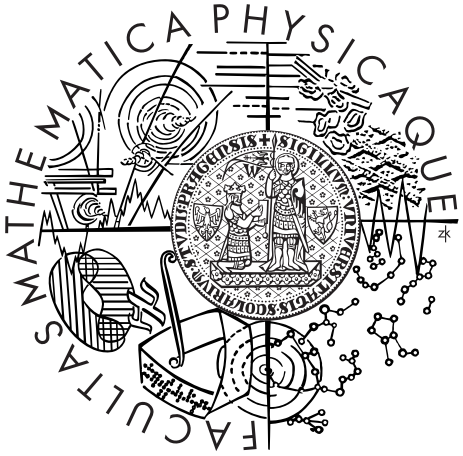
Charles University in Prague
Faculty of Mathematics and Physics

DOCTORAL THESIS



Jiří Dvořák

Statistical inference for spatial and space-time Cox point processes

Department of Probability and Mathematical Statistics

Supervisor of the doctoral thesis:	RNDr. Michaela Prokešová, Ph.D.
Study programme:	Mathematics
Specialization:	m-4 Probability and Mathematical Statistics

Prague 2014

I would like to express my gratitude to all those who directly or indirectly helped me carry out my research in the past four years and write this thesis. First of all, I am grateful to my supervisor, RNDr. Michaela Prokešová, Ph.D., for guiding me both in research and other parts of academic life with never-ending enthusiasm, patience and support.

I am also thankful to prof. Eva B. V. Jensen (Aarhus University, Denmark) for reading a draft of this thesis and providing invaluable comments and insight which helped me considerably improve the text of the thesis.

Finally, I would like to express my special gratitude to my family who supported my decision to pursue the Ph.D. degree and stood by me for the whole duration of my studies. I am especially thankful to my wife Hanka who showed endless patience, love and tolerance over the years in both personal and professional matters.

The work was supported by the Charles University Grant Agency, project no. 664313, Czech Science Foundation, project no. P201/10/0472, and the grant SVV-2014-260105.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, 29 May 2014

Název práce: Statistika prostorových a časoprostorových Coxových bodových procesů

Autor: Jiří Dvořák

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí disertační práce: RNDr. Michaela Prokešová, Ph.D.

Abstrakt: Odhady parametrů pro modely prostorových a časoprostorových bodových procesů jsou v posledních letech předmětem aktivního výzkumu. Pro modelování shlukových procesů jsou nejčastěji využívány Coxovy bodové procesy. Odhady metodou maximální věrohodnosti či bayesovské odhady jsou pro tuto třídu modelů obvykle výpočetně velmi náročné, a proto bylo v literatuře navrženo několik alternativních metod odhadu založených na momentových vlastnostech uvažovaných procesů.
V první části práce podáváme přehled dostupných momentových metod odhadu parametrů pro stacionární prostorové Coxovy bodové procesy a pomocí simulační studie porovnáváme jejich úspěšnost. Dále rozebíráme možnosti zobecnění těchto metod pro nestacionární prostorové bodové procesy.
Ve druhé části práce se zaměřujeme na odhady metodou minimálního kontrastu pro nestacionární časoprostorové shot-noise Coxovy bodové procesy a zabýváme se využitím projekcí procesu do prostorové a časové domény k postupnému odhadu parametrů jednotlivých částí modelu. Navrhujeme vícekrokovou metodu odhadu využívající projekcí procesu a také vylepšenou metodu, jež řeší potenciální problém s překrýváním jednotlivých shluků.
Pro obě metody ukazujeme konzistenci a asymptotickou normalitu výsledných odhadů v asymptotice rostoucích pozorovacích oken a provádíme jejich srovnání s dříve publikovanou metodou pomocí simulační studie. Díky tomu jsou obě metody připraveny k použití v praxi.

Klíčová slova: Coxův bodový proces, časoprostorový bodový proces, momentové metody odhadu parametrů, metoda minimálního kontrastu, projekce procesu

Title: Statistical inference for spatial and space-time Cox point processes

Author: Jiří Dvořák

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michaela Prokešová, Ph.D.

Abstract: Fitting of parametric models to spatial and space-time point patterns has been a very active research area in the last few years. Concerning clustered patterns, the Cox point process is the model of choice. To avoid the computationally demanding maximum likelihood estimation or Bayesian inference, several estimation methods based on the moment properties of the processes in question

were proposed in the literature.
We give overview of the state-of-the-art moment estimation methods for stationary spatial Cox point processes and compare their performance in a simulation study. We also discuss generalization of such methods for inhomogeneous spatial point processes.
In the core part of the thesis we focus on minimum contrast estimation for inhomogeneous space-time shot-noise Cox point processes and investigate the possibility to use projections to the spatial and temporal domain to estimate different parts of the model separately. We propose a step-wise estimation procedure based on projection processes and also a refined method which remedies the problem of possible cluster overlapping.
We establish consistency and asymptotic normality of the estimators for both methods under the increasing window asymptotics and compare their performance on middle-sized observation windows by means of a simulation study. This makes the methods ready-to-use in practice.

Keywords: Cox point process, space-time point process, moment estimation methods, minimum contrast estimation, projection process

Contents

List of Notation

$\mathbb{Z}$	set of integers
$\mathbb{N}$	set of positive integers
$\mathbb{N}_0$	set of non-negative integers
$\mathbb{R}$	set of real numbers (one-dimensional Euclidean space)
$\mathbb{R}^d$	d-dimensional Euclidean space
$\mathcal{B}^d$	Borel sets in $\mathbb{R}^d$
$\mathcal{B}_0^d$	bounded Borel sets in $\mathbb{R}^d$
$\mathcal{B}(x, r)$	the closed ball centered at $x \in \mathbb{R}^d$ with radius $r \geq 0$
ω_d	volume of a unit ball in $\mathbb{R}^d$
σ_d	surface area of a unit ball in $\mathbb{R}^d$
$\lvert A \rvert$	appropriate Hausdorff measure of the set A
$A \oplus r$	set A dilated by distance r
$A \ominus r$	set A eroded by distance r
A_u	set A shifted by the vector u
$\lVert u \rVert$	Euclidean norm of the vector u
u^T	transposition of the vector u
o	the origin in $\mathbb{R}^d$
I	indicator function
$\mathbf{1}$	identity matrix (of appropriate dimension)
$(\Omega, \mathcal{A}, \mathbb{P})$	abstract probability space
X	point process
$X(B)$	number of points of X occuring in the set B
W	observation window
α_k	kth-order factorial moment measure
λ_k	kth-order intensity function
λ	intensity or (first-order) intensity function
$\lambda^{(k)}$	k-th order derivative of λ w.r.t. parameters
g	pair-correlation function
K	K-function (reduced second-order moment function)
λ_o, λ_x	intensity function of Palm distribution at location o or x
$o_P(1)$	sequence of random variables converging to 0 in probability
$O_P(1)$	sequence of random variables bounded in probability
SOIRS	second-order intensity reweighted stationary (process)
MSE	mean squared error
MCEK, MCEg	minimum contrast estimation using K-function and pair-correlation function g, respectively
CLE, PLE	composite likelihood and Palm likelihood estimation

List of Tables

Introduction

Many fields of science deal with data that are point patterns, such as positions of trees in a rain forest (Waagepetersen, 2007), maps of disease cases (Diggle et al., 2007, Beneš et al., 2011) or the locations of point-like defects in industrial materials (Ohser and Mücklich, 2000). The necessity to analyse such datasets lead to significant development of the point process theory in the past decades, see e.g. Diggle (1983), Ripley (1988), Stoyan et al. (1995), Diggle (2003) or Illian et al. (2008).

Also, spatial point processes play a fundamental role in the random set theory (Stoyan et al., 1995) and image analysis (Serra, 1982). Other areas of application of point process models include—but are not limited to—astronomy (Babu and Feigelson, 1996, Kerscher, 2000), forest fire modelling (Peng et al., 2005, Møller and Díaz-Avalos, 2010) and earthquake modelling (Musmeci and Vere-Jones, 1992, Ogata, 1998). Note that the last four references concerning earthquake and forest fire modelling in fact deal with space-time point process models, the time being either discrete or continuous variable.

Parametric point process models enable detailed statistical inference which is necessary for analysis of real datasets. The preferred model for clustered point patterns is the Cox point process. Usually, finding maximum likelihood estimates for Cox point process models is computationally very demanding – to obtain the estimates it is needed to repetitively evaluate mean values of complicated high-dimensional integrals which are part of the normalizing constant in the likelihood as illustrated in Equation (2.2) in Chapter 2. One can of course take advantage of Markov chain Monte Carlo methods or other techniques (Møller and Waagepetersen, 2004) but this does not reduce the computational complexity of the estimation procedure.

For wide range of Cox process models, such as stationary Poisson-Neyman-Scott process (Neyman and Scott, 1958) or stationary shot-noise Cox process (Møller, 2003), the second-order moment characteristics, such as the pair-correlation function or the K-function (Illian et al., 2008), are available in a closed form. The same holds also for non-stationary versions of the processes, under the assumption of so-called second-order intensity reweighted stationarity (SOIRS), see Baddeley et al. (2000). This enables inference for the parameters determining the range and strength of clustering.

The estimation methods based on moment properties of the process were first introduced for stationary spatial Cox point processes and are traditionally called moment estimation methods in the literature. They provide a faster, simulation-free alternative to the maximum likelihood estimation. They include minimum contrast estimation (Diggle, 1983, 2003), composite likelihood method (Guan, 2006) and Palm likelihood method (Tanaka et al., 2008). In Section 2.1 we give a detailed description of the methods.

For all the above-mentioned moment estimation methods—with the exception of the minimum contrast estimation based on the pair-correlation function—consistency and asymptotic normality of the estimators have been proved under suitable assumptions, see the references given in Section 2.1. However, the theoretical expressions for the asymptotic variance of the estimators are too compli-

cated to be used for direct comparison of the efficiency of the estimators. Moreover, the convergence may be too slow and the bias can be a more important issue than the variance for certain point processes observed on medium sized windows. Therefore we decided to assess the empirical performance of the discussed moment estimation procedures by a simulation study. The study was published in Dvořák and Prokešová (2012) and the results are presented in Section 2.1.5.

The moment estimation methods for stationary spatial Cox point processes can be extendend also to SOIRS processes. Brief discussion is provided in Section 2.2 with more details given in the paper Prokešová et al. (2014).

The core of this thesis lies in Chapter 3 which concerns estimation for SOIRS space-time Cox point process models. The temporal coordinate plays a distinct role and thus the estimation methods from purely spatial case cannot be employed directly (e.g. using the Euclidean distance to describe differences between space-time events is not appropriate or natural). Hence, devoted space-time methods should be used for statistical inference.

The discussion on high computational demands of maximum likelihood estimation also applies here for SOIRS space-time Cox point processes and, in general, higher number of parameters appear in the model and need to be estimated, thus increasing dimensionality of the respective optimization problems. Hence we restrict our attention to minimum contrast estimation and we investigate the possibility of using step-wise estimation and dimensionality-reducing techniques to estimate different parts of the model separately. Namely we focus on using the projections of the space-time process with a particular non-trivial structure into the spatial and temporal domain, respectively (see Section 1.4 for definition of the projection processes). This work is inspired by the paper Møller and Ghorbani (2012).

In Section 3.1 we propose a step-wise estimation procedure using the projection processes and minimum contrast estimation, see also Prokešová and Dvořák (2014). We establish consistency and asymptotic normality of the resulting estimator under so-called increasing window asymptotics (Section 3.1.3). It means that the asymptotic regime is governed by an increasing sequence of compact observation windows which grow unboundedly, i.e. the data are observed in always larger and larger region and longer and longer time interval. We also compare our proposed method to the original method of Møller and Ghorbani (2012) by means of a simulation study presented in Section 3.1.5 and conclude that for most of the models considered in the simulation study our method performs better in terms of the relative mean squared error.

The method using projection processes suffers from the problem of cluster overlapping, see the discussion at the end of Section 3.1.5. To remedy this, in Section 3.2 we propose a refined estimation method. It avoids projection of the space-time point process to the temporal domain, i.e. the situation where the most information is lost by the projection and which may sometimes result in severe overlapping of clusters which were originally distinct in the space-time domain. The overlapping negatively affects mainly the estimates of parameters governing the clustering in the temporal domain.

Again, we establish consistency and asymptotic normality of the refined estimator under the increasing window asymptotics and under appropriate assumptions on moment properties and mixing properties of the point process in question,

see Section 3.2.3. These results provide theoretical justification of the estimation method in the sense that using more information from larger and larger observation windows increases the precision of the estimates. The asymptotic results also enable construction of confidence regions and hypotheses testing for the model parameters in practical applications.

We also illustrate in the simulation study presented in Section 3.2.4 that the problem of cluster overlapping has been sucessfully overcome by the refined method and that the precision of estimates of the temporal clustering parameters has improved considerably when compared to the method using projection processes.

The main contribution of this thesis to the field of statistical inference for point processes lies in the development of novel estimation methods for non-stationary space-time shot-noise Cox point process models and in establishing asymptotic properties of the estimators. This enables construction of confidence regions and hypotheses testing and makes the proposed methods ready-to-use in practical situations.

Finally, Chapter 4 discusses briefly several interesting open questions raised by the work described in this thesis and indicates possible directions of further research. However, answering such questions lies outside the scope of this thesis.

1. Background on spatial and space-time point processes

In this chapter we introduce the basic concepts and notation used in the subsequent chapters.

We restrict our attention to point processes in the Euclidean space $\mathbb{R}^d$ for some positive integer d. This is the most common setting in applications, with either $d = 2$ or $d = 3$. When discussing space-time point processes we consider point processes on the space $\mathbb{R}^d \times \mathbb{R}$, i.e. Euclidean space with one distinct coordinate having the role of time.

For introduction to the theory of point processes on more general spaces (complete separable metric spaces), see e.g. Daley and Vere-Jones (2008).

1.1 General theory

This section gives several standard and fundamental definitions relating to the point process theory. For more details, the classical textbooks such as Daley and Vere-Jones (2008), Illian et al. (2008) or Møller and Waagepetersen (2004) can be consulted.

For ease of exposition we now focus on spatial point processes. The case of space-time point processes will be discussed later in Section 1.4.

Definition 1.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an abstract probability space and $\mathcal{N}$ the system of locally finite subsets of $\mathbb{R}^d$, i.e. $\mathcal{N} = \left\{N \subset \mathbb{R}^d : \#(N \cap B) < \infty \; \forall B \in \mathcal{B}_0^d\right\}$. Let $\mathcal{N}$ be equipped with the σ-algebra $\mathfrak{N} = \sigma\left\{U_{B,m} : m \in \mathbb{N}_0, B \in \mathcal{B}_0^d\right\}$ where $U_{B,m} = \{N \in \mathcal{N} : \#(N \cap B) = m\}$. **Point process** X **in** $\mathbb{R}^d$ is a measurable mapping $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{N}, \mathfrak{N})$. Sample realization of X is called a **point pattern**.

Examples of point patterns with different type of interaction between points (repulsion, aggregation, no interaction) are shown in Figure 1.1. They illustrate the diversity of possible point patterns and the underlying point processes.

Our definition has a formally equivalent alternative: X can be viewed as a locally finite random counting measure satisfying $X(\{y\}) \leq 1 \; \forall y \in \mathbb{R}^d$. To see this it is enough to identify a locally finite set $A \subset \mathbb{R}^d$ with the measure $\sum_{x \in A} \delta_x$ where δ_x is the Dirac measure with atom at point $x \in \mathbb{R}^d$.

In the subsequent chapters it will be more convenient to treat point processes as random locally finite subsets of $\mathbb{R}^d$ and their realizations as configurations of points in $\mathbb{R}^d$. However, we will also take advantage of the convenient notation corresponding to the measure-theoretic approach. Namely the symbol $X(B)$ will denote the number of points of X occuring in the set B

Note also that the Definition 1.1 allows only so-called *simple* point processes, i.e. no two points may occur at the same location. For more general definition of a point process as a locally finite random counting measure (admitting multiplicity of points) see Daley and Vere-Jones (2008).

When observing a point pattern in applications it is not possible to observe a realization of the process in the whole $\mathbb{R}^d$. Thus, a fundamental role is played

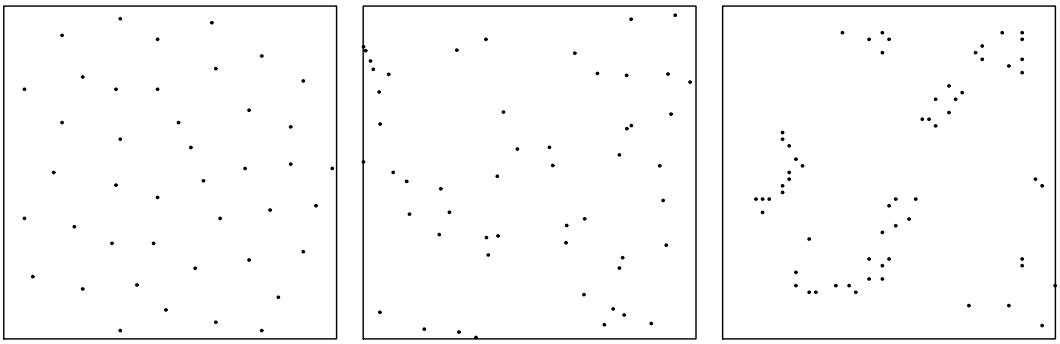


Figure 1.1: Point patterns illustrating different types of interaction between points of the process, i.e. repulsion, no interaction and aggregation. *Left:* the locations of the centers of 42 biological cells observed under optical microscopy in a histological section (*cells* dataset). *Middle:* simulated realization of a Poisson point process, see Example 1.12 below. *Right:* the locations of 62 seedlings and saplings of California redwood trees in a square sampling region (*redwood* dataset). The two real point patterns are standard datasets available in the `spatstat` package for `R` (Baddeley and Turner, 2005).

by the *observation window*. It is a compact set $W \in \mathcal{B}^d$ (with positive Lebesgue measure) which corresponds to the region of study. The shape, size and position of the observation window W is (and always must be) an integral part of the recorded dataset, together with the locations of individual points.

Definition 1.2. A point process X on $\mathbb{R}^d$ is called **stationary** if its distribution is invariant w.r.t. translations in $\mathbb{R}^d$, i.e. the distribution of the shifted process $X + y = \{x + y : x \in X\}$ is the same as the distribution of X itself for all $y \in \mathbb{R}^d$.

Stationary point processes are also often called *homogeneous* in purely spatial contexts, as noted in Daley and Vere-Jones (2008, Chap. 15.1).

Definition 1.3. A point process X on $\mathbb{R}^d$ is called **isotropic** if its distribution is invariant w.r.t. rotations around the origin in $\mathbb{R}^d$, i.e. the distribution of the rotated process $\mathcal{O}X = \{\mathcal{O}x : x \in X\}$ is the same as the distribution of X itself for all rotations $\mathcal{O}$ around the origin in $\mathbb{R}^d$.

Definition 1.4. Let X be a point process on $\mathbb{R}^d$. Its **kth-order factorial moment measure** α_k is defined as

$$\alpha_k(A) = \mathbb{E}\left(\sum_{u_1,\ldots,u_k \in X}^{\neq} I\left[(u_1,\ldots,u_k) \in A\right]\right),\ A \in \left(\mathcal{B}^d\right)^k,$$

where $\neq$ denotes that the summation is over k-tuples of distinct points of X and I is the indicator function.

Note that the measure $\alpha_1(\cdot) = \mathbb{E}\, X(\cdot)$ is often called the *intensity measure*.

Definition 1.5. Consider a point process X on $\mathbb{R}^d$. If its kth-order factorial moment measure α_k has a density w.r.t. the Lebesgue measure on $\left(\mathbb{R}^d\right)^k$ it is

denoted λ_k and called the **kth-order product density of X** or the **kth-order intensity function of X**.

The first-order intensity function λ_1 is often called simply the *intensity function* and denoted λ. For stationary point processes on $\mathbb{R}^d$ the intensity function is constant. Hence, with a slight abuse of notation common in the point process literature, one may write $\lambda(x) = \lambda > 0,\ x \in \mathbb{R}^d$. The constant λ is usually called the *intensity* of the stationary point process X. Its non-parametric estimate is easily obtained as

$$\widehat{\lambda} = X(W)/|W|. \tag{1.1}$$

For an inhomogeneous point process a non-parametric estimate of the intensity function λ can be obtained by kernel smoothing, i.e.

$$\widehat{\lambda}(u) = \sum_{x \in X \cap W} \frac{h_b(u-x)}{w_{b,W}(x)},\ u \in W,$$

where h_b is a kernel function with bandwidth $b > 0$ and the edge correction factors $w_{b,W}$ are given by

$$w_{b,W}(x) = \int_W h_b(u-x)\,\mathrm{d}u,$$

so that $\int_W \widehat{\lambda}(u)\,\mathrm{d}u = X(W)$. This implies approximate unbiasedness of the estimate $\widehat{\lambda}$. Also, if a parametric model for the intensity function λ is available it is possible to obtain parametric estimates of λ as illustrated in Section 2.2.

The following *Campbell theorem* (Daley and Vere-Jones, 2008, Sec. 9.5) is one of the fundamental tools of point process theory and is often used in direct calculations of certain expectations.

Theorem 1.6. *Let X be a point process and assume that its kth-order factorial moment measure α_k and the kth-order intensity function λ_k exist. For any integrable Borel-measurable function h it holds that*

$$\begin{aligned}\mathbb{E} \sum_{x_1,\ldots,x_k \in X}^{\neq} h(x_1,\ldots,x_k) &= \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} h(u_1,\ldots,u_n)\,\alpha_k\,(\mathrm{d}u_1,\ldots,\mathrm{d}u_k) \\ &= \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} h(u_1,\ldots,u_n)\lambda_k(u_1,\ldots,u_n)\,\mathrm{d}u_1\ldots\mathrm{d}u_k.\end{aligned}$$

Definition 1.7. Let X be a point process on $\mathbb{R}^d$. If both λ and λ_2 exist we define the **pair-correlation function** g (sometimes called simply the **g-function**) by the formula

$$g(x,y) = \frac{\lambda_2(x,y)}{\lambda(x)\lambda(y)},\ x,y \in \mathbb{R}^d : \lambda(x) > 0, \lambda(y) > 0.$$

If the point process X is stationary with intensity $\lambda > 0$ we can write, again with slight abuse of notation, the pair-correlation function as a function of a single argument:

$$g(x,y) = \frac{\lambda_2(x,y)}{\lambda(x)\lambda(y)} = \frac{\lambda_2(0,y-x)}{\lambda^2} = g(y-x),\ x,y \in \mathbb{R}^d.$$

Similarly, if the point process X is both stationary and isotropic the pair-correlation function can be written as a function of a single scalar argument:

$$g(x, y) = g(\|y - x\|),\ x, y \in \mathbb{R}^d.$$

In the case of stationary isotropic process the pair-correlation function can be easily plotted. This enables visual assessment of the tendencies of the process towards clustering or repulsion on different scales. The value $g(r) = 1$ corresponds to no interaction at the spatial lag r (cf. the Poisson process in Example 1.12) while values $g(r) > 1$ and $g(r) < 1$ indicate clustering and repulsion at the given scale, respectively. See Section 2.1 for examples.

Definition 1.8. Let X be a stationary point process on $\mathbb{R}^d$ with intensity $\lambda > 0$ and $A \in \mathcal{B}_0^d$ be an arbitrary set with positive Lebesgue measure $|A|$. The **reduced second-order moment function** K or simply the **K-function** is defined by the formula

$$K(r) = \mathbb{E} \sum_{x,y \in X}^{\neq} \frac{I\left[x \in A, \|x - y\| \leq r\right]}{\lambda^2 |A|},\ r \geq 0.$$

Note that the definition does not depend on the choice of A. This can be easily shown using the Campbell-Mecke theorem (Møller and Waagepetersen, 2004, App. C.2). The K-function is sometimes also called *Ripley's K-function* as it was introduced in the paper by Ripley (1976).

Equivalent alternative definition of the K-function of a stationary process can be made using the *Palm distribution* of the process (Daley and Vere-Jones, 2008, Chap. 13). This approach provides a simple interpretation of the values of K-function: the quantity $\lambda K(r)$ is the mean number of further points within distance r from the typical point of the process X.

For a stationary process X for which λ_2 exists (and thus the pair-correlation function g is properly defined) there is a simple relationship between K and g:

$$K(r) = \int_{B(o,r)} g(u)\,\mathrm{d}u,\ r \geq 0, \tag{1.2}$$

where $B(o, r)$ is the closed ball with radius r centered in the origin o of $\mathbb{R}^d$. If the point process X is also isotropic the formula can be further simplified using spherical coordinates and the values of g can be recovered from the K-function by differentiation:

$$K(r) = \sigma_d \int_0^r s^{d-1} g(s)\,\mathrm{d}s,\ r \geq 0, \tag{1.3}$$

where σ_d is the surface area of a unit sphere in $\mathbb{R}^d$.

It is possible to extend the definition of the K-function to inhomogeneous point processes fulfilling a certain property (translation invariance of the pair-correlation function). The following definition is due to Baddeley et al. (2000).

Definition 1.9. Let X be a point process on $\mathbb{R}^d$ with well-defined pair-correlation function $g(x, y), x, y \in \mathbb{R}^d$. If there is a function $g' : \mathbb{R}^d \to \mathbb{R}$ such that

$$g(x, y) = g'(y - x),\ x, y \in R^d,$$

then X is called **second-order intensity reweighted stationary** (SOIRS). In this case, the **inhomogeneous K-function** is defined as

$$K(r) = \int_{B(o,r)} g'(u)\,\mathrm{d}u,\ r \geq 0.$$

With a slight abuse of notation we usually write simply $g(y-x)$ instead of $g'(y-x)$, as is common in the literature.

Note that stationary point processes with well-defined pair-correlation function possess the SOIRS property. Another example of SOIRS point processes are processes obtained by location-dependent thinning from stationary processes, as described below in Definition 1.10. However, stationary and thinned processes are not the only types of processes with the SOIRS property. Baddeley et al. (2000, Sec. 2.1) give an example of such a process.

The definition of the inhomogeneous K-function makes sense only for SOIRS processes. If the process in question is stationary (and hence SOIRS), the inhomogeneous K-function coincides with the original Ripley's K-function from Definition 1.8.

The inhomogeneous K-function in fact expresses weighted average of the number of points in the vicinity of the points of the process. The weights constitute of reciprocals of the intensity of the process at the respective locations.

The plot of K-function of a particular process can be used to demonstrate the tendencies towards clustering or repulsion. Values higher than those of the benchmark Poisson process indicate clustering while lower values indicate repulsive interactions between points.

A non-parametric estimate of the K-function for a stationary or SOIRS point process X observed in a compact observation window W can be obtained by counting pairs of points in the following way:

$$\widehat{K}(r) = \frac{1}{|W|} \sum_{x,y \in X \cap W}^{\neq} \frac{I\left[\|x-y\| \leq r\right]}{w(x,y)\widehat{\lambda}(x)\widehat{\lambda}(y)},\ r \geq 0, \tag{1.4}$$

where $\widehat{\lambda}$ is an estimate of the intensity function and $w(x,y)$ is an edge correction factor (for a detailed discussion on different choices of edge correction factors see Gabriel, 2014).

On the other hand, a non-parametric estimate of the pair-correlation function of an isotropic stationary or SOIRS point process is obtained by kernel smoothing, i.e.

$$\widehat{g}(r) = \frac{1}{|W|} \sum_{x,y \in X \cap W}^{\neq} \frac{h_b(\|x-y\| - r)}{\sigma_d r^{d-1} w(x,y)\widehat{\lambda}(x)\widehat{\lambda}(y)},\ r \geq 0, \tag{1.5}$$

where $w(x,y)$ is an edge correction factor and h_b is a kernel function (e.g. the Epanechnikov kernel) with bandwidth $b > 0$. The bandwidth needs to be specified by the user or a data-driven cross-validation procedure for choosing optimal bandwidth may be employed (Guan, 2007*a*,*b*).

Figure 1.2 shows the empirical estimates of the K-function and the pair-correlation function g for the point patterns from Figure 1.1, i.e. a regular pattern exhibiting repulsion between points, simulated realization of a Poisson process

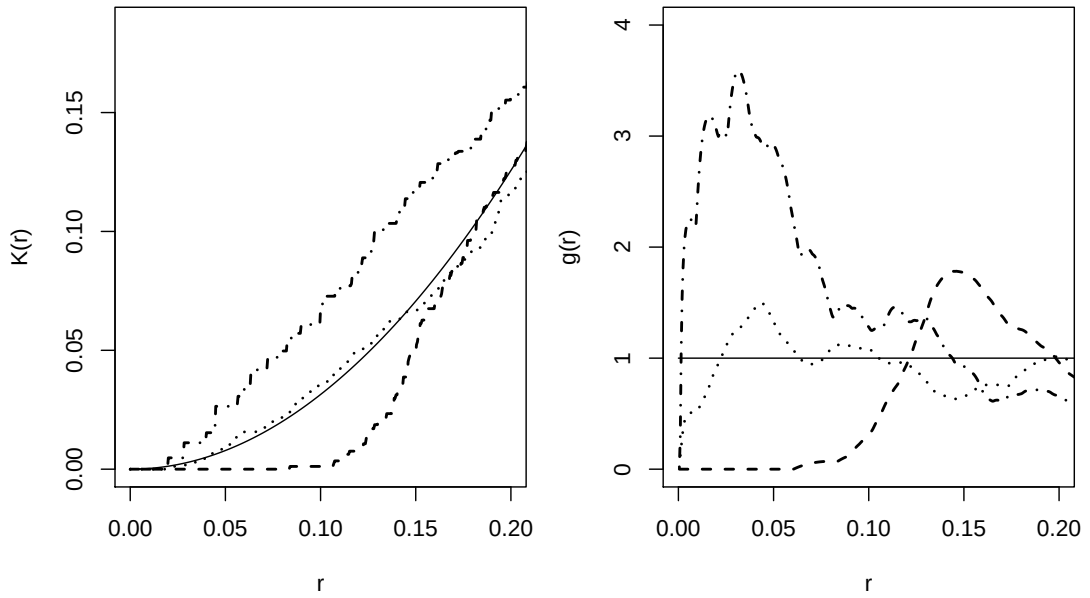


Figure 1.2: Empirical estimates of the K-function (left) and the pair-correlation function g (right) for the point patterns from Figure 1.1, i.e. a regular pattern (*cells* dataset, dashed line), a simulated realization of a Poisson process (no interaction between points, dotted line) and a clustered point pattern (*redwood* dataset, dot-dashed line). For comparison, the theoretical value of K and g corresponding to a Poisson process is also plotted (solid line).

with no interaction between points and an aggregated point pattern showing clustering of the points. We assume that the point patterns result from stationary point processes. For comparison, the theoretical value of the K-function and g-function corresponding to a Poisson process is also plotted. We recall that values of K and g above (below) the values for a Poisson process indicate clustering (repulsion) of the points.

A common way to introduce inhomogeneity to a point process model is so-called *location-dependent thinning* of a stationary point process. This procedure is described in the following definition.

Definition 1.10. Consider a stationary point process X_0 on $\mathbb{R}^d$ with intensity $\lambda > 0$. Furthermore, let $f : \mathbb{R}^d \to [0, 1]$ be a function on $\mathbb{R}^d$. Let X be the point process defined in the following way:

$$X = \{x \in X_0 : U(x) \leq f(x)\},$$

where $U(x)$ are independent, identically distributed random variables independent of X_0 with uniform distribution on $[0, 1]$. Then X is a **thinned version of** X_0 with **retention probabilities** $f(x)$. We usually call f the **inhomogeneity function**.

Direct consequence of this definition is the following fact. If the kth-order intensity function $\lambda_{0,k}$ of the stationary process X_0 exists, so does the kth-order

intensity function λ_k of the thinned process X and it has the form

$$\lambda_k(u_1, \dots, u_k) = \lambda_{0,k}(u_1, \dots, u_k) \prod_{i=1}^{k} f(u_i),\ u_1, \dots, u_k \in \mathbb{R}^d.$$

This in turn means that the pair-correlation functions of X and X_0 are the same. Furthermore, this implies translation invariance of the pair-correlation function and hence SOIRS property of the thinned process X.

Even though the location-dependent thinning is the most popular type of inhomogeneity considered in applications, other types of inhomogeneity are possible, see e.g. Prokešová (2010).

For stationary point processes the factorial moment measures of order at least 2 are invariant under diagonal shifts. This is the cornerstone of the following definition.

Definition 1.11. Let X be a stationary point process with intensity $\lambda > 0$. The **reduced second-order factorial moment measure** α_2^{red} is uniquely determined by the disintegration formula

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(u_1, u_2)\, \alpha_2\,(\mathrm{d}u_1, \mathrm{d}u_2) = \lambda \int_{\mathbb{R}^d} h(u_1, u_2 + u_1)\, \alpha_2^{\mathrm{red}}\,(\mathrm{d}u_2)\ \mathrm{d}u_1 \tag{1.6}$$

for any integrable Borel-measurable function h.

We denote λ_o the density of α_2^{red} (assuming it exists) and call it the **Palm intensity**.

The term *Palm intensity* arises from the fact that λ_o is in fact the intensity function of the Palm distribution $\mathcal{P}_0$ of the process X. For details, see Daley and Vere-Jones (2008, Prop. 13.2.VI).

If the second-order intensity function λ_2 of X exists the disintegration formula (1.6) implies that

$$\lambda_2(x, y) = \lambda_2(0, y - x) = \lambda \cdot \lambda_o(y - x),\ x, y \in \mathbb{R}^d,$$

providing a simple expression for λ_o in terms of intensity functions up to second order.

Similarly, in the SOIRS case we can decompose λ_2 in the following way:

$$\lambda_2(x, y) = \lambda(x)\lambda(y)g(y - x) = \lambda(x)\lambda_x(y) = \lambda(y)\lambda_y(x),\ x, y \in \mathbb{R}^d,$$

where $\lambda_x(\cdot)$ is the intensity function of the Palm distribution of X conditioned by the event that a point of X occurs in the location x. For more details on Palm theory see Daley and Vere-Jones (2008, Sec. 13).

1.2 Models of clustered point patterns

This section introduces classical models suitable for modelling clustered point patterns. These models will be used in the subsequent chapters.

Example 1.12. (Poisson process, Møller and Waagepetersen (2004, Sec. 3.1)) The Poisson process is a benchmark model which exhibits no interaction between the points of the process. It is characterized by its intensity function λ (which needs to be locally integrable) and the following properties:

- for every $B \in \mathcal{B}^d$ for which $\alpha_1(B) = \int_B \lambda(u)\,\mathrm{d}u < \infty$ it holds that $X(B)$ is a random variable with Poisson distribution with mean $\alpha_1(B)$,
- for every $k = 2, 3, \ldots$ and pairwise disjoint sets $B_1, \ldots, B_k \in \mathcal{B}^d$ the random variables $X(B_1), \ldots, X(B_k)$ are independent.

Stationary Poisson process with intensity λ is often referred to as *homogeneous* Poisson process or, mostly in ecological or biological applications, the model of *complete spatial randomness*.

The moment properties of the Poisson process are very simple. The most important for us are the following, which hold both for homogeneous and inhomogeneous Poisson process:

$$\begin{aligned}
&\lambda_k(u_1, \ldots, u_k) = \prod_{i=1}^{k} \lambda(u_i),\ u_1, \ldots, u_k \in \mathbb{R}^d,\\
&g(x, y) = 1,\ x, y \in \mathbb{R}^d,\\
&K(r) = \omega_d\, r^d,\ r \geq 0,
\end{aligned}$$

where $\omega_d = |B(o, 1)|$ is the volume of a unit ball in $\mathbb{R}^d$.

Note that the definition of the Poisson process in fact relies only on α_1. It is possible to state the definition without assuming that the intensity function λ exists but it is not useful for us. In the following chapters we will always assume that the processes in question have an intensity function.

Independence properties of the Poisson process single it out as a benchmark process towards which the tendencies for clustering or regularity of other processes are compared. Values of g- or K-function higher (lower) than those corresponding to the Poisson process indicate clustering (repulsion) of the points.

Example 1.13. (Cox process, Møller and Waagepetersen (2004, Sec. 5.1))
A natural generalization of the Poisson process (which is characterized by its intensity function λ and two properties stated in Example 1.12) is so-called Cox process or doubly-stochastic process for which the intensity function is random. This model was first introduced by Cox (1955) and for its flexibility it is the model of choice in most applications for modelling clustering.

Let $\Lambda = \{\Lambda(x), x \in \mathbb{R}^d\}$ be a non-negative random field such that $x \to \Lambda(x)$ is a locally integrable function with probability 1. If the conditional distribution of point process X given $\Lambda = \lambda$ is a Poisson process on $\mathbb{R}^d$ with intensity function λ, then X is said to be a Cox process with the driving field Λ.

We remark here that when observing a single point pattern in practice, it is impossible to distinguish whether the data are generated by a Cox process (with random intensity function $\Lambda = \lambda$ in this particular case) or simply by Poisson process (with intensity function λ). This is a problem for choosing the right model when analyzing an observed point pattern.

For a general Cox point process we can express the kth-order intensity function λ_k (if it exists) as the following expectation:

$$\lambda_k(u_1, \ldots, u_k) = \mathbb{E}\left[\prod_{i=1}^{k} \Lambda(u_i),\ u_1, \ldots, u_k \in \mathbb{R}^d\right].$$

As a consequence, a Cox point process is stationary if and only if the driving field Λ is stationary.

For most specific models of Cox point processes the values of the pair-correlation function g are ≥ 1 but Møller and Waagepetersen (2004, Sec. 5.6.2) provide an example showing that this is not always the case.

Example 1.14. (log-Gaussian Cox process, Møller and Waagepetersen (2004, Sec. 5.6)) This class of Cox point processes provides broad range of models for aggregated point patterns. Let $Z(x), x \in \mathbb{R}^d$ be a Gaussian random field. A Cox point process driven by the random intensity $\Lambda(x) = \exp\{Z(x)\}, x \in \mathbb{R}^d$ is called a log-Gaussian Cox process. Clusters of points then appear at locations with high values of Λ.

The model was introduced independently by Coles and Jones (1991) in the field of astronomy and by Møller et al. (1998) in statistics. It is fully characterized by the mean and covariance function

$$m(x) = \mathbb{E}\, Z(x), \qquad c(x,y) = \mathbb{C}\mathrm{ov}(Z(x), Z(y)),\ x, y \in \mathbb{R}^d.$$

The range of possible mean and covariance functions makes this class of models very flexible and useful in applications, see Chap. 5.6 in Møller and Waagepetersen (2004) and the references therein.

The intensity function and the pair-correlation function can be expressed as simple functions of m and c, namely

$$\lambda_1(x) = \exp\{m(x) + c(x,x)/2\}, \qquad g(x,y) = \exp\{c(x,y)\},\ x, y \in \mathbb{R}^d.$$

This will enable parametric model fitting for log-Gaussian Cox processes in Section 2.1.

Example 1.15. (Poisson-Neyman-Scott pr., Møller and Waagepetersen (2004, Sec. 5.3)) Unlike log-Gaussian Cox process, the definition of Poisson-Neyman-Scott point process works with *individual clusters* of points in a specific "mother–daughter" representation. A model of this type was first presented in Neyman and Scott (1958).

Consider a stationary Poisson process C on $\mathbb{R}^d$ with intensity $\kappa > 0$. This process represents the *mother points*. Each mother point $c \in C$ produces a process X_c of *daughter points*. Conditional on C, let $X_c, c \in C$, be independent Poisson processes with intensity functions

$$\lambda_c(x) = \nu\, k(x - c),\ x \in \mathbb{R}^d,$$

where $\nu > 0$ is a parameter and k is a probability density function. This means that the total number of points $X_c(\mathbb{R}^d)$ in cluster X_c has Poisson distribution with mean ν.

The process X then consists of the daughter points only, i.e. $X = \bigcup_{c \in C} X_c$. Note that X is stationary and in fact it is a Cox process with the driving field Λ of the form

$$\Lambda(x) = \sum_{c \in C} \nu\, k(x - c),\ x \in \mathbb{R}^d.$$

Such driving field Λ is stationary and locally integrable. It is isotropic if and only if k is isotropic.

We remark here that the Poisson-Neyman-Scott processes considered here constitute a special class of more general *Neyman-Scott processes* (Neyman and Scott, 1958). In the general case, the daughter points are distributed independently around each mother point according to the density k but the distribution of $X_c(\mathbb{R}^d)$ does not need to be Poisson. Hence, in general Neyman-Scott processes are not Cox point processes.

For a stationary Poisson-Neyman-Scott process X it holds that

$$\lambda = \kappa\,\nu, \qquad g(x,y) = 1 + h(y-x)/\kappa,\ x, y \in \mathbb{R}^d,$$

where $h(u) = \int k(\tau)k(u+\tau)\,\mathrm{d}\tau$.

Two most common models of this class are so-called *modified Thomas process* (with k being the probability density function of radially symmetric normal distribution with standard deviation $\sigma > 0$, see Thomas, 1949) and *Matérn process* (with k corresponding to the uniform distribution on the ball with radius r centered at the origin, see Matérn, 1986).

Apart from location-dependent thinning, there are different types of inhomogeneity that can be considered for the Neyman-Scott processes, see e.g. Mrkvička (2014) or Prokešová (2010).

