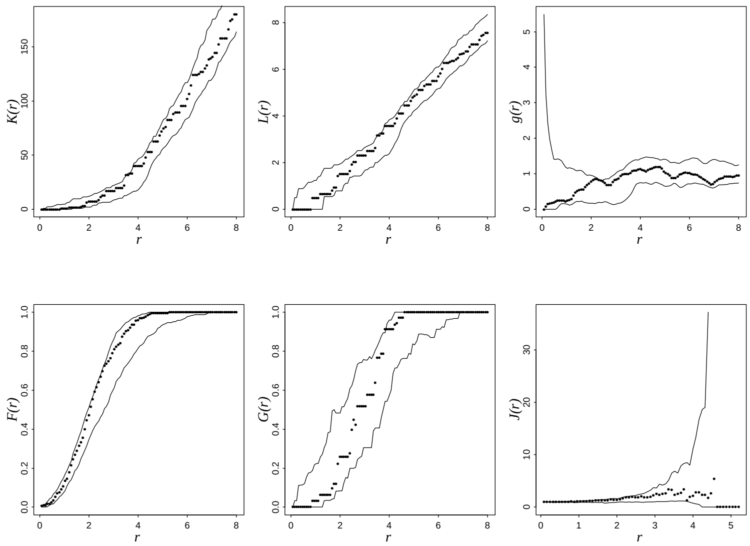

Figure 8.4: Envelopes for a Strauss process with parameters $(\hat{\beta}_t, \hat{\gamma}_t, \hat{R})$. Summary statistics for the Spanish towns data are shown as dots.

In Figure 8.6, we have used the same procedure, but this time the initial parameters are $\hat{\beta}_t$ and $\hat{\gamma}_t$ which were found as the pseudo likelihood estimates. The leftmost plots have been made with trust-regions $\delta_\beta = 0.031$ and $\delta_\gamma = 0.3$ and for the rightmost, $\delta_\beta = 0.003$ and $\delta_\gamma = 0.03$. Again we see that $\beta \approx 0.115$ and $\gamma \approx 0.44$.

In Table 8.1, we summarize what can be seen from Figures 8.5 and 8.6. We

	Figure 8.5 $\beta = 0.188$, $\gamma = 0.3$		Figure 8.6 $\beta = 0.112, \gamma = 0.44$	
Trust-regions	$\delta_\beta = 0.031$ $\delta_\gamma = 0.3$	$\delta_\beta = 0.003$ $\delta_\gamma = 0.03$	$\delta_\beta = 0.031$ $\delta_\gamma = 0.3$	$\delta_\beta = 0.003$ $\delta_\gamma = 0.03$
$\mathbb{E}\hat{\beta}$	0.115	0.115	0.115	0.115
$\mathbb{E}\hat{\gamma}$	0.439	0.439	0.440	0.442
Var$\hat{\beta}$	0.0000285	0.0000069	0.0000251	0.0000054
Var$\hat{\gamma}$	0.000339	0.000101	0.000277	0.000101
Number of iterations	200	200	200	736
Number of iterations before use	6	32	0	0

Table 8.1: Values of β and γ and variance of these.

see that $\mathbb{E}\hat{\beta} = 0.115$ for the first figure. This value has been found using only the sample values after the 6th and the 32nd iterations respectively, to avoid bias from the initial parameter. For this figure, we also see that $\mathbb{E}\hat{\gamma}$ for the left as well as for right plot yields 0.439. What is significant between these two columns, is the variance of the estimated values. For the leftmost plots, the variance is roughly 3.5 times as large. This is of course due to the fact that the trust-regions are larger for the leftmost plots — something that also applies for Figure 8.6. For this figure, the value of $\mathbb{E}\hat{\beta}$ differs slightly, as does the $\mathbb{E}\hat{\gamma}$-value. However, as the value of $\mathbb{E}\hat{\beta}$ and $\mathbb{E}\hat{\gamma}$ has been calculated from 736 iterations, we use them as our MCMCMLE. Thus $\hat{\beta}_{MLE} = 0.115$ and $\hat{\gamma}_{MLE} = 0.44$.

The estimation has been done several times with different start values, and also with different sample sizes. The result has always been values similarly to those mentioned as the mean values, and hence we settle on the Strauss model with parameters $\hat{R} = 3.4$, $\hat{\beta} = 0.115$, and $\hat{\gamma} = 0.44$.

Model Checking

We compare the summary statistics for the data with envelopes found by simulating Strauss processes with parameters $(\beta, \gamma, R) = (0.115, 0.44, 3.4)$

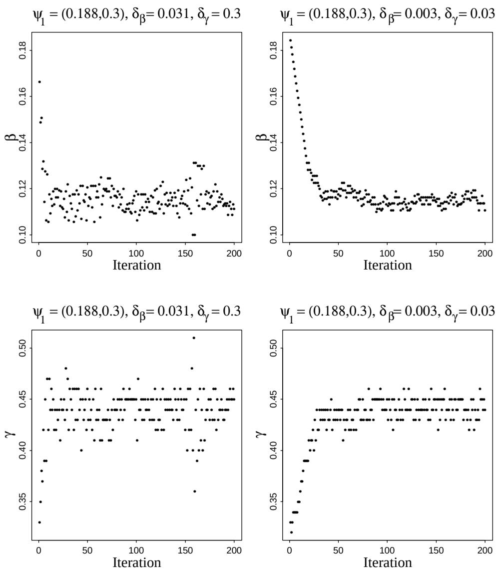

Figure 8.5: The resulting values of β and γ after the use of MCMC maximum likelihood estimation, starting with $\beta_0 = 0.188$ and $\gamma_0 = 0.3$. It is notable that the values are centered around $\beta = 0.115$ and $\gamma = 0.44$.

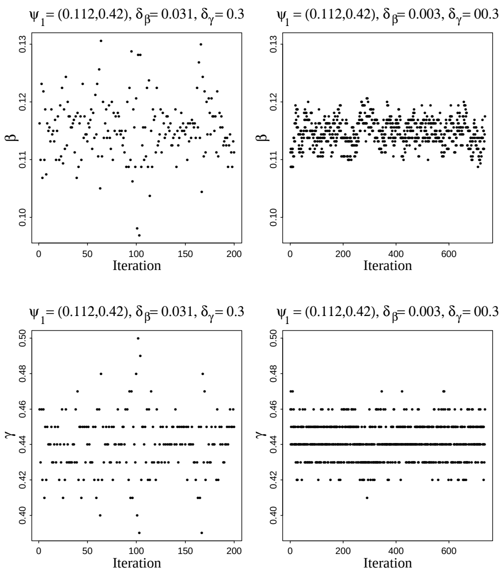

Figure 8.6: The resulting values of γ and β after using MCMC maximum likelihood estimation, starting with $\beta = \beta_t = 0.112$ and $\gamma = \gamma_t = 0.44$. Note that the values are scattered around $\gamma = 0.44$ and $\beta = 0.115$.

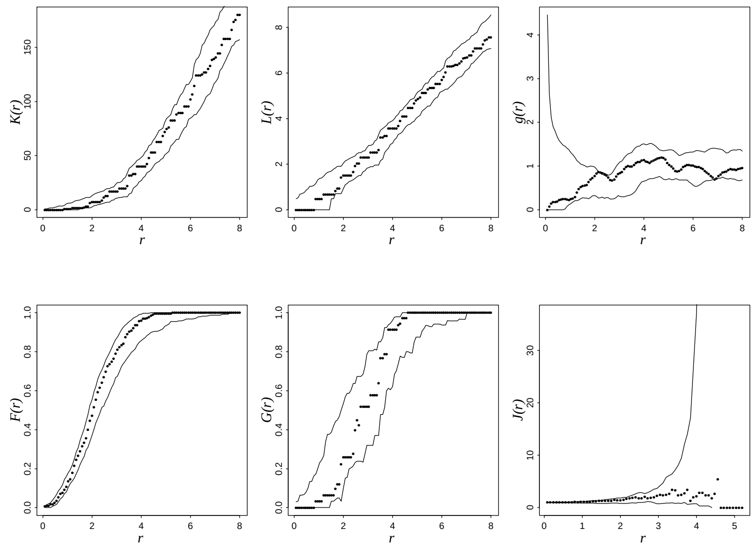

Figure 8.7: Summary statistics for the Strauss model with estimated parameter values calculated with MCMC maximum likelihood estimation. The dotted line is for the Spanish towns data, and the full lines are the envelopes for the simulated Strauss processes.

Figure 8.7 shows a plot of these statistics. Here, only the g-function has 1% outside the envelopes, while there is 0% outside for the rest of the plots.

Conclusion of Strauss Model Fitting

Comparing the two Strauss models we have now estimated, it is hard to favor one for the other. The only notable difference is that the method using MCMC maximum likelihood estimation is far more computationally expensive than pseudo-likelihood estimation. Trying to find a difference in output for the two methods, we take a look at the average number of points and neighbours for the Strauss models. These values are to be compared with the corresponding values for the Spanish towns data, which are 69 points and 30 neighbours. The pseudo-likelihood method gives with 100.000 samples a mean number of points of 71, and the corresponding value of neighbours is 30, whereas the MCMC maximum likelihood method gives respectively 70 points and 31 neighbours. Hence, there is no significant difference in output between the two methods.

8.1.3 Fitting a Multi-scale Model

We now consider the possibility that the Spanish towns data can be described even better by a multi-scale process.

Pseudo-likelihood Estimation

For fixed values of R_1 we find the best value of R_2 and *vice-versa*. Among these, the pseudo-likelihood is maximized for $R_1 = 1.68$ and $R_2 = 3.4$. Another R_1 value for which the pseudo-likelihood is large, is $R_1 = 2.36$.

With $R_1 = 1.68$ and $R_2 = 3.4$, the maximum pseudo likelihood estimates become $\hat{\beta} = 0.146, \hat{\gamma}_1 = 0.13$ and $\hat{\gamma}_2 = 0.47$.

Simulating multi-scale-processes with these parameters, provides us with a set of envelopes, for which 2% of the data is outside for the g-function. For the other statistics, the data is inside the envelopes. Hence we accept this model.

