# Semi-parametric models for multivariate point pattern data

Hessellund, Kristian Bjørn

*DOI (link to publication from Publisher):*
10.5278/vbn.phd.eng.00082

*Publication date:*
2020

*Document Version*
Publisher's PDF, also known as Version of record

# SEMI-PARAMETRIC MODELS FOR MULTIVARIATE POINT PATTERN DATA

BY
KRISTIAN BJØRN HESSELLUND

DISSERTATION SUBMITTED 2020

AALBORG UNIVERSITY
DENMARK

# Semi-parametric models for multivariate point pattern data

Ph.D. Dissertation
Kristian Bjørn Hessellund

Dissertation submitted December 14, 2020

| | |
|---|---|
| Dissertation submitted: | December 14, 2020 |
| PhD supervisor: | Professor Rasmus Waagepetersen<br>Aalborg University |
| Assistant PhD supervisor: | Professor Jesper Møller<br>Aalborg University |
| PhD committee: | Associate Professor Torben Tvedebrink (chairman)<br>Aalborg University |
| | Professor Erik Thorlund Parner<br>University of Aarhus |
| | Associate Professor Ottmar Cronie<br>University of Gothenburg |
| PhD Series: | Faculty of Engineering and Science, Aalborg University |
| Department: | Department of Mathematical Sciences |

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-855-1

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk



Printed in Denmark by Rosendahls, 2021

# Abstract

This thesis considers spatial point patterns, which are realizations of stochastic processes called spatial point processes. Point patterns are collections of data points, where each data point indicates the location of the point. This thesis only consider point pattern data that are observed in two-dimensional planar regions, but the developed methodology in this thesis can handle point pattern data that are observed in a general $d$-dimensional space. The topic of point processes is applied in many types of scientific areas, including epidemiology, agriculture, criminology or pathology, where the data points in these cases are given by the locations of infected persons in a region, trees in a forest, street crimes in a city or cells in a metastasis. Besides knowing the locations of the data points, additional information may be linked to each data point. For instance, a data point may also indicate that it is of a specific type. The data are then expanded to be a multivariate point pattern, which is a realization of a multivariate point process. The majority of the current literature regarding multivariate point pattern analysis is restricted to the bivariate case, but in the recent years new multivariate point processes have been developed to analyze multivariate point patterns. In some cases, a point pattern occurs in a very complex heterogeneous environment, which makes it difficult to model the point pattern using point processes. One example is a point pattern created from locations of different types of street crimes committed in a city. Modeling such a point pattern using a point process model is complex, since the locations of the street crimes depend in a complex way on the urban structure of the city and the population density. This thesis focuses on developing new methodology to analysis multivariate point patterns that are observed in such complex heterogeneous environments.

When analyzing a multivariate point pattern an obvious starting point is to study the first order properties of each point type. To study the first order properties of a point process the intensity function is used, which basically determines the probability of observing a point at any given location. The intensity function can thereby be used to study where the points are most dense in a point pattern. If the intensity function is constant, the point process is called homogeneous. Otherwise the point process is called inhomogeneous.

In this thesis we only consider inhomogeneous multivariate point processes. If additional information is available, like spatial covariates, one could assume a parametric model for the intensity function and then study how the point pattern is affected by such covariates. In paper A a semi-parametric model is assumed for the intensity functions. The semi-parametric model for the intensity functions consists of a complex non-parametric component, that is common to all types of points, and a type-specific component that depends on spatial covariates. We use the assumption of a semi-parametric model to study how spatial covariates affects the occurrences of a multivariate point process. Paper A analyzes how different types of street crimes in Washington D.C. depend on demographic and socio-economic spatial covariates, where the non-parametric component is assumed to take into account the urban structure of the city and the population density. Paper A proposes methodology to infer the effects of spatial covariates without specifying the complex non-parametric component.

Next step in a multivariate point pattern analysis is usually to study the second order properties between the types of points and within each point type. More specifically, one studies if the points have a positive, negative or no spatial dependence between each other. To conclude on the spatial dependence within each point type (and between two types of points), one can apply different kind of functions called (cross) summary statistics. One popular summary statistic is the pair correlation function (PCF). For a given distance between two locations, the PCF describes how the presence of a point in the first location affect the probability of observing a point in the second location. This heuristic interpretation of the PCF can easily be extended to the cross PCF by letting the two points be of different types. To analyze the dependence between the points, the cross PCF and PCF can be estimated non-parametrically. Alternatively, one can assume a multivariate point process model for the multivariate point pattern, in which case the (cross) PCFs are given by parametric models. In paper B the multivariate point pattern is assumed to be a realization of a so called multivariate log Gaussian Cox process (LGCP). LGCP models are point processes with stochastic intensity functions that are suitable for modeling clustered point patterns. Moreover, paper B still assumes a semi-parametric model for each intensity function, which now consists of a complex non-parametric component, a parametric component that depends on covariates and a stochastic component. Paper B proposes methodology to estimate the parameters in the LGCP without estimating the complex non-parametric component in the intensity functions. To exemplify the methodology, Paper B analyzes the second order properties of different street crimes in Washington D.C. along with different types of cells in a lymph node metastasis.

This thesis contains two parts. The first part is an introduction to relevant state-of-the-art, background theory regarding point processes and composite

likelihood estimating functions, which are specific types of likelihood functions that are well-suited for statistical inference of the point processes. These topics are presented to the reader with the intention, that the reader more easily can understand the content and relevance of the second part of the thesis, which is a collection of two papers.

# Resumé

Denne afhandling beskæftiger sig med data af typen punktmønstre, som er realisationer af stokastisk processer ved navn punktprocesser. Punktmønstre er en samling af datapunkter, hvor hvert datapunkt typisk angiver placeringen eller tidspunktet af punktet. Denne afhandling afgrænser sig til kun at fokusere på rumlige punktmønstre der observeres i afgrænsede områder af det to-dimensionelle plan. Metoderne, som er foreslået i denne afhandling, kan dog håndtere punktmønstre, der er observeret i et generelt $d$-dimensionelt plan. Emnet om punktprocesser anvendes i mange typer af videnskabelige fagområder, blandt andet epidemiologi, landbrug, kriminologi eller patologi, hvor datapunkterne her er givet ved lokationerne af smittede personer i en region, træer i en skov, forbrydelser i en by eller celler i en metastase. Udover placeringerne af datapunkterne kan man også koble mere information til hvert datapunkt. Eksempelvis kan hvert datapunkt også angive, at det er af en bestemt type. Dermed udvides datamængden til at være et multivariat punktmønster, hvilket kan betragtes som en realisation af en multivariat punktproces. Størstedelen af den eksisterende litteratur omhandlende analyse af multivariate punktmønstre er dog restringeret til det bivariate tilfælde, mens der i de senere år er udviklet nye multivariate punktprocesser til at analysere multivariate punktmønstre. Nogle punktmønstre forekommer i et meget komplekst heterogent område, hvorfor det er meget svært at modellere punktmønstrene ved hjælp af punktprocesser. Et eksempel herpå er et punktmønster dannet af lokationerne af forskellige typer forbrydelser begået i en by. Det er et komplekst problem at modellere et sådant punktmønster ved hjælp af en punktproces, da placeringerne af de forskellige forbrydelser afhænger af byens urbane struktur og populationstætheden. Denne afhandling fokuserer på at udvikle nye metoder til at analysere multivariate punkmønstre, som er observeret i komplekse heterogene områder.

Når man analyserer et multivariat punktmønster, vil man ofte starte med at studere førsteordens egenskaberne for hver punkttype. Til at studere førsteordens egenskaberne af en punktproces anvendes intensitetsfunktionen, som beskriver sandsynligheden for at observere et punkt i enhver given placering. Intensitetsfunktionen kan dermed anvendes til at fortælle, hvor

punkterne er mest tætte i en punktproces. Hvis intensitetsfunktionen er konstant, så kaldes punktprocessen for homogen. Hvis det omvendte tilfælde gør sig gældende, så kaldes punktprocessen for inhomogen. I denne afhandling fokuserer vi kun på inhomogene multivariate punktprocesser. Hvis der er yderligere information tilgængelig i form af rumlige kovariater, så kan man antage en parametrisk model for intensitetsfunktionen og dermed undersøge, hvordan et punktmønster afhænger af disse kovariater. I artikel A antages der en semi-parametrisk model for intensitetsfunktionerne. Den semi-parametriske model for intensitetsfunktionerne består af en kompleks ikke-parametrisk komponent, der er fælles for alle punkttyper, og en typespecifik komponent som afhænger af tilgængelige rumlige kovariater. Vi bruger antagelsen om en semi-parametrisk model til at analysere, hvordan rumlige kovariater påvirker forekomsten af en given multivariat punktproces. Et eksempel herpå kan findes i artikel A. Her analyseres hvordan forskellige typer af forbrydelser i Washington D.C. afhænger af tilgængelige demografiske og socio-økonomiske rumlige kovariater, hvor det antages at den komplekse ikke-parametriske komponent i intensitetsfunktionerne tager højde for byens urbane struktur og populationstætheden. Artikel A foreslår en metode til at kunne estimere parametrene for kovariaterne uden at skulle specificere den komplekse ikke-parametriske komponent.

Næste skridt i analysen af et multivariat punktmønster er ofte at studere andenordens egenskaberne mellem punkttyperne og indenfor hver punkttype. Her undersøger man, om punkterne har en positiv, negativ eller ingen rumlig afhængighed mellem hinanden. Til at afgøre den rumlige afhængighed mellem punkterne kan forskellige (kryds) summary statistics anvendes. En populær summary statistic er parkorrelationsfunktionen (PCF). For en given afstand mellem to lokationer, så beskriver PCF'en, hvordan tilstedeværelsen af et punkt i den ene lokation påvirker sandsynligheden for at observere et punkt i den anden lokation. Denne heuristiske beskrivelse af PCF'en kan nemt udvides til kryds PCF'en ved at lade de to punkter være af forskellig type. Til at analysere afhængigheden mellem punkterne kan kryds PCF'en og PCF'en estimeres ikke-parametrisk. Alternativt kan man antage en multivariat punktproces model for det multivariate punktmønster, hvorved (kryds) PCF'erne er givet på en parametrisk form. I artikel B antages et multivariat punktmønster at være en realisation af en såkaldt multivariat log Gaussisk Cox punktproces (LGCP). LGCP modeller er punktprocesser med en stokastisk intensitetsfunktion, der er velegnet til at modellere punktmønstre, hvor punkterne klynger sig sammen. Artikel B antager desuden stadig en semi-parametrisk model for hver intensitetsfunktion, som nu består af en kompleks ikke-parametrisk komponent, en parametriske komponent som afhænger af kovariater og en stokastisk komponent. Artikel B foreslår en metode til at estimere parametrene i LGCP'en uden at skulle estimere den komplekse ikke-parametriske komponent i intensitetsfunktionerne. Denne

metode er eksemplificeret i artikel B, hvor andenordens egenskaberne analyseres for forskellige forbrydelser i Washington D.C. samt forskellige typer af celler i en lymfemetastase.

Afhandlingen indeholder to dele. Den første del består af en introduktion til relevant state-of-the-art, baggrundsteori om punktprocesser og composite likelihood estimationsfunktioner, som er specifikke typer af likelihood funktioner, der er velegnet til at lave statistisk inferens af punktprocesser. Disse emner præsenteres for læseren med den hensigt, at læseren nemmere kan forstå indholdet og relevansen af andel del af afhandlingen, som er en samling af to artikler.

# Contents

# Preface

This thesis contains the scientific research conducted by me and my collaborators during my employment as a Ph.D. student at the Department of Mathematical Sciences at Aalborg University. My work is supported by "Centre of Stochastic Geometry and Advanced Bioimaging", which is funded by the Villum Foundation.

The objective of this thesis is to develop methodology to analyze first and second order properties of multivariate point pattern data with complex intensity functions. The developed methodology is applicable for various types of multivariate point pattern data. However, the methodology is demonstrated to be particularly well-suited for street crime data and cancer cell data, since such data sets can be recognized as a realizations of multivariate point processes with complex intensity functions. The thesis consists of two parts: Part I introduces relevant state-of-the-art, data problems that relate to this thesis, background theory on point processes and theory regarding composite likelihood estimating functions. Part II is a collection of two papers:

A: Hessellund, K. B., Xu, G., Guan, Y. and Waagepetersen, R. (2019). Semiparametric multinomial logistic regression for multivariate point pattern data. *Journal of the American Statistical Association*, accepted for publication.

B: Hessellund, K. B., Xu, G., Guan, Y. and Waagepetersen, R. (2020). Second order semi-parametric inference for multivariate log Gaussian Cox processes. Submitted to *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. Available on arXiv:2012.02155.

Notice that some background material are presented in both papers, which makes the papers completely self-contained and thereby independent of each other.

I would like to thank my supervisor Rasmus Waagepetersen for always competent and motivating supervision. The main part of my knowledge in the topic of statistics can be attributed to you. Our travels together to conferences and especially Miami have been a true pleasure. Moreover, I would like

to thank Yongtao Guan and Ganggang Xu for excellent cooperation and for hosting me during my stay in Miami. I would also like to thank my fellow Ph.D. students for our social activities. At last, a special thanks to my Anne for taking care of our daughter Vera, cheering me with helpful comments and supporting me during the loss of my mother. Without you I would not have been able to complete my three years as a Ph.D. student.

Kristian Bjørn Hessellund
Aalborg University, December 14, 2020

# Part I

# Introduction

# Background

## 1 Introduction

In this thesis the data of interest are two-dimensional spatial point patterns, which are finite collections of data points within bounded planar regions called observation windows. To gain knowledge of what mechanisms that generate the occurrence of the points we consider a point pattern as a realization of a spatial point process. There exists several types of spatial point process models that have the ability of generating point patterns with either positive, negative or no dependence between the points. The detection of such characteristics in a point pattern is a main part of a point pattern analysis. The topic of point processes is applied in a broad range of scientific fields, including epidemiology (Diggle et al. (2005); Liang et al. (2017)), forestry and plant ecology (Brix and Møller (2001); Grabarnik and Särkkä (2009)) or astrophysics (Stoica et al. (2007)). To get a characterization of a point pattern one could estimate a so-called summary statistic using non-parametric methods. Alternatively, one may assume a point process model for the point pattern and then infer the point pattern characteristics parametrically. In Section 3 we describe some relevant point process models, while Section 4 contains a method to infer the point process models.

In this thesis we focus on the setup of multivariate point processes, i.e. we consider point pattern data where the points can be of different type. When analyzing a multivariate point pattern one would initially like to study the first order properties of each point type. The intensity function is used to summarize the first order properties of a point process. The intensity function essentially determines the probability of observing a point at any given spatial location. A more detailed description of the intensity can be found in Section 2.1. If the intensity is constant, the point process is called homogene. Otherwise the point process is called inhomogene. In a first order point pattern analysis one may try to gain knowledge of where the point pattern is most dense. There exist different kernel methods for the intensity to study the density of a point pattern, including Diggle (1985) and Baddeley et al.

(2000). Guan (2008) extended the method by Diggle (1985) and proposed a non-parametric estimator for the intensity that depends on spatial covariates. However, kernel smoothing methods depend heavily on a smoothing parameter called bandwidth. Diggle (1985), Guan (2008) and Cronie and Van Lieshout (2018) developed bandwidth selection methods that depend on knowledge of the observation window. These bandwidth selection methods can be problematic to apply in the cases where the observation window is highly irregular or even unknown. As an alternative to kernel methods of the intensity, one may study how the point process is affected by covariates, if spatial covariates are available. To study the relation between a point process and covariates, the intensity is assumed to follow a parametric model that depends on the spatial covariates, in which case there exist methodology to estimate the effect of the covariates (Waagepetersen (2007); Choiruddin et al. (2018)). In more complex situations one may presume a more complex model for the intensity function. Diggle (1990) assume a multiplicative model for the intensity in order to describe possible raised incidence near a prespecified point. The assumption of a multiplicative intensity was further applied in a case-control study (Diggle and Rowlingson, 1994) and for analysis of golden plover birds (Guan et al., 2008).

Subsequently, one may like to study the second order properties between and within each type of points. A key question is whether the points tend to spatially depend on each other, and if so, is the dependency positive or negative. To study the spatial dependence between points of the same type, a simple approach is to conduct an exploratory analysis of the point pattern based on non-parametric estimates of a second order summary statistic. In particular, Ripley's $K$-function (Ripley, 1976) or the pair correlation function (PCF) (Møller and Waagepetersen, 2004) are popular choices of second order summary statistics. For a given distance and a constant intensity, the $K$-function determines the expected number of points within that distance from an arbitrary point in the point process. For a given distance between two locations, the PCF basically describes how the presence of a point in the first location affects the probability of observing a point in the second location. A detailed description of the PCF can be found in Section 2.1.

To study the spatial dependence between two types of points one may conduct a simple exploratory analysis of the bivariate point pattern using a non-parametric estimate of the cross pair correlation function (cross PCF). The cross PCF is a straightforward extension of the PCF to the bivariate case. A detailed description of the cross PCF can be found in Section 2.2. However, the non-parametric estimator for the cross PCF suffers from strong bias at small spatial lags when positively dependent point patterns are studied (Stoyan and Stoyan, 1994). When analyzing second order properties of an inhomogeneous point process, the second order inference depends on knowledge of the intensity. If spatial covariates are not available to model the inten-

sity, then the intensity must be estimated using kernel smoothing methods. Since kernel smoothing methods depends heavily on a bandwidth, different conclusions of a point pattern analysis might occur depending on the choice of bandwidth. With this aspect in mind, the method of modeling the intensity function parametrically is more favorable, since the parametric model does not depend on a bandwidth. On the other hand, a non-parametric estimate of the intensity is less restrictive in the sense, that the intensity is not assumed to follow a specific parametric model. As pointed out by Diggle et al. (2007) non-parametric estimation of both the intensity and the $K$-function using the same data is an ill-posed problem. As an example, a single realization of a homogeneous Cox process (see Section 3.2 for details) is indistinguishable from a realization of an inhomogeneous Poisson process (see Section 3.1 for details). Instead, Diggle et al. (2007) accommodate this issue by case-control methodology, where the intensity of the case process is assumed to be proportional to the intensity of the control process. Hence the intensity of the case process is estimated using a non-parametric estimate of the control intensity.

Another difficulty for multivariate point process analysis is that cross summary statistics only consider pairs of point types. Thus, the complexity of a multivariate point pattern analysis increases quadratic by the number of point types, when the point pattern data is analyzed non-parametrically. This phenomena is referred to as curse of dimensionality. In the light of that, the literature regarding multivariate point pattern analysis is mostly restricted to the bivariate case (Diggle and Milne (1983); Högmander and Särkkä (1999)). For this reason, a proper multivariate point process model contributes to the analysis, since it can give more sparse characterizations of the dependence structure between the point types. In the recent years there has been a growing interest in analysing multivariate point pattern data using multivariate point process models. Grabarnik and Särkkä (2009) use a multivariate Gibbs point process with hierarchical interactions and estimate the parameters using pseudo-likelihood. Rajala et al. (2018) model a multivariate point pattern using a multivariate Gibbs point process with additional group lasso for model selection. Waagepetersen et al. (2016) model a multivariate point pattern using a multivariate log Gaussian Cox process and estimate the parameters using a least squares method. Recently, Choiruddin et al. (2019) modified the estimation approach in Waagepetersen et al. (2016) by using regularization techniques in order to do model selection for a large number of parameters.

In some cases a multivariate point pattern is observed in a highly heterogeneous environment, which makes the intensity functions complex to model. One example of such data is the locations of different types of street crimes in Washington D.C. (see Section 1.1). Another example is the locations of different cell types in a lymph node metastasis (see Section 1.1). The existing methodology for multivariate point pattern analysis with complex

intensity functions is insufficient. This thesis aims to fill this gap by developing new methods to conduct appropriate first and second order analysis of such point pattern data. The following chapter serves to give background knowledge for better understanding of paper A and paper B along with a brief summary of the findings in the papers. Section 1.1 briefly describes some data of interest and what kind of difficulties that are connected to the data. Section 2 describes general uni- and multivariate point process theory. Section 3 describes some relevant point process models. Section 4 describes the theory of estimating functions with a focus on composite likelihood estimating functions. At last, Section 5 summarizes the main findings in paper A and paper B.

## 1.1 Data examples

**Washington D.C. street crimes**

Police departments in cities all over the world collect street crime data on a daily basis. The data is often public available on the internet to inform the citizens what type of crime is committed and where the crime took place. In a fixed span of time one can aggregate the locations of street crimes and consider the data as a spatial point pattern. Methods to analyze street crime data are quite important from the police departments perspective, since such methods can help reducing the occurrences of street crime but also to allocate policing resources more properly. One popular tool to analyze street crimes is to identify the hot spots, i.e the micro-places where street crimes tend to cluster. The most common way to identify the hot spots is by using non-parametric methods (Ratcliffe (2004); Gorr and Lee (2015)). Also point processes have been applied to model the behaviour of street crimes. In particular, so-called self-exciting point processes have been applied to model burglaries (Mohler et al. (2011); Reinhart and Greenhouse (2018)) and gang rivalry (Egesdal et al., 2010), since self-exciting point processes have the ability to model the behaviour that crimes trigger new crimes.

In Paper A and Paper B we consider locations of street crimes committed in Washington D.C. aggregated over January and February 2017. We consider the 5378 data points as a six-variate point pattern, where the six types with numbers in parantheses are: Burglary (259), Assault with deadly weapon (332), Motor vehicle theft (335), Theft from automobile (1832), Robbery (366) and Other theft (2254). Figure 1 show the street crimes committed in Washington D.C. along with a map of the city. For point pattern data like street crimes, the intensity function is rather complex due to the dependence on the layout of the city and the population density. Moreover, the intensity can vary considerably when moving from one street to another. When analyzing the street crime locations two key questions occur: 1) How does the street crimes

relate to some demographical spatial covariates? 2) Does the different types of street crimes spatially depend on each other, and if so, is the dependence positive or negative?

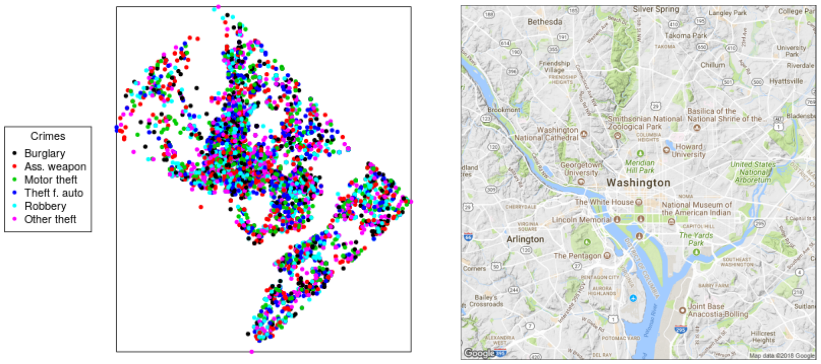


**Fig. 1:** Left: Street crime locations in Washington D.C. Right: Map of Washington D.C.

To accommodate the issue regarding the layout of the city, one possibility is to build a linear network space, where the edges correspond to streets. Such setup has already been applied to determine "hot routes" in criminology (Tompson et al., 2008) but also in the field of neurology (Baddeley et al., 2014). Clearly, when the space is changed from a planar regular region to a linear network, the proper distance between points is not Euclidean distance but rather the distance along the network. As showed in Okabe and Yamada (2001), a summary statistic can be defined on a linear network if the Euclidean distance is replaced with the shortest path distance. Hence we can address question 2) using non-parametric methods. Alternatively, Anderes et al. (2018) developed isotropic covariance functions for Gaussian random fields on linear networks, where these covariance functions are isotropic in the sense that they depend on the so-called resistance metric. Thus, one could build log Gaussian Cox point process models on a linear network from Gaussian random fields with the covariance functions proposed by Anderes et al. (2018) to address the two key questions. However, difficulties occur when a linear network is applied to approximate the point space. For instance, a linear network is not adequate for approximating the street map since the crimes can also occur in parks or at open spaces. Thereby, a linear network does not take into account where in the park or the open space the crimes take place.

Instead we address the two key questions by assuming a semi-parametric multiplicative model for the intensity function. More specifically, we assume that the intensity for all types of street crimes has a complex non-parametric component and a parametric component for each crime type that depends on

covariates. The non-parametric component takes into account the layout of the city and population density. Section 4.3 and Section 4.4 briefly describe the methodologies that are applied to analyze Washington D.C. street crimes.

**Lymph node metastasis**

The recent development of super-resolution microscopy techniques now provide spatial locations of cells and molecules at the nanoscale. The accessible microscopy data can be analyzed using spatial point processes. The literature on spatial point process analysis of microsropy data is almost restricted to simple non-parametric analyses. Rossy et al. (2013) study possible clustering of Lck cells of the immune system using non-parametric estimates of the $K$-function. Yong and Hancock (2018) use non-parametric estimates of the (cross) $K$-function to study the spatial dependence between and within types of proteins in a plasma membrane. In those papers the analyses are insufficient due to the neglection of inhomogeneity in the point pattern data. In connection to parametric analysis, Bell and Grunwald (2004) used a Strauss point process with mixed effects to study replicated point pattern data from three groups: normal, schizo-affective and schizophrenic. The point pattern data consisted of locations of pyramidal neurons in the cortex in the human brain. This data set has earlier been analyzed by Diggle et al. (1991) and later by Mateu (2001).

In Paper B we consider locations of four types of cells in a lymph node metastasis. The four types of cells are Stroma cells (Stroma), Cytotoxic T-lymphocytes cells (CD8), Hypoxic tumor cells (Hypoxic) and Normoxic tumor cells (Normoxic). When taking a closer look at Figure 2 we see a clear

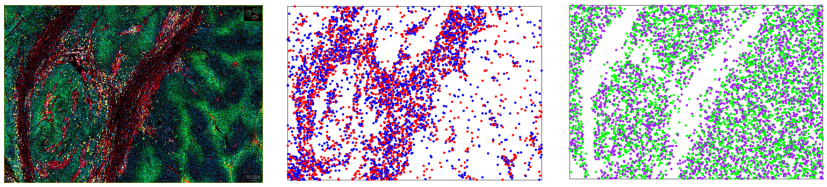

**Fig. 2:** Left: Fluorescence image of lymph node metastasis. Middle: locations of stroma (red) and CD8 (blue) cells. Right: locations of Hypoxic (purple) and Normoxic (green) cells

segregation between the two bivariate point patterns (Stroma, CD8) and (Hypoxic, Normoxic). Thus, we conduct two distinct bivariate analyses in paper B. In both analyses we study the second order properties of each bivariate point pattern. In each of the two bivariate point patterns we see a clear large-scale trend, which is difficult to model with simple parametric models. In addition, no covariate information is available. Section 4.4 briefly describes the methodology that is applied to the lymph node metastasis data.

# 2 Point process theory

This section briefly describes point process theory with a focus on the $n$th order joint intensity in the univariate case and the $(m,n)$th order cross intensity in the multivariate case.

## 2.1 Univariate point processes

A univariate point process $X$ on a metric space $S$ is a random countable subset of $S$. A more rigorous definition of point processes can be found in Daley and Vere-Jones (2003). We denote by $x$ a realization of a point process $X$ and $n(x)$ is the cardinality of $x$. We only focus on point processes whose realizations are locally finite subsets of $S$, i.e. $n(x \cap B)$ is finite for $B \subseteq S$ with $|B| < \infty$. We usually denote by $\mathbf{u}$ or $\mathbf{v}$ points in $S$. The construction of $X$ can be carried out for a general metric space $S$. We, however, restrict the space to be $S = \mathbb{R}^d$ with associated Euclidean distance $d(\mathbf{u},\mathbf{v}) = ||\mathbf{u}-\mathbf{v}||$ as the metric. In practice we observe $X$ in a bounded observation window $W \subset \mathbb{R}^d$.

For characterization of a point process we introduce certain functions called $n$th order joint intensity functions and denote them by $\lambda^{(n)}$, where $n \geq 1$. For each $n \geq 1$ assume that $X$ has a $n$th order joint intensity function $\lambda^{(n)}$ that fulfills:

$$\mathbb{E} \sum_{\mathbf{u}_1,\ldots,\mathbf{u}_n \in X}^{\neq} \mathbf{1}\left[\mathbf{u}_1 \in B_1, \ldots, \mathbf{u}_n \in B_n\right] = \int_{\prod_{k=1}^n B_k} \lambda^{(n)}(\mathbf{u}_1,\ldots,\mathbf{u}_n)\mathrm{d}\mathbf{u}_1 \cdots \mathrm{d}\mathbf{u}_n, \tag{1}$$

where $B_k \subseteq \mathbb{R}^d$ for $k = 1,\ldots,n$. The inequality sign $\neq$ over the summation means that $\mathbf{u}_1,\ldots,\mathbf{u}_n$ are pairwise distinct. For $n = 1$ we let $\lambda^{(1)} = \lambda$ and the function is then the intensity function. Using heuristic arguments, we can think of $\lambda^{(n)}(\mathbf{u}_1,\ldots,\mathbf{u}_n)\prod_{k=1}^n \mathrm{d}\mathbf{u}_k$ as the joint probability for observing $n$ points from $X$ jointly in $n$ infinitesimally small balls with centres $\mathbf{u}_1,\ldots,\mathbf{u}_n$ and volumens $\mathrm{d}\mathbf{u}_1,\ldots,\mathrm{d}\mathbf{u}_n$. We can generalize (1) using standard measure theoretical arguments for any non-negative function $h$ on $(\mathbb{R}^d)^n$:

$$\mathbb{E} \sum_{\mathbf{u}_1,\ldots,\mathbf{u}_n \in X}^{\neq} h(\mathbf{u}_1,\ldots,\mathbf{u}_n) = \int_{(\mathbb{R}^d)^n} h(\mathbf{u}_1,\ldots,\mathbf{u}_n)\lambda^{(n)}(\mathbf{u}_1,\ldots,\mathbf{u}_n)\mathrm{d}\mathbf{u}_1 \cdots \mathrm{d}\mathbf{u}_n. \tag{2}$$

The equality (2) is a central result in this thesis and in general point process theory. It is often referred to as Campbells formulae.

Although there exists several summary functions to describe the higher order properties of a point process (see e.g. Møller and Waagepetersen (2004) for an expansive overview), we only consider the normalized $n$'th order joint intensity function $g^{(n)}$ in paper A and paper B. This summary function is

central for characterization of $n$th order properties of a point process, and is for $n \geq 2$ defined as:

$$g^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) = \lambda^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) / \lambda(\mathbf{u}_1) \cdots \lambda(\mathbf{u}_n), \tag{3}$$

provided that $\lambda^{(n)}$ and $\lambda$ exists. We define $g^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) = 0$ if $\lambda(\mathbf{u}_i) = 0$ for some $i = 1, \ldots, n$. A special case is $n = 2$, where the function is then the pair correlation function (PCF) and is denoted by $g^{(2)} = g$. Using heuristic arguments, we can interpret $\lambda(\mathbf{u}_2)g(\mathbf{u}_1, \mathbf{u}_2)$ as the conditional intensity of observing a point from $X$ in $\mathbf{u}_2$, given that $\mathbf{u}_1 \in X$. Thereby, the PCF reveals how the presence of a point at $\mathbf{u}_1$ affects the intensity of further points of $X$ at $\mathbf{u}_2$.

In general, if $\lambda(\cdot) = \lambda$ is constant, we say that $X$ is a homogeneous point process. Otherwise, we say that $X$ is an inhomogeneous point process. Furthermore, we say that $X$ is a stationary point process if its distribution is invariant under translations. Moreover, we say that $X$ is isotropic if its distribution is invariant under rotations around the origin in $\mathbb{R}^d$. Consequently, an isotropic $X$ has a second order joint intensity that depends on the distance between the points, i.e. $\lambda^{(2)}(\mathbf{u}_1, \mathbf{u}_2) = \lambda^{(2)}(\|\mathbf{u}_1 - \mathbf{u}_2\|)$. Following Baddeley et al. (2000) a less restrictive property is second order intensity reweighted stationary (SOIRS), that is if $X$ has a second order joint intensity that depends on the difference between pairs of points, i.e. $\lambda^{(2)}(\mathbf{u}_1, \mathbf{u}_2) = \lambda^{(2)}(\mathbf{u}_1 - \mathbf{u}_2)$.

In point process theory we consider a density for one point process with respect to another. Let $X_1$ and $X_2$ be point processes defined on the same space $S$ and define $N_{lf} = \{x \in S \colon n(x \cap B) < \infty \text{ for all bounded } B \subseteq S\}$. Following Møller and Waagepetersen (2004), the distribution of $X_1$ is absolutely continuous with respect to the distribution of $X_2$ if and only if $\mathrm{P}(X_1 \in F) > 0$ implies $\mathrm{P}(X_2 \in F) > 0$ for $F \subseteq N_{lf}$. If we assume that $\mathrm{P}(X_1 \in F) > 0$ implies $\mathrm{P}(X_2 \in F) > 0$, then by Radon-Nykodym theorem (see Billingsley (1995)), there exists a function $f$ that fulfills:

$$\mathrm{P}(X_1 \in F) = \mathrm{E}\left[1[X_2 \in F] f(X_2)\right].$$

The function $f$ is then called the density of $X_1$ with respect to $X_2$. Examples of such densities are considered in Section 3, where some relevant point process models are introduced.

## 2.2 Multivariate point processes

In order to extend the univariate setup for point processes to a multivariate setup, consider a so-called marked point process. A marked point process $X$ is a point process on $\mathbb{R}^d$ with random "marks" $m_{\mathbf{u}} \in M$ attached to each point $\mathbf{u} \in X$, where $M$ is a given space. The space $M$ can be of different type, however we focus on the space $M = \{1, \ldots, p\}$, i.e. the marks specify $p$

different types of points. In this context, $X$ then becomes a multi-type point process. Equivalently, we can define $\mathbf{X} = (X_1, \ldots, X_p)$ as a multivariate point process, where each $X_i$ is a univariate point process and the indices defines the types of points.

Besides characterizing the $n$th order properties of each point process $X_i$ using $\lambda_i^{(n)}$, another key question is how the point processes interact with each other. To characterize the interaction between a pair, say $X_i$ and $X_j$, we introduce the $(m,n)$th order cross intensity function and denote it by $\lambda_{ij}^{(m,n)}$. For each $m, n \geq 1$ we assume that the pair $(X_i, X_j)$ has a $(m,n)$th order cross intensity function $\lambda_{ij}^{(m,n)}$ that fulfills:

$$\begin{aligned}
&\mathbb{E} \sum_{\mathbf{u}_1,\ldots,\mathbf{u}_m \in X_i}^{\neq} \sum_{\mathbf{v}_1,\ldots,\mathbf{v}_n \in X_j}^{\neq} \mathbf{1}\left[\mathbf{u}_1 \in A_1, \ldots, \mathbf{u}_m \in A_m, \mathbf{v}_1 \in B_1, \ldots, \mathbf{v}_n \in B_n\right] \\
&= \int_{\prod_{k=1}^m A_k} \int_{\prod_{l=1}^n B_l} \lambda_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) \mathrm{d}\mathbf{u}_1 \cdots \mathrm{d}\mathbf{u}_m \mathrm{d}\mathbf{v}_1 \cdots \mathrm{d}\mathbf{v}_n,
\end{aligned}$$

where $A_k, B_l \subseteq \mathbb{R}^d$ for $k = 1, \ldots, m$ and $l = 1, \ldots, n$. When $i = j$ and $\mathbf{u}_k \neq \mathbf{v}_l$ for $k = 1, \ldots, m$ and $l = 1, \ldots, n$, the function is reduced to $\lambda_{ii}^{(m,n)} = \lambda_i^{(m+n)}$. Moreover, when $m, n = 1$ and $i \neq j$ the function is referred to as the cross intensity and it is denoted by $\lambda_{ij}^{(1,1)} = \lambda_{ij}$. We can think of $\lambda_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) \prod_{k=1}^m \mathrm{d}\mathbf{u}_k \prod_{l=1}^n \mathrm{d}\mathbf{v}_l$ as the joint probability for observing $m$ points from $X_i$ and $n$ points from $X_j$ jointly in $n+m$ infinitesimally small balls with centres $\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n$ and volumens $\mathrm{d}\mathbf{u}_1, \ldots, \mathrm{d}\mathbf{u}_m$, $\mathrm{d}\mathbf{v}_1, \ldots, \mathrm{d}\mathbf{v}_n$.

Several summary functions for univariate point processes can be extended to describe higher order cross properties of a pair $(X_i, X_j)$. However, we focus on the normalized $(m,n)$th order cross intensity function, denoted by $g_{ij}^{(m,n)}$, when characterizing the $(m,n)$th order cross dependence between $(X_i, X_j)$. This cross summary function is defined as:

$$g_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) = \frac{\lambda_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n)}{\lambda_i(\mathbf{u}_1) \cdots \lambda_i(\mathbf{u}_m) \lambda_j(\mathbf{v}_1) \cdots \lambda_j(\mathbf{v}_n)},$$

provided that $\lambda_{ij}^{(m,n)}$, $\lambda_i$ and $\lambda_j$ exists. We define $g_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$ if $\lambda_i(\mathbf{u}_k) = 0$ for $k = 1, \ldots, m$ or $\lambda_j(\mathbf{v}_l) = 0$ for $l = 1, \ldots, n$. When $m, n = 1$ and $i \neq j$ the function is then the cross pair correlation function (cross PCF) and it is denoted by $g_{ij}^{(1,1)} = g_{ij}$. Heuristic arguments to interpret the PCF can easily be extended for the interpretation of the cross PCF. Notice, if $X_i$ and $X_j$ are independent then $g_{ij} \equiv 1$. Thus, we can use the cross PCF as a summary

function to determine whether $X_i$ and $X_j$ are positively spatially correlated ($g_{ij} > 1$) or negatively spatially correlated ($g_{ij} < 1$).

From the univariate setup concepts like stationarity, isotropy and SOIRS can in a straightforward manner be extended to the multivariate setup. In addition, Campbells formulae can easily be extended for the cross joint intensity function.

# 3 Point process models

In the last decades there has been an increasing focus on developing new spatial point process models (see review by Møller and Waagepetersen (2017) and their references) due to the enhance in computational power and computational methods, e.g. Markov Chain Monte Carlo (MCMC). Popular choices of point process models include Markov point processes to model point patterns with interaction (van Lieshout, 2000), Determinantal point processes to model repulsive point patterns (Lavancier et al., 2015) and log Gaussian Cox point processes to model clustered point patterns (Møller et al., 1998). This section introduce the relevant point process models for paper A and paper B.

## 3.1 Poisson point processes

The Poisson point process plays a key role in spatial statistics as the model of complete spatial randomness. Hence, Poisson processes are often considered as "reference" models, when univariate point processes are studied. Furthermore, Poisson processes also play a fundamental role when more complex point process models are constructed (see Section 3.3). A basic result for a Poisson process shows that $\lambda^{(n)}(\mathbf{u}_1, \dots, \mathbf{u}_n) = \prod_{k=1}^{n} \lambda(\mathbf{u}_k)$, hence Poisson processes only depend on their first order properties. Thus, $g(\mathbf{u}_1, \mathbf{u}_2) = 1$ for a Poisson process. Given a point process $X$ with $g(\mathbf{u}_1, \mathbf{u}_2) > 1$ (or $g(\mathbf{u}_1, \mathbf{u}_2) < 1$), this indicates that pairs of points from $X$ are more (or less) likely to occur at locations $\mathbf{u}_1$ and $\mathbf{u}_2$ relative to a Poisson process with same intensity as $X$. Consider $X_1$ and $X_2$ as finite Poisson processes on $S$ with intensities $\lambda_1$ and $\lambda_2$. Assume that the distribution of $X_1$ is absolutely continuous with respect to the distribution of $X_2$. It can then be verified, that $X_1$ has a density $f$ respect to $X_2$, where $f$ is given by:

$$f(x) = \exp\left(\int \lambda_2(\mathbf{u}) - \lambda_1(\mathbf{u}) \mathrm{d}\mathbf{u}\right) \prod_{\mathbf{u} \in x} \lambda_1(\mathbf{u}) / \lambda_2(\mathbf{u}), \tag{4}$$

where $x \subset S$ is a finite point pattern.

A multivariate Poisson process can be constructed in a straightforward manner by defining a marked Poisson process with $M = \{1, \dots, p\}$ as the mark space.

## 3.2 Cox point processes

Cox point processes are a class of point processes that generalize Poisson point processes by the property that the intensity functions are stochastic itself. In that sense Cox processes can be understood as "doubly" stochastic Poisson processes. To establish a more precise characterization of Cox processes consider a non-negative random field $\{\Lambda(\mathbf{u}) \colon \mathbf{u} \in S\}$, i.e. a collection of non-negative random variables with indices in $S$. We say that $X$ is a Cox process driven by $\Lambda$, if $X$ conditioned on $\Lambda = \lambda$ is an inhomogeneous Poisson process with intensity $\lambda$. A fundamental result for Cox processes shows that $\operatorname{Var}[N(X \cap B)] \geq \mathrm{E}[N(X \cap B)]$, where $B \subset S$. This result reveals that Cox processes are "overdispersed" relative to a Poisson process, which makes Cox processes suitable for modeling of clustered point patterns. In general, the $n$th order joint intensity is given by:

$$\lambda^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) = \mathrm{E}[\Lambda(\mathbf{u}_1) \cdots \Lambda(\mathbf{u}_n)] < \infty, \tag{5}$$

provided that $\lambda^{(n)}$ exists. Following (3) the PCF for a Cox process is given by $g(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{E}[\Lambda(\mathbf{u}_1)\Lambda(\mathbf{u}_2)] / \mathrm{E}[\Lambda(\mathbf{u}_1)]\,\mathrm{E}[\Lambda(\mathbf{u}_2)]$, which is useful knowledge when one wants to obtain an expression of the PCF for a specific type of Cox process. Consider $X$ as a Cox process restricted to $B \subseteq S$ with $|B| < \infty$. Then following (4), a density of $X$ with respect to a Poisson process with unit intensity is given by:

$$f(x) = \mathrm{E}\left[\exp\left(|B| - \int_B \Lambda(\mathbf{u})\mathrm{d}\mathbf{u}\right) \prod_{\mathbf{u} \in x \cap B} \Lambda(\mathbf{u})\right], \tag{6}$$

where $x$ is a finite point pattern on $B$. Notice that (6) is difficult to calculate due to $\exp\left(\int_B \Lambda(\mathbf{u})\mathrm{d}\mathbf{u}\right)$. To overcome the issue of calculating $\exp\left(\int_B \Lambda(\mathbf{u})\mathrm{d}\mathbf{u}\right)$ for statistical inference of a Cox process, we instead describe an alternative in Section 4. The next section describes one popular choice of a Cox process model called a log Gaussian Cox process.

## 3.3 Log Gaussian Cox point processes

One highly popular type of Cox processes is the log Gaussian Cox process (LGCP), where the log random field is a Gaussian random field. In paper B LGCP models play a main role, since we assume that the multivariate point pattern can be modeled as a multivariate LGCP.