1.3 Shot-noise Cox processes

In the present section we introduce the class of so-called *shot-noise Cox processes*, see Møller and Waagepetersen (2004, Sec. 5.4). Since this class of models is crucial for the subsequent chapters we devote a full section to it, as opposed to the previous examples.

Following Møller (2003) we generalize the Poisson-Neyman-Scott processes from Example 1.15 in the following way. Let the driving field Λ be of the form

$$\Lambda(x) = \sum_{(r,v)\in\Phi} r\,k(x,v),\ x \in \mathbb{R}^d,$$

where Φ is a Poisson process on $(0,\infty)\times\mathbb{R}^d$ with intensity measure U and k is now a smoothing kernel, i.e. a non-negative function integrable in both coordinates. Under some basic integrability assumptions $\Lambda(x)$ is an almost surely locally integrable random field and the corresponding point process X is a well-defined Cox process, see Møller (2003) or Hellmund et al. (2008).

Note that the Poisson-Neyman-Scott processes in Example 1.15 fit into the current framework and constitute an important (sub)class of shot-noise Cox processes.

The process X driven by the random field $\Lambda(x)$ is stationary if the kernel k is function of the difference of its arguments, i.e. $k(x,v) = k(v-x)$, and the intensity measure U has the following product form:

$$U(\mathrm{d}r, \mathrm{d}v) = \mu\,V(\mathrm{d}r)\,\mathrm{d}v. \tag{1.7}$$

Here $\mu > 0$ and $V(\mathrm{d}r)$ is an arbitrary measure on $(0,\infty)$ satisfying the integrability condition $\int_0^\infty \min(1,r)V(\mathrm{d}r) < \infty$.

A shot-noise Cox point process can be viewed as a generalized cluster process. We can rewrite X as $X = \bigcup_{(r,v)\in\Phi} X_v$ where X_v is the cluster centered at location v. Conditionally on Φ, the cluster processes X_v are independent Poisson processes with intensity function $r\,k(\cdot, v)$, i.e. r corresponds to the *weight* of the cluster. Note that even for a compact set $A \in \mathcal{B}^d$ the number of cluster centers in A may be infinite almost surely, see Møller (2003) and Hellmund et al. (2008) for details. This also brings severe complications to simulation of such processes. However, these can be overcome as described in Møller (2003, App. D).

In the stationary case and under the condition $\int_0^\infty r\,V(\mathrm{d}r) < \infty$, almost surely only finitely many clusters will have at least one point. If we condition by the positions of the points only and assume for simplicity that k is a probability density function, then the shifted cluster processes $(X_v - v)$ are independent and identically distributed, and the number of points in a cluster has a mixed Poisson distribution with the mixing distribution governed by the measure V.

Thus, the measure V determines the distribution of the number of points in a cluster. By choosing an appropriate measure V we can obtain much more variable number of points in individual clusters than for Poisson-Neyman-Scott processes in Example 1.15.

Now we give three examples of stationary shot-noise Cox processes. Figure 1.3 shows sample realizations from the given models.

Example 1.16. (Poisson-Neyman-Scott process) If $V(\mathrm{d}r) = \delta_1(\mathrm{d}r)$ is the Dirac measure concentrated in 1 the class of Poisson-Neyman-Scott processes is recovered, see Example 1.15. The cluster centers form a stationary Poisson process on $\mathbb{R}^d$ with intensity μ given by the decomposition (1.7). Furthermore, let $k(u) = c\,\tilde{k}(u)$ where $c > 0$ and $\tilde{k}$ is a probability density function. Then, the mean number of points in individual clusters is $c = \int k(u)\,\mathrm{d}u$ and the points within the cluster are distributed independently, according to the probability density function $\tilde{k}$, around the cluster center.

Figure 1.3 (left) shows a sample realization from the Thomas process, i.e. the Poisson-Neyman-Scott process with k being the probability density function of a bivariate zero-mean radially symmetric Gaussian distribution, see also Example 1.15.

Example 1.17. (Gamma shot-noise Cox process) Let V be defined by $V(\mathrm{d}r) = r^{-1}\exp\{-\theta r\}\,\mathrm{d}r$, where $\theta > 0$ is a parameter. Note that V is not integrable in the neighbourhood of 0. As a consequence, the corresponding shot-noise Cox process X is not a cluster process in the classical sense (Illian et al., 2008, Sec. 6.3) since the number of "clusters" in any compact set is infinite. However, because the weights of the majority of the clusters are very small, X is still a well-defined Cox process.

The name *gamma shot-noise Cox process* refers to the fact that V is the Lévy measure of a gamma distributed random variable (Hellmund et al., 2008, Sec. 4).

Figure 1.3 (middle) shows a sample realization from this model, where k is the probability density function of a bivariate zero-mean radially symmetric Gaussian distribution. Clearly, the variability in cluster weights is higher than for the Poisson-Neyman-Scott process in Example 1.16.

Since this model is frequently used in the simulation studies in the subsequent chapters we give a more detailed figure exemplifying the role of individual

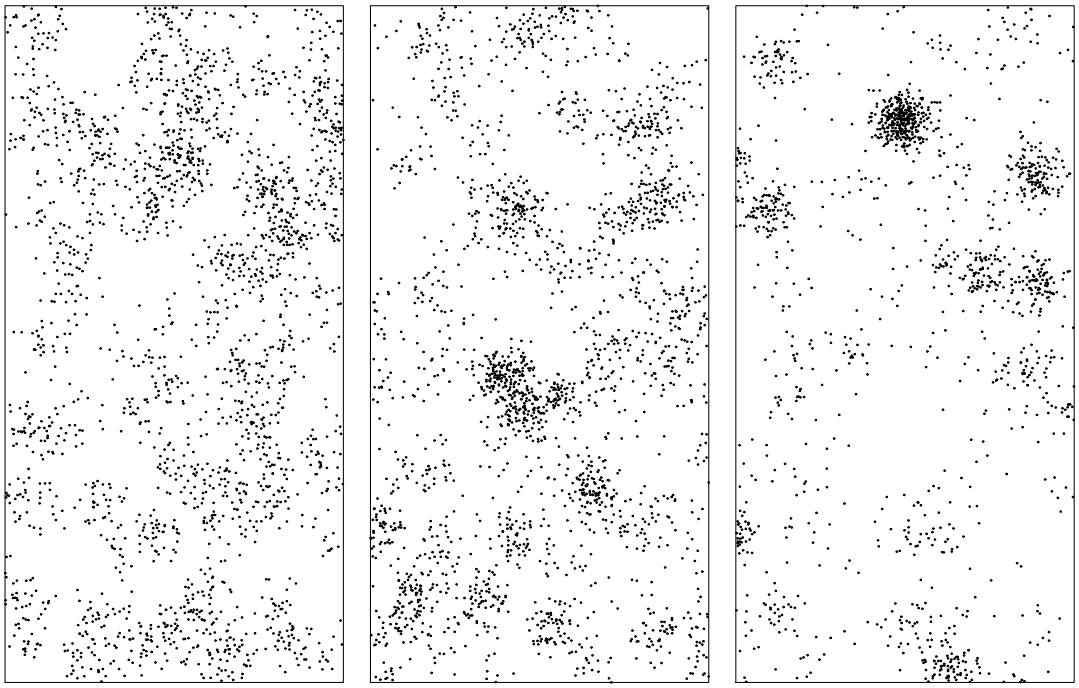


Figure 1.3: Sample realizations from different stationary planar shot-noise Cox processes with the same intensity λ and with the same probability density funcion governing the displacement of points around the cluster centers. Left: Poisson-Neyman-Scott process, middle: gamma shot-noise Cox process, right: inverse-Gaussian shot-noise Cox process. For details see Examples 1.16–1.18.

model parameters. We choose k to be the probability density function of a bivariate zero-mean radially symmetric Gaussian distribution with standard deviation $\sigma > 0$. For sample realizations of gamma shot-noise Cox process from models with different combinations of parameter values see Figure 1.4.

Example 1.18. (Inverse-Gaussian shot-noise Cox process) Let V be defined by the Lévy measure of the inverse-Gaussian distribution:

$$V(\mathrm{d}r) = \frac{1}{\sqrt{\pi}} r^{-3/2} \exp\{-\theta r\} \, \mathrm{d}r,$$

where $\theta > 0$ is a parameter. Then we obtain even more variable point patterns, as shown in Figure 1.3 (right).

The moment properties of shot-noise Cox processes are easily available, see Hellmund et al. (2008, Sec. 4). In particular, for processes satisfying condition (1.7) we have

$$\begin{aligned} \lambda(x) =& \mu \int_0^\infty r V(\mathrm{d}r) \int_{\mathbb{R}^d} k(x, v) \, \mathrm{d}v, \ x \in \mathbb{R}^d, \\ g(x, y) =& 1 + \frac{\mu}{\lambda(x)\lambda(y)} \int_0^\infty r^2 V(\mathrm{d}r) \int_{\mathbb{R}^d} k(x, w) k(y, w) \, \mathrm{d}w, \ x, y \in \mathbb{R}^d. \end{aligned}$$

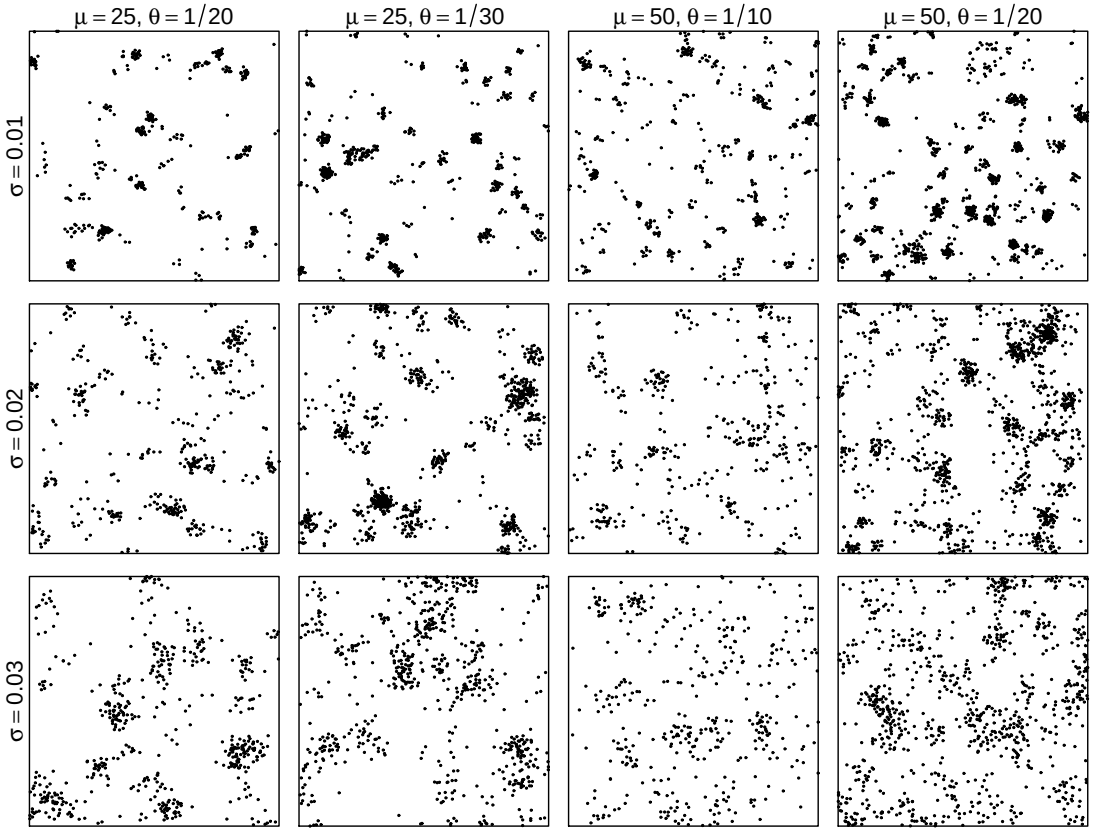


Figure 1.4: Sample realizations of stationary gamma shot-noise Cox processes with different combinations of parameter values. It is clearly seen that the smaller the value of σ the tighter are the clusters. Parameter μ controls the total intensity of points. Lower values of θ result in heavier clusters with more points on average. For details, see Example 1.17.

Note that in both equations we have a product of separate integrals for V and k. This will be important for estimation procedures presented in the subsequent chapters.

To simplify our notation we will set $V_1 = \int_0^\infty rV(\mathrm{d}r)$ and $V_2 = \int_0^\infty r^2V(\mathrm{d}r)$. For the parametric models presented in Examples 1.16–1.18 these integrals are simple functions of the model parameters, namely $V_1 = 1, 1/\theta, 1/\sqrt{\theta}$ and $V_2 = 1, 1/\theta^2, 1/(2\theta^{3/2})$, respectively.

When we apply location-dependent thinning with inhomogeneity function f to a stationary shot-noise Cox process specified by μ, V and k, we obtain a new SOIRS shot-noise Cox process with the same μ and V but with a new kernel $\tilde{k}(x,y) = f(x)k(y-x)$. In the subsequent chapters, however, we prefer the parametrization using the inhomogeneity function f and the homogeneous kernel $k(y-x)$ as opposed to the inhomogeneous kernel function $\tilde{k}$. The advantage of this parametrization is that the values of f cancel out in the formula for the pair-correlation function g and we obtain

$$g(x,y) = 1 + \frac{V_2}{\mu\,(V_1)^2} \frac{\int_{\mathbb{R}^d} k(x,w)k(y,w)\,\mathrm{d}w}{\int_{\mathbb{R}^d} k(x,w)\,\mathrm{d}w \int_{\mathbb{R}^d} k(y,w)\,\mathrm{d}w}, \; x,y \in \mathbb{R}^d.$$

The specific form of the pair-correlation function, and hence also the K-func-

tion, enables parameter estimation for shot-noise Cox processes as described in the following chapters.

1.4 Space-time point processes

As a space-time point process we consider a point process in $\mathbb{R}^d \times \mathbb{R}$, i.e. the temporal coordinate is a continuous variable. The Definition 1.1 can be easily generalized to this scenario. However, space-time models with discrete time may be considered as well, see e.g. Møller and Díaz-Avalos (2010).

In this section we introduce the properties and characteristics of space-time point processes which will be useful in the subsequent chapters. For a more detailed discussion on space-time point processes see Diggle (2007).

A space-time point process consists of "points" both with spatial and temporal coordinate. These are sometimes called *events* to emphasize the difference from a point in the spatial domain. We will denote a point of the space-time point process X as $(u, t) \in X$ meaning that an event of the process occured at the time t at the spatial location u.

We stress here that the temporal dynamics of a space-time process consists of new events occuring at a given location and at a given time. It does not describe temporal evolution of a system of moving points.

At this moment we emphasize the principal difference between a space-time point process, i.e. a point process on $\mathbb{R}^d \times \mathbb{R}$ where one coordinate plays a distinct role, and a spatial point process on $\mathbb{R}^{d+1}$.

- The temporal coordinate is not interchangeable with the spatial coordinates and, for example, it is not appropriate to talk about isotropy of a space-time point process w.r.t. rotations of $\mathbb{R}^{d+1}$. However, isotropy in the spatial part is a valid and useful property. For a possible definition of isotropy for space-time point processes see Definition 1.20 below.

- Using Euclidean distance to describe differences between space-time events is not appropriate or even natural.

- In applications, the temporal dynamics is often very different from the behaviour of the process in the spatial domain.

These reasons, among others, indicate that it is not suitable to simply adapt techniques from the spatial point process theory and that dedicated space-time approach to statistical inference and model-fitting is required.

When observing a space-time point pattern we consider the space-time observation window in the product form $W \times T$, where $W \in \mathcal{B}^d$ is a compact set with positive Lebesgue measure and $T \subset \mathbb{R}$ is a time interval or union of time intervals. It is possible to consider more general space-time observation windows but this is the most common scenario in applications – a fixed study region in space observed over some period of time.

Let us stress here that we assume that T has a positive length and the space-time events are observed with their precise time of occurence. This is the main difference from another type of data frequently analysed in space-time modelling – snapshots of the study region in discrete time instances.

Many of the definitions stated above extend easily to the space-time setting. For example the definition of a space-time point process, stationarity or factorial moment measures are obtained simply by considering the space $\mathbb{R}^{d+1}$ instead of $\mathbb{R}^d \times \mathbb{R}$. Unless specified otherwise, we do not introduce a new notation for space-time processes and their characteristics and we use the notation from the purely spatial case. It will be always clear from context whether a spatial or a space-time process is considered at the moment.

For clarity we give the following definition as an example of simple extension from the purely spatial case.

Definition 1.19. Let X be a point process on $\mathbb{R}^d \times \mathbb{R}$. Its k**th-order factorial moment measure** α_k is defined as

$$\alpha_k(A) = \mathbb{E}\left(\sum_{(u_1,t_1),\ldots,(u_k,t_k)\in X}^{\neq} I\left[((u_1,t_1),\ldots,(u_k,t_k)) \in A\right] \right),\ A \in \left(\mathcal{B}^{d+1}\right)^k,$$

where $\neq$ denotes that the summation is over k-tuples of distinct points of X and I is the indicator function.

Also, the definitions of the kth-order intensity functions and the pair-correlation function follow in the same way and a version of the Campbell theorem holds for space-time point processes. On the other hand, isotropy of a space-time point process is a tricky concept. Gabriel and Diggle (2009) propose the following definition of isotropy.

Definition 1.20. A space-time point process is second-order intensity reweighted stationary and isotropic if its intensity function is bounded away from zero and its pair-correlation function $g((u,t),(v,s))$ depends only on the differences $\|v-u\|$ and $|s-t|$.

The definition of a *space-time K-function* is not fully agreed upon in the literature. We focus now on the inhomogeneous version of the K-function under the SOIRS assumption as it is used in the subsequent chapters and the stationary version is covered as a special case.

Gabriel and Diggle (2009) suggest a definition based on a space-time version of the formula (1.3), namely

$$K_{\mathrm{ST}}(r,t) = \sigma_d \int_0^t \int_0^r s^{d-1} g(s,\tau)\,\mathrm{d}s\,\mathrm{d}\tau,\ r \geq 0, t \geq 0.$$

However, we prefer the definition proposed by Møller and Ghorbani (2012) which corresponds to the formula (1.2):

$$K(r,t) = \int\int I\left[\|s\| \leq r, |\tau| \leq t\right] g(s,\tau)\,\mathrm{d}s\,\mathrm{d}\tau \tag{1.8}$$

$$= \sigma_d \int_{-t}^t \int_0^r s^{d-1} g(s,\tau)\,\mathrm{d}s\,\mathrm{d}\tau, r \geq 0, t \geq 0. \tag{1.9}$$

The two definitions differ only by a multiplicative constant $1/2$ but the latter keeps the favourable interpretation in the stationary case: $\lambda\, K(r,t)$ is the

expected number of further points within distance r and time lag t from the origin given that the process X has a point at the origin (Møller and Ghorbani, 2012, Sec. 2.2). Therefore we choose to work with this definition of (inhomogeneous) space-time K-function when analyzing stationary or SOIRS space-time point processes.

Based on the ideas of Møller and Ghorbani (2012) we now define the spatial and temporal projection process. Assume that a space-time point process X is observed in an observation window $W \times T$ where $W \in \mathcal{B}^d$ is a compact set with area $|W| > 0$ and T is a bounded time interval with positive length or a union of such intervals.

We further assume that the second-order intensity function λ_2 of X exists. This implies that for any pair of distinct points $(u,t) \neq (v,s)$ from X (with both $u, v \in W$ or both $t, s \in T$) we have $u \neq v$ and $t \neq s$ almost surely. Thus we may disregard the multiple points in the observed spatial and temporal projection processes and define them as

$$X_s = \left\{u \in \mathbb{R}^d : \exists\, t \in T \text{ such that } (u,t) \in X\right\},$$
$$X_t = \left\{t \in \mathbb{R} : \exists\, u \in W \text{ such that } (u,t) \in X\right\}.$$

It follows from the boundedness of W and T that X_s and X_t are properly defined point processes on $\mathbb{R}^d$ and on $\mathbb{R}$, respectively. The crucial point is that we do not project from the whole $\mathbb{R}^d \times \mathbb{R}$ to $\mathbb{R}^d$, say, but merely from $\mathbb{R}^d \times T$ to $\mathbb{R}^d$ where T is a bounded set.

Heuristic interpretation of X_s might be as follows. We select those events from X which occur in the time interval T; X_s then consists of spatial locations of such points (we neglect the time coordinate). Similarly for X_t. Figure 1.5 shows an example of a space-time point process together with the corresponding spatial and temporal projection processes.

The moment characteristics of the projection processes X_s, X_t are fully determined by the characteristics of the space-time process X, as illustrated in the following proposition.

Proposition 1.21. *Let the kth-order intensity function λ_k of the space-time process X exist. Then the kth-order intensity functions $\lambda_{s,k}$ of X_s and $\lambda_{t,k}$ of X_t exist and are of the form*

$$\lambda_{s,k}(u_1,\dots,u_k) = \int_T \dots \int_T \lambda_k\left((u_1,t_1),\dots,(u_k,t_k)\right) \mathrm{d}t_1 \dots \mathrm{d}t_k,\ u_1,\dots,u_k \in \mathbb{R}^d,$$
$$\lambda_{t,k}(t_1,\dots,t_k) = \int_W \dots \int_W \lambda_k\left((u_1,t_1),\dots,(u_k,t_k)\right) \mathrm{d}u_1 \dots \mathrm{d}u_k,\ t_1,\dots,t_k \in \mathbb{R}.$$

Proof. The statement follows easily from the definition of the kth-order intensity function as the density of the kth-order factorial moment measure, the Campbell theorem and the Fubini theorem. □

As before, we will use the simple notation λ_s and λ_t for the first-order intensity function of X_s and X_t, respectively. This concludes the overview of the background material required for the subsequent chapters.

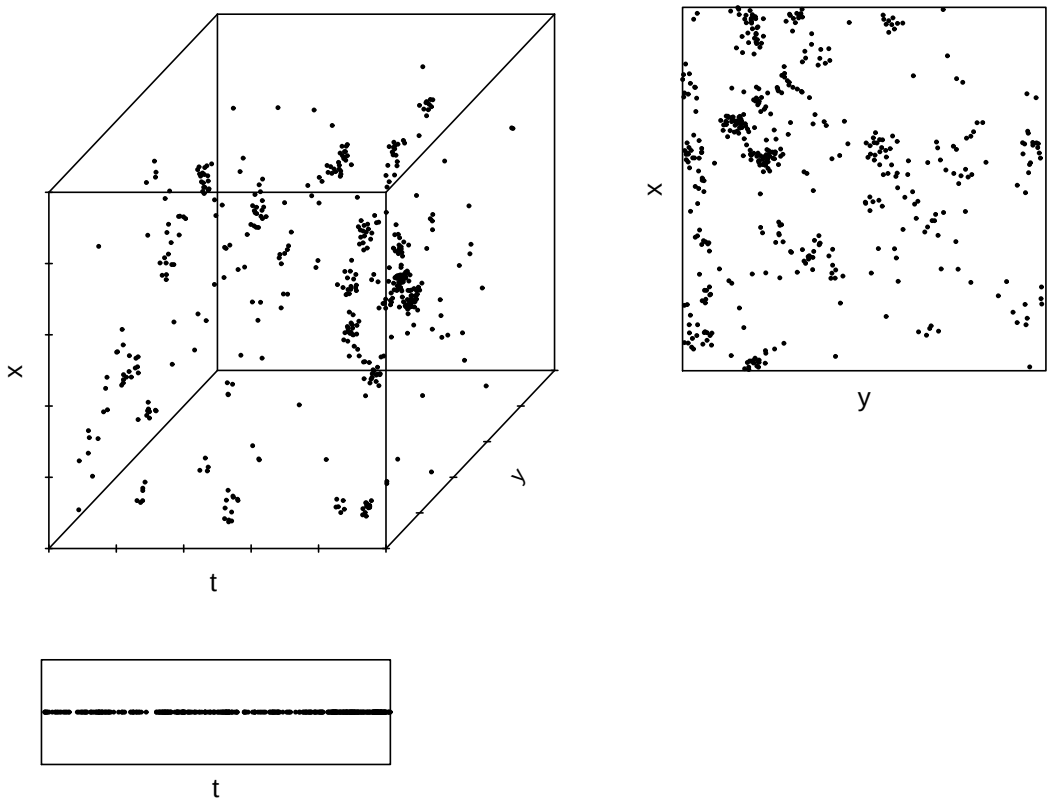


Figure 1.5: Example of a sample realization of a space-time point process X (upper left) together with the corresponding temporal projection process X_t (bottom) and spatial projection process X_s (right).

2. Parameter estimation for spatial Cox point processes

In this chapter we review several state-of-the-art estimation methods for parametric models of spatial Cox point processes. The methods are based on second-order properties of the process, i.e. the second-order intensity function λ_2 and the characteristics derived from it (pair-correlation function, K-function or the Palm intensity function). As explained in the examples in Sections 1.2 and 1.3 these characteristics are for many Cox process models available in reasonably tractable form as functions of an unknown (vector) model parameter θ and thus the optimization of the respective estimation criteria is numerically feasible.

The author's contribution to the problem of parameter estimation for spatial Cox point processes lies in the simulation studies comparing the performance of different estimation methods both in the stationary case, see Section 2.1.5 and the paper Dvořák and Prokešová (2012), and in the non-stationary setting, see Prokešová et al. (2014). Results of the latter study are not presented here in order to maintain reasonable extent of the thesis.

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote a point pattern observed in a compact observation window W, i.e. the realization of the process $X \cap W$, where $x_i, i = 1, \dots, n$, denote locations of the observed points. We assume a parametric model for X and the vector of unknown parameters will be denoted θ.

Let us first discuss the classical maximum likelihood estimation. To use this method we will consider the probability density function $f(\mathbf{x}; \theta)$ of X with respect to the distribution of stationary unit-rate Poisson process restricted to W, i.e. with $\lambda(u) = I[u \in W]$. The maximum likelihood estimate $\hat{\theta}$ is then obtained as the value of θ maximizing $f(\mathbf{x}; \theta)$.

For a Poisson process with intensity function $\lambda(u; \theta),\ u \in W$, the probability density function is

$$f(\mathbf{x}; \theta) = \exp\left\{|W| - \int_W \lambda(u; \theta)\, \mathrm{d}u\right\} \prod_{i=1}^{n} \lambda(x_i; \theta). \tag{2.1}$$

As long as a suitable parametrization of $\lambda(\cdot; \theta)$ is available the density $f(\cdot; \theta)$ has a tractable form and the estimate $\hat{\theta}$ can be easily obtained.

For a Cox process driven by random intensity function $\Lambda(u; \theta),\ u \in W$, the probability density function is

$$f(\mathbf{x}; \theta) = \mathbb{E}_\theta \left[\exp\left\{|W| - \int_W \Lambda(u; \theta)\, \mathrm{d}u\right\} \prod_{i=1}^{n} \Lambda(x_i; \theta)\right], \tag{2.2}$$

where $\mathbb{E}_\theta$ denotes the expectation w.r.t. the distribution of the process with parameter θ.

We remark here that asymptotic results for the maximum likelihood estimation were established in Jensen (2005) for certain Cox processes on a real line. However, we are not aware of any such results for spatial Cox processes.

In order to obtain the maximum likelihood estimate of θ it is needed to repetitively evaluate the expectation including a complicated integral term with respect to possible values of the random conditional intensity function $\Lambda(\cdot; \theta)$. One

can of course take advantage of MCMC or other techniques and use approximations of the likelihood function $f(\mathbf{x};\theta)$, see e.g. Møller and Waagepetersen (2004, Chap. 10). Nevertheless, this approach is usually computationally very demanding and thus faster, simulation-free alternatives were sought.

Three such methods will be described below, both in the context of stationary and non-stationary Cox point processes. The methods are in fact second-order moment estimation methods because all of them are based on the second-order moment characteristics of the process X such as λ_2, K, g or λ_o.

2.1 Stationary case

In the following we will assume that X is a stationary Cox point process characterized by its second-order intensity function $\lambda_2(\cdot;\theta)$ or by some other second-order characteristic. However, we stress here that in general a point process is not determined by its second-order properties. Baddeley and Silverman (1984) provide an example of a cell process with rather regular structure but containing sparse regions of clustered points – the K-function of their process coincides with the K-function of the Poisson process. Despite clear differences between the processes that can be easily revealed visually, higher-order functionals are needed to distinguish the processes from each other.

2.1.1 Minimum contrast estimation

This estimation method was in the context of spatial statistics first introduced in Diggle (1983, Chap. 5). It can be based either on the K-function (this version will be denoted MCEK in the fottowing) or the pair-correlation function g (denoted MCEg), see e.g. Diggle (2003, Chap. 6). In the version using the g-function this method requires isotropy of the process X in addition to its stationarity. We recall that in this case the g-function is a function of a scalar argument.

The vector of parameters θ is estimated by minimizing the discrepancy measure

$$\int_{r_0}^{r_1} [\widehat{K}(u)^c - K(u;\theta)^c]^2 \,\mathrm{d}u \qquad \text{or} \qquad \int_{r_0}^{r_1} [\widehat{g}(u)^c - g(u;\theta)^c]^2 \,\mathrm{d}u \tag{2.3}$$

between the non-parametric estimate $\widehat{K}$ or $\widehat{g}$ given by Equations (1.4) and (1.5) and its theoretical value $K(\cdot;\theta)$ or $g(\cdot;\theta)$, respectively.

The constants c, r_0 and r_1 are used to control the sampling fluctuations in the estimates of K and g. In Diggle (2003, Sec. 6.1.1) it is recommended that for fitting aggregated planar point patterns using the K-function the constant $c = 0.25$ is used and that for data on a unit square r_1 should not be larger than 0.25. The remaining constant r_0 can be set to 0 or a small positive value, e.g. the minimum observed interpoint distance. For the pair-correlation function g there is no standard recommendation available.

We remark that it is possible to use a weight function instead of the exponent c in order to stabilize the variance of the estimate of K or g. For example Guan (2009) considers the discrepancy measure

$$\int_{r_0}^{r_1} w(u)[\widehat{g}(u) - g(u;\theta)]^2 \,\mathrm{d}u$$

and suggests that a weight function $w(u)$ inversely proportional to the variance of $\widehat{g}(u)$ is used.

Minimum contrast estimators are popular partly due to the fact that they have been implemented for quite some time in the `spatstat` package for `R` (Baddeley and Turner, 2005) for three important cluster point process models: Thomas process, Matérn cluster point process and isotropic log-Gaussian Cox process.

The asymptotic properties of the minimum contrast estimators using the K-function are discussed in Heinrich (1992), Guan and Sherman (2007) and Waagepetersen and Guan (2009). In Heinrich (1992) strong consistency and asymptotic normality was proved for Poisson cluster processes. In Guan and Sherman (2007) asymptotic normality was shown under a strong mixing assumption which is fulfilled for any point process with finite dependence range, e.g. the Matérn cluster process, and also for a broad range of log-Gaussian Cox processes (Guan and Sherman, 2007, Sec. 2.2). Waagepetersen and Guan (2009) extend the asymptotic results to a two-step estimation procedure for inhomogeneous Cox processes which uses minimum contrast with the K-function to obtain estimates of the clustering parameters.

2.1.2 Composite likelihood method

Composite likelihood approach is a general statistical methodology, see Lindsay (1988). In the context of point processes it is based on adding together individual log-likelihoods for single points or pairs of points of the process X to form a composite log-likelihood.

Several versions of composite likelihood have been suggested for estimation of different types of spatial point processes (Baddeley and Turner, 2000, Guan, 2006, Møller and Waagepetersen, 2007). Composite likelihood suitable for estimation of Cox processes was introduced in Guan (2006). It uses the second-order intensity function $\lambda_2(\cdot, \theta)$ to obtain the probability density function for two points of X occurring at locations x and y:

$$f(x, y; \theta) = \frac{\lambda_2(x - y; \theta)}{\int_W \int_W \lambda_2(u - v; \theta)\, \mathrm{d}u\, \mathrm{d}v}, \; x, y \in W.$$

After adding the individual log-likelihoods the composite log-likelihood is obtained:

$$\log CL(\theta) = \sum_{x \neq y \in X \cap W, ||x-y|| < R} \log \frac{\lambda_2(x - y; \theta)}{\int_W \int_W I(||u - v|| < R) \lambda_2(u - v; \theta)\, \mathrm{d}u\, \mathrm{d}v}. \tag{2.4}$$

Here only the pairs of points within distance $R > 0$ are considered. This is the only parameter that needs to be chosen a priori by the user, as opposed to the three parameters that need to be specified for the minimum contrast method. Disregarding the pairs of points separated by distance larger than R is motivated by the fact that distant pairs of points are often nearly independent. They do not carry much information about the parameters but increase variability of the estimator.

Numerical maximization of $\log CL(\theta)$ in the form presented above can be computationally demanding. This issue can be solved by application of the inner-region correction for the edge effects. It leads to a simplified formula

$$\log CL(\theta) = \sum_{x \in X \cap (W \ominus R), y \in X \cap W, 0 < ||x-y|| < R} \log \frac{\lambda_2(x-y;\theta)}{\lambda^2 |W \ominus R| K(R)}, \tag{2.5}$$

where $W \ominus R = \{w \in W : B(w,R) \subset W\}$, i.e. $W \ominus R$ is the window W eroded by distance R. Vector $\hat{\theta}$ maximizing $\log CL(\theta)$ is then taken for the estimate of θ. Consistency and asymptotic normality of the composite likelihood estimator are proved in Guan (2006) under suitable mixing assumptions.

The composite likelihood method of Guan (2006) has been recently implemented in the `spatstat` package for `R` for several popular Poisson-Neyman-Scott process models and for the log-Gaussian Cox processes, making it an easy to use, viable alternative to the minimum contrast estimation.

2.1.3 Palm likelihood method

The Palm likelihood estimator was introduced in Tanaka et al. (2008) and uses a very "geometrical" approach. It is based on the process of differences among the points of the observed point process X. Let us start with considering the processes

$$Y_x = \{y - x, x \neq y \in X \cap W\},\ x \in X \cap W.$$

Each Y_x is again a point process (now inhomogeneous) with intensity function $\lambda_o(\cdot;\theta)$ – the Palm intensity of the original process X.

Ignoring for the moment the interactions in the process Y_x, i.e. approximating Y_x by a Poisson process, the log-likelihood of $Y_x \cap B(o,R)$ is the following (up to a constant – cf. the formula (2.1)):

$$\sum_{y \in X \cap W, 0 < ||x-y|| < R} \log \lambda_o(x-y;\theta)) - \int_{\mathbb{R}^d} I(||u|| < R) \lambda_o(u;\theta)\, \mathrm{d}u.$$

Here we consider only pairs of points within distance $R > 0$ following the same reasoning as in the case of the composite likelihood estimator.

Treating all the $Y_x, x \in X \cap W$, as independent, identically distributed replications, we sum the individual log-likelihoods over $x \in X \cap W$ and get the so-called Palm log-likelihood:

$$\log PL(\theta) = \sum_{x \neq y \in X \cap W, ||x-y|| < R} \log \lambda_o(x-y;\theta) - X(W) \int_{\mathbb{R}^d} I(||u|| < R) \lambda_o(u;\theta)\, \mathrm{d}u. \tag{2.6}$$

Applying the inner-region correction we get the following expression for the Palm log-likelihood which is more useful for the estimation in practice:

$$\log PL(\theta) = \sum_{x \in X \cap (W \ominus R), y \in X \cap W, 0 < ||x-y|| < R} \log \lambda_o(x-y;\theta) - X(W \ominus R) \lambda K(R). \tag{2.7}$$

Again, the maximizer of $\log PL(\theta)$ is taken for the estimate of θ.

An alternative way of arriving at the Palm likelihood goes as follows. Let

$$Y(R) = \{y - x : x \neq y \in X \cap W, \|y - x\| < R\}$$

be the point process of differences of points in X observed in W with the mutual distance smaller than R. Evidently, $Y(R)$ is a point process contained in $B(o, R)$. The intensity function of this point process can be determined as follows. Let A be a Borel subset of $B(o, R)$. Then,

$$\mathbb{E}\,|Y(R) \cap A| = \int_W \int_W I[y - x \in A]\lambda\lambda_o(y - x; \theta)\,\mathrm{d}x\,\mathrm{d}y = \int_A \gamma_W(u)\lambda\lambda_o(u; \theta)\,\mathrm{d}u,$$

where $\gamma_W(u) = |W \cap (W + u)|$ is the set covariance of the window W, see Stoyan et al. (1995, p. 126) for further details. The point process $Y(R)$ has thus the intensity function, concentrated on $B(o, R)$, of the form

$$\lambda_R(u) = \gamma_W(u)\lambda\lambda_o(u; \theta),\ u \in B(o, R).$$

The Palm log-likelihood

$$\begin{aligned}\log PL(\theta) = &\sum_{x \neq y \in X \cap W, (y-x) \in B(o,R)} \log\left(X(W)\lambda_o(y - x; \theta)\right) \\ &- X(W) \int_{B(o,R)} \lambda_o(u; \theta)\,\mathrm{d}u, \end{aligned} \tag{2.8}$$

is obtained by treating $Y(R)$ as an inhomogeneous Poisson process with intensity function $\lambda_R(u)$, replacing the intensity λ of the original point process X by the observed intensity $X(W)/|W|$ and approximating $\gamma_W(u)$, $u \in B(o, R)$, by $|W|$. This is a reasonable approximation for R substantially smaller than the size of the observation window W. Note that both versions of the Palm log-likelihood given in Equations (2.6) and (2.8) are equivalent. In fact, they are equal up to an additive constant not depending on θ and differ only in the underlying reasoning.

Note that even though the Palm likelihood estimation was derived by using the process of differences it is a second-order moment method as well because it is based on the second-order characteristic λ_o of the observed point process X. Let us also remark that its use is not restricted to simple models such as the Thomas process. If the Palm intensity function λ_o is in such form that direct maximization of $\log PL(\theta)$ is not possible (e.g. for general Neyman-Scott processes), numerical algorithms are available for finding $\operatorname{argmax} \log PL(\theta)$, see Tanaka et al. (2008, Section 3).

Strong consistency and asymptotic normality of the Palm likelihood method are proved in Prokešová and Jensen (2013) under suitable mixing assumptions. These are fulfilled for a broad range of Poisson-Neyman-Scott processes and log-Gaussian Cox processes.

Moreover we would like to stress here that since the Palm intensity is defined as a function of the argument $u \in \mathbb{R}^d$ and not just its norm $\|u\|$, the Palm likelihood estimation is not restricted to isotropic point processes only. It can be used for estimation of any suitably parametrized stationary Cox point process model in $\mathbb{R}^d$ (the same holds for the composite likelihood estimation). A worked out example of an anisotropic Thomas process in $\mathbb{R}^2$ can be found in Prokešová and Jensen (2013).

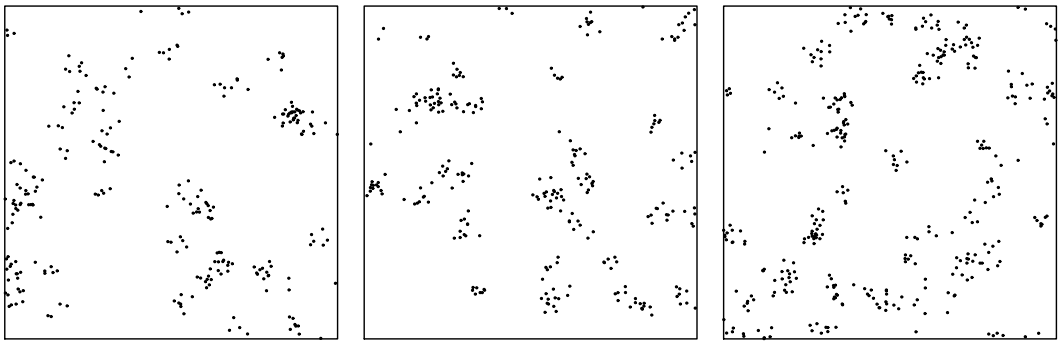


Figure 2.1: Sample realizations of stationary Thomas process described in Example 2.1. Parameter values $\nu = 50, \kappa = 6, \sigma = 0.02$. The observation window W is the unit square $[0,1]^2$.

2.1.4 Examples

In this section we give two detailed examples illustrating the use of the moment estimation methods described above. We focus on planar point process models used later in the simulation study in Section 2.1.5.

Example 2.1. (Thomas process) First we consider a stationary Poisson-Neyman-Scott process X in $\mathbb{R}^2$ as described in Example 1.15. Three sample realizations of the Thomas process are shown in Figure 2.1.

We recall the notation: $\kappa > 0$ denotes the intensity of the mother points and $\nu > 0$ is the mean number of points in a cluster. Furthermore, we choose k to be the probability density function of a bivariate zero-mean radially symmetric Gaussian distribution with standard deviation $\sigma > 0$. The vector of parameters is thus $\theta = (\kappa, \nu, \sigma)$. In this case the moment characteristics of X are as follows:

$$\begin{aligned}
\lambda &= \kappa\,\nu, \\
g(u) &= 1 + \frac{1}{4\pi\kappa\sigma^2} \exp\left\{-\frac{\|u\|^2}{4\sigma^2}\right\},\ u \in \mathbb{R}^2, \\
K(r) &= \pi r^2 + \frac{1}{\kappa}\left(1 - \exp\left\{-\frac{r^2}{4\sigma^2}\right\}\right),\ r \geq 0, \\
\lambda_o(u) &= \kappa\,\nu + \frac{\nu}{4\pi\sigma^2} \exp\left\{-\frac{\|u\|^2}{4\sigma^2}\right\},\ u \in \mathbb{R}^2.
\end{aligned}$$

Note that the pair-correlation function and the K-function do not depend on the parameter ν. This implies that ν is not identifiable by the minimum contrast method.