Letting $R_1 = 2.36$ and $R_2 = 3.4$ yield the estimates $\hat{\beta} = 0.146, \hat{\gamma}_1 = 0.25$, and $\hat{\gamma}_2 = 0.50$ for which, envelopes of 99 simulations has $1\%, 3\%, 1\%, 0\%$, and 1% of the data are outside. This is also acceptable, and we hence have two sets of parameters which describe the data using a multi-scale model.

It is hard to favor one over the other, but it seems like the former is best, if referring to the summary statistics alone. We thus consider this as our MPLE.

MCMC Maximum Likelihood Estimation

We let the initial parameter for the MCMCMLE procedure (see Section 7.3.1) be $\beta = 0.146, \gamma_1 = 0.13$, and $\gamma_2 = 0.47$, with $R_1 = 1.68$ and $R_2 = 3.4$, letting the trust-regions be $\delta_\beta = 0.003, \delta_{\gamma_1} = 0.03$, and $\delta_{\gamma_2} = 0.03$. We discard 20.000 samples as burn-in and use a spacing of 20 samples to draw 1000 samples for the estimation. With $j = 500$ we obtain $\mathbb{E}\hat{\beta} = 0.127, \mathbb{E}\hat{\gamma}_1 = 0.18$ and $\mathbb{E}\hat{\gamma}_2 = 0.52$ and with these estimates, envelopes for 99 simulated multi-scale processes, shows that $1\%, 0\%, 1\%, 0\%$, and 1% of the data is outside the envelopes.

Doing the same with $R_1 = 2.36, R_2 = 3.4$, and initial parameters $\beta = 0.146, \gamma_1 = 0.25, \gamma_2 = 0.5$, yields $\mathbb{E}\hat{\beta} = 0.119, \mathbb{E}\hat{\gamma}_1 = 0.71, \mathbb{E}\hat{\gamma}_2 = 0.5$, but for these parameters, $9\%, 6\%, 3\%, 7\%$ and 2% of the data is outside the envelopes, and hence we reject this model.

Conclusion of the Multi-Scale Model Fitting

Drawing 10.000 samples from each of the above models, we find for the MPLE $\mathbb{E}x(\mathcal{X}) = 69.7$, $\mathbb{E}S(x; R_1) = 1.4$, and $\mathbb{E}S(x; R_2) = 28.7$ and for the MCMCMLE are $\mathbb{E}x(\mathcal{X}) = 68.8$, $\mathbb{E}S(x; R_1) = 1.9$, and $\mathbb{E}S(x; R_2) = 29.9$. Comparing these with $x(\mathcal{X}) = 69, S(x; R_1) = 2, S(x; R_2) = 30$ for the Spanish towns data, we conclude that the MCMC maximum likelihood method, returning parameters closer to the real data, gives better estimates. So we let $\beta = 0.127, \gamma_1 = 0.18$, and $\gamma_2 = 0.52$ be the estimated parameters for the multi-scale model.

8.1.4 Fitting the Soft-core model

As a final model for the Spanish towns we consider the soft-core model, which is described in Section 5.4. The most obvious distinction between a multi-scale model and the soft-core model is that in the former we have all pairs of events being neighbours.

In analysing this model we restrict attention to the pseudo-likelihood estimation, since the parameters obtained in the previous analyses has been almost as good as those found by using MCMCMLE.

The value of σ, which maximizes the pseudo-likelihood function is found in a manner similar to that of finding R for the Strauss process. We settle on the value $\hat{\sigma} = 2.4$, and find that $\hat{\beta} = 0.104$ is the MPLE. In order to check the model with these parameters, we once again simulate 99 processes and calculate envelopes.

Here, only 1% of the data is outside the envelopes for the F- and the J-function, and thus we accept this model as well.

8.1.5 Summary

For the Spanish towns data, we have found three models which describe the data in an appropriate way

- **the Strauss process** with parameters $\beta = 0.115, \gamma = 0.44$ and $R = 3.4$.
- **the multi-scale process** with parameters $\beta = 0.127, \gamma_1 = 0.18, \gamma_2 = 0.52, R_1 = 1.68$, and $R_2 = 3.4$.
- **the Soft-core process** with parameters $\sigma = 2.4$ and $\beta = 0.104$.

8.2 Short-leaf Pines

This dataset (shown in Figure 8.8) consists of the locations of 127 short-leaf pines in an area of 290 by 387 metres.

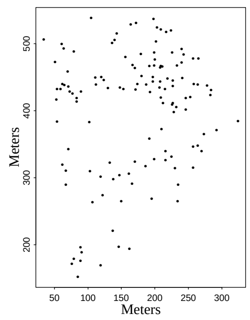

Figure 8.8: Positions of short-leaf pines in a 290×387 meters area.

Testing if the data can be described by a Poisson point process, it is apparent by Figures 8.8 and 8.9 that the data could follow a cluster model if we assume stationarity. Alternatively we could consider an inhomogeneous process, but with no data available to describe the inhomogeneity we refrain from doing so.

Estimating the Matérn Cluster Process

An appropriate choice of model could be the Matérn cluster process which is described in Example 5.11. In order to specify the model, we have to estimate the intensity of the mother process λ_Y, the number of daughter points to be distributed, and the radius R within which daughter points are distributed. To deal with the missing data problem, we use an area that is enlarged by $2R$ in both x and y, so that we assume the data to follow a Matérn cluster process on $W_{\oplus R}$ and not on W only.

Here we use contrast methods, where we approximate the discrepancy (see Section 7.5) by using the summary statistic F. Thus we wish to minimize

$$D(\lambda_Y, R) \quad = \quad \int_{\varepsilon}^{r_u} (\hat{F}(r) - \overline{F}(r; \lambda_Y, R))^2 \mathrm{d}r,$$

where $\overline{F}$ is the sample mean of F and as usual we approximate the integral by a finite sum.

Using the above method, we find that $\hat{\theta} = (\hat{\lambda}_Y, \hat{R}) = (0.00014, 51)$, implying that $\nu = 7.4$. With these parameters, 99 simulations produce envelopes for which 7% and 1% of the data is outside for the g and the G-functions (and 0% otherwise).

Simulating another 1000 realisations with these parameters yields an average

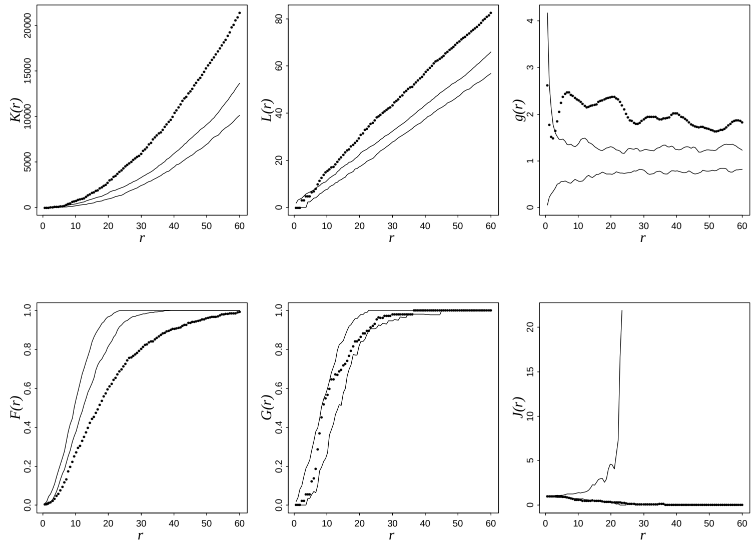

Figure 8.9: Envelopes for 99 Poisson processes, with summary statistics for the short-leaf pines data.

number of events of 126.4, which is very close to the actual number of events in the data, and hence we accept the model.

Estimating the Thomas Process

For the Thomas process we know both the theoretical K- and g-functions, and we can hence apply contrast methods as described in Section 7.5. We choose in this example to minimize the discrepancy for the K-function, and thus by finding the values λ_Y and σ that minimizes (7.15), we hope that the theoretical K-function looks like the function for the data. The estimates are found to be $\hat{\theta} = (\hat{\lambda}_Y, \hat{\sigma}) = (0.00005, 36.3)$, and as can be seen in Figure 8.10, these values actually corresponds quite nicely to the K-function for the data. In the same figure, we have made a plot of the theoretical g-function with the g-function for the data. Given the data, this also seems like the best fit.

With the parameters estimated, we find that the mean number of offspring per mother point should be 21.3.

In order to simulate a Thomas process with these parameters, we again consider the dataset to be a missing data problem. However, in theory, there can be interaction to infinity because of the distribution of the daughter points. Letting R be the 99% quantile of $N(0, \sigma^2)$, we accept to make a small error by simulating on $W_{\oplus R}$. For the model checking, 99 simulations provided envelopes,

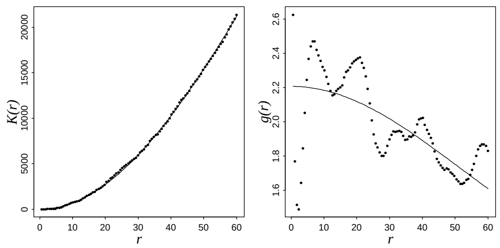

Figure 8.10: The theoretical K- and g-functions for the Thomas process (solid lines), and the corresponding functions for the data.

with data being completely inside.

Here the average number of events in 1000 samples is 126.2.

Conclusion

The short-leaf pines data can be fitted by either a Matérn cluster process with parameters $\hat{\theta} = (\hat{\lambda}_Y, \hat{R}) = (0.00014, 51)$ or a Thomas cluster process with $\hat{\theta} = (\hat{\lambda}_Y, \hat{\sigma}) = (0.00005, 36.3)$.

The two models suggest a different number of clusters and hence also a different number of offspring per cluster. For the Matérn cluster process the number of clusters are estimated to be 17 with about 7 offspring per cluster, while the Thomas cluster process has 6 clusters with a mean of 21 offspring per cluster. The two processes have comparable radii for the offspring – the Matérn cluster process has events distributed within radius 51, while the Thomas cluster process has 95% of the events within a radius of 70 metres from the mother point.