### Univariate LGCP

Following Møller et al. (1998) we consider a Gaussian random field $\{V(\mathbf{u}) \colon \mathbf{u} \in S\}$ with mean function $m(\mathbf{u}) = \mathrm{E}[V(\mathbf{u})]$ and covariance function $c_V(\mathbf{u}, \mathbf{v}) = \operatorname{Cov}[V(\mathbf{u}), V(\mathbf{v})]$, where $\mathbf{u}, \mathbf{v} \in S$. If $X$ is a Cox process driven by $\Lambda(\cdot) =$

$\exp(V(\cdot))$, then we say that $X$ is a LGCP. If we denote by $\sigma(\mathbf{u})^2 = \text{Var}[V(\mathbf{u})]$, we can rewrite $V(\mathbf{u})$ as:

$$V(\mathbf{u}) = \sigma(\mathbf{u})U(\mathbf{u}) + m(\mathbf{u}), \tag{7}$$

where $\{U(\mathbf{u}) : \mathbf{u} \in S\}$ is a zero mean unit variance Gaussian random field with covariance function $c_U(\mathbf{u}, \mathbf{v}) = \text{Cov}[U(\mathbf{u}), U(\mathbf{v})]$. It follows directly from (5) and the moment generating functions for Gaussian random variables that the intensity function and PCF for a LGCP are given by:

$$\lambda(\mathbf{u}) = \exp\left(m(\mathbf{u}) + \frac{1}{2}\sigma(\mathbf{u})^2\right) \quad \text{and}$$
$$g(\mathbf{u}, \mathbf{v}) = \exp\left(\sigma(\mathbf{u})c_U(\mathbf{u}, \mathbf{v})\sigma(\mathbf{v})\right).$$

For convenience we often assume that $X$ is SOIRS, i.e. $\sigma(\mathbf{u}) = \sigma$ is constant and $c_U(\mathbf{u}, \mathbf{v}) = c_U(\mathbf{u} - \mathbf{v})$. For pairwise distinct $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^d$ the $n$th order joint intensity function for a LGCP is given by:

$$\lambda^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) = \prod_{i=1}^{n} \lambda(\mathbf{u}_i) \prod_{1 \leq i \leq j \leq n} g(\mathbf{u}_i - \mathbf{u}_j).$$

Notice that the distribution of a LGCP is fully characterized by the intensity and PCF. Even dough LGCP models are appealing due to their simplicity and the intuitive interpretation of $\sigma$ and $m$, one could be interested in constructing an even more flexible Cox process model. Such a model is introduced in Section 3.4.

**Multivariate LGCP**

To extend univariate LGCP models to the multivariate case we consider a multivariate Cox process $\mathbf{X} = (X_1, \ldots, X_p)^T$, where $p \geq 2$. A multivariate Cox process $\mathbf{X}$ is a straightforward extension of the univariate case, where each $X_i$ is a Cox process driven by a random intensity $\Lambda_i$ and conditioned on $\Lambda_i = \lambda_i$, each $X_i$ is an independent inhomogeneous Poisson process with intensity $\lambda_i$. Following Waagepetersen et al. (2016) we consider $\mathbf{Y}(\mathbf{u}) = (Y_1(\mathbf{u}), \ldots, Y_q(\mathbf{u}))^T$ and $\mathbf{U}(\mathbf{u}) = (U_1(\mathbf{u}), \ldots, U_p(\mathbf{u}))^T$, where $\{Y_l(\mathbf{u}) : \mathbf{u} \in S\}$, $l = 1, \ldots, q$, and $\{U_i(\mathbf{u}) : \mathbf{u} \in S\}$, $i = 1, \ldots, p$, are zero mean unit variance Gaussian random fields with covariance functions $c_{Y_l}(\mathbf{u}, \mathbf{v}) = \text{Cov}[Y_l(\mathbf{u}), Y_l(\mathbf{v})]$ and $c_{U_i}(\mathbf{u}, \mathbf{v}) = \text{Cov}[U_i(\mathbf{u}), U_i(\mathbf{v})]$, respectively. We can expand (7) to the multivariate case:

$$V_i(\mathbf{u}) = m_i(\mathbf{u}) + \sum_{l=1}^{q} \alpha_{il}(\mathbf{u})Y_l(\mathbf{u}) + \sigma_i(\mathbf{u})U_i(\mathbf{u}),$$

where the function $m_i(\mathbf{u})$ is deterministic and may depend on spatial covariates. We assume independence within and between $\mathbf{Y}(\mathbf{u})$ and $\mathbf{U}(\mathbf{u})$. Moreover, we assume that $\mathbf{X}$ is second order cross intensity reweighted isotropic, i.e. $c_{Y_l}(\mathbf{u},\mathbf{v}) = c_{Y_l}(||\mathbf{u}-\mathbf{v}||)$, $c_{U_i}(\mathbf{u},\mathbf{v}) = c_{U_i}(||\mathbf{u}-\mathbf{v}||)$, $\sigma_i(\mathbf{u}) = \sigma_i$ and $\alpha_{il}(\mathbf{u}) = \alpha_{il}$. The purpose of $Y_l$ is to represent latent variables that affect all the components in $\mathbf{X}$, while the concept of $U_i$ is to model the clustering within each component $X_i$, $i = 1,\ldots,p$. A multivariate LGCP $\mathbf{X}$ is then a multivariate Cox process, where each component has a random intensity function given by:

$$\Lambda_i(\mathbf{u}) = \exp\left(V_i(\mathbf{u})\right). \tag{8}$$

Using the moment generating function for Gaussian random variables the intensity is given by:

$$\lambda_i(\mathbf{u}) = \mathrm{E}\left[\Lambda_i(\mathbf{u})\right] = \exp\left(m_i(\mathbf{u}) + \sum_{l=1}^{q}\frac{\alpha_{il}^2}{2} + \frac{\sigma_i^2}{2}\right).$$

Similarly, the cross intensity between $X_i$ and $X_j$ is given by:

$$\begin{aligned}\lambda_{ij}(\mathbf{u},\mathbf{v}) &= \mathrm{E}\left[\Lambda_i(\mathbf{u})\Lambda_j(\mathbf{v})\right] \\ &= \exp\left(m_i(\mathbf{u}) + \sum_{l=1}^{q}\frac{\alpha_{il}^2}{2} + \frac{\sigma_i^2}{2}\right)\exp\left(m_j(\mathbf{v}) + \sum_{l=1}^{q}\frac{\alpha_{jl}^2}{2} + \frac{\sigma_j^2}{2}\right) \\ &\times \exp\left(\sum_{l=1}^{q}\alpha_{il}\alpha_{jl}c_{Y_l}(||\mathbf{u}-\mathbf{v}||) + \mathbf{1}[i=j]\sigma_i^2 c_{U_i}(||\mathbf{u}-\mathbf{v}||)\right).\end{aligned}$$

Hence, the cross PCF between $X_i$ and $X_j$ is given by:

$$g_{ij}(||\mathbf{u}-\mathbf{v}||) = \exp\left(\sum_{l=1}^{q}\alpha_{il}\alpha_{jl}c_{Y_l}(||\mathbf{u}-\mathbf{v}||) + \mathbf{1}[i=j]\sigma_i^2 c_{U_i}(||\mathbf{u}-\mathbf{v}||)\right).$$

If $i \neq j$ and $\sum_{l=1}^{q}\alpha_{il}\alpha_{jl}c_{Y_l}(||\mathbf{u}-\mathbf{v}||) > 0$ ($\sum_{l=1}^{q}\alpha_{il}\alpha_{jl}c_{Y_l}(||\mathbf{u}-\mathbf{v}||) < 0$) this indicates that $X_i$ and $X_j$ are positively (negatively) spatially correlated at lag $||\mathbf{u}-\mathbf{v}||$, while $\sum_{l=1}^{q}\alpha_{il}\alpha_{jl} = 0$ indicates that $X_i$ and $X_j$ are independent. Paper B assume a multivariate LGCP for the multivariate point pattern data along with a semi-parametric model for the intensity. Thus, paper B propose methodology to infer the parameters $\alpha_{il}$ and $\sigma_i^2$, $i = 1,\ldots,p$ and $l = 1,\ldots,q$ without estimating the complex non-parametric component in the semi-parametric intensity.

## 3.4 Tukey-Cox point processes

Although univariate LGCP models are appealing due to their simplicity and intuitive interpretation of the parameters $\sigma^2$ and $m$, one could be interested

in expanding the univariate LGCP models to even more flexible models. Following Xu and Genton (2017) the Tukey $t$-and-$h$ transformation is given by:

$$\tau_{t,h}(z) = \exp\left(\frac{h^2}{2}z\right)\exp(tz)/t - t,$$

where $\tau_{t,h}$ is a strictly monotone function of $z$, $h \geq 0$ and $t \in \mathbb{R}$. Let $Z \sim N(0,1)$. Then we can define $V = \tau_{t,h}(Z)$ as a stochastic variable that follows a Tukey $t$-and-$h$ distribution. We can interpret $t$ as the parameter that controls the skewness of the distribution, while $h$ controls the tails of the distribution. Inspired by the Tukey $t$-and-$h$ distribution we can define:

$$V(\mathbf{u}) = \frac{\gamma(\mathbf{u})}{2}Z^2(\mathbf{u}) + \sigma(\mathbf{u})Z(\mathbf{u}) + m(\mathbf{u}),$$

where $Z$ is a zero mean unit variance Gaussian random field with covariance function $c(\mathbf{u}_i, \mathbf{u}_j) = \text{Cov}(Z(\mathbf{u}_i), Z(\mathbf{u}_j))$. Following the unpublished paper by Genton et al. (2020), we can then define by $X$ a Tukey Cox point process (TCP), if $X$ is a Cox process with random intensity $\Lambda(\mathbf{u}) = \exp(V(\mathbf{u}))$. Clearly, if $\gamma(\cdot) = 0$, then $X$ reduces to a LGCP. Furthermore, if $\sigma(\cdot) = \sigma_0$ is fixed and $\gamma(\cdot) > 0$ (or $\gamma(\cdot) < 0$), then the resulting TCP is more (less) clustered than a LGCP with the same $\sigma_0$ and intensity as the TCP. In that sense, a TCP is a more flexible model compared to a LGCP. For convenience, we assume that $X$ is SOIRS and isotropic, i.e. $\gamma(\mathbf{u}) = \gamma$ and $\sigma(\mathbf{u}) = \sigma$ are constant and $c(\mathbf{u}_i, \mathbf{u}_j) = c(r)$ where $r = ||\mathbf{u}_i - \mathbf{u}_j||$. Following Genton et al. (2020), we can use the moment generating functions for Gaussian random variables and obtain the intensity for a TCP:

$$\lambda(\mathbf{u}) = \text{E}[\Lambda(\mathbf{u})] = \frac{1}{\sqrt{1-\gamma}}\exp\left(\frac{\sigma^2}{2(1-\gamma)} + m(\mathbf{u})\right). \tag{9}$$

In general, we can derive the $n$th order joint intensity $\lambda^{(n)}$ for a TCP $X$ under some mild assumptions. Consider $\mathbf{Z} = (Z(\mathbf{u}_1), \ldots, Z(\mathbf{u}_n))^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{0} = (0, \ldots, 0)^T$ is a $n \times 1$ mean vector and $\boldsymbol{\Sigma}$ is a $n \times n$ covariance matrix with entries $\Sigma_{ij} = c(||\mathbf{u}_i - \mathbf{u}_j||)$. Denote by $|M|$ the determinant of a matrix $M$. If we assume that $\boldsymbol{\Sigma}$ is positive definite and $\gamma \leq 1/n$, then following Genton et al. (2020) we can derive $\lambda^{(n)}$ for $X$ using the moment generating functions:

$$\begin{aligned}\lambda^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) &= \text{E}[\Lambda(\mathbf{u}_1)\ldots\Lambda(\mathbf{u}_n)] \\ &= |\mathbf{I}_n - \gamma\boldsymbol{\Sigma}|^{1/2}\exp\left(\mathbf{m}_n + \frac{1}{2}\mathbf{s}^T(\mathbf{I}_n - \gamma\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\mathbf{s}\right),\end{aligned} \tag{10}$$

where $\mathbf{s} = (\sigma, \ldots, \sigma)^T$, $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{m}_n = \sum_{i=1}^{n} m(\mathbf{u}_i)$. Notice, that we are able to rewrite (10) as:

$$\lambda^{(n)}(\mathbf{u}_1, \ldots, \mathbf{u}_n) = \prod_{k=1}^{n} \lambda(\mathbf{u}_k) \left( \frac{(1-\gamma)^n}{|\mathbf{I}_n - \gamma\boldsymbol{\Sigma}|} \right)^{1/2} \exp\left( \frac{1}{2}\mathbf{s}^T(\mathbf{I}_n - \gamma\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\mathbf{s} - \frac{n}{2}\frac{\sigma^2}{1-\gamma} \right). \tag{11}$$

From (11) we can derive the PCF for a TCP $X$, which is a parametric model given by:

$$g(r; \gamma, \sigma^2) = \left( 1 - \left[ \frac{\gamma}{1-\gamma} c(r) \right]^2 \right)^{-1/2} \exp\left( \frac{\sigma^2}{(1-\gamma)^2} \frac{c(r) + \frac{\gamma}{1-\gamma} c(r)^2}{1 - \left( \frac{\gamma}{1-\gamma} c(r) \right)^2} \right), \tag{12}$$

where $\gamma \leq 1/2$ and $\sigma^2 \in \mathbb{R}_+$.

One subject of this thesis has been to propose a method to infer the parameters $\gamma$ and $\sigma^2$ in a TCP. Clearly, one need to consider at least second order moments to make $\gamma$ and $\sigma^2$ identifiable, since the estimation of $\gamma$ and $\sigma^2$ cannot be estimated separately using (9). One option is to follow Waagepetersen and Guan (2009) and estimate $\gamma$ and $\sigma^2$ using a two-step estimation approach. Using this approach one may estimate the intensity function parametrically and then plug in the intensity estimate into a so-called second order estimating function (see Waagepetersen and Guan (2009) for details), which crucially depends on the PCF. However, consider (12) at lag zero and fix $\sigma^2 = \sigma_0^2$. Then $c(0) = 1$ and we can rewrite (12) as:

$$g(0, \gamma, \sigma_0^2) = \left( \frac{(1-\gamma)^2}{1-2\gamma} \right)^{1/2} \exp\left( \sigma_0^2 \frac{1}{(1-\gamma)(1-2\gamma)} \right). \tag{13}$$

When taking a closer look at (13), this is a convex function of $\gamma$ with limits:

$$\lim_{\gamma \to 1/2} g(0, \gamma, \sigma_0^2) = \infty \text{ and } \lim_{\gamma \to -\infty} g(0, \gamma, \sigma_0^2) = \infty.$$

This suggets that one needs to impose some restrictions on the parameter space of $\gamma$ in order to make $\gamma$ identifiable. Figure 3 shows the values of the PCF for a TCP at lag 0 with fixed $\sigma_0 = 0.5$ and values of $\gamma$ between $-3$ and 0.2. From Figure 3 we see that it is not clear how to restrict the parameter space of $\gamma$ in order to make an one-to-one correspondance between $\gamma$ and $g(0; \gamma, \sigma_0^2)$, since the global minimum of the PCF depends on the values of $\sigma^2$ and $\gamma$. As a conclusion, the identifiability issue regarding $\gamma$ remains a subject for further research such that TCP models can be applied as point process models to analyze point pattern data.

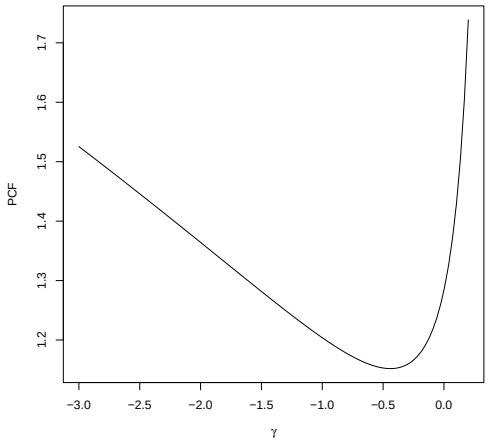


**Fig. 3:** Plot of the true PCF for a specific choice of TCP at lag 0 with $\sigma_0 = 0.5$ and values of $\gamma$ between $-3$ and 0.2.

# 4 Estimating functions

This section provides an alternative to Monte Carlo methods and simulation based inference for Cox processes (see Møller and Waagepetersen (2004) for details).

## 4.1 General theory on estimating functions

In this section we introduce the theory of estimating functions, which generalizes the concept of maximum likelihood estimation. In the following we consider a random vector $\mathbf{Z} = (Z_1, \ldots, Z_m)$. Associated with $\mathbf{Z}$ is a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ and $\mathcal{P} = \{\mathrm{f}_{\mathbf{Z}}(\cdot; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}$ as a parameterized class of statistical models, where $\boldsymbol{\theta}$ is a parameter in the parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^K$. Given an observed sample $\mathbf{z} = (z_1, \ldots, z_m)$ one can define a likelihood function of $\boldsymbol{\theta}$ as the joint probability density at the observed data $\mathbf{z}$:

$$L(\boldsymbol{\theta}; \mathbf{z}) = \mathrm{f}_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\theta}). \tag{14}$$

By maximizing (14) one can obtain an estimate of $\boldsymbol{\theta}$, which is known as the maximum likelihood estimator (MLE). In fact, if $l(\boldsymbol{\theta}; \mathbf{z}) = \log(L(\boldsymbol{\theta}; \mathbf{z}))$ is differentiable with respect to $\boldsymbol{\theta}$, one can obtain the MLE by solving:

$$\mathrm{s}(\boldsymbol{\theta}; \mathbf{z}) = \mathbf{0},$$

where $\mathrm{s} = \mathrm{d}l/\mathrm{d}\boldsymbol{\theta}$ is called the score function. Notice, a sufficient condition for a local MLE $\hat{\boldsymbol{\theta}}$ is that the Hessian matrix $H = \mathrm{d}\mathrm{s}/\mathrm{d}\boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}$ is negatively semi-definite.

To generalize the concept of maximum likelihood estimation we consider a so-called estimating function $\boldsymbol{\psi}$, which is a vector function on $\mathbb{R}^K \times \mathbb{R}^m$ with components $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^T$. A strict measure theoretic definition of estimating functions can be found in Song (2007). To identify the parameters the number of components of $\boldsymbol{\psi}$ need to be of same cardinality as the dimension of $\boldsymbol{\Theta}$. In general, by solving the equation:

$$\boldsymbol{\psi}(\mathbf{z}; \boldsymbol{\theta}) = \mathbf{0} \tag{15}$$

we can obtain an estimate for the parameter $\boldsymbol{\theta}$ of interest. Equation (15) is called an estimating equation. If $\boldsymbol{\psi}$ and $\tilde{\boldsymbol{\psi}}$ provide the same estimates for given $\mathbf{z}$, we can consider $\boldsymbol{\psi}$ and $\tilde{\boldsymbol{\psi}}$ as equivalent estimating functions. Furthermore, if $\mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\psi}(\mathbf{z}; \boldsymbol{\theta})\right] = \mathbf{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we call $\boldsymbol{\psi}$ an unbiased estimating function.

For some data sets it is difficult to specify or compute a full parametric density function. Then one may develop estimating functions that are simpler than the score function but still sensible for estimation of $\boldsymbol{\theta}$.

## 4.2 General theory on composite likelihood

One alternative to MLE is composite likelihood estimation, which is constructed by a multiplum of a collection of likelihood components. The likelihood components can be marginal or conditional probability density functions, where the dependence between the components is ignored. The resulting estimating function is derived as the derivative of the sum of log-likelihood components. Even though the estimating function is derived from a misspecified model, the function is still unbiased and thereby sensible for parameter estimation.

To outline the concept of composite likelihoods we consider a random vector $\mathbf{Z} = (Z_1, \ldots Z_m)$. Furthermore, we associate $\mathbf{Z}$ with a joint probability density function $\mathrm{f}_{\mathbf{Z}}(\cdot; \boldsymbol{\theta})$, where the parameter $\boldsymbol{\theta}$ is in a parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^K$. Given a sample $\mathbf{z} = (z_1, \ldots, z_m)$ and a subset of indices $B \subseteq \{1, \ldots, N\}$ we can construct the following likelihood function:

$$L_{\mathbf{Z}_B}(\boldsymbol{\theta}, \mathbf{z}) = \mathrm{f}_{\mathbf{Z}_B}\left(\mathbf{z}_B; \boldsymbol{\theta}\right),$$

where $\mathbf{Z}_B = (Z_k)_{k \in B}$. Now define a set of indices $\{B_1, \ldots, B_N\}$, where $B_k \subseteq \{1, \ldots, N\}$. Following Varin et al. (2011) we define by:

$$L_{cl}(\boldsymbol{\theta}; \mathbf{z}) = \prod_{k=1}^{N} L_{\mathbf{Z}_{B_k}}(\boldsymbol{\theta}; \mathbf{z}_{B_k})^{w_k}$$

a composite likelihood function, where $w_k$ are non-negative weights. If $w_k = w_l$ for all $k, l = 1, \ldots, N$, the weights are simply omitted. Although many types of composite likelihood functions can be constructed, we only focus on first and second order composite likelihoods, i.e:

$$L_1(\boldsymbol{\theta}; \mathbf{z}) = \prod_{k=1}^{N} \mathrm{f}_{\mathbf{Z}_{B_k}}(\mathbf{z}_{B_k}; \boldsymbol{\theta}), \quad \text{and} \quad L_2(\boldsymbol{\theta}; \mathbf{z}) = \prod_{k \neq l} \mathrm{f}_{\mathbf{Z}_{B_{kl}}}(\mathbf{z}_{B_{kl}}; \boldsymbol{\theta}),$$

where $B_k = \{k\}$ and $B_{kl} = \{k, l\}$. The first order composite likelihood function is often referred to as the independence likelihood, since this function consists of a product of marginal likelihood functions. As a consequence, the first order composite likelihood can only infer marginal parameters. However, the second order composite likelihood function can be used to infer parameters that relates to dependence between $Z_k$ and $Z_l$.

## 4.3 First order composite likelihood in paper A

The theory of composite likelihood functions plays a central role in this thesis, since the theory is applied in paper A and paper B to construct new methods to infer the first and second order properties in multivariate point pattern data with complex intensity functions. As mentioned in Section 1.1 we analyze in paper A how street crimes in Washington D.C. relate to demographical spatial covariates. As already pointed out, the street crime occurrences depend in a complex way on the population density and the urban structure of the city. Denote by $\mathbf{X} = (X_1, \ldots, X_p)$ a multivariate spatial point process, where $X_i$ is a spatial point process of type $i$, $i = 1, \ldots, p$. To accommodate the difficulty of modeling the complex intensity functions we assume in paper A that each $X_i$ has a semi-parametric intensity function of the form:

$$\lambda_i(\mathbf{u}; \boldsymbol{\gamma}_i) = \lambda_0(\mathbf{u}) \exp\left(\boldsymbol{\gamma}_i^T \mathbf{z}(\mathbf{u})\right), \tag{16}$$

where $\lambda_0(\cdot)$ is a complex background intensity that is assumed to be common for all the types of points. In case of Washington D.C. street crimes, $\lambda_0$ takes into account the population density and the urban structure of the city. Moreover, $\mathbf{z}(\mathbf{u}) \in \mathbb{R}^{(q+1)}$ is a vector of spatial covariates at location $\mathbf{u}$, while $\boldsymbol{\gamma}_i \in \mathbb{R}^{(q+1)}$ is a vector of regression parameters.

Since $\lambda_0(\cdot)$ is completely unspecified the parameters $\boldsymbol{\gamma}_i$ are not identifiable. To overcome this issue we instead choose a baseline, say $X_p$, and estimate contrasts $\boldsymbol{\beta}_i = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_p$, for $i = 1, \ldots, p-1$. In paper A we estimate each $\boldsymbol{\beta}_i$ by constructing a conditional probability at some location $\mathbf{u}$. We condition on that an event is observed at $\mathbf{u}$. The probability that $\mathbf{u}$ is from type $i$

is then given by:

$$\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}) = \frac{\lambda_i(\mathbf{u};\boldsymbol{\gamma}_i)}{\sum_{k=1}^{p}\lambda_k(\mathbf{u};\boldsymbol{\gamma}_k)} = \begin{cases} \frac{\exp\left(\boldsymbol{\beta}_i^T\mathbf{z}(\mathbf{u})\right)}{1+\sum_{k=1}^{p-1}\exp\left(\boldsymbol{\beta}_k^T\mathbf{z}(\mathbf{u})\right)}, & i=1,\dots p-1 \\ \frac{1}{1+\sum_{k=1}^{p-1}\exp\left(\boldsymbol{\beta}_k^T\mathbf{z}(\mathbf{u})\right)}, & i=p. \end{cases}$$

Notice that the probabilities do not depend on the complex background intensity. In paper A we apply the theory of first order composite likelihood function to define the multinomial conditional composite likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{p}\prod_{\mathbf{u}\in X_i\cap W}\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}).$$

The resulting log multinomial conditional composite likelihood function is then:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{p}\sum_{\mathbf{u}\in X_i\cap W}\boldsymbol{\beta}_i^T\mathbf{z}(\mathbf{u}) - \log\left(1+\sum_{k=1}^{p-1}\exp\left(\boldsymbol{\beta}_k^T\mathbf{z}(\mathbf{u})\right)\right). \tag{17}$$

Since (17) is a concave function of $\boldsymbol{\beta}$, we maximize (17) by solving the estimation equation $\mathrm{e}(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\mathrm{e}(\boldsymbol{\beta}) = (\mathrm{e}_1(\boldsymbol{\beta})^T,\dots,\mathrm{e}_{p-1}(\boldsymbol{\beta})^T)^T, \tag{18}$$

and

$$\mathrm{e}_i(\boldsymbol{\beta}) = \sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u}) - \sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W}\frac{\mathbf{z}(\mathbf{u})\exp\left(\boldsymbol{\beta}_i^T\mathbf{z}(\mathbf{u})\right)}{1+\sum_{k=1}^{p-1}\exp\left(\boldsymbol{\beta}_k^T\mathbf{z}(\mathbf{u})\right)}, \quad i=1,\dots,p-1.$$

We show some asymptotic properties of $\widehat{\boldsymbol{\beta}}$ in paper A when $\mathbf{X}$ is observed on a sequence of increasing windows $W_n$. The main asymptotic results of $\hat{\boldsymbol{\beta}}$ are presented in Section 5.1.

## 4.4 Second order composite likelihood in paper B

As pointed out earlier, the first order composite likelihood function cannot infer parameters that relate to dependence between random events. Instead we define a second order composite likelihood function. To study the full between and within dependence structure of a multivariate point process, we assume in paper B that the multivariate point pattern data can be modeled as a multivariate LGCP model $\mathbf{X} = (X_1,\dots,X_p)$. We assume that $\mathbf{X}$ is second order cross intensity reweighted stationary and isotropic, which means that $g_{ij}(\mathbf{u},\mathbf{v}) = g_{ij}(r)$ for all $i,j = 1,\dots,p$, where $r = ||\mathbf{u}-\mathbf{v}||$. In paper B we deal

with the complex intensity functions for each $X_i$ by combining the random intensity functions (8) with the semi-parametric model (16), i.e:

$$\Lambda_i(\mathbf{u}) = \lambda_0(\mathbf{u})\exp\left(\boldsymbol{\gamma}_i^T\mathbf{z}(\mathbf{u})\right)\exp\left(m_i + \sum_{l=1}^{q}\alpha_{il}Y_l(\mathbf{u}) + \sigma_i U_i(\mathbf{u})\right),$$

where $m_i = -\sum_{l=1}^{p}\alpha_{il}^2/2 - \sigma_i^2/2$. Following paper B, the intensity of $X_i$ is:

$$\lambda_i(\mathbf{u}) = \mathrm{E}\left[\Lambda_i(\mathbf{u})\right] = \lambda_0(\mathbf{u})\exp\left(\boldsymbol{\gamma}_i^T\mathbf{z}(\mathbf{u})\right).$$

In a similar way, we can calculate the cross intensity function between $X_i$ and $X_j$:

$$\begin{aligned}\lambda_{ij}(\mathbf{u},\mathbf{v}) = \mathrm{E}\left[\Lambda_i(\mathbf{u})\Lambda_j(\mathbf{v})\right] =& \lambda_0(\mathbf{u})\lambda_0(\mathbf{v})\exp\left(\boldsymbol{\gamma}_i^T\mathbf{z}(\mathbf{u})\right)\exp\left(\boldsymbol{\gamma}_j^T\mathbf{z}(\mathbf{v})\right)\\ &\times\exp\left(\sum_{l=1}^{q}\alpha_{il}\alpha_{jl}c_{Y_l}(r) + \mathbf{1}[i=j]\sigma_i^2 c_{U_i}(r)\right).\end{aligned}$$

We use the exponential correlation function for $c_{Y_l}(r)$ and $c_{U_i}(r)$ in paper B, i.e $c_{Y_l}(r) = \exp(-r/\xi_l)$ and $c_{U_i}(r) = \exp(-r/\varphi_i)$, where $\xi_l$, $l = 1,\ldots,q$ and $\varphi_i$, $i = 1,\ldots,p$ are correlation scale parameters. Let $\boldsymbol{\theta}$ be the concatenation of $\boldsymbol{\alpha}_{\cdot k} = (\alpha_{1k},\ldots,\alpha_{pk})^T$, $k = 1,\ldots,p$, $\boldsymbol{\sigma}^2 = (\sigma_1^2,\ldots,\sigma_p^2)^T$, $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_q)^T$ and $\boldsymbol{\varphi} = (\varphi_1,\ldots,\varphi_p)^T$. The cross PCF between $X_i$ and $X_j$ is then of the form:

$$g_{ij}(r;\boldsymbol{\theta}) = \exp\left(\sum_{l=1}^{q}\alpha_{il}\alpha_{jl}\exp\left(\frac{-r}{\xi_l}\right) + \mathbf{1}[i=j]\sigma_i^2\exp\left(\frac{-r}{\varphi_i}\right)\right).$$

Now assume that $\boldsymbol{\beta}$ is known and estimated using the approach described in Section 4.3 and denote by $f_i(\mathbf{u}) = \exp(\boldsymbol{\beta}_i^T\mathbf{z}(\mathbf{u}))$. In paper B we adapt the ideas from paper A to construct second order conditional probabilities in a similar way. We condition on that two events are observed at $\mathbf{u}$ and $\mathbf{v}$ where $\mathbf{u} \neq \mathbf{v}$. The conditional probability that $\mathbf{u}$ is of type $i$ and $\mathbf{v}$ is of type $j$ is then given by:

$$\mathrm{p}_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\theta}) = \frac{\lambda_{ij}(\mathbf{u},\mathbf{v})}{\sum_{k,l}\lambda_{kl}(\mathbf{u},\mathbf{v})} = \frac{f_i(\mathbf{u})f_j(\mathbf{v})g_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\theta})}{\sum_{k,l}f_k(\mathbf{u})f_l(\mathbf{v})g_{kl}(\mathbf{u},\mathbf{v};\boldsymbol{\theta})}. \tag{19}$$

Notice that (19) does not depend on $\lambda_0(\cdot)\lambda_0(\cdot)$. To estimate $\boldsymbol{\theta}$ we then maximize the following second order conditional composite likelihood:

$$L(\boldsymbol{\theta}) = \prod_{ij}\prod_{\substack{\mathbf{u}\in X_i\cap W\\ \mathbf{v}\in X_j\cap W}}^{\neq} 1[||\mathbf{u}-\mathbf{v}|| \leq R]\mathrm{p}_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\theta}), \tag{20}$$

where $R > 0$ is a user-specific tuning parameter. Since we do not have any closed-form solution for this maximization problem, we instead estimate $\boldsymbol{\theta}$ using an iterative Newton-Raphson method as described in paper B.

# 5 Summary of findings in part II

This section summarizes the main findings in paper A and paper B.

## 5.1 Summary of findings in paper A

In paper A we show some asymptotic properties of $\widehat{\boldsymbol{\beta}}$ when $\mathbf{X}$ is observed on a sequence of increasing windows $W_n$. Let $|W_n|$ be the size of $W_n$. Denote by $\mathrm{e}^{(n)}(\boldsymbol{\beta})$ the multinomial conditional composite likelihood estimating function in (18) evaluated on $W_n$ and let $\widehat{\boldsymbol{\beta}}_n$ be the sequence of estimators that solves the estimating equation $\mathrm{e}^{(n)}(\boldsymbol{\beta}) = \mathbf{0}$. Furthermore, let $\boldsymbol{\beta}_i^* = \gamma_i^* - \gamma_p^*$, where $\gamma_i^*$ is the true value of the semi-parametric intensity function for $X_i$. Under mild conditions (C1-C4 in paper A) we show in paper A that there exists a sequence of solutions $\widehat{\boldsymbol{\beta}}_n$ that solves $\mathrm{e}^{(n)}(\boldsymbol{\beta}) = \mathbf{0}$ for which:

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}^*,$$

as $n \to \infty$. To show consistency of $\widehat{\boldsymbol{\beta}}_n$ we use Theorem 2 in Waagepetersen and Guan (2009) and verify their conditions W1-W4 (see Section C.1 in paper A for details).

Furthermore, the asymptotic covariance matrix of $\boldsymbol{\beta}$ is of the form:

$$\mathrm{S}(\boldsymbol{\beta}^*)^{-1}\Sigma(\boldsymbol{\beta}^*, g)\mathrm{S}(\boldsymbol{\beta}^*)^{-1},$$

where $\mathrm{S}(\boldsymbol{\beta}^*) = \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}^T} - \mathrm{e}(\boldsymbol{\beta}^*)\right]$ is the sensitivity matrix and $\Sigma(\boldsymbol{\beta}^*, g) = \mathrm{Var}\left[\mathrm{e}(\boldsymbol{\beta}^*)\right]$ is the covariance matrix of $\mathrm{e}(\boldsymbol{\beta}^*)$. The expressions for $\mathrm{S}(\boldsymbol{\beta}^*)$ and $\Sigma(\boldsymbol{\beta}^*, g)$ can be found in paper A. Let $\bar{\mathrm{S}}(\boldsymbol{\beta}^*) = \mathrm{S}(\boldsymbol{\beta}^*)/|W_n|$ and $\bar{\Sigma}(\boldsymbol{\beta}^*, g) = \Sigma(\boldsymbol{\beta}^*, g)/|W_n|$. Under mild conditions (C1-C4 and N1-N3 in paper A) we show in paper A that:

$$|W_n|^{-1/2}\bar{\Sigma}_n^{-1/2}(\boldsymbol{\beta}^*, g)\bar{\mathrm{S}}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{d} N\left(\mathbf{0}, \mathrm{I}_{(p-1)q}\right).$$

To show asymptotic normality of $\boldsymbol{\beta}_n$ we consider a first order Taylor expansion of $\mathrm{e}^{(n)}(\boldsymbol{\beta}) = \mathbf{0}$ around $\boldsymbol{\beta}^*$. Subsequently, we follow Theorem 3.1 in (Biscio and Waagepetersen, 2019) and verify conditions $\mathcal{H}1 - \mathcal{H}4$ (see Section C.2 in paper A for details).

Notice that $\Sigma(\boldsymbol{\beta}^*, g)$ depends on the underlying cross PCFs. Moreover, the estimation of cross PCFs requires consistent estimates of the intensity functions. Since $\lambda_0$ is unknown, we are not available to provide consistent estimates of the intensity functions under model (16). However, when taking a closer look at $\Sigma(\boldsymbol{\beta}^*, g)$ we found in paper A that the expressions of $\Sigma(\boldsymbol{\beta}^*, g)$ can be computed using estimates of ratios of cross PCFs, i.e.

$$g_{ij,kl}(\mathbf{u}, \mathbf{v}) = g_{ij}(\mathbf{u}, \mathbf{v})/g_{kl}(\mathbf{u}, \mathbf{v}), \quad i, j = 1, \ldots, p, \tag{21}$$

where the types $k$ and $l$ are fixed. Consider now the quantity:

$$F_{ij}(r;b,\boldsymbol{\beta}) = \sum_{\substack{\mathbf{u}\in X_i\cap W \\ \mathbf{v}\in X_j\cap W}}^{\neq} \frac{k_b(||\mathbf{u}-\mathbf{v}||-r)}{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta})}, \tag{22}$$

where $k_b(\cdot) = k(\cdot/b)/b$ is a kernel function defined on a bounded interval in $\mathbb{R}$ and $b$ is a bandwidth. In paper A we define:

$$\hat{g}_{ij,kl}(r;b,\widehat{\boldsymbol{\beta}}) = F_{ij}(r;b,\widehat{\boldsymbol{\beta}})/F_{kl}(r;b,\widehat{\boldsymbol{\beta}})$$

as an estimator of (21). Let $b_n$ be a sequence of bandwidths and denote by $\hat{g}_{ij,kl}(r;b_n,\widehat{\boldsymbol{\beta}}_n)$ a sequence of estimators:

$$\hat{g}_{ij,kl,n}(r;b,\widehat{\boldsymbol{\beta}}) = F_{ij,n}(r;b,\widehat{\boldsymbol{\beta}})/F_{kl,n}(r;b,\widehat{\boldsymbol{\beta}}),$$

where $F_{ij,n}$'s are given by (22) with $W = W_n$. Under mild conditions (C2 and K1-K3 in paper A) we show in paper A that:

$$\hat{g}_{ij,kl,n}(r;b,\widehat{\boldsymbol{\beta}}) \xrightarrow{p} g_{ij,kl}(r), \text{ as } n\to\infty \text{ for } i,j,k,l = 1,\dots,p. \tag{23}$$

The proof of (23) can be found in Section C.3 in paper A.

## 5.2 Summary of findings in paper B

In paper B the parameters $\boldsymbol{\alpha}_{il}$ are not identifiable due to the unknown $\lambda_0$. In fact, we are only able to estimate contrasts $\boldsymbol{\alpha}_{il} - \boldsymbol{\alpha}_{pl}$, $i = 1,\dots,p-1$, where we choose $X_p$ as a baseline. In order to make $\boldsymbol{\alpha}_{il}$ identifiable and thereby infer the PCFs and cross PCFs in a multivariate LGCP we need to impose some restrictions on the $\alpha_{il}$s. We impose the sum-to-zero constraint $\sum_{i=1}^{p}\alpha_{il} = 0$, $l = 1,\dots,q$, in paper B to ensure identifiability. With abuse of notation, we denote by $\boldsymbol{\alpha}$ both the matrix $[\alpha_{il}]_{il}$ and the vectorized version where the matrix is laid out column-wise $(\boldsymbol{\alpha}_{\cdot 1}^T,\dots,\boldsymbol{\alpha}_{\cdot q}^T)^T \in \mathbb{R}^{pq}$. We propose an optimization algorithm in paper B to infer $\boldsymbol{\theta}$, where we accommodate the sum-to-zero constraint on $\boldsymbol{\alpha}_{il}$ by the change of variable $\mathrm{B}\boldsymbol{\psi} = \boldsymbol{\alpha}$, where $\boldsymbol{\psi}$ is a $(p-1)\times q$ matrix and

$$\mathrm{B}^T = \begin{bmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{bmatrix}$$

is a $(p-1)\times p$ matrix. We then estimate $\boldsymbol{\psi}$, which is identifiable. Hence, we get $\hat{\alpha} = \mathrm{B}\hat{\boldsymbol{\psi}}$.

One limitation of paper B is that we did not establish asymptotic results for the parameter estimates in the multivariate LGCP model. To establish such asymptotic results one would need the expression of the sensitivity and the variance of the second order conditional composite likelihood estimating function. In fact, to provide an unbiased estimate of the variance of the second order conditional composite likelihood estimating function, one need to compute a certain complex quadruple sum. To compute such an quadruple sum is both time consuming and also numerically unstable. Instead, we in paper B combine the negation of the log of (20) with Lasso penalization (Tibshirani, 1996) to enable the parameter estimates of $\alpha_{il}$ to be exactly zero. In that way, we can determine whether two point processes are independent. In paper B we define the regularized object function as:

$$-\log\left(L(\boldsymbol{\theta})\right) + \eta \sum_{i=1}^{p}\sum_{l=1}^{q} |\alpha_{il}|, \quad \mathrm{C}\boldsymbol{\alpha} = \mathbf{0},$$

where $\eta$ is a penalty parameter and C is a $q \times pq$ matrix given by:

$$\mathrm{C} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{bmatrix}.$$

Notice that we use the vectorized version of $\boldsymbol{\alpha}$. In paper B we apply the augmented Lagrangian algorithm proposed by Shi et al. (2016) to compute $\hat{\boldsymbol{\alpha}}$ (see Section D in paper B for details).

We also propose a $K$-fold cross validation criterion in paper B to determine the value of $q$ and $\eta$. For each $i, j$, let $M_{ij}$ denote the set of pairs $(\mathbf{u}, \mathbf{v})$, where $\mathbf{u} \in X_i$, $\mathbf{v} \in X_j$ and $0 < ||\mathbf{u} - \mathbf{v}|| < R$, where $R$ is specified in (20). We randomly split $M_{ij}$ into $K$ subsets $M_{ij,1}, \ldots, M_{ij,K}$ of equal sizes. For each $k = 1, \ldots, K$ we obtain a parameter estimate $\hat{\boldsymbol{\theta}}_k$ by maximizing:

$$l_k(\boldsymbol{\theta}) = \sum_{i,j} \overset{\neq}{\sum_{(\mathbf{u},\mathbf{v})\in M_{ij,-k}}} \log\Big(\mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})\Big) + \eta \sum_{i=1}^{p}\sum_{l=1}^{q} |\alpha_{il}|, \quad \mathrm{C}\boldsymbol{\alpha} = \mathbf{0},$$

where $M_{ij,-k} = \cup_{l\neq k} M_{ij,l}$. The $k$th cross validation score is then:

$$CV_k(q, \eta) = \sum_{i\neq j} \overset{\neq}{\sum_{(\mathbf{u},\mathbf{v})\in M_{ij,k}}} \log\Big(\mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \hat{\boldsymbol{\theta}}_k)\Big).$$

Let $\overline{CV}(q, \eta)$ be the average of $CV_k(q, \eta)$, $k = 1, \ldots, K$. Then one can choose $q$ and $\eta$ that minimizes $\overline{CV}(q, \eta)$. In paper B we first select $q$ and secondly $\eta$, since $q$ determines the overall model complexity, while the choice of $\eta > 0$ may introduce additional sparsity for a given choice of $q$.

# References


Anderes, E., Møller, J., and Rasmussen, J. (2018). Isotropic covariance functions on graphs and their edges. *Annals of Statistics*, 48:2478–2503.

Baddeley, A., Jammalamadaka, A., and Nair, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society, Series C*, 63:673–694.

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54:329–350.

Bell, M. L. and Grunwald, G. K. (2004). Mixed models for the analysis of replicated spatial point patterns. *Biostatistics*, 5:633–648.

Billingsley, P. (1995). *Probability and Measure*. Wiley, New York.

Biscio, C. A. N. and Waagepetersen, R. P. (2019). A general central limit theorem and subsampling variance estimator for $\alpha$-mixing point processes. *Scandinavian Journal of Statistics*. Appeared online, 1-23.

Brix, A. and Møller, J. (2001). Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scandinavian journal of statistics*, 28:471–488.

Choiruddin, A., Coeurjolly, J.-F., and Letué, F. (2018). Convex and non-convex regularization methods for spatial point processes intensity estimation. *Electronic Journal of Statistics*, 12:1210–1255.

Choiruddin, A., Cuevas-Pacheco, F., Couerjolly, J.-F., and Waagepetersen, R. (2019). Regularized estimation for highly multivariate log Gaussian Cox processes. *Statistics and Computing*, 30:649–662.

Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity function. *Biometrika*, 105:455–462.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, 2. edition.

Diggle, P., Zheng, P., and Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society: Series C*, 54(3):645–658.

Diggle, P. J. (1985). A Kernel Method for Smoothing Point Process Data. *Applied Statistics*, 34:138–147.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A*, 153:349–362.

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*, 63:550–557.

Diggle, P. J., Lange, N., and Benes, F. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association*, 86:618–625.

Diggle, P. J. and Milne, R. K. (1983). Bivariate Cox processes: some models for bivariate spatial point processes. *Journal of the Royal Stastistical Society, Series B*, 45:11–21.

Diggle, P. J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, 157:433–440.

Egesdal, M., Fathauer, C., Louie, K., Neuman, J., Mohler, G., and Lewis, E. (2010). Statistical and stochastic modeling of gang ryvalry in Los Angeles. *SIAM Undergraduate Research Online*, 3:72–94.