The score equation we get by differentiation of the log Palm likelihood (2.7) with respect to ν will yield

$$\widehat{\nu} = \frac{N}{\kappa K(R) X(W \ominus R)}, \tag{2.9}$$

where N denotes the number of terms in the sum in (2.7) and the K-function does not depend on ν.

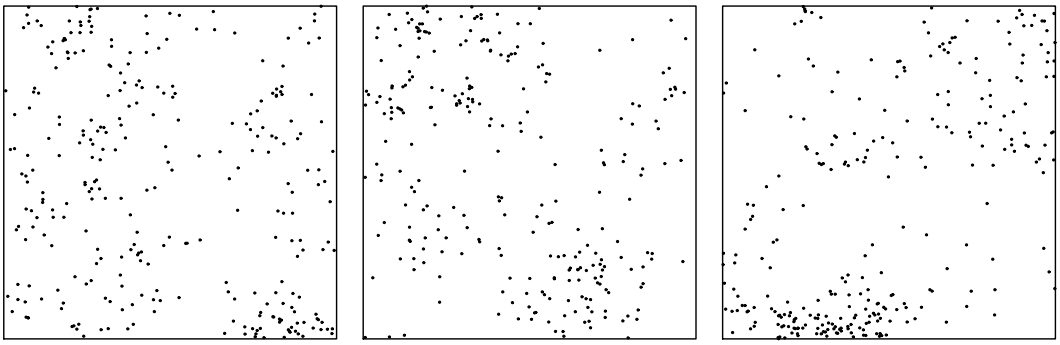


Figure 2.2: Sample realizations of stationary log-Gaussian Cox process described in Example 2.2. Parameter values $\beta = 10, \sigma^2 = 1$ and ν chosen so that the intensity of the point process is 300, the same as for the Thomas process in Figure 2.1. The observation window W is the unit square $[0,1]^2$.

For the composite likelihood the score equation we get by differentiation of (2.5) with respect to ν is identity $\frac{2N}{\nu} = \frac{2N}{\nu}$ and the parameter ν is not identifiable. The same applies to the minimum contrast estimation using both the K- and g-function because none of them depends on ν. Thus the parameter ν must be estimated using the formula for the intensity function, i.e. $\widehat{\lambda} = \widehat{\kappa}\,\widehat{\nu}$. Here we take $\widehat{\lambda} = X(W)/|W|$.

Example 2.2. (log-Gaussian Cox process with exponential covariance function) Now we consider the log-Gaussian Cox process X in $\mathbb{R}^2$ as described in Example 1.14. Three sample realizations of the log-Gaussian Cox process are shown in Figure 2.2.

Let $Z(x), x \in \mathbb{R}^2$ be a Gaussian random field with constant mean ν and exponential covariance function $c(x,y) = \sigma^2 \exp\{-\beta \|x-y\|\}$, where $\sigma^2 > 0$ and $\beta > 0$ are parameters. The vector of parameters is thus $\theta = (\nu, \beta, \sigma^2)$ in this case. The moment characteristics of the process X are of the form

$$\begin{aligned}
\lambda &= \exp\left\{\nu + \frac{\sigma^2}{2}\right\}, \\
g(u) &= \exp\left\{\sigma^2 \exp\{-\beta\|u\|\}\right\},\ u \in \mathbb{R}^2, \\
K(r) &= 2\pi \int_0^r r \exp\left\{\sigma^2 \exp\{-\beta s\}\right\} \mathrm{d}s,\ r \geq 0, \\
\lambda_o(u) &= \exp\left\{\nu + \frac{\sigma^2}{2}\right\} \exp\left\{\sigma^2 \exp\{-\beta\|u\|\}\right\},\ u \in \mathbb{R}^2.
\end{aligned}$$

Again, the pair-correlation function and the K-function do not depend on the parameter ν. This implies that ν is not identifiable by the minimum contrast method and the composite likelihood method and has to be calculated from the estimate of intensity by $\log\left(\frac{X(W)}{|W|}\right) - \frac{\sigma^2}{2}$.

Also, the K-function does not have a simple analytic form. Despite this it is possible to use it for minimum contrast estimation taking advantage of numerical integration in order to obtain values of the theoretical K-function at given parameter values.

For the Palm likelihood method we get the estimate

$$\widehat{\nu} = \log \frac{N}{K(R)X(W \ominus R)} - \frac{\sigma^2}{2}, \tag{2.10}$$

where N denotes the number of terms in the sum in (2.7) and the K-function does not depend on ν.

2.1.5 Simulation study I.

All of the moment estimation procedures reviewed in this chapter besides MCEg were proved to be consistent and asymptotically normal under suitable mixing conditions, see the references in the respective sections above. However, the theoretical expressions for the asymptotic variance of the estimators are too complicated to be used for direct comparison of the efficiency of the estimators. Moreover, the convergence may be too slow and the bias can be a more important issue than the variance for certain point processes observed on medium sized windows. Therefore we decided to assess the empirical performance of the discussed moment estimation procedures by a simulation study. The study was published in Dvořák and Prokešová (2012) and the results are presented below.

To compare the empirical performance of the estimators we chose two different types of cluster processes – the Thomas process and the log-Gaussian Cox process with exponential covariance function, see Examples 2.1 and 2.2.

To assess the performance of the estimators on middle-sized to large point patterns exhibiting different degree of clustering we chose eight combinations of parameter values for the Thomas process ($\kappa = 25$ or 50, $\nu = 4$ or 6 and $\sigma = 0.02$ or 0.04, representing relatively strong and weak clustering, respectively) and for the log-Gaussian Cox process ($\beta = 10$ or 20 representing relatively strong and weak dependence, respectively, $\sigma^2 = 1$ and ν calculated so that the intensity of the process is 100, 150, 200 and 300 as in the case of the Thomas process).

For each process and each combination of parameters we generated 500 independent realizations and re-estimated the parameters using the three moment estimation methods (i.e. using formulas (2.3), (2.5) and (2.7)).

Parameter estimation – computational details

Realizations of the Thomas process and the log-Gaussian Cox process, with appropriate parameters, in a unit square window W were simulated using the package `spatstat` for `R` (Baddeley and Turner, 2005).

For simulation of the Thomas process realizations we used the algorithm presented in Møller (2003, Sec. 4.1) which introduces *edge effects* and the distribution of the resulting process $\widetilde{X}$ is only an approximation of the (desired) distribution of X itself. Some point may be missing in $\widetilde{X}$ which would occur in X. It is possible to calculate the mean number of missing points in a single realization (Møller, 2003). For the combinations of parameters considered in this study it is always smaller than $4 \cdot 10^{-4}$, i.e. negligible compared to the mean number of points in a single realization.

When simulating realizations of the log-Gaussian Cox process we need to use a finite representation of the random field Z in the observation window W. We

chose a regular grid of 25×25 points and instead of the Gaussian random field Z we simulated a Gaussian random vector representing values of Z in the given lattice points. Then we approximated values of the random field Z in W by the value in the nearest lattice point. Hence, the distribution of the resulting process $\widetilde{X}$ is only an approximation of the (desired) distribution of X itself. However, the discretization of the random field values is fine enough for the approximation to be reasonably accurate and useful for the study.

We used the minimum contrast method (MCE) with both the K-function (MCEK) and the pair-correlation function g (MCEg). The tuning parameters in the case of the K-function were chosen according to the recommendation in Diggle (2003, Section 6.1.1). The same values of parameters were used for the g-function, i.e. $c = 0.25$, $R = 0.25$ and r equal to the minimum interpoint distance observed in the given realization. The estimation was conducted by the functions provided by `spatstat`.

The composite likelihood (CLE) and Palm likelihood (PLE) estimation methods have only one tuning parameter, R, determining the maximum distance of pairs of points that will be taken into account. The distance R has to be chosen carefully so that the variance of the estimators is reduced but not much information about the interactions in the point pattern is lost. Since the observation window W is a unit square we chose the values of R to be 0.1, 0.2 and 0.3, respectively.

To take into account the interpoint interactions when choosing the value of R we also consider $R = r_{data}$, where r_{data} corresponds to a range of correlation. It is determined as the smallest r for which $\widehat{g}(r) < 1$ holds, where $\widehat{g}$ is a non-parametric estimate of the pair-correlation function g.

For CLE and PLE we applied the inner-region correction for the edge effects, see (2.5) and (2.7). It is straightforward and can be used for data observed in irregular windows. Other edge corrections could be used as well, for example the torus correction could perform better in the case of a rectangular window but it is difficult to use in practice for irregularly shaped windows.

The estimation for CLE and PLE was conducted in software `Mathematica 7`. Maximization of the appropriate log-likelihood functions was performed using a combination of two methods.

First, derivatives of the log-likelihood function with respect to the unknown model parameters were calculated and a vector of parameter values was found for which all the derivatives were equal to zero using a default Newton's method.

Then, if the numerical method diverged (i.e. the estimate lied out of a generously long interval containing the true value of the parameter), the parameters are estimated again by direct maximization of the log-likelihood function using simulated annealing.

This procedure is motivated by the fact that the estimation using derivatives is fast but somewhat numerically unstable, while direct maximization is more computationally demanding but in general less likely to diverge. The combination of these two methods was a good compromise between the computational time and numerical stability.

For the minimum contrast methods and the composite likelihood method the value of the non-identifiable parameter ν was estimated from the observed intensity of the point process as described in the examples in Section 2.1.4.

Results

Main results are summarized in Tables 2.1 and 2.2 for the Thomas process and the log-Gaussian Cox process, respectively. They show results for two variants of MCE (based on K-function and pair-correlation function) and for CLE and PLE with different choices of the tuning parameter R. Complete results are available in Dvořák and Prokešová (2011).

The tables show relative mean biases of the estimators and relative mean squared errors (MSEs) (by relative we mean divided by the true value of the estimated parameter or by its square for the MSE). All the statistics were obtained from the middle 95% of the estimates from 500 replications.

Neglecting the 5% of the most extreme estimates was motivated by the fact that in certain situations the estimation methods can be numerically unstable. If such a situation was encountered in practice the estimates would easily be identified as unrealistic and one would alter the parameters of the underlying optimization methods or opt for an alternative estimation method. However, due to the extent of the computations involved, this was not possible in this simulation study. Note that if all the estimates were used to calculate the statistics these would be severly distorted by the numerical instability and hence completely uninformative for a prospective researchers willing to find a method of choice for their particular dataset.

Thomas process

Parameter σ. Among the three parameters of the Thomas process model σ is the one estimated with most accuracy by any of the compared methods. Particularly MCEg and CLE with $R = 0.1$ give very precise estimates. From the two MCE methods MCEK is always worse than MCEg. PLE with $R = 0.1$ (as well as CLE with $R = 0.2$) is comparable with MCEK for strong clustering. For weak clustering PLE is inferior to the other methods.

The reason for this is the strong positive bias (around 25% for $R = 0.1$) of PLE in case of weak clustering. PLE generally overestimates the parameter σ while MCE always underestimates it. CLE with $R = 0.1$ is virtually unbiased for strong clustering and positively biased for weak clustering. For strong clustering CLE with $R = 0.1$ generally showed the smallest variance followed by MCEg, PLE and MCEK. For weak clustering variance of CLE and PLE with $R = 0.1$ becomes larger and MCEg becomes the most stable estimator.

An important observation is the rapid deterioration of the quality of the CLE and PLE estimates with increasing value of the tuning parameter R. Increase in the variability of the estimators with growing R is to be expected, since with the employed inner-region edge correction and with growing R we lose growing percentage of the data. However, for CLE and $R = 0.3$ the estimator of σ is not just bad, it is even numerically unstable, and in case of weak clustering and higher intensity of the point process the same happens also for $R = 0.2$.

An explanation for this could be found in the exact form of the score equations for CLE – the estimating functions as described in Example 2.1 in Section 2.1.4 have a steep step around the correct value of the parameter σ and then become virtually constant (and nonzero). Thus for a small number of observed pairs of

points from X and a clustered point pattern (with few distant pairs observed) the estimating equation can lead to a severe overestimation of the parameter σ. Therefore it is extremely important when estimating the interaction parameters for processes with weaker clustering by CLE or PLE to choose the tuning constant R not too large. Definitely not larger than the range of correlation and reasonably small with respect to the size of the observation window so that not too much data is lost by the edge correction.

Parameter κ. For the other interaction parameter κ estimated directly by all three methods the best estimates are provided by PLE (with $R = 0.1$, for weak clustering also with $R = 0.2$) in the majority of cases. The exception are processes with low intensity and strong clustering, i.e. with a few number of tight well defined clusters in the observed point pattern – here MCEg gives better results. Like with the other parameters the MCEK method is consistently worse than MCEg and the worst results are generally provided by CLE (with $R = 0.1$).

Note that the good results of PLE ($R = 0.1$) are mainly caused by the comparatively small variability of the estimates because all the PLE estimates with $R = 0.1$ have considerable negative bias. The bias becomes lower for $R = 0.2$ and thus the best estimates for processes with low intensity and weak interaction (i.e. a few loose clusters in the observation window) are obtained by PLE with $R = 0.2$. This behaviour is again implied by the exact form of the estimating functions used. Note that for CLE the situation is somewhat similar as for the parameter σ – CLE with $R = 0.1$ provides nearly unbiased estimates, whereas with $R = 0.2$ we have considerable positive bias and with $R = 0.3$ the estimates become useless. Nevertheless the variability of CLE is generally higher than variability of the other estimators.

When comparing the MCEg and MCEK methods the main factor is also variability of the estimators - the MCEg method is consistently less variable than MCEK (the same situation as with the parameter σ).

Parameter ν. The last parameter ν is by all the methods determined by means of the observed number of points of X and the values of the already estimated parameters. It is true not just for the MCE and CLE where $\widehat{\nu} = X(W)/(|W|\widehat{\kappa})$ but also for the PLE method. Namely (2.9) is derived just from the comparison of the expected and observed number N of pairs of points from X.

Thus it is natural that the quality of the estimates of ν for MCE and CLE follows the pattern from estimation of κ with MCEg being better than MCEK and CLE providing even worse estimates. The PLE estimator uses mean number of pairs of points from X which could provide more exact estimates than just $X(W)$. On the other hand, the formula (2.9) includes both of the parameters κ and σ and thus can be influenced more by the bad quality of those estimates. The simulation study shows that in reality for weak clustering PLE (with $R = 0.1$) provides the best estimates of ν – again mainly due to low variability of the estimates since they are negatively biased. For strong clustering the estimates provided by MCE are better.

Procedure using r_{data}. The estimated values of r_{data} are for strong clustering concentrated around 0.1 as expected, with a fair proportion being smaller than 0.1 (see Dvořák and Prokešová (2011) for the histograms of r_{data}). For CLE the estimates of σ using $R = r_{data}$ are comparable with those obtained with $R = 0.1$; they have only slightly larger bias and variance. The bias is nevertheless still very small (smaller than for MCE) and the variance is only a bit larger than for MCEg. Overall, the procedure with r_{data} for CLE estimation of σ provides very good estimates – better than for MCEK. Estimates of the other interaction parameter κ by CLE with r_{data} are worse but comparable with estimates produced by CLE with $R = 0.1$.

For PLE the estimates of the interaction parameters with $R = r_{data}$ are also worse than for fixed value $R = 0.1$. By closer inspection of the simulated data we see that particularly for the parameter κ the cases with $r_{data} < 0.1$ produce significantly biased estimates.

For weak clustering the estimated values of r_{data} are scattered between 0.1 and 0.2 with a few cases taking the value of 0.25 which was the upper bound for the estimated range of interaction (since the estimate of g function is not very stable for values larger than 0.25). For estimation of σ CLE with $R = r_{data}$ has significantly larger bias and variance than with $R = 0.1$ and PLE produces impractical estimates. For estimation of κ CLE with $R = r_{data}$ produces sometimes better estimates than with $R = 0.1$ but they are still worse that those produced by any other method. PLE with $R = r_{data}$ produces generally worse estimates of κ than with $R = 0.1$ but these are typically still better than those produced by the other methods.

In conclusion we can say that the method using $R = r_{data}$ is not superior to the fixed value of the tuning parameter $R = 0.1$ for CLE and PLE. Partially it can be explained by the larger variability caused by the changing value of R but very often also the bias of the estimators is larger than with the fixed value of $R = 0.1$. The question of the ideal choice of R for a given point process model and a given observation window is not a simple one. Even the amount of information expressed by the mean number of observed pairs of points with distance $\leq R$ for a clustered point process X need not be a monotone function of R (see Prokešová and Jensen (2013, Section 3)). More sophisticated methods of the adaptive choice of R must be used to produce a considerable improvement in the quality of the estimates than just a simple choice $R =$ range of interaction.

Overall performance. In conclusion we can say that quality of all the compared estimators improves with higher intensity of the process X and stronger interactions in the point process (i.e. tighter clusters). Concerning minimum contrast estimation the version using pair-correlation function always yields better estimates than the version using the K-function. When using the CLE and PLE methods for estimation of the interaction parameters it is important to use reasonably small values of the tuning parameter R which provide reasonably good estimates. High values of R lead to numerical instability of the methods and impractical estimates – this is especially important for CLE and, to a lesser extent, also for PLE.

To address the overall performance, the minimum contrast estimation using the pair-correlation function provides the best estimates of the interaction para-

					MCE		CLE				PLE			
		κ	ν	σ	K	g	.1	.2	.3	r_{data}	.1	.2	.3	r_{data}
Bias	$\widehat{\sigma}$	25	4	.02	-.044	-.029	.009	-.034	-.019	.028	.068	.023	.086	.118
				.04	-.064	-.067	.077	.208	34.7	.188	.287	.494	1.27	1.22
		50	6	.02	-.028	-.027	.001	-.045	-.086	.021	.096	.014	.052	.106
				.04	-.044	-.047	.093	25.2	160	.238	.233	.849	3.55	2.41
	$\widehat{\kappa}$	25	4	.02	.103	.070	-.013	.271	.917	-.097	-.143	.083	.161	-.260
				.04	.148	.159	.098	.196	.987	-.087	-.331	-.178	.042	-.107
		50	6	.02	.060	.022	.071	.318	.818	.004	-.122	.115	.139	-.157
				.04	.109	.102	.064	.405	1.02	.075	-.293	-.133	.024	-.087
	$\widehat{\nu}$	25	4	.02	-.074	-.067	.101	-.145	-.303	.386	-.416	-.431	-.375	-.365
				.04	-.094	-.117	.098	.022	-.187	.896	-.252	-.272	-.289	-.272
		50	6	.02	-.054	-.038	-.057	-.199	-.338	-.002	-.405	-.434	-.344	-.378
				.04	-.068	-.076	-.035	-.230	-.260	-.011	-.257	-.274	-.254	-.266
MSE	$\widehat{\sigma}$	25	4	.02	.035	.014	.016	.032	.094	.023	.031	.045	.135	.051
				.04	.050	.034	.061	.317	$1{\cdot}10^4$	.168	.211	.764	6.85	5.47
		50	6	.02	.022	.009	.006	.018	.036	.011	.026	.026	.055	.027
				.04	.035	.023	.036	$2{\cdot}10^4$	$1{\cdot}10^5$	.332	.077	2.45	32.6	16.8
	$\widehat{\kappa}$	25	4	.02	.125	.073	.207	.364	3.08	.244	.108	.186	.225	.233
				.04	.270	.217	.429	.576	5.67	.321	.170	.085	.192	.186
		50	6	.02	.091	.052	.097	.304	1.65	.111	.039	.100	.099	.109
				.04	.189	.148	.184	.562	4.48	.268	.093	.039	.145	.150
	$\widehat{\nu}$	25	4	.02	.075	.044	.266	.142	.250	1.38	.192	.225	.185	.166
				.04	.144	.116	.582	.435	.410	42.0	.076	.112	.159	.133
		50	6	.02	.055	.031	.062	.101	.200	.108	.173	.211	.140	.160
				.04	.130	.103	.133	.151	.219	.179	.071	.088	.133	.125

Table 2.1: Summary of simulation results – Thomas process.

meters (in the sense of the relative MSE) although for point patterns with strong clustering when estimating σ CLE (with $R = 0.1$) yields fully comparable and sometimes better quality estimates.

For point patterns with weak clustering PLE (with $R = 0.2$ and $R = 0.1$) yields the best estimates of κ. However, this is due to the minimal variability of the PLE estimator which has a serious negative bias in this case. The second best (according to MSE) estimator is MCEg which has half-size bias (i.e. acceptable 10%).

Parameter ν is calculated from the intensity (of points or pairs of points) of X and as such it depends on the quality of the other parameter estimates. The best values of MSE were achieved for strong clustering by MCEg and for weak clustering by (again significantly biased) PLE.

Log-Gaussian Cox process

Parameter β. For the log-Gaussian Cox process considered in this study β is the scale parameter of the covariance function of the Gaussian driving field and the hardest one to estimate. Unlike in the case of Thomas process for minimum contrast estimation MCEK showed always better performance than MCEg. Nevertheless for stronger interaction ($\beta = 10$) the best results are obtained by

CLE with $R = 0.1$ which has both minimum bias and minimum variability. As in the case of Thomas process it is important also here to have reasonably small value of the tuning parameter R, i.e. $R = 0.1$. For larger values of R the CLE (and also PLE) deteriorates fast and the estimates become useless.

Quality of the estimates of course improves with growing intensity λ; the growing amount of information seems to be best used by CLE since for high intensity ($\lambda = 300$) CLE (with $R = 0.1$) outperforms MCEK even in the case of weak dependence. CLE (with $R = 0.1$) generally slightly underestimates β but the absolute value of bias is much smaller than for any other compared estimator and this holds uniformly for any log-Gaussian Cox process considered in the study. When the variability of this estimator becomes small enough (thanks to the sufficient amount of data in the large λ case) it becomes superior to MCE estimation even in the case of weak dependence.

To understand this results better it is good to note that the case $\beta = 10$ which we call stronger interaction/dependence means that values of $g(u) - 1$ (where g is the pair-correlation function) are significantly positive for a larger range of values u than for the case of $\beta = 20$. Thus the observed point pattern of X for $\beta = 10$ is much more variable and may contain a few clusters (or sometimes none if λ is low) with large scale and large number of points (such as the one in the lower part of the right panel in Figure 2.2) whereas for $\beta = 20$ the observed point pattern is more homogeneous with large number of smaller clusters (both in terms of scale and number of points). The large clusters from X with $\beta = 10$ make the estimation of the functional characteristics K and g less stable than in the case of $\beta = 20$ or the Thomas process, which influences negatively the quality of the MCE estimation. Obviously the CLE method is less influenced by the occurrence of these large clusters.

The same fact may explain higher efficiency of MCEK when compared with MCEg. Even in the case of weaker dependence ($\beta = 20$) the clusters observed in X are highly variable in terms of the number of points and there are always some with a fairly high number of points. In such a situation the estimate of the K-function as a cumulative function is more stable than the estimate of the g-function (which corresponds to the density of K-function).

The performance of PLE is inferior to the other methods and for weak dependence ($\beta = 20$) the estimator is unusable. For stronger dependence in the point pattern the performance for PLE with $R = 0.1$ is comparable with the MCE methods, for larger values of R the estimator becomes unusable again.

Parameter σ^2. The other interaction parameter σ^2 is best estimated by the MCE methods. The estimates are only slightly biased and the variance is lower than for the other estimators. Note that the sign of the bias depends on the value of β – for $\beta = 10$ (i.e. point patterns with a few large and heavy clusters) σ^2 is underestimated while for $\beta = 20$ it is overestimated by both MCEK and MCEg. The performance of the two MCE methods is very similar.

Worse but still reasonable results are obtained for PLE with $R = 0.1$ and strong dependence in the point pattern. In the case of weak dependence the PLE method needs higher intensity λ of the point process to provide usable estimates. The worst results are provided by CLE. Both CLE and PLE have considerable negative bias (larger for the case of strong dependence) but PLE

shows consistently lower variability than CLE which makes it superior.

As in the case of estimation of β both PLE and CLE perform reasonably well only with the tuning parameter $R = 0.1$, for larger values both the bias and variability increase to impractical values.

Parameter ν. The intensity parameter ν is the easiest one to estimate and it is estimated very well by any of the compared methods. Remember that MCE and CLE estimate ν by the same formula (see Example 2.2 in Section 2.1.4) which is influenced by the value of the estimated parameter σ^2. Thus it follows from the properties of the estimates of σ^2 that the MCE estimates of ν has to be superior to the CLE estimates and that MCEK and MCEg perform equally well. Note moreover that the CLE estimates with $R = 0.2, 0.3$ are still very good even though the quality of the estimates of σ^2 was not good at all. Since $\widehat{\nu}$ is a linear function of $\widehat{\sigma}^2$ the influence of the quality of the estimates of interaction parameters on the estimate of the intensity parameter ν is much smaller than for the Thomas process case.

Another interesting observation is that the efficiency of the PLE estimate which is determined by formula (2.10) is worse than for the other methods and improves with larger value of the tuning parameter R. The formula (2.10) uses the number of observed pairs of points with distance smaller than R and provides negatively biased estimates – the smaller the value of R the larger the bias of the estimates. Although the PLE estimates of ν are still fairly exact the conclusion is that the intensity parameter ν is better estimated by the simpler formula using just the observed number of points of the point process X.

Procedure using r_{data}. The procedure using $R = r_{data}$ was even less successfull for the log-Gaussian Cox process than for the Thomas process. This could be explained partially by a more uniform distribution of the estimated values of r_{data} over the whole range between 0 and 0.25 (see Dvořák and Prokešová (2011) for the histograms of r_{data}). Moreover, when $\beta = 10$ we get quite often the estimate $r_{data} = 0.25$ and for such R both CLE and PLE become impractical.

For PLE the use of data dependent R produces impractical estimates of β and estimates of σ^2 with larger variance and for strong interaction also with larger bias than the procedure with fixed $R = 0.1$.

For CLE the estimate of the interaction parameter β using $R = r_{data}$ has in case of strong dependence generally smaller bias than with fixed $R = 0.1$ but the variance is larger and in total the estimate is worse than with fixed $R = 0.1$. Still it is superior to MCE. In case of weak dependence the small bias is the same as with fixed $R = 0.1$ and in total larger variance makes the estimate inferior to MCE. The estimate of σ^2 by CLE with $R = r_{data}$ is worse than the other estimates and for small values of the intensity it is even impractical.

Overall performance. The quality of all the compared estimators improves with higher intensity of the point process X and in the majority of cases also with weaker dependence (i.e. larger value of the parameter β). However, there is one important exception – the CLE with $R = 0.1$ provides more precise estimates of β for processes with stronger dependence. In this case CLE is much better than the other estimation methods. A plausible explanation is that the score

				MCE		CLE				PLE			
		λ	β	K	g	.1	.2	.3	r_{data}	.1	.2	.3	r_{data}
Bias	$\widehat{\beta}$	100	10	.281	.498	-.070	.190	.645	.087	-.221	1.76	2.44	.747
			20	.146	.216	.090	.380	.401	.095	.527	1.80	4.41	.359
		300	10	.259	.360	-.143	-.036	.258	-.062	-.224	1.51	2.25	.847
			20	.140	.180	-.051	.437	.272	.092	.417	.798	.629	.484
	$\widehat{\sigma}^2$	100	10	-.065	-.016	-.311	-.559	-.688	.032	-.218	-.467	-.570	-.357
			20	.050	.059	-.140	-.357	-.603	.536	-.295	-.269	-.525	-.122
		300	10	-.046	-.023	-.241	-.535	-.722	-.509	-.160	-.384	-.484	-.375
			20	.043	.072	-.131	-.336	-.668	-.196	-.131	-.212	-.422	-.129
	$\widehat{\nu}$	100	10	-.001	-.006	.026	.047	.064	-.018	-.146	-.052	-.016	-.111
			20	-.005	-.006	.017	.042	.068	-.067	-.094	-.079	-.005	-.147
		300	10	.000	-.003	.019	.041	.061	.037	-.115	-.043	-.013	-.050
			20	-.006	-.008	.012	.027	.059	.014	-.086	-.062	-.011	-.082
MSE	$\widehat{\beta}$	100	10	.538	.825	.112	.756	2.16	.649	.513	7.59	11.0	3.35
			20	.309	.357	.428	.983	.798	.576	2.14	8.67	51.5	2.21
		300	10	.235	.299	.085	.206	.566	.089	.268	6.12	14.0	3.57
			20	.124	.132	.091	.631	.394	.220	1.13	2.10	2.00	1.24
	$\widehat{\sigma}^2$	100	10	.146	.151	.380	.548	.774	2.90	.184	.436	.623	.430
			20	.196	.198	.476	.561	.756	4.16	.294	.519	.714	.529
		300	10	.060	.058	.184	.485	.736	.480	.132	.267	.450	.285
			20	.064	.063	.136	.354	.688	.233	.121	.226	.402	.138
	$\widehat{\nu}$	100	10	.005	.005	.007	.008	.012	.054	.038	.012	.011	.041
			20	.004	.004	.008	.010	.012	.066	.016	.016	.009	.055
		300	10	.002	.002	.003	.006	.007	.006	.021	.006	.005	.009
			20	.001	.001	.002	.003	.006	.003	.011	.006	.003	.011

Table 2.2: Summary of simulation results – log-Gaussian Cox process, $\sigma^2 = 1$ in all cases.

equations of CLE are less influenced by the high variability of the observed point pattern (caused by the variability of a few large and heavy clusters) than are the estimates of the functions K and g used for MCE. For point processes with $\beta = 20$ and small enough intensity ($\lambda < 300$) the observed clusters in the point pattern are more homogeneous and the MCEK becomes the best (in the sense of the relative MSE) estimator of β followed by MCEg.

The parameter σ^2 is best estimated by the minimum contrast methods, which show very similar performance, followed by PLE with $R = 0.1$. As in the other cases small value of the tuning parameter R is crucial for estimation of the interaction parameters by both PLE and CLE; for larger $R = 0.2, 0.3$ both the bias and variability of the PLE and CLE become too large.

Estimators of the intensity parameter ν are very precise for any of the compared methods. The important conclusion from the simulation study is that for log-Gaussian Cox processes a simple estimate by means of the observed (first order) intensity of the point process is superior to the more complicated estimate provided by PLE which uses the observed intensity of pairs of points from X.

Conclusions and further remarks

A short conclusion of the simulation study is that the minimum contrast method using the pair-correlation function g provides the best estimates in the majority of cases (see the next paragraph for an exceptional situation). It seems that the functional criterion (2.3) is able to get more information from the point pattern than the composite and Palm likelihood. On the other hand, when it comes to more complicated models like those considered in Tanaka et al. (2008) where the pair-correlation function does not have a closed form anymore but is expressed by means of an integral the maximization of (2.3) may be numerically more demanding. Then the more feasible Palm or composite likelihood estimation can be a good alternative.

There is one important situation where the estimators' performance was different. For the estimation of the interaction range β for log-Gaussian Cox model the best results were provided by composite likelihood and MCEK was superior to MCEg. A plausible explanation is that the population of clusters observed in the log-Gaussian Cox process is much more variable (both in terms of scale and number of points in the clusters) than the one observed in the Poisson-Neyman-Scott processes. A small (and highly variable) number of large and heavy clusters observed in the point pattern makes the estimates of the K-function unstable and even more so for g. In such a case the simpler estimating equations implied by CLE are probably more stable and provide better estimates of the interaction range parameter β.

In biological applications which study e.g. the interactions among members of a plant community or seed dispersal curves the interaction range can be a parameter of primary interest. Our results suggest that especially in the case of a nonhomegenous population of observed clusters composite likelihood should be the preferred estimation method.

Another interesting issue we observed in the simulation study is a rather extreme version of the bias-variance trade-off when estimating the number of clusters (i.e. parameter κ) for Thomas process with weak clustering (i.e. loose clusters). Observed point patterns are very homogeneous in this case and the individual clusters are completely unrecognisable. Thus it is understandable that for κ we got estimates with the worst performance (compared to other parameters) and as the results show the amount of data we used is still not enough to estimate this parameter safely. The smallest MSE was exhibited by PLE in spite of the heavy bias it had, followed by only slightly biased MCEg. Thus if one is not prepared to accept the granted serious bias provided by PLE it is better to prefer MCEg also for the estimation of κ.

The results we obtained about the better efficiency of the MCEg method in comparison with the MCEK method for Poisson-Neyman-Scott processes are in agreement with the results published in Brix (1999) for the so called G-shot-noise Cox processes. Even though these processes are not distinguishable from the Poisson-Neyman-Scott processes by just the g- or K-function the MCE method can be used for the parameter estimation of these models. Brix (1999, Sec. 4.2.1) states that MCEK can be unstable in some cases and that MCEg should be the preferred method.

Further we experimented with a simple data-based procedure for the choice of the tuning parameter R for CLE and PLE methods. The conclusion is that such

a choice does not improve the quality of the estimates in comparison with the a priori chosen fixed value $R = 0.1$. More sophisticated methods of the adaptive choice of R must be used to produce a considerable improvement in the quality of the estimates, such as e.g. the cross-validation procedure suggested in Guan (2006, Sec. 2.4).

2.2 Non-stationary case

For non-stationary spatial point process models the methods of parameter estimation described in the previous section cannot be employed directly. The main obstacle is the non-constant (first-order) intensity function $\lambda(\cdot)$ which appears e.g. in non-parametric estimators of K- and g-function. If $\lambda(\cdot)$ was known *a priori* one could easily adapt the three moment estimation methods also to the non-stationary case, at least for certain class of models with tractable second-order moment characteristics.

A solution to this estimation problem was proposed in Waagepetersen (2007) in the case of inhomogeneous Poisson-Neyman-Scott processes. The suggested two-step procedure consists of estimation of the inhomogeneity parameters using Poisson log-likelihood estimating function in the first step and then, conditionally on the estimated parameter values from the first step, estimation of the clustering parameters using minimum contrast method on the inhomogeneous K-function, cf. Section 2.1.1.

We stress here that the described method directly corresponds to minimum contrast estimation in the stationary case (which is, however, usually not considered to be a two-step method). In both cases one needs to first estimate the intensity λ or the intensity function $\lambda(\cdot)$ and plug-in the estimate into the non-parametric estimator of the K-function given in Equation (1.4). Finally, the clustering parameters are estimated using minimum contrast method. The main difference is that in the stationary case a natural estimator of λ is available, see Equation (1.1). In the non-stationary case the estimation of $\lambda(\cdot)$ is more involved.

The two-step estimation method proposed in Waagepetersen (2007) can be used for inhomogeneous Cox point processes with tractable second-order moment characteristics. It can be also generalized so that it uses other moment estimation methods in the second step, specifically one of the three methods described in the previous section.

In the following we consider a non-stationary spatial Cox point process X with $\lambda(\cdot;\beta)$ and $g(\cdot;\theta)$ fulfilling the SOIRS property. Here β and θ are (vectors of) unknown parameters. We assume that it is possible to separate the inhomogeneity and interaction parameters so that we avoid overspecification of the model.

Note that such a parametrization may not be the most natural one but it is often available and simplifies the description of the estimation procedures. As an example, consider a non-stationary process X obtained by location-dependent thinning of a stationary Thomas process in $\mathbb{R}^2$ with parameters κ, ν and σ, see Example 1.15. Let $f(x;\beta)$ be a non-constant function used for the thinning, properly scaled so that $\max_{x\in W} f(x;\beta) = 1$, where W is the observation window. Suppose that $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ is a p-dimensional vector. For this model λ

and g take on the form

$$\lambda(x) = \kappa\,\nu\,f(x,\beta),\ u \in W,$$
$$g(x,y) = 1 + \frac{1}{4\pi\,\kappa\,\sigma^2}\exp\left\{-\frac{\|x-y\|^2}{4\sigma^2}\right\},\ x,y \in W.$$

Now we can set $\beta_0 = \kappa\,\nu$. Then, the intensity function λ is parametrized by the inhomogeneity parameter $(\beta_0, \beta_1, \ldots, \beta_p)$ and the pair-correlation function is parametrized by the interaction parameter (κ, σ).

We describe now the first estimation step – estimation of the inhomogeneity parameters β. Different versions of the second step, i.e. estimation of the interaction parameters θ, are described in the corresponding subsections below.

For estimating β, Waagepetersen (2007) suggests to follow the approach of Schoenberg (2005), i.e. treat the point process X as a Poisson process with the same intensity function $\lambda(\cdot;\beta)$ and maximize the corresponding log-likelihood function

$$L_1(\beta) = \sum_{x \in X \cap W} \log \lambda(x;\beta) - \int_W \lambda(u;\beta)\,\mathrm{d}u \tag{2.11}$$

with respect to β. By doing this we disregard any interaction between points and hence we loose efficiency of the estimator. However, under appropriate assumptions it is possible to establish consistency (Schoenberg, 2005) and asymptotic normality of the estimator (Waagepetersen and Guan, 2009). This estimation procedure can be also regarded as the first-order composite likelihood method.

It is possible to formulate this estimation procedure in the framework of estimating equations (Mukhopadhyay, 2004). The estimate $\widehat{\beta}$ of β is obtained by solving the equation

$$U_1(\beta) = \sum_{x \in X \cap W} \frac{\lambda^{(1)}(x;\beta)}{\lambda(x;\beta)} - \int_W \lambda^{(1)}(u;\beta)\,\mathrm{d}u = 0, \tag{2.12}$$

where $\lambda^{(1)}$ denotes the derivative of λ w.r.t. the vector β. Note that $U_1(\beta)$ is the score function of the Poisson log-likelihood (2.11). By Campbell theorem, $U_1(\beta)$ is an unbiased estimating equation in the sense that $\mathbb{E}\,U_1(\beta^*) = 0$ where β^* is the actual value of β governing the distribution of X.

We remark here that Waagepetersen (2007) showed that the estimate of β obtained by solving the Equation (2.12) differs negligibly from the estimate obtained by a more complicated and computationally much more demanding second-order estimating equation, corresponding to the score function of the full composite likelihood (2.4) in the inhomogeneous case. This justifies the use of the first-order intensity function for the estimation of β and it appears reasonable to estimate the interaction parameters θ conditionally on fixed $\widehat{\beta}$. See also Prokešová et al. (2014) for this discussion.

An alternative method suitable for estimation of β if λ is of log-linear form is the variational approach proposed by Coeurjolly and Møller (2014). Strong consistency and asymptotic normality of their estimator is shown in the paper, together with a simple and efficient implementation. This makes their approach a viable alternative to the classical Poisson log-likelihood (2.11).

2.2.1 Minimum contrast estimation

As stated above, the two-step estimation procedure using minimum contrast estimation on the K-function was first proposed for inhomogeneous SOIRS Poisson-Neyman-Scott processes in Waagepetersen (2007). The usefulness of the method was illustrated by application to the rain forest data (a popular dataset available in `spatstat`) and a simulation study. However, no asymptotic results for estimates of the clustering parameters were provided.

Following the idea in the above-mentioned paper Guan (2009) introduced a two-step estimation method with (weighted) minimum contrast using pair-correlation function g in the second step, again under SOIRS assumption. A simulation study presented in the paper shows that in certain situations this version produces more stable estimates than using the K-function in the second step. The method is also illustrated in an application to the rain forest dataset.

Finally, Waagepetersen and Guan (2009) prove joint asymptotic normality of the estimates $(\widehat{\beta}, \widehat{\theta})$ in the so-called increasing window asymptotics for spatial SOIRS Cox processes under appropriate moment assumptions and specific mixing conditions. The estimation procedure is formulated in the framework of estimating equations as follows:

$$\begin{aligned} m(\beta, \theta) &= \int_{r_0}^{r_1} [\widehat{K}(u;\beta)^c - K(u;\theta)^c]^2 \,\mathrm{d}u, \\ U_2(\beta, \theta) &= -|W| \frac{\partial m(\beta, \theta)}{\partial \theta} \\ &= 2c|W| \int_{r_0}^{r_1} [\widehat{K}(u;\beta)^c - K(u;\theta)^c] K(u;\theta)^{c-1} K^{(1)}(u;\theta) \,\mathrm{d}u, \end{aligned}$$

where m is the appropriate contrast criterion, W is the observation window, $\widehat{K}(\cdot;\beta)$ is the non-parametric estimate of the K-function (1.4) given the value of β and $K^{(1)}$ is the derivative of $K(\cdot;\theta)$ with respect to the vector of parameters θ. Finding the estimates $(\widehat{\beta}, \widehat{\theta})$ thus corresponds to solving

$$U(\beta, \theta) = (U_1(\beta), U_2(\beta, \theta)) = 0,$$

where U_1 is given by (2.12). We will use a similar formulation of estimation procedures in the next chapter.