8.3 Biological Cell Data

In this section we consider a dataset that consists of 42 events representing the centers of biological cells at the early stage of the development of an adult insect. The dataset is shown in Figure 8.11. For more information on this dataset we refer to Crick & Lawrence (1975) that provide the biological background and to Ripley (1977) and Diggle (1983) who considers it from a statistical point of view. Clearly, this pattern is highly regular with some kind of hard-core of $0.08 - 0.10$, and the Dirichlet cells are all comparable in size and shape. Further, we notice a striking lack of events in the corners of the plot which may indicate that the observation window is not the unit square. Disregarding

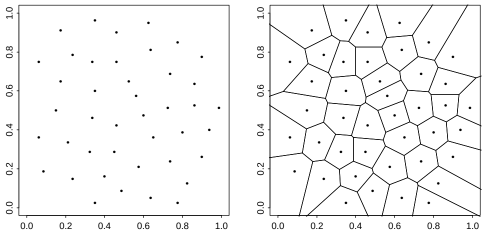

Figure 8.11: Centers of 42 biological cells (left) and Dirichlet cells (right).

this observation we still try to fit a model on the unit square, mainly because the software we have implemented only allows for rectangular areas.

Fitting the SSI and the Matérn II Models

Motivated by the obvious hard-core we try fitting first an SSI process and then a Matérn type II process to the data, disregarding that the Matérn type II process might seem less biologically appropriate. To speed up computations and get a better fit, we modify the SSI process so that it stops after placing $40-45$ events. For the Matérn process we use intensity 1.000 of the generating process that produces just about 42 events after thinning. The summary statistics for these models are shown in Figures 8.12 and 8.13. Both these models are rejected and from the plot of g in particular we see that it appears that we need both a hard-core and also repulsion out to a range of about 0.15.

Fitting a Multi-scale Model

We hence try fitting a multi-scale process with two levels. Using pseudo-likelihood estimation gives $\hat{R}_1 = 0.084$ and $\hat{R}_2 = 0.1066$ and essentially $\gamma_1 = \gamma_2 = 0$ which does not provide a very good fit. Examining the plots it appears that we should rather take R_2 somewhere between 0.12 and 0.15. Trying a few values of R_2 in that range we settle on the results shown in Figure 8.14 which is not accepted but appears to be close.

Fitting a Hybrid Hard/Soft–core Model

Following the example of Ripley (1977) we now try fitting a hybrid hard- and soft-core model of finite interaction range (see also Example 5.22).

Figure 8.15 shows the summary statistics for the values found by Ripley (1977)

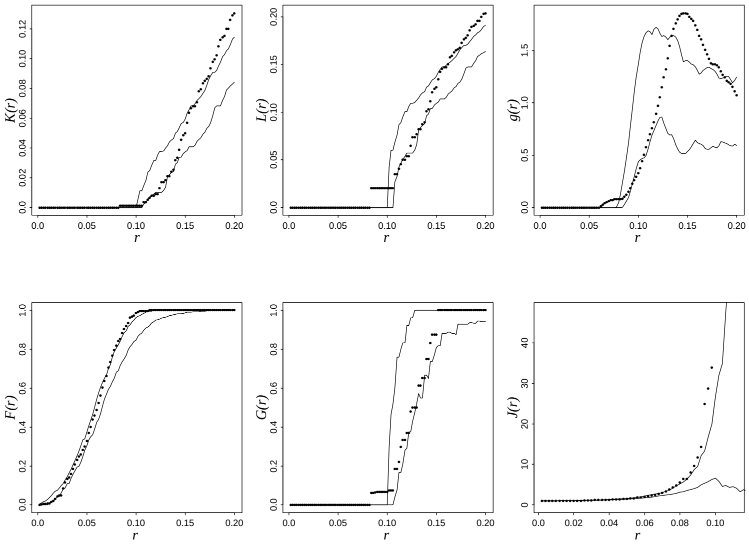

Figure 8.12: Envelopes from 99 samples from a SSI process with $R = 0.10$ restricted to 45 events.

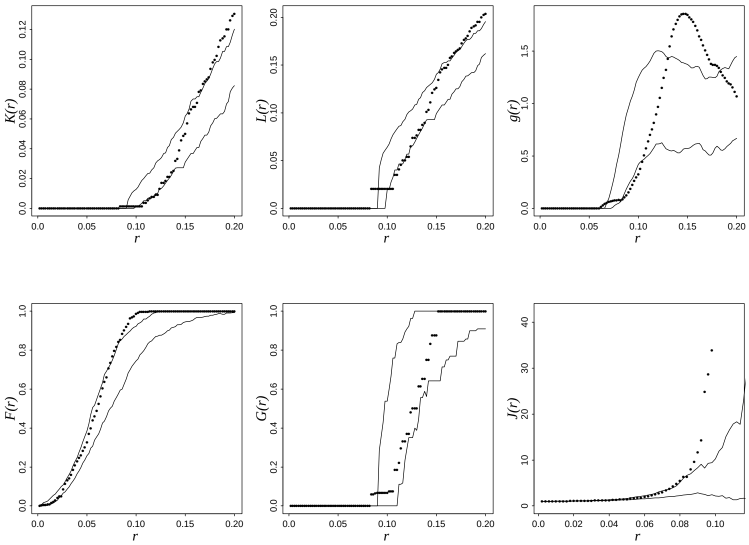

Figure 8.13: Envelopes from 99 samples from a Matérn type II process with $R = 0.10$ thinned from a Poisson point process with intensity 1000.

presumably using ad-hoc methods, and even though we reject these values based on the plots we acknowledge that they are quite good especially considering that Ripley makes a hard-core of 0.1 despite the fact that a single pair of events only have a distance of 0.836. Diggle (1983) also fits this model to the data but by using contrast methods. Figure 8.16 shows the result which is remarkably worse than the result obtained for the values found by Ripley. Finally, we modify the values found by Ripley based on the plots and after a few ad-hoc attempts to determine how a small change in one of the parameters affects the plots we settle on the values used in Figure 8.17 which are acceptable.

Fitting Ord's Process

The final type of model we will try to fit for this dataset is Ord's process (see Example 5.30). Seeing that all the bounded Dirichlet cells in Figure 8.11 have an area in the range of $0.016 - 0.026$ with a mean of 0.021 we wish to make an Ord's process that discourage cells of other sizes. Letting $A(\xi, x)$ denote the area of $C(\xi \mid x)$ we therefore consider a density given by

$$p(x) \quad \propto \quad \beta^{x(\mathcal{X})} \prod_{\xi \in x} \begin{cases} 1 & \text{if } |0.021 - A(\xi \mid x)| < \varepsilon \\ e^{-\rho|0.021 - A(\xi|x)|} & \text{otherwise.} \end{cases}$$

Encouraged by the data we let $\varepsilon = 0.04$, and let $\beta = 45$ which gives a mean of 42 event, and try with $\rho = 20$. The result is not accepted and the reason for this is Figure 8.18 where 5 samples from this process are shown along with the dataset. We notice that neither size nor shape of the Dirichlet cells is regular enough in the samples. It might be possible to fix this problem by using other parameters, but we also notice that the events are not situated in the center of their Dirichlet cells in the samples so we probably need to incorporate this in the model as well. Due to time constrains this is left undone.

8.4 Swedish Pines

In this section we analyse a dataset of Swedish pine saplings, see Ripley (1981). The dataset consists of 71 saplings located in a square of 10×10 meters, as seen in Figure 8.19. Analyses of this dataset can be found many places, among these are Ripley (1981) and Baddeley & Turner (1998), giving results on which we will comment later.

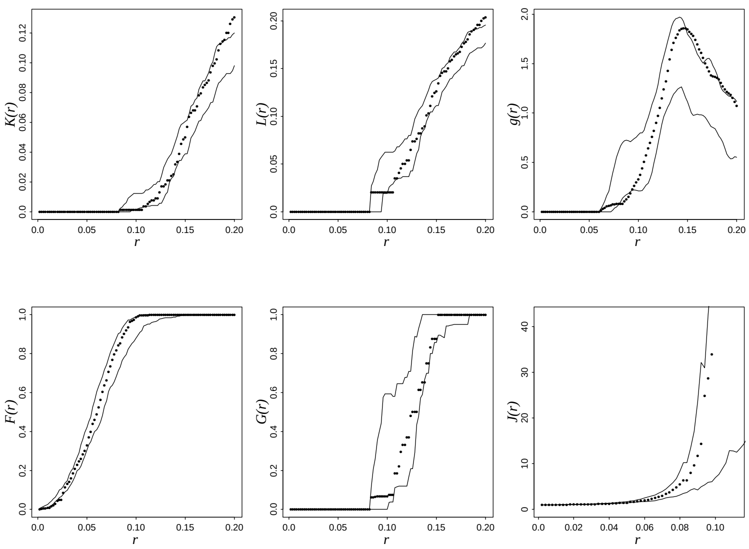

Figure 8.14: Envelopes from 99 samples from a multi-scale process with $\beta = 5000, \gamma_1 = 0.00, \gamma_2 = 0.10, R_1 = 0.084$, and $R_2 = 0.130$.

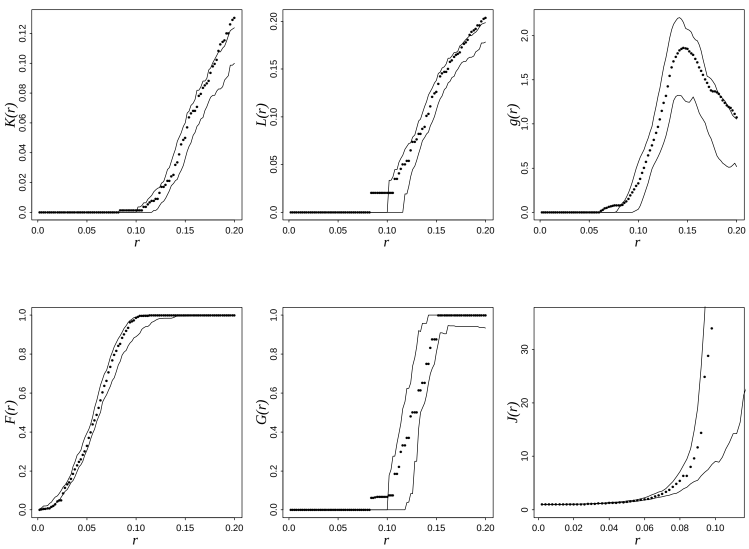

Figure 8.15: Envelopes from 99 samples from a hard/soft process with Ripley's suggestions $\beta = 20000, \rho = 80, R_1 = 0.100, R_2 = 0.140$.