Genton, M. G., Hessellund, K. B., Møller, J., Waagepetersen, R., and Xu, G. (2020). Tukey-Cox point processes. unpublished.

Gorr, W. L. and Lee, Y. (2015). Early warning system for temporary crime hot spots. *Journal of Quantitative Crimonology*, 31:25–47.

Grabarnik, P. and Särkkä, A. (2009). Modelling the spatial structure of forest stands by multivariate point processes with hierarchical interactions. *Ecological modelling*, 220:1232–1240.

Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, 103:1238–1247.

Guan, Y., Waagepetersen, R., and Beale, C. (2008). Second-order analysis of inhomogeneous spatial point processes with proportional intensity functions. *Journal of the American Statistical Association*, 103:769–777.

Högmander, H. and Särkkä, A. (1999). Multitype spatial point patterns with hierarchical interactions. *Biometrics*, 55:1051–1058.

Lavancier, F., Møller, J., and Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Stastistical Society, Series B*, 77:853–877.

Liang, S., Carlin, B. P., and Gelfand, A. E. (2017). Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. *The Annals of Applied Statistics*, 3:943–962.

Mateu, J. (2001). Parametric procedures in the analysis of replicated pairwise interaction point patterns. *Biometrical Journal*, 43:375–394.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106:100–108.

Møller, J., Syversveen, A. R., and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451–482.

Møller, J. and Waagepetersen, R. (2017). Some recent developments in stastistics for spatial point patterns. *Annual Review of Statistics and Its Application*, 4:317–342.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, 1. edition.

Okabe, A. and Yamada, I. (2001). The K-function method on a network and its computational implementation. *Geographical Analysis*, 33:271–290.

Rajala, T., Murrell, D. J., and Olhede, S. C. (2018). Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection. *Journal of the Royal Statistical Society, Series C*, 67:1237–1273.

Ratcliffe, J. (2004). Crime mapping and the training needs of law enforcement. *European Journal on Criminal Policy and Research*, 10:65–83.

Reinhart, A. and Greenhouse, J. (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C*, 67:1305–1329.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266.

Rossy, J., Owen, D. M., Williamson, D. J., Yang, Z., and Gaus, K. (2013). Conformational states of the kinase Lck regulate clustering in early T cell signaling. *Nature Immunology*, 14:82–89.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Annals of Applied Statistics*, 10:1019–1040.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics and Applications*. Springer - Verlag New York.

Stoica, R. S., Martinez, V. J., and Saar, E. (2007). A three dimensional object point process for detecting of cosmic filaments. *Journal of the Royal Statistical Society, Series C (Applied statistics)*, 56:459–477.

Stoyan, D. and Stoyan, H. (1994). *Fractals, random shapes and point fields: methods of geometrical statistics*. Chichester: Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tompson, L., Partridge, H., and Shepherd, N. (2008). Hot routes: Developing a new technique for the spatial analysis of crime. *Crime Mapping: A Journal of Research and Practice*, 1:77–96.

van Lieshout, M. N. M. (2000). *Markov point processes and their applications*. London: Imperial College Press.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Waagepetersen, R. (2007). An estimation function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics*, 63:252–258.

Waagepetersen, R. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:685–702.

Waagepetersen, R., Guan, Y., Jalilian, A., and Mateu, J. (2016). Analysis of multispecies point patterns by using multivariate log-Gaussian Cox processes. *Journal of the Royal Statistical Society. Series C.*, 65:77–96.

Xu, G. and Genton, M. (2017). Tukey g-and-h random fields. *Journal of the American Statistical Association*, 112:1236–1249.

Yong, Z. and Hancock, J. F. (2018). Electron microscopy combined with spatial analysis: quantitative mapping of the nano-assemblies of plasma membrane-associating proteins and lipids. *Biophysics Reports*, 4:320–328.

# Part II

# Papers

# Paper A

## Semi-parametric multinomial logistic regression for multivariate point pattern data

Kristian Bjørn Hessellund, Ganggang Xu, Yongtao Guan and Rasmus Waagepetersen

The paper is accepted for publication in
*Journal of the American Statistical Association*

*The layout has been revised.*

# Abstract

*We propose a new method for analysis of multivariate point pattern data observed in a heterogeneous environment and with complex intensity functions. We suggest semi-parametric models for the intensity functions that depend on an unspecified factor common to all types of points. This is for example well suited for analyzing spatial covariate effects on events such as street crime activities that occur in a complex urban environment. A multinomial conditional composite likelihood function is introduced for estimation of intensity function regression parameters and the asymptotic joint distribution of the resulting estimators is derived under mild conditions. Crucially, the asymptotic covariance matrix depends on ratios of cross pair correlation functions of the multivariate point process. To make valid statistical inference without restrictive assumptions, we construct consistent non-parametric estimators for these ratios. Finally, we construct standardized residual plots, predictive probability plots, and semi-parametric intensity plots to validate and to visualize the findings of the model. The effectiveness of the proposed methodology is demonstrated through extensive simulation studies and an application to analyzing the effects of socio-economic and demographical variables on occurrences of street crimes in Washington DC.*


# 1 Introduction

Multivariate point pattern data with many types of points are becoming increasingly common. Ecologists collect large data sets on locations and species of plants and animals, while police authorities gather ever-increasing data sets on times, locations, and types of crimes. In epidemiology, multivariate point pattern data sets concern geo-referenced occurrences of different types of disease or bacteria. While the literature of bivariate point patterns is fairly well-developed (see e.g. the review in Waagepetersen et al. (2016)), much less work has been done on the statistical analysis of point patterns with more than two types of points. Diggle et al. (2005) and Baddeley et al. (2014) considered four- and six-variate multivariate Poisson processes and more recently Jalilian et al. (2015) and Waagepetersen et al. (2016) considered five- and nine-variate multivariate Cox processes. Rajala et al. (2018) and Choiruddin et al. (2019) consider penalized estimation for respectively multivariate Gibbs and log Gaussian Cox point processes for data sets containing locations of more than 80 species of rain forest trees.

This paper is concerned with statistical modeling of the first-order intensity functions of a multivariate spatial point process with an arbitrary number of types of points. For clarity of exposition we discuss our proposal in relation to the specific problem of street crime analysis where we focus on the spatial aspects of street crimes aggregated over a time span of interest,

see also the data example in Section 6. To model street crime activities as a multivariate point process poses three major challenges: (1) to handle the high complexity of the first-order intensity function for each type of points; (2) to relate the street crime locations to available spatial covariates; (3) to take into account spatial correlations within and between different types of crimes. The first challenge arises because street crime activities depend in a complicated way on the layout of the city (streets, squares, malls,...) as well as the typically unknown population density at any location. Moreover, the intensity of crime activities may also change abruptly from one area to neighboring areas. The second challenge arises because it is of great interest to police and criminologists to gain information on how street crime occurrences are related to demography, socio-economic variables, and other covariates. Such information is, for example, helpful to assess the validity of competing theories concerning the causes of the occurrence of crime in space (Weisburd et al., 1993; Cohen et al., 2007; Haberman, 2017), see also Section 3.1. To properly assess the effects of covariates it is necessary to take into account the spatial correlation between street crimes, which leads to the third challenge.

To address the aforementioned first two challenges, we propose a semi-parametric regression model for the first-order intensity functions. Specifically, we propose a multiplicative model where the intensity function for each type of points is a product of a non-parametric component common to all types of points and a parametric component that models the influence of the covariates on the intensity function. The common non-parametric component models background factors such as population density or variation in intensity due to the layout of a city. To fit the model we propose a conditional composite likelihood function that does not depend on the non-parametric factor and is formally equivalent to multinomial logistic regression. We derive the asymptotic joint distribution of the resulting estimators and provide an estimator of the asymptotic covariance matrix. A few papers have considered building full parametric models for clustered multivariate point processes (Jalilian et al., 2015; Waagepetersen et al., 2016; Rajala et al., 2018). However, these parametric models impose restrictive assumptions that are difficult to verify in practice and fitting the models can be rather challenging when the number of point types is large.

Our approach is inspired by the case-control methodology introduced in Diggle and Rowlingson (1994) and further considered in Guan et al. (2008), Zimmerman et al. (2012) and Xu et al. (2019). However, we do not restrict attention to the bivariate case considered in these references. Our approach also has some resemblance to Diggle et al. (2005) who considered spatially varying risks of occurrence of one type of bacteria relative to the occurrence of other types. We, however, estimate relative risks using parametric models depending on covariates, where Diggle et al. (2005) applied non-parametric

kernel estimation. Diggle and Rowlingson (1994), Guan et al. (2008), Zimmerman et al. (2012), and Xu et al. (2019) further assume independence between different types of points and that points of at least one type forms a Poisson process while Diggle et al. (2005) and Zimmerman et al. (2012) assume that all the different types of points form Poisson processes which are independent. According to the third challenge mentioned above, we do not assume that any of the point processes are Poisson and we do not assume independence between different types of points. This significantly expands the applicability of the proposed methodology to ever-growing multivariate point pattern data collected in the big data era.

Our analysis of the street crime data clearly shows that the different types of street crimes are not distributed as Poisson processes and are also not independent of each other, see Figure A.4 for details. Thus, the inferential procedures considered in the existing work cited above, including Diggle and Rowlingson (1994), would not be valid even in the bivariate case. Table A.2 in our simulation study demonstrates that ignoring spatial correlations among different point patterns will lead to severe under-coverage of the resulting confidence intervals. Table A.4 of our crime data analysis also suggests that failure to take into account spatial correlations within and between types of points may result in misleading interpretations of the effects of some covariates.

Our theoretical investigation reveals that the asymptotic covariance matrix of our proposed estimator depends on the so-called pair correlation functions (PCFs) and cross PCFs of the multivariate point process, neither of which can be consistently estimated due to the common non-parametric component included in the model of the first-order intensity functions. A major novelty of our approach is our discovery that the asymptotic covariance matrix can be consistently estimated by an estimator expressed in terms of ratios of the PCFs and cross PCFs, but not the individual PCFs and cross PCFs themselves. In contrast to the individual PCFs and cross PCFs, it is possible to estimate these ratios consistently under the proposed model. However, the naive use of kernel estimators for PCF/cross-PCF ratios can still lead to serious under-coverage of the resulting confidence intervals. To further improve the quality of statistical inference, we developed a novel regularized non-parametric estimator for these ratios by imposing some mild shape constraints. To the best of our knowledge, no such regularized estimator has been studied in the literature. Consequently, valid statistical inferences can be performed for the estimated regression coefficients without restrictive parametric assumptions.

The proposed semi-parametric regression model for the first-order intensity functions allows us to study relative risks given by the ratios of the first-order intensity functions. Our estimators of ratios of PCFs and cross PCFs allow us to generalize this concept to the second-order setting. The application to street crime data in Section 6 shows that practical insights can be

gained by studying these PCF and cross PCF ratios. This is another novelty of our work. A final novel feature of our semi-parametric model is that we can combine information for all types of points to estimate the non-parametric component and subsequently obtain semi-parametric estimates of the intensity function for each type of points. This provides a more precise alternative to the usual non-parametric kernel intensity function estimator that is applied to each type of points separately.

The rest of the paper is organized as follows. Section 2 provides an overview of multivariate point processes with a focus on intensity and cross pair correlation functions. The semi-parametric model and its inference are introduced in Section 3 and theoretical investigations are given in Section 4. Simulation studies are presented in Section 5 and an application to Washington DC street crime data is given in Section 6. Concluding remarks are given in Section 7 and all technical proofs are collected in the supplementary material.

# 2 Background on multivariate point processes

Denote by $X = (X_1, \dots, X_p)$ a multivariate spatial point process, where $X_i$ is a random subset of $\mathbb{R}^d$ with the property that $X_i \cap B$ is of finite cardinality for all bounded $B \subseteq \mathbb{R}^d$ and $i = 1, \dots, p$. We assume that each $X_i$ is observed in a bounded window $W \subset \mathbb{R}^d$ and $X_i \cap X_j = \emptyset$ for any $i \neq j$. Assume that for each $m \geq 1$ and $i = 1, \dots, p$, there exists a non-negative function $\lambda_i^{(m)}(\cdot)$ such that

$$\mathrm{E} \sum_{\mathbf{u}_1,\dots,\mathbf{u}_m \in X_i}^{\neq} 1\left[\mathbf{u}_1 \in A_1, \dots, \mathbf{u}_m \in A_m\right] = \int_{\prod_{j=1}^m A_j} \lambda_i^{(m)}(\mathbf{u}_1, \dots, \mathbf{u}_m) \mathrm{d}\mathbf{u}_1 \cdots \mathrm{d}\mathbf{u}_m,$$

where $A_j \subset \mathbb{R}^d$, and $\sum^{\neq}$ indicates that $\mathbf{u}_1, \dots, \mathbf{u}_m$ are pairwise distinct. The function $\lambda_i^{(m)}(\cdot)$ is called the $m$'th order joint intensity function of $X_i$. When $m = 1$, the function $\lambda_i^{(1)}(\cdot)$ is referred to as the intensity and is denoted $\lambda_i(\cdot)$. Assume further that for each $n, m \geq 1$ and $i, j = 1, \dots, p$, there exists a non-negative function $\lambda_{ij}^{(m,n)}(\cdot, \cdot)$ such that

$$\begin{aligned} &\mathrm{E} \sum_{\mathbf{u}_1,\dots\mathbf{u}_m \in X_i}^{\neq} \sum_{\mathbf{v}_1,\dots,\mathbf{v}_n \in X_j}^{\neq} 1[\mathbf{u}_1 \in A_1, \dots, \mathbf{u}_m \in A_m, \mathbf{v}_1 \in B_1, \dots, \mathbf{v}_n \in B_n] \qquad \text{(A.1)} \\ &= \int_{\prod_{j=1}^m A_j} \int_{\prod_{j=1}^n B_j} \lambda_{ij}^{(m,n)}\left(\mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{v}_1, \dots, \mathbf{v}_n\right) \mathrm{d}\mathbf{u}_1 \cdots \mathrm{d}\mathbf{u}_m \mathrm{d}\mathbf{v}_1 \cdots \mathrm{d}\mathbf{v}_n, \end{aligned}$$

where $A_k \subset \mathbb{R}^d$ and $B_l \subset \mathbb{R}^d$ for $k = 1, \dots, m$ and $l = 1, \dots, n$. The function $\lambda_{ij}^{(m,n)}(\cdot, \cdot)$ is referred to as the $(m, n)$'th order cross intensity function between

$X_i$ and $X_j$, $i, j = 1, \ldots, p$. The normalized (cross) joint intensities $g_i^{(m)}(\cdot)$ and $g_{ij}^{(m,n)}(\cdot,\cdot)$ are defined as

$$g_i^{(m)}(\mathbf{u}_1, \ldots \mathbf{u}_m) = \lambda_i^{(m)}(\mathbf{u}_1, \ldots, \mathbf{u}_m) / \prod_{l=1}^{m} \lambda_i(\mathbf{u}_l), \text{ and}$$

$$g_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) = \frac{\lambda_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n)}{\prod_{l=1}^{m} \lambda_i(\mathbf{u}_l) \prod_{k=1}^{n} \lambda_j(\mathbf{v}_k)}, \tag{A.2}$$

provided the denominators on the right hand sides are positive (otherwise we define $g_i^{(m)}(\mathbf{u}_1, \ldots, \mathbf{u}_m) = 0$ and $g_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) = 0$). For $i \neq j$, $g_{ij}^{(1,1)}(\cdot,\cdot)$ is referred to as the cross pair correlation function (cross PCF) and $g_{ii}^{(1,1)}(\cdot,\cdot)$ coincides with $g_i^{(2)}(\cdot,\cdot)$ which is known as the pair correlation function (PCF). From now on, we write $g_i(\cdot,\cdot)$ for $g_i^{(2)}(\cdot,\cdot)$ and $g_{ij}(\cdot,\cdot)$ for $g_{ij}^{(1,1)}(\cdot,\cdot)$. The notion of cross joint intensities and their normalized versions can be generalized in an obvious way to joint cross intensities $\lambda_{i_1 i_2 \cdots i_k}^{(n_1,\ldots,n_k)}$ and normalized cross joint intensities $g_{i_1 i_2 \cdots i_k}^{(n_1,\ldots,n_k)}$ for $X_{i_1}, \ldots, X_{i_k}$ for any $k \geq 1$, $\{i_1, \ldots, i_k\} \subseteq \{1, 2, \ldots, p\}$, and integers $n_1, \ldots, n_k \geq 1$.

Suppose that a point from $X_i$ is observed at $\mathbf{u}$. Then $\lambda_j(\mathbf{v}) g_{ij}(\mathbf{u}, \mathbf{v})$ can be interpreted as the conditional intensity of $X_j$ at $\mathbf{v}$ given that $\mathbf{u} \in X_i$. Thus the cross PCF informs on how presence of a point in $\mathbf{u}$ affects the intensity of further points in $X_j$. In the special case when $X_i$ and $X_j$ are independent, $g_{ij}(\mathbf{u}, \mathbf{v}) \equiv 1$. If $X = (X_1, \ldots, X_p)$ consists of independent Poisson processes, we call $X$ a multivariate Poisson process. Then $\lambda_i^{(m)}(\mathbf{u}_1, \ldots, \mathbf{u}_m) = \prod_{l=1}^{m} \lambda_i(\mathbf{u}_l)$ and $\lambda_{ij}^{(m,n)}(\mathbf{u}_1, \ldots, \mathbf{u}_m, \mathbf{v}_1, \ldots, \mathbf{v}_n) = \prod_{l=1}^{m} \lambda_i(\mathbf{u}_l) \prod_{k=1}^{n} \lambda_j(\mathbf{v}_k)$. Consequently, $g_{ij}(\mathbf{u}, \mathbf{v}) = 1$, $i, j = 1, \ldots, p$, for a multivariate Poisson process which is the reference model of complete spatial independence.

Throughout the paper, we assume that the multivariate point process is second-order cross-intensity reweighted isotropic meaning that $g_{ij}(\mathbf{u}, \mathbf{v})$ depends only on the distance $\|\mathbf{u} - \mathbf{v}\|$. For this reason, we abuse notation and denote by $g_{ij}(r)$ the value of $g_{ij}(\mathbf{u}, \mathbf{v})$ when $\|\mathbf{u} - \mathbf{v}\| = r$. We often refer to so-called Campbell's formulae. For example, by standard measure theoretical arguments, the definition of $\lambda_i^{(m)}(\cdot)$ implies

$$\mathrm{E} \sum_{\mathbf{u}_1, \ldots, \mathbf{u}_m \in X_i}^{\neq} f(\mathbf{u}_1, \ldots, \mathbf{u}_m) = \int_{(\mathbb{R}^d)^m} f(\mathbf{u}_1, \ldots, \mathbf{u}_m) \lambda_i^{(m)}(\mathbf{u}_1, \ldots, \mathbf{u}_m) \mathrm{d}\mathbf{u}_1 \cdots \mathrm{d}\mathbf{u}_m$$

for any non-negative function $f$ on $(\mathbb{R}^d)^m$. Similar Campbell formulae hold for the cross joint intensities.

# 3 Semi-parametric multinomial logistic regression

In this section, we detail the proposed semi-parametric model and the multinomial logistic regression approach to statistical inference. Formal asymptotic considerations are deferred to Section 4.

## 3.1 Semi-parametric model

For spatial point pattern data in an environment like a city, the intensity function can be rather complex due to the city layout and variations in population density. To overcome this difficulty, we follow Diggle and Rowlingson (1994) and assume that for each point pattern $X_i$, the intensity function takes the multiplicative form

$$\lambda_i(\mathbf{u};\boldsymbol{\gamma}_i) = \lambda_0(\mathbf{u}) \exp\left[\boldsymbol{\gamma}_i^{\mathsf{T}} \mathbf{z}(\mathbf{u})\right], i = 1, \ldots, p, \tag{A.3}$$

where $\lambda_0(\cdot)$ is an unknown background intensity function, $\mathbf{z}(\mathbf{u})$ is a $q$-dimensional vector of spatial covariates at location $\mathbf{u}$, and $\boldsymbol{\gamma}_i \in \mathbb{R}^q$ is the vector of regression parameters. The background intensity $\lambda_0(\cdot)$ can be interpreted as the spatial effects of latent factors such as the urban structure and population density and is assumed to be common for all point types. The model (A.3) is also closely related to the Cox regression model widely used for the conditional intensity in survival analysis (Cox, 1972).

In case of crime, several competing theories regarding causes of crime exist (Weisburd et al., 1993; Haberman, 2017). The crime general theory asserts that general factors drive crimes regardless of crime type. Accordingly, the proportions of crime types should be roughly constant across space. The crime specific theory instead asserts that different crimes depend on different factors, including environmental factors, which should lead to a more segregated occurrence of crime types with some crimes being more frequent in some areas than others. Our background intensity accommodates the effects of environmental factors with a common effect for all crimes. Next, based on (A.3) we can derive conditional probabilities which precisely model the proportions of crime types for each location $\mathbf{u}$ and how they depend on spatial covariates, see (A.5) in Section 3.2.

Following Cohen et al. (2007), crime relevant spatial covariates may be categorized as crime attractors or crime displacements covariates. For example, distances to places like bars, parking lots, and music venues can be viewed as crime attractor covariates. Another example is the indicator of neighborhoods where policing of minor offenses are not strictly enforced. The spatial intensity of policing is an example of a crime displacement covariate since increased police activity in one location merely displaces crime to other locations rather than reducing crime overall (Ratcliffe, 2002). In Section 6 we model Washington DC street crime by demographical covariates along with

the distance to the nearest police station as a crime displacement covariate. The demographical covariates are not as such crime attractors but can be used to study whether the socio-economic status of a neighborhood has an impact on the occurrence of crimes.

The parameters $\gamma_i$ are not identifiable since subtracting $\mathbf{k}^\mathsf{T}\mathbf{z}(\mathbf{u})$ from the log-linear model for some $\mathbf{k} \in \mathbb{R}^q$ while redefining $\lambda_0(\mathbf{u}) := \lambda_0(\mathbf{u}) \exp[\mathbf{k}^\mathsf{T}\mathbf{z}(\mathbf{u})]$ does not change the intensities $\lambda_i$. To address this issue, we pick a baseline process, say $X_p$, and define identifiable parameters $\boldsymbol{\beta}_i = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_p$ for $i = 1, \ldots, p-1$. Further, without loss of generality, we may assume $\boldsymbol{\gamma}_p = 0$ in which case $\lambda_0(\cdot)$ becomes the intensity of the baseline process. Using the new parameterization, we can evaluate the effects of the covariates $\mathbf{z}(\cdot)$ relative to the baseline process $X_p$ similar to matched case-control studies and Cox regression in survival analysis.

Although estimation of the $\lambda_i(\cdot)$ is not our primary concern, note that given estimates $\hat{\lambda}_0(\cdot)$ and $\widehat{\boldsymbol{\beta}}_i$, we may estimate $\lambda_i(\cdot)$ by

$$\hat{\lambda}_i(\mathbf{u}) = \hat{\lambda}_0(\mathbf{u}) \exp\left[\widehat{\boldsymbol{\beta}}_i^\mathsf{T}\mathbf{z}(\mathbf{u})\right]. \tag{A.4}$$

If type $i$ points are rare, this estimate may be advantageous compared to an intensity estimate based only on type $i$ points since we can borrow strength by estimating $\lambda_0(\cdot)$ using all types of points, see also Section 6.2.

In terms of criminology research, a solid amount of literature states that crime is clustered in micro-places called hot spots (see Haberman, 2017, and the references therein). Identification of hot spots may help police departments to allocate their resources properly (Buerger et al., 1995) and hot spot policing reduces crime (Braga et al., 2014). Numerous non-parametric methods have been developed to identify the hot spots, including kernel density estimation (Ratcliffe, 2004; Gorr and Lee, 2015). The estimator (A.4) adds to the existing hot spot detection methods by enhancing non-parametric kernel estimation with additional information from spatial covariates. Forecasting future occurrences of crime is another challenge to police departments. The Broken Windows theory of crime (Wilson and Kelling, 1982) states that the tolerance of 'soft' crimes in a neighborhood attracts criminals, hence the presence of 'soft' crimes can be used to forecast 'serious' crimes, see Cohen et al. (2007) and Gorr and Lee (2015). By a straightforward expansion of model (A.3) to a space-time setup, one could use the estimator (A.4) to forecast 'serious' crimes using an estimate of current soft crime intensity as a covariate.

## 3.2 Multinomial logistic regression

We tackle the estimation of model (A.3) by conditional composite likelihood where we use the reparametrization in terms of the $\boldsymbol{\beta}_i$ from the previous section. Conditioned on that an event is observed at location $\mathbf{u}$, under

model (A.3), the probability that it is from the point process $X_i$ is

$$\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}) = \frac{\lambda_i(\mathbf{u};\boldsymbol{\gamma}_i)}{\sum_{k=1}^{p}\lambda_k(\mathbf{u};\boldsymbol{\gamma}_k)} = \begin{cases} \frac{\exp\left[\boldsymbol{\beta}_i^{\mathsf{T}}\mathbf{z}(\mathbf{u})\right]}{1+\sum_{k=1}^{p-1}\exp\left[\boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{z}(\mathbf{u})\right]}, & i=1,\ldots,p-1, \\ \frac{1}{1+\sum_{k=1}^{p-1}\exp\left[\boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{z}(\mathbf{u})\right]}, & i=p, \end{cases} \tag{A.5}$$

which does not depend on the background intensity $\lambda_0(\cdot)$. To estimate $\boldsymbol{\beta}$, we define the multinomial conditional composite likelihood as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{p}\prod_{\mathbf{u}\in X_i\cap W}\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}).$$

This is formally equivalent to a multinomial logistic regression likelihood function. It is a composite likelihood function because it ignores possible dependencies between types of points given their locations. The log multinomial conditional composite likelihood function is of the form

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{p}\sum_{\mathbf{u}\in X_i\cap W}\left[\boldsymbol{\beta}_i^{\mathsf{T}}\mathbf{z}(\mathbf{u}) - \log\left(1+\sum_{k=1}^{p-1}\exp\left[\boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{z}(\mathbf{u})\right]\right)\right], \tag{A.6}$$

and the conditional composite likelihood estimator is defined as $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$.

## 3.3 Estimation of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$

In this section, we consider the problem of estimating the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$, which is challenging due to the highly complex between- and within-type correlation structure of the multivariate point process.

We denote by $\mathrm{E}(\cdot)$ and $\mathrm{Var}(\cdot)$, expectation and variance with respect to the data generating distribution of $X = (X_1,\ldots,X_p)$, where we assume the intensity function of $X_i$ is of the form (A.3) with the parameters $\boldsymbol{\gamma}_i$ given by some specific values $\boldsymbol{\gamma}_i^* \in \mathbb{R}^q$ and we let $\boldsymbol{\beta}_i^* = \boldsymbol{\gamma}_i^* - \boldsymbol{\gamma}_p^*$ for $i = 1,\ldots,p-1$. In this section and the rest of the paper we will refer to the 'pooled' point process $X^{\mathrm{pl}} = \cup_{k=1}^{p}X_i$, whose intensity function and PCF are

$$\lambda^{\mathrm{pl}}(\mathbf{u};\boldsymbol{\gamma}) = \sum_{k=1}^{p}\lambda_k(\mathbf{u};\boldsymbol{\gamma}_k) \text{ and} \tag{A.7}$$

$$g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta},g) = \sum_{l=1}^{p}\sum_{l'=1}^{p}\mathrm{p}_l(\mathbf{u};\boldsymbol{\beta}_l)\mathrm{p}_{l'}(\mathbf{v};\boldsymbol{\beta}_l)g_{ll'}(\mathbf{u},\mathbf{v}).$$

The "$g$" inside $g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta},g)$ signifies the dependence on the $g_{ll'}$. We use in the following the short forms $\lambda_k^*(\cdot)$, $\mathrm{p}_l^*(\cdot)$, $\lambda^{\mathrm{pl}}(\cdot)$, and $g^{\mathrm{pl}}(\cdot,\cdot)$ for $\lambda_k(\cdot;\boldsymbol{\gamma}_i^*)$, $\mathrm{p}_l(\cdot;\boldsymbol{\beta}_l^*)$, $\lambda^{\mathrm{pl}}(\cdot;\boldsymbol{\gamma}^*)$, and $g^{\mathrm{pl}}(\cdot,\cdot;\boldsymbol{\beta}^*,g)$.

It is trivial to see that $\ell(\boldsymbol{\beta})$ in (A.6) is a concave function of $\boldsymbol{\beta}$ and thus maximizing $\ell(\boldsymbol{\beta})$ is equivalent to solving the estimating equation $\mathbf{e}(\boldsymbol{\beta}) = \mathbf{0}$ where

$$\mathbf{e}(\boldsymbol{\beta}) = \left[\mathbf{e}_1(\boldsymbol{\beta})^\mathsf{T}, \ldots, \mathbf{e}_{p-1}(\boldsymbol{\beta})^\mathsf{T}\right]^\mathsf{T}, \text{ with} \tag{A.8}$$

$$\mathbf{e}_i(\boldsymbol{\beta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_i}\ell(\boldsymbol{\beta}) = \sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u}) - \sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W}\frac{\mathbf{z}(\mathbf{u})\exp\left[\boldsymbol{\beta}_i^\mathsf{T}\mathbf{z}(\mathbf{u})\right]}{1+\sum_{k=1}^{p-1}\exp\left[\boldsymbol{\beta}_k^\mathsf{T}\mathbf{z}(\mathbf{u})\right]}, \tag{A.9}$$

for $i = 1, \ldots, p-1$. According to standard estimating equation theory (see, for example, Crowder, 1986) and formally justified by Theorem 2 in Section 4.1, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is of the form

$$\left[\mathbf{S}(\boldsymbol{\beta}^*)\right]^{-1}\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g)\left[\mathbf{S}(\boldsymbol{\beta}^*)\right]^{-1},$$

where $\mathbf{S}(\boldsymbol{\beta}^*) = \mathrm{E}\left[-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}^\mathsf{T}}\mathbf{e}(\boldsymbol{\beta}^*)\right]$ is the so-called sensitivity matrix and $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g) = \mathrm{Var}\left[\mathbf{e}(\boldsymbol{\beta}^*)\right]$ is the covariance matrix of $\mathbf{e}(\boldsymbol{\beta}^*)$. The "$g$" inside $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g)$ emphasizes that $\mathrm{Var}\left[\mathbf{e}(\boldsymbol{\beta}^*)\right]$ depends on the underlying cross PCFs.

The explicit forms of $\mathbf{S}(\boldsymbol{\beta}^*)$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g)$ are derived in Section A of the supplementary material. The $(i,j)$'th block of $\mathbf{S}(\boldsymbol{\beta}^*)$ is of the form

$$\mathbf{S}(\boldsymbol{\beta}^*)_{ij} = \begin{cases} \int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\left[1-\mathrm{p}_i^*(\mathbf{u})\right]\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u} & i=j, \\ -\int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_j^*(\mathbf{u})\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u} & i\neq j, \end{cases} \tag{A.10}$$

for $i, j = 1, \ldots, p-1$ with $\mathbf{Z}(\mathbf{u},\mathbf{v}) = \mathbf{z}(\mathbf{u})\mathbf{z}(\mathbf{v})^\mathsf{T}$. The $(i,j)$'th block of $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g)$ corresponding to $\mathrm{Cov}\left[\mathbf{e}_i(\boldsymbol{\beta}^*), \mathbf{e}_j(\boldsymbol{\beta}^*)\right]$ takes the form

$$\begin{aligned}&\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g)_{ij} = \\ &\mathbf{S}(\boldsymbol{\beta}^*)_{ij} + \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_j^*(\mathbf{v})g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)T_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}, g)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v},\end{aligned} \tag{A.11}$$

where the function $T_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)$ is defined as

$$1 + \frac{g_{ij}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} - \sum_{l=1}^{p}\frac{\left[\mathrm{p}_l^*(\mathbf{v})g_{il}(\mathbf{u},\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u})g_{jl}(\mathbf{u},\mathbf{v})\right]}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)}. \tag{A.12}$$

By Campbell's formulae we can approximate $\mathbf{S}(\boldsymbol{\beta}^*)$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*, g)$ by $\widehat{\mathbf{S}}(\boldsymbol{\beta}^*)$ and $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}^*, g)$, whose $(i,j)$'th blocks are defined as

$$\widehat{\mathbf{S}}(\boldsymbol{\beta}^*)_{ij} = \begin{cases} \sum_{\mathbf{u}\in X^{\mathrm{pl}}}\mathbf{Z}(\mathbf{u},\mathbf{u})\left[1-\mathrm{p}_i^*(\mathbf{u})\right]\mathrm{p}_i^*(\mathbf{u}) & i=j, \\ -\sum_{\mathbf{u}\in X^{\mathrm{pl}}}\mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i^*(\mathbf{u})\mathrm{p}_j^*(\mathbf{u}) & i\neq j, \end{cases} \tag{A.13}$$

$$\widehat{\mathbf{\Sigma}}(\boldsymbol{\beta}^*, g)_{ij} = \widehat{\mathbf{S}}(\boldsymbol{\beta}^*)_{ij} + \sum_{\mathbf{u},\mathbf{v}\in X^{\mathrm{pl}}:\|\mathbf{u}-\mathbf{v}\|\leq R}^{\neq} \mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i^*(\mathbf{u})\mathrm{p}_j^*(\mathbf{v})T_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g), \tag{A.14}$$

for $i, j = 1, \ldots, p-1$. Here $R$ denotes a 'correlation range' such that $g_{ij}(r) \approx 1$ for $r > R$. In practice we replace $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}$ in (A.13)-(A.14) and the notion "$g$" emphasizes their dependence on the underlying cross PCFs, which will be replaced by non-parametric estimators discussed in the next sections.

## 3.4 Naive kernel estimation of cross PCF ratios

The empirical covariance matrix (A.14) depends critically on cross PCFs which need to be estimated. The definition of a cross PCF in (A.2) suggests that its estimation requires consistent estimators of the intensity functions which are not available under the model (A.3), since $\lambda_0(\cdot)$ is unknown. However, a closer look at (A.12) reveals that for computation of (A.12) it suffices to estimate the cross PCFs up to a common multiplicative factor, or, equivalently, to estimate ratios of cross PCFs, i.e.

$$g_{ij,kl}(\mathbf{u},\mathbf{v}) = g_{ij}(\mathbf{u},\mathbf{v})/g_{kl}(\mathbf{u},\mathbf{v}), \quad i,j = 1,\ldots,p, \tag{A.15}$$

for some arbitrary fixed pair of types of points $k$ and $l$. These ratios are also of great interest in their own right as they measure the strength of correlation among two types of points relative to the strength of correlation between two other types of points. Consider the quantity

$$F_{ij}(r; b, \boldsymbol{\beta}) = \sum_{\substack{\mathbf{u}\in X_i\cap W \\ \mathbf{v}\in X_j\cap W}}^{\neq} \frac{k_b(||\mathbf{u}-\mathbf{v}|| - r)}{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta})}, \tag{A.16}$$

where $k_b(\cdot) = k(\cdot/b)/b$ with $k(\cdot)$ being a kernel function defined on a bounded interval in $\mathbb{R}$ and $b > 0$ is a bandwidth. Using Campbell's formula together with equation (A.5), it follows that under model (A.3),

$$\mathrm{E}[F_{ij}(r; b, \boldsymbol{\beta}^*)] = \int_{W^2} \lambda^{\mathrm{pl}}(\mathbf{u})\lambda^{\mathrm{pl}}(\mathbf{v})g_{ij}(\mathbf{u},\mathbf{v})k_b(||\mathbf{u}-\mathbf{v}|| - r)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v},$$

where $\lambda^{\mathrm{pl}}$ was defined in (A.7). Under suitable conditions and appropriately chosen bandwidth $b$, it is reasonable to expect that $F_{ij}(r; b, \widehat{\boldsymbol{\beta}}) \approx c(r)g_{ij}(r)$, where

$$c(r) = \int_{W^2} \lambda^{\mathrm{pl}}(\mathbf{u})\lambda^{\mathrm{pl}}(\mathbf{v})k_b(||\mathbf{u}-\mathbf{v}|| - r)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v},$$

is a multiplicative factor which, as desired, does not depend on $ij$. Consequently,

$$\hat{g}^{\mathrm{n}}_{ij,kl}(r; b, \widehat{\boldsymbol{\beta}}) = F_{ij}(r; b, \widehat{\boldsymbol{\beta}})/F_{kl}(r; b, \widehat{\boldsymbol{\beta}}) \tag{A.17}$$

becomes an estimator of (A.15).

Note that the estimator (A.17) does not depend on the unknown background intensity $\lambda_0(\cdot)$. The superscript "n" stands for "naive" kernel estimator (a regularized estimator will be introduced in the next section). Our Theorem 3 in Section 4.2 states that under mild conditions, (A.17) is consistent for $g_{ij,kl}(r)$. The naive plug-in estimator $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ is then obtained by replacing $\boldsymbol{\beta}^*$ and the cross PCFs in (A.12) by $\widehat{\boldsymbol{\beta}}$ and the estimators (A.17) of cross PCF ratios. For the rest of the paper, we use the PCF of the baseline process $X_p$ as the fixed denominator in (A.15), letting $k = l = p$.

## 3.5 Regularized cross PCF ratio estimators

Even though Theorem 3 in Section 4.2 shows that the naive kernel estimator (A.17) is consistent under mild conditions, the finite sample performance of the plug-in estimators $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ may be unsatisfactory due to high variabilities of the $\hat{g}^{\mathrm{n}}_{ij,pp}(\cdot)$'s. In particular, our numerical experiments suggest that when the number of observed points is small, some diagonal elements of the $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ may be negative, resulting in negative estimated variances for some components of $\widehat{\boldsymbol{\beta}}$.

We notice that this phenomenon is mainly caused by the existence of a large number of negative values of $T_{ii}(\mathbf{u}, \mathbf{v}; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ when $\|\mathbf{u} - \mathbf{v}\|$ is large, leading to negative values in the diagonal of $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})_{ii}$ as defined in (A.14). This issue can be resolved or alleviated by imposing constraints on the cross PCFs. In this paper, we impose the following constraints

$$g_{ij}(r) \leq \sqrt{g_{ii}(r) g_{jj}(r)} \quad \text{for } r \geq R^*, \quad i, j = 1, \ldots, p, \tag{A.18}$$

for some $R^* \geq 0$. Intuitively, condition (A.18) means that for lags $r \geq R^*$, the spatial correlation between different point processes is weaker than the (geometric) average of spatial correlation within each individual point process. Condition (A.18) is not necessarily true for any multivariate point process but is indeed valid with $R^* = 0$ for a large class of multivariate log Gaussian Cox processes (Waagepetersen et al., 2016) (see also Section 5) and for a large subclass of the multivariate shot-noise Cox processes proposed in Jalilian et al. (2015).

To enforce the constraint (A.18) on the naive kernel estimators, let $\widehat{\mathbf{G}}^{\mathrm{n}}_r$ be a $p \times p$ matrix whose $(i, j)$'th element is $\hat{g}^{\mathrm{n}}_{ij,pp}(r; b, \widehat{\boldsymbol{\beta}})$ for some $r > R^*$. The regularized non-parametric estimators, denoted as $\hat{g}^{\mathrm{r}}_{ij,pp}(r; b, \widehat{\boldsymbol{\beta}})$, are collected in the matrix $\widehat{\mathbf{G}}^{\mathrm{r}}_r$ obtained by

$$\widehat{\mathbf{G}}^{\mathrm{r}}_r = \arg \min_{\Theta = [\theta_{ij}]_{ij}} \left\| \Theta - \widehat{\mathbf{G}}^{\mathrm{n}}_r \right\|_F^2, \text{ with } \theta_{ij} = \theta_{ji}, \theta_{pp} = 1, \theta_{ij}^2 \leq \theta_{ii}\theta_{jj}, \tag{A.19}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

It can be shown (Section B in the supplementary material) that for $\|\mathbf{u}-\mathbf{v}\| > R^*$, the plug-in estimator with $\hat{g}^{\mathrm{r}}_{ij,pp}(\cdot)$'s satisfies

$$\min_{1\le i\le p} T_{ii}\left(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}},\hat{g}^{\mathrm{r}}\right) \ge 1 - \max_{1\le l\le p} \hat{g}^{\mathrm{r}}_{ll,pp}(\|\mathbf{u}-\mathbf{v}\|;b,\widehat{\boldsymbol{\beta}})/g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}},\hat{g}^{\mathrm{r}}).$$

In contrast, using the naive $\hat{g}^{\mathrm{n}}_{ij,pp}(\cdot)$'s, we can only achieve the lower bound

$$1 - \frac{2\max_{1\le l,l'\le p} \hat{g}^{\mathrm{n}}_{ll',pp}(\|\mathbf{u}-\mathbf{v}\|;b,\widehat{\boldsymbol{\beta}}) - \min_{1\le l\le p} \hat{g}^{\mathrm{n}}_{ll,pp}(\|\mathbf{u}-\mathbf{v}\|;b,\widehat{\boldsymbol{\beta}})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}},\hat{g}^{\mathrm{n}})}.$$

Note that the first lower bound above can be much larger than the second lower bound, which partly explains why the regularized cross PCF ratio estimators would produce much fewer large negative $T_{ii}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}},\hat{g}^{\mathrm{r}})$ when $\|\mathbf{u}-\mathbf{v}\| > R^*$, leading to a better covariance matrix estimator. In Section E in the supplementary material, we give a more detailed demonstration through numerical examples.

**Remark 1**
Our numerical investigations suggest that the regularized estimator is quite robust to the choice of $R^*$. The simplest choice is to set $R^* = 0$. Otherwise we recommend to use $R^* = \arg\min_{r\ge 0}\{\max_i P_{ii}(r) > 0.05\}$, where $P_{ii}(r)$ is the percentage of pairs $(\mathbf{u},\mathbf{v})$ that give $T_{ii}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}},\hat{g}^{\mathrm{n}}) < 0$ within the set $\{(\mathbf{u},\mathbf{v}) : \mathbf{u},\mathbf{v}\in X^{\mathrm{pl}} \text{ and } \|\mathbf{u}-\mathbf{v}\| \in (r-h, r+h)\}$. In other words, when the percentage of negative $T_{ii}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}},\hat{g}^{\mathrm{n}})$'s exceeds 5% around the distance $R^*$ for any $i = 1,\dots,p$, the restriction (A.18) will be enforced for $r > R^*$.

# 4 Asymptotic properties

In this section we study asymptotic properties of $\widehat{\boldsymbol{\beta}}$ when $X$ is observed on a sequence of increasing windows $W_n$. Denote by $\mathbf{e}^{(n)}(\boldsymbol{\beta})$ the multinomial estimating function (A.8) evaluated on $W_n$ and by $\widehat{\boldsymbol{\beta}}_n$ the sequence of estimators obtained as solutions to $\mathbf{e}^{(n)}(\boldsymbol{\beta}) = \mathbf{0}$. The quantities $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$, $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, g)$ and $\mathbf{S}_n(\boldsymbol{\beta}^*)$ are defined as in Section 3.3 with $W = W_n$ for the last two. We also define 'averaged' versions, $\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}^*, g) = \boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, g)/|W_n|$ and $\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*) = \mathbf{S}_n(\boldsymbol{\beta}^*)/|W_n|$. Finally, $\|\mathbf{A}\|_{\max} = \max_{ij} a_{ij}$ denotes the maximum norm of $\mathbf{A} = [a_{ij}]_{ij}$.

## 4.1 Consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$

The following conditions are sufficient to establish the consistency of $\widehat{\boldsymbol{\beta}}_n$.

C1 $W_1 \subset W_2 \subset \ldots$ and $|\bigcup_{l=1}^{\infty} W_l| = \infty$.