2.2.2 Composite likelihood method

Different versions of the composite likelihood method for pairs of points were suggested in the inhomogeneous case by Waagepetersen (2007), Jalilian et al. (2013) and Prokešová et al. (2014). For example, the log-likelihood function in Prokešová et al. (2014) takes on the form

$$\log CL(\theta) = \sum_{x,y \in X \cap W, 0 < \|y-x\| < R} \Big[\log \Big(\lambda(x;\widehat{\beta}) \lambda(y;\widehat{\beta}) g(y-x;\theta) \Big) \tag{2.13}$$

$$- \log \left(\int_W \int_W \lambda(u;\widehat{\beta}) \lambda(v;\widehat{\beta}) g(v-u;\theta) I(\|v-u\| < R) \,\mathrm{d}u \,\mathrm{d}v \right) \Big], \tag{2.14}$$

cf. Equation (2.4). The log-likelihood with a fixed value of $\widehat{\beta}$ from the first step is then maximized with respect to the interaction parameters θ. As before, $R > 0$ is a tuning parameter. This variant of the two-step maximization is computationally much less demanding than maximization of the full composite likelihood with respect to the complete parameter (β, θ) (Prokešová et al., 2014).

Asymptotic properties of the estimator based on (2.13) under the increasing window asymptotics can be established using the methodology in Prokešová et al. (2014). For the version of Jalilian et al. (2013) the asymptotic properties are discussed in the appendix of their paper.

2.2.3 Palm likelihood method

It is not a straightforward task to generalize Palm likelihood method to the non-stationary case. Prokešová et al. (2014) describe three different approaches addressing this issue. To give an example, one version of the Palm log-likelihood takes on the form

$$\log PL(\theta) = \sum_{x,y \in X \cap W, 0 < \|y-x\| < R} \log\Big(\lambda(y;\widehat{\beta}) g(y-x;\theta)\Big) - \sum_{x \in X \cap W} \int_{B(x,R)} \lambda(u;\widehat{\beta}) g(u-x;\theta)\,\mathrm{d}u,$$

where again $R > 0$ is a tuning constant. The paper Prokešová et al. (2014) formulates the conditions under which the joint asymptotic normality of $\left(\widehat{\beta}, \widehat{\theta}\right)$ holds for different versions of the proposed inhomogeneous Palm likelihood, under the increasing window asymptotics. The author's contribution to the paper Prokešová et al. (2014) lies in performing the simulation study which accompanies the paper (not presented here).

3. Parameter estimation for space-time Cox point processes

As discussed in Section 1.4, space-time point processes in $\mathbb{R}^d \times \mathbb{R}$ should not just be considered processes in $\mathbb{R}^{d+1}$. One coordinate plays a distinct role and hence devoted space-time methods should be used for statistical inference.

The parametric models for spatial point processes presented in Sections 1.2 and 1.3 can be extended to the space-time setting. The discussion on parameter estimation at the beginning of Chapter 2 also applies here. For models with tractable form of the pair-correlation function (the K-function or the Palm intensity, respectively) the moment estimation methods may be used. Due to low computational demands they are preferred both for stationary and non-stationary processes.

In general one may expect space-time models to be parametrized by a higher number of parameters than purely spatial models considered in Chapter 2. This may cause trouble with the respective optimization procedures – optimization over a higher-dimensional parameter space is more computationally demanding and in some cases requires approximation of higher-dimensional integrals. Therefore we investigated the possibility to use dimensionality-reducing techniques (namely, using projection processes introduced in Section 1.4) to estimate different parts of the model separately. This, however, requires a particular structure of the model, as discussed below. Under certain assumptions the second-order characteristics of the projected processes X_s and X_t have a tractable form and can be used for step-wise estimation.

We focus here on the non-stationary case as it is more challenging and also more useful for practical applications. Also, we discuss here only the minimum contrast estimation – in this case the problems mentioned above are most prominent. The minimum contrast criterion involves higher-dimensional integration and needs to be optimized with respect to a high-dimensional parameter. Also, numerical stability of the necessary non-parametric estimates of K or g is an important issue.

In the following we present two estimation methods for inhomogeneous space-time shot-noise Cox processes with a particular structure – minimum contrast estimation using projection processes and also a refined version which remedies some of the drawbacks of the first method. We remark that inference for other space-time models with similar second-order structure can be made using the same methods.

3.1 Method using projection processes

Most of the work presented in this section is also summarized in the recent paper Prokešová and Dvořák (2014), except for the discussion on asymptotic properties of the respective estimators. For ease of exposition we focus on space-time processes on $\mathbb{R}^2 \times \mathbb{R}$ which are the most common type of space-time processes encountered in practice.

Let X_0 be a stationary space-time shot-noise Cox process on $\mathbb{R}^2 \times \mathbb{R}$ (see

Section 1.3), observed on a compact observation window $W \times T$ and specified by the constant $\mu > 0$, the measure V on $\mathbb{R}^+$ and the homogeneous kernel function $k(u,t)$ on $\mathbb{R}^2 \times \mathbb{R}$ such that k integrates to 1 (k is a probability density function on $\mathbb{R}^2 \times \mathbb{R}$). We require this to avoid overparametrization of the model.

We denote $\lambda_{0,k}$ the k-th order intensity function of X_0. Similarly, we denote $\lambda_{0,s,k}$ the k-th order intensity function of the spatial projection $X_{0,s}$ of X_0, and $\lambda_{0,t,k}$ the k-th order intensity function of the temporal projection $X_{0,t}$ of X_0. Throughout this section we assume that $\lambda_{0,2}$ exists and is bounded so that the pair-correlation function of X_0 is properly defined.

Let X be the SOIRS process obtained by location-dependent thinning from X_0 using the inhomogeneity function $f : \mathbb{R}^2 \times \mathbb{R} \to [0,1]$. As before, we denote by $\lambda_k, \lambda_{s,k}$ and $\lambda_{t,k}$ the k-th order intensity functions of X and of the projection processes X_s and X_t, respectively.

Following Gabriel and Diggle (2009) and Møller and Ghorbani (2012) we adopt a pragmatic assumption about the separability of the first-order intensity function λ (for a more detailed discussion on this assumption see also Section 4.1.1). We assume that

$$f(u,t) = f_1(u) f_2(t),\ u \in \mathbb{R}^2, t \in \mathbb{R}, \tag{3.1}$$

where $f_1 : \mathbb{R}^2 \to [0,1]$ is the spatial inhomogeneity function and $f_2 : \mathbb{R} \to [0,1]$ is the temporal inhomogeneity function. Moreover, we assume

$$\max_{u \in \mathbb{R}^2} f_1(u) = 1 = \max_{t \in \mathbb{R}} f_2(t),$$

which implies $\max_{\mathbb{R}^2 \times \mathbb{R}} f(u,t) = 1$. Again, this assumption prevents overparametrization of the model.

Further following an example of a structured space-time Poisson cluster process in Møller and Ghorbani (2012, Sec. 5) we assume a product structure of the kernel function k, i.e.

$$k(u,t) = k_1(u) k_2(t),\ u \in \mathbb{R}^2, t \in \mathbb{R}, \tag{3.2}$$

where k_1 and k_2 are probability density functions on $\mathbb{R}^2$ and $\mathbb{R}$, respectively.

These assumptions do not imply spatio-temporal separability of the process X (the second-order intensity function does not have a product structure, as shown below) but they allow us to introduce a tractable estimation procedure.

In many situations, e.g. for the epidemiological data discussed in the paper by Møller and Ghorbani (2012), the assumption (3.2) about space-time separability of the smoothing kernel k can be justified by arguing that temporal dynamics of the underlying (biological or other) process is not influenced by spatial variation. Also, in practical applications it may be difficult to decide whether the observed data arise from a process with separable or non-separable smoothing kernel. The main obstacle is that usually only a few points of the process occur in individual clusters. This makes (3.2) a reasonable assumption in many situations.

In the following we take advantage of the notation

$$K_1(v-u) = \int_{\mathbb{R}^2} k_1(u-w) k_1(v-w)\, \mathrm{d}w,\ u, v \in \mathbb{R}^2,$$
$$K_2(s-t) = \int_{\mathbb{R}} k_2(t-\tau) k_2(s-\tau)\, \mathrm{d}\tau,\ s, t \in \mathbb{R},$$

where K_1 is the probability density function of a difference of two independent, identically distributed random variables with density k_1, and similarly for K_2.

The moment characteristics of X are derived easily from the model assumptions and the formulae in Section 1.3. The intensity function is

$$\lambda(u,t) = f_1(u)f_2(t)\mu V_1,\ (u,t) \in \mathbb{R}^2 \times \mathbb{R}, \tag{3.3}$$

and the pair-correlation function is

$$g((u,t),(v,s)) = 1 + \frac{V_2}{\mu(V_1)^2} K_1(v-u)K_2(s-t),\ (u,t),(v,s) \in \mathbb{R}^2 \times \mathbb{R}. \tag{3.4}$$

Obviously neither g nor λ_2 has a space-time product structure and our process X has nontrivial spatio-temporal interactions.

Let us now consider the projection processes X_t and X_s, obtained by projecting X from $W \times \mathbb{R}$ to $\mathbb{R}$ and from $\mathbb{R}^2 \times T$ to $\mathbb{R}^2$, respectively. For their intensity functions λ_t and λ_s we get from Proposition 1.21 that

$$\lambda_t(t) = \mu V_1 f_2(t) \int_W f_1(w)\,\mathrm{d}w,\ t \in \mathbb{R}, \qquad \lambda_s(u) = \mu V_1 f_1(u) \int_T f_2(\tau)\,\mathrm{d}\tau,\ u \in \mathbb{R}^2. \tag{3.5}$$

From Proposition 1.21 and Equations (3.3)–(3.5) we get for their pair-correlation functions that

$$g_t(t,s) = 1 + C_t \frac{V_2}{\mu(V_1)^2} K_2(s-t),\ t,s \in \mathbb{R}, \tag{3.6}$$

$$g_s(u,v) = 1 + C_s \frac{V_2}{\mu(V_1)^2} K_1(v-u),\ u,v \in \mathbb{R}^2, \tag{3.7}$$

where the constants C_t, C_s are defined by

$$C_t = \frac{1}{(\int_W f_1(w)\,\mathrm{d}w)^2} \int_W \int_W f_1(u)f_1(v)K_1(v-u)\,\mathrm{d}u\,\mathrm{d}v, \tag{3.8}$$

$$C_s = \frac{1}{(\int_T f_2(\tau)\,\mathrm{d}\tau)^d} \int_T \int_T f_2(s)f_2(t)K_2(s-t)\,\mathrm{d}s\,\mathrm{d}t. \tag{3.9}$$

Thus eventhough the process X is non-separable the pair-correlation function of the temporal projection process depends on the "spatial" part of the model (i.e. f_1 and k_1) only through the constant C_t and analogically the pair-correlation function of the spatial projection process depends on f_2 and k_2 only through C_s.

Another second-order characteristic used in the sequel is the K-function. For the projection processes we have

$$K_s(r) = \int_{\|u\| \le r} g_s(u)\,\mathrm{d}u,\ r \ge 0, \qquad K_t(t) = \int_{-t}^{t} g_t(s)\,\mathrm{d}s,\ t \ge 0. \tag{3.10}$$

3.1.1 Model parametrization

Suppose that the underlying Poisson measure Φ has the intensity measure of the form $\mu V(\mathrm{d}r)\,\mathrm{d}(v,s)$ where $\mu > 0$ is a scalar parameter and the measure V is parametrized by the (scalar) parameter θ.

We assume a particular form of the inhomogeneity function f:

$$f(u,t;\beta_s,\beta_t) = f_1(z_1(u)\beta_s^T)\, f_2(z_2(t)\beta_t^T), \tag{3.11}$$

where $z_1(u)$ and $z_2(t)$ are vectors of spatial and temporal covariates, respectively, and, with a slight abuse of notation, f_1, f_2 are positive, strictly increasing functions on $\mathbb{R}$. The vectors β_s, β_t denote the unknown inhomogeneity parameters. Note that (3.11) implies (3.1) and thus enables us to work with characteristics of the projected processes. Also, it allows us to incorporate covariate information contained in the vectors $z_1(u)$ and $z_2(t)$.

We remark that the estimation procedure proposed below does not require us to impose a parametric model on f_1 and f_2. It is possible to use non-parametric methods such as kernel smoothing to estimate the first-order intensity function of X. However, for the discussion of the asymptotic properties of the estimators we require this parametrization.

We further set $\beta_0 = \log(\mu\, V_1)$. This is motivated by the popular parametrization used in practice where the intensity function is a log-linear function of the model parameters. Then the intensity function λ of X is parametrized by the vector $\beta = (\beta_0, \beta_s, \beta_t)$. Hence, the intensity functions λ_s and λ_t of the projection processes X_s and X_t, given in Equation (3.5), are also parametrized by the vector β.

Let the spatial smoothing kernel k_1 (and hence K_1) be parametrized by a vector of parameters $\widetilde{\psi}$ and similarly the temporal smoothing kernel k_2 (and hence K_2) be parametrized by the vector $\widetilde{\xi}$. Now we set

$$\begin{aligned}\psi_0 =& C_s \frac{V_2}{\mu\,(V_1)^2},\\ \xi_0 =& C_t \frac{V_2}{\mu\,(V_1)^2},\end{aligned}$$

and $\psi = (\psi_0, \widetilde{\psi})$ and similarly $\xi = (\xi_0, \widetilde{\xi})$. Henceforth

$$\begin{aligned} g_s(u;\psi) &= 1 + \psi_0 K_1(u;\widetilde{\psi}), & K_s(r;\psi) &= \pi r^2 + \psi_0 \int_{\|u\|<r} K_1(u;\widetilde{\psi})\,\mathrm{d}u,\\ g_t(s;\xi) &= 1 + \xi_0 K_2(s;\widetilde{\xi}), & K_t(t;\xi) &= 2t + \xi_0 \int_{|s|<t} K_2(s;\widetilde{\xi})\,\mathrm{d}s. \end{aligned}$$

Below we also use the following notation: $\lambda^{(1)}$ and $\lambda^{(2)}$ are the first and second-order derivatives of the intensity function of X w.r.t. the parameter β; $K_t^{(1)}(t;\xi), K_t^{(2)}(t;\xi)$ are the first and second-order derivatives of $K_t(t;\xi)$ w.r.t. ξ; $K_s^{(1)}(r;\psi), K_s^{(2)}(r;\psi)$ are the first and second-order derivatives of $K_s(r;\psi)$ w.r.t. ψ (assuming that the appropriate derivatives exist).

Finally, we denote by $\beta^*, \psi^*, \xi^*, \mu^*, \theta^*$ the "true" parameter values governing the distribution of the process X.

3.1.2 Step-wise estimation procedure

In this section we introduce a step-wise estimation procedure for X analogical to the estimation of spatial SOIRS Cox point processes (Waagepetersen and Guan,

2009). We assume that the process X is observed in a compact space-time observation window of the form $W \times T$ where $W \subset \mathbb{R}^2$ and $T \subset \mathbb{R}$ have a positive Lebesgue measure (area and length, respectively).

In the estimation procedure we take advantage of the special spatio-temporal structure of X. In the first step the inhomogeneous intensity function λ is estimated (either parametrically or non-parametrically). In the second and third step, conditionally on the knowledge of λ, the second-order characteristics of X are estimated from the data and used for minimum contrast estimation in order to estimate the parameters of the temporal and spatial projection process, respectively. Finally, the parameters of the underlying Poisson measure (i.e. μ and θ) are estimated from the previous estimates and the total number of points observed in $W \times T$.

First step – estimating the first-order space-time intensity function

From (3.3) and (3.5) we get for any $(u,t) \in W \times T$ that

$$\lambda_s(u)\lambda_t(t) = \lambda(u,t) \cdot \int_W \int_T \mu V_1 f_1(v) f_2(s) \,\mathrm{d}s \,\mathrm{d}v = \lambda(u,t) \cdot \mathbb{E}\, X(W \times T).$$

A natural (unbiased) estimator of the mean number of points in $X \cap (W \times T)$ is the actual observed number of points in $X \cap (W \times T)$. Therefore we define the estimate of the space-time intensity function $\widehat{\lambda}$ by

$$\widehat{\lambda}(u,t) = \frac{\widehat{\lambda}_s(u)\widehat{\lambda}_t(t)}{X(W \times T)},\ (u,t) \in W \times T, \tag{3.12}$$

where $\widehat{\lambda}_s$, $\widehat{\lambda}_t$ are estimates of the intensity functions of the projection processes. Thus, the dimensionality of the problem is reduced.

The intensities λ_s, λ_t may be estimated either non-parametrically or a model for the thinning functions f_1, f_2 may be specified, see (3.11). The non-parametric estimation of λ_s is usually implemented by the kernel estimate

$$\widehat{\lambda}_s(u) = \sum_{x \in X_s} h_b(u-x)/w_{b,W}(x),\ u \in W,$$

where $h_b(u) = h(\frac{u}{b})\, b^{-2}$ is a kernel with bandwidth $b > 0$, i.e. h is a fixed probability density function. The edge correction factors $w_{b,W}$ are defined by

$$w_{b,W}(x) = \int_W h_b(u-x) \,\mathrm{d}u,$$

so that $\int_W \widehat{\lambda}_s(u) \,\mathrm{d}u = X_s(W)$ which implies the approximate unbiasedness of the estimate $\widehat{\lambda}_s$. An analogical kernel estimate can be used for $\widehat{\lambda}_t$. For further information about kernel estimation for point processes see e.g. Møller and Waagepetersen (2004, Sec 4.3).

From the (non-parametric) estimates of λ_s and λ_t we also get the estimates of the thinning functions f_1, f_2. Namely, from (3.5), we have

$$\widehat{f_1}(u) = \widehat{\lambda}_s(u) / \left(\max_W \widehat{\lambda}_s(v)\right), \qquad \widehat{f_2}(t) = \widehat{\lambda}_t(t) / \left(\max_T \widehat{\lambda}_t(s)\right),$$

if we assume $\max_{u\in W} f_1(u) = 1 = \max_{t\in T} f_2(t)$.

The product estimation (3.12) makes the non-parametric estimate of the intensity more stable in comparison with a fully space-time kernel estimate. However, we stress that the accuracy of $\widehat{\lambda}, \widehat{\lambda}_s, \widehat{\lambda}_t$ is not important just for its own sake but even more because the inverted values of the intensity functions are used in the estimates of the inhomogeneous pair-correlation functions and other second-order characteristics used in the following estimation steps. Therefore it is recommendable to use the more stable parametric estimate of λ when possible (see Baddeley et al. (2000) for the detailed discussion of this issue).

Ignoring for the moment the inter-point interactions, the inhomogeneity parameter $\beta = (\beta_0, \beta_s, \beta_t)$ may be estimated by means of the Poisson log-likelihood score function similar to (2.12). For X the Poisson log-likelihood score function is given by

$$U_1(\beta) = \sum_{(u,t)\in X\cap(W\times T)} \frac{\lambda^{(1)}(u,t;\beta)}{\lambda(u,t;\beta)} - \int_{W\times T} \lambda^{(1)}(v,s;\beta)\,\mathrm{d}v\,\mathrm{d}s. \tag{3.13}$$

The estimate $\widehat{\beta}$ is obtained as a solution of the vector equation $U_1(\beta) = 0$.

We remark here that one may try to take advantage of the product estimation (3.12) just as in the non-parametric case and use variants of the Poisson log-likelihood score function (3.13) separately for (suitably parametrized) intensity functions λ_s, λ_t of the projection processes X_s, X_t. However, the resulting estimating equations corresponding to β_s and β_t, respectively, exactly match those obtained from (3.13). Thus, no improvement in precision of the estimates of β_s and β_t can be achieved by using the projection processes instead of the space-time process X itself.

Second step – estimating the clustering parameters of the temporal projection process

To estimate the parameter ξ we will use minimum contrast estimation for the projection process X_t. For this purpose we need non-parametric estimates of the K-function or g-function (or semi-parametric estimates if a parametric model was imposed on the inhomogeneity function f). In the general setting they are given by Equations (1.4) and (1.5). However, for the sake of completeness we give here the formula for non-parametric estimate for the projection process X_t, too:

$$\widehat{K}_t(t) = \frac{1}{|T|} \sum_{s,\tau\in X_t}^{\neq} \frac{I(|s-\tau|\le t)}{w_2(s,\tau)\widehat{\lambda}_t(s)\widehat{\lambda}_t(\tau)},$$

where $w_2(s,\tau)$ is the temporal edge-correction factor which is equal to 1 if both ends of the interval with length $2|s-\tau|$ and center s lie within T, and is equal to 2 otherwise (Møller and Ghorbani, 2012). For further discussion on different choices of edge-correction factors see Gabriel (2014).

Non-parametric estimate of the pair-correlation function g_t can be constructed analogously, with appropriate edge-correction factor, following Equation (1.5). The disadvantage of using $\widehat{g}_t$ is the necessity of choosing a suitable kernel and its bandwidth $b > 0$. No tuning constants are necessary for estimation of K.

For this reason we recommend using the minimum contrast estimation with the K-function for estimation of the temporal clustering parameters so as to avoid the need of choosing the kernel and an optimal (or at least sensible) value of the bandwidth b. Therefore we estimate ξ by minimizing the contrast

$$\int_{t_0}^{t_1} (\widehat{K}_t(t)^{c_2} - K_t(t;\xi)^{c_2})^2 \,\mathrm{d}t, \tag{3.14}$$

where $c_2 > 0$ is the variance-stabilizing exponent, usually taking on values $c_2 = 1/2$ or $1/4$, and $0 \le t_0 < t_1$ are fixed constants.

Third step – estimating the clustering parameters of the spatial projection process

To estimate the parameter ψ we will use minimum contrast estimation for the projection process X_s. For this purpose we need non-parametric or semi-parametric estimates of the K-function or g-function (if k_1 and hence also g_s is isotropic). We give here the formula for the estimation of the K-function of the projection process X_s:

$$\widehat{K}_s(r) = \frac{1}{|W|} \sum_{x,y \in X_s}^{\neq} \frac{I(\|x-y\| \le r)}{w_1(x,y)\widehat{\lambda}_s(x)\widehat{\lambda}_s(y)},$$

where w_1 is an edge-correction factor. In the planar case $w_1(x,y)$ can be either the translation edge-correction factor $\frac{|W \cap W_{x-y}|}{|W|}$, where W_{x-y} denotes the set W shifted by the vector $x-y$, or the isotropic edge correction factor. For a detailed discussion on different choices of edge-correction factors see Gabriel (2014).

Non-parametric or semi-parametric estimate of the pair-correlation function g_s can be constructed analogously, with appropriate edge-correction factor, following Equation (1.5).

The disadvantage of using $\widehat{g}_s$ is the necessity of choosing appropriate kernel h and its bandwidth $b > 0$. No tuning constants are necessary for estimation of K. This is the reason why minimum contrast estimation using the K-function is more popular, see e.g. Waagepetersen (2007), Waagepetersen and Guan (2009), Møller and Ghorbani (2012). However, it was shown (Guan, 2009, Dvořák and Prokešová, 2012) that for the planar Poisson-Neyman-Scott cluster processes, both homogeneous and inhomogeneous, the minimum contrast estimation with the pair-correlation function provides better estimates of the clustering parameters than minimum contrast with K-function. Therefore we suggest to estimate the spatial clustering parameters by minimizing the contrast

$$\int_{r_0}^{r_1} (\widehat{g}_s(u)^{c_3} - g_s(u;\psi)^{c_3})^2 \,\mathrm{d}u, \tag{3.15}$$

where $c_3 > 0$ is the variance-stabilizing exponent. The usual choices are $c_3 = 1/2$ or $1/4$. Similarly as before, $0 \le r_0 < r_1$ are fixed constants.

Alternatively, the parameter ψ can be estimated by minimizing the contrast

$$\int_{r_0}^{r_1} (\widehat{K_s}(u)^{c_3} - K_s(u;\psi)^{c_3})^2 \,\mathrm{d}u.$$

This version is useful for studying asymptotic properties of the estimator, see below. We remark here that in general, for estimates obtained by minimum contrast estimation with the g-function asymptotic results have not been established yet. This is mainly due to the lack of convergence results for the empirical estimates of the g-function.

Final step – estimating the parameters of the underlying Poisson measure

Once the estimates of ψ and ξ have been computed, we can plug them into formulas (3.8), (3.9) and from ψ_0 or ξ_0 obtain the estimate of $\alpha = \frac{V_2}{\mu(V_1)^2}$. Finally, we calculate $\widehat{\theta}$ and $\widehat{\mu}$ from $\widehat{\alpha}$ and the equation

$$X(W \times T) = \mu V_1 \int_W \widehat{f}_1(u)\,\mathrm{d}u \int_T \widehat{f}_2(t)\,\mathrm{d}t, \tag{3.16}$$

where $X(W \times T)$ plays role of the estimate of $\mathbb{E}\,X(W \times T) = \int_{W\times T} \lambda(u,t)\,\mathrm{d}u\,\mathrm{d}t$. The actual form of the calculations depends on the precise form of V_1 and V_2 depending on θ.

Note that we can actually obtain two different estimates of α – one by using $\widehat{\psi_0}$ and $\widehat{C}_s$ and another by using $\widehat{\xi_0}$ and $\widehat{C}_t$. According to the results of the simulation study in Section 3.1.5 it is preferable to use $\widehat{\psi_0}$ and $\widehat{C}_s$ since they lead to more stable and substantially better estimates of μ and θ. A plausible explanation for this fact is the smaller information loss when projecting X to X_s (i.e. from $\mathbb{R}^2 \times T$ to $\mathbb{R}^2$) in comparison with projecting X to X_t. This leads to comparatively better estimate of ψ_0 in comparison with ξ_0.

Reformulation of the estimation procedure

To set up the stage for the discussion on the asymptotic properties of the estimators we first formulate the estimation procedure in terms of estimating equations. We focus here on estimation of β, ξ and ψ. For the parameters μ and θ the calculations depend on the particular form of the measure V (more precisely, on the form of V_1 and V_2) and it is complicated to discuss the asymptotics in the general setting. Also, the inhomogeneity parameter β and the clustering parameters ξ and ψ are likely to be of main interest in practical applications.

We assume the parametric form of the first-order intensity function (3.11). In the first step, $\widehat{\beta}$ is obtained using the Poisson log-likelihood score function (3.13), i.e. by solving the estimating equation

$$U_1(\beta) = \sum_{(u,t)\in X\cap(W\times T)} \frac{\lambda^{(1)}(u,t;\beta)}{\lambda(u,t;\beta)} - \int_{W\times T} \lambda^{(1)}(v,s;\beta)\,\mathrm{d}v\,\mathrm{d}s = 0. \tag{3.17}$$

In the second step we use $\widehat{\beta}$ to calculate the semi-parametric estimates $\widehat{K}_t(t;\widehat{\beta})$ of $K_t(t;\xi)$. We use the translation edge-correction here (Gabriel, 2014) as it is convenient for our discussion on the asymptotic properties of the estimators. We then minimize the discrepancy

$$m_{2,\widehat{\beta}}(\xi) = \int_{t_0}^{t_1} (\widehat{K}_t(t;\widehat{\beta})^{c_2} - K_t(t;\xi)^{c_2})^2\,\mathrm{d}t. \tag{3.18}$$

Assuming differentiability of $m_{2,\widehat{\beta}}(\cdot)$ this corresponds to solving the estimating equation

$$\begin{aligned} U_2(\widehat{\beta},\xi) =& -|T|\frac{\partial m_{2,\widehat{\beta}}(\xi)}{\partial \xi} \\ =& 2c_2|T|\int_{t_0}^{t_1}\left[\widehat{K}_t(t;\widehat{\beta})^{c_2} - K_t(t;\xi)^{c_2}\right] K_t(t;\xi)^{c_2-1}K_t^{(1)}(t;\xi)\,\mathrm{d}t = 0. \end{aligned}$$

Similarly, in the third step we use $\widehat{\beta}$ to calculate the semi-parametric estimates $\widehat{K}_s(r;\widehat{\beta})$ of $K_s(r;\psi)$. Again, we use the translation edge-correction method. We then minimize the discrepancy

$$m_{3,\widehat{\beta}}(\psi) = \int_{r_0}^{r_1}(\widehat{K_s}(u;\widehat{\beta})^{c_3} - K_s(u;\psi)^{c_3})^2\,\mathrm{d}u. \tag{3.19}$$

Assuming differentiability of $m_{3,\widehat{\beta}}(\cdot)$ this corresponds to solving the estimating equation

$$\begin{aligned} U_3(\widehat{\beta},\psi) =& -|W|\frac{\partial m_{3,\widehat{\beta}}(\psi)}{\partial \psi} \\ =& 2c_3|W|\int_{r_0}^{r_1}\left[\widehat{K}_s(r;\widehat{\beta})^{c_3} - K_s(r;\psi)^{c_3}\right] K_s(r;\psi)^{c_3-1}K_s^{(1)}(r;\psi)\,\mathrm{d}r = 0. \end{aligned}$$

Altogether, the described estimation procedure can be formulated as solving the vector estimating equation

$$U(\beta,\xi,\psi) = (U_1(\beta), U_2(\beta,\xi), U_3(\beta,\psi)) = 0$$

to obtain the parameter estimates $\widehat{\beta}, \widehat{\xi}$ and $\widehat{\psi}$.

3.1.3 Asymptotic properties

In this section we briefly discuss the asymptotic properties of the estimators considered above under so-called *increasing window asymptotics*, i.e. when the data is available from an increasing sequence of compact observation windows. For a detailed discussion on the asymptotics see Section 3.2.3. Here we formulate the consistency and asymptotic normality results but omit the proofs as they precisely correspond to the proofs detailed in Section 3.2.3.

We consider an increasing sequence of compact observation windows $W_n \times T_n$, $n \geq 1$, such that $W_n \times T_n \nearrow \mathbb{R}^2 \times \mathbb{R}$. For simplicity we choose

$$W_n = [an, bn] \times [cn, dn], \qquad T_n = [en, fn],$$

where $a < 0 < b, c < 0 < d$ and $e < 0 < f$ are fixed constants. More general shapes of the observation windows are also possible. However, the important property of the sequence of observation windows necessary in the proofs is that, for any $h \in \mathbb{R}^2$ and $k \in \mathbb{R}$,

$$\lim_{n\to\infty}\frac{|W_n \times T_n|}{|(W_n \cap W_{n,h}) \times (T_n \cap T_{n,k})|} = 1,$$

where $W_{n,h}$ denotes the set W_n shifted by the vector $h \in \mathbb{R}^2$, and similarly for $T_{n,k}$. Also, we require that $|\partial(W_n \times T_n)|/|W_n \times T_n| \to 0, n \to \infty$, where $\partial(W_n \times T_n)$ is the boundary of $W_n \times T_n$.

Let $(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)$ be the estimated parameter values calculated from $W_n \times T_n$, i.e. the solution of the equation

$$U_n(\beta, \xi, \psi) = (U_{n,1}(\beta), U_{n,2}(\beta, \xi), U_{n,3}(\beta, \psi)) = 0,$$

where

$$\begin{aligned}
U_{n,1}(\beta) =& \sum_{(u,t) \in X \cap (W_n \times T_n)} \frac{\lambda^{(1)}(u, t; \beta)}{\lambda(u, t; \beta)} - \int_{W_n \times T_n} \lambda^{(1)}(v, s; \beta) \, \mathrm{d}v \, \mathrm{d}s, \\
U_{n,2}(\beta, \xi) =& 2c_2 |T_n| \int_{t_0}^{t_1} \left[\widehat{K}_{t,n}(t; \beta)^{c_2} - K_t(t; \xi)^{c_2} \right] K_t(t; \xi)^{c_2 - 1} K_t^{(1)}(t; \xi) \, \mathrm{d}t, \\
U_{n,3}(\beta, \psi) =& 2c_3 |W_n| \int_{r_0}^{r_1} \left[\widehat{K}_{s,n}(r; \beta)^{c_3} - K_s(r; \psi)^{c_3} \right] K_s(r; \psi)^{c_3 - 1} K_s^{(1)}(r; \psi) \, \mathrm{d}r,
\end{aligned}$$

and $\widehat{K}_{t,n}$ and $\widehat{K}_{s,n}$ are the semi-parametric estimates of K_t and K_s, calculated using $X_t \cap T_n$ and $X_s \cap W_n$, respectively.

Note that, for all $n \geq 1$, we use the same temporal projection process X_t (projected from the fixed spatial region W) to define $U_{n,2}$ and the same spatial projection process X_s (projected from the fixed time interval T) to define $U_{n,3}$. If we used e.g. T_n to define projection processes $X_s^{(n)}$, the resulting asymptotic regime for $U_{n,3}$ would be a combination of the increasing window asymptotics and so-called *infill asymptotics*, i.e. the intensity function of $X_s^{(n)}$ (at any location) would be increasing, unbounded function of n. Moreover, second-order moment characteristics of $X_s^{(n)}$ converge to those of a Poisson process and thus in the limit they provide no information about the clustering parameters. For more detailed discussion on this issue see Section 3.2.3.

Following Waagepetersen and Guan (2009) we approximate $U_{n,2}(\beta^*, \xi^*)$ and $U_{n,3}(\beta^*, \psi^*)$ by

$$\begin{aligned}
\widetilde{U}_{n,2}(\beta^*, \xi^*) &= 2c_2^2 |T_n| \int_{t_0}^{t_1} \left[\widehat{K}_{t,n}(t; \beta^*) - K_t(t; \xi^*) \right] K_t(t; \xi^*)^{2c_2 - 2} K_t^{(1)}(t; \xi^*) \, \mathrm{d}t, \\
\widetilde{U}_{n,3}(\beta^*, \psi^*) &= 2c_3^2 |W_n| \int_{r_0}^{r_1} \left[\widehat{K}_{s,n}(r; \beta^*) - K_s(r; \psi^*) \right] K_s(r; \psi^*)^{2c_3 - 2} K_s^{(1)}(r; \psi^*) \, \mathrm{d}r.
\end{aligned}$$

For $\widetilde{U}_{n,2}$, this is based on a Taylor series expansion of the function x^{c_2}, applied on $\widehat{K}_{t,n}(t; \beta)^{c_2} - K_t(t; \xi)^{c_2}$, and similarly for $\widetilde{U}_{n,3}$.

We further define

$$\begin{aligned}
\Sigma_{n,11} &= |W_n \times T_n|^{-1} \operatorname{\mathsf{Var}} \left(U_{n,1}(\beta^*) \right), \\
\widetilde{\Sigma}_{n,22} &= |T_n|^{-1} \operatorname{\mathsf{Var}} \left(\widetilde{U}_{n,2}(\beta^*, \xi^*) \right), \\
\widetilde{\Sigma}_{n,33} &= |W_n|^{-1} \operatorname{\mathsf{Var}} \left(\widetilde{U}_{n,3}(\beta^*, \psi^*) \right),
\end{aligned}$$

$$
\begin{aligned}
J_n(\beta,\xi,\psi) &= -\frac{\partial}{\partial(\beta,\xi,\psi)^T}\, U_n(\beta,\xi,\psi) \\
&= -\begin{pmatrix} \frac{\partial}{\partial\beta^T}\, U_{n,1}(\beta) & \frac{\partial}{\partial\beta^T}\, U_{n,2}(\beta,\xi) & \frac{\partial}{\partial\beta^T}\, U_{n,3}(\beta,\psi) \\ 0 & \frac{\partial}{\partial\xi^T}\, U_{n,2}(\beta,\xi) & 0 \\ 0 & 0 & \frac{\partial}{\partial\psi^T}\, U_{n,3}(\beta,\psi) \end{pmatrix} \\
&= \begin{pmatrix} J_{n,11}(\beta) & J_{n,12}(\beta,\xi) & J_{n,13}(\beta,\psi) \\ 0 & J_{n,22}(\beta,\xi) & 0 \\ 0 & 0 & J_{n,33}(\beta,\psi) \end{pmatrix}
\end{aligned}
$$

and

$$
\begin{aligned}
I_{n,11} &= \frac{1}{|W_n \times T_n|} \int_{W_n \times T_n} \frac{\lambda^{(1)}(v,s;\beta^*)^T \lambda^{(1)}(v,s;\beta^*)}{\lambda(v,s;\beta^*)}\, \mathrm{d}v\, \mathrm{d}s, \\
I_{n,12} &= -2c_2^2 \int_{t_0}^{t_1} H_{n,2}(t;\beta^*) K_t(t;\xi^*)^{2c_2-2} K_t^{(1)}(t;\xi^*)\, \mathrm{d}t, \\
I_{n,13} &= -2c_3^2 \int_{r_0}^{r_1} H_{n,3}(r;\beta^*) K_s(r;\psi^*)^{2c_3-2} K_s^{(1)}(r;\psi^*)\, \mathrm{d}r, \\
I_{22} &= 2c_2^2 \int_{t_0}^{t_1} K_t(t;\xi^*)^{2c_2-2} K_t^{(1)}(t;\xi^*)^T K_t^{(1)}(t;\xi^*)\, \mathrm{d}t, \\
I_{33} &= 2c_3^2 \int_{r_0}^{r_1} K_s(r;\psi^*)^{2c_3-2} K_s^{(1)}(r;\psi^*)^T K_s^{(1)}(r;\psi^*)\, \mathrm{d}r,
\end{aligned}
$$

where

$$
\begin{aligned}
H_{n,2}(t;\beta^*) &= \mathbb{E}\, \frac{\partial}{\partial\beta^T} \widehat{K}_{t,n}(t;\beta)|_{\beta=\beta^*} \\
&= -2 \int_{T_n} \int_{T_n} \frac{I\{|s-\tau| < t\}}{|T_n \cap T_{n,s-\tau}|} \frac{\lambda_t^{(1)}(s;\beta^*)}{\lambda_t(s;\beta^*)} g_t(s-\tau;\xi^*)\, \mathrm{d}s\, \mathrm{d}\tau, \\
H_{n,3}(t;\beta^*) &= \mathbb{E}\, \frac{\partial}{\partial\beta^T} \widehat{K}_{s,n}(r;\beta)|_{\beta=\beta^*} \\
&= -2 \int_{W_n} \int_{W_n} \frac{I\{\|u-v\| < r\}}{|W_n \cap W_{n,u-v}|} \frac{\lambda_s^{(1)}(u;\beta^*)}{\lambda_s(u;\beta^*)} g_s(u-v;\psi^*)\, \mathrm{d}u\, \mathrm{d}v.
\end{aligned}
$$

Now we can formulate the consistency theorem, inspired by the paper Waagepetersen and Guan (2009).

Theorem 3.1. *Apart from the model assumptions formulated above, let the following conditions be met:*

(A1) the inhomogeneity function f is twice continuously differentiable as a function of β,

(A2) $\exists C_1 < \infty$ such that $\|z_1(u)\| < C_1, \|z_2(t)\| < C_1, u \in \mathbb{R}^2, t \in \mathbb{R}$,

(A3) I_{22} and I_{33} are positive definite matrices and $\liminf \omega_{n,11} > 0$, where $\omega_{n,11}$ is the smallest eigenvalue of $I_{n,11}$,

(A4) $\widetilde{\Sigma}_{n,22}$ and $\widetilde{\Sigma}_{n,33}$ converge to positive definite matrices $\widetilde{\Sigma}_{22}$ and $\widetilde{\Sigma}_{33}$, respectively,

(A5) $\lambda^{(2)}(u,t;\beta), \frac{\lambda^{(1)}(u,t;\beta)}{\lambda(u,t;\beta)}, \frac{\lambda^{(2)}(u,t;\beta)}{\lambda(u,t;\beta)}$ *are bounded, continuous functions of* (u,t,β),

(A6) $K_t(t;\xi), K_t^{(1)}(t;\xi), K_t^{(2)}(t;\xi)$ *exist and are continuous functions of* (t,ξ),

(A7) $K_s(r;\psi), K_s^{(1)}(r;\psi), K_s^{(2)}(r;\psi)$ *exist and are continuous functions of* (r,ψ),

(A8) $t_0 \geq 0$ *for* $c_2 \geq 2$, *otherwise* $t_0 > 0$; *similarly,* $r_0 \geq 0$ *for* $c_3 \geq 2$, *otherwise* $r_0 > 0$,

(A9) $\lambda_{0,2}$ *and* $\lambda_{0,3}$ *exist and are bounded and the second-order reduced factorial cumulant measure of* X_0 *has finite total variation,*

(A10) $\exists C_2 < \infty$ *such that for all* $s_1, s_2 \in \mathbb{R}$:

$$\int_{\mathbb{R}} |\lambda_{0,t,4}(0, s_1, \tau, s_2 + \tau) - \lambda_{0,t,2}(0, s_1)\, \lambda_{0,t,2}(0, s_2)|\, \mathrm{d}\tau < C_2,$$

(A11) $\exists C_3 < \infty$ *such that for all* $u_1, u_2 \in \mathbb{R}^2$:

$$\int_{\mathbb{R}^2} |\lambda_{0,s,4}(0, u_1, v, u_2 + v) - \lambda_{0,s,2}(0, u_1)\, \lambda_{0,s,2}(0, u_2)|\, \mathrm{d}v < C_3.$$

Then there is a sequence $\left\{(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)\right\}_{n\geq 1}$ *for which*

$$U_n(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n) = 0$$

with probability tending to 1 and the vector

$$M_n = \begin{pmatrix} |W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*) \\ |T_n|^{1/2}(\widehat{\xi}_n - \xi^*) \\ |W_n|^{1/2}(\widehat{\psi}_n - \psi^*) \end{pmatrix}$$

is bounded in probability, i.e. $\forall \varepsilon > 0\ \exists \delta > 0 : \mathbb{P}(\|M_n\| > \delta) \leq \epsilon$ *for* n *sufficiently large.*

The theorem establishes existence of a consistent sequence $\{(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)\}_{n\geq 1}$ such that $(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)$ corresponds to a root of the estimating function $U_n(\beta, \xi, \psi)$ with probability tending to 1 (this also means that the estimating function actually *has* at least one root with probability tending to 1). If U_n has precisely one root for each n, we can find the consistent sequence $\{(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)\}_{n\geq 1}$ simply by finding the roots.