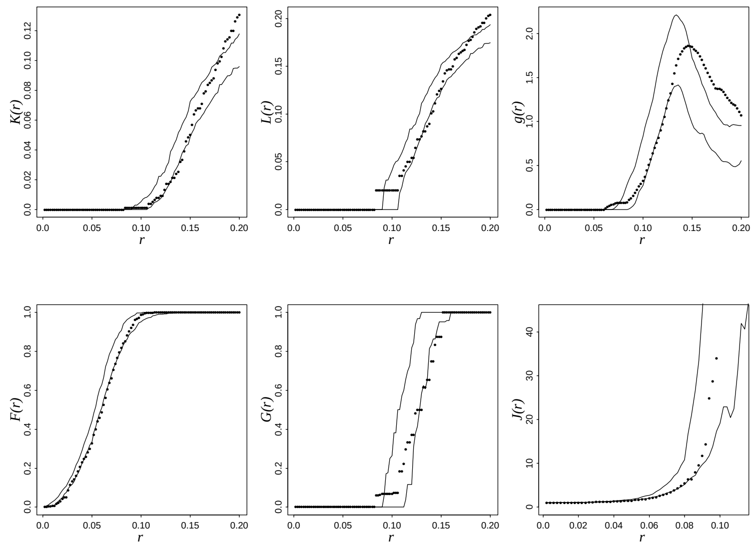

Figure 8.16: Envelopes from 99 samples from a hard/soft process with Diggle's suggestions $\beta = 20000, \rho = 87, R_1 = 0.090, R_2 = 0.130$.

Model Selection

Checking the data for CSR, we calculate the envelopes of summary statistics for 99 simulated Poisson point processes, for which the result is shown in Figure 8.20. The CSR assumption is rejected, and we continue the analysis with an examination of more regular models.

Fitting a Strauss Process

First we check if the data can be described by a Strauss model. Using pseudo-likelihood estimation we find that $(\hat{\beta}, \hat{\gamma}, \hat{R}) = (3.28, 0.20, 0.75)$. For comparison Ripley (1981) gives estimates $(\hat{\gamma}, \hat{R}) = (0.20, 0.70)$ and Ripley (1988) gives $(\hat{\gamma}, \hat{R}) = (0.15, 0.70)$ neither with any estimate of β.

As model validation we simulate 99 Strauss processes with the estimated parameters and calculate the belonging envelopes for all summary statistics. Here we find the number of points from the data located outside these envelopes to be 9% for the J-function and 1% for the remaining functions. This is borderline between acceptance and rejection so we continue the analysis a bit further.

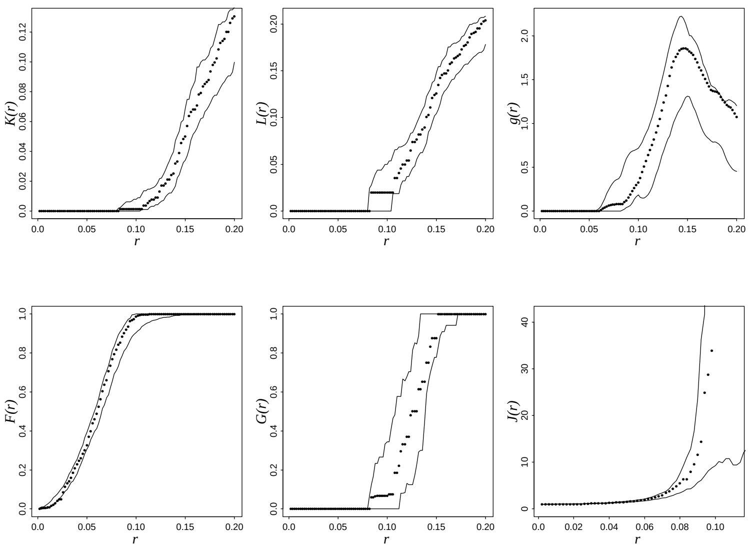

Figure 8.17: Envelopes from 99 samples from a hard/soft process with $\beta = 20000, \rho = 70, R_1 = 0.080, R_2 = 0.145$.

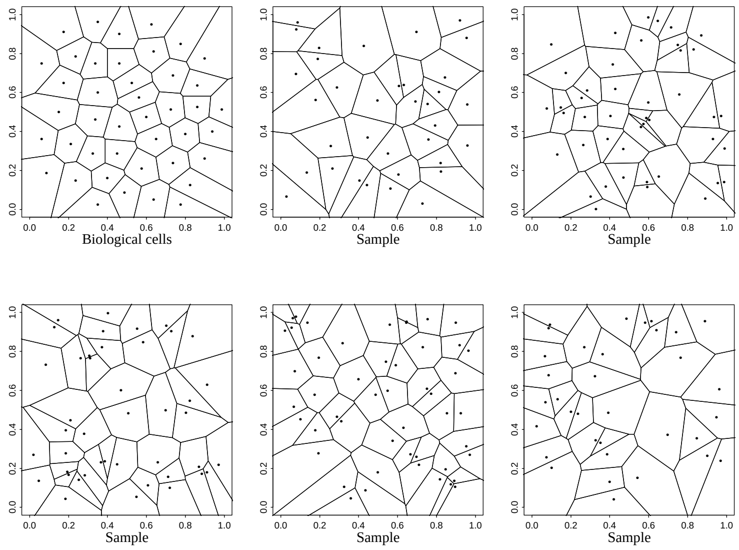

Figure 8.18: Dirichlet cells for the dataset (upper left) and for 5 samples from an Ord's process.

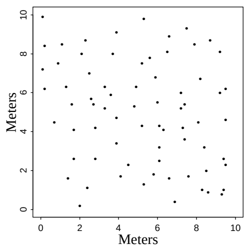

Figure 8.19: The point pattern of the 71 Swedish pines.

Fitting a two-step Multi-scale Process

Using pseudo-likelihood estimation for a two-step multi-scale process we find the parameter estimate

$$(\hat{\beta}, \hat{\gamma}_1, \hat{\gamma}_1, \hat{R}_1, \hat{R}_2) \quad = \quad (11.58, 0.11, 0.42, 0.75, 1.00)$$

which result in the plots shown in Figure 8.21 and hence is accepted.

Fitting Ord's Process

The final process that we will try to fit to the data set is Ord's process, see Example 5.30. This is also done by Baddeley & Turner (1998) and it is suggested as a further examination of the dataset by Ripley (1981).

In this analysis we follow the approach by Baddeley & Turner (1998) using a 'Strauss-type' kernel as the function depending on the area (A) of the Dirichlet cells. This function is given by

$$g(C(\xi \mid x)) = \begin{cases} 1 & \text{if } A > A_0 \\ \gamma & \text{if } A \leq A_0, \end{cases}$$

with parameters β, A_0, and γ. We could use pseudo-likelihood estimation to estimate these but since this is already done in Baddeley & Turner (1998) giving $(\hat{A}_0, \hat{\beta}, \hat{\gamma}) = (1.1, 1.70, 0.43)$ we save ourselves the trouble of implementing the necessary algorithm. However, making 99 samples with these values results in an average number of events of only 51 compared to the 71 events in the dataset. We therefore change $\hat{\beta}$ to 6.0 which give an average number of events of 70 for 99 samples and make the envelopes seen in Figure 8.22. These parameters are rejected, and looking at the plot of the g-function in particular we notice that the reason for this is a lack of inter-event distances around 0.5 meters in the dataset. Any further attempt to fit an Ord's process to this dataset should therefore try to accommodate this fact.

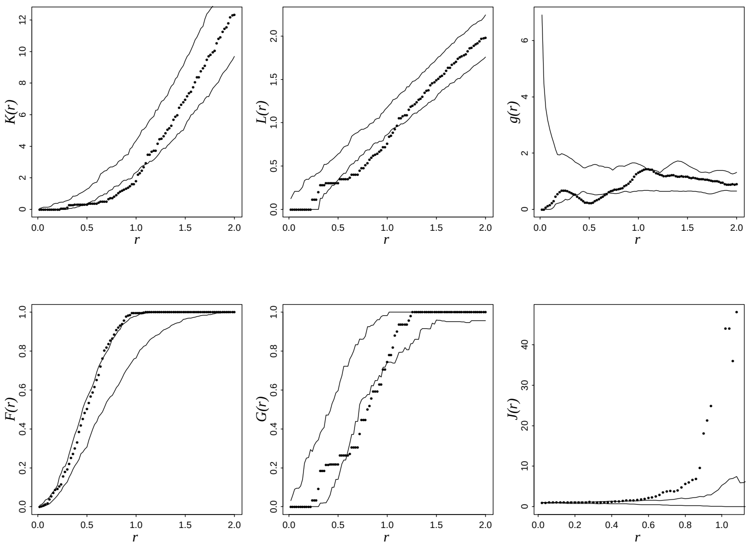

Figure 8.20: Envelopes for a Poisson point process with the corresponding summary statistics for the Swedish pines.

8.5 Norwegian Spruce

This is a dataset containing 134 positions of trees of the Norwegian spruce-type. These are situated in Saxonia (Germany) and have been measured in a 56 by 38 metres area. An analyse of the data set can be seen in Stoyan & Stoyan (1994). As can be seen in Figure 8.23, the data does not seem to be clustered, and since we are dealing with trees, we might expect a hard-core between events.

Model Selection

First we test the dataset for CSR, see Figure 8.24, whereby it can be seen that the hypothesis of CSR clearly must be rejected. It can in particular be seen from the G-function, that we are dealing with a more regular kind of point process with a hard-core of about $R = 1$ metre.

Fitting a Strauss Model

Using pseudo-likelihood estimation we obtain the estimate

$$(\hat{\beta}, \hat{\gamma}, \hat{R}) = (0.128, 0.135, 1.65).$$

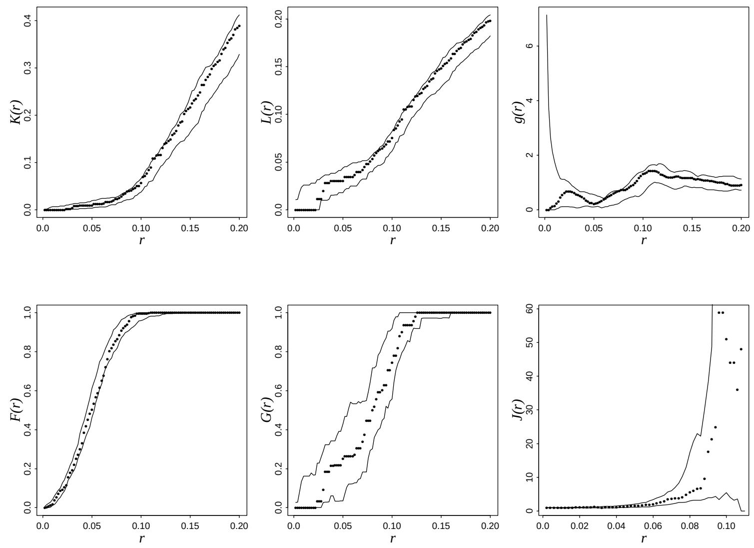

Figure 8.21: Envelopes for the multi-scale process for the Swedish pine saplings data set.