C2 There exists an $0 < K_1 < \infty$ such that $\|\mathbf{z}(\mathbf{u})\|_{\max}$, $\lambda_i^*(\mathbf{u})$ and $g_{ij}(\mathbf{u}, \mathbf{v})$ are bounded above by $K_1$ for all $\mathbf{u}, \mathbf{v} \in \cup_{l=1}^{\infty} W_l$ and $i, j = 1, \ldots, p$.

C3 There exists an $0 < K_2 < \infty$ so that $\int_{\mathbb{R}^d} |g_{ij}(\mathbf{0}, \mathbf{u}) - 1| \mathrm{d}\mathbf{u} < K_2$ for all $i, j = 1, \ldots, p$.

C4 $\liminf_{n\to\infty} \lambda_{\min} \left[ |W_n|^{-1} \int_{W_n} \mathbf{Z}(\mathbf{u}, \mathbf{u}) \lambda_i^*(\mathbf{u}) \mathrm{p}_p(\mathbf{u}; \boldsymbol{\beta}^*) \mathrm{d}\mathbf{u} \right] > 0$ for $i = 1, \ldots, p - 1$, where $\lambda_{\min}[A]$ denotes the minimal eigenvalue of a matrix $A$.

C1-C3 are mild conditions that have been widely used in the literature. C4 ensures that the averaged sensitivity matrix $\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)$ is invertible for sufficiently large $n$, which is commonly used in the estimating equation literature. Heuristically speaking, C4 requires that sufficient information regarding $\boldsymbol{\beta}^*$ need to be accumulated across space and it could be violated if $\mathbf{z}(\cdot)$ is close to constant.

**Theorem 1**

*Under conditions C1-C4, there exists a sequence of solutions $\widehat{\boldsymbol{\beta}}_n$ to the estimating equation $\mathbf{e}_n(\boldsymbol{\beta}) = \mathbf{0}$ for which*

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}^*, \text{ as } n \to \infty.$$

The proof of Theorem 1 is given in Section C.1 of the supplementary material.

Next, we proceed to establish asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$. Following Biscio and Waagepetersen (2019), we define an $\alpha$-mixing coefficient by regarding $X$ as a marked point process with points in $\mathbb{R}^d$ and marks in $M = \{1, \ldots, p\}$. That is, a point $\mathbf{u}$ in $X_i$ corresponds to a marked point $(\mathbf{u}, i)$. We then for sets $A \subseteq \mathbb{R}^d$ and $B \subseteq M$, define $X_{A,B} = X \cap A \times B$ as the set of marked points in $X$ whose 'point parts' fall in $A$ and whose marks fall in $B$.

To define the $\alpha$-mixing coefficient for $X$ we first define an $\alpha$-mixing coefficient for two $\sigma$-algebras $\mathcal{F}$ and $\mathcal{G}$ on a common probability space,

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup\{|\mathrm{P}(F \cap G) - \mathrm{P}(F)\mathrm{P}(G)| \colon F \in \mathcal{F}, G \in \mathcal{G}\}.$$

Define $d(\mathbf{u}, \mathbf{v}) = \max\{|u_i - v_i| \colon 1 \le i \le d\}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. The marked point process $\alpha$-mixing coefficient of $X$ is then for $s, c_1, c_2 \ge 0$ given by

$$\alpha_{c_1,c_2}^X(s) = \sup\{\alpha(\sigma(X_{E_1,M}), \sigma(X_{E_2,M})) \colon \\ E_1 \subset \mathbb{R}^d, E_2 \subset \mathbb{R}^d, |E_1| \le c_1, |E_2| \le c_2, d(E_1, E_2) \ge s\},$$

where $|A|$ is the Lebesgue measure of $A$ and $d(A, B) = \inf\{d(\mathbf{u}, \mathbf{v}) \colon \mathbf{u} \in A, \mathbf{v} \in B\}$. This coefficient measures the dependence between $X \cap E_1 \times M$ and $X \cap E_2 \times M$, where $E_1$ and $E_2$ are arbitrary Borel subsets of $\mathbb{R}^d$ with volumes less than $c_1$ and $c_2$ and separated by the distance $s$.

The following extra conditions are needed to establish asymptotic normality.

N1 There exists $\varepsilon > 0$ such that $\alpha_{2,\infty}^{X}(s) = O(1/s^{d+\varepsilon})$.

N2 There exist an integer $m > 2d/\varepsilon + 2$ and $C_g$ such that $g_{i_1 i_2 \cdots i_k}^{(n_1,n_2,\ldots,n_k)}(\cdot,\ldots,\cdot) \leq C_g$ for any $\{i_1,\ldots,i_k\} \subseteq \{1,2,\ldots,p\}$, and integers $n_1 + \cdots + n_k \leq m$.

N3 It holds that $\liminf_{n\to\infty} \lambda_{\min}\left[\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}^*, g)\right] > 0$.

N1 is a standard mixing condition that, e.g., holds for multivariate log Gaussian Cox processes with PCFs of bounded range (meaning $g_{ij}(r) = 1$ when $r$ is larger than some $0 \leq R < \infty$) or Poisson cluster point processes with sufficiently quickly decaying cluster densities. Condition N2 of bounded normalized joint cross intensities is satisfied for most multivariate point process models. N3 is a standard condition which ensures that the variance of $|W_n|^{-1}\mathbf{e}^{(n)}(\boldsymbol{\beta})$ is not degenerate for sufficiently large $n$.

**Theorem 2**
*Under conditions C1-C4 and N1-N3, as $n \to \infty$, we have that*

$$|W_n|^{1/2}\bar{\boldsymbol{\Sigma}}_n^{-1/2}(\boldsymbol{\beta}^*, g)\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{d} N\left(0, \mathbf{I}_{(p-1)q}\right).$$

The proof of Theorem 2 is given in Section C.2 of the supplementary material.

Theorem 2 implies that the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$ is of the form

$$|W_n|^{-1}\left[\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\right]^{-1}\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}^*, g)\left[\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\right]^{-1} = \left[\mathbf{S}_n(\boldsymbol{\beta}^*)\right]^{-1}\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, g)\left[\mathbf{S}_n(\boldsymbol{\beta}^*)\right]^{-1},$$

where the left hand side suggests that the variance of $\widehat{\boldsymbol{\beta}}_n$ is of order $|W_n|^{-1}$. Based on Theorem 2, one can make statistical inference regarding $\boldsymbol{\beta}^*$ and other quantities of interest. For example, as in classical multinomial regression models, one may be interested in the probability of a certain event at a given location, i.e., $\mathrm{p}_i^*(\mathbf{u})$, or the log-odds $\log\frac{\mathrm{p}_i^*(\mathbf{u})}{\mathrm{p}_p^*(\mathbf{u})} = \mathbf{z}(\mathbf{u})^{\mathsf{T}}\boldsymbol{\beta}_i^*$ for $i = 1,\ldots,p-1$.

Denote by $\mu(\boldsymbol{\beta}^*)$ a parameter of interest where $\mu : \mathbb{R}^{(p-1)q} \to \mathbb{R}$ is differentiable. A simple application of the Delta method gives for $0 < \alpha < 1$ the $100\alpha\%$ approximate confidence interval for $\mu(\boldsymbol{\beta}^*)$,

$$\mu(\widehat{\boldsymbol{\beta}}) \pm z_{1-\alpha/2}\sqrt{\left[\boldsymbol{\mu}^{(1)}(\widehat{\boldsymbol{\beta}})\right]^{\mathsf{T}}\left[\widehat{\mathbf{S}}_n(\widehat{\boldsymbol{\beta}})\right]^{-1}\widehat{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{\beta}}, \hat{g}_n^{\mathrm{r}})\left[\widehat{\mathbf{S}}_n(\widehat{\boldsymbol{\beta}})\right]^{-1}\boldsymbol{\mu}^{(1)}(\widehat{\boldsymbol{\beta}})}, \quad \text{(A.20)}$$

where $z_\alpha$ is the $100\alpha$th percentile of a standard normal distribution, $\boldsymbol{\mu}^{(1)}(\boldsymbol{\beta}) = \mathrm{d}\mu(\boldsymbol{\beta})/\mathrm{d}\boldsymbol{\beta}$, and estimators of $\boldsymbol{\beta}$ and cross PCFs have been plugged into (A.13) and (A.14), see also Sections 3.4-3.5 and Section 4.2.

### 4.2 Asymptotic properties of $\hat{g}^{\mathrm{n}}_{ij,kl}(r;b,\widehat{\boldsymbol{\beta}})$ and $\hat{g}^{\mathrm{r}}_{ij,kl}(r;b,\widehat{\boldsymbol{\beta}})$

Let $W_n$ and $b_n$ be sequences of observation windows and bandwidths, respectively. Denote by $\hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n)$ a sequence of estimators that is given by

$$\hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n) = F_{ij,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n)/F_{kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n),$$

where the $F_{ij,n}$'s are defined as in (A.16) with $W = W_n$. In this subsection, we show that $\hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n)$ is a consistent estimator of $g_{ij,kl}(r)$ for any $i,j = 1,\ldots,p$, under the following conditions.

K1 For $i,j = 1,\ldots,p$, the cross joint intensity $g^{(2,2)}_{ij}$ is translation invariant: $g^{(2,2)}_{ij}(\mathbf{u}_1,\mathbf{u}_2,\mathbf{v}_1,\mathbf{v}_2) = g^{(2,2)}_{ij}(\mathbf{0},\mathbf{u}_2-\mathbf{u}_1,\mathbf{v}_1-\mathbf{u}_1,\mathbf{v}_2-\mathbf{u}_1)$, $\mathbf{u}_1,\mathbf{u}_2,\mathbf{v}_1,\mathbf{v}_2 \in \cup_{l=1}^{\infty} W_l$, and there exists $K_3 < \infty$ so that $\int_{\mathbb{R}^d} |g^{(2,2)}_{ij}(\mathbf{0},\mathbf{u},\mathbf{v},\mathbf{w}+\mathbf{u}) - g_{ij}(\mathbf{0},\mathbf{v})g_{ij}(\mathbf{0},\mathbf{w})|\mathrm{d}\mathbf{u} < K_3$ for all $\mathbf{u},\mathbf{v},\mathbf{w} \in \cup_{l=1}^{\infty} W_l$.

K2 There exists $K_4 < \infty$ so that $g^{(m,n)}_{ij}(\mathbf{u}_1,\ldots,\mathbf{u}_m,\mathbf{v}_1,\ldots,\mathbf{v}_n) < K_4$ for all $\mathbf{u}_m,\mathbf{v}_n \in \cup_{l=1}^{\infty} W_l$ with $m+n<4$ and $i,j = 1,\ldots,p$.

K3 The kernel function $k(\cdot)$ has a compact support $[-1,1]$ and the bandwidth $b_n$ satisfies that (a) $b_n \to 0$; and (b) $|W_n|b_n \to \infty$ as $|W_n| \to \infty$.

**Theorem 3**
*Under conditions C2 and K1 - K3, one has that*

$$\hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}) \xrightarrow{p} g_{ij,kl}(r),\ \textit{as } n\to\infty,\ \textit{for } i,j,k,l = 1,\ldots,p. \tag{A.21}$$

*If we further assume that constraint* (A.18) *holds true, then*

$$\hat{g}^{\mathrm{r}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}) \xrightarrow{p} g_{ij,kl}(r),\ \textit{as } n\to\infty,\ \textit{for } i,j,k,l = 1,\ldots,p. \tag{A.22}$$

The proof of Theorem 3 is given in Section C.3 of the supplementary material.

## 5 Simulation studies

In this section we assess the finite sample performance of the proposed methodology through simulation studies. To evaluate our estimators we need to simulate from a model with known forms of the intensity functions and of the ratios of cross PCFs. This precludes the use of multivariate Gibbs processes as considered e.g. in Rajala et al. (2018) and we consider instead a Cox process model. Specifically, the multivariate point patterns are simulated

from a multivariate log-Gaussian Cox process where for $i = 1, \ldots, p$, $X_i$ has a random intensity function of the form

$$\Lambda_i(\mathbf{u}) = \lambda_0(\mathbf{u}) \exp[\gamma_{i0} + \gamma_{i1}\mathbf{z}(\mathbf{u})] \exp\left[\alpha_i Y(\mathbf{u}) + \sigma_i U_i(\mathbf{u}) - \alpha_i^2/2 - \sigma_i^2/2\right], \tag{A.23}$$

where $\lambda_0(\cdot)$ is the inhomogeneous background intensity, $\mathbf{z}(\cdot)$ is a spatial covariate, and $Y(\cdot)$ and $U_i(\cdot)$ are independent zero-mean unit variance Gaussian random fields. The spatial correlation functions of $Y(\cdot)$ and $U_i(\cdot)$ are assumed to be exponential $c_Y(\mathbf{u}, \mathbf{v}) = \exp(-||\mathbf{u} - \mathbf{v}||/\xi)$ and $c_{U_i}(\mathbf{u}, \mathbf{v}) = \exp(-||\mathbf{u} - \mathbf{v}||/\varphi_i)$ with scale parameters $\xi$ and $\varphi_i$. Conditional on the $\Lambda_i$, the $X_i$ are independent Poisson processes. This model has a natural interpretation and can generate both positive and negative correlations between different types of points.

The process $Y(\cdot)$ can be viewed as an unobserved factor that affects all types of points and hence induces spatial correlations both within and between different types of points. The latent Gaussian process $U_i(\cdot)$ is a type-specific factor that only affects the $i$'th type of points. Conditional on $\lambda_0(\cdot)$ and $\mathbf{z}(\cdot)$, $\mathrm{E}[\Lambda_i(\mathbf{u})] = \lambda_0(\mathbf{u}) \exp[\gamma_{i0} + \gamma_{i1}\mathbf{z}(\mathbf{u})]$ and the cross PCF between $X_i$ and $X_j$ is of the form

$$g_{ij}(r; \boldsymbol{\theta}) = \exp\left[\alpha_i \alpha_j \exp(-r/\xi) + 1[i = j]\sigma_i^2 \exp(-r/\varphi_i)\right], \tag{A.24}$$

where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_p, \xi, \sigma_1^2, \ldots, \sigma_p^2, \varphi_1, \ldots, \varphi_p) \in \mathbb{R}^{3p+1}$. For $i \neq j$, $\alpha_i \alpha_j > 0$ ($< 0$) implies positive (negative) correlation between points from $X_i$ and $X_j$ whereas $\alpha_i \alpha_j = 0$ implies that $X_i$ and $X_j$ are independent given $\lambda_0(\cdot)$ and $\mathbf{z}(\cdot)$.

## 5.1 Simulation settings

More specifically, we consider the multivariate log-Gaussian Cox process with $p = 4$ and observed within a sequence of increasing square windows $W_l = [0, l] \times [0, l]$, $1 \leq l \leq 2$. The baseline intensity function in (A.23) is $\lambda_0(\mathbf{u}) = \exp\left[0.5V(\mathbf{u}) - 0.5^2/2\right]$, where $V(\mathbf{u})$ is a realization of zero-mean unit variance Gaussian random field with the exponential correlation function and a scale parameter 0.05. The spatial covariate $\mathbf{z}(\mathbf{u})$ is chosen as an independent copy of $V(\mathbf{u})$, see Figure A.1(a)-(b).

The parameters for the multivariate log-Gaussian Cox process are listed in Table A.1, where the intercept parameters $\gamma_{i0}^*$, $i = 1, \ldots, p$, are chosen so that there are on average $N_i$ points in the point pattern $X_i$ in $W_1$ with the $N_i$'s specified in Table A.1. We use $X_p$ as the baseline point process and consider three parameters of interest: the intercepts $\beta_{0i}^* = \gamma_{0i}^* - \gamma_{0p}^*$, the slopes $\beta_{1i}^* = \gamma_{1i}^* - \gamma_{1p}^*$, and the log-odds $\theta_i^*(\mathbf{u}) = \log \frac{\mathrm{p}_i(\mathbf{u}; \boldsymbol{\beta}^*)}{\mathrm{p}_p(\mathbf{u}; \boldsymbol{\beta}^*)} = \beta_{0i}^* + \beta_{1i}^* \mathbf{z}(\mathbf{u})$,

**Table A.1:** The true parameters for the multivariate LGCP.

| $X$ | $\alpha_i$ | $\sigma_i^2$ | $\xi$ | $\varphi_i$ | $\gamma_{i0}^*$ | $\gamma_{i1}^*$ | $N_i$ | $X$ | $\alpha_i$ | $\sigma_i^2$ | $\xi$ | $\varphi_i$ | $\gamma_{i0}^*$ | $\gamma_{i1}^*$ | $N_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0.5 | 0.5 | 0.1 | 0.05 | 5.17 | 0 | 150 | $X_2$ | -0.4 | 0.5 | 0.1 | 0.05 | 5.44 | 0.3 | 200 |
| $X_3$ | 0.6 | 0.5 | 0.1 | 0.05 | 5.88 | -0.6 | 300 | $X_4$ | -0.3 | 0.5 | 0.1 | 0.05 | 6.13 | 0.6 | 400 |

for $i = 1, \ldots, p-1$. The log-odds $\theta_i^*(\mathbf{u})$ represent the elevated (or reduced) likelihood of a point in $X_i$ at location $\mathbf{u}$ with an observed covariate $\mathbf{z}(\mathbf{u})$ relative to the probability of a point in $X_p$ at $\mathbf{u}$. For the log odds we consider $\mathbf{z}(\mathbf{u}) = 0.5$. The $\alpha_i$'s are chosen such that there are positive and negative spatial correlations among the $X_i$'s. The resulting PCFs and cross PCFs show (Figure A.1(c)) strong between- and within- spatial dependence.

In the following Section 5.2 we evaluate estimation accuracies for the parameters of interest and the coverage probabilities of their associated confidence intervals. The performances of the non-parametric cross PCF estimators proposed in Sections 3.4-3.5 are further considered in Section E of the supplementary material.

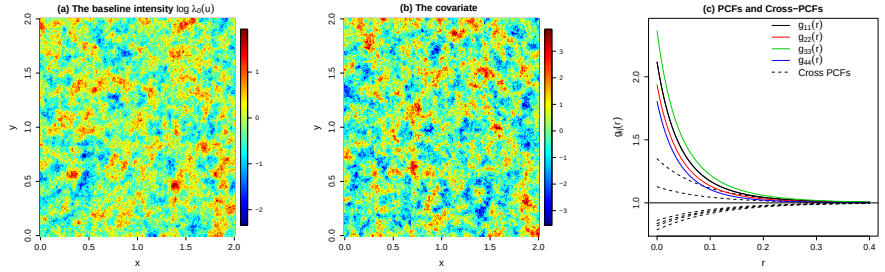


**Fig. A.1:** The log-background intensity (left panel); The spatial covariate (middle panel); The true PCFs and cross PCFs (right panel).

## 5.2 Estimation accuracies and coverage probabilities

The log odds $\theta_i^*(\mathbf{u})$ are estimated by replacing the $\boldsymbol{\beta}_i$'s in the definition of the $\theta_i^*(\mathbf{u})$'s by their estimates $\widehat{\boldsymbol{\beta}}_i$. Four types of confidence intervals are investigated, denoted $\text{CI}_{\hat{g}^{\text{n}}}$, $\text{CI}_{\hat{g}^{\text{r}}}$, $\text{CI}_{g^{\text{Poisson}}}$, and $\text{CI}_{g^{\text{true}}}$. All confidence intervals are constructed using (A.20) with the sensitivity and the covariance matrices estimated using equations (A.13) and (A.14) with $R = 0.4$ but with different choices of cross PCF estimators. The $\text{CI}_{\hat{g}^{\text{n}}}$ and $\text{CI}_{\hat{g}^{\text{r}}}$ use respectively the "naive" and "regularized" kernel cross PCF ratio estimators (A.17) and (A.19). The $R^*$ used for the "regularized" kernel estimators is obtained with the data-driven procedure in Remark 1. The $\text{CI}_{g^{\text{Poisson}}}$ is obtained by assuming $g_{ij}(\cdot) \equiv 1$ for $i, j = 1, \ldots, p$, and $\text{CI}_{g^{\text{true}}}$ is constructed using the true $g_{ij}(\cdot)$'s. The coverage probabilities of $\text{CI}_{g^{\text{true}}}$ serve as bench marks while

**Table A.2:** Estimation accuracies and coverage probabilities of confidence intervals.

| | | | | $\mathrm{CI}_{\hat{g}^{\mathrm{n}}}$ | | $\mathrm{CI}_{\hat{g}^{\mathrm{r}}}$ | | $\mathrm{CI}_{g^{\mathrm{Poisson}}}$ | | $\mathrm{CI}_{g^{\mathrm{true}}}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | SE | 90% | 95% | 90% | 95% | 90% | 95% | 90% | 95% |
| $W_1$ | $\widehat{\beta}_{01}$ | -0.002 | 0.246 | 66.1 | 71.8 | 87.0 | 92.6 | 47.6 | 54.6 | 89.3 | 93.8 |
| | $\widehat{\beta}_{02}$ | 0.002 | 0.155 | 66.5 | 72.2 | 93.6 | 97.1 | 62.6 | 70.5 | 90.5 | 94.9 |
| | $\widehat{\beta}_{03}$ | 0.002 | 0.254 | 67.0 | 74.4 | 84.7 | 90.8 | 39.3 | 45.5 | 89.7 | 94.4 |
| | $\widehat{\beta}_{11}$ | -0.001 | 0.135 | 88.6 | 94.4 | 88.2 | 94.4 | 68.5 | 77.6 | 90.4 | 95.9 |
| | $\widehat{\beta}_{12}$ | 0.002 | 0.105 | 89.5 | 94.4 | 89.1 | 94.6 | 75.7 | 83.3 | 90.5 | 95.4 |
| | $\widehat{\beta}_{13}$ | -0.001 | 0.127 | 87.4 | 93.5 | 86.9 | 92.6 | 63.9 | 73.3 | 89.6 | 94.5 |
| | $\hat{\theta}_1$ | -0.003 | 0.246 | 68.4 | 75.4 | 87.0 | 93.0 | 44.3 | 52.2 | 89.8 | 94.5 |
| | $\hat{\theta}_2$ | -0.008 | 0.157 | 70.5 | 77.6 | 92.2 | 96.0 | 61.4 | 70.6 | 89.7 | 95.3 |
| | $\hat{\theta}_3$ | -0.002 | 0.261 | 72.6 | 80.7 | 86.2 | 91.3 | 44.1 | 51.3 | 90.8 | 94.5 |
| $W_2$ | $\widehat{\beta}_{01}$ | -0.001 | 0.131 | 82.1 | 89.2 | 86.7 | 92.3 | 46.1 | 52.7 | 88.0 | 93.8 |
| | $\widehat{\beta}_{02}$ | -0.006 | 0.080 | 83.3 | 90.4 | 92.3 | 95.7 | 62.0 | 68.5 | 89.6 | 94.7 |
| | $\widehat{\beta}_{03}$ | 0.005 | 0.137 | 81.5 | 88.3 | 86.2 | 92.3 | 34.8 | 42.6 | 87.9 | 94.0 |
| | $\widehat{\beta}_{11}$ | -0.002 | 0.067 | 91.2 | 96.0 | 91.6 | 96.0 | 71.1 | 80.4 | 91.6 | 96.4 |
| | $\widehat{\beta}_{12}$ | -0.001 | 0.054 | 89.7 | 95.5 | 89.7 | 95.5 | 78.1 | 85.8 | 90.5 | 95.6 |
| | $\widehat{\beta}_{13}$ | -0.001 | 0.067 | 88.7 | 95.4 | 88.8 | 95.4 | 63.6 | 72.6 | 89.2 | 95.4 |
| | $\hat{\theta}_1$ | -0.002 | 0.130 | 83.9 | 88.7 | 88.0 | 92.4 | 45.4 | 52.4 | 88.8 | 94.2 |
| | $\hat{\theta}_2$ | -0.006 | 0.083 | 84.0 | 89.7 | 91.2 | 96.0 | 59.4 | 69.0 | 89.2 | 95.2 |
| | $\hat{\theta}_3$ | 0.005 | 0.143 | 83.7 | 88.2 | 86.0 | 91.8 | 40.2 | 47.9 | 88.7 | 93.9 |

$\mathrm{CI}_{g^{\mathrm{Poisson}}}$ may reveal potential problems of using multivariate Poisson point process models in presence of spatial correlations. Summary statistics based on 1000 simulations are given in Table A.2 and also illustrated in Figure A.2. The "Bias" columns in Table A.2 show that the parameter estimates are close to unbiased. Further, as predicted by Theorem 2, the standard errors are approximate halved when the observation window is increased from $W_1$ to the four times larger $W_2$. The coverage probabilities of $\mathrm{CI}_{g^{\mathrm{true}}}$ are all close to the nominal levels, suggesting that statistical inferences based on Theorem 2 are valid provided all cross PCF functions are correctly specified. On the contrary, in almost all cases, $\mathrm{CI}_{g^{\mathrm{Poisson}}}$ suffers from severe undercoverage that may lead to wrong conclusions in practical applications. Confidence intervals based on the "naive" kernel estimator of cross PCF ratios, i.e. $\mathrm{CI}_{\hat{g}^{\mathrm{n}}}$, achieve nominal levels for all slope parameters but suffer from serious undercoverage for intercepts and the log-odds when the observation window is small ($W_1 = [0,1] \times [0,1]$). The undercoverage of $\mathrm{CI}_{\hat{g}^{\mathrm{n}}}$ becomes much less severe when the window expands to $W_2 = [0,2] \times [0,2]$. Finally, confidence intervals based on the "regularized" cross PCF ratio estimators, i.e. $\mathrm{CI}_{\hat{g}^{\mathrm{r}}}$, can effectively correct the undercoverage of $\mathrm{CI}_{\hat{g}^{\mathrm{n}}}$ and achieve nominal levels for all parameters of interest. This suggests that it is important to apply the modification proposed in Section 3.5 for practical applications with only limited sample sizes.

Figure A.2 paints a more complete picture of how estimation accuracies

and coverage probabilities change as $W_l$ expands. The root mean squared error (RMSE) of all estimators decrease as the window size increases, supporting our theoretical findings in Section 4.1. Figure A.2 also reveals that while the coverage probabilities of $\text{CI}_{\hat{g}^{\text{n}}}$ for intercepts and log-odds are getting closer to the nominal level as $W_l$ expands, the undercoverage of $\text{CI}_{g^{\text{Poisson}}}$ does not improve at all. This emphasizes the importance of taking into account spatial correlations to make valid statistical inferences. Lastly, the coverage probabilities of $\text{CI}_{\hat{g}^{\text{r}}}$ are close to the nominal level for all parameters and window sizes and only slightly worse than those of $\text{CI}_{g^{\text{true}}}$. Therefore, we recommend $\text{CI}_{\hat{g}^{\text{r}}}$ for practical use.

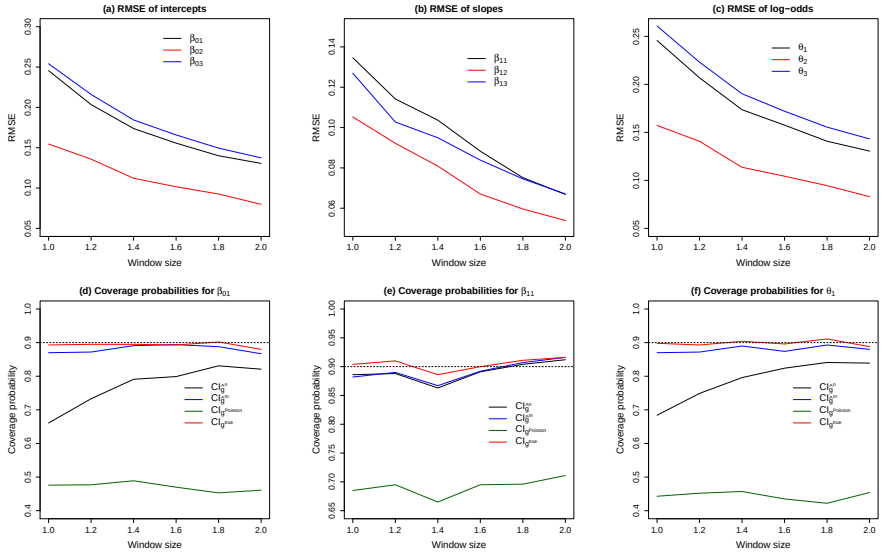


**Fig. A.2:** Top panels: the root mean squared errors (RMSE) of multinomial composite likelihood estimators; Bottom panels: coverage probabilities of various confidence intervals. Observation windows range from $W_1$ to $W_2$.

# 6 Washington DC street crime data

Figure A.3 shows spatial locations of nine types of street crimes committed in Washington DC in January and February 2017. The data set is publicly available from the website [1]. Nine types of street crime are included: (1) Other theft, (2) Robbery, (3) Theft from automobile, (4) Motor vehicle theft, (5) Assault with weapon, (6) Sex abuse, (7) Arson, (8) Burglary and (9) Homicide. The numbers of each crime type are $n_1 = 2254$, $n_2 = 366$, $n_3 = 1832$,

[1] `http://opendata.dc.gov/datasets/crime-incidents-in-2017`

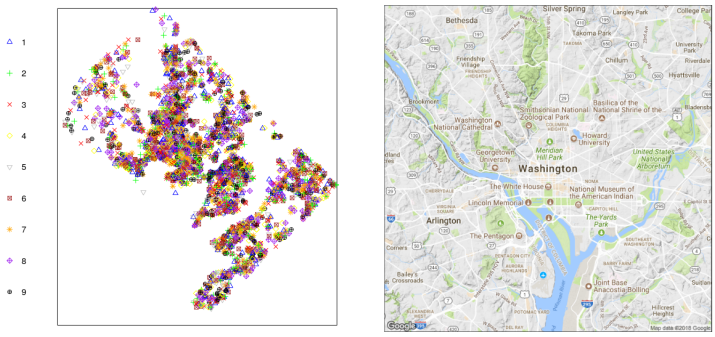


**Fig. A.3:** Left: street crimes locations ($n = 5378$); Right: a map of Washington DC.

**Table A.3:** List of spatial covariates.

| Name | Definition |
|---|---|
| 1. % African | Square root of percentage of African American residents |
| 2. % Hispanic | Square root of percentage of Hispanic residents |
| 3. % Male | Square root of percentage of male residents with age 18-24 |
| 4. % HouseRent | Percentage of housing units occupied by renters |
| 5. % Bachelor | Percentage of residents over age 25 with a bachelor's degree |
| 6. MedIncome | Logarithm of median annual per capita income (in $1000) |
| 7. Pdist | Logarithm of the distance to the nearest police station |

$n_4 = 335$, $n_5 = 332$, $n_6 = 44$, $n_7 = 1$, $n_8 = 259$ and $n_9 = 14$. We omit the rare street crimes "Sex abuse", "Arson" and "Homicide". Using spatial covariates similar to those suggested in Reinhart and Greenhouse (2018), the first 6 spatial covariates listed in Table A.3 are obtained from US census data and are constant within each of 179 census tracts partitioning Washington DC, see also Section 6.3. We calculated ourselves the last covariate (distance to nearest police station) which varies smoothly across the city. Square root and log transformations have been applied to some covariates to achieve approximate normal distributions.

## 6.1 Inference regarding regression coefficients and cross PCFs

Using model (A.3), we assume that the intensity of each street crime is given by

$$\lambda_i(\mathbf{u};\boldsymbol{\gamma}_i) = \lambda_0(\mathbf{u})\exp\left[\gamma_{i0} + \gamma_{i1}z_1(\mathbf{u}) + \cdots + \gamma_{i7}z_7(\mathbf{u})\right], i = 1,\ldots,5,8.$$

where the $z_k(\cdot)$'s are listed in Table A.3. The common first street crime "Other theft" is used as the baseline. The regression parameters are estimated by maximizing the composite likelihood (A.6). The asymptotic standard errors and $p$-values are computed with $R = 3$ km and either of two types of cross PCFs: using the "regularized" kernel estimator $\hat{g}^{\mathrm{r}}$ proposed in

Section 3.5 with $b = 0.2$km, or assuming all $g_{ij}(\cdot) \equiv 1$ ( "Poisson") for any $i, j = 1, \ldots, 5, 8$. The $R^*$ used for the "regularized" kernel estimators is obtained through the data-driven procedure outlined in Remark 1. Estimated regression coefficients, standard deviations, and $p$-values are summarized in Table A.4, and estimated PCF ratios and cross PCF ratios are illustrated in Figure A.4.

Figure A.4(a) indicates that within and between clustering for crimes types other than "Other theft" is less strong than for "Other theft" up to around 250 meters. After that some crime types appear to be more clustered than "Other theft" but the difference in clustering strength vanishes around 3km distance. In particular, Figure A.4 suggests that a multivariate Poisson model is not appropriate for street crime data.

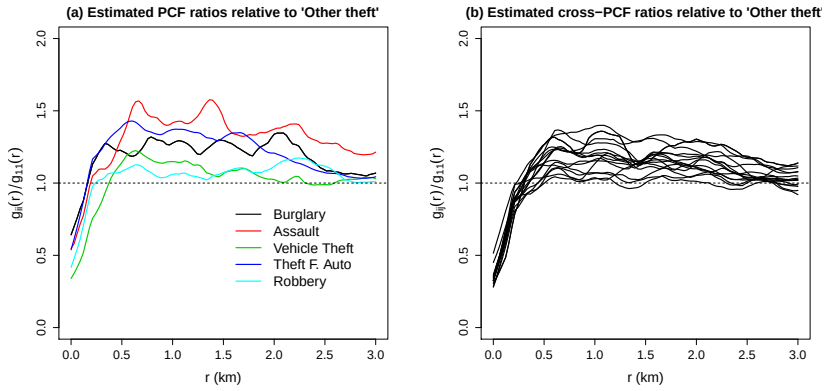


**Fig. A.4:** (a) Estimated PCF ratios $g_{ii}(r)/g_{11}(r)$ for $i = 2, \ldots, 5, 8$; (b) estimated cross PCF ratios $g_{ij}(r)/g_{11}(r)$ for $i, j = 2, \ldots, 5, 8$ and $i \neq j$.

In Table A.4, the Poisson model as expected always gives smaller standard errors for all coefficients. As a result, more regression coefficients appear to be statistically significant at the $\alpha = 0.05$ level (highlighted in blue) compared to those for the proposed method where cross PCFs are estimated from the data. In some cases, the two methods reach contradictory conclusions. For example, the covariate "% HouseRent" is significant under the Poisson model ($p$-value 0.028) when comparing "Theft from auto" to the baseline process "Other theft", while the proposed model asserts otherwise with a $p$-value of 0.352. In such cases, considering the strong spatial correlations displayed in Figure A.4, we argue that the proposed method is more reliable.

Based on the proposed method, all estimated coefficients for "% HouseRent" are negative and many of them are significant, suggesting that when "% HouseRent" is large, "Other theft" becomes relatively more frequent compared to all other crime types. Second, no covariate elevates or reduces the relative risk of "Robbery" compared to "Other theft" and no covariate other

than "% HouseRent" is significant for the relative risk between "Motor vehicle theft" and "Other theft". Third, "Theft from automobile" tend to occur more often in a neighborhood with more African American/Hispanic population, less young male percentage and residents with relatively low education level, as compared to "Other theft". Fourth, "Assault with weapon" is more likely to occur in a neighborhood with low young male population and low income levels compared to "Other theft". Finally, compared to "Other theft", "Burglary" tends to occur more in areas with low African American population, low education level and larger distance to the police station.

Returning to the discussion regarding crime general and crime specific theories in Section 3.1, our results clearly show that the relative risks of different crime types depend significantly on subsets of the covariates considered. This also means that the conditional probabilities (A.5) depend significantly on the covariates which results in a clear spatial segregation regarding the relative risks of different crimes, see Figure A.5 below. These results support the crime specific theory.

## 6.2 Conditional probability maps and intensity estimation

For any location $\mathbf{u}$, using the fitted $\widehat{\boldsymbol{\beta}}$, we can compute $\mathrm{p}_i(\mathbf{u}, \widehat{\boldsymbol{\beta}})$ for $i = 1, \ldots, p$, using (A.5). This enables us to create the conditional probability maps in Figure A.5 which show $\mathrm{p}_i(\mathbf{u}, \widehat{\boldsymbol{\beta}})$, $i = 1, \ldots, 5, 8$ computed at the 5378 observed crime locations. Recall that given a street crime occurs at location $\mathbf{u}$, $\mathrm{p}_i(\mathbf{u}, \widehat{\boldsymbol{\beta}})$ is the fitted probability that the crime is of the $i$'th type. The strong spatial patterns in these conditional probabilities are remarkable. For instance, in the southeast part of the city (southeast to the Anacostia River), given a crime occurs, it is much more likely to be of type "Robbery" or "Assault" than in other parts of the city. In contrast, "Theft from automobile" is more likely to be reported in the middle and northern parts of the city while the hot spot for "Other theft" is located in the middle-west part of the city.

Figure A.6 shows semi-parametric kernel estimates of the six crime intensities using (A.4) where $\lambda_0$ is estimated using the kernel estimate

$$\hat{\lambda}_0(\mathbf{u}) = \frac{1}{p} \sum_{i=1}^{p} \sum_{\mathbf{v} \in X_i} \exp[-\widehat{\boldsymbol{\beta}}_i^{\mathsf{T}} \mathbf{z}(\mathbf{v})] k[(\mathbf{u} - \mathbf{v})/b]/b^2 \tag{A.25}$$

where $k$ is a two-dimensional kernel and the bandwidth $b = 3.37$km is chosen according to the data-driven criterion of Cronie and Van Lieshout (2018). Compared to the conditional probability plots given in Figure A.5 that demonstrate relative compositions of different types of crimes at a given location, the marginal intensities provide additional information on how often each type of crime occurs in the same location. Both plots can be useful

**Table A.4:** Estimated coefficients, standard errors, and $p$-values for street crime data.

| Street crime | Covariate | Coef. | Std. err. $\hat{g}^{\mathrm{r}}$ | Std. err. Poisson | P-values $\hat{g}^{\mathrm{r}}$ | P-values Poisson |
|---|---|---|---|---|---|---|
| Robbery ($n_2 = 366$) | % African | 0.894 | 0.867 | 0.697 | 0.302 | 0.199 |
| | % Hispanic | 0.669 | 0.685 | 0.499 | 0.329 | 0.180 |
| | % Male | 0.141 | 1.183 | 0.962 | 0.905 | 0.884 |
| | % HouseRent | -0.783 | 0.442 | 0.352 | 0.077 | **0.026** |
| | % Bachelor | -1.130 | 0.970 | 0.760 | 0.244 | 0.137 |
| | MedIncome | -0.071 | 0.371 | 0.304 | 0.847 | 0.814 |
| | Pdist | 0.176 | 0.108 | 0.086 | 0.102 | **0.040** |
| Theft from automobile ($n_3 = 1832$) | % African | 2.318 | 0.813 | 0.346 | **0.004** | **<0.0001** |
| | % Hispanic | 2.369 | 0.760 | 0.286 | **0.002** | **<0.0001** |
| | % Male | -2.332 | 1.049 | 0.500 | **0.026** | **<0.0001** |
| | % HouseRent | -0.412 | 0.444 | 0.188 | 0.352 | **0.028** |
| | % Bachelor | 2.936 | 0.891 | 0.417 | **0.001** | **<0.0001** |
| | MedIncome | -0.461 | 0.339 | 0.164 | 0.174 | **0.004** |
| | Pdist | 0.071 | 0.107 | 0.047 | 0.508 | 0.131 |
| Motor vehicle theft ($n_4 = 335$) | % African | -0.451 | 0.872 | 0.702 | 0.605 | 0.520 |
| | % Hispanic | -0.556 | 0.724 | 0.533 | 0.443 | 0.297 |
| | % Male | -0.139 | 1.174 | 0.962 | 0.906 | 0.885 |
| | % HouseRent | -1.295 | 0.443 | 0.355 | **0.003** | **0.0003** |
| | % Bachelor | -1.767 | 0.993 | 0.785 | 0.075 | **0.024** |
| | MedIncome | -0.174 | 0.361 | 0.300 | 0.630 | 0.563 |
| | Pdist | 0.205 | 0.113 | 0.089 | 0.070 | **0.022** |
| Assult with weapon ($n_5 = 332$) | % African | 1.346 | 1.004 | 0.806 | 0.180 | 0.095 |
| | % Hispanic | -0.101 | 0.794 | 0.541 | 0.898 | 0.851 |
| | % Male | -2.76 | 1.358 | 1.132 | **0.042** | **0.0145** |
| | % HouseRent | -1.229 | 0.494 | 0.377 | **0.013** | **0.001** |
| | % Bachelor | -0.619 | 1.124 | 0.839 | 0.582 | 0.461 |
| | MedIncome | -0.798 | 0.391 | 0.314 | **0.041** | **0.011** |
| | Pdist | 0.145 | 0.122 | 0.088 | 0.235 | 0.100 |
| Burglary ($n_8 = 259$) | % African | -2.332 | 1.187 | 0.801 | **0.050** | **0.003** |
| | % Hispanic | -0.029 | 0.983 | 0.583 | 0.977 | 0.961 |
| | % Male | 0.776 | 1.555 | 1.039 | 0.618 | 0.455 |
| | % HouseRent | -1.930 | 0.670 | 0.376 | **0.001** | **<0.0001** |
| | % Bachelor | -3.374 | 1.327 | 0.875 | **0.011** | **0.001** |
| | MedIncome | -0.352 | 0.432 | 0.300 | 0.415 | 0.240 |
| | Pdist | 0.359 | 0.168 | 0.105 | **0.033** | **0.0006** |

in practice for the police department to better allocate limited resources to effective fight different types of crimes.

## 6.3 Residual analysis

In this subsection, we perform a residual analysis for the fitted model. We divide the data according to the 179 census tracts in Washington DC, denoted as $A_1, A_2, \ldots, A_K$, $K = 179$, we define the raw residual for the $i$'th type of

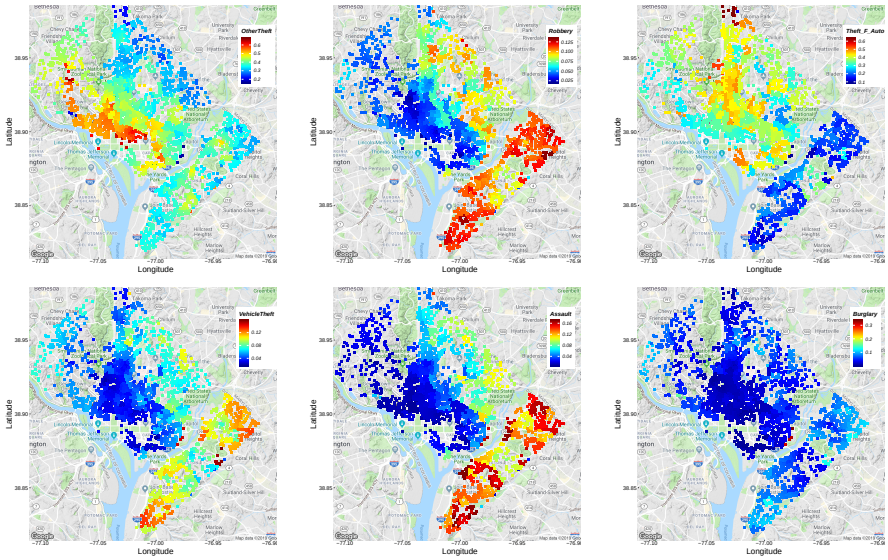


**Fig. A.5:** Estimated conditional probability maps for Washington DC.

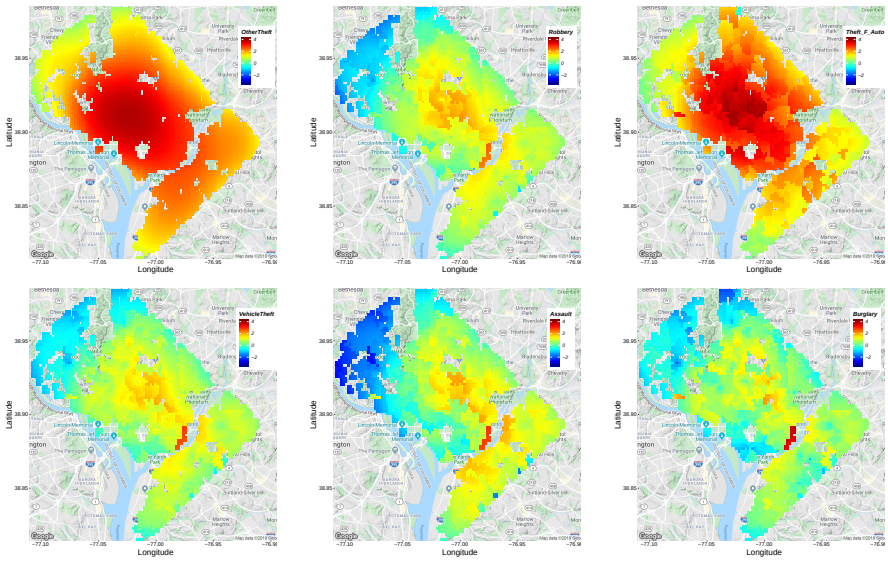


**Fig. A.6:** Semi-parametric log-intensity (per km$^2$) maps for crime data in Washington DC.

street crime in $A_k$ as

$$\hat{\varepsilon}_{i,k}(\widehat{\boldsymbol{\beta}}) = \sum_{\mathbf{u} \in X_i} I(\mathbf{u} \in A_k) - \sum_{\mathbf{u} \in X^{\mathrm{pl}} \cap A_k} \mathrm{p}_i(\mathbf{u}; \widehat{\boldsymbol{\beta}}), \tag{A.26}$$

for $i = 1, \ldots, p$ and $k = 1, \ldots, K$. Equation (A.26) is essentially a restricted version (within $A_k$) of the intercept component of $\mathbf{e}_i(\widehat{\boldsymbol{\beta}})$ defined in (A.9). By definition of $\widehat{\boldsymbol{\beta}}$, $\mathbf{e}_i(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, implying $\sum_{k=1}^{K} \hat{\varepsilon}_{i,k} = 0$ for $i = 1, \ldots, p$. If the model fits the data reasonably well, one should expect most $\hat{\varepsilon}_{i,k}$ to be relatively close to 0.