We now give a few comments regarding the assumptions of Theorem 3.1. Condition *(A1)* is not restrictive and covers the situation most popular in applications where f is assumed to be a log-linear function of β. The assumption *(A2)* of bounded covariates is easily justifiable from a practical point of view.

Concerning condition *(A3)*, I_{22} is a positive definite matrix if there are distinct values $t_0 < \tau_1 < \tau_2 < \ldots < \tau_q < t_1$ (where q is the number of elements of the vector ξ^*) such that the matrix with rows $K_t^{(1)}(\tau_i; \xi^*)$ has full rank, see also the example in Waagepetersen and Guan (2009, Sec. 3.3). It is not possible to discuss

in full generality the condition on the smallest eigenvalue of $I_{n,11}$ since it depends on the behaviour of the covariates over the whole $\mathbb{R}^2 \times \mathbb{R}$. The same applies to the condition *(A4)* on the limiting behaviour of the matrices $\widetilde{\Sigma}_{n,22}$ and $\widetilde{\Sigma}_{n,33}$.

(A5) in fact adds continuity assumptions on the first-order intensity function as a function of the vector (u, t, β) and hence continuity assumptions on the covariates $z_1(u)$ and $z_2(t)$, respectively, see Equation 3.11.

Conditions *(A6)* and *(A7)* depend on the particular form of the smoothing kernels k_1 and k_2, respectively. They are satisfied e.g. for Gaussian kernels. In case of the uniform (Matérn-type) kernels the situation is somewhat complicated and the dimension plays an important role. Namely, straightforward calculations show that for a uniform circular kernel k_1 in $\mathbb{R}^2$ the condition *(A7)* is fulfilled but for a uniform kernel k_2 in $\mathbb{R}$ the assumption *(A6)* does not hold.

Assumption *(A8)* is a technical nuisance but in applications it is possible to use very small positive values of t_0 and r_0 with no harm. Either a fixed positive value or perhaps the minimum observed interpoint distance in the given dataset may be used.

The condition *(A9)* relates to the stationary (unthinned) version X_0 of the process. It follows from the formulae in Hellmund et al. (2008, Sec. 4) that if the smoothing kernel k is bounded and $\int_0^\infty r^k V(\mathrm{d}r) < \infty$ for some $k \in \mathbb{N}$ then $\lambda_{0,k}$ is bounded and all reduced factorial cumulant measures up to order k have finite total variation. Hence *(A9)* is fulfilled if these properties hold for $k = 3$.

We also remark that a sufficient condition for *(A11)*, formulated in terms of the second and fourth-order intensity functions of the space-time process X_0, is that there is a constant $\widetilde{C}_3$ such that for all $u_1, u_2 \in \mathbb{R}^2$ and $s_1, s_2, s_3 \in \mathbb{R}$ the following holds:

$$\int_{\mathbb{R}^2} |\lambda_{0,4}((0,0),(u_1,s_1),(v,s_2),(u_2+v,s_3)) - \lambda_{0,2}((0,0),(u_1,s_1))\,\lambda_{0,2}((0,0),(u_2,s_3-s_2))|\,\mathrm{d}v < \widetilde{C}_3.$$

Similar sufficient condition can be formulated also for *(A10)*.

We now proceed to the formulation of the asymptotic normality results for the estimators considered above.

For general σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ let

$$\alpha(\mathcal{F}_1, \mathcal{F}_2) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$$

denote the standard strong mixing coefficient (Doukhan, 1994).

First we discuss the properties of the estimator $\widehat{\beta}_n$ based on the space-time process X. For a Borel set $A \in \mathcal{B}(\mathbb{R}^2 \times \mathbb{R})$ denote $\mathcal{F}^X(A)$ the σ-algebra generated by $X \cap A$.

For $h > 0$ let $A_{ijk} = [ih, (i+1)h) \times [jh, (j+1)h) \times [kh, (k+1)h)$, $(i, j, k) \in \mathbb{Z}^3$, and

$$\alpha^F_{p_1,p_2}(m) = \sup\Bigg\{\alpha(\mathcal{F}^X(S_1), \mathcal{F}^X(S_2)) : S_1 = \bigcup_{M_1} A_{ijk}, S_2 = \bigcup_{M_2} A_{ijk}, |M_1| \le p_1, |M_2| \le p_2, d(M_1, M_2) \ge m, M_1, M_2 \subset \mathbb{Z}^3\Bigg\},$$

where $|M|$ is the cardinality of the set $M \subseteq \mathbb{Z}^3$ and $d(M_1, M_2)$ denotes the minimal distance between M_1 and M_2 in the grid $\mathbb{Z}^3$.

Theorem 3.2. *Apart from the model assumptions formulated above and* (A1)–(A11)*, suppose there exist* $\delta > 0$ *and* $\nu \in \mathbb{N}, 0 < \delta < \nu$*, such that*

(B1) $\lambda_{0,2+\nu}((u_1, t_1), \ldots, (u_{2+\nu}, t_{2+\nu})) < \infty$*,*

(B2) there exist $h > 0$ *and* $d > 3 \cdot \frac{2+\delta}{\delta}$ *such that* $\alpha^F_{2,\infty}(m) = \mathcal{O}(m^{-d})$*,*

(B3) the matrix $\Sigma_{n,11}$ *converges to a positive-definite matrix* Σ_{11}*.*

Then

$$|W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*) I_{n,11} \Sigma_{n,11}^{-1/2} \xrightarrow{d} N(0, \mathbf{1}),$$

where $\xrightarrow{d}$ *denotes convergence in distribution and* $\mathbf{1}$ *is the identity matrix of appropriate dimension.*

Now we focus on the properties of the estimator $\widehat{\xi}_n$ based on the temporal projection process X_t. For a Borel set $B \in \mathcal{B}(\mathbb{R})$ denote $\mathcal{F}^{X_t}(B)$ the σ-algebra generated by $X_t \cap B$.

For $h > 0$ let $B_i = [ih, (i+1)h)$, $i \in \mathbb{Z}$, and

$$\alpha^F_{t,p_1,p_2}(m) = \sup \Bigg\{ \alpha(\mathcal{F}^{X_t}(S_1 \oplus t_1), \mathcal{F}^{X_t}(S_2 \oplus t_1)) : S_1 = \bigcup_{M_1} B_i, S_2 = \bigcup_{M_2} B_i, \\ |M_1| \le p_1, |M_2| \le p_2, d(M_1, M_2) \ge m, M_1, M_2 \subset \mathbb{Z} \Bigg\},$$

where $|M|$ is the cardinality of the set $M \subseteq \mathbb{Z}$ and $d(M_1, M_2)$ denotes the minimal distance between M_1 and M_2 in the grid $\mathbb{Z}$. Also, $S_i \oplus t_1$ denotes the set S_i dilated by the distance t_1 where t_1 is the upper limit used in the minimum contrast criterion (3.18).

Theorem 3.3. *Apart from the model assumptions formulated above and* (A1)–(A11)*, suppose there exist* $\delta > 0$ *and* $\nu \in \mathbb{N}, 0 < \delta < \nu$*, such that*

(C1) $\lambda_{0,4+2\nu}((u_1, t_1), \ldots, (u_{4+2\nu}, t_{4+2\nu})) < \infty$*,*

(C2) there exist $h > 0$ *and* $d > 1 \cdot \frac{2+\delta}{\delta}$ *such that* $\alpha^F_{t,2,\infty}(m) = \mathcal{O}(m^{-d})$*.*

Then

$$|T_n|^{1/2}(\widehat{\xi}_n - \xi^*) I_{22} \widetilde{\Sigma}_{n,22}^{-1/2} \xrightarrow{d} N(0, \mathbf{1}).$$

Next we discuss the properties of the estimator $\widehat{\psi}_n$ based on the spatial projection process X_s. For a Borel set $C \in \mathcal{B}(\mathbb{R}^2)$ denote $\mathcal{F}^{X_s}(C)$ the σ-algebra generated by $X_s \cap C$.

For $h > 0$ let $C_{ij} = [ih, (i+1)h) \times [jh, (j+1)h)$, $(i,j) \in \mathbb{Z}^2$, and

$$\alpha^F_{s,p_1,p_2}(m) = \sup \Bigg\{ \alpha(\mathcal{F}^{X_s}(S_1 \oplus r_1), \mathcal{F}^{X_s}(S_2 \oplus r_1)) : S_1 = \bigcup_{M_1} C_{ij}, S_2 = \bigcup_{M_2} C_{ij}, \\ |M_1| \le p_1, |M_2| \le p_2, d(M_1, M_2) \ge m, M_1, M_2 \subset \mathbb{Z}^2 \Bigg\},$$

where $|M|$ is the cardinality of the set $M \subseteq \mathbb{Z}^2$ and $d(M_1, M_2)$ denotes the minimal distance between M_1 and M_2 in the grid $\mathbb{Z}^2$. Also, $S_i \oplus r_1$ denotes the set S_i dilated by the distance r_1 where r_1 is the upper limit used in the minimum contrast criterion (3.19).

Theorem 3.4. *Apart from the model assumptions formulated above and* (A1)–(A11)*, suppose there exist* $\delta > 0$ *and* $\nu \in \mathbb{N}, 0 < \delta < \nu$*, such that*

(C1) $\lambda_{0,4+2\nu}((u_1,t_1),\ldots,(u_{4+2\nu},t_{4+2\nu})) < \infty$*,*

(D2) there exist $h > 0$ *and* $d > 2 \cdot \frac{2+\delta}{\delta}$ *such that* $\alpha^F_{s,2,\infty}(m) = \mathcal{O}(m^{-d})$*.*

Then

$$|W_n|^{1/2}(\widehat{\psi}_n - \psi^*) I_{33} \widetilde{\Sigma}^{-1/2}_{n,33} \xrightarrow{d} N(0, \mathbf{1}).$$

We remark that the reason for presenting the asymptotic normality results as three separate theorems is that due to the different normalization required for each estimation step we cannot prove joint asymptotic normality for the vector $(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)$ using the current methodology. For details see the discussion at the end of Section 3.2.3. However, this is not a serious issue if the estimates of the individual parameters are not highly correlated (for example, for the models considered in the simulation study in Section 3.1.5 the estimates of the clustering parameters ξ and ψ are nearly independent – details not presented here).

Also note that the matrices $\Sigma_{n,11}$, $\widetilde{\Sigma}_{n,22}$ and $\widetilde{\Sigma}_{n,33}$ can be computed as discussed in Sec. 3.2 and App. B of Waagepetersen and Guan (2009). A plug-in approach can then be used in practice to estimate these matrices (together with $I_{n,11}$, I_{22} and I_{33}) in order to construct confidence regions for the estimates.

Now we formulate sufficient conditions for the mixing properties required in the asymptotic normality theorems above. They are formulated as conditions on the tail behaviour of the smoothing kernel k. These are in general much easier to verify than the mixing conditions themselves.

Lemma 3.5. *Let* X_0 *be a stationary shot-noise Cox process in* $\mathbb{R}^2 \times \mathbb{R}$ *with well-defined first-order moment measure and smoothing kernel* $k(u,t) = k_1(u)k_2(t)$*,* $u \in \mathbb{R}^2, t \in R$*, satisfying*

$$\sup_{(u,t) \in [-\frac{m}{2}, \frac{m}{2}]^3} \left\{ \int_{\mathbb{R}^3 \setminus [-m,m]^3} k_1(v-u)k_2(s-t) \, \mathrm{d}(v,s) \right\} = \mathcal{O}(m^{-d-3}). \tag{3.20}$$

Then X_0 *satisfies condition* (B2)*.*

If k_2 *satisfies*

$$\sup_{s \in [-\frac{m}{2}, \frac{m}{2}]} \left\{ \int_{\mathbb{R} \setminus [-m,m]} k_2(s-\tau) \, \mathrm{d}\tau \right\} = \mathcal{O}(m^{-d-1}), \tag{3.21}$$

then $X_{0,t}$ satisfies condition (C2).

Also, if k_1 satisfies

$$\sup_{u\in[-\frac{m}{2},\frac{m}{2}]^2}\left\{\int_{\mathbb{R}^2\setminus[-m,m]^2} k_1(v-u)\,\mathrm{d}v\right\} = \mathcal{O}(m^{-d-2}), \tag{3.22}$$

then $X_{0,s}$ satisfies condition (D2).

Proof. The proofs of the three parts of the statement follow the same arguments. Hence, we give proof only for the last part concerning $X_{0,s}$.

Recall that r_1 is the upper limit in the minimum contrast criterion (3.19). For a given $h>0$ we define $n = mh - \frac{h}{2} - r_1$ for $m\in\mathbb{N}$ large enough so that $n > \frac{h}{2}+r_1$. We consider the sets $E_1 = C_{0,0}\oplus r_1 - (h/2,h/2)$, $E_2 = \mathbb{R}^2\setminus[-n,n]^2$, $E_3 = [-n/2,n/2]^2$. In this way the sets E_1 and E_2 are disjoint and the values of random variables $Z_{0,0}$ and $Z_{i,j}$ (for (i,j) being at distance at least m from $(0,0)$) depend on points occuring in disjoint sets. Here $Z_{i,j}$ are values of a random field Z, defined on the integer lattice $\mathbb{Z}^2$ and used in the proof of asymptotic normality. For details, see the proof of Theorem 3.11.

Using the cluster representation of X_0 we define two independent space-time processes X_1 and X_2 as follows:

$$X_1 = \bigcup_{(r,v,s)\in\Phi, v\in E_3, s\in\mathbb{R}} X_{v,s}, \quad X_2 = X\setminus X_1,$$

where $X_{v,s}$ denotes the process of daughter points corresponding to the parent point at location $(v,s)\in\mathbb{R}^2\times\mathbb{R}$. Let $X_{1,s}$ and $X_{2,s}$ denote the spatial projection processes corresponding to X_1 and X_2, respectively, i.e.

$$X_{1,s} = \left\{u\in\mathbb{R}^2 : \exists\, t\in T \text{ such that } (u,t)\in X_1\right\},$$
$$X_{2,s} = \left\{u\in\mathbb{R}^2 : \exists\, t\in T \text{ such that } (u,t)\in X_2\right\}.$$

Following the standard arguments like those in Waagepetersen and Guan (2009, App. E) we get

$$\begin{aligned}
\alpha(\mathcal{F}^{X_s}(E_1),\mathcal{F}^{X_s}(E_2)) &\le 5\,\mathbb{E}\,|X_{1,s}\cap E_2| + 5\,\mathbb{E}\,|X_{2,s}\cap E_1| \\
&= 5\,\mathbb{E}\,|X_1\cap(E_2\times T)| + 5\,\mathbb{E}\,|X_2\cap(E_1\times T)| \\
&= 5\mu V_1|T|\left(\int_{E_3}\int_{E_2} k_1(u-v)\,\mathrm{d}u\,\mathrm{d}v + \int_{\mathbb{R}^2\setminus E_3}\int_{E_1} k_1(u-v)\,\mathrm{d}u\,\mathrm{d}v\right) \\
&\le const\cdot\left(n^2\sup_{v\in[-n/2,n/2]^2}\int_{\mathbb{R}^2\setminus[-n,n]^2} k_1(u-v)\,\mathrm{d}u\right. \\
&\quad \left. + (h+2r_1)^2\sup_{v\in E_1}\int_{\mathbb{R}^2\setminus[-n/2,n/2]^2} k_1(u-v)\,\mathrm{d}u\right).
\end{aligned}$$

As long as m is large enough so that $E_1\subset[-n/4,n/4]^2$ we get from (3.22) that both terms on the right-hand side are $\mathcal{O}(m^{-d})$. This implies *(D2)* for $\alpha^F_{s,1,\infty}(m)$.

To check the required property for $\alpha^F_{s,2,\infty}(m)$ we now need to consider the sets $E_1 = (C_{0,0}\cup C_{i,j})\oplus r_1 - (h/2,h/2)$ for some $(i,j)\in\mathbb{Z}^2$, $E_2 = (\mathbb{R}^2\setminus[-n,n]^2)\setminus([-n,n]^2+(ih,jh))$ and $E_3 = [-n/2,n/2]^2\cup([-n/2,n/2]^2+(ih,jh))$. We define

the processes X_1 and X_2 as before, but now using the new definition of E_3. Similarly as before and using stationarity of X_0 we get

$$\alpha(\mathcal{F}^{X_s}(E_1), \mathcal{F}^{X_s}(E_2)) \le const \cdot \Bigg(n^2 \sup_{v \in [-n/2,n/2]^2} \int_{\mathbb{R}^2 \setminus [-n,n]^2} k_1(u-v)\,\mathrm{d}u$$

$$+ (h+2r_1)^2 \sup_{v \in (C_{0,0} \oplus r_1 - (h/2,h/2))} \int_{\mathbb{R}^2 \setminus [-n/2,n/2]^2} k_1(u-v)\,\mathrm{d}u \Bigg).$$

Again, for m large enough so that $(C_{0,0} \oplus r_1 - (h/2, h/2)) \subset [-n/4, n/4]^2$ we get from (3.22) that both terms on the right-hand side are $\mathcal{O}(m^{-d})$. This implies *(D2)* for $\alpha^F_{s,2,\infty}(m)$ and finishes the proof. □

Clearly, the inhomogeneous process X obtained by location-dependent thinning from the stationary shot-noise Cox process X_0 inherits the mixing properties of X_0. This is because the dependence between any two regions cannot increase by the location-dependent thinning which treats every point of the process independently from the others. Hence, the condition (3.20) ensures that *(B2)* is fulfilled also for the inhomogeneous process X, and similarly for the latter conditions and the assumptions *(C2)* and *(D2)*, respectively.

We remark here that in the paper Waagepetersen and Guan (2009) the mixing condition used in their Theorem 1 (for a point process Y defined on $\mathbb{R}^2$) is much stronger than our assumption *(D2)*. Their assumption is formulated for the mixing coefficient $\alpha^Y_{a,\infty}$ of the point process Y rather than for the corresponding random field:

$$\alpha^Y_{p_1,p_2}(m) = \sup \left\{ \alpha(\mathcal{F}^Y(A), \mathcal{F}^Y(B)) : |A| \le p_1, |B| \le p_2, d(A,B) \ge m \right\},$$

where $\mathcal{F}^Y(A)$ denotes the σ-algebra generated by $Y \cap A$, $d(A,B)$ denotes the Hausdorff distance between A and B and the supremum is taken over all Borel subsets A, B of $\mathbb{R}^2$. The assumption used in Waagepetersen and Guan (2009) can be formulated in our notation as follows:

(d2) there is a constant $a > 8r_1^2$ such that $\alpha^Y_{a,\infty}(m) = \mathcal{O}(m^{-d})$ for some $d > 2 \cdot \frac{2+\delta}{\delta}$.

When we take $Y = X_{0,s}$ the condition *(d2)* implies our assumption *(D2)*. However, it is unnecessarily strong and no simple conditions are available for Poisson-Neyman-Scott processes or shot-noise Cox processes which would secure fulfillment of *(d2)*. The condition (3.22) is incorrectly claimed in Waagepetersen and Guan (2009, App. E) to imply *(d2)*. On the other hand, we have seen above that it is sufficient for our assumption *(D2)*.

Recall now the notation used in the proof of Lemma 3.5. The error in Waagepetersen and Guan (2009, App. E) was in considering $E_1 = [-h, h]^2$ for some $h > 0$ and further assuming that any Borel set A with fixed volume a fits into such E_1. However, for $\alpha^Y_{a,\infty}(m)$ to be of order $\mathcal{O}(m^{-d})$ a universal set E_1 would be needed which would cover all Borel sets of volume $\le a$. Obviously, this is not possible as the set A may be arbitrarily "thin" and so for any fixed square E_1 there will always be a set A which is not covered by E_1. Therefore, the tail condition (3.22) can only assure *(D2)* for the mixing coefficient of the random field and not *(d2)* for the mixing coefficient of the point process.

3.1.4 Comparison to a previously published method

In the paper of Møller and Ghorbani (2012) a method for estimation of the parameters of an inhomogeneous space-time shot-noise Cox process has been proposed to analyze a particular data example. However, properties of the procedure have not been investigated further. The method can be easily generalized from the original context (essentially the Poisson cluster process as described here in Example 1.15) to our more general setting of shot-noise Cox processes and used to estimate the parameters of the model considered above. In particular, we adopt the same assumptions on separability of the inhomogeneity function f and of the kernel function k.

The approach of Møller and Ghorbani (2012) is based on specifically reweighted versions of the pair-correlation function. Functions g_1 and g_2 are defined as

$$g_1(u) = \frac{1}{|T|^2}\int_T\int_T g(u, s-t)\,\mathrm{d}s\,\mathrm{d}t, \qquad g_2(t) = \frac{1}{|W|^2}\int_W\int_W g(u-v, t)\,\mathrm{d}u\,\mathrm{d}v, \tag{3.23}$$

where $u \in \mathbb{R}^2$, $t \in \mathbb{R}$, and g is the pair-correlation function (3.4). These are not equivalent to the pair-correlation functions g_s, g_t but they resemble them by having the property

$$\int\int h_1(u,v)g_1(u-v)\,\mathrm{d}u\,\mathrm{d}v = \frac{1}{|T|^2}\,\mathbb{E}\sum_{(u,s),(v,t)\in X:s,t\in T}^{\neq}\frac{h_1(u,v)}{\lambda(u,s)\lambda(v,t)}$$

for any non-negative Borel function h_1 defined on $\mathbb{R}^2 \times \mathbb{R}^2$, and

$$\int\int h_2(s,t)g_2(s-t)\,\mathrm{d}s\,\mathrm{d}t = \frac{1}{|W|^2}\,\mathbb{E}\sum_{(u,s),(v,t)\in X:u,v\in W}^{\neq}\frac{h_2(s,t)}{\lambda(u,s)\lambda(v,t)}$$

for any non-negative Borel function h_2 defined on $\mathbb{R} \times \mathbb{R}$. The crutial difference is in the reweighting by the values of $\lambda(u,t)$ as opposed to $\lambda_s(u)$ or $\lambda_t(t)$, c.f. for example g_s which has a similar property

$$\int\int h_1(u,v)g_s(u-v)\,\mathrm{d}u\,\mathrm{d}v = \frac{1}{|T|^2}\,\mathbb{E}\sum_{u,v\in X_s}^{\neq}\frac{h_1(u,v)}{\lambda_s(u)\lambda_s(v)}$$

and accordingly for g_t and λ_t.

Note that the inhomogeneity function $f(u,t)$ does not appear neither in the formula (3.4) nor in (3.23). This results in somewhat simpler form of g_1, g_2 for our model compared to g_s, g_t which depend on the complicated constants C_s, C_t. On the other hand, this makes the estimation of g_1, g_2 technically more complicated as these cannot be estimated directly from the projection processes X_s, X_t. Values of the space-time intensity function $\lambda(u,t)$ in the points of the process X are needed to estimate g_1, g_2 rather than the intensity functions of the projection processes used to estimate g_s, g_t.

In the same way as in (3.10), functions resembling a K-function can be defined as

$$K_1^{MG}(r) = \int_{\|u\|\le r} g_1(u)\,\mathrm{d}u, \qquad K_2^{MG}(t) = \int_{-t}^{t} g_2(s)\,\mathrm{d}s. \tag{3.24}$$

3.1.5 Simulation study II.

To assess the performance of the estimation procedure introduced in Section 3.1.2 we apply it to realizations of inhomogeneous space-time gamma shot-noise Cox process (see Example 1.17) with parameters μ and θ, observed on the unit cube $W \times T = [0,1]^2 \times [0,1]$. We choose the kernel function $k(u,t)$ to be the product of a spatial Gaussian kernel function k_1 with standard deviation $\sigma > 0$ (probability density function of a zero-mean bivariate radially symmetric normal distribution) and a uniform temporal kernel function k_2 on a bounded interval $[0, t_*]$, i.e.

$$k_2(t) = \frac{1}{t_*} I(0 \leq t \leq t_*),\ t \in \mathbb{R}. \tag{3.25}$$

The uniform temporal kernel was chosen because it has only a single parameter which considerably affects the shape of the pair-correlation function g and the K-function of the temporal projection process X_t. On the other hand, this choice violates the assumption *(A6)* of Theorem 3.1 and hence the asymptotic results from the Section 3.1.3 cannot be used for this model. An alternative choice might be a one-sided triangular kernel on interval $[0, t_*]$. We consider the two-parametric truncated exponential kernel used in Møller and Ghorbani (2012, Sec. 5.3), i.e.

$$k_2(t) = \frac{\alpha}{1 - \exp\{-\alpha t_*\}} \exp\{-\alpha t\} I(0 \leq t \leq t_*),\ t \in \mathbb{R}, \alpha > 0, \tag{3.26}$$

to be unreasonably complicated for datasets which include only a few points in a cluster on average (recall that we observe only a realization of the point process in question and not the underlying driving field Λ itself, i.e. only very limited information about the shape of the kernel is available). While the range parameter t_* influences the shape of the K- or g-function of X_t for all the three kernels in a similar way and can be identified from their non-parametric or semi-parametric estimates (see Figure 3.1 for a detailed example with the uniform kernel), the shape parameter α of the truncated exponential kernel has only very small influence on the shape of the K- or g-function, as illustrated in Figure 3.2.

At first we generate realizations of a homogeneous version of the process (with the intensity $\frac{\mu}{\theta}$) and then apply the location dependent thinning as described in Section 1.1. We use the inhomogeneity function $f((u_1, u_2), t)$ proportional to $\exp\{\beta_1 u_1 + \beta_2 u_2 + \beta_3 t\}$, properly scaled to fulfill the condition $\max_{W \times T} f = 1$. Here u_1 and u_2 are the spatial coordinates and t is the temporal coordinate of the point (event) $((u_1, u_2), t)$. The intensity function of the thinned process is therefore $\frac{\mu}{\theta} f((u_1, u_2), t)$.

The correspondence to the model parametrization stated above is the following: $\beta_s = (\beta_1, \beta_2), \beta_t = \beta_3, \widetilde{\psi} = \sigma, \widetilde{\xi} = t_*, \psi_0 = \frac{C_s}{\mu}$ and $\xi_0 = \frac{C_t}{\mu}$.

In the first estimation step we use the Poisson log-likelihood score function (3.13) to estimate the parameters β_1, β_2 and β_3 from the space-time process X.

The clustering parameters σ and t_* are estimated in the next steps by the minimum contrast estimation. Parameter σ of the spatial kernel k_1, together with ψ_0, is estimated using the pair-correlation function by minimizing the contrast (3.15) with the value r_0 chosen to be the minimal observed interpoint distance in X_s (which is a standard choice in similar situations in the literature) and $r_1 = 4\sigma$. The value of 4σ corresponds to the practical range of interaction of the

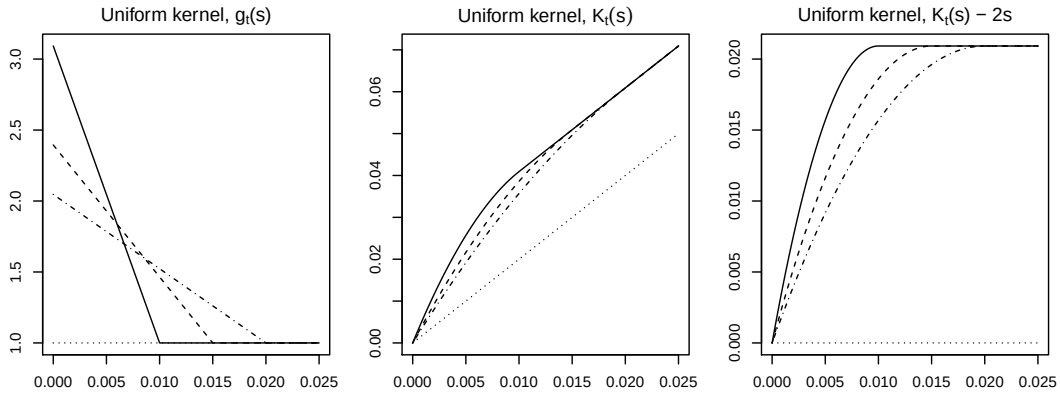


Figure 3.1: Graphs of theoretical characteristics of the gamma shot-noise Cox process considered in Section 3.1.5, i.e. using uniform temporal smoothing kernel (3.25). The corresponding parameter values are $\mu = 50, \beta_1 = 0.5, \beta_2 = -1, \sigma = 0.02$. For details on model parametrization, see text. Note that other model parameters – θ and β_3 – do not affect g_t or K_t in the case of gamma shot-noise Cox process, see Equations (3.6), (3.8) and Section 1.3. The curves correspond to the choice $t_* = 0.010$ (solid line), $t_* = 0.015$ (dashed line), $t_* = 0.20$ (dot-dashed line). Also, the curves corresponding to Poisson process are shown for comparison (dotted line).

considered point processes since for $|u| > 4\sigma$ we get from (3.7) that $g(u) \sim 1$. Using larger values of r_1 would result in no further gain of information, only in larger variability of the estimates caused by higher variability of $\widehat{g}(u)$ for values with $|u| > 4\sigma$.

The variance stabilizing exponent $c_3 = 1/2$ was chosen based on the results of a smaller pilot study in which different versions of this estimator were examined. This choice resulted in lower mean squared errors than the exponent $c_3 = 1/4$.

Minimization of a slightly modified version of the contrast (3.14), i.e.

$$\int_{t_0}^{t_1} ((\widehat{K}_t(t) - 2t)^{1/4} - (K_t(t) - 2t)^{1/4})^2 \, \mathrm{d}t, \tag{3.27}$$

is used to estimate the temporal clustering parameter t_* together with ξ_0. Note that $2t$ is the value of the K-function for the (temporal) Poisson process and thus we are using the contrast for the deviance between the K-function of the clustered model and the benchmark value of $2t$ for the model of complete spatial randomness. According to a smaller pilot study this contrast resulted in better quality estimates compared to using (3.14).

For X_t the range of interaction is equal to t_* – see (3.6). Thus the bounds of the integration interval are chosen to be $t_0 = 0.05t_*$ and $t_1 = 0.95t_*$. This choice is preferable to $t_0 = 0$ and $t_1 = t_*$ since it results in more stable estimates. The results of the pilot study also implied the choice of the variance stabilizing exponent $c_2 = 1/4$, even though the corresponding estimator with $c_2 = 1/2$ performed almost equally well.

At this point it is possible to compute estimates of C_s and C_t as these are functions of the previously estimated parameters. Then, these values can be used together with the estimates of ψ_0 and ξ_0 to estimate $\alpha = \frac{V_2}{\mu(V_1)^2}$ in two different

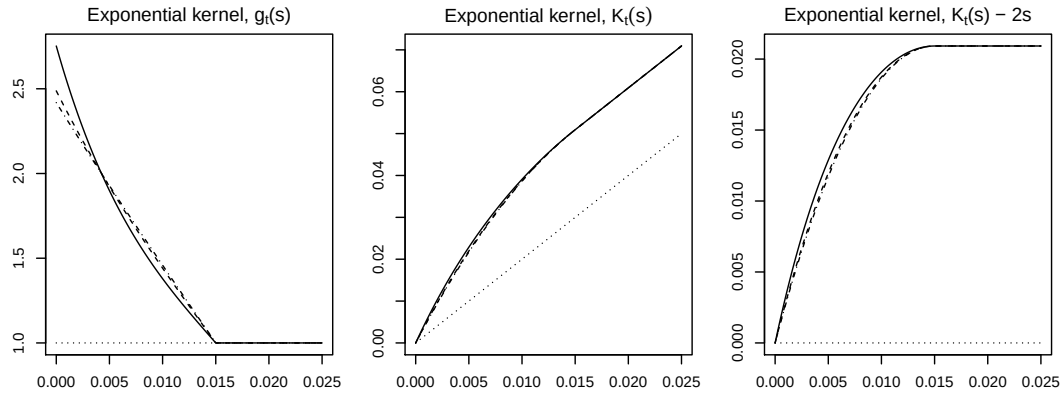


Figure 3.2: Graphs of theoretical characteristics of the gamma shot-noise Cox process using truncated exponential temporal smoothing kernel (3.26). The corresponding parameter values are $\mu = 50, \beta_1 = 0.5, \beta_2 = -1, \sigma = 0.02, t_* = 0.015$. For details on model parametrization, see text. Note that other model parameters – θ and β_3 – do not affect g_t or K_t in the case of gamma shot-noise Cox process, see Equations (3.6), (3.8) and Section 1.3. The curves correspond to the choice $\alpha = 120$ (solid line), $\alpha = 60$ (dashed line), $\alpha = 30$ (dot-dashed line). Also, the curves corresponding to Poisson process are shown for comparison (dotted line).

ways. For the gamma shot-noise Cox process considered in this study α has a particularly simple form: $\alpha = 1/\mu$.

Finally, estimates of the remaining parameters μ and θ are calculated using Equation (3.16) and the estimate of α based on either $\widehat{\psi}_0$ or $\widehat{\xi}_0$. Hence $\widehat{\mu} = 1/\widehat{\alpha}$ and

$$\widehat{\theta} = \frac{\widehat{\mu} \int_W \widehat{f}_1(u)\,\mathrm{d}u \int_T \widehat{f}_2(t)\,\mathrm{d}t}{X(W \times T)}.$$

In the study it turned out that estimates of μ and θ based on the value of $\widehat{\psi}_0$ are more stable in the sense of mean squared error. The estimates of C_s and C_t are rather robust w.r.t. the estimated values of $\widehat{t_*}$ and $\widehat{\sigma}$, respectively, and the factor most influencing the precision of $\widehat{\alpha}$ is the precision of $\widehat{\psi}_0$ and $\widehat{\xi}_0$ themselves. The value of ψ_0 estimated from the spatial projection process X_s using (3.15) proved to be more precise than its temporal counterpart $\widehat{\xi}_0$. Thus, the estimates of μ and θ based on the value of $\widehat{\psi}_0$ are reported in the results of the study rather than those based on $\widehat{\xi}_0$ obtained from (3.27).

To study properties of the estimators under different cluster size distributions we chose the values of μ and θ to be 50, 75 or 100 and 1/20, 1/30 or 1/40, respectively. Different degree of clustering is obtained by taking the values of σ either 0.01 or 0.02 and the values of t_* 0.015 or 0.030. For the inhomogeneity function we use the parameter values $\beta_1 = 0.5$, $\beta_2 = -1$ and $\beta_3 = 0.7$. These choices result in the mean number of points $\mathbb{E}\,X(W \times T)$ ranging from approx. 350 to 1430, depending on the particular combination of parameters.

For each combination of parameters we generated 500 independent realizations from our model and re-estimated the parameters. For simulations we used the algorithm presented in Møller (2003, Sec. 4.1) for which *edge effects* and *truncation* occur and may cause some points to be missing in the simulated (ap-

true values				rel. bias				rel. MSE			
μ	θ	t_*	σ	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\sigma}$	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\sigma}$
50	1/20	0.015	0.01	0.169	0.104	-0.028	0.099	0.097	0.247	0.022	0.089
			0.02	0.181	0.192	-0.033	-0.047	0.116	0.326	0.022	0.010
		0.030	0.01	0.180	0.110	-0.047	0.094	0.103	0.285	0.036	0.080
			0.02	0.155	0.215	-0.062	-0.042	0.100	0.365	0.042	0.011
50	1/40	0.015	0.01	0.150	0.119	-0.032	0.127	0.080	0.265	0.012	0.094
			0.02	0.144	0.170	-0.031	-0.047	0.087	0.275	0.014	0.007
		0.030	0.01	0.131	0.152	-0.062	0.103	0.065	0.281	0.028	0.079
			0.02	0.154	0.159	-0.055	-0.042	0.094	0.281	0.025	0.007
100	1/20	0.015	0.01	0.089	0.107	-0.028	0.016	0.041	0.151	0.019	0.039
			0.02	0.121	0.154	-0.035	-0.041	0.071	0.210	0.018	0.013
		0.030	0.01	0.065	0.089	-0.063	0.023	0.038	0.147	0.038	0.042
			0.02	0.138	0.205	-0.067	-0.052	0.078	0.240	0.038	0.013
100	1/40	0.015	0.01	0.081	0.087	-0.031	0.017	0.032	0.144	0.013	0.032
			0.02	0.089	0.117	-0.024	-0.031	0.057	0.164	0.012	0.010
		0.030	0.01	0.070	0.105	-0.046	0.036	0.029	0.128	0.029	0.040
			0.02	0.124	0.126	-0.054	-0.041	0.060	0.154	0.027	0.010

Table 3.1: Relative biases and relative mean squared errors of the estimates of the inhomogeneous space-time gamma shot-noise Cox process. Estimates obtained using projection processes.

proximating) process $\widetilde{X}$ compared to the (true) process X itself. For details see Møller (2003). However, we have calculated the mean number of missing points in one realization for the models considered in this study. We conclude that the mean number of points missing due to edge effects and truncation is less than 0.052 for all combinations of parameter values. Hence, the error caused by approximating X by $\widetilde{X}$ is negligible when compared to the mean number of points in a single realization.

All the computations were performed in `R`, with the use of auxiliary functions provided in the package `spatstat`, see Baddeley and Turner (2005).

Results

Table 3.1 shows relative mean biases of the estimators and relative mean squared errors (MSEs). By relative we mean divided by the true value of the corresponding parameter or by its square for the MSE. Note that only the results for the extreme combinations of parameters μ and θ are reported since the performance of the estimators showed consistent behaviour within the above described range of parameter values.

The characteristics were obtained from the middle 95 % of the estimates from 500 replications. Neglecting the 5 % of the most extreme estimates was motivated by the fact that in certain situations the estimation methods can be numerically unstable. If such a situation was encountered in practice the estimates would be easily identified as unrealistic and one would alter the parameters of the minimum

contrast procedure or the underlying optimization method. However, due to the extent of the computations involved, this was not possible in this simulation study.

Table 3.1 indicates that all the estimators perform fairly well. It shows that the precision of $\widehat{\mu}$, in terms of the relative mean squared error, improves with the increasing mean number of points of X (implied by higher values of μ/θ) and so does its relative bias in most of the cases.

The estimates of θ depend linearily on $\widehat{\mu}$ but their precision is negatively influenced by their dependence on the actual number of points in $X \cap (W \times T)$. This results in higher variability of the estimates and, accordingly, higher rel. MSE of $\widehat{\theta}$. The positive bias of $\widehat{\mu}$ naturally leads to the positive bias of $\widehat{\theta}$.

Parameter t_* is the only parameter systematically underestimated in the simulation study. Moreover, it is apparent that the parameter t_* is better estimated on the smaller scale, i.e. when the true value of the parameter is $t_* = 0.015$. The obvious explanation is the problem of overlapping of spatially disjoint clusters from the original process X in the projected process X_t. When t_* is too large many disjoint clusters from X appear to be one large cluster in X_t and the efficiency of the estimator decreases. Also, the estimates are slightly less variable for the combinations of parameters with lower θ, i.e. when the realizations are dominated by few separated clusters with high number of points, as opposed to the situations with less variable cluster size distribution corresponding to the higher values of θ. Cluster overlapping plays more important role in the latter case.

The estimates of σ exhibit lower rel. MSE for $\sigma = 0.02$ than for 0.01. Moreover, small positive bias is observed for $\sigma = 0.01$ and small negative bias for $\sigma = 0.02$.

As a conclusion, the clustering parameters σ and t_* are estimated fairly well by the proposed method using projection processes. The parameters μ and θ describing the distribution of the cluster sizes are more difficult to estimate. Nevertheless, they can still be reasonably estimated by the proposed method.

Results – alternative method

For the combinations of parameters reported in detail in Table 3.1 we performed the estimation using the alternative method of Møller and Ghorbani (2012) described in Section 3.1.4. The same realizations of the process X as above were used. Thus, the results are directly comparable.

Based on the results of a pilot study we chose to estimate the spatial clustering parameters by minimizing (3.15) with g_s replaced by K_1^{MG}, $q = 1/4, r_1 = 4\sigma$ and r_0 equal to the minimum observed interpoint distance in X. We use K_1^{MG} for the estimation rather than g_1 because K_1^{MG} was used in the original paper by Møller and Ghorbani (2012). As in the previous section the modified contrast (3.15) is minimized w.r.t. the parameters σ and ψ_0' which is a function of μ, t_*, V_1, V_2 and $|T|$.