With these parameters, we simulate 99 processes and calculate envelopes, resulting in 14% of the data being outside the envelopes for the K/L-function, while 15% are outside for the g-function. Only a negligible percentage of the data is outside for the rest of the summary statistics. What is more important, is that by looking at the g-function (see Figure 8.25), it seems that there is also interaction further out than $R = 1.65$. Hence we continue by fitting a two-step multi-scale model.

Fitting a two-step Multi-scale Model

Using pseudo-likelihood estimation we find the parameter

$$(\hat{\beta}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{R}_1, \hat{R}_2) = (0.578, 0.092, 0.489, 1.65, 3.58)$$

With this model, 99 simulations yields envelopes with 1% and 6% of the data outside, for the K/L- and g-functions respectively, and 0% for the rest. Even though we accept this model we still try a three-step multi-scale model to see if this produces an even better fit.

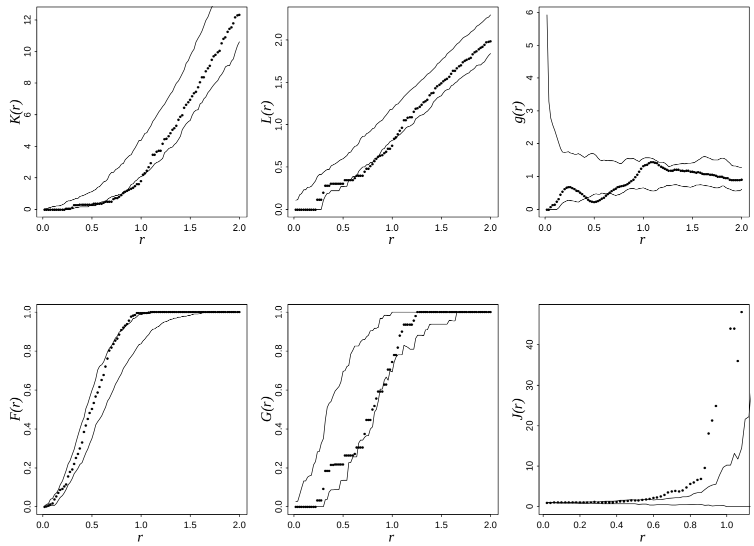

Figure 8.22: Envelopes for the 99 simulated Ord's processes compared with the swedish pine data.

Three-step Multi-scale Model

Again using pseudo-likelihood estimation we find the parameter

$$(\hat{\beta}, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{R}_1, \hat{R}_2, \hat{r}_3) \quad = \quad (9.68, 0.041, 0.406, 0.689, 1.65, 3.58, 6.41)$$

The summary statistics for these parameters are shown in Figure 8.26, and as can be seen, fairly much of the data is outside the envelopes (11% for the K/L-function and 3% for the g-function). By looking at the g-function, we see that the R-values might be too large, and hence we try to adjust these a

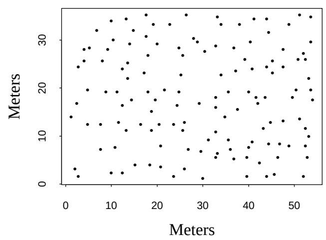

Figure 8.23: Spruce-data. Shows the locations of 134 trees of the Spruce-specie.

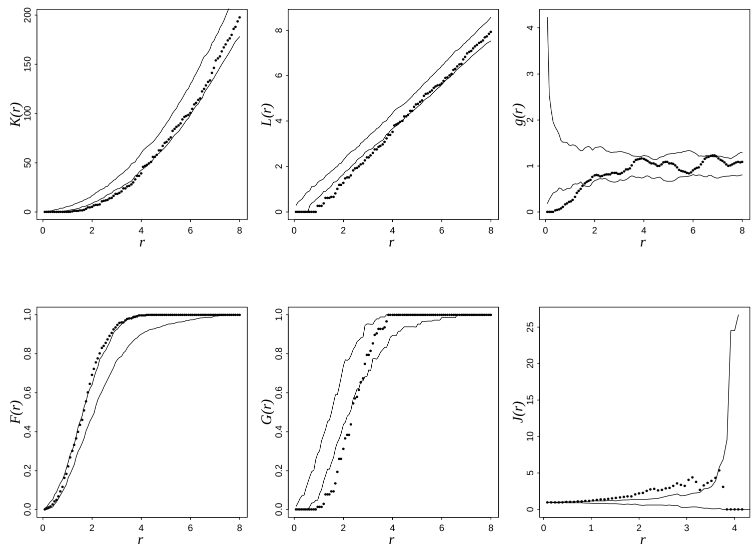

Figure 8.24: Envelopes for 99 simulated Poisson point processes (full line), together with summary statistics for the spruce-data (dots).

little, such that $(R_1 = 1.6, R_2 = 3.52, R_3 = 6.32)$. The corresponding MPLEs are $\hat{\beta} = 4.76, \hat{\gamma}_1 = 0.057, \hat{\gamma}_2 = 0.446$, and $\hat{\gamma}_3 = 0.736$. With these parameters, 99 simulations only yield 3% and 2% of the data outside for the K/L and F-functions and thus we accept this model.

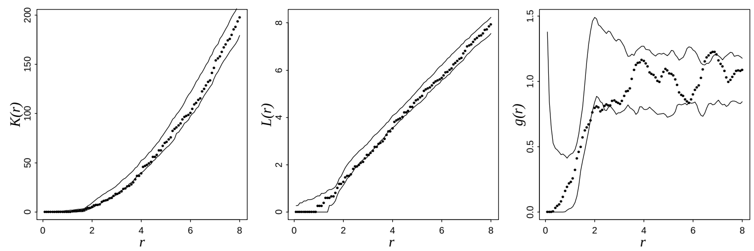

Figure 8.25: Envelopes for the spruce data fitting the Strauss model with $(\hat{R} = 1.65, \hat{\beta} = 0.128, \hat{\gamma} = 0.135)$.

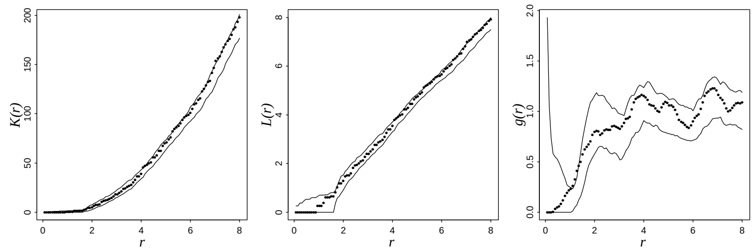

Figure 8.26: Envelopes for the spruce data fitting a three-step multi-scale model with ($\hat{R}_1 = 1.65, \hat{R}_2 = 3.6, \hat{R}_3 = 6.41, \hat{\beta} = 9.68, \hat{\gamma}_1 = 0.041, \hat{\gamma}_2 = 0.406, \hat{\gamma}_3 = 0.689$).

8.6 Bamboo Forest

This analysis is based on a sample of 201 wild bamboos located in a square of 10×10 metres. The dataset is shown in Figure 8.27, where 86 bamboos with the label 1 corresponds to dead (blighted) bamboos and the remaining 115 with label 2 are the living bamboos.

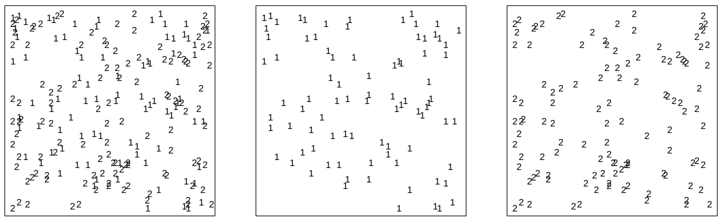

Figure 8.27: Plot of the bamboo data. The number 1 corresponds to blighted bamboos whereas the ones labeled 2 are living bamboos.

First we determine if the dataset can be described by complete spatial randomness, see Figure 8.28.

The result of this plot is that there are respectively 12%, 8%, 0%, 2%, and 2% outside the envelopes for the $K/L, g, F, G$, and J -functions. Hence we reject the assumption of complete spatial randomness. We then treat the two types as separate point patterns and find that both the blighted and the living bamboos can be accepted to follow a homogeneous Poisson model.

Hence the two types can each be described by a homogeneous Poisson point process, but when combined it seems like there is some attraction between the two different types. Searching for a model that does not depart much from Poisson and may model mild attraction, lead us to considering if the Cox process could be a possible choice, for instance the bivariate linked Cox process treated in Diggle (1983).

Due to time constrain we will not extend this analysis further, but refer the reader to Ogata & Tanemura (1985) where approximate likelihood for binary marked point processes including an analyse of these bamboos, are treated.

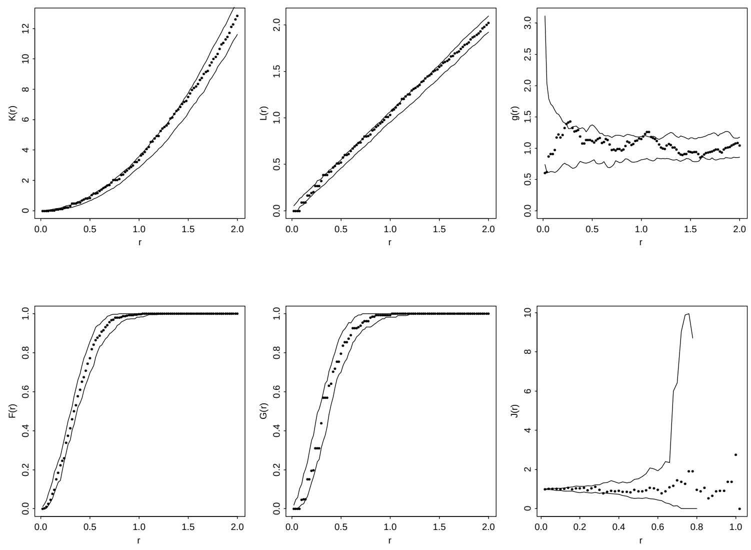

Figure 8.28: Summary statistics for the bamboo data, with envelopes calculated from a sample of 99 simulated homogeneous Poisson processes.