Use the same arguments leading to (A.14), the variance of $\hat{\varepsilon}_{i,k}(\boldsymbol{\beta}^*)$ can be estimated by

$$\hat{\sigma}^2_{i,k}(\boldsymbol{\beta}^*, g) = \sum_{\mathbf{u} \in X^{\mathrm{pl}} \cap A_k} [1 - \mathrm{p}^*_i(\mathbf{u})]\, \mathrm{p}^*_i(\mathbf{u}) + \sum_{\mathbf{u}, \mathbf{v} \in X^{\mathrm{pl}} \cap A_k}^{\mathbf{u} \neq \mathbf{v}} \mathrm{p}^*_i(\mathbf{u}) \mathrm{p}^*_i(\mathbf{v}) T_{ii}(\mathbf{u}, \mathbf{v}; \boldsymbol{\beta}^*, g),$$

where $T_{ii}(\mathbf{u}, \mathbf{v}; \boldsymbol{\beta}^*, g)$ is defined in (A.12). Consequently, by replacing $\boldsymbol{\beta}$ and cross PCFs by their estimates, the standardized residual can be defined as $\hat{\varepsilon}_{i,k}(\widehat{\boldsymbol{\beta}}) = \hat{\varepsilon}_{i,k}(\widehat{\boldsymbol{\beta}}) / \hat{\sigma}_{i,k}(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{r}})$, for $i = 1, \ldots, p$ and $k = 1, \ldots, K$.

Standardized residuals for all census tracts in Washington DC are illustrated in Figure A.7. One census tract that does not have any reported street crime activities in January and February 2017 is indicated by the black color. Most standardized residuals are inside the range of $[-3, 3]$ for all six types of street crimes, indicating an adequate model fit. Finally, the apparent strong spatial correlations among the residuals further support the use of the proposed method.

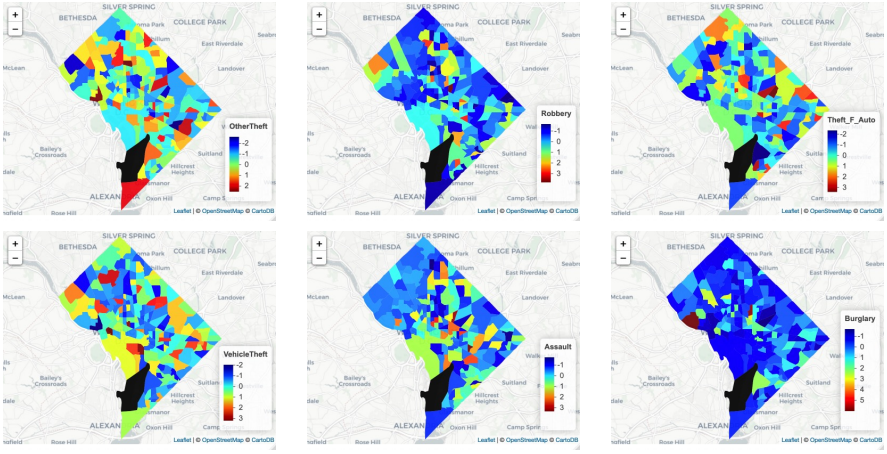


**Fig. A.7:** Standardized residuals for 179 census tracts for six types of street crimes.

## 6.4 Goodness-of-fit assessment

In addition to the graphical residual analysis in Section 6.3, it is useful to have a numerical summary of the overall goodness-of-fit of the fitted model.

In this section, we propose a Monte Carlo test procedure inspired by the goodness-of-fit tests proposed in Dong and Yu (2019a,b). To do so, we view the covariate vector $\mathbf{z}(\mathbf{u})$ as a realization of a random vector $\mathbf{Z}(\mathbf{u}) = (Z_1(\mathbf{u}), \cdots, Z_q(\mathbf{u}))^\mathsf{T}$ and consider our data as a collection of marked points $(\mathbf{u}, \mathbf{Z}(\mathbf{u}), Y(\mathbf{u}))$ where $\mathbf{u} \in X^{\mathrm{pl}}$ denotes a crime scene and $Y(\mathbf{u}) \in \{1, \ldots, p\}$ is the type of crime committed at $\mathbf{u}$. We can define an empirical conditional distribution function as

$$\widehat{F}_{Y|Z_j}(y|z) = \frac{1}{N_j(z)} \sum_{\mathbf{u} \in X^{\mathrm{pl}}} I\left[Y(\mathbf{u}) \leq y, z_j(\mathbf{u}) \leq z\right], \quad y = 1, \cdots, p, z \in \mathbb{R}, \tag{A.27}$$

where $N_j(z) = \sum_{\mathbf{u} \in X^{\mathrm{pl}}} I\left[z_j(\mathbf{u}) \leq z\right]$, $j = 1, \cdots, q$. This is an estimate of

$$F_{Y|Z_j}(y|z) = \mathrm{E} \frac{1}{M_j(z)} \sum_{\mathbf{u} \in X^{\mathrm{pl}}} I\left[Y(\mathbf{u}) \leq y, Z_j(\mathbf{u}) \leq z\right] \tag{A.28}$$

where $M_j(z) = \mathrm{E}N_j(z)$. Under our model (A.3), one can show that an alternative estimator of (A.28) is given by

$$\widehat{F}^*_{Y|Z_j}(y|z) = \frac{1}{N_j(z)} \sum_{\mathbf{u} \in X^{\mathrm{pl}}} \left[\sum_{i=1}^{y} \mathrm{p}_i(\mathbf{u}; \widehat{\boldsymbol{\beta}})\right] I\left[z_j(\mathbf{u}) \leq z\right], \quad y = 1, \cdots, p, z \in \mathbb{R}, \tag{A.29}$$

where $\mathrm{p}_i(\mathbf{u}; \widehat{\boldsymbol{\beta}})$ is defined in (A.5) with $\widehat{\boldsymbol{\beta}}$ obtained from (A.6).

Following Dong and Yu (2019a), if model (A.3) is appropriate, one would expect $\widehat{F}_{Y|Z_j}(y|z)$ and $\widehat{F}^*_{Y|Z_j}(y|z)$ to be close for any $z$ and $j = 1, \cdots, q$. Therefore, we can define for each covariate a test statistic as

$$T_j(z) = \sum_{i=1}^{p} \left|\widehat{F}_{Y|Z_j}(y|z) - \widehat{F}^*_{Y|Z_j}(y|z)\right| \Delta_{j,z}(i), \quad j = 1, \cdots, q, \tag{A.30}$$

where $\Delta_{j,z}(i) = \widehat{F}_{Y|Z_j}(i|z) - \widehat{F}_{Y|Z_j}(i-1|z)$, $i = 1, \cdots, p$.

It remains to evaluate the distribution of the $T_j(z)$'s. Dong and Yu (2019a) suggests using a bootstrap exploiting that their pairs of covariate vectors and response variables are independent and identically distributed. This is not possible in our situation where the $(\mathbf{Z}(\mathbf{u}), Y(\mathbf{u}))$'s are not independent. In the following, we pursue some model-based bootstrap alternatives where we replace the unknown background intensity $\lambda_0(\cdot)$ by its non-parametric estimate (A.25) and try out some simple models for the correlation structure.

The simplest choice is the multivariate Poisson model, where we assume the $X_i$'s are independent inhomogeneous Poisson processes with intensity functions $\lambda_i(\cdot)$'s. Based on $B$ simulations from this model, one can compute $T^{\mathrm{Poisson}}_{j,1}(z), \cdots, T^{\mathrm{Poisson}}_{j,B}(z)$ from which point-wise 95% percentiles can be estimated. Figure A.8 shows the observed test statistic $T_j(z)$ as a function of $z$

for three covariates and the corresponding 95% percentiles based on simulations from the Poisson model (plots for the remaining covariates are similar and shown in the supplementary material). The absence of between or within spatial correlation for the Poisson model means that the simulated parameter estimates based on (A.6) vary too little compared to their variation under the true data generating mechanism where spatial correlation is present as suggested in Figures A.4 and A.7. It is therefore not surprising that some observed $T_j(z)$'s are above the 95% percentile based on Poisson simulations.

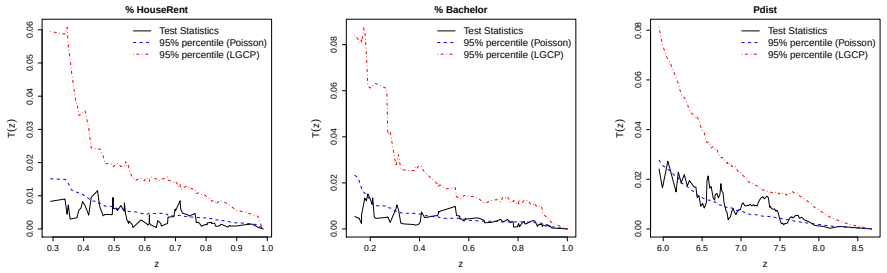


**Fig. A.8:** Estimated $T_j(z)$, $j = 4, 5, 7$ in (A.30) for the covariates % HouseRent, % Bachelor, and Pdist, together with 95% percentiles computed from the Poisson model and the LGCP model (plots for the other covariates are given in the supplementary material).

To partially account for spatial correlation we next consider a second special case of model (A.3) where all point processes are independent log-Gaussian Cox processes (LGCPs), each with an exponential covariance function. Plugging in the kernel estimate of $\lambda_0(\cdot)$, we then estimate the correlation parameters for each LGCP separately using standard minimum contrast methods. Figure A.8 shows that all observed $T_j(z)$'s are well below the 95% percentiles based on simulations of the fitted LGCPs. Thus large values of the observed $T_j(z)$'s can be explained by sampling variation even when we only take into account correlation within each type of points and not between.

Plugging in the non-parametric estimate of $\lambda_0(\cdot)$ is not optimal but seems to be the only alternative at the moment to fit parametric models for the correlation structure. Developing a parametric model for the full correlation structure is beyond the scope of this paper.

# 7 Concluding remarks

We propose a flexible semi-parametric model for multivariate point pattern data. The non-parametric component of the model takes into account features of the multivariate intensity function that are difficult to model or specify while the parametric part facilitates a study of effects of covariates on

relative risks of occurrence of different types of points. Interesting conditional probability maps can be obtained from the parametric part and the intensity of a specific type of points can be estimated using the full data set by combining the parametric estimate of the relative risk with an estimate of the non-parametric part.

Our multinomial logistic composite likelihood estimation approach does not require knowledge of the non-parametric model component. It is moreover well founded theoretically since we established the asymptotic properties of the estimation approach in a very general setting that does not require any independence assumptions, neither within nor between the different types of points.

Our non-parametric estimation approach allows us to estimate cross PCFs up to a common multiplicative factor. This is sufficient for estimating the covariance matrix of regression parameter estimates and for inferring ratios of cross PCFs. However, to infer individual cross PCFs, it seems necessary to introduce parametric models for the cross PCFs. We plan to pursue this in future work. There is also room for improving the kernel estimate (A.25) which can be criticized for ignoring the layout of the city.

Our methodology is applicable in very diverse fields. Our example application is within criminology where the estimated conditional probability maps disclose a remarkable structure in the occurrence of various types of street crimes in Washington DC. Other obvious areas of applications are disease mapping in epidemiology and studies of spatial distributions of plant and animal species in ecology. Our approach can further be extended to space-time multivariate point pattern data, which have attracted much interest in various research areas including criminology, see e.g. the thorough review in the recent paper Reinhart and Greenhouse (2018).

## SUPPLEMENTARY MATERIAL

The supplementary material for this paper contains further simulation studies and plots, proofs, and auxiliary results.

# A Sensitivity and covariance matrices for $\mathbf{e}(\boldsymbol{\beta})$

**Theorem A.1**
The sensitivity matrix of the estimating function $\mathbf{e}(\boldsymbol{\beta})$ is a symmetric $(p-1)q \times (p-1)q$ matrix $\mathbf{S}(\boldsymbol{\beta})$, where the diagonal blocks are given by:

$$\mathbf{S}(\boldsymbol{\beta})_{ii} = \int_W \mathbf{Z}(\mathbf{u},\mathbf{u}) \left[1 - \mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\right] \mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u}$$

for $i = 1, \ldots, p-1$ and the off-diagonal blocks are given by:

$$\mathbf{S}(\boldsymbol{\beta})_{ij} = -\int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\mathrm{p}_j(\mathbf{u};\boldsymbol{\beta})\lambda^{\mathrm{pl}}(\mathbf{u})\mathrm{d}\mathbf{u}$$

for distinct $i, j = 1, \ldots, p-1$. When $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ these results simplify to

$$\mathbf{S}(\boldsymbol{\beta}^*)_{ii} = \int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\left[1 - \mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\right]\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u} \quad \text{and}$$

$$\mathbf{S}(\boldsymbol{\beta}^*)_{ij} = -\int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_j(\mathbf{u};\boldsymbol{\beta}^*)\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u}.$$

*Proof.* Some straightforward algebra yields that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_i^\mathsf{T}}\mathbf{e}_i(\boldsymbol{\beta}) \\
&= \nabla_{\boldsymbol{\beta}_i^\mathsf{T}} \sum_{\mathbf{u}\in X_i\cap W} \mathbf{z}(\mathbf{u}) - \sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W} \mathbf{z}(\mathbf{u})\nabla_{\boldsymbol{\beta}_i^\mathsf{T}}\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}) \\
&= -\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_i\cap W} \mathbf{Z}(\mathbf{u},\mathbf{u})\frac{\exp\left[\boldsymbol{\beta}_i^\mathsf{T}\mathbf{z}(\mathbf{u})\right]\left(1+\sum_{k=1}^{p-1}\exp\left[\boldsymbol{\beta}_k^\mathsf{T}\mathbf{z}(\mathbf{u})\right] - \exp\left[\boldsymbol{\beta}_i^\mathsf{T}\mathbf{z}(\mathbf{u})\right]\right)}{(1+\sum_{k=1}^{p-1}\exp\left[\boldsymbol{\beta}_k^\mathsf{T}\mathbf{z}(\mathbf{u})\right])^2} \\
&= -\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W} \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\left[1-\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\right]. \qquad \text{(A.31)}
\end{aligned}$$

The expectation of (A.31) negated is by Campbell's formula:

$$\mathbf{S}(\boldsymbol{\beta})_{ii} = \mathrm{E}\left[-\frac{\partial}{\partial\boldsymbol{\beta}_i^\mathsf{T}}\mathbf{e}_i(\boldsymbol{\beta})\right] = \int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\left[1-\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\right]\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\lambda^{\mathrm{pl}}(\mathbf{u})\mathrm{d}\mathbf{u}.$$

Similarly, for the off-diagonal blocks of the Hessian matrix of $\ell(\boldsymbol{\beta})$, we have that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_j^\mathsf{T}}\mathbf{e}_i(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}_j^\mathsf{T}} \sum_{\mathbf{u}\in X_i\cap W} \mathbf{z}(\mathbf{u}) - \sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W} \mathbf{z}(\mathbf{u})\nabla_{\boldsymbol{\beta}_j^\mathsf{T}}\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}) \\
&= \sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W} \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\mathrm{p}_j(\mathbf{u};\boldsymbol{\beta}). \qquad \text{(A.32)}
\end{aligned}$$

The expectation of (A.32) negated is

$$\mathbf{S}(\boldsymbol{\beta})_{ij} = \mathrm{E}\left[-\frac{\partial}{\partial\boldsymbol{\beta}_j^\mathsf{T}}\mathbf{e}_i(\boldsymbol{\beta})\right] = -\int_W \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\mathrm{p}_j(\mathbf{u};\boldsymbol{\beta})\lambda^{\mathrm{pl}}(\mathbf{u})\mathrm{d}\mathbf{u}.$$

Finally, when $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, we have that $\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\lambda^{\mathrm{pl}}(\mathbf{u}) = \lambda_i^*(\mathbf{u})$ due to (A.5), which completes the proof of Theorem A.1. □

**Theorem A.2**
The variance-covariance matrix of $\mathbf{e}(\boldsymbol{\beta}^*)$ is a $(p-1)q \times (p-1)q$ matrix with blocks given by (A.11) for $i, j = 1, \ldots, p-1$.

*Proof.* It is straightforward to prove that $\mathrm{E}[\mathbf{e}(\boldsymbol{\beta}^*)] = \mathbf{0}$ by showing that

$$\begin{aligned}\mathrm{E}[\mathbf{e}_i(\boldsymbol{\beta}^*)] &= \int_W \mathbf{z}(\mathbf{u})\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u} - \sum_{l=1}^{p}\int_W \mathbf{z}(\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\lambda_l^*(\mathbf{u})\mathrm{d}\mathbf{u} \\ &= \int_W z(\mathbf{u})\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u} - \int_W z(\mathbf{u})\lambda_i^*(\mathbf{u})\mathrm{d}\mathbf{u} = \mathbf{0},\end{aligned}$$

where the second last equality follows from the fact that $\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\lambda^{\mathrm{pl}}(\mathbf{u}) = \lambda_i^*(\mathbf{u})$ due to (A.5). The diagonal blocks of $\mathrm{Var}\left[\mathbf{e}(\boldsymbol{\beta}^*)\right]$ are then given by $\mathrm{Var}\left[\mathbf{e}_i(\boldsymbol{\beta}^*)\right] = \mathrm{E}\left[\mathbf{e}_i(\boldsymbol{\beta}^*)\mathbf{e}_i(\boldsymbol{\beta}^*)^{\mathsf{T}}\right]$ for $i = 1, \ldots, p-1$, where $\mathbf{e}_i(\boldsymbol{\beta}^*)\mathbf{e}_i(\boldsymbol{\beta}^*)^{\mathsf{T}}$ is

$$\begin{aligned}&\mathbf{e}_i(\boldsymbol{\beta}^*)\mathbf{e}_i(\boldsymbol{\beta}^*)^{\mathsf{T}} = \\ &\sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u})\sum_{\mathbf{v}\in X_i\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}} - \sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u})\sum_{l=1}^{p}\sum_{\mathbf{v}\in X_l\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}}\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*) - \\ &\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W}\mathbf{z}(\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\sum_{\mathbf{v}\in X_i\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}} + \\ &\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W}\mathbf{z}(\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\sum_{l=1}^{p}\sum_{\mathbf{v}\in X_l\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}}\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*),\end{aligned} \tag{A.33}$$

where the first term in (A.33) is

$$\sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u})\sum_{\mathbf{v}\in X_i\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}} = \sum_{\mathbf{u},\mathbf{v}\in X_i\cap W}^{\neq}\mathbf{Z}(\mathbf{u},\mathbf{v}) + \sum_{\mathbf{u}\in X_i\cap W}\mathbf{Z}(\mathbf{u},\mathbf{u}),$$

and the second and third terms are of the form

$$\begin{aligned}&\sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u})\sum_{l=1}^{p}\sum_{\mathbf{v}\in X_l\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}}\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*) = \\ &\sum_{\substack{l=1\\ l\neq i}}^{p}\sum_{\substack{\mathbf{u}\in X_i\cap W\\ \mathbf{v}\in X_l\cap W}}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*) + \\ &\sum_{\mathbf{u},\mathbf{v}\in X_i\cap W}^{\neq}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*) + \sum_{\mathbf{u}\in X_i\cap W}\mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\end{aligned} \tag{A.34}$$

and the fourth term is

$$\begin{aligned}
&\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W}\mathbf{z}(\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\sum_{l'=1}^{p}\sum_{\mathbf{v}\in X_{l'}\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}}\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*) = \\
&\sum_{\substack{l,l'=1\\ l\neq l'}}^{p}\sum_{\substack{\mathbf{u}\in X_l\cap W\\ \mathbf{v}\in X_{l'}\cap W}}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*)+ \\
&\sum_{l=1}^{p}\sum_{\mathbf{u},\mathbf{v}\in X_l\cap W}^{\neq}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*)+ \\
&\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap W}\mathbf{Z}(\mathbf{u},\mathbf{u})\left[\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\right]^2.
\end{aligned} \tag{A.35}$$

The variance of $\mathbf{e}_i(\boldsymbol{\beta}^*)$ is by Campbell's formula

$$\begin{aligned}
&\mathrm{Var}\left[\mathbf{e}_i(\boldsymbol{\beta}^*)\right] \\
&= \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_i^*(\mathbf{v})g_{ii}(\mathbf{u},\mathbf{v})\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v} \\
&- \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*)\lambda_i^*(\mathbf{u})\sum_{l=1}^{p}\lambda_l^*(\mathbf{v})g_{il}(\mathbf{u},\mathbf{v})\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v} \\
&- \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\lambda_i^*(\mathbf{v})\sum_{l=1}^{p}\lambda_l^*(\mathbf{u})g_{il}(\mathbf{u},\mathbf{v})\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v} \\
&+ \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\mathrm{p}_i(\mathbf{v};\boldsymbol{\beta}^*)\sum_{l,l'=1}^{p}\lambda_l^*(\mathbf{u})\lambda_{l'}^*(\mathbf{v})g_{ll'}(\mathbf{u},\mathbf{v})\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v}+\mathbf{S}_{ii}(\boldsymbol{\beta}^*)\mathbf{d}\mathbf{u} \\
&= \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_i^*(\mathbf{v})g_{ii}(\mathbf{u},\mathbf{v})\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v} \\
&- \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_i^*(\mathbf{v})\left[\sum_{l=1}^{p}[\mathrm{p}_l(\mathbf{v};\boldsymbol{\beta}^*)g_{il}(\mathbf{u},\mathbf{v})+\mathrm{p}_l(\mathbf{u};\boldsymbol{\beta}^*)g_{il}(\mathbf{u},\mathbf{v})]\right]\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v} \\
&+ \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_i^*(\mathbf{v})g^{\mathrm{pl}}(\mathbf{u},\mathbf{v})\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v}+\mathbf{S}_{ii}(\boldsymbol{\beta}^*) \\
&= \mathbf{S}_{ii}(\boldsymbol{\beta}^*)+\int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_i^*(\mathbf{v})[g^{\mathrm{pl}}(\mathbf{u},\mathbf{v})+g_{ii}(\mathbf{u},\mathbf{v})]\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v} \\
&- \int_{W^2}\mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_i^*(\mathbf{v})\left[\sum_{l=1}^{p}[\mathrm{p}_l(\mathbf{v};\boldsymbol{\beta}^*)g_{il}(\mathbf{u},\mathbf{v})+\mathrm{p}_l(\mathbf{u};\boldsymbol{\beta}^*)g_{il}(\mathbf{u},\mathbf{v})]\right]\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v}
\end{aligned}$$

The calculation of $\mathrm{Cov}[\mathbf{e}_i(\boldsymbol{\beta}^*),\mathbf{e}_j(\boldsymbol{\beta}^*)]$ is similar to (A.33) with the first term replaced by $\sum_{\mathbf{u}\in X_i\cap W}\mathbf{z}(\mathbf{u})\sum_{\mathbf{v}\in X_j\cap W}\mathbf{z}(\mathbf{v})^{\mathsf{T}} = \sum_{\substack{\mathbf{u}\in X_i\cap W\\ \mathbf{v}\in X_j\cap W}}\mathbf{z}(\mathbf{u})\mathbf{z}(\mathbf{v})^{\mathsf{T}}$. Following the same calculations as for the diagonal blocks of $\mathrm{Var}[\mathbf{e}(\boldsymbol{\beta}^*)]$, the off-

diagonal blocks are given by

$$
\begin{aligned}
&\operatorname{Cov}\left[\mathbf{e}_i(\boldsymbol{\beta}^*), \mathbf{e}_j(\boldsymbol{\beta}^*)\right] \\
&= \mathbf{S}_{ij}(\boldsymbol{\beta}^*) + \int_{W^2} \mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_j^*(\mathbf{v}) \left[g^{\mathrm{pl}}(\mathbf{u},\mathbf{v}) + g_{ij}(\mathbf{u},\mathbf{v})\right] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&- \int_{W^2} \mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_j^*(\mathbf{v}) \left[\sum_{l=1}^{p} \left[\mathrm{p}_l(\mathbf{v};\boldsymbol{\beta}^*) g_{il}(\mathbf{u},\mathbf{v}) + \mathrm{p}_l(\mathbf{u};\boldsymbol{\beta}^*) g_{jl}(\mathbf{u},\mathbf{v})\right]\right] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.
\end{aligned}
$$

The result now follows by the definition of $T_{ij}$ in (A.12). □

# B Lower bound for $T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)$

The proof of lower bounds for $T_{ii}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}}, g^{\mathrm{n}})$ and $T_{ii}(\mathbf{u},\mathbf{v};\widehat{\boldsymbol{\beta}}, g^{\mathrm{r}})$ are identical to the proof of $T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)$, therefore, we only provide proofs of $T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)$.

## B.1 Lower bounds under constraint (A.18)

$$
\begin{aligned}
&T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g) \\
&= 1 + \frac{g_{ii}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} - \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{v}) g_{il}(\mathbf{u},\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u}) g_{il}(\mathbf{u},\mathbf{v})\right]}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&= 1 + \frac{g_{ii}(\mathbf{u},\mathbf{v}) \sum_{l=1}^{p} \left[\mathrm{p}_l^*(\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u})\right]/2}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} - \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u})\right] g_{il}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&\geq 1 + \frac{g_{ii}(\mathbf{u},\mathbf{v}) \sum_{l=1}^{p} \left[\mathrm{p}_l^*(\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u})\right]/2}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&- \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u})\right] \sqrt{g_{ii}(\mathbf{u},\mathbf{v}) g_{ll}(\mathbf{u},\mathbf{v})}}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&= 1 + \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{u}) + \mathrm{p}_l^*(\mathbf{v})\right] \left(g_{ii}(\mathbf{u},\mathbf{v}) - 2\sqrt{g_{ii}(\mathbf{u},\mathbf{v}) g_{ll}(\mathbf{u},\mathbf{v})}\right)}{2 g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&= 1 + \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{u}) + \mathrm{p}_l^*(\mathbf{v})\right] \left[\left(\sqrt{g_{ii}(\mathbf{u},\mathbf{v})} - \sqrt{g_{ll}(\mathbf{u},\mathbf{v})}\right)^2 - g_{ll}(\mathbf{u},\mathbf{v})\right]}{2 g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&\geq 1 - \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{u}) + \mathrm{p}_l^*(\mathbf{v})\right] g_{ll}(\mathbf{u},\mathbf{v})}{2 g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)} \\
&\geq 1 - \frac{\max_{1\leq l\leq p} g_{ll}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*, g)}.
\end{aligned}
$$

Since the right-hand side of the last inequality does not depend on $i$, we have that

$$\min_{1\leq i\leq p} T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g) \geq 1 - \frac{\max_{1\leq l\leq p} g_{ll}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)}.$$

### B.2 Lower bounds without constraint (A.18)

$$\begin{aligned} T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g) &= 1 + \frac{g_{ii}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)} - \sum_{l=1}^{p} \frac{\left[\mathrm{p}_l^*(\mathbf{v}) g_{il}(\mathbf{u},\mathbf{v}) + \mathrm{p}_l^*(\mathbf{u}) g_{il}(\mathbf{u},\mathbf{v})\right]}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)} \\ &\geq 1 + \frac{g_{ii}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)} - \frac{2\max_{1\leq l\leq p} g_{il}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \min_{1\leq i\leq p} T_{ii}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g) &\geq 1 + \frac{\min_{1\leq i\leq p} g_{ii}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)} - \frac{2\max_{1\leq i\leq p}\max_{1\leq l\leq p} g_{il}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)} \\ &\geq 1 - \frac{2\max_{1\leq l,l'\leq p} g_{ll'}(\mathbf{u},\mathbf{v}) - \min_{1\leq i\leq p} g_{ii}(\mathbf{u},\mathbf{v})}{g^{\mathrm{pl}}(\mathbf{u},\mathbf{v};\boldsymbol{\beta}^*,g)}. \end{aligned}$$

## C Proofs regarding consistency and asymptotic normality

In the following proofs we several times refer to auxiliary lemmas stated in Section D.

### C.1 Proof of Theorem 1

For ease of notation we use the abbreviations $\mathrm{p}_i(\mathbf{u})$ and $\mathrm{p}_i^*(\mathbf{u})$ for $\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})$ and $\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)$, respectively. To prove Theorem 1 we invoke Theorem 2 in Waagepetersen and Guan (2009) with $V_n = |W_n|^{1/2}\mathbf{I}_q$ where $\mathbf{I}_q$ is the $q\times q$ identity matrix and we define $\bar{\mathbf{J}}(\boldsymbol{\beta}) = -\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}^\mathsf{T}}\mathbf{e}_n(\boldsymbol{\beta})$ to be the 'average' observed information over $W_n$. It suffices to verify the following conditions

W1 There exists a $t' > 0$ such that $t'_n \geq t'$ for all sufficiently large $n$ where

$$t'_n = \inf_{\|\mathbf{x}\|=1} \mathbf{x}^\mathsf{T} \bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\,\mathbf{x}.$$

W2 As $n\to\infty$, $\|\bar{\mathbf{J}}_n(\boldsymbol{\beta}) - \bar{\mathbf{S}}_n(\boldsymbol{\beta})\|_{\max}$ converges to zero in probability for any $\boldsymbol{\beta}\in\mathbb{R}^q$.

W3 For any $\delta > 0$, $\sup_{\||W_n|^{-1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|\leq\delta} \|\bar{\mathbf{J}}_n(\boldsymbol{\beta}) - \bar{\mathbf{J}}_n(\boldsymbol{\beta}^*)\|_{\max} \to 0$ in probability as $n\to\infty$.

W4 As $n \to \infty$, the sequence $|W_n|^{-1/2}\mathrm{e}^{(n)}(\boldsymbol{\beta}^*)$ is bounded in probability.

*Proof of* W1:
Consider a unit length $(p-1)q \times 1$ vector $\mathbf{x} = [\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_{p-1}^\mathsf{T}]^\mathsf{T}$ with sub-vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\mathsf{T}$ for $i = 1, \ldots, p-1$. By (A.10),

$$\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*) = |W_n|^{-1} \int_{W_n} \mathbf{A}(\mathbf{u}) \otimes \mathbf{Z}(\mathbf{u}, \mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u},$$

where $\otimes$ denotes the Kronecker product and $\mathbf{A}(\mathbf{u}) = [a_{ij}(\mathbf{u})]_{ij}$ is a $(p-1) \times (p-1)$ symmetric matrix of the form

$$\mathbf{A}(\mathbf{u}) = \tag{A.36}$$
$$\begin{bmatrix} \left[1 - \mathrm{p}_1^*(\mathbf{u})\right] \mathrm{p}_1^*(\mathbf{u}) & -\mathrm{p}_1^*(\mathbf{u})\mathrm{p}_2^*(\mathbf{u}) & \cdots & -\mathrm{p}_1^*(\mathbf{u})\mathrm{p}_{p-1}^*(\mathbf{u}) \\ -\mathrm{p}_1^*(\mathbf{u})\mathrm{p}_2^*(\mathbf{u}) & \left[1 - \mathrm{p}_2^*(\mathbf{u})\right] \mathrm{p}_2^*(\mathbf{u}) & & -\mathrm{p}_2^*(\mathbf{u})\mathrm{p}_{p-1}^*(\mathbf{u}) \\ \vdots & & \ddots & \vdots \\ -\mathrm{p}_1^*(\mathbf{u})\mathrm{p}_{p-1}^*(\mathbf{u}) & \cdots & & \left[1 - \mathrm{p}_{p-1}^*(\mathbf{u})\right] \mathrm{p}_{p-1}^*(\mathbf{u}) \end{bmatrix}.$$

Defining now $b_i(\mathbf{u}) = \mathbf{z}(\mathbf{u})^\mathsf{T}\mathbf{x}_i$,

$$\begin{aligned} \mathbf{x}^\mathsf{T}\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\mathbf{x} &= |W_n|^{-1} \int_{W_n} \sum_{i,j=1}^{p-1} a_{ij}(\mathbf{u}) b_i(\mathbf{u}) b_j(\mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u} \\ &= |W_n|^{-1} \int_{W_n} \mathbf{b}^\mathsf{T}(\mathbf{u}) \mathbf{A}(\mathbf{u}) \mathbf{b}(\mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u}, \end{aligned}$$

where $\mathbf{b}(\mathbf{u}) = \left[b_1(\mathbf{u}), \ldots, b_{p-1}(\mathbf{u})\right]^\mathsf{T}$. Using Lemma D.1 in the supplementary material, it immediately follows that

$$\begin{aligned} \mathbf{b}^T(\mathbf{u})\mathbf{A}(\mathbf{u})\mathbf{b}(\mathbf{u}) \geq \mathrm{p}_p^*(\mathbf{u}) \sum_{i=1}^{p-1} \mathbf{b}_i^2(\mathbf{u}) \mathrm{p}_i^*(\mathbf{u}) &= \mathrm{p}_p^*(\mathbf{u}) \sum_{i=1}^{p-1} \mathbf{x}_i^\mathsf{T} \left[\mathbf{z}(\mathbf{u})\mathbf{z}(\mathbf{u})^\mathsf{T} \mathrm{p}_i^*(\mathbf{u})\right] \mathbf{x}_i \\ &= \mathbf{x}^\mathsf{T}\mathbf{Z}_D(\mathbf{u})\mathbf{x}, \end{aligned}$$

where $\mathbf{Z}_D(\mathbf{u})$ is the $(p-1)q \times (p-1)q$ block-diagonal matrix with diagonal blocks $\mathbf{z}(\mathbf{u})\mathbf{z}(\mathbf{u})^\mathsf{T}\mathrm{p}_i^*(\mathbf{u})\mathrm{p}_p^*(\mathbf{u})$. Therefore, it follows that

$$\begin{aligned} \mathbf{x}^\mathsf{T}\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\mathbf{x} &\geq \mathbf{x}^\mathsf{T} \left[|W_n|^{-1} \int_{W_n} \mathbf{Z}_D(\mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u}\right] \mathbf{x} \\ &\geq \lambda_{\min} \left[|W_n|^{-1} \int_{W_n} \mathbf{Z}_D(\mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u}\right] \\ &= \min_{1 \leq i \leq p-1} \left\{ \lambda_{\min} \left[|W_n|^{-1} \int_{W_n} \mathbf{z}(\mathbf{u})\mathbf{z}(\mathbf{u})^\mathsf{T} \mathrm{p}_i^*(\mathbf{u}) \mathrm{p}_p^*(\mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{u}) \mathrm{d}\mathbf{u}\right] \right\} \\ &= \min_{1 \leq i \leq p-1} \left\{ \lambda_{\min} \left[|W_n|^{-1} \int_{W_n} \mathbf{z}(\mathbf{u})\mathbf{z}(\mathbf{u})^\mathsf{T} \lambda_i^*(\mathbf{u}) \mathrm{p}_p^*(\mathbf{u}) \mathrm{d}\mathbf{u}\right] \right\}, \end{aligned}$$

where the second-last equality follows from the block-diagonal structure of the matrix $\mathbf{Z}_D(\mathbf{u})$. The result now follows from C4.

*Proof of* W2:
The proof of W2 follows directly from Lemma D.2 in the supplementary material and a straightforward application of Chebyshev's inequality.

*Proof of* W3:
Consider the following inequality

$$\begin{aligned}&\|\bar{\mathbf{J}}_n(\boldsymbol{\beta}^*) - \bar{\mathbf{J}}_n(\boldsymbol{\beta})\|_{\max} \leq \\ &\|\bar{\mathbf{J}}_n(\boldsymbol{\beta}^*) - \bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\|_{\max} + \|\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*) - \bar{\mathbf{S}}_n(\boldsymbol{\beta})\|_{\max} + \|\bar{\mathbf{S}}_n(\boldsymbol{\beta}) - \bar{\mathbf{J}}_n(\boldsymbol{\beta})\|_{\max}.\end{aligned}$$

The first and third terms on the right hand side tends to zero as $n \to \infty$ by W2. To show that the second term tends to zero for $|W_n|^{-1/2}\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\| \leq \delta$, we use Theorem A.1 in the supplementary material and consider first the diagonal blocks of $\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)$ and $\bar{\mathbf{S}}_n(\boldsymbol{\beta})$:

$$\begin{aligned}&\|\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)_{ii} - \bar{\mathbf{S}}_n(\boldsymbol{\beta})_{ii}\|_{\max} \\ \leq &\frac{1}{|W_n|}\int_{W_n} \|\mathbf{Z}(\mathbf{u},\mathbf{u})\|_{\max}\, \lambda^{\mathrm{pl}}(\mathbf{u})\, \left|[1-\mathrm{p}_i^*(\mathbf{u})]\,\mathrm{p}_i^*(\mathbf{u}) - [1-\mathrm{p}_i(\mathbf{u})]\,\mathrm{p}_i(\mathbf{u})\right| \mathrm{d}\mathbf{u} \\ \leq &\frac{1}{|W_n|}\int_{W_n} K_1^3 p\, \left|[1-\mathrm{p}_i^*(\mathbf{u})]\mathrm{p}_i^*(\mathbf{u}) - [1-\mathrm{p}_i(\mathbf{u})]\,\mathrm{p}_i(\mathbf{u})\right| \mathrm{d}\mathbf{u},\end{aligned}$$

where the last inequality follows from condition C2.

Let $t_{i,\boldsymbol{\beta}}(\mathbf{u}) = [1-\mathrm{p}_i(\mathbf{u})]\,\mathrm{p}_i(\mathbf{u})$. By straightforward calculations it can be shown that $\left\|\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}} t_{i,\tilde{\boldsymbol{\beta}}}(\mathbf{u})\right\|_{\max} \leq C$ for some constant $C > 0$ under condition C2 for any $\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\| \leq |W_n|^{1/2}\delta$, which in turn gives that

$$\frac{\left|t_{i,\boldsymbol{\beta}^*}(\mathbf{u}) - t_{i,\boldsymbol{\beta}}(\mathbf{u})\right|}{|W_n|} \leq \frac{\left\|\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}} t_{i,\tilde{\boldsymbol{\beta}}}(\mathbf{u})\right\|_{\max} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|}{|W_n|} \leq \frac{C}{|W_n|^{1/2}} \frac{\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|}{|W_n|^{1/2}},$$

where the second fraction is bounded by $\delta$ while the first fraction converge to 0 as $n \to \infty$. Therefore, $\|\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)_{ii} - \bar{\mathbf{S}}_n(\boldsymbol{\beta})_{ii}\|_{\max} \to 0$ when $n \to \infty$, $i = 1, \ldots, p$. Similarly, $\left\|\bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)_{ij} - \bar{\mathbf{S}}_n(\boldsymbol{\beta})_{ij}\right\|_{\max} \to 0$ when $n \to \infty$ for any $i \neq j = 1, \ldots, p-1$.

*Proof of* W4:
By Theorem A.2,

$$\begin{aligned}\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, g)_{ij} &= \mathrm{Var}\left[\mathrm{e}^{(n)}(\boldsymbol{\beta}^*)\right] \\ &= \int_{W_n}\int_{W_n} \mathbf{Z}(\mathbf{u},\mathbf{v})\lambda_i^*(\mathbf{u})\lambda_j^*(\mathbf{u}) h_{ij}(\mathbf{u},\mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} + \mathbf{S}_{ij}(\boldsymbol{\beta}^*),\end{aligned}$$

where the functions

$$\begin{aligned} h_{ij}(\mathbf{u},\mathbf{v}) =& g_{ij}(\mathbf{u},\mathbf{v}) + \sum_{l=1}^{p}\sum_{l'=1}^{p} \mathrm{p}_l^*(\mathbf{u})\mathrm{p}_{l'}^*(\mathbf{v}) g_{ll'}(\mathbf{u},\mathbf{v}) \\ & - \sum_{l=1}^{p} \mathrm{p}_l^*(\mathbf{v}) g_{il}(\mathbf{u},\mathbf{v}) - \sum_{l=1}^{p} \mathrm{p}_l^*(\mathbf{u}) g_{jl}(\mathbf{u},\mathbf{v}), \end{aligned} \tag{A.37}$$

$i, j = 1, \ldots, p-1$, can be bounded as follows:

$$\begin{aligned} |h_{ij}(\mathbf{u},\mathbf{v})| =& \Big| g_{ij}(\mathbf{u},\mathbf{v}) - 1 + \sum_{l=1}^{p}\sum_{l'=1}^{p} \mathrm{p}_l^*(\mathbf{u})\mathrm{p}_{l'}^*(\mathbf{v}) \left[g_{ll'}(\mathbf{u},\mathbf{v}) - 1\right] \\ & - \sum_{l=1}^{p} \mathrm{p}_l^*(\mathbf{v}) \left[g_{il}(\mathbf{u},\mathbf{v}) - 1\right] - \sum_{l=1}^{p} \mathrm{p}_l^*(\mathbf{u}) \Big[ g_{jl}(\mathbf{u},\mathbf{v}) - 1 \Big] \Big| \\ & \leq |g_{ij}(\mathbf{u},\mathbf{v}) - 1| + \sum_{l=1}^{p}\sum_{l'=1}^{p} |g_{ll'}(\mathbf{u},\mathbf{v}) - 1| \\ & + \sum_{l=1}^{p} \Big[ |g_{il}(\mathbf{u},\mathbf{v}) - 1| + \Big| g_{jl}(\mathbf{u},\mathbf{v}) - 1 \Big| \Big] \\ & \leq 4 \sum_{l=1}^{p}\sum_{l'=1}^{p} |g_{ll'}(\mathbf{u},\mathbf{v}) - 1| \,. \end{aligned}$$

Therefore, under conditions C2 and C3 and recalling that $g_{ij}$ is isotropic, it is straightforward to show that

$$\begin{aligned} \|\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, g)_{ij}\|_{\max} \leq & K_1^4 \int_{W_n}\int_{W_n} |h_{ij}(\mathbf{u},\mathbf{v})| \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} + K_1^3 |W_n| \\ & \leq 4K_1^4 \int_{W_n}\int_{W_n} \sum_{l=1}^{p}\sum_{l'=1}^{p} |g_{ll'}(\mathbf{u},\mathbf{v}) - 1| \, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} + K_1^3 |W_n| \\ & \leq 4K_1^4 \int_{W_n} \sum_{l=1}^{p}\sum_{l'=1}^{p} K_2 \mathrm{d}\mathbf{u} + K_1^3 |W_n| = (4K_1^4 K_2 p^2 + K_1^3)|W_n|, \end{aligned}$$

which implies that $|W_n|^{-1}\|\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^*, g)_{ij}\|_{\max}$ is asymptotically bounded for any $i, j = 1, \ldots, p-1$. It then follows from Chebychev's inequality that $|W_n|^{-1/2}\mathrm{e}^{(n)}(\boldsymbol{\beta}^*)$ is (element-wise) bounded in probability as $n \to \infty$.