Similarly, based on the results of a pilot study, we chose to estimate the temporal clustering parameters by minimizing (3.27) with K_t replaced by K_2 and $t_0 = 0.05t_*, t_1 = 0.95t_*$. This choices resulted in the lowest values of the mean squared error among all considered alternatives. Again, the modified con-

true values				rel. bias				rel. MSE			
μ	θ	t_*	σ	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\sigma}$	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\sigma}$
50	1/20	0.015	0.01	0.078	0.012	-0.019	-0.002	0.066	0.237	0.029	0.007
			0.02	0.122	0.083	-0.023	-0.001	0.134	0.346	0.029	0.012
		0.030	0.01	0.093	0.049	-0.014	0.005	0.067	0.254	0.060	0.006
			0.02	0.109	0.108	-0.031	-0.005	0.120	0.369	0.055	0.013
50	1/40	0.015	0.01	0.085	0.078	-0.024	-0.005	0.055	0.245	0.015	0.002
			0.02	0.096	0.064	-0.024	0.020	0.110	0.308	0.017	0.184
		0.030	0.01	0.090	0.084	-0.043	-0.004	0.053	0.273	0.035	0.003
			0.02	0.111	0.046	-0.023	0.014	0.118	0.311	0.035	0.183
100	1/20	0.015	0.01	0.055	0.023	-0.021	0.159	0.054	0.179	0.025	1.66
			0.02	0.082	0.045	-0.017	0.086	0.100	0.251	0.023	0.957
		0.030	0.01	0.036	-0.001	-0.037	0.114	0.049	0.185	0.056	1.12
			0.02	0.098	0.102	-0.047	0.155	0.119	0.288	0.050	2.03
100	1/40	0.015	0.01	0.054	0.023	-0.022	0.081	0.043	0.184	0.016	0.741
			0.02	0.093	0.048	-0.026	0.036	0.085	0.216	0.016	0.458
		0.030	0.01	0.053	0.039	-0.021	0.019	0.035	0.172	0.038	0.110
			0.02	0.113	0.045	-0.031	0.030	0.087	0.202	0.032	0.499

Table 3.2: Relative biases and relative mean squared errors of the estimates of the inhomogeneous space-time gamma shot-noise Cox process. Estimates obtained using the method of Møller and Ghorbani (2012).

trast (3.27) is minimized w.r.t. the parameters t_* and ξ_0' which is a function of μ, σ, V_1, V_2 and $|W|$.

Now it is possible to estimate $\alpha = \frac{V_2}{\mu(V_1)^2}$ using either $\widehat{\psi}_0'$ or $\widehat{\xi}_0'$ and estimate the remaining parameters μ and θ as in Section 3.1.5. Again, estimates of μ and θ based on $\widehat{\psi}_0'$ were more stable than those based on $\widehat{\xi}_0'$ due to the better precision of $\widehat{\psi}_0'$. Thus, only the characteristics of the estimates based on $\widehat{\psi}_0'$ are reported here.

Table 3.2 summarizes the estimated characteristics of the alternative estimators, again calculated from the middle 95 % of the estimates from 500 replications. The behaviour of $\widehat{\mu}, \widehat{\theta}$ and t_* in this case follows the same trends as described for the corresponding estimators in the simulation study in Section 3.1.5. But there is one striking difference – very strong dependence of the quality of the estimate of σ on the total intensity of the point pattern. Namely for $\mu = 100$ the estimate of σ performs badly, its variability is much larger than for the method using projection processes. For $\mu = 50$ the estimate of σ performes well, especially for the tight clusters (i.e. true $\sigma = 0.01$) $\widehat{\sigma}$ is very efficient.

The case $\mu = 50$ and $\sigma = 0.01$ is the only case when the alternative method performes better than the method using projection processes for the estimation of σ and consequently (recall that ψ_0' is used to obtain $\widehat{\mu}$ and $\widehat{\theta}$) also for the estimation of μ and θ.

Concerning the estimates $\widehat{t_*}$ from Table 3.2, they are also negatively biased. Their bias is somewhat smaller than for the estimates using projection processes.

However, according to the rel. MSE, our method using projection processes provides better estimates of t_* uniformly for all combinations of the true parameter values.

In conclusion we can say that using the reweighted versions of the spatial and temporal pair-correlation functions seem to increase the variability of the estimates of the corresponding K-functions used for the estimation and consequently also the MSE of the parameter estimates. The only exception is the case with smaller number of very tight clusters in X_s (i.e. $\mu = 50$ and $\sigma = 0.01$).

Problem of information loss

By using only the projection processes for estimation the amount of information we get from the data is reduced. We loose efficiency because of the overlapping in the projection process of originally distinct clusters. Specifically, pairs of points which in the space-time could be far apart may get very close in one or the other projection process and we get extra noise when investigating the clustering features of the projection processes.

This problem gets more prominent with higher number of relatively loose clusters and with larger dimension reduction, that is, when using the temporal projection process X_t. We have seen this phenomenon in the estimation of the constant α – using $\widehat{\psi}_0$ resulting from minimum contrast estimation for X_s provided better quality estimates than using $\widehat{\xi}_0$ resulting from minimum contrast estimation for X_t. Also, increasing the intensity results in more pronounced cluster overlapping, both for the space-time process X and for the projection processes.

On the other hand, for models with tight clusters or for models in which the realizations are dominated by a small number of clusters with high number of points the cluster overlapping is a less serious issue. Also, when constructing the spatial projection process X_s we disregard only the temporal coordinate (one-dimensional information) and hence the cluster overlapping is less prominent than for X_t, or may not occur at all.

If the overlapping in the projection processes is too strong the parameter estimates resulting from the suggested procedure will not be good (particularly the estimates of the parameters of the temporal kernel function) and a different method must be used. Such a method is described in the following section.

3.2 Refined method

This section presents unpublished work of the author and his supervisor. The estimator proposed below remedies the main flaw of the method using projection processes, i.e. the projection to the temporal domain is avoided and hence cluster overlapping is not a serious issue anymore. The method is based on estimation of the space-time K-function $K(R, t)$ with a fixed spatial range $R > 0$ as a function of a single argument and using it in a minimum contrast method to estimate the parameters of the temporal part of the model. In this way, only spatially close pairs of points, likely to belong to the same cluster, are counted when estimating $K(R, t)$ non-parametricaly (recall that when estimating the K-function K_t of the temporal projection process X_t all observed pairs of points within specified temporal lag are counted, no matter what their spatial distance is).

With the aim to discuss asymptotic properties of the proposed estimator we recall the basic groundwork and introduce the necessary notation. For ease of exposition we focus on space-time processes in $\mathbb{R}^2 \times \mathbb{R}$, i.e. the most common type of space-time point processes considered in practice. In the following we quickly recall the considered model and notation (most of which was already used in Section 3.1).

Let X_0 be a stationary shot-noise Cox point process on $\mathbb{R}^2 \times \mathbb{R}$ driven by a random field

$$\Lambda(u,t) = \sum_{(v,s,r)\in\Phi} r\, k(v-u, t-s), \qquad (u,t) \in \mathbb{R}^2 \times \mathbb{R},$$

where Φ is a Poisson measure on $\mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}^+$ with intensity measure U and k is a probability kernel, i.e. $\int_{\mathbb{R}^2\times\mathbb{R}} k(w,\tau)\,\mathrm{d}w\,\mathrm{d}\tau = 1$.

To ensure stationarity of X_0 we assume that $U(\mathrm{d}(r,v,s)) = \mu V(\mathrm{d}r)\,\mathrm{d}(v,s)$, where $\mu > 0$ is a constant and $V(\mathrm{d}r)$ is an arbitrary measure on $\mathbb{R}^+$ satisfying the integrability assumption $\int_{\mathbb{R}^+} \min(1,r) V(\mathrm{d}r) < \infty$, cf. also formula (1.7).

We recall the notation $V_1 = \int_{\mathbb{R}^+} r\, V(\mathrm{d}r)$ and $V_2 = \int_{\mathbb{R}^+} r^2\, V(\mathrm{d}r)$. Then the intensity λ_0 and the pair-correlation function $g_0((u,t),(v,s)) = g_0(v-u, s-t)$ of the stationary process X_0 are (see the formulae in Section 1.3)

$$\begin{aligned} \lambda_0 &= \mu\, V_1, \\ g_0(v-u, s-t) &= 1 + \frac{V_2}{\mu\,(V_1)^2} \int_{\mathbb{R}^2\times\mathbb{R}} \int_{\mathbb{R}^2\times\mathbb{R}} k(w-u, \tau-t) k(w-v, \tau-s)\,\mathrm{d}w\,\mathrm{d}\tau. \end{aligned}$$

We further denote $\lambda_{0,k}$ the k-th order intensity function of X_0 and similarly we denote $\lambda_{0,s,k}$ the k-th order intensity function of the spatial projection $X_{0,s}$ of X_0. Throughout this section we assume that $\lambda_{0,2}$ exists and is bounded so that the pair-correlation function of X_0 is properly defined.

Now let X be the inhomogeneous space-time shot-noise Cox point process obtained from X_0 by location-dependent thinning using an inhomogeneity function $f(u,t)$, scaled such that

$$\max_{(u,t)\in\mathbb{R}^2\times\mathbb{R}} f(u,t) = 1. \tag{3.28}$$

Hence, X has the SOIRS property (see Definition 1.9) and the intensity function λ and the pair-correlation function g of X are

$$\begin{aligned} \lambda(u,t) &= \mu\, V_1\, f(u,t), \\ g(v-u, s-t) &= 1 + \frac{V_2}{\mu\,(V_1)^2} \int_{\mathbb{R}^2\times\mathbb{R}} k(w-u, \tau-t) k(w-v, \tau-s)\,\mathrm{d}w\,\mathrm{d}\tau. \end{aligned}$$

As before, we denote by $\lambda_{s,k}$ the k-th order intensity function of the spatial projection X_s of X.

Space-time separability of the inhomogeneity function is required so that the moment characteristics of the projection process X_s have a tractable form, see Section 3.1. The assumption takes on the form

$$f(u,t) = f_1(u) f_2(t),\ u \in \mathbb{R}^2,\ t \in \mathbb{R}. \tag{3.29}$$

In connection with Equation (3.28) this means that

$$\max_{u\in\mathbb{R}^2} f_1(u) = 1, \qquad \max_{t\in\mathbb{R}} f_2(t) = 1.$$

Moreover, we require the space-time separability of the smoothing kernel, i.e.

$$k(w,\tau) = k_1(w)k_2(\tau)\ u \in \mathbb{R}^2, t \in \mathbb{R}, \tag{3.30}$$

which ensures the special structure of the second-order moment characteristics of the process X. In the following we also take advantage of the notation

$$K_1(v-u) = \int_{\mathbb{R}^2} k_1(u-w)k_1(v-w)\,\mathrm{d}w,$$
$$K_2(s-t) = \int_{\mathbb{R}} k_2(t-\tau)k_2(s-\tau)\,\mathrm{d}\tau.$$

The pair-correlation function g of X takes on the form

$$g(v-u, s-t) = 1 + \frac{V_2}{\mu\,(V_1)^2} K_1(v-u)\,K_2(s-t).$$

and the space-time K-function with the fixed spatial range $R > 0$ is

$$K(R,t) = 2\pi R^2 t + \frac{V_2}{\mu\,(V_1)^2} \int_{\|u\|<R} K_1(u)\,\mathrm{d}u \int_{|s|<t} K_2(s)\,\mathrm{d}s.$$

The K-function of the spatial projection process X_s can be written as

$$K_s(r) = \pi r^2 + C_s \frac{V_2}{\mu\,(V_1)^2} \int_{\|u\|<r} K_1(u)\,\mathrm{d}u,$$

where C_s is defined in (3.9).

3.2.1 Model parametrization

Suppose that the underlying Poisson measure Φ has the intensity measure of the form $\mu V(\mathrm{d}r)\,\mathrm{d}(v,s)$, where $\mu > 0$ is a scalar parameter and the measure V is parametrized by the (scalar) parameter θ.

We assume a particular form of the inhomogeneity function f, see also the assumption (3.29):

$$f(u,t;\beta_s,\beta_t) = f_1(z_1(u)\beta_s^T)\,f_2(z_2(t)\beta_t^T), \tag{3.31}$$

where $z_1(u)$ and $z_2(t)$ are vectors of spatial and temporal covariates, respectively, and, with a slight abuse of notation, f_1, f_2 are positive, strictly increasing functions. The vectors β_s, β_t denote the unknown inhomogeneity parameters.

We further set $\beta_0 = \log(\mu\,V_1)$. Then the intensity function λ of X is parametrized by the vector $\beta = (\beta_0, \beta_s, \beta_t)$.

Let the spatial smoothing kernel k_1 (and hence K_1) be parametrized by a vector of parameters $\widetilde{\psi}$ and similarly the temporal smoothing kernel k_2 (and hence K_2) be parametrized by the vector $\widetilde{\xi}$. Now we set

$$\psi_0 = C_s \frac{V_2}{\mu\,(V_1)^2},$$
$$\xi_0 = \frac{V_2}{\mu\,(V_1)^2} \int_{\|u\|<R} K_1(u;\widetilde{\psi})\,\mathrm{d}u,$$

and $\psi = (\psi_0, \widetilde{\psi})$ and similarly $\xi = (\xi_0, \widetilde{\xi})$. Henceforth

$$K_s(r;\psi) = \pi r^2 + \psi_0 \int_{\|u\|<r} K_1(u;\widetilde{\psi})\,\mathrm{d}u,$$
$$K(R,t;\xi) = 2\pi R^2 t + \xi_0 \int_{|s|<t} K_2(s;\widetilde{\xi})\,\mathrm{d}s.$$

In the following we prefer writing simply $K(t;\xi)$ instead of $K(R,t;\xi)$ as this can cause no confusion to the reader.

Below we also use the following notation: $\lambda^{(1)}$ and $\lambda^{(2)}$ are the first and second-order derivatives of the intensity function of X w.r.t. the parameter β; $K^{(1)}(t;\xi), K^{(2)}(t;\xi)$ are the first and second-order derivatives of $K(t;\xi)$ w.r.t. ξ; $K_s^{(1)}(r;\psi), K_s^{(2)}(r;\psi)$ are the first and second-order derivatives of $K_s(r;\psi)$ w.r.t. ψ (assuming that the appropriate derivatives exist).

Finally, we denote by $\beta^*, \xi^*, \psi^*, \mu^*$ and θ^* the "true" parameter values governing the distribution of the process X.

3.2.2 Step-wise estimation procedure

Assume that the data available is observed in a compact space-time observation window of the form $W \times T$.

The parameter estimates $\widehat{\beta}, \widehat{\xi}$ and $\widehat{\psi}$ are obtained by solving the estimating equation

$$U(\beta,\xi,\psi) = (U_1(\beta), U_2(\beta,\xi), U_3(\beta,\psi)) = 0,$$

where the estimating functions U_1, U_2 and U_3 are defined in the following paragraphs.

The parameters μ and θ of the underlying Poisson measure are estimated from the previously estimated values of $\widehat{\beta}_0, \widehat{\xi}_0$ and $\widehat{\psi}_0$ and the total number of points in the observation window, see Section 3.1.2.

First step We estimate the parameter β of the intensity function λ of X in the same way as in Section 3.1, i.e. by means of the Poisson likelihood score estimating function (Waagepetersen and Guan, 2009). The estimate $\widehat{\beta}$ is thus obtained by solving the estimating equation

$$U_1(\beta) = \sum_{(u,t)\in X\cap(W\times T)} \frac{\lambda^{(1)}(u,t;\beta)}{\lambda(u,t;\beta)} - \int_{W\times T} \lambda^{(1)}(v,s;\beta)\,\mathrm{d}v\,\mathrm{d}s = 0. \tag{3.32}$$

We remark here that the estimating equation is the same as in the method using projection processes, see Equation (3.17).

Note that in order to obtain the formula above it is necessary to change the order of integration and differentiation. However, here and in what follows it is always possible under the assumptions of the theorems below (in the second and third step and also in the proofs below). We omit this remark hereafter.

We also remark here that as far as estimation of β_s and β_t is concerned, it does not matter whether they are estimated together using (3.32) or separately using estimating functions corresponding to the Poisson likelihood for the projected

processes X_s and X_t. The respective estimating functions are identical. However, estimation of β_0 is slightly more complicated when using the projection processes. The following equation can be utilized in that case:

$$\lambda(u, \tau; \beta) = \frac{\lambda_s(u; \beta)\, \lambda_t(\tau; \beta)}{\int_{W \times T} \lambda(u', \tau'; \beta)\, \mathrm{d}v'\, \mathrm{d}\tau'},$$

where the denominator is estimated by $X(W \times T)$. This, together with the estimates $\widehat{\beta}_s$ and $\widehat{\beta}_t$ obtained from the projected processes, allows estimation of β_0.

Second step Conditionally on $\beta = \widehat{\beta}$, we estimate the parameter ξ of K, the K-function of the space-time process X, by minimizing the discrepancy

$$m_{2,\beta}(\xi) = \int_{t_0}^{t_1} \left[\widehat{K}(t; \beta)^{c_2} - K(t; \xi)^{c_2}\right]^2 \mathrm{d}t, \tag{3.33}$$

where $0 \leq t_0 < t_1$ and c_2 are given constants and $\widehat{K}(\cdot; \beta)$ is the empirical estimate of K (unbiased due to the edge-correction factor):

$$\widehat{K}(t; \beta) = \sum_{(u,s),(v,\tau) \in X \cap (W \times T)}^{\neq} \frac{I\{\|u - v\| < R\}\, I\{|s - \tau| < t\}}{\lambda(u, s; \beta)\lambda(v, \tau; \beta)|W \cap W_{u-v}|\, |T \cap T_{s-\tau}|},$$

where W_{u-v} denotes the set W shifted by the vector $u - v$ and similarly for $T_{s-\tau}$. We choose here the translation edge-correction because it is convenient for the discussion on the asymptotic properties below. Note that also other edge-correction factors can be used (Gabriel, 2014).

The estimate $\widehat{\xi}$ is thus obtained by solving the estimating equation $U_2(\widehat{\beta}, \xi) = 0$, where

$$\begin{aligned} U_2(\beta, \xi) =& -|W \times T| \frac{\partial m_{2,\beta}(\xi)}{\partial \xi} \\ =& 2c_2|W \times T| \int_{t_0}^{t_1} \left[\widehat{K}(t; \beta)^{c_2} - K(t; \xi)^{c_2}\right] K(t; \xi)^{c_2 - 1} K^{(1)}(t; \xi)\, \mathrm{d}t, \end{aligned}$$

assuming that K is a differentiable function of ξ.

Third step Conditionally on $\beta = \widehat{\beta}$, we estimate the parameter ψ of K_s, the K-function of the spatial projection process, by minimizing the discrepancy

$$m_{3,\beta}(\psi) = \int_{r_0}^{r_1} \left[\widehat{K}_s(r; \beta)^{c_3} - K_s(r; \psi)^{c_3}\right]^2 \mathrm{d}r, \tag{3.34}$$

where $0 \leq r_0 < r_1$ and c_3 are given constants and $\widehat{K}_s(\cdot; \beta)$ is the empirical estimate of K_s (unbiased due to the translation edge-correction factor):

$$\widehat{K}_s(r; \beta) = \sum_{u,v \in X_s \cap W}^{\neq} \frac{I\{\|u - v\| < r\}}{\lambda_s(u; \beta)\lambda_s(v; \beta)|W \cap W_{u-v}|}.$$

The estimate $\widehat{\psi}$ is thus obtained by solving the estimating equation $U_3(\widehat{\beta}, \psi) = 0$, where

$$\begin{aligned} U_3(\beta, \psi) =& - |W| \frac{\partial m_{3,\beta}(\psi)}{\partial \psi} \\ =& 2c_3 |W| \int_{r_0}^{r_1} \left[\widehat{K}_s(r;\beta)^{c_3} - K_s(r;\psi)^{c_3} \right] K_s(r;\psi)^{c_3 - 1} K_s^{(1)}(r;\psi) \, \mathrm{d}r, \end{aligned}$$

assuming that K_s is a differentiable function of ψ.

3.2.3 Asymptotic properties

The approach used in the following is inspired by the paper Waagepetersen and Guan (2009).

When discussing asymptotic properties of the estimators proposed above, we consider in the following the increasing window asymptotics. Namely, we consider an increasing sequence of compact observation windows $W_n \times T_n, n \geq 1$, such that $W_n \times T_n \nearrow \mathbb{R}^2 \times \mathbb{R}$. A particular example might be

$$W_n = [an, bn] \times [cn, dn], \qquad T_n = [en, fn],$$

where $a < 0 < b, c < 0 < d$ and $e < 0 < f$ are fixed constants. The important property of the sequence of observation windows necessary in the following is that, for any $h \in \mathbb{R}^2$ and $k \in \mathbb{R}$,

$$\lim_{n \to \infty} \frac{|W_n \times T_n|}{|(W_n \cap W_{n,h}) \times (T_n \cap T_{n,k})|} = 1,$$

where $W_{n,h}$ denotes the set W_n shifted by the vector $h \in \mathbb{R}^2$, and similarly for $T_{n,k}$.

Let $(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)$ be the estimated parameter values calculated from $W_n \times T_n$, i.e. the solution of the equation $U_n(\beta, \xi, \psi) = (U_{n,1}(\beta), U_{n,2}(\beta, \xi), U_{n,3}(\beta, \psi)) = 0$, where

$$\begin{aligned} U_{n,1}(\beta) =& \sum_{(u,t) \in X \cap (W_n \times T_n)} \frac{\lambda^{(1)}(u,t;\beta)}{\lambda(u,t;\beta)} - \int_{W_n \times T_n} \lambda^{(1)}(v,s;\beta) \, \mathrm{d}v \, \mathrm{d}s, \\ U_{n,2}(\beta, \xi) =& 2c_2 |W_n \times T_n| \int_{t_0}^{t_1} \left[\widehat{K}_n(t;\beta)^{c_2} - K(t;\xi)^{c_2} \right] K(t;\xi)^{c_2 - 1} K^{(1)}(t;\xi) \, \mathrm{d}t, \\ U_{n,3}(\beta, \psi) =& 2c_3 |W_n| \int_{r_0}^{r_1} \left[\widehat{K}_{s,n}(r;\beta)^{c_3} - K_s(r;\psi)^{c_3} \right] K_s(r;\psi)^{c_3 - 1} K_s^{(1)}(r;\psi) \, \mathrm{d}r, \end{aligned}$$

where $\widehat{K}_n$ and $\widehat{K}_{s,n}$ are the semi-parametric estimates of K and K_s, calculated using $X \cap (W_n \times T_n)$ and $X_s \cap W_n$, respectively.

Note that we use the same spatial projection process X_s (projected from the fixed time interval T) to define $U_{n,3}$ for all $n \geq 1$. If we used T_n to define projection processes $X_s^{(n)}$, the resulting asymptotic regime for $U_{n,3}$ would be a combination of the increasing window asymptotics and the infill asymptotics, i.e. the intensity function of $X_s^{(n)}$ (at any location) would be increasing, unbounded function of n.

Also, the constant $C_s^{(n)} \to 0$ as $T_n \nearrow \mathbb{R}$ (at least for f_2 bounded away from 0, which will be assured in the following by the assumption of bounded covariates),

see Equation (3.9). Hence the pair-correlation function of $X_s^{(n)}$ is constant 1 in the limit and the influence of the parameter ψ is lost – $X_s^{(n)}$ cannot be distinguished (in the limit) from the Poisson process using second-order moment properties only. As a result, if we want to use the spatial projection process for estimation of ψ we need to fix the time interval T used for the projection.

Different normalization of $U_{n,2}$ and $U_{n,3}$ makes the formulas in the following somewhat complicated but it is needed to keep variances of the respective terms controlled.

Following Waagepetersen and Guan (2009) we approximate $U_{n,2}(\beta^*, \xi^*)$ and $U_{n,3}(\beta^*, \psi^*)$ by

$$\widetilde{U}_{n,2}(\beta^*, \xi^*) = 2c_2^2 |W_n \times T_n| \int_{t_0}^{t_1} \left[\widehat{K}_n(t;\beta^*) - K(t;\xi^*)\right] K(t;\xi^*)^{2c_2-2} K^{(1)}(t;\xi^*)\,\mathrm{d}t,$$

$$\widetilde{U}_{n,3}(\beta^*, \psi^*) = 2c_3^2 |W_n| \int_{r_0}^{r_1} \left[\widehat{K}_{s,n}(r;\beta^*) - K_s(r;\psi^*)\right] K_s(r;\psi^*)^{2c_3-2} K_s^{(1)}(r;\psi^*)\,\mathrm{d}r.$$

For $\widetilde{U}_{n,2}$ this is simply a Taylor series expansion of $\left[\widehat{K}_n(t;\beta)^{c_2} - K(t;\xi)^{c_2}\right]$. To see this, consider the Taylor expansion of the function $h(x) = x^{c_2}$, i.e.

$$h(a) - h(b) = h^{(1)}(u)(a-b),$$

where $h^{(1)}$ is the derivative of h and u lies between a and b. Now we rearrange the terms as follows:

$$h(a) - h(b) = h^{(1)}(b)(a-b) + (a-b)\left(h^{(1)}(u) - h^{(1)}(b)\right),$$

and approximate $h(a) - h(b) \approx h^{(1)}(b)(a-b)$. Finally, we take $a = \widehat{K}_n(t;\beta^*)$ and $b = K(t;\xi^*)$. We will show later that

$$|W_n \times T_n|^{-1/2} \left(U_{n,2}(\beta^*, \xi^*) - \widetilde{U}_{n,2}(\beta^*, \xi^*)\right) \to 0$$

in probability as $n \to \infty$ and hence the error of approximation is negligible in the respective calculations. The situation is similar for $\widetilde{U}_{n,3}$ (details omitted).

From the mathematical point of view $\widetilde{U}_{n,2}(\beta^*, \xi^*)$ is easier to handle than $U_{n,2}(\beta^*, \xi^*)$ since we avoid the exponent c_2 for $\widehat{K}_n(t;\beta^*)$ (Waagepetersen and Guan, 2009), and similarly for $\widetilde{U}_{n,3}(\beta^*, \psi^*)$.

Before we formulate the first theorem we further define

$$\widetilde{\Sigma}_n = |W_n \times T_n|^{-1} \,\mathsf{Var}\left(U_{n,1}(\beta^*), \widetilde{U}_{n,2}(\beta^*, \xi^*)\right) = \begin{pmatrix} \widetilde{\Sigma}_{n,11} & \widetilde{\Sigma}_{n,12} \\ \widetilde{\Sigma}_{n,12}^T & \widetilde{\Sigma}_{n,22} \end{pmatrix},$$

$$\widetilde{\Sigma}_{n,33} = |W_n|^{-1} \,\mathsf{Var}\left(\widetilde{U}_{n,3}(\beta^*, \psi^*)\right),$$

$$\begin{aligned} J_n(\beta,\xi,\psi) &= -\frac{\partial}{\partial(\beta,\xi,\psi)^T}\, U_n(\beta,\xi,\psi) \\ &= -\begin{pmatrix} \frac{\partial}{\partial\beta^T}\, U_{n,1}(\beta) & \frac{\partial}{\partial\beta^T}\, U_{n,2}(\beta,\xi) & \frac{\partial}{\partial\beta^T}\, U_{n,3}(\beta,\psi) \\ 0 & \frac{\partial}{\partial\xi^T}\, U_{n,2}(\beta,\xi) & 0 \\ 0 & 0 & \frac{\partial}{\partial\psi^T}\, U_{n,3}(\beta,\psi) \end{pmatrix} \\ &= \begin{pmatrix} J_{n,11}(\beta) & J_{n,12}(\beta,\xi) & J_{n,13}(\beta,\psi) \\ 0 & J_{n,22}(\beta,\xi) & 0 \\ 0 & 0 & J_{n,33}(\beta,\psi) \end{pmatrix} \end{aligned}$$

and

$$
\begin{aligned}
I_n &= \begin{pmatrix} I_{n,11} & I_{n,12} \\ 0 & I_{22} \end{pmatrix}, \\
I_{n,11} &= \frac{1}{|W_n \times T_n|} \int_{W_n \times T_n} \frac{\lambda^{(1)}(v,s;\beta^*)^T \lambda^{(1)}(v,s;\beta^*)}{\lambda(v,s;\beta^*)} \,\mathrm{d}v \,\mathrm{d}s, \\
I_{n,12} &= -2c_2^2 \int_{t_0}^{t_1} H_{n,2}(t;\beta^*) K(t;\xi^*)^{2c_2-2} K^{(1)}(t;\xi^*) \,\mathrm{d}t, \\
I_{n,13} &= -2c_3^2 \int_{r_0}^{r_1} H_{n,3}(r;\beta^*) K_s(r;\psi^*)^{2c_3-2} K_s^{(1)}(r;\psi^*) \,\mathrm{d}r, \\
I_{22} &= 2c_2^2 \int_{t_0}^{t_1} K(t;\xi^*)^{2c_2-2} K^{(1)}(t;\xi^*)^T K^{(1)}(t;\xi^*) \,\mathrm{d}t, \\
I_{33} &= 2c_3^2 \int_{r_0}^{r_1} K_s(r;\psi^*)^{2c_3-2} K_s^{(1)}(r;\psi^*)^T K_s^{(1)}(r;\psi^*) \,\mathrm{d}r,
\end{aligned}
$$

where

$$
\begin{aligned}
H_{n,2}(t;\beta^*) =& \,\mathbb{E} \frac{\partial}{\partial \beta^T} \widehat{K}_n(t;\beta)|_{\beta=\beta^*} \\
=& -2 \int_{W_n \times T_n} \int_{W_n \times T_n} \frac{I\{\|u-v\| < R\} I\{|s-\tau| < t\}}{|W_n \cap W_{n,u-v}| \cdot |T_n \cap T_{n,s-\tau}|} \cdot \\
& \cdot \frac{\lambda^{(1)}(u,s;\beta^*)}{\lambda(u,s;\beta^*)} g(u-v, s-\tau; \xi^*, \psi^*) \,\mathrm{d}u \,\mathrm{d}s \,\mathrm{d}v \,\mathrm{d}\tau, \\
H_{n,3}(r;\beta^*) =& \,\mathbb{E} \frac{\partial}{\partial \beta^T} \widehat{K}_{s,n}(r;\beta)|_{\beta=\beta^*} \\
=& -2 \int_{W_n} \int_{W_n} \frac{I\{\|u-v\| < r\}}{|W_n \cap W_{n,u-v}|} \frac{\lambda_s^{(1)}(u;\beta^*)}{\lambda_s(u;\beta^*)} g_s(u-v;\psi^*) \,\mathrm{d}u \,\mathrm{d}v.
\end{aligned}
$$

Consistency theorem

We can now formulate the consistency theorem, inspired by Waagepetersen and Guan (2009). Note that the numbering of the respective assumptions follows the numbering used in Section 3.1.3. Also, *(A6')* denotes an altered version of the assumption *(A6)* from Section 3.1.3 and the assumption *(A12)* replaces *(A10)* from Section 3.1.3.

Theorem 3.6. *Apart from the model assumptions formulated above, let the following conditions be met:*

(A1) the inhomogeneity function f is twice continuously differentiable as a function of β,

(A2) $\exists C_1 < \infty$ such that $\|z_1(u)\| < C_1, \|z_2(t)\| < C_1, u \in \mathbb{R}^2, t \in \mathbb{R}$,

(A3) I_{22} and I_{33} are positive definite matrices and $\liminf \omega_{n,11} > 0$, where $\omega_{n,11}$ is the smallest eigenvalue of $I_{n,11}$,

(A4) $\widetilde{\Sigma}_{n,22}$ and $\widetilde{\Sigma}_{n,33}$ converge to positive definite matrices $\widetilde{\Sigma}_{22}$ and $\widetilde{\Sigma}_{33}$, respectively,

(A5) $\lambda^{(2)}(u,t;\beta), \frac{\lambda^{(1)}(u,t;\beta)}{\lambda(u,t;\beta)}, \frac{\lambda^{(2)}(u,t;\beta)}{\lambda(u,t;\beta)}$ *are bounded, continuous functions of* (u,t,β),

(A6') $K(t;\xi), K^{(1)}(t;\xi), K^{(2)}(t;\xi)$ *exist and are continuous functions of* (t,ξ),

(A7) $K_s(r;\psi), K_s^{(1)}(r;\psi), K_s^{(2)}(r;\psi)$ *exist and are continuous functions of* (r,ψ),

(A8) $t_0 \geq 0$ *for* $c_2 \geq 2$, *otherwise* $t_0 > 0$; *similarly,* $r_0 \geq 0$ *for* $c_3 \geq 2$, *otherwise* $r_0 > 0$,

(A9) $\lambda_{0,2}$ *and* $\lambda_{0,3}$ *exist and are bounded and the second-order reduced factorial cumulant measure of* X_0 *has finite total variation,*

(A11) $\exists C_3 < \infty$ *such that for all* $u_1, u_2 \in \mathbb{R}^2$:

$$\int_{\mathbb{R}^2} |\lambda_{0,s,4}(0,u_1,v,u_2+v) - \lambda_{0,s,2}(0,u_1)\,\lambda_{0,s,2}(0,u_2)|\,\mathrm{d}v < C_3,$$

(A12) $\exists C_4 < \infty$ *such that for all* $(u_1,t_1),(u_2,t_2) \in \mathbb{R}^2 \times \mathbb{R}$:

$$\int_{\mathbb{R}^2\times\mathbb{R}} |\lambda_{0,4}((0,0),(u_1,t_1),(v,s),(u_2+v,t_2+s)) - \lambda_{0,2}((0,0),(u_1,t_1))\,\lambda_{0,2}((0,0),(u_2,t_2))|\,\mathrm{d}(v,s) < C_4.$$

Then there is a sequence $\left\{(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)\right\}_{n\geq 1}$ *for which*

$$U_n(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n) = 0$$

with probability tending to 1 and the vector

$$L_n = \begin{pmatrix} |W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*) \\ |W_n \times T_n|^{1/2}(\widehat{\xi}_n - \xi^*) \\ |W_n|^{1/2}(\widehat{\psi}_n - \psi^*) \end{pmatrix}$$

is bounded in probability, i.e. $\forall \varepsilon > 0\ \exists \delta > 0 : \mathbb{P}(\|L_n\| > \delta) \leq \epsilon$ *for* n *sufficiently large.*

The theorem establishes existence of a consistent sequence $\{(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)\}_{n\geq 1}$ such that $(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)$ corresponds to a root of the estimating function $U_n(\beta,\xi,\psi)$ with probability tending to 1 (this also means that the estimating function actually *has* at least one root with probability tending to 1). If U_n has precisely one root for each n, we can find the consistent sequence $\{(\widehat{\beta}_n, \widehat{\xi}_n, \widehat{\psi}_n)\}_{n\geq 1}$ simply by finding the roots.

For a discussion on the assumtions of the theorem see Section 3.1.3.

The proof of the theorem is based on the general asymptotic result presented in Waagepetersen and Guan (2009, App. C) and attributed to unpublished lecture notes by Professor Jens L. Jensen, Aarhus University, Denmark. We give here the statement of the result in the following lemma. For the proof see Waagepetersen and Guan (2009, App. C).

Consider a parametrized family of probability measures $\mathbb{P}_\theta, \theta \in \mathbb{R}^p$, and a sequence of estimating functions $u_n : \mathbb{R}^p \to \mathbb{R}^p, n \geq 1$. The distribution of $\{u_n(\theta)\}_{n\geq 1}$ is governed by $\mathbb{P} = \mathbb{P}_{\theta^*}$, where θ^* denotes the "true" parameter value. For a matrix $A = (a_{ij})$, $\|A\|_M = \max_{ij} |a_{ij}|$ and we let $J_n(\theta) = -\partial u_n(\theta)/\partial\theta^T$.

Lemma 3.7. *Assume that there is a sequence of invertible symmetric matrices V_n such that*

(a) $\|V_n^{-1}\|_M \to 0$,

(b) there exists $l > 0$ *such that* $\mathbb{P}(l_n < l) \to 0$ *where*

$$l_n = \inf_{\|\phi\|=1} \left\{ \phi V_n^{-1} J_n(\theta^*) V_n^{-1} \phi^T \right\},$$

(c) for any $d > 0$,

$$\sup_{\|(\theta-\theta^*)V_n\| \le d} \left[\|V_n^{-1} \left\{ J_n(\theta) - J_n(\theta^*) \right\} V_n^{-1}\|_M \right] = \gamma_{nd} \to 0$$

in probability under $\mathbb{P} = \mathbb{P}_{\theta^*}$,

(d) the sequence $u_n(\theta^*)V_n^{-1}$ *is bounded in probability, i.e. for each* $\varepsilon > 0$ *there exists* d *such that* $\mathbb{P}\left(\|u_n(\theta^*)V_n^{-1}\| > d\right) \le \varepsilon$ *for* n *sufficiently large.*

Then for each $\varepsilon > 0$ *there exists* $d > 0$ *such that*

$$\mathbb{P}\left(\exists \tilde{\theta}_n : u_n(\tilde{\theta}_n) = 0 \text{ and } \|(\tilde{\theta}_n - \theta^*)V_n\| < d \right) > 1 - \varepsilon$$

whenever n *is sufficiently large.*

Proof of Theorem 3.6

To prove Theorem 3.6 we apply Lemma 3.7 sequentially as follows. First, we apply it to $u_n = U_{n,1}$ with $V_n = |W_n \times T_n|^{1/2} \cdot \mathbf{1}$, where $\mathbf{1}$ is the identity matrix of appropriate dimension. It follows that there is a sequence $\{\widehat{\beta}_n\}_{n\ge1}$ such that $|W_n \times T_n|^{1/2}\|\widehat{\beta}_n - \beta^*\|$ is bounded in probability and $U_{n,1}(\widehat{\beta}_n) = 0$ with probability tending to 1. This also implies $\widehat{\beta}_n \to 0$ in probability as $n \to \infty$.

Next, we use Lemma 3.7 on $u_n(\cdot) = U_{n,2}(\widehat{\beta}_n, \cdot)$ with $V_n = (|W_n \times T_n|\widetilde{\Sigma}_{n,22})^{1/2}$ to show that there is a sequence $\{\widehat{\xi}_n\}_{n\ge1}$ such that $U_{n,2}(\widehat{\beta}_n, \widehat{\xi}_n) = 0$ with probability tending to 1 and $|W_n \times T_n|^{1/2}\|\widehat{\xi}_n - \xi^*\|$ is bounded in probability (recall that by assumption *(A4)* $\widetilde{\Sigma}_{n,22}^{1/2}$ converges to a positive definite matrix $\widetilde{\Sigma}_{22}^{1/2}$ and hence it does not influence the boundedness in probability of $|W_n \times T_n|^{1/2}\|\widehat{\xi}_n - \xi^*\|\widetilde{\Sigma}_{n,22}^{1/2})$. The difficult part in using this lemma is to show boundedness in probability of $|W_n \times T_n|^{-1/2}U_{n,2}(\widehat{\beta}_n, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2}$ in condition (d). To do this we use a Taylor series expansion

$$\begin{aligned} |W_n \times T_n|^{-1/2}U_{n,2}(\widehat{\beta}_n, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2} =& |W_n \times T_n|^{-1/2}U_{n,2}(\beta^*, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2} \\ &- |W_n \times T_n|^{-1/2}(\widehat{\beta}_n - \beta^*)J_{n,12}(\widetilde{\beta}, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2}, \end{aligned} \tag{3.35}$$

where $\|\widetilde{\beta} - \beta^*\| \le \|\widehat{\beta}_n - \beta^*\|$. In this way we can show boundedness in probability of the two terms on the right-hand side which are easier to handle than $|W_n \times T_n|^{-1/2}U_{n,2}(\widehat{\beta}_n, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2}$. In the first term on the right-hand side only the true parameter values appear and in the second term, the behaviour of $|W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*)$ is controlled thanks to the result obtained above.

Finally, we use Lemma 3.7 in a similar way on $u_n(\cdot) = U_{n,3}(\widehat{\beta}_n, \cdot)$ with $V_n = (|W_n|\widetilde{\Sigma}_{n,33})^{1/2}$ to show that there is a sequence $\{\widehat{\psi}_n\}_{n\geq 1}$ such that $|W_n|^{1/2}\|\widehat{\psi}_n - \psi^*\|$ is bounded in probability and $U_{n,3}(\widehat{\beta}_n, \widehat{\psi}_n) = 0$ with probability tending to 1. To verify the condition (d) of the lemma we use a Taylor series expansion

$$\begin{aligned}|W_n|^{-1/2}U_{n,3}(\widehat{\beta}_n, \psi^*)\widetilde{\Sigma}_{n,33}^{-1/2} =& |W_n|^{-1/2}U_{n,3}(\beta^*, \psi^*)\widetilde{\Sigma}_{n,33}^{-1/2} \\ &- |W_n|^{-1/2}(\widehat{\beta}_n - \beta^*)J_{n,13}(\widetilde{\beta}, \psi^*)\widetilde{\Sigma}_{n,33}^{-1/2},\end{aligned}$$

where $\|\widetilde{\beta} - \beta^*\| \leq \|\widehat{\beta}_n - \beta^*\|$. In this way we can show boundedness in probability of the two terms on the right-hand side which are easier to handle than $|W_n|^{-1/2}U_{n,3}(\widehat{\beta}_n, \psi^*)\widetilde{\Sigma}_{n,33}^{-1/2}$.

Hence, the statement of the Theorem 3.6 is obtained once we verify the conditions of Lemma 3.7 for the three cases described above.

First step, i.e. $|W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*)$. Recall that $u_n(\cdot) = U_{n,1}(\cdot)$, $V_n = |W_n \times T_n|^{1/2} \cdot \mathbf{1}$.

Condition (a) of Lemma 3.7 follows easily from the fact that $|W_n \times T_n| \to \infty$ as $n \to \infty$.