Chapter 9

Conclusion

The aim of this thesis, as set out in the Preface, is to give a comprehensive mathematical and statistical coverage of the theory of spatial point processes. Looking at the material presented in the previous chapters we find that this goal largely has been achieved. Any student or researcher interested in spatial point processes will find all the important parts such as the model types, model selection and validation, simulation, and parametric inference well covered. The theory is treated with a solid mathematical foundation as well as several examples. This said, there are still many possibilities for enhancement. Throughout the thesis there are topics that have only been brushed upon, some of which probably could be the topic of an entire thesis.

There are three topics which we would like to emphasize, two of which have not even been discussed in the thesis so far, that in particular would be of further interest. The first is that of marked point processes. Although we have covered this topic to some extend, we still feel a need for further studies, especially through examples that could give guidelines as to how different kinds of marked point patterns are analysed in practice. The second topic is that of very large point patterns. Most of the theory presented in this thesis will work for any size point pattern in principle, but if we get a very large number of events, many of the algorithms presented will be extremely slow, so some theory specifically aimed at very large point patterns would be in order. The third and last topic is replicated point patterns, i.e. the case where we have replication from the same point process (or at least from related processes). The replication could be a time-series or replication of the same "experiment" at different locations.

Apart from these three topics we would also like to have done a more thorough comparison of the different algorithms both for simulation and for parametric inference. In particular it would be of interest to study how well our modification of Markov chain Monte Carlo maximum likelihood estimation for missing data performs, and to have implemented the perfect simulation algorithm in order to compare it to the Metropolis-Hastings algorithm.

In general we feel that we have presented the theory of how to study spatial point processes quite well but that we to some extend have neglected to present clearly why one should use the theory. The analyses made illustrate how to apply the theory, but lack somewhat in emphasizing the commercial value of the analysis performed.

Appendix A

Implementation

In this appendix we give the code for some of the implementation we have made and used to apply the presented theory in practice.

The less computer intensive parts of the algorithms have been implemented in S-plus, and the computer intensive parts have been implemented in C. This is done to exploit the analyzing and graphical tools of S-plus while retaining speed of computation.

On page 184 and 185 we show the implemention used to calculate the various summary statistics.

On page 186 and 187 we show the implemention of a Metropolis-Hastings algorithm for a Strauss process.

On page 188 we show the implementation of the algorithm for pseudo-likelihood estimation for the Strauss process.