## C.2 Proof of Theorem 2

By a first-order Taylor-expansion of $\mathbf{e}^{(n)}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ around $\boldsymbol{\beta}^*$,

$$\begin{aligned} \mathbf{0} &= \mathbf{e}^{(n)}(\boldsymbol{\beta}^*) - J_n(\tilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\ &= \mathbf{e}^{(n)}(\boldsymbol{\beta}^*) - \left[\mathbf{J}_n(\tilde{\boldsymbol{\beta}}) - \mathbf{S}_n(\boldsymbol{\beta}^*)\right](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \mathbf{S}_n(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \end{aligned}$$

where $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$. Under condition C4, $\bar{\mathbf{S}}_n^{-1}(\boldsymbol{\beta}^*)$ is well defined for sufficiently large $n$ and thus

$$\begin{aligned}\sqrt{|W_n|}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) =& \bar{\mathbf{S}}_n^{-1}(\boldsymbol{\beta}^*) \left[|W_n|^{-1/2}\mathbf{e}^{(n)}(\boldsymbol{\beta}^*)\right] \\ &- \bar{\mathbf{S}}_n^{-1}(\boldsymbol{\beta}^*) \left[\bar{\mathbf{J}}_n(\tilde{\boldsymbol{\beta}}) - \bar{\mathbf{S}}_n(\boldsymbol{\beta}^*)\right] \left[|W_n|^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right] \quad \text{(A.38)} \\ &= \bar{\mathbf{S}}_n^{-1}(\boldsymbol{\beta}^*) \left[|W_n|^{-1/2}\mathbf{e}^{(n)}(\boldsymbol{\beta}^*)\right] + o_p(1),\end{aligned}$$

where the last equality follows from W1-W3 in the proof of Theorem 1 and the conclusion of Theorem 1. Therefore, to prove Theorem 2, it suffices to prove the asymptotic normality of $|W_n|^{-1/2}\mathbf{e}^{(n)}(\boldsymbol{\beta}^*)$.

Let for $\mathbf{l} \in \mathbb{Z}^{\mathrm{d}}$, $C(\mathbf{l})$ denote the unit volume hypercube centered around $\mathbf{l}$ and let $\mathcal{D}_n = \left\{\mathbf{l} \in \mathbb{Z}^d \ : \ C_n(\mathbf{l}) \cap W_n \neq \emptyset\right\}$. Furthermore, define

$$\mathbf{Z}_n(\mathbf{l}) = \left[(\mathbf{Z}_n^{(1)}(\mathbf{l}))^{\mathsf{T}}, \ldots, (\mathbf{Z}_n^{(p-1)}(\mathbf{l}))^{\mathsf{T}}\right]^{\mathsf{T}}$$

where

$$\mathbf{Z}_n^{(i)}(\mathbf{l}) = \sum_{\mathbf{u} \in X_i \cap C(\mathbf{l}) \cap W_n} \mathbf{z}(\mathbf{u}) - \sum_{l=1}^{p} \sum_{\mathbf{u} \in X_l \cap C(\mathbf{l}) \cap W_n} \mathbf{z}(\mathbf{u}) \frac{\exp\left[\boldsymbol{\beta}_i^{*\mathsf{T}} \mathbf{z}(\mathbf{u})\right]}{1 + \sum_{k=1}^{p-1} \exp\left[\boldsymbol{\beta}_k^{*\mathsf{T}} \mathbf{z}(\mathbf{u})\right]}$$

is the restriction of $\mathbf{e}_i(\boldsymbol{\beta}^*)$ to $C(\mathbf{l}) \cap W_n$.

The asymptotic normality of $|W_n|^{-1/2}\mathbf{e}^{(n)}(\boldsymbol{\beta}^*)$ then follows from Theorem 3.1 in Biscio and Waagepetersen (2019) provided the following holds

- $\mathcal{H}1$ $W_1 \subset W_2 \subset \ldots$ and $|\bigcup_{l=1}^{\infty} W_l| = \infty$.
- $\mathcal{H}2$ There exists $\varepsilon > 0$ such that $\alpha_{2,\infty}^X(s) = O(1/s^{d+\varepsilon})$.
- $\mathcal{H}3$ There exists $\tau > 2d/\varepsilon$ such that $\sup_{n \in \mathbb{N}} \sup_{\mathbf{l} \in \mathcal{D}_n} \mathrm{E}\left\|\mathbf{Z}_n(\mathbf{l}) - \mathrm{E}\mathbf{Z}_n(\mathbf{l})\right\|^{2+\tau} < \infty$.
- $\mathcal{H}4$ We have $0 < \liminf_{n\to\infty} \lambda_{\min}\left[\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}^*, g)\right]$, where $\lambda_{\min}(M)$ denotes the smallest eigenvalue of a symmetric matrix $M$.

Conditions $\mathcal{H}1$ and $\mathcal{H}2$ are assumed in conditions C1 and N1, and condition $\mathcal{H}4$ is ensured by condition N3. Therefore, it suffices to show that $\mathcal{H}3$ holds under the conditions of Theorem 2.

*Proof of $\mathcal{H}3$*:
Since $\mathrm{E}\left[\mathbf{Z}_n(\mathbf{l})\right] = \mathbf{0}$ and $m - 2 > 2d/\varepsilon$ it suffices to show that

$$\sup_{n \in \mathbb{N}} \sup_{\mathbf{l} \in \mathcal{D}_n} \mathrm{E}\left\|\mathbf{Z}_n(\mathbf{l})\right\|^m < \infty \quad \text{(A.39)}$$

(i.e. we take $\tau = m-2$). Moreover, letting $Z_{n,j}^{(i)}(\mathbf{l})$ denote the $j$th component of $\mathbf{Z}_n^{(i)}(\mathbf{l})$,

$$\mathrm{E}\,\|\mathbf{Z}_n(\mathbf{l})\|^m \leq (q(p-1))^{m/2} \sum_{i=1}^{p-1} \sum_{j=1}^{q} \mathrm{E}|Z_{n,j}^{(i)}(\mathbf{l})|^m$$

so we just need to show the boundedness of $\mathrm{E}|Z_{n,j}^{(i)}(\mathbf{l})|^m$.

The binomial formula $(x+y)^m = \sum_{k=0}^m \binom{m}{k} x^k y^{m-k}$ gives that

$$\begin{aligned}
&\mathrm{E}\left[|Z_{n,j}^{(i)}(\mathbf{l})|^m\right] \\
&= \sum_{k=0}^{m} \binom{m}{k} \mathrm{E}\left[\left(\sum_{\mathbf{u}\in X_i\cap C(\mathbf{l})\cap W_n} z_j(\mathbf{u})\right)^k \left(-\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap C(\mathbf{l})\cap W_n} z_j(\mathbf{u})\mathrm{p}_i^*(\mathbf{u})\right)^{m-k}\right] \\
&\leq \sum_{k=0}^{m} \binom{m}{k} \mathrm{E}\left[\left(\sum_{\mathbf{u}\in X_i\cap C(\mathbf{l})\cap W_n} K_1\right)^k \left(\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap C(\mathbf{l})\cap W_n} K_1\right)^{m-k}\right] \\
&= K_1^m \sum_{k=0}^{m} \binom{m}{k} \mathrm{E}\left[\left(\sum_{\mathbf{u}\in X_i\cap C(\mathbf{l})\cap W_n} 1\right)^k \left(\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap C(\mathbf{l})\cap W_n} 1\right)^{m-k}\right]
\end{aligned}$$

where the inequality follows from assumption C2.

Regarding the expression inside the expectation,

$$\begin{aligned}
&\left(\sum_{\mathbf{u}\in X_i\cap C(\mathbf{l})\cap W_n} 1\right)^k \left(\sum_{l=1}^{p}\sum_{\mathbf{u}\in X_l\cap C(\mathbf{l})\cap W_n} 1\right)^{m-k} \\
&= \sum_{\mathbf{u}_1\in X_i\cap C(\mathbf{l})\cap W_n} \sum_{\mathbf{u}_2\in X_i\cap C(\mathbf{l})\cap W_n} \cdots \sum_{\mathbf{u}_k\in X_i\cap C(\mathbf{l})\cap W_n} \\
&\quad \sum_{l_1=1}^{p} \sum_{\mathbf{v}_{1,l_1}\in X_{l_1}\cap C(\mathbf{l})\cap W_n} \sum_{l_2=1}^{p} \sum_{\mathbf{v}_{2,l_2}\in X_{l_2}\cap C(\mathbf{l})\cap W_n} \cdots \sum_{l_{m-k}=1}^{p} \sum_{\mathbf{v}_{m-k,l_{m-k}}\in X_{l_{m-k}}\cap C(\mathbf{l})\cap W_n} 1
\end{aligned}$$

The above sum consists of $p^{m-k}$ terms of the form

$$\begin{aligned}
&\sum_{\mathbf{u}_1\in X_i\cap C(\mathbf{l})\cap W_n} \sum_{\mathbf{u}_2\in X_i\cap C(\mathbf{l})\cap W_n} \cdots \sum_{\mathbf{u}_k\in X_i\cap C(\mathbf{l})\cap W_n} \\
&\sum_{\mathbf{v}_{1,l_1}\in X_{l_1}\cap C(\mathbf{l})\cap W_n} \sum_{\mathbf{v}_{2,l_2}\in X_{l_2}\cap C(\mathbf{l})\cap W_n} \cdots \sum_{\mathbf{v}_{m-k,l_{m-k}}\in X_{l_{m-k}}\cap C(\mathbf{l})\cap W_n} 1
\end{aligned}$$

which again can be split into a number of terms according to the possible combinations of ties between the summation indices $\mathbf{u}_j$, $j = 1,\ldots,k$ and $\mathbf{v}_{j',l_{j'}}$,

$j' = 1, \ldots, m-k$, $l_{j'} = 1, \ldots, p$. By Campbell's formula, the expectations of these terms can be evaluated as integrals with respect to cross joint intensities. For the sum where all indices are distinct, the expectation becomes an integral over $C(\mathbf{l})^m$ with respect to the appropriate joint cross intensity $\lambda_{j_1 \cdots j_N}^{(k_1,k_2\ldots,k_N)}$ of total order $m$, where $j_1 = i$, $N \leq p$, $k_1 \geq k$, $\sum_{l=1}^{N} k_l = m$ and $\{j_2, \ldots, j_N\} \subseteq \{1, 2, \ldots, p\} \setminus \{i\}$. For example, if $l_1 = l_2 = \cdots = l_{m-k} = i$ then $N = 1$ and $k_1 = m$ so that the cross joint intensity becomes the $m$th order joint intensity $\lambda_i^{(m)}$ of $X_i$. The joint cross intensity $\lambda_{j_1 \cdots j_N}^{(k_1,k_2\ldots,k_N)}$ is bounded above by $K_1^m C_g$ by conditions C2 and N2.

If not all indices are distinct we obtain lower order integrals involving lower order joint cross densities. These integrals are of a smaller magnitude compared to the case where all indices are distinct. Therefore, under conditions C2 and N2, we have that

$$\begin{aligned} \mathrm{E}\left( \sum_{\mathbf{u} \in X_i \cap C(\mathbf{l}) \cap W_n} 1 \right)^k \left( \sum_{l=1}^{p} \sum_{\mathbf{u} \in X_l \cap C(\mathbf{l}) \cap W_n} 1 \right)^{m-k} &\leq p^{m-k} K K_1^m C_g |C(\mathbf{l})|^m \\ &= p^{m-k} K K_1^m C_g \end{aligned}$$

where $K < \infty$ is an upper bound for the number of combinations of ties mentioned above. This completes the proof of $\mathcal{H}3$.

## C.3 Proof of Theorem 3

The proof of (A.21) in Theorem 3 can be separated into the following steps

A1. $\hat{g}_{ij,kl,n}^{\mathrm{n}}(r; b_n, \boldsymbol{\beta}^*) \xrightarrow{p} g_{ij,kl}(r)$ as $n \to \infty$.

A2. $\hat{g}_{ij,kl,n}^{\mathrm{n}}(r; b_n, \widehat{\boldsymbol{\beta}}) \xrightarrow{p} \hat{g}_{ij,kl,n}^{\mathrm{n}}(r; b_n, \boldsymbol{\beta}^*)$ as $n \to \infty$.

*Proof of A1*:

In the following, we ease the notation by omitting function arguments. We need to show that

$$\hat{g}_{ij,kl,n}^{\mathrm{n}} = \frac{F_{ij,n}}{F_{kl,n}} = \frac{g_{ij}}{g_{kl}} \frac{g_{kl} |W_n|^{-1} F_{ij,n} / \left[ g_{ij} \mathrm{E}(F_{kl,n}) |W_n|^{-1} \right]}{|W_n|^{-1} F_{kl,n} / \left[ \mathrm{E}(F_{kl,n}) |W_n|^{-1} \right]} \xrightarrow{p} g_{ij,kl},$$

as $n \to \infty$, which is equivalent to showing that as $n \to \infty$

$$|W_n|^{-1} \left[ g_{kl} F_{ij,n} - g_{ij} \mathrm{E}(F_{kl,n}) \right] \xrightarrow{p} 0 \quad \text{and} \tag{A.40}$$

$$|W_n|^{-1} \left[ F_{kl,n} - \mathrm{E}(F_{kl,n}) \right] \xrightarrow{p} 0. \tag{A.41}$$

By rewriting (A.40), we can see that

$$\frac{1}{|W_n|}\left[g_{kl}F_{ij,n} - g_{ij}\mathrm{E}(F_{kl,n})\right] = \frac{g_{kl}}{|W_n|}\left[F_{ij,n} - \mathrm{E}(F_{ij,n})\right] + \frac{1}{|W_n|}\left[g_{kl}\mathrm{E}(F_{ij,n}) - g_{ij}\mathrm{E}(F_{kl,n})\right].$$

For the first term on the right hand side it follows from Chebychev's inequality that

$$\mathrm{P}(|g_{kl}|W_n|^{-1}\left[F_{ij,n} - \mathrm{E}(F_{ij,n})\right]| > \varepsilon) < \frac{g_{kl}^2\mathrm{Var}(|W_n|^{-1}F_{ij,n})}{\varepsilon^2}$$

for any $\varepsilon > 0$. Thus the first term converges to zero in probability as $n \to \infty$ due to Lemma D.5 in the supplementary material along with condition C2. The second term can be expanded as

$$\frac{1}{|W_n|}\int_{W_n}\int_{W_n}\left[g_{kl}(r)g_{ij}(||\mathbf{u}-\mathbf{v}||) - g_{ij}(r)g_{ij}(||\mathbf{u}-\mathbf{v}||)\right] \times \left[\sum_{i=1}^{p}\lambda_i^*(\mathbf{u})\right]\left[\sum_{j=1}^{p}\lambda_j^*(\mathbf{v})\right]k_{b_n}(||\mathbf{u}-\mathbf{v}|| - r)\mathbf{d}\mathbf{u}\mathbf{d}\mathbf{v}. \quad \text{(A.42)}$$

By the continuity of $g_{ij}(\cdot)$ and $g_{kl}(\cdot)$, for any $\delta > 0$, we can choose an $\varepsilon > 0$ such that $|g_{kl}(r)g_{ij}(||\mathbf{u}-\mathbf{v}||) - g_{kl}(||\mathbf{u}-\mathbf{v}||)g_{ij}(r)| < \delta$ provided that $|||\mathbf{u}-\mathbf{v}|| - r| < \varepsilon$. By the definition of the kernel function $k(\cdot)$, a bandwidth $b_n < \varepsilon$ implies that the integral is only over $\mathbf{u}$ and $\mathbf{v}$ with $|||\mathbf{u}-\mathbf{v}\| - r| < \varepsilon$. By Lemma D.3 in the supplementary material, this further implies that (A.42) is bounded by $\delta p^2 K_1^2 \tilde{C}_1 r^{d-1}$. Since $\delta > 0$ can be arbitrarily chosen and $b_n \to 0$, we have that (A.42) converges to zero as $n \to \infty$. Finally, (A.41) follows from a simple application of Chebyshev's inequality using the conclusion of Lemma D.5.

*Proof of A2*:

We show that $\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}) \to \hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \boldsymbol{\beta}^*)$ by using a multivariate Taylor approximation:

$$\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}) - \hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \boldsymbol{\beta}^*) = \nabla\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \boldsymbol{\beta}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathsf{T}}, \quad \text{(A.43)}$$

where $\boldsymbol{\beta}'$ is on the line segment that connects $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$. Since the estimator $\widehat{\boldsymbol{\beta}}$ is consistent, i.e. $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathsf{T}} \xrightarrow{p} 0$ as $n \to \infty$, it suffices to show that $\nabla\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \boldsymbol{\beta}')$ is bounded. Some tedious algebra yields that under condition C2,

$$\left\|\nabla F_{ij,n}(r; b_n, \boldsymbol{\beta}')\right\|_{\max} \leq 2\left[\sup_{\mathbf{u}}\|\mathbf{z}(\mathbf{u})\|_{\max}\right]F_{ij,n}(r; b_n, \boldsymbol{\beta}') \leq 2K_1 F_{ij,n}(r; b_n, \boldsymbol{\beta}'),$$

which further gives that

$$\begin{aligned}
&\|\nabla \hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\boldsymbol{\beta}')\|_{\max} \\
&= \left\| \nabla \frac{F_{ij,n}(r;b_n,\boldsymbol{\beta}')}{F_{kl,n}(r;b_n,\boldsymbol{\beta}')} \right\|_{\max} \leq 4K_1 \frac{F_{ij,n}(r;b_n,\boldsymbol{\beta}')}{F_{kl,n}(r;b_n,\boldsymbol{\beta}')} \\
&= 4K_1 \frac{F_{ij,n}(r;b_n,\boldsymbol{\beta}^*)}{F_{kl,n}(r;b_n,\boldsymbol{\beta}^*)} \frac{F_{ij,n}(r;b_n,\boldsymbol{\beta}')}{F_{ij,n}(r;b_n,\boldsymbol{\beta}^*)} \frac{F_{kl,n}(r;b_n,\boldsymbol{\beta}^*)}{F_{kl,n}(r;b_n,\boldsymbol{\beta}')} \\
&= 4K_1 \hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\boldsymbol{\beta}^*) \frac{F_{ij,n}(r;b_n,\boldsymbol{\beta}')/|W_n|}{F_{ij,n}(r;b_n,\boldsymbol{\beta}^*)/|W_n|} \frac{F_{kl,n}(r;b_n,\boldsymbol{\beta}^*)/|W_n|}{F_{kl,n}(r;b_n,\boldsymbol{\beta}')/|W_n|}.
\end{aligned}$$

It follows immediately from A1, that $\hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\boldsymbol{\beta}^*) \xrightarrow{p} g_{ij,kl}(r)$ as $n \to \infty$. Now it remains to show that $\left(F_{ij,n}(r;b_n,\boldsymbol{\beta}^*)/|W_n|\right) / \left(F_{ij,n}(r;b_n,\boldsymbol{\beta}')/|W_n|\right) \xrightarrow{p} 1$ as $n \to \infty$ or equivalently,

$$\frac{1}{|W_n|} \left| F_{ij,n}(r;b_n,\boldsymbol{\beta}^*) - F_{ij,n}(r;b_n,\boldsymbol{\beta}') \right| \xrightarrow{p} 0 \text{ as } n \to \infty. \tag{A.44}$$

Using the inequality $\left\|\nabla F_{ij,n}(r;b_n,\boldsymbol{\beta}')\right\|_{\max} \leq 2K_1 F_{ij,n}(r;b_n,\boldsymbol{\beta}')$ and a Taylor expansion,

$$\begin{aligned}
\frac{1}{|W_n|}|F_{ij,n}(r;b_n,\boldsymbol{\beta}^*) - F_{ij,n}(r;b_n,\boldsymbol{\beta}')| &= \frac{1}{|W_n|}|\nabla F_{ij,n}(r;b_n,\boldsymbol{\beta}'')^{\mathsf{T}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}')| \\
&\leq \frac{2K_1(p-1)q\|\boldsymbol{\beta}^* - \boldsymbol{\beta}'\|_{\max}}{|W_n|} F_{ij,n}(r;b_n,\boldsymbol{\beta}''),
\end{aligned} \tag{A.45}$$

where $\boldsymbol{\beta}''$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}'$. Recalling the definition of $F_{ij,n}$,

$$\begin{aligned}
F_{ij,n}(r;b_n,\boldsymbol{\beta}'') &= \sum_{\substack{\mathbf{u}\in X_i\cap W_n \\ \mathbf{v}\in X_j\cap W_n}}^{\neq} \frac{k_{b_n}(\|\mathbf{u}-\mathbf{v}\|-r)}{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}'')\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta}'')} \\
&= \sum_{\substack{\mathbf{u}\in X_i\cap W_n \\ \mathbf{v}\in X_j\cap W_n}}^{\neq} \frac{k_{b_n}(\|\mathbf{u}-\mathbf{v}\|-r)}{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta}^*)} \times \frac{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta}^*)}{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}'')\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta}'')}.
\end{aligned}$$

Under condition C2, $\|\mathbf{z}(\mathbf{u})\|_{\max} \leq K_1$ ensures that there exists a $c > 0$ such that

$$\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*) \geq c, \qquad i = 1,\ldots,p,$$

for any $\mathbf{u} \in \left\{\mathbf{u} \in \cup_{l=1}^{\infty} W_l : \lambda_0(\mathbf{u}) > 0\right\}$. By consistency of $\hat{\boldsymbol{\beta}}_n$ we have that $\boldsymbol{\beta}' \xrightarrow{p} \boldsymbol{\beta}^*$ and hence $\boldsymbol{\beta}'' \xrightarrow{p} \boldsymbol{\beta}^*$, which implies that with a probability tending

to 1,

$$\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}'') \geq c, \qquad i = 1,\ldots,p,$$

for any $\mathbf{u} \in \left\{\mathbf{u} \in \cup_{l=1}^{\infty} W_l : \lambda_0(\mathbf{u}) > 0\right\}$. Consequently, with a probability tending to 1, we can bound $F_{ij,n}(r; b_n, \boldsymbol{\beta}'')$ as

$$F_{ij,n}(r; b_n, \boldsymbol{\beta}'') \leq \frac{1}{c^2} \sum_{\substack{\mathbf{u}\in X_i\cap W_n \\ \mathbf{v}\in X_j\cap W_n}}^{\neq} \frac{k_{b_n}(\|\mathbf{u}-\mathbf{v}\| - r)}{\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta}^*)\mathrm{p}_j(\mathbf{v};\boldsymbol{\beta}^*)} = \frac{1}{c^2} F_{ij,n}(r; b_n, \boldsymbol{\beta}^*). \quad \text{(A.46)}$$

Using equality (A.41),

$$\begin{aligned}
|W_n|^{-1} F_{ij,n}(r; b_n, \boldsymbol{\beta}^*) &\xrightarrow{p} |W_n|^{-1}\mathrm{E}\left[F_{ij,n}(r; b_n, \boldsymbol{\beta}^*)\right] \\
&= |W_n|^{-1} \int_{W_n^2} \lambda^{\mathrm{pl}}(\mathbf{u})\lambda^{\mathrm{pl}}(\mathbf{v}) g_{ij}(\mathbf{u},\mathbf{v}) k_b(||\mathbf{u}-\mathbf{v}|| - r)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&\leq \frac{p^2 K_1^3}{|W_n|} \int_{W_n^2} k_{b_n}(||\mathbf{u}-\mathbf{v}|| - r)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&\leq p^2 K_1^3 \int_{\mathbb{R}^{\mathrm{d}}} k_{b_n}(||\mathbf{x}|| - r)\mathrm{d}\mathbf{x},
\end{aligned}$$

where the first inequality follows from condition C2. Combining the above inequality with Lemma D.3 gives that, with a probability tending to 1, there exists a constant $C_F$ such that

$$|W_n|^{-1} F_{ij,n}(r; b_n, \boldsymbol{\beta}^*) \leq C_F. \quad \text{(A.47)}$$

Finally, for any $\varepsilon > 0$ and letting $\bar{F}^*_{ij,n} = F_{ij,n}(r; b_n, \boldsymbol{\beta}^*)/|W_n|$,

$$P(\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\|_{\max}\bar{F}^*_{ij,n} > \varepsilon) \leq P(\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\|_{\max} C_F > \varepsilon) + P(\bar{F}^*_{ij,n} > C_F)$$

Thus, since $\boldsymbol{\beta}' \xrightarrow{p} \boldsymbol{\beta}^*$ as $n \to \infty$, (A.44) immediately follows from inequalities (A.45)-(A.47), which completes the proof of A2. Therefore, the proof of (A.21) in Theorem 3 is finished.

To show (A.22) in Theorem 3, note that

$$\begin{aligned}
|\hat{g}^{\mathrm{r}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}_n) - g_{ij,kl}(r)| \leq\ & |\hat{g}^{\mathrm{r}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}_n) - \hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}_n)| \\
&+ |\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}_n) - g_{ij,kl}(r)|,
\end{aligned}$$

so that it is enough to show that

$$|\hat{g}^{\mathrm{r}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}_n) - \hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}}_n)| \xrightarrow{p} 0 \text{ for } r \geq R^* \text{ as } n \to \infty. \quad \text{(A.48)}$$

Combining restriction (A.18) and the convergence $\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}_n}) \xrightarrow{p} g_{ij,kl}(r)$, as $n \to \infty$, we obtain that with a probability tending to 1, $\hat{g}^{\mathrm{n}}_{ij,kl,n}(r; b_n, \widehat{\boldsymbol{\beta}_n})$

satisfies restriction $\hat{g}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}_n}) \leq \sqrt{\hat{g}_{ii,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}_n})\hat{g}_{jj,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}_n})}$ for a sufficiently large $n$, in which case $|\hat{g}^{\mathrm{r}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n) - \hat{g}^{\mathrm{n}}_{ij,kl,n}(r;b_n,\widehat{\boldsymbol{\beta}}_n)| = 0$ by the design of algorithm (A.19). Therefore, (A.48) follows, which completes the proof of (A.22) in Theorem 3.

# D Auxiliary Lemmas

**Lemma D.1**
Define $\mathbf{A}(\mathbf{u})$ as in (A.36). Then for any $\mathbf{b} = (b_1,\ldots,b_{p-1})^{\mathsf{T}} \in \mathbb{R}^{p-1}$,

$$\mathbf{b}^{\mathsf{T}}\mathbf{A}(\mathbf{u})\mathbf{b} \geq \mathrm{p}^*_p(\mathbf{u}) \sum_{i=1}^{p-1} b_i^2 \mathrm{p}^*_i(\mathbf{u}).$$

*Proof.* Using the Cauchy-Schwartz inequality,

$$\begin{aligned}
\mathbf{b}^{\mathsf{T}}\mathbf{A}(\mathbf{u})\mathbf{b} &= \sum_{i=1}^{p-1} b_i^2 \mathrm{p}^*_i(\mathbf{u}) - \left[\sum_{i=1}^{p-1} b_i \mathrm{p}^*_i(\mathbf{u})\right]^2 \\
&= \sum_{i=1}^{p-1} b_i^2 \mathrm{p}^*_i(\mathbf{u}) - \left[\sum_{i=1}^{p-1} b_i \frac{\mathrm{p}^*_i(\mathbf{u})}{1-\mathrm{p}^*_p(\mathbf{u})}\right]^2 [1-\mathrm{p}^*_p(u)]^2 \\
&\geq \sum_{i=1}^{p-1} b_i^2 \mathrm{p}^*_i(\mathbf{u}) - \left[\sum_{i=1}^{p-1} b_i^2 \mathrm{p}^*_i(\mathbf{u})\right] [1-\mathrm{p}^*_p(\mathbf{u})] \\
&= \mathrm{p}^*_p(\mathbf{u}) \sum_{i=1}^{p-1} b_i^2 \mathrm{p}^*_i(\mathbf{u}).
\end{aligned}$$

□

**Lemma D.2**
Assume C1 - C3 holds true. Then as $n \to \infty$, $|W_n|^{-1}\mathrm{Var}^{\odot}\left[\mathbf{J}_n(\boldsymbol{\beta})_{ij}\right] \leq C$ for some constant $0 < C < \infty$ for any $i,j = 1,\ldots,p-1$. Here, for a random matrix $\mathbf{A}$, $\mathrm{Var}^{\odot}(\mathbf{A})$ denotes the element-wise variance of $\mathbf{A}$.

*Proof.* Denote by $\mathbf{A}^{\odot 2}$ the element-wise square of the matrix $\mathbf{A}$ and by $\mathbf{A}\odot\mathbf{B}$ the element-wise product of matrices $\mathbf{A}$ and $\mathbf{B}$. Recall that $\mathrm{E}\left[\mathbf{J}_n(\boldsymbol{\beta})_{ij}\right] = -\mathbf{S}_n(\boldsymbol{\beta})_{ij}$, $i,j = 1,\ldots,p-1$. Clearly, $\mathrm{Var}^{\odot}\left[\mathbf{J}_n(\boldsymbol{\beta})_{ij}\right] = \mathrm{E}\left[\mathbf{J}^{\odot 2}_n(\boldsymbol{\beta})_{ij}\right] - \mathbf{S}^{\odot 2}_n(\boldsymbol{\beta})_{ij}$. Let

$$\begin{aligned}
\mathbf{H}_{ii}(\mathbf{u}) &= \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\left[1-\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\right] \quad \text{and} \\
\mathbf{H}_{ij}(\mathbf{u}) &= \mathbf{Z}(\mathbf{u},\mathbf{u})\mathrm{p}_i(\mathbf{u};\boldsymbol{\beta})\mathrm{p}_j(\mathbf{u};\boldsymbol{\beta}).
\end{aligned}$$

The block elements in $\mathbf{J}^{\odot 2}_n(\boldsymbol{\beta})$ are of the form

$$\mathbf{J}^{\odot 2}_n(\boldsymbol{\beta})_{ij} = \sum_{l,l'=1}^{p} \sum_{\substack{\mathbf{u}\in X_l\cap W_n \\ \mathbf{v}\in X_{l'}\cap W_n}} \mathbf{H}_{ij}(\mathbf{u}) \odot \mathbf{H}_{ij}(\mathbf{v}),$$

for $i, j = 1, \ldots, p-1$, whose expectations are of the form

$$\begin{aligned}\mathrm{E}\left[\mathbf{J}_n^{\odot 2}(\boldsymbol{\beta})_{ij}\right] =& \sum_{l,l'=1}^{p}\sum \int_{W_n}\int_{W_n} \mathbf{H}_{ij}(\mathbf{u}) \odot \mathbf{H}_{ij}(\mathbf{v})\lambda_{ll'}(\mathbf{u},\mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &+ \sum_{l=1}^{p}\int_{W_n} \mathbf{H}_{ij}(\mathbf{u}) \odot \mathbf{H}_{ij}(\mathbf{u})\lambda_l^*(\mathbf{u})\mathrm{d}\mathbf{u}.\end{aligned}$$

Using Theorem A.1, it follows that the squared sensitivity is given by

$$\mathbf{S}_n^{\odot 2}(\boldsymbol{\beta})_{ij} = \sum_{l,l'=1}^{p}\sum \int_{W_n}\int_{W_n} \mathbf{H}_{ij}(\mathbf{u}) \odot \mathbf{H}_{ij}(\mathbf{v})\lambda_l^*(\mathbf{u})\lambda_l^*(\mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$$

for $i, j = 1, \ldots, p-1$. By condition C2, $\|\mathbf{H}_{ij}\|_{\max} \leq K_1^2$ for any $i, j = 1, \ldots, p-1$, which, together with condition C3 and isotropy of $g_{ll'}$, further implies that

$$\begin{aligned}&\mathrm{Var}^{\odot}\left[\mathbf{J}_n(\boldsymbol{\beta})_{ij}\right] \\ &= \int_{W_n^2} \mathbf{H}_{ij}(\mathbf{u}) \odot \mathbf{H}_{ij}(\mathbf{v}) \sum_{l,l'=1}^{p}\sum \left[\lambda_{ll'}(\mathbf{u},\mathbf{v}) - \lambda_l^*(\mathbf{u})\lambda_{l'}(\mathbf{v})\right] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &+ \int_{W_n} \mathbf{H}_{ij}(\mathbf{u}) \odot \mathbf{H}_{ij}(\mathbf{u})\lambda^{\mathrm{pl}}(\mathbf{u})\mathrm{d}\mathbf{u} \\ &\leq \int_{W_n^2} K_1^4 \sum_{l,l'=1}^{p}\sum |\lambda_{ll'}(\mathbf{u},\mathbf{v}) - \lambda_l^*(\mathbf{u})\lambda_{l'}^*(\mathbf{v})| \ \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} + \int_{W_n} K_1^4\lambda^{\mathrm{pl}}(\mathbf{u})\mathrm{d}\mathbf{u} \\ &\leq \sum_{l,l'=1}^{p}\sum \int_{W_n^2} K_1^6 \left|g_{ll'}(\mathbf{u},\mathbf{v}) - 1\right| \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} + \sum_{l=1}^{p}\int_{W_n} K_1^5\mathrm{d}\mathbf{u} \\ &\leq \sum_{l,l'=1}^{p}\sum \int_{W_n} K_1^6K_2\mathrm{d}\mathbf{u} + pK_1^5|W_n| \\ &= |W_n|(p^2K_1^6K_2 + pK_1^5),\end{aligned}$$

which yields that $|W_n|^{-1}\mathrm{Var}^{\odot}\left[\mathbf{J}_n(\boldsymbol{\beta})\right]_{ij} \leq C$ with $C = p^2K_1^6K_2 + pK_1^5$, for any $i, j = 1, \ldots, p-1$. □

**Lemma D.3**
Let $b > 0$ be a bandwidth and $k_b(\cdot) = k(\cdot/b)/b$ with a kernel function $k(\cdot)$ defined on a bounded support in $\mathbb{R}$. Then for $b$ small enough, we have that

$$\begin{aligned}\int_{\mathbb{R}^d} k_b(\|\mathbf{u}\| - r)\mathrm{d}\mathbf{u} &\leq \tilde{C}_1 r^{d-1} \\ \int_{\mathbb{R}^d} k_b^2(\|\mathbf{u}\| - r)\mathrm{d}\mathbf{u} &\leq \tilde{C}_2 \frac{1}{b} r^{d-1},\end{aligned}$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are some positive constants.

*Proof.* Without loss of generality we assume that the kernel function $k(\cdot)$ has a bounded support $[-1,1]$. Using the polar coordinates transformation

$$\int_{\mathbb{R}^d} f(\mathbf{u})\mathrm{d}\mathbf{u} = \int_{\mathbb{S}^{\mathrm{d}-1}} \int_0^\infty f(t\mathbf{v})t^{\mathrm{d}-1}\mathrm{d}t\nu_d(\mathrm{d}\mathbf{v}),$$

where $\nu_d(\cdot)$ is surface measure on the unit sphere $\mathbb{S}^{\mathrm{d}-1}$ in $\mathbb{R}^d$, we have that

$$\begin{aligned} I(r;b) &\equiv \int_{\mathbb{R}^d} k_b(\|\mathbf{u}\|-r)\mathrm{d}\mathbf{u} = \frac{1}{b}\int_{\mathbb{R}^d} k\left(\frac{\|\mathbf{u}\|-r}{b}\right)\mathrm{d}\mathbf{u} \\ &= \frac{1}{b}\int_{\mathbb{S}^{\mathrm{d}-1}}\int_0^\infty k\left(\frac{t-r}{b}\right)t^{d-1}\mathrm{d}t\nu_d(\mathrm{d}\mathbf{v}) \\ &= \frac{\nu_d\left(\mathbb{S}^{\mathrm{d}-1}\right)}{b}\int_0^\infty k\left(\frac{t-r}{b}\right)t^{d-1}\mathrm{d}t. \end{aligned}$$

When substituting the variable $s=(t-r)/b$, we have that

$$\begin{aligned} I(r;b) &= \nu_d\left(\mathbb{S}^{\mathrm{d}-1}\right)\int_{-r/b}^\infty k(s)(bs+r)^{d-1}\mathrm{d}s \\ &\le \nu_d\left(\mathbb{S}^{\mathrm{d}-1}\right)\left[\sup_{s\in[-1,1]}k(s)\right]\int_{\max\{-1,-r/b\}}^1 (bs+r)^{d-1}\mathrm{d}s. \\ &= \nu_d\left(\mathbb{S}^{\mathrm{d}-1}\right)\left[\sup_{s\in[-1,1]}k(s)\right]\frac{1}{db}\Big[(b+r)^d-(b\max\{-1,-r/b\}+r)^d\Big]. \end{aligned}$$

Applying L'Hospital's rule,

$$\lim_{b\to 0}\frac{1}{b}\Big[(b+r)^d-(b\max\{-1,-r/b\}+r)^d\Big] = 2dr^{d-1}.$$

Hence, for $b$ small enough, there exists a constant $\tilde{C}_1$ so that

$$I(r;b)\le \tilde{C}_1 r^{d-1}$$

The second inequality follows the same way. □

**Lemma D.4**
The variance of $F_{ij}(r; b, \boldsymbol{\beta}^*)$ in (A.17) for $i \neq j$ is:

$$
\begin{aligned}
&\mathrm{Var}[F_{ij}(r; b, \boldsymbol{\beta}^*)] = \\
&\int_{W^4} \left[ g_{ij}^{(2,2)}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2) - g_{ij}(\mathbf{u}_1, \mathbf{v}_1) g_{ij}(\mathbf{u}_2, \mathbf{v}_2) \right] \\
&\times t(\mathbf{u}_1, \mathbf{v}_1) t(\mathbf{u}_2, \mathbf{v}_2) \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \mathrm{d}\mathbf{v}_1 \mathrm{d}\mathbf{v}_2 \\
&+ \int_{W^3} g_{ij}^{(1,2)}(\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2) t(\mathbf{u}, \mathbf{v}_1) \frac{\lambda^{\mathrm{pl}}(\mathbf{v}_2)}{\mathrm{p}_i(\mathbf{u}; \boldsymbol{\beta}^*)} k_b(\|\mathbf{u} - \mathbf{v}_2\| - r) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{v}_2 \\
&+ \int_{W^3} g_{ij}^{(2,1)}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}) t(\mathbf{u}_1, \mathbf{v}) \frac{\lambda^{\mathrm{pl}}(\mathbf{u}_2)}{\mathrm{p}_j(\mathbf{v}; \boldsymbol{\beta}^*)} k_b(\|\mathbf{u}_2 - \mathbf{v}\| - r) \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \mathrm{d}\mathbf{v} \\
&+ \int_{W^2} g_{ij}^{(1,1)}(\mathbf{u}, \mathbf{v}) t(\mathbf{u}, \mathbf{v}) \left[ \mathrm{p}_i(\mathbf{u}; \boldsymbol{\beta}^*) \mathrm{p}_j(\mathbf{v}; \boldsymbol{\beta}^*) \right]^{-1} k_b(\|\mathbf{u} - \mathbf{v}\| - r) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v}
\end{aligned}
$$

and for $i = j$:

$$
\begin{aligned}
&\mathrm{Var}[F_{ii}(r; b, \boldsymbol{\beta}^*)] = \\
&\int_{W^4} \left[ g_i^{(4)}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4) - g_i(\mathbf{u}_1, \mathbf{u}_2) g_i(\mathbf{u}_3, \mathbf{u}_4) \right] \\
&\times t(\mathbf{u}_1, \mathbf{u}_2) t(\mathbf{u}_3, \mathbf{u}_4) \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \mathrm{d}\mathbf{u}_3 \mathrm{d}\mathbf{u}_4 \\
&+ 4 \int_{W^3} g_i^{(3)}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) t(\mathbf{u}_1, \mathbf{u}_2) \frac{\lambda^{\mathrm{pl}}(\mathbf{u}_3)}{\mathrm{p}_i(\mathbf{u}_1; \boldsymbol{\beta}^*)} k_b(\|\mathbf{u}_1 - \mathbf{u}_3\| - r) \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \mathrm{d}\mathbf{u}_3 \\
&+ 2 \int_{W^2} g_i(\mathbf{u}_1, \mathbf{u}_2) t(\mathbf{u}_1, \mathbf{u}_2) \left[ \mathrm{p}_i(\mathbf{u}_1; \boldsymbol{\beta}^*) \mathrm{p}_i(\mathbf{u}_2; \boldsymbol{\beta}^*) \right]^{-1} k_b(\|\mathbf{u}_1 - \mathbf{u}_2\| - r) \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2,
\end{aligned}
$$

where $t(\mathbf{u}, \mathbf{v}) = k_b(\|\mathbf{u} - \mathbf{v}\| - r) \lambda^{\mathrm{pl}}(\mathbf{u}) \lambda^{\mathrm{pl}}(\mathbf{v})$.