To verify condition (b) first note that $|W_n \times T_n|^{-1}J_{n,11}(\beta^*)$ is a real symmetric matrix (and hence all its eigenvalues are real) and

$$l_n = \inf_{\|\phi\|=1}\left\{\phi\frac{J_{n,11}(\beta^*)}{|W_n \times T_n|}\phi^T\right\}$$

is equal to the smallest eigenvalue of $|W_n \times T_n|^{-1}J_{n,11}(\beta^*)$ (Hager, 2001). Since the eigenvalues of a matrix are in fact roots of a certain polynomial, Motzkin (1947, Sec. 2) gives uniform continuity of the mapping $g : A \mapsto \inf_{\|\phi\|=1}\{\phi A\phi^T\}$, i.e. $g(A)$ is the smallest eigenvalue of the matrix A.

We show in Lemma 3.10 that $|W_n \times T_n|^{-1}J_{n,11}(\beta^*) - I_{n,11}$ converges to 0 in probability as $n \to \infty$. Taking advantage of the uniform continuity of the mapping g it is easy to show that also $g(|W_n \times T_n|^{-1}J_{n,11}(\beta^*)) - g(I_{n,11})$ converges to 0 in probability as $n \to \infty$, i.e. the difference of the smallest eigenvalues of the respective matrices converges to 0 in probability. This and the assumption *(A3)* verifies condition (b) where we take $l = 1/2 \cdot \liminf \omega_{n,11} > 0$.

Regarding condition (c), it is sufficient to show that

$$\gamma_{nd}^{ij} = \sup_{\|(\theta-\theta^*)|W_n \times T_n|^{1/2}\|\leq d} \frac{|J_{n,11}^{ij}(\theta) - J_{n,11}^{ij}(\theta^*)|}{|W_n \times T_n|}$$

converges in probability to 0 as $n \to \infty$, where $J_{n,11}^{ij}$ is the (i,j)-th element of the matrix $J_{n,11}$. Using the boundedness and continuity assumptions in *(A5)* and *(A9)* (in fact uniform continuity assumptions as the supremum in definition of γ_{nd}^{ij} is taken over a compact set) one can show that the first two moments of γ_{nd}^{ij} can be made arbitrarily close to 0 by choosing n large enough. This implies the required convergence in probability.

Using Campbell theorem and assumptions *(A2)* and *(A9)* it is easy to show for each element of the vector $|W_n \times T_n|^{-1/2}U_{n,1}(\beta^*)$ that its mean is 0 and its variance is bounded from above by the same constant for all n. This implies that $|W_n \times T_n|^{-1/2}U_{n,1}(\beta^*)$ is bounded in probability and hence the condition (d) is verified.

Second step, i.e. $|W_n \times T_n|^{1/2}(\widehat{\xi}_n - \xi^*)$**.** Recall that $u_n(\cdot) = U_{n,2}(\widehat{\beta}_n, \cdot), V_n = (|W_n \times T_n|\widetilde{\Sigma}_{n,22})^{1/2}$.

Condition (a) of Lemma 3.7 follows easily from the fact that $|W_n \times T_n| \to \infty$, $n \to \infty$, and $\widetilde{\Sigma}_{n,22} \to \widetilde{\Sigma}_{22}$ as $n \to \infty$ by assumption *(A4)*. It also secures invertibility of $\widetilde{\Sigma}_{n,22}$, at least for n large enough.

Regarding condition (b), we show in Lemma 3.10 that

$$|W_n \times T_n|^{-1} J_{n,22}(\widehat{\beta}_n, \xi^*) - I_{22} \to 0$$

in probability as $n \to \infty$, and thus

$$|W_n \times T_n|^{-1}\widetilde{\Sigma}_{n,22}^{-1/2} J_{n,22}(\widehat{\beta}_n, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2} - \widetilde{\Sigma}_{n,22}^{-1/2} I_{22}\widetilde{\Sigma}_{n,22}^{-1/2} \to 0$$

in probability as $n \to \infty$.

By assumption *(A3)* I_{22} is a positive definite matrix and hence all its eigenvalues are positive. This implies that lim inf of the smallest eigenvalue of the matrix $\widetilde{\Sigma}_{n,22}^{-1/2} I_{22}\widetilde{\Sigma}_{n,22}^{-1/2}$ is positive. Now it is possible to use the same argument as in the first step to check condition (b).

To verify condition (c) we need to check that for any $d > 0$, $\gamma_{nd} \to 0$ in probability as $n \to \infty$, where

$$\gamma_{nd} = \sup_{\left\|(\xi-\xi^*)|W_n\times T_n|^{1/2}\widetilde{\Sigma}_{n,22}^{1/2}\right\| \le d} \left\| \frac{\widetilde{\Sigma}_{n,22}^{-1/2}(J_{n,22}(\widehat{\beta}_n, \xi) - J_{n,22}(\widehat{\beta}_n, \xi^*))\widetilde{\Sigma}_{n,22}^{-1/2}}{|W_n \times T_n|} \right\|_M .$$

By assumption *(A4)* $\widetilde{\Sigma}_{n,22}$ converges to a (deterministic) positive definite matrix and thus it is sufficient to verify the convergence in probability for

$$\widetilde{\gamma}_{nd'} = \sup_{\left\|(\xi-\xi^*)|W_n\times T_n|^{1/2}\right\| \le d'} \left\| \frac{J_{n,22}(\widehat{\beta}_n, \xi) - J_{n,22}(\widehat{\beta}_n, \xi^*)}{|W_n \times T_n|} \right\|_M$$

(note that it is possible for each $d > 0$ to find d' such that

$$\{\xi : \|(\xi - \xi^*)|W_n \times T_n|^{1/2}\widetilde{\Sigma}_{n,22}^{1/2}\| \le d\} \subseteq \left\{\xi : \left\|(\xi - \xi^*)|W_n \times T_n|^{1/2}\right\| \le d'\right\}$$

for n sufficiently large and check that $\widetilde{\gamma}_{nd'} \to 0$ in probability as $n \to \infty$). As in the first step, we verify the convergence for each element of $\widetilde{\gamma}_{nd'}$ separately. Using the continuity assumptions in *(A6')* the difference in $\widetilde{\gamma}_{nd'}$ can be made arbitrarily small by choosing n large enough and thus the required convergence is obtained and condition (c) is verified.

In view of Equation (3.35) it is sufficient for the condition (d) to be fulfilled that the following quantities are bounded in probability (recall that the term $|W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*)$ is bounded in probability from the first step):

$$|W_n \times T_n|^{-1/2} U_{n,2}(\beta^*, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2},$$
$$|W_n \times T_n|^{-1} J_{n,12}(\widetilde{\beta}, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2},$$

where $\|\widetilde{\beta} - \beta^*\| \le \|\widehat{\beta}_n - \beta^*\|$. Regarding the first term, consider the approximation

$$U_{n,2}(\beta^*, \xi^*) = \widetilde{U}_{n,2}(\beta^*, \xi^*) + V_{n,2}(\beta^*, \xi^*), \tag{3.36}$$

where $\widetilde{U}_{n,2}$ is defined at the beginning of Section 3.2.3. We argue below that $V_{n,2}(\beta^*, \xi^*)$ converges to 0 in probability and hence the approximation is justified. By Campbell theorem (and Fubini theorem) $|W_n \times T_n|^{-1/2}\widetilde{U}_{n,2}(\beta^*, \xi^*)\widetilde{\Sigma}_{n,22}^{-1/2}$ has mean 0 and its variance is an identity matrix (recall the definition of $\widetilde{\Sigma}_{n,22}$). It follows that $\widetilde{U}_{n,2}(\beta^*, \xi^*)$ and hence also $U_{n,2}(\beta^*, \xi^*)$ is bounded in probability.

Regarding the second term, it can be checked that it is bounded in probability by using Lemmas 3.8 and 3.9, assumptions *(A1), (A2), (A4), (A6')* and *(A8)* and the Cauchy-Schwarz inequality and Fubini theorem in appropriate places.

It remains to verify that $V_{n,2}(\beta^*, \xi^*)$ converges to 0 in probability but this can be checked using the same methods and assumptions. Combining these results we get condition (d) for the second step.

Third step, i.e. $|W_n|^{1/2}(\widehat{\psi}_n - \psi^*)$**.** Recall that $u_n(\cdot) = U_{n,3}(\widehat{\beta}_n, \cdot), V_n = (|W_n|\widetilde{\Sigma}_{n,33})^{1/2}$.

For verifying conditions (a)–(d) of Lemma 3.7 the same arguments can be used as in the second step. The only difference is that $|W_n|$ appears instead of $|W_n \times T_n|$ and K_s instead of K. Thus, the details are omitted here. We only remark that e.g. boundedness of $\lambda_{s,k}$ follows from boundedness of λ_k (or $\lambda_{0,k}$, respectively) by Proposition 1.21. □

Technical lemmas

The following lemmas are needed for the proof of Theorem 3.6. The first one is a version of Lemma 1 in Waagepetersen and Guan (2009, App. C).

Lemma 3.8. *Under assumptions* (A2), (A9) *and* (A12) *the variance of*

$$\sum_{(u,t),(v,s)\in X\cap(W_n\times T_n)}^{\neq} \frac{I(\|u-v\|\le r)I(|s-t|\le\tau)h((u,t),(v,s))}{|W_n\times T_n|\lambda(u,t;\beta^*)\lambda(v,s;\beta^*)}$$

is $O(|W_n \times T_n|^{-1})$ *for any bounded function* $h((u,t),(v,s))$ *symmetric in its arguments and for any* $r \ge 0$ *and* $\tau \ge 0$.

For the spatial projection process X_s *a similar result holds, namely under assumptions* (A2), (A9) *and* (A11) *the variance of*

$$\sum_{u,v\in X_s\cap W_n}^{\neq} \frac{I(\|u-v\|\le r)h(u,v)}{|W_n|\lambda_s(u;\beta^*)\lambda_s(v;\beta^*)}$$

is $O(|W_n|^{-1})$ *for any bounded function* $h(u,v)$ *symmetric in its arguments and for any* $r \ge 0$.

Proof. The two parts of this lemma can be proved using the same arguments. For the sake of more concise formulas we focus here on the second part (spatial projection process).

Let $\phi(u,v) = I(\|u-v\| \le r)h(u,v)/\{\lambda_s(u;\beta^*)\lambda_s(v;\beta^*)\}$. Assumption *(A2)* ensures that λ_s is bounded from below by some positive constant and hence ϕ is

bounded from above. Also, ϕ is a symmetric function. By Campbell theorem, the variance of the sum above is equal to

$$
\begin{aligned}
&|W_n|^{-2} \int_{W_n^4} \phi(u,v)\phi(w,z)[\lambda_{s,4}(u,v,w,z) - \lambda_{s,2}(u,v)\lambda_{s,2}(w,z)] \,\mathrm{d}u \,\mathrm{d}v \,\mathrm{d}w \,\mathrm{d}z \\
&+ 4|W_n|^{-2} \int_{W_n^3} \phi(u,v)\phi(v,w)\lambda_{s,3}(u,v,w) \,\mathrm{d}u \,\mathrm{d}v \,\mathrm{d}w \\
&+ 2|W_n|^{-2} \int_{W_n^2} \phi(u,v)^2\lambda_{s,2}(u,v) \,\mathrm{d}u \,\mathrm{d}v.
\end{aligned}
$$

It then follows by direct calculation that each of the three terms is $O(|W_n|^{-1})$. We show this for the last term. By *(A2)* and *(A9)* there exists $M < \infty$ such that $\lambda_{s,2}(u,v) < M$ and $\phi(u,v) < M$ for any $u, v \in \mathbb{R}^2$. It follows that

$$
\begin{aligned}
\left| \int_{W_n^2} \phi(u,v)^2\lambda_{s,2}(u,v) \,\mathrm{d}u \,\mathrm{d}v \right| &\leq M \int_{W_n} \left(\int_{W_n} |\phi(u,v)|^2 \,\mathrm{d}u \right) \mathrm{d}v \\
&\leq M^3 |\mathcal{B}(o,r)| \cdot |W_n|
\end{aligned}
$$

since the expression for $\phi(u,v)$ includes $I(\|u-v\| \leq r)$ and hence

$$
\int_{W_n} |\phi(u,v)|^2 \,\mathrm{d}u \leq M^2 |\mathcal{B}(v,r)| = M^2 |\mathcal{B}(o,r)|.
$$

Thus, the last term of the variance is indeed $O(|W_n|^{-1})$. □

The second lemma is a generalized version of Lemma 2 in Waagepetersen and Guan (2009, App. C).

Lemma 3.9. *Consider sequence $\{\breve{\beta}_n\}_{n\geq 1}$ such that $\breve{\beta}_n \to \beta^*$ in probability as $n \to \infty$. Under the assumptions of Theorem 3.6, for any $c \in \mathbb{R}$,*

$$
\sup_{t \in [t_0, t_1]} |\widehat{K}_n(t; \breve{\beta}_n)^c - K(t; \xi^*)^c|
$$

is $o_P(1)$ for any $0 < t_0 < t_1 < \infty$. If $c \geq 0$ we may take $t_0 = 0$. Similar statement holds also for

$$
\sup_{r \in [r_0, r_1]} |\widehat{K}_{s,n}(r; \breve{\beta}_n)^c - K_s(r; \psi^*)^c|.
$$

Proof. The condition $t_0 > 0$ if $c < 0$ is imposed to avoid division by 0 since $K(0; \xi^*) = 0$.

Due to *(A1)*, *(A2)* and the product structure (3.31) the intensity function λ is bounded and continuous as a function of β. Thus it is possible to show the convergence

$$
\widehat{K}_n(t; \breve{\beta}_n) - \widehat{K}_n(t; \beta^*) \to 0
$$

in probability as $n \to \infty$ for any $t \geq 0$. By Lemma 3.8 we get $\widehat{K}_n(t; \beta^*) \to K(t; \xi^*)$ in probability for any $t \geq 0$ and hence also

$$
\widehat{K}_n(t; \breve{\beta}_n) \to K(t; \xi^*)
$$

in probability as $n \to \infty$ for any $t \geq 0$. Using monotonicity of $\widehat{K}_n(t; \breve{\beta}_n)^c$ and $K(t; \xi^*)^c$ as functions of t, the result follows by arguments similar to those in the proof of the Glivenko-Cantelli theorem, see e.g. van der Vaart (1998, p. 266).

The same type of argument can be used to prove the second part of the lemma. □

The last lemma summarizes some convergence properties needed for the proof.

Lemma 3.10. *Under the conditions of Theorem 3.6, the following holds:*

(a) $|W_n \times T_n|^{-1} J_{n,11}(\beta^*) - I_{n,11} \to 0$ *in probability as* $n \to \infty$.

(b) $|W_n \times T_n|^{-1} J_{n,22}(\widehat{\beta}_n, \xi^*) - I_{22} \to 0$ *in probability as* $n \to \infty$,
$|W_n \times T_n|^{-1} J_{n,22}(\beta^*, \xi^*) - I_{22} \to 0$ *in probability as* $n \to \infty$.

(c) $|W_n|^{-1} J_{n,33}(\widehat{\beta}_n, \psi^*) - I_{33} \to 0$ *in probability as* $n \to \infty$,
$|W_n|^{-1} J_{n,33}(\beta^*, \psi^*) - I_{33} \to 0$ *in probability as* $n \to \infty$.

(d) $|W_n \times T_n|^{-1} J_{n,12}(\beta^*, \xi^*) - I_{n,12} \to 0$ *in probability as* $n \to \infty$.

Proof. (a) By Campbell theorem, $|W_n \times T_n|^{-1} J_{n,11}(\beta^*) - I_{n,11}$ has mean 0. Using assumptions *(A2)* and *(A9)* one can show that the variance of each element of the matrix in question is $O(|W_n \times T_n|^{-1})$. This implies the required convergence in probability.

(b) Note that $|W_n \times T_n|^{-1} J_{n,22}(\widehat{\beta}_n, \xi^*) = I_{22} - V_n$, where

$$\begin{aligned} V_n =& 2c_2 \int_{r_0}^{r_1} \left[\widehat{K}_n(t; \widehat{\beta}_n)^{c_2} - K(t; \xi^*)^{c_2} \right] \cdot \\ & \cdot \left[(c_2 - 1) K(t; \xi^*)^{c_2 - 2} K^{(1)}(t; \xi^*)^T K^{(1)}(t; \xi^*) + K(t; \xi^*)^{c_2 - 1} K^{(2)}(t; \xi^*) \right] \mathrm{d}t. \end{aligned}$$

It is now sufficient to show that $V_n \to 0$ in probability. Denote $s_n(c) = \sup_{t \in [t_0, t_1]} |\widehat{K}_n(t; \widehat{\beta}_n)^c - K(t; \xi^*)^c|$. Then we can write (using any matrix norm)

$$\begin{aligned} \|V_n\| \leq & 2c_2 s_n(c_2) \int_{t_0}^{t_1} \| (c_2 - 1) K(t; \xi^*)^{c_2 - 2} K^{(1)}(t; \xi^*)^T K^{(1)}(t; \xi^*) \\ & + K(t; \xi^*)^{c_2 - 1} K^{(2)}(t; \xi^*) \| \, \mathrm{d}t \\ \leq & \, const. \cdot (t_1 - t_0) s_n(c_2), \end{aligned}$$

since the integrand can be bounded from above by assumptions *(A2)*, *(A6')* and *(A8)*. By Lemma 3.9, $s_n(c_2) \to 0$ in probability and thus $\|V_n\| \to 0$ in probability as $n \to \infty$. This concludes the proof of the first part. Second part of the statement follows similarly.

(c) The proof of the first part follows precisely the arguments in (b), starting with $|W_n|^{-1} J_{n,33}(\widehat{\beta}_n, \psi^*) = I_{33} - V_n$, where V_n is the remainder term, and finishing with the use of Lemma 3.9. Second part of the statement follows similarly.

(d) The statement follows by similar arguments as above, using Lemmas 3.8 and 3.9.

□

Asymptotic normality theorems

In the following we establish asymptotic normality for the sequences $\{(\widehat{\beta}_n, \widehat{\xi}_n)\}_{n\geq 1}$ and $\{\widehat{\psi}_n\}_{n\geq 1}$ from Theorem 3.6, under appropriate assumptions. We separate the statement into two parts because of the different normalization given by $|W_n \times T_n|$ and $|W_n|$, respectively.

For a Borel set $A \in \mathcal{B}(\mathbb{R}^2 \times \mathbb{R})$ denote $\mathcal{F}^X(A)$ the σ-algebra generated by $X \cap A$. For $h > 0$ let $A_{ijk} = [ih, (i+1)h) \times [jh, (j+1)h) \times [kh, (k+1)h)$, $(i, j, k) \in \mathbb{Z}^3$, and

$$\alpha^F_{p_1,p_2}(m) = \sup \Bigg\{ \alpha(\mathcal{F}^X(S_1 \oplus \max(R, t_1)), \mathcal{F}^X(S_2 \oplus \max(R, t_1))) : S_1 = \bigcup_{M_1} A_{ijk},$$
$$S_2 = \bigcup_{M_2} A_{ijk}, |M_1| \leq p_1, |M_2| \leq p_2, d(M_1, M_2) \geq m, M_1, M_2 \subset \mathbb{Z}^3 \Bigg\},$$

where $|M|$ is the cardinality of the set $M \subseteq \mathbb{Z}^3$ and $d(M_1, M_2)$ denotes the minimal distance between M_1 and M_2 in the grid $\mathbb{Z}^3$. Also, $S_i \oplus \max(R, t_1)$ denotes the set S_i dilated by the distance $\max(R, t_1)$ where t_1 is the upper limit used in the minimum contrast criterion (3.34) and R is the fixed spatial range.

We remark that the numbering of the respective assumptions considered in the following corresponds to the numbering used in Section 3.1.3, with *(B3')* denoting an altered version of the assumption *(B3)* from Section 3.1.3.

Theorem 3.11. *Apart from the model assumptions formulated above and the assumptions of the Theorem 3.6, suppose there exist $\delta > 0$ and $\nu \in \mathbb{N}, 0 < \delta < \nu$, such that*

(C1) $\lambda_{0,4+2\nu}((u_1, t_1), \ldots, (u_{4+2\nu}, t_{4+2\nu})) < \infty$,

(B2) there exist $h > 0$ and $d > 3 \cdot \frac{2+\delta}{\delta}$ such that $\alpha^F_{2,\infty}(m) = \mathcal{O}(m^{-d})$,

(B3') the matrix $\widetilde{\Sigma}_n$ converges to a positive-definite matrix $\widetilde{\Sigma}$.

Then

$$|W_n \times T_n|^{1/2} \left\{ (\widehat{\beta}_n, \widehat{\xi}_n) - (\beta^*, \xi^*) \right\} I_n \widetilde{\Sigma}_n^{-1/2} \xrightarrow{d} N(0, \mathbf{1}).$$

Proof. Consider the following Taylor series expansion:

$$(U_{n,1}(\beta^*), U_{n,2}(\beta^*, \xi^*)) = \Big(U_{n,1}(\widehat{\beta}_n), U_{n,2}(\widehat{\beta}_n, \widehat{\xi}_n) \Big) + \left\{ (\widehat{\beta}_n, \widehat{\xi}_n) - (\beta^*, \xi^*) \right\} J_n(\widetilde{\beta}_n, \widetilde{\xi}_n), \tag{3.37}$$

where $(\widetilde{\beta}_n, \widetilde{\xi}_n)$ is between $(\widehat{\beta}_n, \widehat{\xi}_n)$ and (β^*, ξ^*). Now we multiply both sides of the equation by $|W_n \times T_n|^{-1/2} \widetilde{\Sigma}_n^{-1/2}$ and get

$$\begin{aligned} |W_n \times T_n|^{-1/2} \left(U_{n,1}(\beta^*), U_{n,2}(\beta^*, \xi^*)\right) \widetilde{\Sigma}_n^{-1/2} = \\ = |W_n \times T_n|^{-1/2} \Big(U_{n,1}(\widehat{\beta}_n), U_{n,2}(\widehat{\beta}_n, \widehat{\xi}_n) \Big) \widetilde{\Sigma}_n^{-1/2} \\ + |W_n \times T_n|^{1/2} \left\{ (\widehat{\beta}_n, \widehat{\xi}_n) - (\beta^*, \xi^*) \right\} \frac{J_n(\widetilde{\beta}_n, \widetilde{\xi}_n)}{|W_n \times T_n|} \widetilde{\Sigma}_n^{-1/2}. \end{aligned}$$

By Theorem 3.6, $\left(U_{n,1}(\widehat{\beta}_n), U_{n,2}(\widehat{\beta}_n, \widehat{\xi}_n)\right) = 0$ with probability tending to 1. Also,

$$\frac{J_n(\widetilde{\beta}_n, \widetilde{\xi}_n)}{|W_n \times T_n|} - \frac{J_n(\beta^*, \xi^*)}{|W_n \times T_n|} \to 0$$

in probability as $n \to \infty$ by continuity arguments similar to those used for checking condition (c) of Lemma 3.7 in the proof of Theorem 3.6 (see above). Finally, by Lemma 3.10 we have

$$\frac{J_n(\beta^*, \xi^*)}{|W_n \times T_n|} - I_n \to 0$$

in probability. Hence it is sufficient to check asymptotic normality for

$$|W_n \times T_n|^{-1/2} \left(U_{n,1}(\beta^*), U_{n,2}(\beta^*, \xi^*)\right) \widetilde{\Sigma}_n^{-1/2}.$$

In view of the approximation (3.36) we can see that its limiting distribution is the same as limiting distribution of

$$|W_n \times T_n|^{-1/2} (U_{n,1}(\beta^*), \widetilde{U}_{n,2}(\beta^*, \xi^*)) \widetilde{\Sigma}_n^{-1/2}$$

and hence we can focus on the latter one.

Let $h > 0$ be as in assumption *(B2)* and define

$$A_{ijk} = [ih, (i+1)h) \times [jh, (j+1)h) \times [kh, (k+1)h), \ (i,j,k) \in \mathbb{Z}^3.$$

Setting

$$X_{ijk} = \sum_{(u,t) \in X \cap A_{ijk}} \frac{\lambda^{(1)}(u,t;\beta^*)}{\lambda(u,t;\beta^*)} - \int_{A_{ijk}} \lambda^{(1)}(v,s;\beta^*) \, \mathrm{d}v \, \mathrm{d}s, \ (i,j,k) \in \mathbb{Z}^3,$$

we can easily show that

$$|W_n \times T_n|^{-1/2} U_{n,1}(\beta^*) = |W_n \times T_n|^{-1/2} \sum_{(i,j,k) \in \mathbb{Z}^3 : A_{ijk} \subseteq (W_n \times T_n)} X_{ijk} + o_P(1).$$

The remainder term corresponds to those A_{ijk} which hit the boundary of $W_n \times T_n$. The size of the boundary grows at a slower rate than the volume of $W_n \times T_n$ due to our assumption on the rectangular shape of the observation window. This is the key ingredient for showing that the remainder term is in fact $o_P(1)$, i.e. converges to 0 in probability.

Similarly, in $\widetilde{U}_{n,2}(\beta^*, \xi^*)$ we may replace $\widehat{K}_n(\tau;\beta^*)$ by

$$\frac{1}{|W_n \times T_n|} \sum_{(u,t) \in X \cap (W_n \times T_n)} \sum_{(v,s) \in X} \frac{I(0 < \|u-v\| \le R) I(0 < |t-s| \le \tau)}{\lambda(u,t;\beta^*) \lambda((v,s;\beta^*)}$$

and denote

$$\begin{aligned} Y_{ijk} =& 2c_2^2 \sum_{(u,t) \in X \cap A_{ijk}} \int_{t_0}^{t_1} \sum_{(v,s) \in X} \frac{I(0 < \|u-v\| \le R) I(0 < |t-s| \le \tau)}{\lambda(u,t;\beta^*) \lambda((v,s;\beta^*)} \cdot \\ & \cdot K(\tau;\xi^*)^{2c_2-2} K^{(1)}(\tau;\xi^*) \, \mathrm{d}\tau - 2c_2^2 h^3 \int_{t_0}^{t_1} K(\tau;\xi^*)^{2c_2-1} K^{(1)}(\tau;\xi^*) \, \mathrm{d}\tau. \end{aligned}$$

Then, it can be shown that

$$|W_n \times T_n|^{-1/2}\widetilde{U}_{n,2}(\beta^*, \xi^*) = |W_n \times T_n|^{-1/2} \sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk}\subseteq(W_n\times T_n)} Y_{ijk} + o_P(1).$$

We aim at using the Cramér-Wold theorem (Cramér and Wold, 1936) to show asymptotic normality of $|W_n \times T_n|^{-1/2}\left(U_{n,1}(\beta^*), \widetilde{U}_{n,2}(\beta^*, \xi^*)\right)\widetilde{\Sigma}_n^{-1/2}$. To do this, we take two arbitrary non-zero vectors x and y of appropriate dimensions and set

$$Z_{ijk} = X_{ijk}x^T + Y_{ijk}y^T,$$
$$\sigma_n^2 = |W_n \times T_n|^{-1}\,\mathsf{Var}\left(\sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk}\subseteq W_n\times T_n} Z_{ijk}\right) = (x,y)\widetilde{\Sigma}_n(x,y)^T + o(1).$$

In this way we have constructed a random field $\{Z_{ijk}\}$ defined on the integer lattice $\mathbb{Z}^3$.

We will show that

$$(\sigma_n^2|W_n \times T_n|)^{-1/2} \sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk}\subseteq(W_n\times T_n)} Z_{ijk}$$

is asymptotically standard normal. Together with *(B3')* this implies that

$$|W_n \times T_n|^{-1/2} \sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk}\subseteq(W_n\times T_n)} Z_{ijk}$$

converges in distribution to a normally distributed random variable with mean 0 and variance $(x,y)\widetilde{\Sigma}(x,y)^T$. Assumption *(B3')* and the Cramér-Wold theorem then imply that $|W_n \times T_n|^{-1/2}(U_{n,1}(\beta^*), \widetilde{U}_{n,2}(\beta^*, \xi^*))\widetilde{\Sigma}_n^{-1/2}$ is asymptotically standard normal.

To show that $(\sigma_n^2|W_n \times T_n|)^{-1/2}\sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk}\subseteq(W_n\times T_n)} Z_{ijk}$ is asymptotically standard normal we use a classical central limit theorem for random fields on a lattice (Guyon, 1995, Thm. 3.3.1), with the additional assumption of uniform integrability, see also the discussion in Karácsony (2006). Namely, the following conditions must be met for some $\delta > 0$:

(a) $\liminf \sigma_n^2 > 0$,

(b) $|Z_{ijk}|^{2+\delta}$ are uniformly integrable,

(c) $\sum_{m=1}^{\infty} m^2 \alpha_{2,\infty}^F(m)^{\delta/(2+\delta)} < \infty$.

Note that the mixing coefficient $\alpha_{2,\infty}^F$ defined above is the appropriate one to be used for this central limit theorem for random fields on a lattice since it takes into account the prescribed number of blocks A_{ijk}, not their volume. This precisely corresponds to the framework of a random field on a lattice.

Condition (a) above is fulfilled by assumption *(B3')* and condition (b) follows from the moment assumption *(C1)*. Finally, the mixing condition (c) is implied by *(B2)*. This concludes the proof. □

Now we focus on the properties of the estimator $\widehat{\psi}_n$ based on the spatial projection process X_s. For a Borel set $B \in \mathcal{B}(\mathbb{R}^2)$ denote $\mathcal{F}^{X_s}(B)$ the σ-algebra generated by $X_s \cap B$.

For $h > 0$ let $B_{ij} = [ih, (i+1)h) \times [jh, (j+1)h)$, $(i, j) \in \mathbb{Z}^2$, and

$$\alpha^F_{s,p_1,p_2}(m) = \sup \Bigg\{ \alpha(\mathcal{F}^{X_s}(S_1 \oplus r_1), \mathcal{F}^{X_s}(S_2 \oplus r_1)) : S_1 = \bigcup_{M_1} B_{ij}, S_2 = \bigcup_{M_2} B_{ij}, \\ |M_1| \le p_1, |M_2| \le p_2, d(M_1, M_2) \ge m, M_1, M_2 \subset \mathbb{Z}^2 \Bigg\},$$

where $d(M_1, M_2)$ denotes the minimal distance between M_1 and M_2 in the grid $\mathbb{Z}^2$. Also, $S_i \oplus r_1$ denotes the set S_i dilated by the distance r_1 where r_1 is the upper limit used in the minimum contrast criterion (3.33).

Theorem 3.12. *Apart from the model assumptions formulated above and the assumptions of the Theorem 3.6, suppose there exist $\delta > 0$ and $\nu \in \mathbb{N}, 0 < \delta < \nu$, such that*

(C1) $\lambda_{0,4+2\nu}((u_1, t_1), \ldots, (u_{4+2\nu}, t_{4+2\nu})) < \infty$,

(D2) there exist $h > 0$ and $d > 2 \cdot \frac{2+\delta}{\delta}$ such that $\alpha^F_{s,2,\infty}(m) = \mathcal{O}(m^{-d})$.

Then

$$|W_n|^{1/2}(\widehat{\psi}_n - \psi^*) I_{33} \widetilde{\Sigma}^{-1/2}_{n,33} \xrightarrow{d} N(0, \mathbf{1}).$$

Proof. The proof of this theorem follows closely the arguments in the proof of Theorem 3.11. However, there are two important differences. The first one lies in the mixing condition needed for the use of Theorem 3.3.1 in Guyon (1995). The assumption *(D2)* substitutes assumption *(B2)* – note that different rate is now required for the mixing coefficient $\alpha^F_{s,2,\infty}(m)$.

The second important difference lies in the Taylor series expansion at the beginning of the proof, see Equation (3.37). A slightly more delicate approach is necessary in this case. Consider the Taylor series expansion

$$(U_{n,1}(\beta^*), U_{n,3}(\beta^*, \psi^*)) = (U_{n,1}(\widehat{\beta}_n), U_{n,3}(\widehat{\beta}_n, \widehat{\psi}_n)) \\ + \Big\{ (\widehat{\beta}_n, \widehat{\psi}_n) - (\beta^*, \psi^*) \Big\} \begin{pmatrix} J_{n,11}(\widetilde{\beta}_n) & J_{n,13}(\widetilde{\beta}_n, \widetilde{\psi}_n) \\ 0 & J_{n,33}(\widetilde{\beta}_n, \widetilde{\psi}_n) \end{pmatrix},$$

where $(\widetilde{\beta}_n, \widetilde{\psi}_n)$ is between $(\widehat{\beta}_n, \widehat{\psi}_n)$ and (β^*, ψ^*). We now focus on the second part of the vector equation:

$$U_{n,3}(\beta^*, \psi^*) = U_{n,3}(\widehat{\beta}_n, \widehat{\psi}_n) + (\widehat{\beta}_n - \beta^*) J_{n,13}(\widetilde{\beta}_n, \widetilde{\psi}_n) + (\widehat{\psi}_n - \psi^*) J_{n,33}(\widetilde{\beta}_n, \widetilde{\psi}_n).$$

We multiply both sides of the equation (from the right) by $|W_n|^{-1/2} \widetilde{\Sigma}^{-1/2}_{n,33}$ and discuss each term separately.

On the left-hand side, the term $|W_n|^{-1/2} U_{n,3}(\beta^*, \psi^*) \widetilde{\Sigma}^{-1/2}_{n,33}$ depends only on the true parameter values and it can be shown that it has asymptotically a standard normal distribution, using the same techniques as in the proof of Theorem 3.11.

Moving to the right-hand side, $|W_n|^{-1/2}U_{n,3}(\widehat{\beta}_n, \widehat{\psi}_n)\widetilde{\Sigma}_{n,33}^{-1/2}$ equals 0 with probability tending to 1 by Theorem 3.6. We rewrite the remaining terms as follows:

$$|W_n|^{-1/2}(\widehat{\beta}_n - \beta^*)J_{n,13}(\widetilde{\beta}_n, \widetilde{\psi}_n)\widetilde{\Sigma}_{n,33}^{-1/2} = |W_n|^{1/2}(\widehat{\beta}_n - \beta^*)\frac{J_{n,13}(\widetilde{\beta}_n, \widetilde{\psi}_n)}{|W_n|}\widetilde{\Sigma}_{n,33}^{-1/2},$$
$$|W_n|^{-1/2}(\widehat{\psi}_n - \psi^*)J_{n,33}(\widetilde{\beta}_n, \widetilde{\psi}_n)\widetilde{\Sigma}_{n,33}^{-1/2} = |W_n|^{1/2}(\widehat{\psi}_n - \psi^*)\frac{J_{n,33}(\widetilde{\beta}_n, \widetilde{\psi}_n)}{|W_n|}\widetilde{\Sigma}_{n,33}^{-1/2}.$$

Under the assumptions of the theorem one can show that

- $|W_n|^{1/2}(\widehat{\beta}_n - \beta^*)$ converges to 0 in probability, since $|W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*)$ is bounded in probability by Theorem 3.6;
- $|W_n|^{-1}J_{n,13}(\widetilde{\beta}_n, \widetilde{\psi}_n) - I_{n,13}$ converges to 0 in probability, using similar continuity arguments as in the proof of Theorem 3.6;
- the elements of the matrix $I_{n,13}$ are bounded.

Also, by assumption *(A4)*, $\widetilde{\Sigma}_{n,33}$ converges to a positive definite matrix $\widetilde{\Sigma}_{33}$. We conclude that the whole term $|W_n|^{1/2}(\widehat{\beta}_n - \beta^*)\frac{J_{n,13}(\widetilde{\beta}_n,\widetilde{\psi}_n)}{|W_n|}\widetilde{\Sigma}_{n,33}^{-1/2}$ converges to 0 in probability as $n \to \infty$. Hence, this term does not affect the limiting distribution of $|W_n|^{1/2}(\widehat{\psi}_n - \psi^*)$.

To finish the proof it remains to show that $|W_n|^{-1}J_{n,33}(\widetilde{\beta}_n, \widetilde{\psi}_n) - I_{33}$ converges to 0 in probability. This can be done following the arguments in the proof of Theorem 3.11. □

The reason for splitting the asymptotic normality results into two theorems is that the current methodology cannot be employed directly to show joint asymptotic normality for the vector

$$L_n = \begin{pmatrix} |W_n \times T_n|^{1/2}(\widehat{\beta}_n - \beta^*) \\ |W_n \times T_n|^{1/2}(\widehat{\xi}_n - \xi^*) \\ |W_n|^{1/2}(\widehat{\psi}_n - \psi^*) \end{pmatrix}.$$

To see this, we could approximate as above

$$|W_n|^{-1/2}\widetilde{U}_{n,3}(\beta^*, \psi^*) = |W_n|^{-1/2} \sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk} \subseteq (W_n \times T_n)} Y'_{ijk} + o_P(1)$$

for suitably defined Y'_{ijk}. Note that it is not appropriate to use such an approximation for $|W_n \times T_n|^{-1/2}\widetilde{U}_{n,3}(\beta^*, \psi^*)$ as it converges to 0 in probability.

As a next step, we could define

$$Z'_{ijk} = X_{ijk}x^T + Y_{ijk}y^T + Y'_{ijk}y'^T,$$
$$\sigma_n^2 = |W_n \times T_n|^{-1} \mathsf{Var}\left(\sum_{(i,j,k)\in\mathbb{Z}^3 : A_{ijk} \subseteq W_n \times T_n} Z'_{ijk} \right),$$

and study the limiting distribution of

$$(\sigma_n^2|W_n \times T_n|)^{-1/2} \sum_{(i,j,k)\in\mathbb{Z}^3: A_{ijk}\subseteq(W_n\times T_n)} Z'_{ijk}.$$

With this normalization the part of the sum corresponding to $Y'_{ijk}y^T$ converges to 0 in probability. Using the Cramér-Wold theorem (Cramér and Wold, 1936) we would get singular asymptotic variance matrix for the whole vector L_n which is not a satisfactory result.

The other option is to define a triangular array of random fields

$$Z'_{n,ijk} = X_{ijk}x^T + Y_{ijk}y^T + |T_n|^{1/2}Y'_{ijk}y'^T$$

and use the central limit theorem of Karácsony (2006). However, in this case the required uniform integrability condition for $Z'_{n,ijk}$ does not hold. Hence, the current methodology cannot be employed directly to show joint asymptotic normality for the vector L_n.

Also note that the matrices $\widetilde{\Sigma}_n$ and $\widetilde{\Sigma}_{n,33}$ can be computed as discussed in Sec. 3.2 and App. B of Waagepetersen and Guan (2009). A plug-in approach can then be used in practice to estimate these matrices (together with I_n and I_{33}) in order to construct confidence regions for the estimates.

3.2.4 Simulation study III.

Design of the simulation study

To provide a comparison of the performance of the refined estimation method and the method using projection processes on finite observation windows, we performed a simulation study using the same model as in Section 3.1.5. Thus, the process X_0 is a stationary gamma shot-noise Cox point process determined by the parameters of the underlying Poisson measure $\mu > 0$ and $\theta > 0$ and the smoothing kernel k is the product of a bivariate Gaussian kernel k_1 with standard deviation $\sigma > 0$ in the spatial part and a uniform temporal kernel on the interval $[0, t_*]$ for $t_* > 0$. Both k_1 and k_2 are assumed to be scaled so that they integrate to 1 over their respective domains. The inhomogeneous version X of the process X_0 is obtained by location dependent thinning using the inhomogeneity function f which is a log-linear function of the coordinates, see Section 3.1.5. The observation window is $W \times T = [0, 1]^2 \times [0, 1]$.

We used the same combinations of parameter values as in the simulation study in Section 3.1.5 – in fact, the same realizations were used. To inspect the performance of the refined method in the case with even more severe cluster overlapping in the temporal projection process, we consider also $t_* = 0.045$ in addition to $t_* = 0.015$ and $t_* = 0.030$ used in the previous study. For the combinations of parameter values with $t_* = 0.045$ we generated new realizations and re-estimated the parameters using both the refined method and the method using projection processes to provide fair comparison.

To maintain reasonable extent of the simulation study, we consider only the combinations of parameter values with $\sigma = 0.02$. The refined method estimates σ precisely in the same way as the method using projection processes (cluster overlapping is not that serious issue in the spatial projection process compared

to the temporal projection) and the previous study did not indicate situations where the precision of estimates of the other model parameters would depend on the value of σ. The higher value $\sigma = 0.02$ resulted in almost all situations in lower precision of estimates of the other model parameters than $\sigma = 0.01$. Hence, $\sigma = 0.02$ (implying not very tight clusters) is the more interesting case for the comparison.

The inhomogeneity parameters β are estimated using the Poisson likelihood estimating equation (3.32). The parameters of the spatial and temporal smoothing kernels k_1 and k_2 are estimated by minimizing the contrasts (3.34) and (3.33), respectively. Finally, estimates of the parameters μ and θ of the underlying Poisson measure Φ are obtained from the previous estimates and the formula for the total intensity of points in the observation window.

For each combination of parameter values 500 independent realizations were used. All the computations were performed in `R`, with the use of auxiliary functions provided in the package `spatstat`, see Baddeley and Turner (2005).

Choice of tuning constants

In order to fully specify the estimation methods several tuning constants must be chosen. For estimation of the inhomogeneity parameters and the parameter σ of the spatial smoothing kernel k_1 we choose precisely the same values as in the simulation study in Section 3.1.5.