The interested reader may obtain all the implementation we have made by writing hedegard@@bigfoot.com, but it should be understood that documentation is sparse or non-existing.

```
##################################################################
##          S-code for calculating summary statistics            #
##################################################################

ss <- function(x,y,maxr,numb=100,maxx=1,maxy=1,minx=0,miny=0,
               lambdahat=length(x), Lh=T, Fh=T, Gh=T)
# USAGE:
#   INPUT:
#    x: x-values for pattern
#    y: y-values for pattern
#    maxr:      largest r-values for the summary statistics
#    numb:      number of 'measurements'           (default=100)
#    maxx:      upper x-value                      (default=1)
#    maxy:      upper y-value                      (default=1)
#    minx:      lower x-value                      (default=0)
#    miny:      lower y-value                      (default=0)
#    lambdahat: estimate for the number of points  (default=length(x))
#    Lh:        calculate K, L, and g-function     (default=true)
#    Fh:        calculate F-function               (default=true)
#    Gh:        calculate G-function               (default=true)
#   OUTPUT:
#    A list of summary statistics which can be plot against rrange,
#    i.e. plot(rrange,Khat) yields a plot of the K-function etc.
#    Khat:      K-function for the pattern
#    Lhat:      L-function for the pattern
#    ghat:      g-function for the pattern
#    Fhat:      F-function for the pattern
#    Ghat:      G-function for the pattern
#    Jhat:      J-function for the pattern
#    rrange:    the vector to plot against each summary statistic
  {
    dyn.load("/user/hedegard/Projekt/Splus/Functions/ss.o")
# determine the size of the observation window
    xmax <- maxx-minx
    ymax <- maxy-miny
# calculate the intensity
    lambdahat <- lambdahat/(xmax*ymax)
# setup the values for which to calculate summary statistics
    rrange <- seq(maxr/numb,maxr,maxr/numb)
# call to c-code, calculating the K, L, and g-functions
    if (Lh==T) {
      z <- .C("ss_Lg_c",
              as.integer(length(x)),
              as.single(x),
              as.single(y),
              as.integer(length(rrange)),
              as.single(rrange),
              as.single(lambdahat),
              as.single(xmax),
              as.single(ymax),
              Lhat = single(length(rrange)),
              ghat = single(length(rrange)))
      Lhat <- z$Lhat
      ghat <- z$ghat
      Khat <- (Lhat)^2*pi
    } else {Lhat <- NULL; ghat <- NULL; Khat <- NULL}
# Fhat
    if (Fh==T) {
# determine the size of the grid (20 by i, where i >= 20)
      if (xmax < ymax) {xnr <- 20; ynr <- floor(20*(ymax/xmax))}
      else {xnr <- floor(20*(xmax/ymax)); ynr <- 20}
# call to c-code, calculating the F-function
      z <- .C("ss_F_c",
              as.integer(length(x)),
              as.single(x),
              as.single(y),
              as.integer(length(rrange)),
              as.single(rrange),
              Fhat = single(length(rrange)),
              as.integer(xnr),  #grid size
              as.integer(ynr),
              as.single(xmax),
              as.single(ymax),
              as.single(minx),
              as.single(miny))
    Fhat <- z$Fhat
    } else {Fhat <- NULL}
    if (Gh==T) {
# call to c-code, calculating the G-function
      z <- .C("ss_G_c",
              as.integer(length(x)),
              as.single(x),
              as.single(y),
              dist=single(length(x)),
              as.integer(length(rrange)),
              as.single(rrange),
              Ghat = single(length(rrange)),
              as.single(xmax),
              as.single(ymax),
              as.single(maxx),
              as.single(minx),
              as.single(maxy),
              as.single(miny))
      Ghat <- z$Ghat
    } else {Ghat <- NULL}
    if ((Fh==T) & (Gh==T)) {
# calculating the J-function
      Jhat <- (1-Ghat)/(1-Fhat)
    } else {Jhat <- NULL}
# return the summary statistics, and the r-range
    list(Khat=Khat, Lhat = Lhat, ghat = ghat, Fhat = Fhat,
         Ghat=Ghat, Jhat=Jhat, rrange=rrange)
  }
```

```c
/*################################################################
##          C-code for calculating summary statistics           #
################################################################*/

#include <math.h>

const float Pi=3.141593;

/* Epanichnikov kernel */
float EpaKernel(float b, float s)
  {
    return (fabs(s) <= b) * 3 / (4*b) * (1-s*s/(b*b));
  }

/* L and g-function */
void ss_Lg_c(long int *events, float *x, float *y, long int *nr,
             float *rrange, float *lambdahat, float *xmax,
             float *ymax, float *Lhat, float *ghat)
{
  int r,i,j;
  float dx,dy,lh2,b,edge,dist;
  lh2 = *lambdahat * *lambdahat;
  b = 0.15 / sqrt(*lambdahat);

  for (r = 0; r < *nr; r++)
    {
      Lhat[r]=0;
      ghat[r]=0;
      for (i=0; i < (*events -1); i++)
        {
          for (j=i+1; j < *events; j++)
            {
              dx = fabs(x[i] - x[j]);
              dy = fabs(y[i] - y[j]);
              dist = sqrt(dx*dx+dy*dy);
              edge = 1 / ((*xmax-dx)*(*ymax-dy));

              Lhat[r] = Lhat[r] + (dist < rrange[r]) * edge;
              ghat[r] = ghat[r] + EpaKernel(b,(rrange[r]-dist)) * edge;
            }
        }
      Lhat[r] = 2 * Lhat[r] / lh2;
      Lhat[r] = sqrt(Lhat[r] / Pi);
      ghat[r] = ghat[r] / (lh2*Pi*rrange[r]); /* *(*xmax * *ymax)); */
    }
}

/* F-function */
void ss_F_c(long int *events, float *x, float *y,
            long int *nr, float *rrange, float *Fhat,
            long int *gridnrx, long int *gridnry,
            float *xmax, float *ymax, float *minx, float *miny)
{
  int r,i,j,e,sum;
  float gridx,gridy,delta,dx,dy,dist2,mindist,r2;

  for (r=0; r < *nr; r++){
    r2 = rrange[r] * rrange[r];
    sum = 0;
    if (*xmax < *ymax)
      delta = (*xmax-2*rrange[r]) / (20-1);
    else
      delta = (*ymax-2*rrange[r]) / (20-1);
    gridx = rrange[r] + *minx;
    for (i=0; i < *gridnrx; i++){
      gridy = rrange[r] + *miny;
      for (j=0; j < *gridnry; j++){
        mindist = *xmax + *ymax;   /* triangle inequality: */
                                   /* mindist <= xmax + ymax */
        mindist *= mindist;
        for (e=0; e < *events; e++){
          dx = gridx - x[e];
          dy = gridy - y[e];
          dist2 = (dx*dx+dy*dy);
          if (dist2 < mindist)
            mindist = dist2;
        }
        sum += (mindist <= r2);
        gridy += delta;
      }
      gridx += delta;
    }
    Fhat[r] =  (float) sum / (*gridnrx * *gridnry);
  }
}

/* G-function */
void ss_G_c(long int *events, float *x, float *y,
            float *dist, long int *nr, float *rrange,
            float *Ghat, float *xmax, float *ymax,
            float *maxx, float *minx, float *maxy, float *miny)
{
  int r,i,j,e,sum;
  float dx,dy,dist2,mindist,r2;
  int inside,ins;

  for (i=0; i < *events; i++){
    mindist = *xmax + *ymax;
    mindist *= mindist;
    for (j=0; j < *events; j++){
      if (i != j){
        dx = x[i] - x[j];
        dy = y[i] - y[j];
        dist2 = (dx*dx+dy*dy);
        if (dist2 < mindist)
          mindist = dist2;
      }
    }
    dist[i] = mindist;
  }

  for (r=0; r < *nr; r++){
    r2 = rrange[r] * rrange[r];
    sum = 0;
    inside = 0;
    for (e =0; e < *events; e++){
      ins = x[e] >= (*minx + rrange[r]) && x[e] <= (*maxx - rrange[r])
        && y[e] >= (*miny + rrange[r]) && y[e] <= (*maxy - rrange[r]);
      sum += ins && dist[e] <= r2;
      inside += ins;
    }
    if (inside == 0)
      Ghat[r] = 1;
    else
      Ghat[r] = (float) sum / inside ;
  }
}
```

```
#################################################################
##  S-code : Metropolis-Hastings algorithm for Strauss process   #
#################################################################

MHS.PL <- function(x=numeric(), y=numeric(), beta, gamma, radius,
                   samples, spacing, burnin, maxevents,
                   fact=3, maxx=1,maxy=1,minx=0,miny=0)
# USAGE:
#   INPUT:
#    x, y:                        values for start pattern
#    beta, gamma, radius:         parameters
#    samples:                     number of samples to make
#    spacing:                     spacing between each sample
#    burnin:                      burnin before taking first sample
#    maxevents:                   algorithm terminates i no. of event
#                                 ever exceeds maxevents
#    fact:                        simulation is made on fact * reactangle
#    maxx, maxy, minx, miny:      where rectangle = (minx;maxx)*(miny;maxy)
#                                 but only events within rectangle are
#                                 returned.
#   OUTPUT:
#    x,y :                        list of the samples = event in rectangle
#    events:                      number of events per sample
#    maxx, maxy, minx, miny:      the rectangle

{
    if ( fact < 1)
      stop("fact must larger than (or equal to) 1")
    mfact <- (fact-1)/2
    dyn.load("/user/hedegard/Projekt/Splus/Functions/MH-strauss.o")
    xmax <- maxx-minx
    ymax <- maxy-miny
    muX<- fact*fact*xmax*ymax
    n <- length(x)
    ppx <- list()
    ppy <- list()
    ppn <- NULL
    x[(n+1):maxevents] <- -1
    y[(n+1):maxevents] <- -1

    .C("init_Strauss_c",
       as.single(beta),
       as.single(gamma),
       as.single(radius))

    ite <- burnin
    for (t in (1:samples)){
      z <- .C("MH_c",
              n = as.integer(n),
              as.integer(maxevents),
              x = as.single(x),
              y = as.single(y),
              as.integer(ite),
              as.single(fact),
              as.single(mfact),
              as.single(muX),
              as.single(xmax),
              as.single(ymax),
              as.single(minx),
              as.single(miny)
              )
      ite <- spacing
      n <- z$n
      x <- z$x
      y <- z$y
      x1 <- x[1:n]
      y1 <- y[1:n]
      inside <- (x1>minx) & (x1<maxx) & (y1>miny) & (y1<maxy)
      ppx[[t]] <- x1[inside]
      ppy[[t]] <- y1[inside]
      ppn[t] <- length(ppx[[t]])
      print(t)
    }
  list(x=ppx,y=ppy,events=ppn,minx=minx,maxx=maxx,miny=miny,maxy=maxy)
}
```

```
/*################################################################
##  c-code : Metropolis-Hastings algorithm for Strauss process   #
################################################################*/

#include <S.h>
#include <math.h>

#define RANDIN seed_in((long *)NULL)
#define RANDOUT seed_out((long *)NULL)

/* ------------------------------------ */
/*          Parameters                  */
/* ------------------------------------ */
float Beta;
float Gamma;
float Rad2;

/* ------------------------------------ */
/*          Init parameters             */
/* ------------------------------------ */
void init_Strauss_c(float *b, float *g, float *r)
{
  Beta = *b;
  Gamma = *g;
  Rad2 = *r * *r;
}

/* ------------------------------------ */
/*  Papangelou conditional intensity    */
/*  one for each supported model        */
/* ------------------------------------ */

float papangelou_c(long int *n, float *x, float *y, float xnew, float ynew)
{
  int i;
  int nei=0;

  for (i=0; i<*n; i++)
    nei = nei + (pow(x[i]-xnew,2)+ pow(y[i]-ynew,2) < Rad2);

  if (nei == 0)
    return (Beta);
  else
    return (Beta * pow(Gamma,nei));
}

/* ------------------------------------ */
/*  Metropolis-Hastings algorithm       */
/*  for Strauss point processes         */
/* ------------------------------------ */

void MH_c(long int *n, long int *max_n, float *x, float *y, long int *ite, float
 *fact, float *mfact, float *muX, float *xmax, float *ymax, float *minx, float *
miny)
{
  float
    x1,y1,pap;
  long int
    i, eta;

  RANDIN;

  for (i=0; (*n < *max_n) && (i< *ite); i++) {
    if (( unif_rand() > 0.5) || (*n==0)) {
      x1 = *minx + *xmax*(unif_rand()*(*fact)-(*mfact)); /*proposing to add an e
vent*/
      y1 = *miny + *ymax*(unif_rand()*(*fact)-(*mfact));
      pap = papangelou_c(n, x, y, x1, y1);

      if ( unif_rand() < ( pap * *muX / (*n+1) )) {  /*deciding wether to keep
                                                        the new event or not*/
        x[*n] = x1;
        y[*n] = y1;
        *n = *n+1;
      }
    }
    else {                                            /* or deleting an event*/
      eta = floor( *n * unif_rand());
      x1 = x[eta];
      y1 = y[eta];
      x[eta] = x[*n-1];
      y[eta] = y[*n-1];
      *n=*n-1;
      pap = papangelou_c(n, x, y, x1, y1);

      if (unif_rand() > ( (*n+1) / (*muX * pap))) {   /*checking if the event */
        x[*n] = x1;                                   /*really should have been
*/
        y[*n] = y1;                                   /* deleted*/
        *n=*n+1;
      }
    }
  }
  RANDOUT;
}
```

```
###################################################################
##   S-code : Pseudo-likelihood estimation for Strauss process   #
###################################################################

PLEStrauss <- function(x,y, R=0.085, grid=50, rmax=0.1,
                       maxx=1,maxy=1,minx=0,miny=0)
# USAGE:

#   INPUT:
#    x, y:                       point pattern
#    R:                          fixed interaction of R
#    grid:                       determines the number of dummy point (grid^2)
#    rmax:                       pl_A is found: A is rectangle "minus" rmax
#                                rmax=R except when comparing R-values
#    maxx, maxy, minx, miny:     where rectangle = (minx;maxx)*(miny;maxy)

#   OUTPUT:
#    logpl :                     value of logPL
#    beta, gamma:                the MPLE


{
  ins <- (x > minx+rmax) & (x < (maxx-rmax)) & (y > miny+rmax )
  & (y < (maxy-rmax))
  minusx <- x[ins]
  minusy <- y[ins]
  xmax <- maxx-minx
  ymax <- maxy-miny
  n <- length(minusx)

## make dummy points. xnr*ynr grid
  step <- (min(xmax,ymax)-2*rmax)/(grid-1)
    if (xmax < ymax) {xnr <- grid; ynr <- floor((ymax-2*rmax)/step +1)}
    else {xnr <- floor((xmax-2*rmax)/step +1); ynr <- grid}

  dummyx <- rep(seq(minx+rmax,maxx-rmax,step),ynr)
  dummyy <- sort(rep(seq(miny+rmax,maxy-rmax,step),xnr))

  ux <- c(minusx,dummyx)
  uy <- c(minusy,dummyy)

  w <- (xmax-2*rmax)*(ymax-2*rmax) * rep(1/length(ux),length(ux))

  z <- c(rep(1,n),rep(0,length(dummyx)))
  y <- z/w

  v <- rep(0,length(ux))
  for (i in 1:length(v))
    {
      v[i] <- length( x[((x-ux[i])^2 + (y-uy[i])^2) < R^2]) - z[i]
    }

  res <- glm(y~v, family = poisson, link =log, weights =w,
             control=glm.control(maxit=50))

  beta  <- as.single(exp(res$coef[1]))
  gamma <- as.single(exp(res$coef[2]))
  deviance <- res$deviance

  logpl <- - deviance/2 - length(minusx)*(1+log(w[1]))

  list(logpl=logpl,beta=beta,gamma=gamma)
}
```

Bibliography

Baddeley, A. J. & Møller, J. (1989), 'Nearest-neighbour Markov point processes and random sets', *International Statistical Review* **2**, 89–121.

Baddeley, A. J., Møller, J. & Waagepetersen, R. (1999), Non and semi-parametric estimation of interaction in inhomogeneous point patterns.

Baddeley, A. J. & Silverman, B. W. (1984), 'A cautionary example for the use of second order methods for analyzing spatial point patterns', *Biometrics* **40**, 1089–1094.

Baddeley, A. J. & Turner, T. R. (1998), Practical maximum pseudolikelihood for spatial point patterns, Research report 98, Department of Mathematics and Statistics, University of Western Australia.

Baddeley, A. J. & van Lieshout, M. N. M. (1995), 'Area-interaction point processes', *Annals of the Institute of Statistical Mathematics* **46**, 601–619.

Baddeley, A. J., van Lieshout, M. N. M. & Møller, J. (1996), 'Markov properties of cluster processes', *Advances in Applied Probability (SGSA)* **28**, 346–355.

Barndorff-Nielsen, O. (1978), *Information and Exponential Families in Statistical Theory*, John Wiley & Sons.

Besag, J. & Clifford, P. (1989), 'Generalised monte carlo significance tests', *Biometrika* **76**, 633–642.

Besag, J. E. (1974), 'Spatial interaction and the statistical analysis of lattice systems (with discussion)', *Journal of the Royal Statistical Society* **B** **36**, 192–236.

Besag, J. E. (1977), 'Some methods of statistical analysis for spatial data', *Bulletin of the Institute of International Statistics* **47**, 77–92.

Besag, J. E., Milne, R. & Zachary, S. (1982), 'Point process limits of lattice processes', *Journal of Applied Probability* **19**, 210–216.