*Proof.* The variance is

$$
\mathrm{Var}[F_{ij}(r; b, \boldsymbol{\beta}^*)] = \mathrm{E}[F_{ij}(r; b, \boldsymbol{\beta}^*)^2] - \mathrm{E}[F_{ij}(r; b, \boldsymbol{\beta}^*)]^2,
$$

where

$$
\mathrm{E}[F_{ij}(r; b, \boldsymbol{\beta}^*)]^2 = \left[ \int_{W^2} g_{ij}(\mathbf{u}, \mathbf{v}) t(\mathbf{u}, \mathbf{v}) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v} \right]^2.
$$

Define $s_{ij}(\mathbf{u}, \mathbf{v}) = \left[ \mathrm{p}_i(\mathbf{u}; \boldsymbol{\beta}^*) \mathrm{p}_j(\mathbf{v}; \boldsymbol{\beta}^*) \right]^{-1}$ and suppose first that $i \neq j$. Then

we have that

$$
\begin{aligned}
&\mathrm{E}[F_{ij}(r;b,\boldsymbol{\beta}^*)^2] \\
&= \mathrm{E} \sum_{\substack{\mathbf{u}_1 \in X_i \cap W \\ \mathbf{v}_1 \in X_j \cap W}}^{\neq} s_{ij}(\mathbf{u}_1,\mathbf{v}_1) k_b(\|\mathbf{u}_1 - \mathbf{v}_1\| - r) \sum_{\substack{\mathbf{u}_2 \in X_i \cap W \\ \mathbf{v}_2 \in X_j \cap W}}^{\neq} s_{ij}(\mathbf{u}_2,\mathbf{v}_2) k_b(\|\mathbf{u}_2 - \mathbf{v}_2\| - r) \\
&= \mathrm{E} \sum_{\substack{\mathbf{u}_1,\mathbf{u}_2 \in X_i \cap W \\ \mathbf{v}_1,\mathbf{v}_2 \in X_j \cap W}}^{\neq} s_{ij}(\mathbf{u}_1,\mathbf{v}_1) s_{ij}(\mathbf{u}_2,\mathbf{v}_2) k_b(\|\mathbf{u}_1 - \mathbf{v}_1\| - r) k_b(\|\mathbf{u}_2 - \mathbf{v}_2\| - r) \\
&\quad + \mathrm{E} \sum_{\substack{\mathbf{u} \in X_j \cap W \\ \mathbf{v}_1,\mathbf{v}_2 \in X_j \cap W}}^{\neq} s_{ij}(\mathbf{u},\mathbf{v}_1) s_{ij}(\mathbf{u},\mathbf{v}_2) k_b(||\mathbf{u} - \mathbf{v}_1|| - r) k_b(||\mathbf{u} - \mathbf{v}_2|| - r) \\
&\quad + \mathrm{E} \sum_{\substack{\mathbf{u}_1,\mathbf{u}_2 \in X_i \cap W \\ \mathbf{v} \in X_j \cap W}}^{\neq} s_{ij}(\mathbf{u}_1,\mathbf{v}) s_{ij}(\mathbf{u}_2,\mathbf{v}) k_b(\|\mathbf{u}_1 - \mathbf{v}\| - r) k_b(\|\mathbf{u}_2 - \mathbf{v}\| - r) \\
&\quad + \mathrm{E} \sum_{\substack{\mathbf{u} \in X_i \cap W \\ \mathbf{v} \in X_j \cap W}}^{\neq} s_{ij}(\mathbf{u},\mathbf{v})^2 k_b(||\mathbf{u} - \mathbf{v}|| - r)^2,
\end{aligned}
$$

where we recall that $\sum^{\neq}$ means summation over distinct points. If $i = j$ we can rename the indices and have that

$$
\begin{aligned}
&\mathrm{E}[F_{ii}(r;b,\boldsymbol{\beta}^*)^2] \\
&= \mathrm{E} \sum_{\mathbf{u}_1,\mathbf{u}_2 \in X_i \cap W}^{\neq} s_{ii}(\mathbf{u}_1,\mathbf{u}_2) k_b(\|\mathbf{u}_1 - \mathbf{u}_2\| - r) \sum_{\mathbf{u}_3,\mathbf{u}_4 \in X_i \cap W}^{\neq} s_{ii}(\mathbf{u}_3,\mathbf{u}_4) k_b(\|\mathbf{u}_3 - \mathbf{u}_4\| - r) \\
&= \mathrm{E} \sum_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4 \in X_i \cap W}^{\neq} s_{ii}(\mathbf{u}_1,\mathbf{u}_2) s_{ii}(\mathbf{u}_3,\mathbf{u}_4) k_b(\|\mathbf{u}_1 - \mathbf{u}_2\| - r) k_b(\|\mathbf{u}_3 - \mathbf{u}_4\| - r) \\
&\quad + 4\mathrm{E} \sum_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3 \in X_i \cap W}^{\neq} s_{ii}(\mathbf{u}_1,\mathbf{u}_2) s_{ii}(\mathbf{u}_1,\mathbf{u}_3) k_b(\|\mathbf{u}_1 - \mathbf{u}_2\| - r) k_b(\|\mathbf{u}_1 - \mathbf{u}_3\| - r) \\
&\quad + 2\mathrm{E} \sum_{\mathbf{u}_1,\mathbf{u}_2 \in X_i \cap W}^{\neq} s_{ii}(\mathbf{u}_1,\mathbf{u}_2)^2 k_b(\|\mathbf{u}_1 - \mathbf{u}_2\| - r)^2.
\end{aligned}
$$

Lemma D.4 then follows directly from applying Campbell's formula to each expectation. □

**Lemma D.5**
Under conditions C2,K1-K3, the variance of $|W_n|^{-1} F_{ij,n}(r;b_n,\boldsymbol{\beta}^*_{ij})$ converges to zero in probability as $n \to \infty$.

*Proof.* Following Lemma D.4 we have two cases for $\text{Var}\left[F_{ij,n}(r; b_n, \boldsymbol{\beta}_{ij}^*)\right]$ when $i \neq j$ and $i = j$. For $i \neq j$ we write $\text{Var}\left[F_{ij,n}(r; b_n, \boldsymbol{\beta}_{ij}^*)\right]$ as a sum of four terms $T_{1,n}, \ldots, T_{4,n}$ and for $i = j$ we write $\text{Var}\left[F_{ii,n}(r; b_n, \boldsymbol{\beta}_{ii}^*)\right]$ as a sum of three terms $T'_{1,n}, \ldots, T'_{3,n}$. First we consider $i \neq j$ and applying condition C2, translation invariance of $g_{ij}^{(2,2)}$ (condition K1) and a change of variable, it follows that $T_{1,n}$ is bounded as

$$T_{1,n} \leq p^4 K_1^4 \int_{W_n^4} \left| g_{ij}^{(2,2)}(\mathbf{0}, \mathbf{u}, \mathbf{v}, \mathbf{w} + \mathbf{u}) - g_{ij}(\mathbf{0}, \mathbf{v}) g_{ij}(\mathbf{0}, \mathbf{w}) \right| \\ k_{b_n}(\|\mathbf{v}\| - r) k_{b_n}(\|\mathbf{w}\| - r) \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v} \mathrm{d}\mathbf{w}.$$

Using the second part of condition K1, the above upper bound can be simplied to

$$p^4 K_1^4 K_3 |W_n| \int_{W_n} k_{b_n}\left(\|\mathbf{v}\| - r\right) \mathrm{d}\mathbf{v} \int_{W_n} k_{b_n}\left(\|\mathbf{w}\| - r\right) \mathrm{d}\mathbf{w}.$$

Consequently, it follows from Lemma D.3 and condition K3 that

$$\frac{T_{1,n}}{|W_n|^2} \leq \frac{p^4 K_1^4 K_3 \left(\tilde{C}_1 r^{d-1}\right)^2}{|W_n|} \to 0, \text{ as } n \to \infty.$$

Using similar arguments for $T'_{1,n}$ is can be shown that $T'_{1,n}$ tend to zero. Regarding $T_{2,n}$, first note that under condition C2, $\|\mathbf{z}(\mathbf{u})\|_{\max} \leq K_1$ ensures that there exists a $c > 0$ such that

$$p_i(\mathbf{u}; \boldsymbol{\beta}^*) \geq c, \qquad i = 1, \ldots, p, \tag{A.49}$$

for any $\mathbf{u} \in \left\{\mathbf{u} \in \cup_{l=1}^{\infty} W_l : \lambda_0(\mathbf{u}) > 0\right\}$. Applying further condition K2, $T_{2,n}$ is bounded as

$$T_{2,n} \leq \frac{K_4 K_1^3}{c} \int_{W_n^3} k_{b_n}(\|\mathbf{u} - \mathbf{v}_1\| - r) k_{b_n}(\|\mathbf{u} - \mathbf{v}_2\| - r) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v}_1 \mathrm{d}\mathbf{v}_2$$

By Lemma D.3,

$$T_{2,n} \leq \frac{K_4 K_1^3}{c} |W_n| (\tilde{C}_1 r^{d-1})^2.$$

Hence $T_{2,n}/|W_n|^2 \to 0$ as $n \to \infty$. It follows along the same lines that $T_{3,n}$ and $T'_{2,n}$ tend to zero as $n \to \infty$.

Regarding the fourth term $T_{4,n}$, we have

$$T_{4,n} \leq \frac{2K_4 p^2 K_1^2}{|W_n|^2 c^2} \int_{W_n^2} k_{b_n}^2(\|\mathbf{u} - \mathbf{v}\| - r) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v} \leq \frac{2K_4 p^2 K_1^2}{|W_n| c^2} \int_{\mathbb{R}^d} k_{b_n}^2(\|\mathbf{u}\| - r) \mathrm{d}\mathbf{u}$$

Applying Lemma D.3 and the last part of condition K3, $T_{4,n}$ tends to zero as $n \to \infty$. Using similar arguments for $T'_{3,n}$ it can be shown that $T'_{3,n}$ tends to zero as $n \to \infty$. Thus $\text{Var}\left[|W_n|^{-1} F_{ij,n}(r; b_n, \boldsymbol{\beta}^*)\right] \to 0$ as $n \to \infty$. □

# E Performances of kernel estimators of cross PCF ratios

Following the last paragraph in Section 3.4, Figure A.9 illustrates the performances of the "naive" kernel estimator and "regularized" kernel estimator for some ratios $g_{ij}(r)/g_{44}(r)$ under various window sizes. Both estimators are unbiased and the "regularized" kernel estimator has a slightly narrower 95% probability band than the "naive" kernel estimator for the window $W_1$. As the observation window is increased from $W_1$ to $W_2$, all probability bands become much tighter, supporting the theoretical findings in Theorem 3. However, we do not observe appreciable differences between these two nonparametric estimators in terms of estimation accuracies.

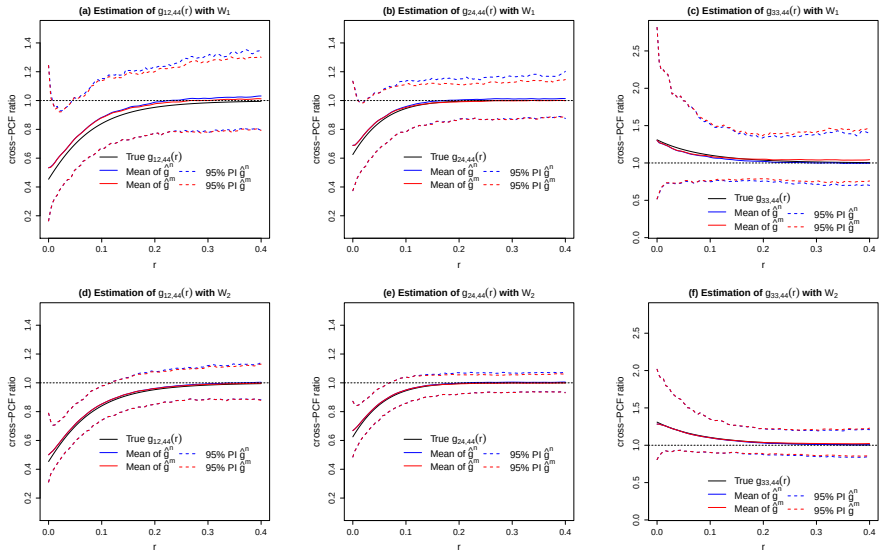


**Fig. A.9:** Means of estimated cross PCF ratios and point-wise 95% probability intervals for cross PCF ratios. Upper row: $W_1$, lower row: $W_2$.

To shed more light on why $\mathrm{CI}_{\hat{g}^{\mathrm{rr}}}$ outperforms $\mathrm{CI}_{\hat{g}^{\mathrm{nn}}}$, we study the diagonal blocks of the covariance matrix estimator (A.14) and focus on the diagonal elements in each $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}^*, g)_{ii}$ corresponding to the intercept, which can be rewritten as

$$\widehat{\tau}_i(\boldsymbol{\beta}^*, g) = \sum_{\mathbf{u} \in X^{\mathrm{pl}}} [1 - \mathrm{p}_i^*(\mathbf{u})]\, \mathrm{p}_i^*(\mathbf{u}) + \sum_{l=1}^{L-1} \varphi_{i,l}\,(\boldsymbol{\beta}^*, g)\,, \tag{A.50}$$

for $i = 1, \ldots, p-1$, where for $L \geq 1$ and $l = 1, \ldots, L-1$,

$$\varphi_{i,l}(\boldsymbol{\beta}^*, g) = \sum_{\mathbf{u},\mathbf{v} \in X^{\mathrm{pl}}}^{\neq} \mathrm{p}_i^*(\mathbf{u}) \mathrm{p}_i^*(\mathbf{v}) T_{ii}(\mathbf{u}, \mathbf{v}; \boldsymbol{\beta}^*, g) 1(r_l < \|\mathbf{u} - \mathbf{v}\| \leq r_{l+1}), \quad \text{(A.51)}$$

for an equally-spaced partition $0 = r_1 < r_2 < \cdots < r_L = R$ of the interval $[0, R]$ and with $T_{ii}(\mathbf{u}, \mathbf{v}; \boldsymbol{\beta}^*, g)$ defined in (A.12).

The $\widehat{\tau}_i(\boldsymbol{\beta}^*, g)$'s are estimators of variances. However, after plugging in $\widehat{\boldsymbol{\beta}}$ and estimated cross PCFs, the resulting $\widehat{\tau}_i(\widehat{\boldsymbol{\beta}}, \hat{g})$'s are not guaranteed to be positive. This issue is especially severe for the "naive" kernel cross PCF ratio estimators. Figure A.10 compares the means of $\varphi_i(r; \boldsymbol{\beta}^*, g)$, $\varphi_i(r; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ and $\varphi_i(r; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{r}})$ based on 1000 simulations together with point-wise 95% probability bands for $i = 1, 2, 3$. While there exist little differences between the means of $\varphi_i(r; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ and $\varphi_i(r; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{r}})$, the low quantiles of $\varphi_i(r; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$ can take very large negative values, which may lead to a small and even negative value of $\widehat{\tau}_i(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{n}})$. In contrast, the lower quantiles of $\varphi_i(r; \widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{r}})$ are always close to 0, and thus the associated $\widehat{\tau}_i(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{r}})$'s are bounded away from negative values. Since the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ takes the form $\left[\widehat{\mathbf{S}}_n(\widehat{\boldsymbol{\beta}})\right]^{-1} \widehat{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{\beta}}, \hat{g}) \left[\widehat{\mathbf{S}}_n(\widehat{\boldsymbol{\beta}})\right]^{-1}$, it is generally the case that larger diagonal elements in $\widehat{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{\beta}}, \hat{g})$ leads to larger estimated variances for $\widehat{\boldsymbol{\beta}}$. Therefore, $\mathrm{CI}_{\hat{g}^{\mathrm{r}}}$ tends to achieve higher coverage probability than that of $\mathrm{CI}_{\hat{g}^{\mathrm{n}}}$.

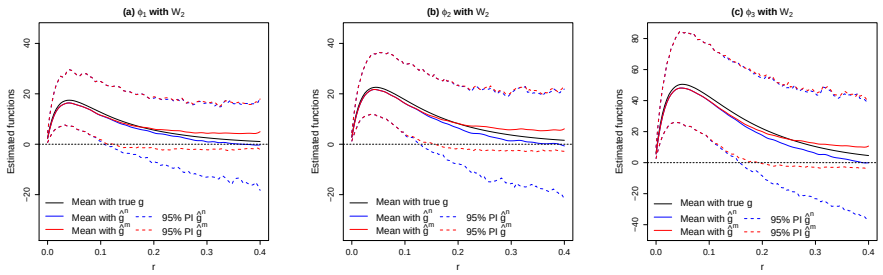


**Fig. A.10:** Means of $\varphi_{i,l}(\boldsymbol{\beta}^*, g)$, $\varphi_{i,l}(\boldsymbol{\beta}, \hat{g}^{\mathrm{n}})$ and $\varphi_{i,l}(\widehat{\boldsymbol{\beta}}, \hat{g}^{\mathrm{r}})$ against $r_l$, $l = 1, \ldots, L$, $i = 1, 2, 3$, and point-wise 95% probability bands.

# F Goodness-of-fit assessments of crime data

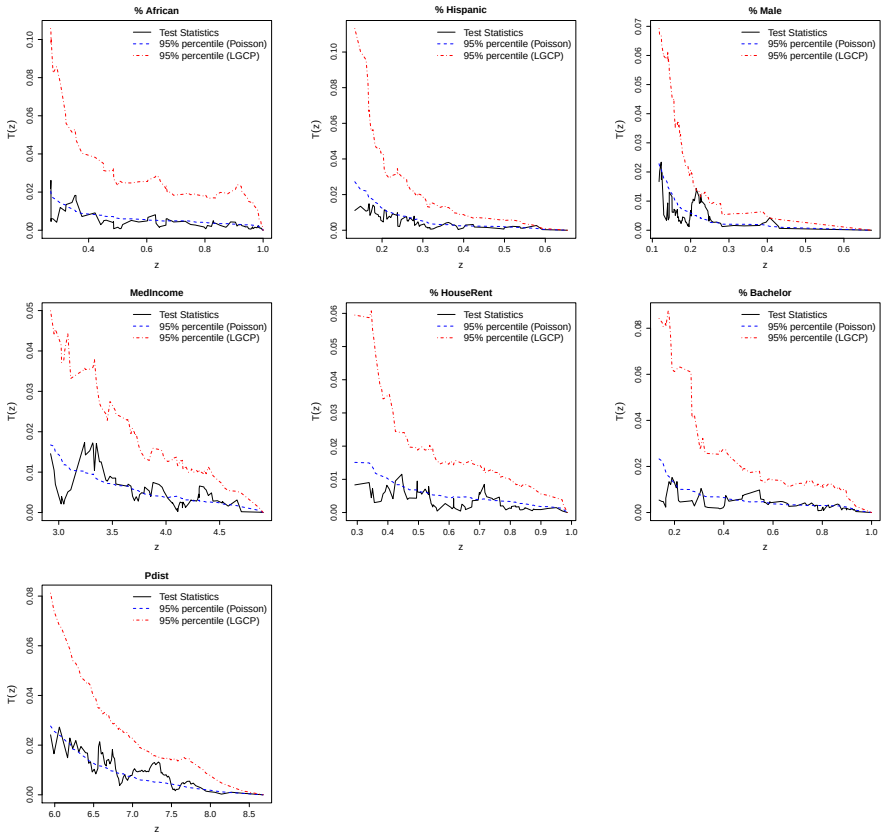


**Fig. A.11:** Estimated $T_j(z)$'s in (A.30) for three covariates together with 95% percentiles computed from the Poisson model and the LGCP model.

# References

Baddeley, A., Jammalamadaka, A., and Nair, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society: Series C*, 63(5):673–694.

Biscio, C. A. N. and Waagepetersen, R. P. (2019). A general central limit theorem and subsampling variance estimator for $\alpha$-mixing point processes. *Scandinavian Journal of Statistics*. Appeared online, 1-23.

Braga, A. A., Papachristos, A. V., and Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly*, 13:633–663.

Buerger, P. J., Cohn, E. G., and Petrosino, A. J. (1995). Defining the hot spots

of crime: Operationalizing theoretical concepts for field research. *Crime and Place*, pages 237–257.

Choiruddin, A., Cuevas-Pacheco, F., Couerjolly, J.-F., and Waagepetersen, R. (2019). Regularized estimation for highly multivariate log gaussian cox processes. *Statistics and Computing*, 30:649–662.

Cohen, J., Gorr, W. L., and Olligschlaeger, A. M. (2007). Leading Indicators and Spatial Interactions: A Crime-Forecasting Model for Proactive Police Deployment. *Geographical Analysis*, 29:105–127.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220.

Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity function. *Biometrika*, 105:455–462.

Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory*, 2(3):305–330.

Diggle, P., Zheng, P., and Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society: Series C*, 54(3):645–658.

Diggle, P. J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, 157:433–440.

Dong, J. and Yu, Q. (2019a). Joint distribution and marginal distribution methods for checking assumptions of generalized linear model. *Communications in Statistics-Theory and Methods*, 0(0):1–21.

Dong, J. and Yu, Q. (2019b). Marginal distribution test for checking proportional hazards model assumption. *Journal of Statistical Planning and Inference*, 201:58–70.

Gorr, W. L. and Lee, Y. (2015). Early warning system for temporary crime hot spots. *Journal of Quantitative Crimonology*, 31:25–47.

Guan, Y., Waagepetersen, R., and Beale, C. (2008). Second-order analysis of inhomogeneous spatial point processes with proportional intensity functions. *Journal of the American Statistical Association*, 103:769–777.

Haberman, C. (2017). Overlapping hot spots?: Examination of the spatial heterogeneity of hot spots of different crime types. *Criminology and Public Policy*, 16:633–660.

Jalilian, A., Guan, Y., Mateu, J., and Waagepetersen, R. (2015). Multivariate product-shot-noise Cox point process models. *Biometrics*, 71:1022–1033.

Rajala, T., Murrell, D. J., and Olhede, S. C. (2018). Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection. *Journal of the Royal Statistical Society: Series C*, 67:1237–1273.

Ratcliffe, J. (2002). Burglary reduction and the myth of displacement. *Trends and Issues in Crime and Criminal Justice*, 232:1–6.

Ratcliffe, J. (2004). Crime mapping and the training needs of law enforcement. *European Journal on Criminal Policy and Research*, 10:65–83.

Reinhart, A. and Greenhouse, J. (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C*, 67:1305–1329.

Waagepetersen, R. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:685–702.

Waagepetersen, R., Guan, Y., Jalilian, A., and Mateu, J. (2016). Analysis of multispecies point patterns by using multivariate log-gaussian cox processes. *Journal of the Royal Statistical Society. Series C.*, 65:77–96.

Weisburd, D., Maher, L., Sherman, L., Buerger, M., Cohn, E., and Petrosino, A. (1993). Contrasting crime general and crime specific theory: The case of hot spots of crime. *Advances in Criminological Theory*, 4:45–70.

Wilson, J. Q. and Kelling, G. L. (1982). Broken Windows: The Police and Neighborhood Safety. *Atlantic Monthly*, 249:29–38.

Xu, G., Waagepetersen, R., and Guan, Y. (2019). Stochastic quasi-likelihood for case-control point pattern data. *Journal of the American Statistical Association*, 114:631–644.

Zimmerman, D., Sun, P., and Fang, X. (2012). Likelihood-based estimation of spatial intensity and variation in disease risk from locations observed with error. *Statistics and Its Interface*, 5:207–219.

# Paper B

## Second order semi-parametric inference for multivariate log Gaussian Cox processes

Kristian Bjørn Hessellund, Ganggang Xu, Yongtao Guan and Rasmus Waagepetersen

The paper is submitted to
*Journal of the Royal Statistical Society: Series C (Applied Statistics)*

*The layout has been revised.*

# Abstract

*This paper introduces a new approach to inferring the second order properties of a multivariate log Gaussian Cox process (LGCP) with a complex intensity function. We assume a semi-parametric model for the multivariate intensity function containing an unspecified complex factor common to all types of points. Given this model we exploit the availability of several types of points to construct a second-order conditional composite likelihood to infer the pair correlation and cross pair correlation functions of the LGCP. Crucially this likelihood does not depend on the unspecified part of the intensity function. We also introduce a cross validation method for model selection and an algorithm for regularized inference that can be used to obtain sparse models for cross pair correlation functions. The methodology is applied to simulated data as well as data examples from microscopy and criminology. This shows how the new approach outperforms existing alternatives where the intensity functions are estimated non-parametrically.*


# 1 Introduction

A multivariate or multi-type point pattern is a marked point pattern where the marks belong to a finite set corresponding to different types of points. Equivalently, a multivariate point pattern can be viewed as a finite collection of ordinary point patterns, where each of these point patterns consists of points of a specific type. In this paper we consider point pattern data from biology and criminology. In the former case the point pattern represents locations of different types of cells in a tumor and in the latter case crime scenes of different types of crimes. An obvious key point of interest is then to study possible associations between the points of different types.

If consistent estimates of the intensity functions are available, and under assumptions of second-order intensity reweighted stationarity (Baddeley et al., 2000) or intensity-reweighted moment stationarity (van Lieshout, 2011), an immediate approach is to compute non-parametric cross summary statistics such as cross $K$, cross pair correlation, or cross $J$ functions (Møller and Waagepetersen, 2003; Baddeley et al., 2014; Cronie and van Lieshout, 2016). Parametric estimation of cross associations is also possible, see e.g. Jalilian et al. (2015); Waagepetersen et al. (2016); Choiruddin et al. (2019) who used parametric models of intensity and pair correlation functions, or Rajala et al. (2018) who specified a full model in terms of a multivariate Markov point process.

In some cases it is not straightforward to estimate the intensity function. For the cells data considered in this paper, the intensities of each type appear to be very heterogeneous, possibly varying within regions corresponding to different types of tissue. However, it is not straightforward to delienate these

regions. For the crime data considered, the intensity functions depend in a complex manner on the urban structures and the population density.

In case of bivariate case-control processes, Diggle et al. (2007) suggested a semi-parametric model where complex features of the case and control intensity functions were captured by a common non-parametric factor. This factor was estimated non-parametrically from the control point process and next used in a semi-parametric estimate of the intensity function for the case process. Finally this estimate was plugged into an estimate of the $K$-function for the case process. Using the control process to estimate the non-parametric part of the case intensity function mitigates the problem of confounding of possible clustering in the case process with variations in the case intensity function. However, sensitivity to the choice of bandwidth for the non-parametric estimation remains. Also the case and control processes were assumed to be independent whereby the cross pair correlation function is restricted to be one. Diggle et al. (2007) assumed the control process to be Poisson and Henrys and Brown (2009) relaxed this assumption by allowing both case and control processes to be clustered. They however retained independence between the two processes. Guan et al. (2008) used the same framework as Diggle et al. (2007) but used a second-order conditional composite likelihood to fit a parametric model to the case pair correlation function. The composite likelihood notably did not depend on the non-parametric part of the case intensity function and hence avoided choosing a bandwidth for non-parametric estimation.

In the context of multivariate point processes, Hessellund et al. (2019) used a semi-parametric model for the multivariate intensity function assuming a multiplicative structure where for each type of points, the intensity function is a product of a common background intensity and a log-linear factor modeling effects of covariates. Hence focus is on estimating differences between the intensity functions (for different types of points) that can be explained in terms of the covariates. In the bivariate case this model coincides with the one used in Diggle et al. (2007). However, Hessellund et al. (2019) did not impose any restrictive assumptions regarding the correlations within each type of points or between different types of points. While the main focus in Hessellund et al. (2019) was inference for the intensity function, they also obtained non-parametric estimates of ratios of cross pair correlation functions. They were, however, not able to obtain estimates of the individual cross pair correlation functions.

Our objective in this paper is to infer the full within and between correlation structure of a multivariate point process. To do so we adopt the parametric log Gaussian Cox process (LGCP) model for the correlation structure proposed in Waagepetersen et al. (2016) and further considered in Choiruddin et al. (2019). This model is flexible and has a very natural interpretation in terms of latent structures. However, to deal with complex intensity func-

tions, we replace the parametric model for the intensity function used in Waagepetersen et al. (2016) with the semi-parametric model for the intensity function from Hessellund et al. (2019). In this way we combine the strengths of two modelling approaches.

The presence of a non-parametric factor in the intensity function means that ingenuity is needed for fitting the parametric part of the model. We generalize the approach for the bivariate case in Guan et al. (2008) and obtain a second order conditional composite likelihood function which only depends on the parametric parts of the model and hence does not require knowledge of the non-parametric component. Compared with Guan et al. (2008) we consider an arbitrary number of point processes and do not assume that any of the point processes are Poisson nor that any two point processes are uncorrelated.

Some key questions we want to address for a particular data set are whether some point processes are uncorrelated and if not, whether they are negatively or positively correlated. We address these questions by a model selection approach where the models considered represent different types of correlation structures. Absence of correlation between point processes requires that certain parameters must be zero. To enable selection of models with parameters set to zero we combine our semi-parametric composite likelihood with a Lasso penalization (Tibshirani, 1996) which precisely facilitates that some parameters can be estimated to be exactly zero. A similar approach was considered by Choiruddin et al. (2019) in the context of least squares estimation for a multivariate LGCP with a full parametric model for the multivariate intensity function.

The rest of the paper is organized as follows. Section 2 gives a brief overview of multivariate point processes with focus on the intensity functions and cross intensity functions. Next the semi-parametric model for the intensity function and the multivariate LGCP model is described. Section 3 introduces the second order conditional composite likelihood function, an optimization algorithm based on the proximal Newton method, and a cross validation method for model selection. Section 4 contains simulation studies and Section 5 applies our methodology to cells and crimes data sets. Some concluding remarks are given in Section 6.

# 2 Semi-parametric modelling of a multivariate point process

## 2.1 Background on intensity functions

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a multivariate spatial point process, where $X_i$ is a spatial point process on $\mathbb{R}^d$ representing points of type $i$, $i = 1, \ldots, p$. Each

$X_i$ is hence a random subset of $\mathbb{R}^d$ such that the cardinality of $X_i \cap B$ is finite almost surely for any bounded $B \subset \mathbb{R}^d$. In practice we observe $\boldsymbol{X}$ in a spatial window $W$, where the window $W \subset \mathbb{R}^d$ is bounded with area $|W| > 0$. We will assume there exist for each $i, j = 1, \ldots, p$, non-negative functions $\rho_i(\cdot)$ and $\rho_{ij}(\cdot)$ so that the so-called Campbell's formulae:

$$\mathrm{E} \sum_{\mathbf{u} \in X_i} h_1(\mathbf{u}) = \int h_1(\mathbf{u}) \rho_i(\mathbf{u}) \mathrm{d}\mathbf{u} \tag{B.1}$$

$$\mathrm{E} \sum_{\mathbf{u} \in X_i, \mathbf{v} \in X_j}^{\neq} h_2(\mathbf{u}, \mathbf{v}) = \int h_2(\mathbf{u}, \mathbf{v}) \rho_{ij}(\mathbf{u}, \mathbf{v}) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v}, \tag{B.2}$$

hold for any non-negative functions $h_1(\cdot)$ and $h_2(\cdot, \cdot)$, where $\sum^{\neq}$ means that the sum is over pairwise distinct pairs $(\mathbf{u}, \mathbf{v})$. The function $\rho_i(\cdot)$ is called the intensity function of $X_i$. If $i = j$, then $\rho_{ii}(\cdot)$ is called the second order intensity function of $X_i$, while if $i \neq j$, $\rho_{ij}(\cdot)$ is called the cross intensity function between $X_i$ and $X_j$. The normalized cross intensity function, called cross pair correlation function (cross PCF), is denoted by $g_{ij}(\cdot)$ and defined by: $\rho_i(\mathbf{u}) \rho_j(\mathbf{v}) g_{ij}(\mathbf{u}, \mathbf{v}) = \rho_{ij}(\mathbf{u}, \mathbf{v})$. If $i = j$ we just call $g_{ii}(\cdot)$ the pair correlation function (PCF) for $X_i$. If $X_i$ and $X_j$ are independent, then $g_{ij}(\mathbf{u}, \mathbf{v}) = 1$ while $g_{ij}(\mathbf{u}, \mathbf{v}) > 1$ ($< 1$) is indicative of positive (negative) association between $X_i$ and $X_j$ (or between points in $X_i$ in the case $i = j$). Hence the cross PCFs provide useful insight regarding the dependence within and between the point processes. We assume that $X$ is second order cross intensity reweighted stationary and isotropic, i.e., with an abuse of notation, $g_{ij}(\mathbf{u}, \mathbf{v}) = g_{ij}(r)$, $i, j = 1, \ldots, p$, where $r = ||\mathbf{u} - \mathbf{v}||$.

## 2.2 Semi-parametric regression model for the intensity

It may sometimes be difficult to specify a simple parametric model for the intensity functions. One may then resort to non-parametric estimation of the intensity functions but the results depend heavily on the choice of smoothing bandwidth where different data driven methods may result in very different results, see e.g. simulation studies in Cronie and van Lieshout (2018) and Shaw et al. (2020). We instead consider a semi-parametric model where a background intensity function $\rho_0(\cdot)$ captures complex variation in the intensity function that is common to all the point processes $X_1, \ldots, X_p$. For the cells data considered in Section 5.1, $\rho_0(\cdot)$ may capture variations in tissue composition that influence occurrence of different types of cells while for the crime data in Section 5.2, $\rho_0(\cdot)$ captures variation in population density and dependence of the intensities on the urban structure. More specifically, following Hessellund et al. (2019), we consider the multiplicative model:

$$\rho_i(\mathbf{u}; \boldsymbol{\gamma}_i) = \rho_0(\mathbf{u}) \exp(\boldsymbol{\gamma}_i^{\mathsf{T}} \mathbf{z}(\mathbf{u})) \tag{B.3}$$

for the intensity of $X_i$ where $\mathbf{z}(\mathbf{u})$ denotes a vector of spatial covariates at location $\mathbf{u}$ and $\boldsymbol{\gamma}_i$ is a regression parameter vector.

Since $\rho_0(\cdot)$ is left completely unspecified, the parameters $\boldsymbol{\gamma}_i$ are not identifiable: replacing the $l$th entry $\gamma_{il}$ in $\boldsymbol{\gamma}_i$ by $\gamma_{il} - K$ for $i = 1, \ldots, p$ while replacing $\rho_0(\cdot)$ by $\rho_0(\cdot)\exp(Kz_l(u))$ does not change the model when $\rho_0(\cdot)$ is unspecified. Hessellund et al. (2019) proposed methodology for estimating contrasts $\boldsymbol{\beta}_i = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_p$ where $\boldsymbol{\beta}_p = 0$. Alternatively, one could impose sum-to-zero constraints $\sum_l \boldsymbol{\beta}_{il} = 0$ on the $\boldsymbol{\beta}_i$.

Given the semi-parametric model for the intensity functions and its associated estimation procedure we specify in the next section a model for the correlation structure of the multivariate point process.

## 2.3 Multivariate log Gaussian Cox model

Following the setup in Waagepetersen et al. (2016), we assume that $X_i$ for $i = 1, \ldots, p$, is a Cox process with random intensity given by:

$$\Lambda_i(\mathbf{u}) = \rho_0(\mathbf{u})\exp(\boldsymbol{\gamma}_i^{\mathsf{T}}\mathbf{z}(\mathbf{u}))\exp\left(\mu_i + \sum_{k=1}^{q}\alpha_{ik}Y_k(\mathbf{u}) + \sigma_i U_i(\mathbf{u})\right), \tag{B.4}$$

where the $Y_k$ and $U_i$ are independent zero mean unit variance Gaussian random fields and $\mu_i = -\sum_{k=1}^{q}\alpha_{ik}^2/2 - \sigma_i^2/2$. We interpret the $Y_k$ as latent random factors that influence all types of points. Hence the different types of points may be correlated due to possible dependence on the $Y_k$. Moreover, each $U_i$ is a type-specific random factor that only affects the $i$th type of points. Hence $U_i$ models random clustering within each $X_i$.

Consider for a moment the ideal situation where the $Y_k$ are observed (non-random). Following the same considerations as for the $\boldsymbol{\gamma}_i$ in the previous section we should then impose restrictions $\alpha_{pl} = 0$ or $\sum_{i=1}^{p}\alpha_{il} = 0$, $l = 1, \ldots, q$, in order to ensure identifiability. In case of unobserved $Y_k$ and hence less information, the need for a constraint is not less pertinent. In the following we impose the sum-to-zero constraint $\sum_{i=1}^{p}\alpha_{il} = 0$, $l = 1, \ldots, q$ which treats all $X_i$ symmetrically.

The intensity function of $X_i$ is $\rho_i(\mathbf{u}) = \mathrm{E}[\Lambda_i(\mathbf{u})] = \rho_0(\mathbf{u})\exp(\boldsymbol{\gamma}_i^{\mathsf{T}}\mathbf{z}(\mathbf{u}))$, which follows from the moment generating function of a Gaussian random variable. Similarly,

$$\begin{aligned}\rho_{ij}(\mathbf{u},\mathbf{v}) = \mathrm{E}[\Lambda_i(\mathbf{u})\Lambda_j(\mathbf{v})] =& \rho_0(\mathbf{u})\rho_0(\mathbf{v})\exp(\boldsymbol{\gamma}_i^{\mathsf{T}}\mathbf{z}(\mathbf{u}))\exp(\boldsymbol{\gamma}_j^{\mathsf{T}}\mathbf{z}(\mathbf{u})) \\ &\times \exp\left(\sum_{k=1}^{q}\alpha_{ik}\alpha_{jk}c_{Y_k}(\mathbf{u},\mathbf{v}) + 1[i=j]\sigma_i^2 c_{U_i}(\mathbf{u},\mathbf{v})\right),\end{aligned}$$

where $c_{Y_k}(\mathbf{u},\mathbf{v}) = \mathrm{Corr}[Y_k(\mathbf{u}), Y_k(\mathbf{v})]$ and $c_{U_i}(\mathbf{u},\mathbf{v}) = \mathrm{Corr}[U_i(\mathbf{u}), U_i(\mathbf{v})]$.

For $c_{Y_k}(\cdot)$ and $c_{U_i}(\cdot)$ we use exponential correlation functions, i.e. $c_{Y_k}(\mathbf{u},\mathbf{v}) = \exp(-||\mathbf{u}-\mathbf{v}||/\xi_k)$ and $c_{U_i}(\mathbf{u},\mathbf{v}) = \exp(-||\mathbf{u}-\mathbf{v}||/\varphi_i)$ with correlation scale

parameters $\xi_k$ and $\varphi_i$. Other parametric correlation models might of course be used instead, depending on the application. Denote by $\boldsymbol{\theta}$ the concatenation of $\boldsymbol{\alpha}_{\cdot k} = (\alpha_{1k}, \ldots, \alpha_{pk})^\mathsf{T}$, $k = 1, \ldots, q$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^\mathsf{T}$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_p^2)^\mathsf{T}$, and $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)^\mathsf{T}$. The cross PCF between $X_i$ and $X_j$ is then given by the parametric model:

$$g_{ij}(r; \boldsymbol{\theta}) = \exp\Big( \sum_{k=1}^{q} \alpha_{ik}\alpha_{jk} \exp(-r/\xi_k) + 1[i = j]\sigma_i^2 \exp(-r/\varphi_i) \Big). \tag{B.5}$$

If $\sum_{k=1}^{q} \alpha_{ik}\alpha_{jk}\exp(-r/\xi_k)$ is greater (smaller) than 0, this implies positive (negative) spatial correlation between points from $X_i$ and $X_j$ at the lag $r$. If for example $\alpha_{ik}\alpha_{jk} = 0$ for all $k = 1, \ldots, q$, then $X_i$ and $X_j$ are independent.

The number $q$ of latent common fields controls the complexity of the model and will be chosen according to a cross validation criterion detailed in Section 3.3. In Waagepetersen et al. (2016) and Choiruddin et al. (2019), estimation of $\boldsymbol{\theta}$ for a chosen $q$ was based on a least squares criterion where non-parametric estimates of the pair correlation function acted as 'dependent' variables. These non-parametric estimates were based on fully specified regression models for the log intensity functions that are not available in our current setting. Section 3 therefore introduces a second order conditional composite likelihood function for estimation of $\boldsymbol{\theta}$ that does not require knowledge of $\rho_0(\cdot)$.

# 3 Second order conditional composite likelihood

We assume initially that the $\boldsymbol{\beta}_i$ are known and thus suppress dependence on these in the notation. The idea is to condition on the union of all points regardless of type and for each $\mathbf{u} \neq \mathbf{v} \in \cup_{i=1}^{p} X_i$ consider the conditional probability (see Section A in the supplementary material) that $\mathbf{u}$ is of type $i$ and $\mathbf{v}$ is of type $j$:

$$\mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) = \frac{\rho_{ij}(\mathbf{u}, \mathbf{v})}{\sum_{k,l} \rho_{kl}(\mathbf{u}, \mathbf{v})} = \frac{f_i(\mathbf{u}) f_j(\mathbf{v}) g_{ij}(r; \boldsymbol{\theta}_{ij})}{\sum_{k,l} f_k(\mathbf{u}) f_l(\mathbf{v}) g_{kl}(r; \boldsymbol{\theta}_{kl})}, \tag{B.6}$$

where $f_i(\mathbf{u}) = \exp(\boldsymbol{\beta}_i^\mathsf{T} \mathbf{z}(\mathbf{u}))$, $i = 1, \ldots, p$. Note that $\rho_0(\mathbf{u})\rho_0(\mathbf{v})$ cancels out in (B.6) so that the probabilities do not depend on the unspecified $\rho_0(\cdot)$. We then estimate $\boldsymbol{\theta}$ by maximizing the second order conditional composite likelihood function given by:

$$L(\boldsymbol{\theta}) = \prod_{i,j} \prod_{\mathbf{u} \in X_i \cap W, \mathbf{v} \in X_j \cap W}^{\neq} 1_R[\mathbf{u}, \mathbf{v}] \mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}), \tag{B.7}$$

where $1_R[\mathbf{u}, \mathbf{v}] = 1[||\mathbf{u} - \mathbf{v}|| \leq R]$ and $R > 0$ is a user-specified tuning parameter. Specifying an $R < \infty$ is useful for reducing computing time and

can also improve the statistical efficiency by omitting pairs of points $\mathbf{u}$ and $\mathbf{v}$ that are distant from each other and hence do not provide much information on the correlation structure. As a rule of thumb, $R$ should be chosen so that $g_{ij}(r;\boldsymbol{\theta}) \approx 1$ for $||\mathbf{u}-\mathbf{v}|| > R$. Methods for choosing $R$ are discussed in Lavancier et al. (2019).

The cross PCFs (B.5) and hence the second order conditional composite likelihood function are invariant to simultaneous interchange of columns $\boldsymbol{\alpha}_{\cdot k} = (\alpha_{ik})_k$ and corresponding correlation scale parameters $\xi_k$ as well as to multiplication by $-1$ of $\boldsymbol{\alpha}_{\cdot k}$. This lack of identifiability is not of much concern since we are not interested in the individual $\alpha_{ij}$'s but rather the resulting correlation structure which is invariant to the aforementioned transformations.

Following the idea of two-step estimation in Waagepetersen and Guan (2009), we replace the parameters $\boldsymbol{\beta}_i$ by consistent estimates $\hat{\boldsymbol{\beta}}_i$ obtained using the method proposed in Hessellund et al. (2019) that does not require knowledge of $\boldsymbol{\theta}$.

## 3.1 Optimization

We denote by $l_-(\boldsymbol{\theta})$ the negation of the log of (B.7) and turn the estimation of $\boldsymbol{\theta}$ into a minimization problem. In order to minimize $l_-(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we consider a cyclic block descent algorithm. The strategy is to update $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$, $\sigma^2$ and $\boldsymbol{\varphi}$ in turn until a convergence criterion is met. In the following we will, with a convenient abuse of notation, use $\boldsymbol{\alpha}$ to denote both the matrix $[\alpha_{ij}]_{ij}$ and the vectorized version where the matrix is laid out column-wise $(\boldsymbol{\alpha}_{\cdot 1}^\mathsf{T}, \ldots, \boldsymbol{\alpha}_{\cdot q}^\mathsf{T})^\mathsf{T}$. It will be clear from the context which interpretation of $\boldsymbol{\alpha}$ is relevant. Denote by $\boldsymbol{\theta}^{(n)} = ((\boldsymbol{\alpha}^{(n)})^\mathsf{T}, (\boldsymbol{\xi}^{(n)})^\mathsf{T}, (\sigma^{2(n)})^\mathsf{T}, (\boldsymbol{\varphi}^{(n)})^\mathsf{T})^\mathsf{T}$ the current value of $\boldsymbol{\theta}$. We update each parameter using a quasi Newton-Raphson iteration with additional line search. This is equivalent to minimizing a certain least squares problem. We give the details of this since this is also needed for solving a regularized version of our estimation problem, see Section 3.2.