For estimation of the parameter t_* of the temporal smoothing kernel k_2 we use the minimum contrast criterion (3.33) with the following choices of tuning constants:

- we choose the fixed spatial range $R = 4\sigma$ which corresponds to the practical range of correlation of the (spatial) Gaussian kernel. Based on a smaller pilot study, this choice provides more precise estimates of μ and θ than $R = 3\sigma$ (possible interpretation is that whole clusters are taken into account and this provides relevant information about the total intensity of points and variability of the cluster weights). However, for t_* it is possible to achieve a slight improvement in the precision for most combinations of parameter values by choosing $R = 3\sigma$ (possible interpretation – spatially smaller, central part of the clusters play the key role in this case; this provides enough information about t_* and results in lower variability of $\widehat{K}(t)$);

- the variance stabilizing exponent c_2 is set to be $1/4$. Based on a pilot study, the alternative choice $c_2 = 1/2$ results in slight improvement in the estimates precision in most situations but in certain cases ($\mu = 100, \theta = 1/20$) it yields much worse estimates than $c_2 = 1/4$. Hence, we aim for robustness of the estimation method by selecting the smaller value;

- $t_0 = 0.05t_*$, the same value as in the study in Section 3.1.5. This resulted in better precision of estimates in the pilot study than the alternatives;

- we choose the value t_1 to be $1.10t_*$. We inspected the following choices in the pilot study: $0.95t_*, 1.00t_*, 1.05t_*, 1.10t_*$. The results indicated that the precision of estimates of t_* is better for higher values of t_0. The influence

true values			rel. bias			rel. MSE		
μ	θ	t_*	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$
50	1/20	0.015	0.095	0.157	-0.0005	0.064	0.367	0.0057
		0.030	0.085	0.188	0.0023	0.064	0.395	0.0063
		0.045	0.108	0.219	0.0047	0.063	0.368	0.0064
	1/40	0.015	0.079	0.162	0.0013	0.050	0.335	0.0016
		0.030	0.085	0.146	0.0042	0.058	0.359	0.0023
		0.045	0.113	0.243	-0.0070	0.065	0.452	0.0028
100	1/20	0.015	0.045	0.120	0.0001	0.035	0.220	0.0024
		0.030	0.058	0.155	-0.0024	0.032	0.225	0.0029
		0.045	0.060	0.146	-0.0021	0.037	0.196	0.0034
	1/40	0.015	0.032	0.090	0.0005	0.031	0.159	0.0010
		0.030	0.052	0.081	0.0103	0.029	0.155	0.0318
		0.045	0.029	0.094	0.0206	0.029	0.173	0.0494

Table 3.3: Relative biases and relative mean squared errors of the estimates for the inhomogeneous space-time gamma shot-noise Cox process. Estimates obtained using the refined method. Characteristics estimated from all 500 realizations.

of the choice of t_0 on the precision of estimates of other parameters is negligible.

Results

Table 3.3 shows relative mean biases of the estimators and relative mean squared errors (MSEs) (by relative we mean divided by the true value of the estimated parameter or by its square for the MSE). The characteristics were obtained using all the estimates from 500 replications, as opposed to the study reported in Section 3.1.5 where only the middle 95 % of the estimates were used due to lower numerical stability of the optimization precedures. In the present study the convergence of the numerical optimization methods was not an issue.

Note that only the results for the extreme combinations of parameters μ and θ are reported since the performance of the estimators showed consistent behaviour within the above-described range of parameter values. We also omit the estimator of σ as it is precisely the same as reported in Section 3.1.5.

The overall performance of the estimator has improved compared to the method using projection processes. The most notable change is that $\widehat{t_*}$ is now essentially unbiased, i.e. the negative bias occuring for the previous method has been removed. Only two combinations of parameter values exhibit a relative bias larger than 1 % ($\mu = 100, \theta = 1/40$ and $t_* = 0.030$ and 0.045, respectively). This is caused by numerical divergence of the estimator for 2 and 4 realizations, respectively (note that the characteristics in Table 3.3 are estimated from all 500 realizations). Also, huge improvement has been achieved in terms of the relative MSE of $\widehat{t_*}$.

Regarding $\widehat{t_*}$, very precise estimates are obtained even for combinations of parameter values with $t_* = 0.045$, i.e. much wider temporal smoothing kernel than

the ones considered in the simulation study in Section 3.1.5. Both relative bias and relative MSE are reasonably small. This is also true for $\widehat{\mu}$ and $\widehat{\theta}$. Hence, the issue of possible cluster overlapping if the estimation using projection processes was used is successfully remedied.

The estimators of μ and θ exhibit a positive bias which is lower for models with higher value of μ and also for models with lower θ if $\mu = 100$. The same holds true for the relative MSEs of $\widehat{\mu}$ and $\widehat{\theta}$.

The precision of $\widehat{\mu}$ and $\widehat{\theta}$ does not depend on the value of t_* in most situations (no apparent trend). On the other hand, it improves considerably with increasing intensity of the process, i.e. with increasing value of μ and $1/\theta$.

Compared to the previous method using projection processes, the refined method provides estimates of μ and θ with smaller relative bias. Also, relative MSE of $\widehat{\mu}$ is considerably reduced. On the other hand, relative MSE of $\widehat{\theta}$ cannot be compared to the method using projection processes in a straightforward way – the characteristics were estimated using 95 % and 100 % of realizations, respectively, and thus higher values of relative MSE shown for the refined method do not imply that it is inferior to the previous method.

To provide fair comparison, we also show the corresponding table of relative biases and relative MSEs estimated from the middle 95 % of the estimates, as was the case in Section 3.1.5, see Table 3.4. In order to stress the improvement when using the refined method, we also show the table corresponding to the method using projection processes (it is extracted from Table 3.1 with added lines corresponding to $t_* = 0.045$).

Table 3.4 indicates that the bias of $\widehat{t_*}$ has been removed by the refined method. The positive bias of $\widehat{\mu}$ and $\widehat{\theta}$ occurs for both methods but it is reduced approximately by one half (or even more in some cases) for the refined method. Also, when comparing the refined method to the method using projection processes, the relative MSE of $\widehat{t_*}$ is reduced approximately by a factor of 10, relative MSE of $\widehat{\mu}$ by one half and of $\widehat{\theta}$ by one quarter or one third, depending on a particular combination of parameter values.

To conclude, the refined method constitutes a significant improvement compared to the method using projection processes. The parameter t_* of the temporal smoothing kernel is estimated very precisely while $\widehat{\mu}$ and $\widehat{\theta}$ still exhibit positive bias, also resulting in higher relative MSEs for the considered models. Especially the parameter θ of the underlying Poisson measure is difficult to estimate – probably due to the fact that it expresses itself mainly in the distribution of cluster weights and not in the spatio-temporal arrangement of points of the process in question.

true values			rel. bias (95 %)			rel. MSE (95 %)		
μ	θ	t_*	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$
50	1/20	0.015	0.070	0.070	0.0003	0.048	0.201	0.0038
		0.030	0.063	0.108	-0.0005	0.048	0.258	0.0038
		0.045	0.082	0.144	-0.0011	0.046	0.236	0.0041
	1/40	0.015	0.062	0.087	-0.0006	0.038	0.213	0.0011
		0.030	0.063	0.064	0.0010	0.042	0.212	0.0016
		0.045	0.092	0.152	-0.0054	0.046	0.268	0.0019
100	1/20	0.015	0.041	0.058	-0.0008	0.026	0.138	0.0018
		0.030	0.044	0.094	-0.0020	0.025	0.144	0.0021
		0.045	0.048	0.096	-0.0036	0.026	0.130	0.0024
	1/40	0.015	0.030	0.048	-0.0004	0.021	0.111	0.0007
		0.030	0.040	0.031	-0.0032	0.021	0.104	0.0009
		0.045	0.028	0.039	-0.0033	0.020	0.112	0.0012

true values			rel. bias (95 %)			rel. MSE (95 %)		
μ	θ	t_*	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$	$\widehat{\mu}$	$\widehat{\theta}$	$\widehat{t_*}$
50	1/20	0.015	0.181	0.192	-0.0327	0.116	0.326	0.0217
		0.030	0.155	0.215	-0.0623	0.100	0.365	0.0423
		0.045	0.187	0.253	-0.0936	0.115	0.332	0.0674
	1/40	0.015	0.144	0.170	-0.0305	0.087	0.275	0.0139
		0.030	0.154	0.159	-0.0555	0.094	0.281	0.0254
		0.045	0.171	0.239	-0.0806	0.102	0.344	0.0411
100	1/20	0.015	0.121	0.154	-0.0350	0.071	0.210	0.0178
		0.030	0.138	0.205	-0.0673	0.078	0.240	0.0380
		0.045	0.133	0.185	-0.0909	0.076	0.188	0.0615
	1/40	0.015	0.089	0.117	-0.0235	0.057	0.164	0.0124
		0.030	0.124	0.126	-0.0540	0.060	0.154	0.0272
		0.045	0.102	0.127	-0.0771	0.051	0.152	0.0419

Table 3.4: Relative biases and relative mean squared errors of the estimates for the inhomogeneous space-time gamma shot-noise Cox process. Estimates obtained using the refined method (top part) and the method using projection processes (bottom part). Characteristics estimated from the middle 95 % of estimates.

4. Directions of future research

The topic of statistical inference for spatial and space-time point processes is still lively and evolving and the research in this field is far from being finished. The results presented in this thesis contribute to the field but at the same time raise a number of follow-up questions and interesting topics which call for further inquiry. This chapter provides a brief discussion on a few such topics and possible directions of future research.

4.1 Separability assumptions for space-time processes

The estimation procedures for space-time shot-noise Cox processes presented in Chapter 3 take advantage of the particular model structure, see Equations (3.1) and (3.2). These separability assumptions enable us to work with the first- and second-order moment properties of the projection processes X_t and X_s in a tractable form.

The assumption (3.2) of space-time separability of the smoothing kernel k can sometimes be justified in practice, see the discussion in Section 3.1. Also, in applications it may be difficult to decide whether the observed data arise from a process with separable or non-separable smoothing kernel. The main obstacle is that usually only a few points of the process occur in individual clusters.

On the other hand, the assumption (3.1) of space-time separability of the inhomogeneity function f used for location dependent thinning is mostly technical and can be considered a nuisance. However, without this assumption the form of the first- and second-order moment characteristics of the projection processes X_t and X_s would be very complicated and could not be used for parameter inference.

4.1.1 First-order separability

While the assumption (3.1) of separability of the inhomogeneity function f, i.e.

$$f(u,t) = f_1(u)f_2(t),\ u \in \mathbb{R}^2, t \in \mathbb{R},$$

enables us to use the projection processes for statistical inference, there is a way how it could be eliminated. The key idea was already presented in Section 3.2 when describing the refined estimation method which avoids projection to the temporal domain. Hence, the refined method does not rely at all on the temporal projection process X_t.

We propose to extend this idea also to the spatial part of the model. To avoid projection to the spatial domain, i.e. instead of using the spatial projection process X_s, we suggest to estimate the values of the space-time K-function $K(r, \widetilde{T})$ with a fixed temporal lag $\widetilde{T} > 0$ as a function of a single (spatial) argument. The non-parametric or semi-parametric estimate of $K(r, \widetilde{T})$ can be used for minimum contrast estimation in order to estimate parameters of the spatial smoothing kernel.

The main advantage of this approach is that we do not need to work with moment characteristics of the projected processes X_t and X_s and hence we can drop the first-order separability assumption (3.1) for the inhomogeneity function f.

For defining the estimation procedure, we allow for a general (non-negative) function f such that $\max_{W\times T} f = 1$. Assume now a parametric model for the shot-noise Cox process X similar to those used in Sections 3.1 and 3.2 but with

$$\psi_0 = \frac{V_2}{\mu\,(V_1)^2} \int_{|s|<\widetilde{T}} K_2(s;\widetilde{\xi})\,\mathrm{d}s.$$

The third step of the refined method described in Section 3.2 is now replaced by minimizing the discrepancy

$$m_{3,\beta}(\psi) = \int_{r_0}^{r_1} \left[\widehat{K}_{\widetilde{T}}(r;\beta)^{c_3} - K_{\widetilde{T}}(r;\psi)^{c_3}\right]^2 \mathrm{d}r,$$

conditionally on $\beta = \widehat{\beta}$, where

$$\widehat{K}_{\widetilde{T}}(r;\beta) = \sum_{(u,s),(v,\tau)\in X\cap(W\times T)}^{\neq} \frac{I\left\{\|u-v\| < r\right\} I\{|s-\tau| < \widetilde{T}\}}{\lambda(u,s;\beta)\lambda(v,\tau;\beta)|W\cap W_{u-v}|\,|T\cap T_{s-\tau}|}$$

is the empirical estimate of $K_{\widetilde{T}}(r) = K(r,\widetilde{T})$ (omitting now the dependence on the model parameters to avoid confusion).

To study asymptotic properties of the estimator under the increasing-window asymptotics we can formulate the estimation procedure in the framework of estimating equations just as we did in Sections 3.1 and 3.2. Assume now that the inhomogeneity function f is of the form

$$f(z(u,t)\widetilde{\beta}^T),\ u\in\mathbb{R}^2, t\in\mathbb{R},$$

where $z(u,t)$ is the vector of space-time covariates, $\widetilde{\beta}$ is the vector of unknown inhomogeneity parameters and f is positive, strictly increasing, twice continuously differentiable function. Setting again $\beta_0 = \log(\mu V_1)$, the first-order intensity function of the process X is parametrized by the vector $\beta = (\beta_0, \widetilde{\beta})$.

We hypothesise that with such parametrization it is possible to prove consistency and asymptotic normality results for the proposed estimator similar to those presented in Sections 3.1 and 3.2, under appropriate assumptions. However, it still remains to construct the proofs in detail.

Another advantage of the proposed approach is that the asymptotic normality results can be formulated in a single theorem because in the three estimation steps the same normalization by $|W_n \times T_n|$ is used. In this way, certain clumsiness of formulation of the asymptotic normality theorems presented in the previous chapter is eliminated.

4.1.2 Separability of the smoothing kernel

The estimation procedures described in Chapter 3 depend strongly on the assumption (3.2) of separability of the space-time smoothing kernel k, i.e.

$$k(u,t) = k_1(u)k_2(t),\ u\in\mathbb{R}^2, t\in\mathbb{R},$$

where k_1 and k_2 are probability density functions on $\mathbb{R}^2$ and $\mathbb{R}$, respectively. The important consequence is the special structure (3.4) of the space-time pair-correlation function and hence also the space-time K-function:

$$K(r,t) = 2\pi r^2 t + \frac{V_2}{\mu\,(V_1)^2} \int_{|u|<r} K_1(u)\,\mathrm{d}u \int_{|s|<t} K_2(s)\,\mathrm{d}s,\ r \geq 0, t \geq 0. \tag{4.1}$$

This enables us to separate the parameters of the spatial kernel k_1 and the temporal kernel k_2 and estimate them sequentially. Without this assumption both the method using projection processes and the refined method cannot be used.

In practical applications one should check the separability assumption (3.2) before using the estimation methods presented in Chapter 3. To do this, Møller and Ghorbani (2012) proposed a functional summary statistic $F(r,t)$, $r,t > 0$, which is constant if (3.2) is fulfilled. It is defined as

$$F(r,t) = \frac{K(r,t) - 2\pi r^2 t}{(K_1^{MG}(r) - \pi r^2)(K_2^{MG}(t) - 2t)},\ r,t > 0,$$

where K_1^{MG} and K_2^{MG} are defined in (3.24). Non-parametric or semi-parametric estimates of K_1^{MG} and K_2^{MG} can be obtained in a straightforward way (Møller and Ghorbani, 2012, Sec. 4.1) and hence the empirical estimate of F is obtained as

$$\widehat{F}(r,t) = \frac{\widehat{K}(r,t) - 2\pi r^2 t}{(\widehat{K}_1^{MG}(r) - \pi r^2)(\widehat{K}_2^{MG}(t) - 2t)},\ r,t > 0. \tag{4.2}$$

If the estimated surface $\widehat{F}(r,t)$ is approximately constant, we may consider the separability condition (3.2) fulfilled.

This method for checking (3.2) is based on visual assessment of the surface $\widehat{F}(r,t)$ or, if a parametric model has been fitted to the data, plotting of appropriate envelopes from simulated realizations from the fitted model. The latter, however, is in fact a model validation procedure (relying on a specified parametric form of the smoothing kernel etc.) rather than procedure for checking the assumption (3.2) alone.

For this reason we focus now on visual assessment of the surface $\widehat{F}(r,t)$. From the author's point of view the summary statistic F defined above is not suited for visual assessment due to high variability of its empirical estimate $\widehat{F}$ – the terms in the denominator of (4.2) increase substantially the variability of $\widehat{F}$.

To remedy this, we propose to perform visual assessment of sections of the surface $\widehat{K}(r,t)$ in order to avoid division by empirical estimates of the respective summary statistics as was the case in (4.2). The idea is based on the observation that, for distinct values $0 < t_1 < \ldots < t_n$, the functions

$$\begin{aligned} F_1'(r) &= K(r,t_1) - 2\pi r^2 t_1,\ r \geq 0, \\ &\vdots \\ F_n'(r) &= K(r,t_n) - 2\pi r^2 t_n,\ r \geq 0, \end{aligned}$$

are the same (up to a multiplicative constant) if the separability assumption (3.2) is fulfilled. This is a direct consequence of the structure (4.1). In order to eliminate the multiplicative constants we propose to specify an interval

$[R_0, R_1], 0 \leq R_0 < R_1 < \infty$, and define

$$F_1(r) = \frac{F_1'(r)}{\int_{R_0}^{R_1} F_1'(s)\,\mathrm{d}s},\ r \in [R_0, R_1],$$

$$\vdots$$

$$F_n(r) = \frac{F_n'(r)}{\int_{R_0}^{R_1} F_n'(s)\,\mathrm{d}s},\ r \in [R_0, R_1].$$

Under the assumption (3.2), the functions $F_1, \ldots, F_n$ are all equal. By pluging in the empirical estimates of $K(r,t)$ we may easily obtain estimates of $F_1, \ldots, F_n$ and visually assess if they are the same. If not, the assumption (3.2) should be dropped. Naturally, the variability of the estimated values can make the decision very difficult.

Also, we can do the same for the perpendicular sections of the surface $K(r,t)$, i.e. specify the values $0 < r_1 < \ldots < r_m$ and define functions $G_i(t)$ based on the values of $K(r_i, t), i = 1, \ldots, m$. This can provide complementary information.

However, based on a preliminary analysis, the variability of the estimates can still be high and it may be difficult to decide whether the plotted curves $\widehat{F}_i(r)$ (or $\widehat{G}_i(t)$) are the same. To simplify the decision-making process we recommend using interactive sequential plotting of the functions $\widehat{F}_i(r)$ (or $\widehat{G}_i(t)$), i.e. plotting the functions one-by-one with the possibility to step backwards and forwards, to see if there is an apparent trend in the values. Such a functionality is provided e.g. by the command `Manipulate` in the `Mathematica` software.

In an attempt to justify this approach we have inspected the theoretical curves $F_i(r)$ and $G_i(t)$ for a particular ambit process model for which the separability assumption (3.2) is violated. The differences in the curves were visible but relatively small in scale and we observed that one could not reliably detect the differences if empirical estimates were used instead of the theoretical values. However, it remains a question of further investigation whether for other models the proposed approach can sucessfully detect violations of the separability assumption (3.2).

4.2 Other moment estimation methods for space-time processes

We hypothesise that for inhomogeneous space-time shot-noise Cox point process models considered in Chapter 3, i.e. models fulfilling the separability assumptions (3.1) and (3.2), it is possible to develop step-wise estimation procedures based on projection processes X_t and X_s similar to the one described in Section 3.1 but using composite likelihood or Palm likelihood approach instead of minimum contrast estimation.

The first- and second-order moment characteristics of such processes are available in a closed form, assuming a suitable parametric form of the smoothing kernel, see Equations (3.5), (3.6) and (3.7). Hence, it is likely that the composite likelihood or the Palm likelihood method can be employed for the projection processes X_t and X_s and that the asymptotic properties of the resulting estimators can be established using the methodology described briefly in Section 3.1 and in detail in Section 3.2.

It is possible that for certain space-time point process models such estimators would prove to be more efficient on finite observation windows than the minimum contrast estimator. We observed this in the spatial setting in the paper Dvořák and Prokešová (2012). However, investigation of such estimators still remains a question of further research.

4.3 Asymptotics for minimum contrast estimation with the g-function

As discussed in Section 2.1.1, paramater estimation for certain classes of Cox point process models is possible using minimum contrast estimation with either the K-function or the pair-correlation function g (for isotropic processes for which g is a function of a scalar argument). For this, one needs empirical estimates of K or g. While K can be estimated in a straightforward way by counting pairs of points within certain distance and with proper weights, empirical estimation of g is usually performed by kernel smoothing.

The need to specify the smoothing kernel and its bandwidth is the reason why the minimum contrast estimation based on the K-function (denoted here MCEK) has seen much more use than the version based on the pair-correlation function g (denoted here MCEg). However, it has been shown by simulation studies that in many situations using MCEg results in better precision and lower bias of the estimates than using MCEK, see Dvořák and Prokešová (2012) and Guan (2009). Also, cross-validation procedures have been developed for choosing optimal bandwidth of the smoothing kernel, see Guan (2007a,b). This makes MCEg a viable alternative to MCEK.

Asymptotic properties (consistency, asymptotic normality) of the estimates obtained by MCEK have been studied and established under the increasing window asymptotics in different settings and for different classes of Cox point processes in Heinrich (1992), Guan and Sherman (2007) and Waagepetersen and Guan (2009), and also in the previous chapter of this thesis. These results provide theoretical justification of the estimation method in the sense that using more information from larger and larger observation windows increases the precision of the estimates.

For estimates obtained by MCEg such asymptotic results have not been established yet, mainly due to the lack of convergence results for the empirical estimates of the pair-correlation function. This has been remedied by the recent paper Heinrich and Klein (2014) which formulates conditions under which consistency and asymptotic normality of the empirical estimates of the pair-correlation function hold.

The method of parameter estimation by MCEg can be reformulated to the framework of estimating equations (Mukhopadhyay, 2004). Hence, it is possible to establish consistency of the estimates by the methodology of estimating equations, using e.g. results of Crowder (1986). In the field of spatial statistics these results were used e.g. in the work of Guan (2006) or Prokešová and Jensen (2013).

Asymptotic normality of the estimates can be established under α-mixing assumption (Doukhan, 1994) using an appropriate central limit theorem for random

fields. The first formulation of such a theorem is due to Bolthausen (1982). It was later extended to non-stationary random fields in Guyon (1995) and to triangular arrays of non-stationary random fields in Karácsony (2006). These results were used in the context of parameter estimation for point processes in the work of Waagepetersen and Guan (2009) and Coeurjolly and Møller (2014).

Combining the results from the three above-mentioned areas (convergence of the empirical estimates of the pair-correlation function, estimating equations methodology and central limit theorems for random fields) it is likely that asymptotic properties for the MCEg estimator can be established under suitable conditions. This would provide theoretical justification of MCEg as a method for estimation of clustering parameters and set the basis for construction of confidence regions and statistical tests of parameter hypotheses.

4.4 Model validation for shot-noise Cox processes

After fitting a parametric model to the observed point pattern—either spatial or space-time—a goodness-of-fit analysis should naturally follow in order to reveal possible inadequacy of the model. We discuss here solely the inhomogeneous scenario as the methods can be easily applied in the stationary case.

According to the discussion in Jalilian and Vahidi-Asl (2011, Sec. 1), model validation for planar Cox processes is currently performed by a comparison between a tractable second-order characteristic (such as the K-function) of the fitted model and its empirical estimate (Waagepetersen, 2007). In addition to the visual comparison, simulation-based tests can be performed based on maximum or integral deviation measures or envelopes, see e.g. Myllymäki et al. (2013) and the references therein.

However, one should not be using the same characteristic both for model fitting and validation as this may result in incorrect conclusions. Since the moment methods commonly used for estimation for Cox processes are based on second-order properties of the processes in question (see Chapter 2) plots of e.g. the pair-correlation function or the K-function cannot be used for validation of the fitted model.

Also, different Cox processes can have the same second-order properties as illustrated in Møller and Waagepetersen (2004, Ex. 5.7). Thus higher-order characteristics should be considered for checking the goodness-of-fit.

One possible approach was introduced in Jalilian and Vahidi-Asl (2011) who proposed a residual analysis based on Laplace functionals. We remark that the Laplace functional is defined for all point processes and uniquely characterizes the distribution of the process, see Daley and Vere-Jones (2008, Sec. 9.4). However, not all point processes have the Laplace functional in a tractable form.

The residual analysis described in Jalilian and Vahidi-Asl (2011) was originally developed for inhomogeneous Poisson-Neyman-Scott processes but can be extended to shot-noise Cox processes and any point process with tractable Laplace functional (Jalilian and Vahidi-Asl, 2011). The method uses Q-Q plots and the so-called *four-panel diagnostic plots* corresponding to those introduced in Baddeley et al. (2005). The advantage is that these plots can reveal (Jalilian and

Vahidi-Asl, 2011, Sec. 4.3):

- inadequacy of the model in describing spatial trend of data,
- covariates which must be included in the model,
- outliers in the data,
- overall disagreement, including spatial trend and clustering features, between data and the fitted model.

On the other hand, the method requires choosing type (scale) of the residuals and also choosing an appropriate collection of sets to be used in the analysis (the problem is that the residual set function is not additive (Jalilian and Vahidi-Asl, 2011, Sec. 4.2). Moreover, the diagnostic relies mostly on visual examination of the plots and simulation-based approach must be used in order to obtain envelopes for the Q-Q plots (for possible testing).

In order to streamline the model validation we propose to use a functional statistic based on Minkowski functionals (Schneider, 1993, Sec. 4.2) to perform global envelope testing (Myllymäki et al., 2013). The approach is based on the ideas presented in Parker et al. (2013).

Let X be a planar point process observed in a compact set W. For a given value $r > 0$ we center a disc with radius r on each point and analyze the topology of the set (the union of the discs)

$$Z(r) = \bigcup_{x \in X \cap W} \mathcal{B}(x, r).$$

We calculate the Minkowski functionals of $Z(r)$, i.e. the area $A(r)$ of $Z(r)$, together with its perimeter $P(r)$ and its Euler-Poincaré characteristic $\chi(r)$, see Parker et al. (2013). For each radius r the Minkowski functionals $A(r), P(r), \chi(r)$ depend on the locations of all observed points simultaneously and hence they incorporate information about interactions of all orders.

For the analysis we specify an interval $[r_0, r_1]$, $0 \leq r_0 < r_1 < \infty$, and calculate $A(r), P(r)$ and $\chi(r)$ for all values of r in $[r_0, r_1]$, hence obtaining the tree functional statistics. Now we concatenate them together and use this aggregate functional statistic for simulation-based goodness-of-fit testing. For this purpose e.g. the methods presented in Myllymäki et al. (2013) can be used.

Based on the simulation study described in Parker et al. (2013) we expect the Euler-Poincaré characteristic $\chi(r)$ to be the most informative of the three Minkowski functionals and the most capable of distinguishing between different Cox process models. This is motivated by the fact that $\chi(r)$ can vary more freely with increasing r than $A(r)$ or $P(r)$. Consequently, it may be beneficial to base the analysis solely on $\chi(r)$.

Note that the procedure outlined above can indicate disparities between the fitted model and the observed data but it provides no insight into the cause of poor fit. While the residuals of Jalilian and Vahidi-Asl (2011) can be viewed as a diagnostic tool, our procedure can be used only for model validation. However, it provides means of formal hypothesis testing and is much more straightforward than the residual analysis since only the values of r_0 and r_1 need to be specified.

To conclude with, we remark that adjusting the model validation procedures to space-time setting also remains a question of further research.

Conclusion

In the present thesis we have considered the problem of statistical inference for parametric spatial and space-time Cox point process models, both stationary and those fulfilling the SOIRS property (second-order intensity reweighted stationarity). We focused on moment estimation methods, i.e. we avoided the computationally demanding maximum likelihood estimation and Bayesian inference.

We have reviewed the state-of-the-art moment estimation methods for the stationary spatial Cox point processes, i.e. the minimum constrast estimation and the composite likelihood and Palm likelihood approaches, see Chapter 2. We compared the performance of these estimation methods for particular stationary shot-noise and log-Gaussian Cox process models by means of a simulation study and conclude that there is no uniformly best estimator and it depends on the particular model and the parameter of main interest which estimation method should be used. We also briefly discuss two-step estimation procedures for inhomogeneous SOIRS Cox processes.

The core of the thesis lies in Chapter 3 where we proposed a step-wise estimation procedure for inhomogeneous space-time shot-noise Cox processes with a particular non-trivial structure. The method takes advantage of the projections of the space-time process into the spatial and temporal domain, respectively, in order to reduce dimensionality of the problem and estimate different parts of the model separately. We have established consistency and asymptotic normality of the resulting estimator under the increasing window asymptotics and compared its performance to a previously published estimation method in a simulation study. The results indicated that at least for the models considered in the simulation study our method using projection processes produces better precision estimates than the previous method.

To remedy the issue of cluster overlapping in the projections—a particular drawback of the method using projection processes—we proposed a refined estimation method which avoids projection to the temporal domain where the problem of information loss is the most prominent. We have established consistency and asymptotic normality also for the refined method, again under the increasing window asymptotics. The main challenge in showing the asymptotic properties—both for the refined method and for the method using projection processes—was in the different normalization needed for different estimation steps.

We have also shown in a simulation study that the refined method does not suffer from the problem of cluster overlapping (at least when the clusters are reasonably tight in the spatial domain and the number of observed points is not extremely high) and hence it genuinely remedies the drawback of the method using projection processes. When comparing performance of the refined method to the method using projection processes, the negative bias of the estimates of the temporal clustering parameter has been removed and the relative mean squared error has been considerably reduced. Hence, substantial improvement in precision of the estimates is achieved when using the refined method.

In the final chapter we discussed several interesting open questions raised by the work described in this thesis and we indicated possible directions of further research. However, answering such questions lies outside the scope of this thesis.

In conclusion, the main contribution of this thesis lies in the two estimation methods proposed for inhomogeneous space-time shot-noise Cox processes. We have studied both the asymptotic properties of the estimators (which lay ground for construction of confidence regions and hypotheses testing) and their empirical performance for point patterns observed on middle-sized windows (which provides the prospective users a better insight and hints on choosing the necessary tuning constants). This makes the proposed methods ready-to-use in practical situations.

Bibliography

Babu, G. and Feigelson, E. (1996), *Astrostatistics*, Chapman & Hall, London.

Baddeley, A. J., Møller, J. and Waagepetersen, R. P. (2000), 'Non- and semi-parametric estimation of interaction in inhomogeneous point patterns', *Stat Neerlandica* **54**, 329–350.

Baddeley, A. J. and Silverman, B. W. (1984), 'A cautionary example on the use of second-order methods for analyzing point patterns', *Biometrics* **40**(4), 1089–1093.

Baddeley, A. J., Turner, R., Møller, J. and Hazelton, M. (2005), 'Residual analysis for spatial point processes (with discussion)', *J R Stat Soc B* **67**(5), 617–666.

Baddeley, A. and Turner, R. (2000), 'Practical maximum pseudolikelihood for spatial point patterns', *Aust NZ J Stat* **42**, 283–322.

Baddeley, A. and Turner, R. (2005), 'Spatstat: An R package for analyzing spatial point patterns', *J Stat Softw* **12**(6), 1–42.

Beneš, V., Bodlák, K., Møller, J. and Waagepetersen, R. P. (2011), 'A case study on point process modelling in disease mapping', *Image Anal Stereol* **24**(3).

Bolthausen, E. (1982), 'On the central limit theorem for stationary mixing random fields', *Ann Probab* **10**(4), 1047–1050.

Brix, A. (1999), 'Generalized gamma measures and shot-noise Cox processes', *Adv Appl Probab* **31**(4), 929–953.

Coeurjolly, J.-F. and Møller, J. (2014), 'Variational approach for spatial point process intensity estimation', *Bernoulli* (to appear).

Coles, P. and Jones, B. (1991), 'A lognormal model for the cosmological mass distribution', *Mon Not R Astron Soc* **248**, 1–13.

Cox, D. R. (1955), 'Some statistical models related with series of events', *J Roy Stat Soc B* **17**, 129–164.

Cramér, H. and Wold, H. (1936), 'Some theorems on distribution functions', *J Lond Math Soc* **s1-11**(4), 290–294.

Crowder, M. (1986), 'On consistency and inconsistency of estimating equations', *Economet Theor* **2**, 305–330.

Daley, D. J. and Vere-Jones, D. (2008), *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*, 2nd edn, Springer-Verlag, New York.

Diggle, P. J. (1983), *Statistical Analysis of Spatial Point Patterns*, Mathematics in biology, Academic Press, London.

Diggle, P. J. (2003), *Statistical Analysis of Spatial Point Patterns*, Mathematics in biology, 2nd edn, Arnold, London.

Diggle, P. J. (2007), Spatio-temporal point processes: methods and applications, *in* B. Finkenstädt, L. Held and V. Isham, eds, 'Statistical methods for spatio-temporal systems', Chapman and Hall/CRC, Boca Raton.

Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G. and Gooding, S. (2007), 'Second-order analysis of inhomogeneous spatial point processes using case–control data', *Biometrics* **63**(2), 550–557.

Doukhan, P. (1994), *Mixing: Properties and Examples*, Lecture notes in statistics, Springer-Verlag, New York.

Dvořák, J. and Prokešová, M. (2011), 'Moment estimation methods for stationary spatial Cox processes – a simulation study'. [Online; accessed 24. 2. 2014, http://www.karlin.mff.cuni.cz/ dvorak/CoxEstimation/ CoxEstimation_SimulationStudy.pdf].

Dvořák, J. and Prokešová, M. (2012), 'Moment estimation methods for stationary spatial Cox processes – a comparison', *Kybernetika* **48**(5), 1007–1026.

Gabriel, E. (2014), 'Estimating second-order characteristics of inhomogeneous spatio-temporal point processes', *Methodol Comput Appl Probab* **16**(2), 411–431.

Gabriel, E. and Diggle, P. J. (2009), 'Second-order analysis of inhomogeneous spatio-temporal point process data', *Stat Neerl* **63**(1), 43–51.

Guan, Y. (2006), 'A composite likelihood approach in fitting spatial point process models', *J Am Stat Assoc* **101**(476), 1502–1512.

Guan, Y. (2007*a*), 'A composite likelihood cross-validation approach in selecting bandwidth for the estimation of the pair correlation function', *Scand J Statist* **34**(2), 336–346.

Guan, Y. (2007*b*), 'A least-squares cross-validation bandwidth selection approach in pair correlation function estimations', *Statist Probab Lett* **77**(18), 1722–1729.

Guan, Y. (2009), 'A minimum contrast estimation procedure for estimating the second-order parameters of inhomogeneous spatial point processes', *Stat Interface* **2**, 91–99.

Guan, Y. and Sherman, M. (2007), 'On least squares fitting for stationary spatial point processes', *J Roy Stat Soc B* **69**(1), 31–49.

Guyon, X. (1995), *Random Fields on a Network*, Springer, New York.

Hager, W. (2001), 'Minimizing a quadratic over a sphere', *SIAM J Optimiz* **12**(1), 188–208.

Heinrich, L. (1992), Minimum contrast estimates for parameters of spatial ergodic point processes, *in* 'Trans. 11th Prague Conference on Random Processes, Information Theory and Statistical Decision Functions', Academic Publishing House, Prague, pp. 479–492.

Heinrich, L. and Klein, S. (2014), 'Central limit theorems for empirical product densities of stationary point processes', *Stat Inference Stoch Process*. To appear.

Hellmund, G., Prokešová, M. and Jensen, E. B. V. (2008), 'Lévy-based Cox point processes', *Adv Appl Probab* **40**, 603–629.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008), *Statistical Analysis and Modelling of Spatial Point Patterns*, Wiley, Chichester.

Jalilian, A., Guan, Y. and Waagepetersen, R. (2013), 'Decomposition of variance for spatial Cox processes', *Scand J Stat* **40**(1), 119–137.

Jalilian, A. H. and Vahidi-Asl, M. Q. (2011), 'Residual analysis for inhomogeneous Neyman–Scott processes', *Scand J Stat* **38**(4), 617–630.

Jensen, A. T. (2005), Statistical inference for doubly stochastic Poisson processes, PhD thesis, Department of Applied Mathematics and Statistics, University of Copenhagen.

Karácsony, Z. (2006), 'A central limit theorem for mixing random fields', *Miskolc Math Notes* **7**(2), 147–160.

Kerscher, M. (2000), Statistical analysis of large-scale structure in the Universe, *in* K. R. Mecke and D. Stoyan, eds, 'Statistical Physics and Spatial Satistics: The Art of Analyzing and Modeling Spatial Structures and Pattern Formation', Vol. 554 of *Lecture Notes in Physics, Berlin Springer Verlag*, pp. 36—71.

Lindsay, B. G. (1988), 'Composite likelihood methods', *Contemp Math* **80**, 221–239.

Matérn, B. (1986), *Spatial Variation*, Vol. 36 of *Lecture Notes in Statistics*, Springer-Verlag, Berlin.

Møller, J. and Waagepetersen, R. P. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman & Hall/CRC, Boca Raton.

Møller, J. and Waagepetersen, R. P. (2007), 'Modern statistics for spatial point processes', *Scand J Stat* **34**(4), 643–684.

Møller, J. (2003), 'Shot noise Cox processes', *Adv Appl Probab* **35**, 614–640.

Møller, J. and Díaz-Avalos, C. (2010), 'Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires', *Scand J Stat* **37**(1), 2–25.

Møller, J. and Ghorbani, M. (2012), 'Aspects of second-order analysis of structured inhomogeneous spatio-temporal point processes', *Stat Neerl* **66**(4), 472–491.

Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998), 'Log Gaussian Cox processes', *Scand J Stat* **25**, 451–482.

Motzkin, T. (1947), 'From among n conjugate algebraic integers, $n-1$ can be approximately given', *Bulletin of the American Mathematical Society* **53**(2), 156–162.

Mrkvička, T. (2014), 'Distinguishing different types of inhomogeneity in Neyman–Scott point processes', *Methodol Comput Appl Probab* **16**(2), 385–395.

Mukhopadhyay, P. (2004), *An Introduction to Estimating Functions*, Alpha Science International, Harrow.

Musmeci, F. and Vere-Jones, D. (1992), 'A space-time clustering model for historical earthquakes', *Ann I Stat Math* **44**(1), 1–11.

Myllymäki, M., Mrkvička, T., Seijo, H. and Grabarnik, P. (2013), 'Global envelope tests for spatial processes', *ArXiv e-prints* .

Neyman, J. and Scott, E. L. (1958), 'Statistical approach to problems of cosmology', *J Roy Stat Soc B* **20**, 1–43.

Ogata, Y. (1998), 'Space-time point-process models for earthquake occurrences', *Ann I Stat Math* **50**(2), 379–402.

Ohser, J. and Mücklich, F. (2000), *Statistical Analysis of Microstructures in Materials Science*, John Wiley & Sons, Chichester.

Parker, J., Sherman, E., van de Raa, M., van der Meer, D., Samelson, L. E. and Losert, W. (2013), 'Automatic sorting of point pattern sets using Minkowski functionals', *Phys Rev E* **88**, 022720–1–022720–8.

Peng, R. D., Schoenberg, F. P. and Woods, J. (2005), 'A space-time conditional intensity model for evaluating a wildfire hazard index', *J Am Stat Assoc* **100**, 26–35.

Prokešová, M. (2010), 'Inhomogeneity in spatial Cox point processes – location dependent thinning is not the only option', *Image Anal Stereol* **29**, 133–141.

Prokešová, M. and Dvořák, J. (2014), 'Statistics for inhomogeneous space-time shot-noise Cox processes', *Methodol Comput Appl Probab* **16**(2), 433–449.

Prokešová, M., Dvořák, J. and Jensen, E. B. V. (2014), 'Two-step estimation procedures for inhomogeneous shot-noise Cox processes'. Submitted.

Prokešová, M. and Jensen, E. B. V. (2013), 'Asymptotic Palm likelihood theory for stationary point processes', *Ann I Stat Math* **65**(2), 387–412.

Ripley, B. D. (1976), 'The second-order analysis of stationary point processes', *J Appl Probab* **13**, 255—266.

Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge.

Schneider, R. (1993), *Convex Bodies: The Brunn-Minkowski Theory*, Cambridge University Press.

Schoenberg, F. P. (2005), 'Consistent parametric estimation of the intensity of a spatial–temporal point process', *J Stat Plan Infer* **128**(1), 79–93.

Serra, J. (1982), *Image Analysis and Mathematical Morphology*, Academic Press, London.

Stoyan, D., Kendall, W. S. and Mecke, J. (1995), *Stochastic Geometry and its Applications*, Wiley Series in Probability and Mathematical Statistics, 2nd edn, Wiley, Chichester.

Tanaka, U., Ogata, Y. and Stoyan, D. (2008), 'Parameter estimation and model selection for Neyman-Scott point processes', *Biometrical J* **50**(1), 43–57.

Thomas, M. (1949), 'A generalization of Poisson's binomial limit for use in ecology', *Biometrika* **36**, 18–25.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Waagepetersen, R. P. (2007), 'An estimating function approach to inference for inhomogeneous Neyman-Scott processes', *Biometrics* **63**, 252–258.

Waagepetersen, R. P. and Guan, Y. (2009), 'Two-step estimation for inhomogeneous spatial point processes', *J R Stat Soc B* **71**, 685–702.