Billingsley, P. (1995), *Probability and Measure*, Wiley.

Brix, A. & Møller, J. (1998), Space-time multi type log Gaussian Cox processes with a view to modelling weed data, Technical Report R-98-2012, Department of Mathematical Sciences, Aalborg University.

Campbell, N. R. (1909), 'The study of discontinuous phenomena', *Proc. Camb. Phil. Soc.* **15**, 117–136.

Carter, D. S. & Prenter, P. M. (1972), 'Exponential spaces and counting processes', *Zeitschrift für Wahrsheinlichkeitstheorie und verwandte Gebiete* **21**, 1–19.

Chan, K. S. & Geyer, C. J. (1994), 'Discussion of the paper "markov chains for exploring posterior distributions" by luke tierney', *Ann. Statist.* **22**, 1747–1758.

Chatfield, C. (1996), *The Analysis of Time Series*, Thomson Science and Professional.

Clifford, P. & Nicholls, G. (1994), Comparison of birth-and-death and Metropolis-Hastings Markov chain Monte Carlo for the Strauss process. Department of Statistics, Oxford University. Manuscript.

Cowie, S. R. (1967), 'The cummulative frequency nearest-neibour method for the identification of spatial patterns', *Dept. of Geography, University of Bristol Seminar Paper* **A10**.

Cox, D. R. (1955), 'Some statistical properties related with series of events', *Journal of the Royal Statistical Society* **B 17**, 129–164.

Cox, D. R. & Isham, V. (1980), *Point Processes*, Chapman and Hall, London.

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

Crick, F. & Lawrence, P. (1975), 'Compartments and polychones in insect development', *Science* .

Daley, D. J. & Vere-Jones, D. (1988), *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York.

Diggle, P. J. (1983), *Statistical Analysis of Spatial Point Patterns*, Academic Press, London.

Diggle, P. J., Besag, J. E. & Gleaves, J. T. (1976), 'Statistical analysis of spatial point patterns by means of distans methods', *Biometrics* **32**, 659–667.

Fiksel, T. (1984), 'Estimation of parameterized pair potentials of marked and nonmarked Gibbsian point processes', *Elektronische Informationsverarbeitung und Kypernetik* **20**, 270–278.

Fiksel, T. (1988), 'Estimation of interaction partials of gibbsian point processes', *Statistics* **19**, 77–86.

Fill, J. A. (1998), 'An interruptible algorithm for perfect sampling via Markov chains', *Annals of Applied Probability* **8**, 131–162.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Georgii, H.-O. (1988), *Gibbs Measures and Phase Transition*, Walter de Gruyter, Berlin.

Geyer, C. J. (1992), 'Pratical markov chain monte carlo', *Statistical Science* **7**(4), 473–511.

Geyer, C. J. (1999), Likelihood inference for spatial point processes, *in* O. E. Barndorff-Nielsen, W. S. Kendall & M. N. M. van Lieshout, eds, 'Stochastic Geometry, Likelihood and Computation', Chapman and Hall, London, pp. 79–140.

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

Geyer, C. J. & Thompson, E. A. (1992), 'Constrained Monte Carlo maximum likelihood for dependent data', *Journal of the Royal Society of Statistics* **B 54**, 657–699.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Glass, L. & Tobler, W. R. (1971), 'Uniform distribution of objects in a homogeneous field: Cities on a plain', *Nature* **233(5314)**, 67–68.

Grandell, J. (1976), *Doubly stochastic Poisson processes*, Springer Verlag, Berlin.

Häggström, O., van Lieshout, M. N. M. & Møller, J. (1999), 'Characterisation results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes', *Bernoulli* . (To appear).

Hammersley, J. & Clifford, P. (1971), Markov fields on finite graphs and lattices. Unpublished manuscript.

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Heikkinen, J. & Arjas, E. (1998), 'Non-parametric bayesian estimation of a spatial poisson intensity', *Scandinavian Journal of Statistics* **25**, 435–450.

Jagers, P. (1973), 'On palm probabilities', *Probab. Theory Relat. Fields* **26**, 17–32.

Jensen, J. L. & Künsch, H. R. (1994), 'On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes', *Annals of the Institute of Statistical Mathematics* **46**, 475–486.

Jensen, J. L. & Møller, J. (1991), 'Pseudolikelihood for exponential family models of spatial point processes', *Annals of Applied Probability* **3**, 445–461.

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

Kendall, W. (1998), Perfect simulation for the area-interaction point process, *in* L. Accardi & C. Heyde, eds, 'Probability Towards 2000', Springer, New York, pp. 218–234.

Kendall, W. & Møller, J. (1999), 'Perfect metropolis–hastings simulation of locally stable point processes'.

Kingman, J. F. C. (1993), *Poisson Processes*, Oxford University Press.

Lantuéjoul, C. (1994), 'Nonconditional simulation of stationary isotropic multigaussian random functions', *Geostatistical Simulations* . (eds. M. Armstrong and P. Dowd) Kluwer Academic Publishers, Dordrecht.

Matérn, B. (1960), 'Spatial variation', *Meddelanden fran Statens Skogsforskningsinstitut* **49**(5).

Matheron, G. (1975), *Random Sets and Integral Geometry*, Wiley, New York.

Mecke, J. (1967), 'Stationäre zufällige Maße auf lokalkompakten Abelschen Gruppen', *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **9**, 36–58.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equations of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, 1087–1092.

Meyn, S. P. & Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Møller, J. (1989), 'On the rate of convergence of spatial birth-and-death processes', *Annals of the Institute of Statistical Mathematics* **3**, 565–581.

Møller, J. (1994), *Lectures on Random Voronoi Tessellations*, Lecture Notes in Statistics, Springer-Verlag, New York.

Møller, J. (1999), Markov chain Monte Carlo and spatial point processes, *in* O. Barndorff-Nielsen, W. Kendall & M. van Lieshout, eds, 'Stochastic Geometry, Likelihood and Computation', Vol. 80, Chapman and Hall, pp. 141–172.

Møller, J. & Schladitz, K. (1999), 'Extensions of fill's algorithm for perfect simulation', *Journal of the Royal Statistical Society* **B 61**. (To appear).

Møller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998), 'Log Gaussian Cox processes', *Scandinavian Journal of Statistics* **25**, 451–482.

Moyeed, R. A. & Baddeley, A. J. (1991), 'Stochastic approximation of the MLE for a spatial point pattern', *Scandinavian Journal of Statistics* **18**, 39–50.

Neymann, J. & Scott, E. L. (1958), 'Statistical approach to problem of cosmology', *J. Roy. Statist. Soc.* **B20**, 1–43.

Nguyen, X. X. & Zessin, H. (1979), 'Integral and differential characterizations of Gibbs processes', *Mathematische Nachrichten* **88**, 105–115.

Ogata, Y. & Tanemura, M. (1984), 'Likelihood analysis of spatial point patterns', *Journal of the Royal Statistical Society* **B 46**, 496–518.

Ogata, Y. & Tanemura, M. (1985), 'Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method', *Biometrics* **41**, 421–433.

Penttinen, A. (1984), *Modelling Interaction in Spatial Point Patterns: Parameter Estimation by the Maximum Likelihood Method*, Number 7 in Jyväskylä Studies in Computer Science, Economics, and Statistics.

Peskun, P. (1973), 'Optimum monte-carlo sampling using markov chains', *Biometrika* **60**, 607–612.

Preston, C. (1977), 'Spatial birth-and-death processes', *Bulletin of the International Statistical Institute* **46**, 371–391.

Preston, C. J. (1976), *Random Fields*, Lecture Notes in Mathematics, 534. Springer-Verlag, Berlin-Heidelberg.

Propp, J. G. & Wilson, D. B. (1996), 'Exact sampling with coupled Markov chains and applications to statistical mechanics', *Random Structures and Algorithms* **9**, 223–252.

Quine, M. P. & Watson, D. F. (1984), 'Radial simulation of n-dimensional poisson processes', *Journal of Applied Probability* **21**, 548–557.

Ripley, B. D. (1976), 'The second-order analysis of stationary point processes', *Journal of Applied Probability* **13**, 255–266.

Ripley, B. D. (1977), 'Modelling spatial patterns (with discussion)', *Journal of the Royal Statistical Society* **B 39**, 172–212.

Ripley, B. D. (1979), 'Simulating spatial patterns: dependent samples from a multivariate density. Algorithm AS 137.', *Applied Statistics* **28**, 109–112.

Ripley, B. D. (1981), *Spatial Statistics*, Wiley, New York.

Ripley, B. D. (1987), *Stochastic Simulation*, Wiley, New York.

Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge.

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

Roberts, G. O. & Rosenthal, J. S. (1996), Geometric ergodicity and hybrid markov chains. Statistical Laboratory, University of Cambridge.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton University Press.

Ruelle, D. (1969), *Statistical Mechanics: Rigorous Results*, W.A. Benjamin, Reading, Massachusetts.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall.

Stoyan, D., Kendall, W. S. & Mecke, J. (1995), *Stochastic Geometry and its Applications*, second edn, Wiley, Chichester.

Stoyan, D. & Stoyan, H. (1994), *Fractals, Random Shapes and point Fields*, Wiley, Chichester.

Strauss, D. J. (1975), 'A model for clustering', *Biometrika* **63**, 467–475.

Takacs, R. (1986), 'Estimator for the pair-potential of a Gibbsian point process', *Statistics* **17**, 429–433.

Tanemura, M. (1988), 'Random packing and random tessellation in relation to the dimension of space', *Journal of Microscopy* **151**, 247–255.

Thönnes, E. (1999), 'Perfect simulation of some point processes for the i impatient user.', *Advances in Applied Probability* . (To appear).

Venables, W. & Ripley, B. (1994), *Modern Applied Satistics with S-PLUS*, Springer.

Widom, B. & Rowlinson, J. S. (1970), 'A new model for the study of liquid-vapor phase transitions', *Journal of Chemical Physics* **52**, 1670–1684.

Wood, A. T. A. & Chan, G. (1994), 'Simulation of stationary Gaussian processes in $[0,1]^d$', *Journal of Computational and Graphical Statistics* **3**, 409–432.

Younes, L. (1988), 'Estimation and annealing for Gibbsian fields', *Annales de l'Institut Henri Poincaré. Probabilitiés et Statistiques* **24**, 269–294.

Index