We denote by $\tilde{\boldsymbol{\theta}}$ a temporary parameter vector that keeps track of the updates leading from $\boldsymbol{\theta}^{(n)}$ to $\boldsymbol{\theta}^{(n+1)}$ and initialize $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(n)}$. Denote by $\tilde{\boldsymbol{\tau}} \in \{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\xi}}, \tilde{\sigma}^2, \tilde{\boldsymbol{\varphi}}\}$ the parameter vector to be updated and by $\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau})$ the vector obtained by replacing $\tilde{\boldsymbol{\tau}}$ in $\tilde{\boldsymbol{\theta}}$ by $\boldsymbol{\tau}$. Consider a quadratic approximation of $l_-(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))$ with respect to $\boldsymbol{\tau}$ around $\tilde{\boldsymbol{\theta}}$:

$$q(\boldsymbol{\tau}) = l_-(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})^\mathsf{T}\mathrm{e}(\tilde{\boldsymbol{\tau}}) + \frac{1}{2}(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})^\mathsf{T}\mathrm{H}(\tilde{\boldsymbol{\tau}})(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}}). \tag{B.8}$$

Here (omitting for convenience the arguments $\mathbf{u}, \mathbf{v}$)

$$\mathrm{e}(\boldsymbol{\tau}) = \nabla_{\boldsymbol{\tau}} l_-(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau})) = \sum_{i,j} \sum_{\substack{\mathbf{u}\in X_i\cap W \\ \mathbf{v}\in X_j\cap W}}^{\neq} 1_R \left( \frac{\sum_{k,l} \nabla_{\boldsymbol{\tau}} \rho_{kl}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))}{\sum_{k,l} \rho_{kl}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))} - \frac{\nabla_{\boldsymbol{\tau}} \rho_{ij}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))}{\rho_{ij}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))} \right)$$

is the gradient with respect to $\boldsymbol{\tau}$ and

$$\mathrm{H}(\boldsymbol{\tau}) = \mathrm{E}[\nabla^2_{\boldsymbol{\tau}} l_-(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))] = \int_{W^2} 1_R \mathrm{Cov}\left(Z(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))\right) \sum_{i,j} \rho_{ij}(\tilde{\boldsymbol{\theta}}(\boldsymbol{\tau})) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$$

is the expected Hessian with respect to $\boldsymbol{\tau}$, where $Z(\mathbf{u}, \mathbf{v}, \tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))$ denotes a random vector which takes values $\nabla_{\boldsymbol{\tau}} \log(\rho_{ij}(\mathbf{u}, \mathbf{v}; \tilde{\boldsymbol{\theta}}(\boldsymbol{\tau})))$ with probabilities $\mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \tilde{\boldsymbol{\theta}}(\boldsymbol{\tau}))$ (Lemma B.2 in the supplementary material). We estimate $\mathrm{H}(\tilde{\boldsymbol{\tau}})$ by

$$\hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}}) = \sum_{i,j} \sum_{\mathbf{u} \in X_i \cap W, \mathbf{v} \in X_j \cap W}^{\neq} 1_R[\mathbf{u}, \mathbf{v}] \mathrm{Cov}\left(Z(\mathbf{u}, \mathbf{v}, \tilde{\boldsymbol{\theta}})\right),$$

which is unbiased by (B.2). Since $\hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}})$ is a symmetric, positive semi-definite matrix, the eigendecomposition implies that $\hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}})^{1/2} = UD^{1/2}U^{\mathsf{T}}$, where $D$ is the diagonal matrix of the (all non-negative) eigen values of $\hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}})$ and $U$ is the matrix with the eigen vectors as columns. Assuming that all the eigen values are positive, following Section C in the supplementary material, the minimizer $\hat{\boldsymbol{\tau}}$ of (B.8) is a solution of a least squares problem:

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \left( \frac{1}{2} ||Y - X\boldsymbol{\tau}||^2 \right) = \left(X^{\mathsf{T}}X\right)^{-1} X^{\mathsf{T}}Y, \tag{B.9}$$

where $Y = \hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}})^{1/2} \left(-\hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}})^{-1}\mathrm{e}(\tilde{\boldsymbol{\tau}}) + \tilde{\boldsymbol{\tau}}\right)$ and $X = \hat{\mathrm{H}}(\tilde{\boldsymbol{\tau}})^{1/2}$. Introducing a line search, we update $\boldsymbol{\tau}^{(n+1)} = \tilde{\boldsymbol{\tau}} + t(\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\tau}})$, for some $t > 0$ and also update $\tilde{\boldsymbol{\theta}}$ by replacing $\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau}^{(n)}$ by $\boldsymbol{\tau}^{(n+1)}$. When all components of $\tilde{\boldsymbol{\theta}}$ have been updated we let $\boldsymbol{\theta}^{(n+1)} = \tilde{\boldsymbol{\theta}}$.

As mentioned in Section 2.3 we impose a sum to zero constraint on each $\boldsymbol{\alpha}_{\cdot k}$, i.e. $\sum_{i=1}^{p} \alpha_{ik} = 0$, $k = 1, \ldots, q$. The constraint is easily accommodated by the change of variable $\mathbf{B}\boldsymbol{\psi} = \boldsymbol{\alpha}$, where $\boldsymbol{\psi}$ is a $(p-1) \times q$ matrix and $\mathbf{B}^{\mathsf{T}} = [\mathrm{I}_{p-1}\ \textbf{-1}]$ is a $(p-1) \times p$ matrix, where $\mathrm{I}_{p-1}$ is the $(p-1) \times (p-1)$ identity matrix and $\textbf{-1} = [\text{-}1, \ldots, \text{-}1]^{\mathsf{T}} \in \mathbb{R}^p$. Under the sum to zero constraint, the relation between $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ is one-to-one. Thus in case of $\boldsymbol{\tau} = \boldsymbol{\alpha}$ we update the unconstrained parameter $\boldsymbol{\psi}$ using, by the chain rule, the gradient $\mathbf{B}^{\mathsf{T}}\mathrm{e}(\tilde{\boldsymbol{\theta}})$ and the Hessian $\mathbf{B}^{\mathsf{T}}\mathrm{H}(\tilde{\boldsymbol{\theta}})\mathbf{B}$ and finally let $\boldsymbol{\alpha}^{(n+1)} = \mathbf{B}\boldsymbol{\psi}^{(n+1)}$.

The cyclical block updating is iterated until relative function convergence,

$$\left| \left[ l_- \left( \boldsymbol{\theta}^{(n+1)} \right) - l_- \left( \boldsymbol{\theta}^{(n)} \right) \right] / l_- \left( \boldsymbol{\theta}^{(n)} \right) \right| < \varepsilon, \tag{B.10}$$

for some $\varepsilon > 0$ in which case we set $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(n+1)}$. Algorithm 1 gives a brief overview of the cyclical block descent algorithm.

**Algorithm 1** Cyclical block descent algorithm

1: Simulate initial parameters $\hat{\boldsymbol{\psi}}^{(0)}$, $\hat{\boldsymbol{\xi}}^{(0)}$, $\hat{\sigma}^{2(0)}$ and $\hat{\boldsymbol{\varphi}}^{(0)}$
2: $n := 0$
3: ***repeat***
4: $\quad \tilde{\boldsymbol{\psi}} := \boldsymbol{\psi}^{(n)}$, $\tilde{\boldsymbol{\xi}} := \boldsymbol{\xi}^{(n)}$, $\tilde{\sigma}^2 := \sigma^{2(n)}$ and $\tilde{\boldsymbol{\varphi}} := \boldsymbol{\varphi}^{(n)}$
5: $\quad$ update $\tilde{\boldsymbol{\psi}}$, $\tilde{\boldsymbol{\xi}}$, $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\varphi}}$ in turn using (B.9) combined with line search
6: $\quad \boldsymbol{\psi}^{(n+1)} := \tilde{\boldsymbol{\psi}}$, $\boldsymbol{\xi}^{(n+1)} := \tilde{\boldsymbol{\xi}}$, $\sigma^{2(n+1)} := \tilde{\sigma}^2$, $\boldsymbol{\varphi}^{(n+1)} := \tilde{\boldsymbol{\varphi}}$, and $\boldsymbol{\alpha}^{(n+1)} := B\tilde{\boldsymbol{\psi}}^{(n+1)}$
7: $\quad n := n + 1$
8: ***until*** relative convergence criterion (B.10)
9: ***return*** $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(n)}$

## 3.2 Optimization with lasso regularization

The overall model complexity is controlled by the number $q$ of latent fields. Nevertheless, for any $q$, more sparse submodels could be obtained by restricting some $\alpha_{ij}$, $i = 1, \ldots, p$, $j = 1, \ldots, q$ to zero. Of course if all entries in a column $\boldsymbol{\alpha}_{\cdot k}$ are restricted to zero this just corresponds to reducing $q$ by one. In order to look for sparse submodels for a given $q$ we extend the estimation approach by introducing a lasso regularization on $\boldsymbol{\alpha}$. We express the sum to zero constraint for $\boldsymbol{\alpha}$ by $\mathbf{C}\boldsymbol{\alpha} = \mathbf{0}$, where $\mathbf{C} = [\mathbf{D}_1 \cdots \mathbf{D}_q]$ is a $q \times pq$ matrix that consists of submatrices $\mathbf{D}_i$, $i = 1, \ldots q$, of dimension $q \times p$. Each submatrix $\mathbf{D}_i$ consists of ones on the $i$th row and zeros otherwise. Here $\boldsymbol{\alpha}$ should be interpreted as the vector obtained by concatenating the $\boldsymbol{\alpha}_{\cdot k}$, cf. Section 3.1. Note that the regularization is not relevant in the bivariate case $p = 2$ since in this case, by the sum to zero constraint, $\alpha_{1k} = 0$ implies $\alpha_{2k} = 0$ which just corresponds to reducing $q$ by 1.

The regularized object function becomes:

$$l_{-}(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{p} \sum_{j=1}^{q} |\alpha_{ij}|, \quad \mathbf{C}\boldsymbol{\alpha} = \mathbf{0}, \tag{B.11}$$

where $\lambda \sum_{i=1}^{p} \sum_{j=1}^{q} |\alpha_{ij}|$ is a lasso penalty that can lead to exact zero components in the estimate of $\boldsymbol{\alpha}$. We minimize this using a cyclical block descent algorithm which only differs from the one in Section 3.1 by the update $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left( \frac{1}{2} ||Y - X\boldsymbol{\alpha}||^2 + \lambda \sum_{i=1}^{p} \sum_{j=1}^{q} |\alpha_{ij}| \right)$ subject to $\mathbf{C}\boldsymbol{\alpha} = \mathbf{0}$. To compute $\hat{\boldsymbol{\alpha}}$ under the sum to zero constraint, we use the augmented Lagrangian algorithm suggested in Shi et al. (2016). Details are given in Section D in the supplementary material. In Section 3.3 we propose a cross validation procedure to choose $\lambda$.

### 3.3 Determination of $q$ and $\lambda$

We choose the values of $q$ and $\lambda$ according to a $K$-fold ($K \geq 2$) cross validation criterion constructed so that it targets selection of an appropriate cross correlation structure. Let for each $i, j$, $M_{ij}$ denote the set of pairs $(\mathbf{u}, \mathbf{v})$ with $\mathbf{u} \in X_i$, $\mathbf{v} \in X_j$, and $0 < \|\mathbf{u} - \mathbf{v}\| \leq R$. We randomly split $M_{ij}$ into $K$ equally sized subsets $M_{ij,1}, \ldots M_{ij,K}$. We then obtain for each $k = 1, \ldots, K$, a parameter estimate $\hat{\boldsymbol{\theta}}_k$ by maximizing the regularized conditional composite likelihood

$$l_k(\boldsymbol{\theta}) = \sum_{i,j} \sum_{(\mathbf{u},\mathbf{v}) \in M_{ij,-k}}^{\neq} \log \mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) + \lambda \sum_{i=1}^{p} \sum_{j=1}^{q} |\alpha_{ij}|, \quad \mathbf{C}\boldsymbol{\alpha} = \mathbf{0},$$

for the training data set consisting of the leave one fold out sets $M_{ij,-k} = \cup_{l \neq k} M_{ij,l}$. The $k$th cross validation score based on the validation sets $M_{ij,k}$, $i \neq j$, is then

$$CV_k(q, \lambda) = \sum_{i \neq j} \sum_{(\mathbf{u},\mathbf{v}) \in M_{ij,k}} \log \mathrm{p}_{ij}(\mathbf{u}, \mathbf{v}; \hat{\boldsymbol{\theta}}_k).$$

We here omit the $M_{ii,k}$ to focus the cross validation on the fit of the cross correlation structure. To reduce the sensitivity to Monte Carlo variation, one may compute cross validation scores $CV_{kl}(q, \lambda)$, $l = 1, \ldots, L$, based on $L$ independent $K$-fold random splits of the data and use the average $\overline{CV}(q, \lambda)$ of the $CV_{kl}(q, \lambda)$, $k = 1, \ldots, K$, $l = 1, \ldots, L$. According to standard practice in the statistical learning literature (Hastie et al., 2013) we use $K$ in the range of 5 to 10.

Consider the case $\lambda = 0$ which is relevant for example when $p = 2$. The most obvious choice of $q$ is the one that minimizes the cross validation score, $q_{\min} = \arg\min_q \overline{CV}(q, 0)$. We denote this the minimum (MIN) rule. However, due to sensitivity to Monte Carlo error, a so-called one standard error rule has been proposed (Hastie et al., 2013) that promotes more sparse solutions. Let $\mathrm{SD}(q, 0)$ denote the standard deviation of a cross validation score $CV_{kl}(q, 0)$ obtained from a single validation set. In the current framework, the one standard error (1-SE) rule selects the smallest $q$ ($q_{\text{1-SE}}$) for which $\overline{CV}(q, 0) \leq \overline{CV}(q_{\min}, 0) + \mathrm{SE}(q_{\min}, 0)$, where $\mathrm{SE}(q, 0) = \mathrm{SD}(q_{\min}, 0)/\sqrt{KL}$ is the standard error of $\overline{CV}(q_{\min}, 0)$.

For joint selection of $(q, \lambda)$ the immediate choice would be the minimizer of $\overline{CV}(q, \lambda)$. However, computing $\overline{CV}(q, \lambda)$ over a two-dimensional grid of $q$ and $\lambda$ values is very time consuming. Instead we use a two-step approach where we first determine $q_{\min}$ as in the previous paragraph and next choose the $\lambda$ that minimizes $\overline{CV}(q_{\min}, \lambda)$ over values of $\lambda$. Thus the initial selection of $q$ determines the overall model complexity while the subsequent possible selection of a $\lambda > 0$ may introduce additional sparsity given $q_{\min}$.

### 3.4 Model assessment

Assuming the model (B.3) for the intensity functions, Hessellund et al. (2019) obtained a consistent non-parametric estimate of any ratio $g_{ij}(r)/g_{lk}(r)$ of cross PCFs, $r > 0$, $i, j, l, k = 1, \ldots, p$. Similarly we can obtain semi-parametric estimates of these ratios based on our semi-parametric estimates of the cross PCFs. If the assumed multivariate LGCP is valid, the non-parametric and the semi-parametric estimates of cross PCF ratios should not differ much. In our data examples in Section 5 we informally assess the models by visual comparison of the two types of estimates. We also conduct a so-called global envelope goodness-of-fit test (Myllymäki et al., 2016) based on the difference between the two types of estimates over spatial lags $r \in [0, R]$. This requires simulation under a null model. For this we use the fitted multivariate LGCP where we replace the unknown background intensity $\rho_0$ by a non-parametric estimate introduced in Hessellund et al. (2019), see also Section E in the supplementary material.

## 4 Simulation study

To study the performance of the package consisting of the second order conditional composite likelihood, the optimization algorithm in Section 3.1 with additional lasso regularization in Section 3.2, and the cross validation procedure, we conduct simulation studies based on two different settings for a five-variate LGCP $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^\top$ on $W = [0,1]^2$. In both cases we simulate one covariate $Z(\cdot)$ and a background intensity $\rho_0(\cdot) = 400\exp(0.5V(\cdot) - 0.5^2/2)$, where $Z$ and $V$ are zero mean unit variance Gaussian random fields with exponential and Gaussian correlation functions, i.e. $\mathrm{Corr}(Z(\mathbf{u}), Z(\mathbf{v})) = \exp(-||\mathbf{u}-\mathbf{v}||/0.05)$ and $\mathrm{Corr}(V(\mathbf{u}), V(\mathbf{v})) = \exp(-(||\mathbf{u}-\mathbf{v}||/0.2)^2)$. The realizations of $Z$ and $\rho_0$ are shown in Figure B.1.

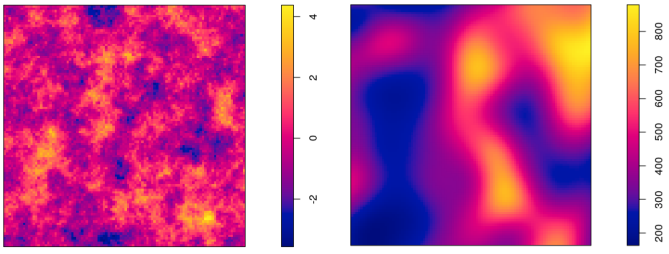


**Fig. B.1:** Left: simulated covariate $Z$. Right: simulated $\rho_0$.

Table B.1 shows the values used for all parameters except $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ as well as the expected number of points $N$ for each point process. Regarding $q$, $\boldsymbol{\alpha}$

**Table B.1:** Simulation settings for $\boldsymbol{X}$ in each setup $q = 0, 2$ (excluding $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$).

| $\boldsymbol{X}$ | $\gamma_1$ | $\gamma_2$ | $\sigma$ | $\varphi$ | $N$ | $\boldsymbol{X}$ | $\gamma_1$ | $\gamma_2$ | $\sigma$ | $\varphi$ | $N$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 0.1 | -0.1 | 0.71 | 0.02 | 550 | $X_4$ | 0.4 | 0.1 | 0.71 | 0.03 | 750 |
| $X_2$ | 0.2 | -0.2 | 0.71 | 0.02 | 619 | $X_5$ | 0.5 | 0.2 | 0.71 | 0.04 | 830 |
| $X_3$ | 0.3 | 0 | 0.71 | 0.03 | 677 | | | | | | |

and $\boldsymbol{\xi}$ we take $q = 0$ for the first setting resulting in a case with independent components $X_1, \ldots, X_5$. In the second setting we let $q = 2$ and choose $\boldsymbol{\alpha}$ as specified in the table left in Figure B.2. We moreover let $\xi_1 = 0.02$ and $\xi_2 = 0.03$. The resulting PCFs and cross PCFs are shown in the middle and right plots in Figure B.2. In the case $q = 2$ we have a positive spatial

| $\boldsymbol{X}$ | $\boldsymbol{\alpha}_1$ | $\boldsymbol{\alpha}_2$ |
|---|---|---|
| $X_1$ | 0.5 | -1 |
| $X_2$ | 0.5 | 0 |
| $X_3$ | -1 | 0 |
| $X_4$ | 0 | 0.5 |
| $X_5$ | 0 | 0.5 |

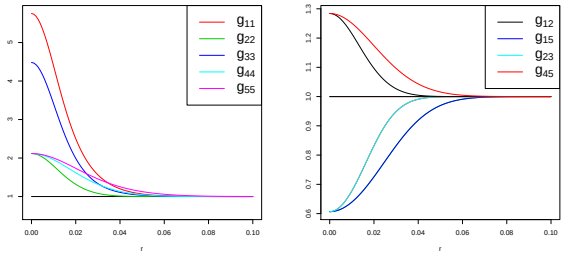


**Fig. B.2:** Left: $\boldsymbol{\alpha}$. Middle: true PCFs. Right: true cross PCFs (note $g_{13} = g_{23}$ and $g_{14} = g_{15}$).

dependence between $X_1$ and $X_2$ and between $X_4$ and $X_5$, while there is a negative spatial dependence between $X_3$ and $(X_1, X_2)$ and between $X_1$ and $(X_4, X_5)$.

For our second order conditional composite likelihood we specified exponential correlation models for the fields $Y_k$, $k = 1, \ldots, q$ and $U_i$, $i = 1, \ldots, 5$. In practice it is rarely the case that the true correlation models correspond exactly to the specified ones. To reflect this we simulate the $Y_k$ and $U_i$ using Gaussian correlation functions, i.e. $\text{Corr}(Y_k(\mathbf{u}), Y_k(\mathbf{v})) = \exp(-(||\mathbf{u} - \mathbf{v}||/\xi_k)^2)$ and $\text{Corr}(U_i(\mathbf{u}), U_i(\mathbf{v})) = \exp(-(||\mathbf{u} - \mathbf{v}||/\varphi_i)^2)$. Hence the model applied is misspecified for the simulated data. For each setting we generate 100 simulated realizations of $\mathbf{X}$.

In both settings we select $q$ among the values $0, 1, \ldots, 5$, and next, for the chosen $q$, $\lambda$ among the values $10, 8, 6, 5, 4, 3, 2, 1, 0.5, 0.25, 0$, using cross validation with $K = 5$ and $L = 10$, and we consider results using both the MIN and the 1-SE approach for the selection of $q$. We also consider the case where $q = 7$ is fixed in order to assess the effect of regularization in an over-parametrized setting. For the second order conditional composite likelihood we only consider distinct pairs of points $\mathbf{u}, \mathbf{v}$ with $\|\mathbf{u} - \mathbf{v}\| \leq R = 0.1$.

All first order parameters are estimated using the approach in Hessellund et al. (2019). The initial parameters for $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$, $\sigma^2$ and $\boldsymbol{\varphi}$ are simulated as $\alpha_{ik} \sim$ Unif(-0.25,0.25), $\xi_k, \varphi_i \sim$ Unif(0.01,0.04) and $\sigma_i^2 \sim$ Unif(0.4,0.6). The parameter $\varepsilon$ in the relative function convergence criterion (B.10) is set to $10^{-5}$ and the convergence parameters $\tilde{\varepsilon}$ and $\tilde{\tilde{\varepsilon}}$ for the regularized optimization ((B.15) in the supplementary material) are set to $10^{-10}$.

We measure the performance for each selected model using mean integrated squared error (MISE) aggregated over respectively all PCFs and all cross PCFs, i.e.

$$\text{MISE}_{\text{between}}(\hat{\boldsymbol{\theta}}) = \sum_{i<j} \text{E}\left[\int_{0.01}^{0.1} \left(g_{ij}(r;\hat{\boldsymbol{\theta}}_{ij}) - g_{ij}(r;\boldsymbol{\theta}_{ij})\right)^2 \text{d}r\right], \tag{B.12}$$

while $\text{MISE}_{\text{within}}(\hat{\boldsymbol{\theta}})$ and $\text{MISE}_{\text{total}}(\hat{\boldsymbol{\theta}})$ are defined in the same way but with sum over $i = j$ or $i \leq j$.

We compare the performance of the proposed method with two non-parametric approaches. For the first approach, referred to as 'simple', we estimate the intensity functions non-parametrically using the `spatstat` (Baddeley et al., 2015) procedure `density.ppp` with bandwidths selected using the method introduced in Cronie and van Lieshout (2018). Next the PCFs and cross PCFs are estimated using the `spatstat` procedures `pcfinhom` and `crosspcfinhom` with the intensity functions replaced by the non-parametric estimates. For the PCF and cross PCF estimation we manually specify reasonable bandwidths based on the knowledge of the true PCFs and cross PCFs (note that this is in favor of the non-parametric approach). The second approach is an adaption to the multivariate case of the method proposed in Diggle et al. (2007) (see Section F in the supplementary material for details). To measure the performances of the non-parametric approaches we simply replace the fitted parametric cross PCFs in (B.12) by the non-parametric estimate.

## 4.1 Five-variate LGCP with zero common latent fields

In the case $q = 0$, the MIN rule only selects the true value $q = 0$ for 1% of the simulated data sets while values of $q = 1, 2, 3$ are selected for 99% of the simulations, see left Table B.2. Using the 1-SE rule, $q = 0$ is selected in 77% of the cases and a value of $q$ bigger than 1 is only selected in two cases. The reason that the MIN rule frequently selects $q$ larger than zero may be that in fact neither of the models with $q = 0, \ldots, 3$ are severely overparametrized. E.g. with $q = 3$ the in total 15 PCFs and cross PCFs are parametrized using just 25 parameters, i.e. less than 2 parameters on average for each PCF or cross PCF. Hence overfitting that can be detected by the cross validation procedure mainly occurs for $q = 4, 5$. The middle third column in Table B.2 (left) shows

**Table B.2:** (Left: true $q = 0$. Right: true $q = 2$) Distribution of $q$ chosen by MIN and 1-SE rules, 95% probability interval for selected $\lambda$s, and averages over simulated data sets of percentages of estimated $\alpha_{ik}$s equal to zero.

| $q$ | MIN | 1-SE | $\lambda$ | $(\%\alpha_{ik} = 0)$ | MIN | 1-SE | $\lambda$ | $(\%\alpha_{ik} = 0)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 77 | - | - | 0 | 0 | - | - |
| 1 | 32 | 21 | (0;0.61) | 2 | 2 | 39 | 0 | 0 |
| 2 | 56 | 2 | (0;0.41) | 7 | 60 | 61 | (0;0.38) | 1 |
| 3 | 11 | 0 | (0;0.88) | 6 | 36 | 0 | (0;0.25) | 0.6 |
| 4 | 0 | 0 | - | - | 2 | 0 | 0 | 0 |
| 7 | - | - | (0;4.52) | 52 | - | - | (0;2) | 26 |

95% probability intervals for the selected $\lambda$s when $q = 1, 2, 3$ and the last column shows the average percentages of $\alpha_{ik}$'s that are estimated to be 0. These columns show that when a larger $q$ is selected then also a larger $\lambda$ is selected leading to a higher percentage of zeros in the estimated $\boldsymbol{\alpha}$. This makes sense since larger $q$ means more superfluous parameters and hence more need for regularization. In the case $q = 7$ the selected $\lambda$s tend to be markedly larger than for the smaller $q$s up to 3. Also 52% of the $\alpha_{ik}$ are estimated to be zero in the case $q = 7$ while the percentages are quite small for $q$ up to 3. For $q = 1, 2, 3$ the selected $\lambda$ was zero (meaning no regularization) in 59%, 57%, and 55% of the cases indicating that $q = 1, 2, 3$ already leads to a rather sparse setup and explaining the small percentages of $\alpha_{ik}$ estimated to be zero. Figure B.3 shows the average of $\overline{CV}(q, 0)$ over all simulated data sets and confirms that the CV scores are quite similar across different $q$.

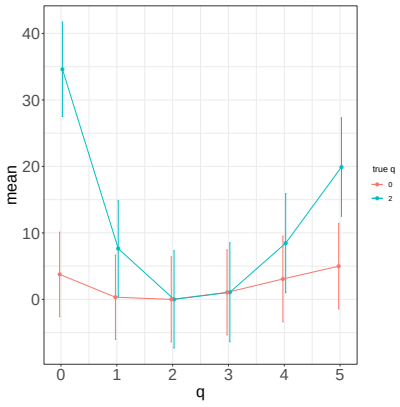


**Fig. B.3:** Averages over simulated data sets of $\overline{CV}(q, 0)$ scores with minimum average CV-score subtracted. The bars show the average of the standard errors $SE(q, 0)$ obtained for $\overline{CV}(q, 0)$ for each simulated data set. Red is for $q = 0$ while blue is for $q = 2$.

Figure B.4 shows means and 95% pointwise probability intervals for esti-

**Table B.3:** MISE using $(q_{\min}, 0)$, $(q_{\min}, \lambda)$, $(q_{1\text{-SE}}, 0)$, simple, or Diggle's non-parametric approach when true $q = 0$.

| | $(q_{\min}, 0)$ | $(q_{\min}, \lambda)$ | $(q_{1\text{-SE}}, 0)$ | simple | Diggle |
|---|---|---|---|---|---|
| $\text{MISE}_{\text{total}}$ | $3.77 \cdot 10^{-4}$ | $3.78 \cdot 10^{-4}$ | $3.81 \cdot 10^{-4}$ | $2.22 \cdot 10^{-3}$ | $2.63 \cdot 10^{-3}$ |
| $\text{MISE}_{\text{within}}$ | $1.02 \cdot 10^{-3}$ | $1.02 \cdot 10^{-3}$ | $1.11 \cdot 10^{-3}$ | $4.58 \cdot 10^{-3}$ | $3.36 \cdot 10^{-3}$ |
| $\text{MISE}_{\text{between}}$ | $5.45 \cdot 10^{-5}$ | $5.34 \cdot 10^{-5}$ | $2.01 \cdot 10^{-5}$ | $1.04 \cdot 10^{-3}$ | $2.27 \cdot 10^{-3}$ |

mates of a subset of the PCFs and cross PCFs obtained for the simulated data sets with $q$ selected among $0, 1, 2, 3, 4, 5$ using either MIN or 1-SE and with $\lambda = 0$. We only show estimates with no regularization since the regularized estimates are very similar. The means are quite similar for MIN and 1-SE and the MIN and 1-SE estimates are close to unbiased for the cross PCFs, while a moderate bias is present for the PCF. This bias is not unexpected as we specify the wrong parametric model. However, the simple non-parametric estimates are strongly biased in all cases.

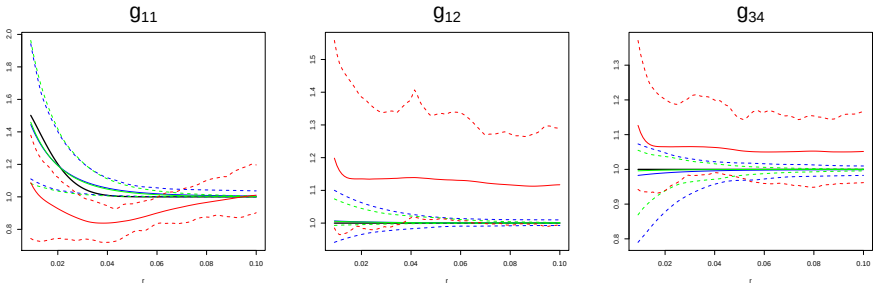


**Fig. B.4:** (true $q = 0$) Blue, green and red solid lines indicate pointwise means of estimates for selected cross PCFs using MIN, 1-SE, or simple non-parametric estimation. The dotted lines indicate the corresponding 95% pointwise probability intervals. Black solid lines indicate true cross PCFs.

Table B.3 gives total, within and between MISEs with different strategies for choosing $(q, \lambda)$ and for the two non-parametric approaches. The non-parametric approaches are clearly outperformed by the semi-parametric method. The results for MIN with $\lambda = 0$ or $\lambda$ selected are very similar and also similar to 1-SE in case of $\text{MISE}_{\text{within}}$. However, $\text{MISE}_{\text{between}}$ for cross PCFs is more than twice as big for MIN compared to 1-SE. This is not so surprising since 1-SE chooses the true $q = 0$ most of the time while MIN tends to choose larger values of $q$. The between MISEs are on the other hand on a much smaller scale than the within MISEs.

## 4.2 Five-variate LGCP with two common latent fields

In case of $q = 2$ both MIN and 1-SE performs quite well in the sense that the chosen $q$'s differ at most by one from the true $q$ in 98% (MIN) or in 100% (1-SE) of the cases and the true $q = 2$ is chosen in 60% (MIN) or 61% (1-SE) of the cases. The $\lambda$ column in Table B.2 (right) shows 95% probability intervals for the selected $\lambda$s. For $q = 1, 4$ the cross validation always selected $\lambda = 0$. The selected $\lambda$s for $q = 2, 3$ are in general small and 80% ($q = 2$) or 95% ($q = 3$) of the $\lambda$s were selected to be zero. These results indicate that regularization is not pertinent in this case where the true $\boldsymbol{\alpha}$ is not particularly sparse. Also the percentages of $\alpha_{ik}$ estimated to be zero are very small for $q = 1, 2, 3, 4$. In case of the overparametrized model $q = 7$ we on the other hand do see an effect of regularization with larger selected $\lambda$s and on average 26% of the $\alpha_{ik}$s estimated to be zero.

Figure B.5 shows means and 95% probability intervals for selected estimated PCFs and cross PCFs obtained with MIN or 1-SE without regularization. In both cases MIN and 1-SE produce some bias for the PCFs, which is expected as we specify the wrong model. As in the case $q = 0$, the non-parametric estimates are more biased than the semi-parametric estimates.

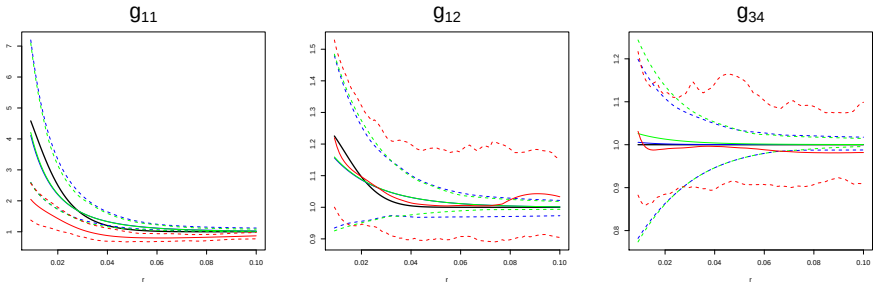


**Fig. B.5:** (true $q = 2$) Blue, green and red solid lines indicate pointwise means of estimates for selected cross PCFs using MIN, 1-SE, or simple non-parametric estimation. The dotted lines indicate the corresponding 95% pointwise probability intervals. Black solid lines indicate true cross PCFs.

Table B.4 shows that MIN and 1-SE perform very similar regarding MISE. In case of $\text{MISE}_{\text{between}}$, MIN and 1-SE are somewhat better than the simple approach but much better than Diggle's approach. On the other hand, MIN and 1-SE are somewhat better than Diggle's approach but much better than the simple approach in terms of $\text{MISE}_{\text{within}}$. Overall ($\text{MISE}_{\text{total}}$) the semi-parametric method outperforms the non-parametric methods.

**Table B.4:** MISE using $(q_{\min}, 0)$, $(q_{\min}, \lambda)$, $(q_{1\text{-SE}}, 0)$, the simple or Diggles non-parametric approach when true $q = 2$.

| | $(q_{\min}, 0)$ | $(q_{\min}, \lambda)$ | $(q_{1\text{-SE}}, 0)$ | simple | Diggle |
|---|---|---|---|---|---|
| $\text{MISE}_{\text{total}}$ | $1.64{\cdot}10^{-3}$ | $1.64{\cdot}10^{-3}$ | $1.65{\cdot}10^{-3}$ | $7.25{\cdot}10^{-3}$ | $4.91{\cdot}10^{-3}$ |
| $\text{MISE}_{\text{within}}$ | $4.43{\cdot}10^{-3}$ | $4.42{\cdot}10^{-3}$ | $4.40{\cdot}10^{-3}$ | $2.04{\cdot}10^{-2}$ | $8.63{\cdot}10^{-3}$ |
| $\text{MISE}_{\text{between}}$ | $2.43{\cdot}10^{-4}$ | $2.43{\cdot}10^{-4}$ | $2.71{\cdot}10^{-4}$ | $6.76{\cdot}10^{-4}$ | $3.05{\cdot}10^{-3}$ |

# 5 Data examples

In the following we apply our new methodology to point patterns of cells in tumor tissue and crime scenes in Washington DC.

## 5.1 Lymph node metastasis

Figure B.6 shows a fluorescence image of a lymph node metastasis as well as point patterns of locations of four types of cells extracted from the image using machine learning techniques. The four types of cells (with abbreviated names and numbers of cells in parantheses) are Hypoxic tumor cells (Hypoxic, 11733), Normoxic tumor cells (Normoxic, 18469), Stroma cells (Stroma, 6015), and Cytotoxic T-lymphocytes (CD8, 1466). For better visualization we only show random subsets obtained by independent thinnings of the points. Our aim is to characterize the point patterns in terms of their intensity functions and their PCFs and cross PCFs.

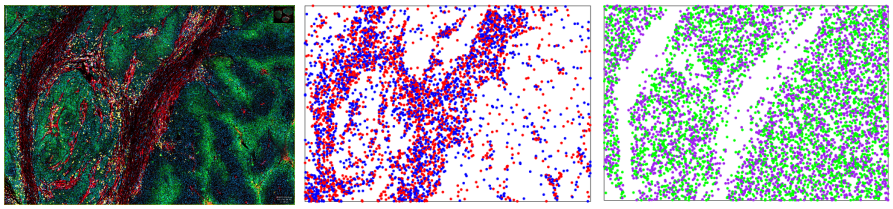

**Fig. B.6:** Left: fluorescence image of a lymph node metastasis. Middle: bivariate point pattern of CD8 (blue) and 50% independently thinned Stroma (red) cells. Right: bivariate point pattern of 80% independently thinned Hypoxic (purple) and 80% independently thinned Normoxic (green) cells (data kindly provided by Arnulf Mayer, Dept. of Radiation Oncology, University Medical Center, Mainz, Germany).

Figure B.7 shows non-parametric estimates of the intensity functions for the four point patterns. These plots show a strong segregation between the patterns of Stroma and CD8 cells versus the tumor cells. In the following we study the more subtle variation within the bivariate point patterns of Stroma and CD8 respectively Normoxic and Hypoxic. There are no spatial covariates available for this data set so the intensity functions will be propor-

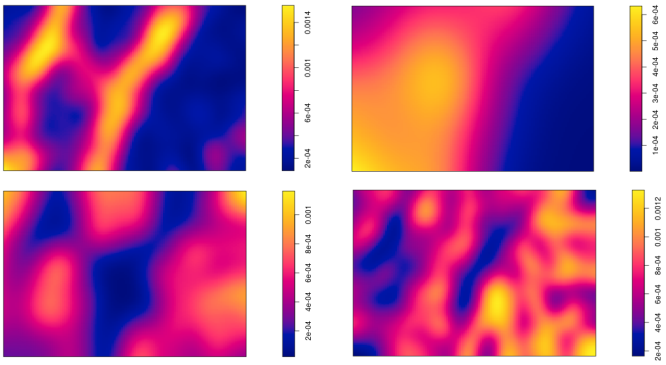

**Fig. B.7:** Kernel estimates of the intensity functions for Stroma (upper left), CD8 (upper right), Hypoxic (lower left) and Normoxic (lower right) with bandwidths 121.2, 359.5, 173,9 and 115.7, respectively.

tional to the common component $\rho_0(\cdot)$ both for the pairs Stroma,CD8 and Normoxic,Hypoxic. Since the point patterns are of high cardinality we reduce computing time by working with independent thinnings of the point patterns. The PCFs and cross PCFs are invariant to independent thinning while the intensity functions are only changed by a multiplication with the thinning probability. In the following we present a detailed analysis of the Stroma-CD8 point pattern. The analysis for the Normoxic-Hypoxic tumor cells is quite similar and is presented in Section G of the supplementary material.

### Stroma and CD8

For Stroma and CD8 we use all CD8 points and independently thin the Stroma points with a thinning probability of 50%. The point patterns clearly show some large scale trends (Figure B.7) that are not easily fitted by simple parametric models. We instead assume the model (B.3), choose CD8 as the baseline, and following Hessellund et al. (2019) estimate $\boldsymbol{\beta} = (\beta_{\text{Str}}, \beta_{\text{CD8}})^{\mathsf{T}} = (\gamma_{\text{Str}} - \gamma_{\text{CD8}}, \gamma_{\text{CD8}} - \gamma_{\text{CD8}})^{\mathsf{T}}$ by $\hat{\beta}_{\text{Str}} = \log(3007/1466) = 0.72$ and $\hat{\beta}_{\text{CD8}} = 0$. We next choose $q$ among the values $\{0, 1, 2\}$ using a 5-fold cross validation as described in Section 3.3, where we resample $L = 10$ times. We choose the maximal interpoint distance $R$ for pairs of points to be 400 $\mu$m which corresponds to approximately 15% of the largest observation window side length. According to the left panel in Figure B.8 we choose $q = 1$ that minimizes the cross validation score. The right panel in Figure B.8 shows a non-parametric estimate of $\rho_0$ using the estimator introduced in Hessellund et al. (2019) with bandwidth chosen as described in Section E in the supplementary material.

According to the parameter estimates in Table B.5 and the resulting PCFs

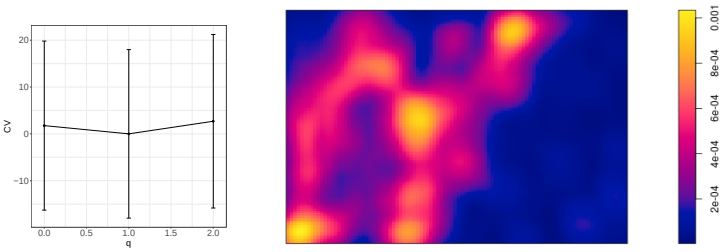


**Fig. B.8:** Left: CV-scores (minus minimum CV-score) with standard errors. Right: non-parametric estimate of $\rho_0$ with bandwidth = 194.5.

**Table B.5:** Parameter estimates for Stroma and CD8 for $q = 1$.

| | $\hat{\alpha}$ | $\hat{\xi}$ | $\hat{\sigma}$ | $\hat{\varphi}$ | | $\hat{\alpha}$ | $\hat{\xi}$ | $\hat{\sigma}$ | $\hat{\varphi}$ |
|---|---|---|---|---|---|---|---|---|---|
| Stroma | 0.52 | 63.4 | 0.32 | 97.7 | CD8 | -0.52 | 63.4 | 0.78 | 193.6 |

and cross PCFs shown in the left panel of Figure B.9 both Stroma and CD8 are randomly clustered point processes. The clustering is partly negatively correlated (cf. $\hat{\alpha}$ and the fitted cross PCF in Figure B.9) and partly independent (cf. $\hat{\sigma}$) between Stroma and CD8. The strongest clustering is found for CD8 due to the higher value of $\hat{\sigma}_{\text{CD8}}$ than $\hat{\sigma}_{\text{Str}}$, see also the fitted PCFs in Figure B.9.

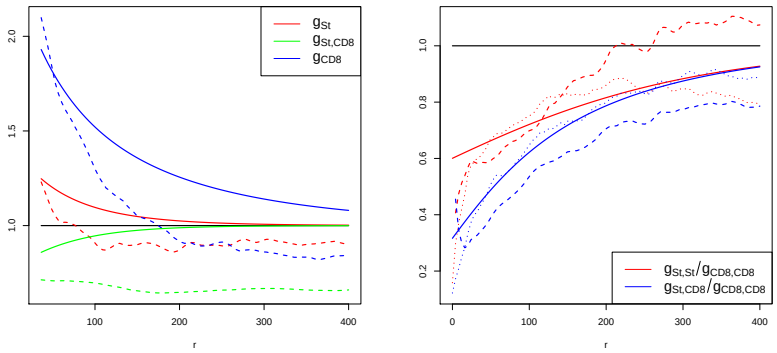


**Fig. B.9:** Left: estimated (cross) PCFs using the semi-parametric model (solid) and simple approach (dashed). Right: estimated (cross) PCF ratio using semi-parametric model (solid), simple approach (dashed) and consistent approach (dotted).

The total estimated variances for the log random intensity functions of Stroma and CD8 are rather moderate, respectively 0.37 and 0.88, while the empirical variance of $\log \hat{\rho}_0$ over the observation window is 1.15. In this sense,

the majority of the variation in the random intensity functions (especially for Stroma) is explained by $\rho_0$.

Following Section 3.4, the right panel in Figure B.9 compares semi-parametric estimates of cross PCF ratios $g_{\text{Str}}/g_{\text{CD8}}$ and $g_{\text{Str,CD8}}/g_{\text{CD8}}$ with the non-parametric estimates introduced in Hessellund et al. (2019). The agreement seems reasonable and this is confirmed by global envelope $p$-values of 0.05 in case of $g_{\text{Str}}/g_{\text{CD8}}$ and 0.09 for $g_{\text{Str,CD8}}/g_{\text{CD8,CD8}}$, see also the global envelope plots in Section H in the supplementary material.

Figure B.9 (left) also shows simple non-parametric PCF and cross PCF estimates which are generally smaller than the semi-parametric estimates. In particular, the non-parametric estimate of the cross PCF suggests a strong negative correlation between Stroma and CD8 for all spatial lags considered. As discussed in Shaw et al. (2020) this might be due to that the selected bandwidths imply too little smoothing in the non-parametric intensity estimates (upper plots in Figure B.7). Figure B.9 (right) further shows that the simple estimates of cross PCF ratios deviate more from the consistent non-parametric estimates than the semi-parametric estimates.

## 5.2 Washington DC street crimes

It is of great interest for criminologists and police authorities to study the spatial patterns of crime scenes since this can lead to better understanding of factors affecting crime and more efficient policing strategies. In this section we focus on the spatial correlation between six common types of street crimes committed in Washington DC in January and February 2017. These are extracted from a larger data set publicly available from the website [1] The

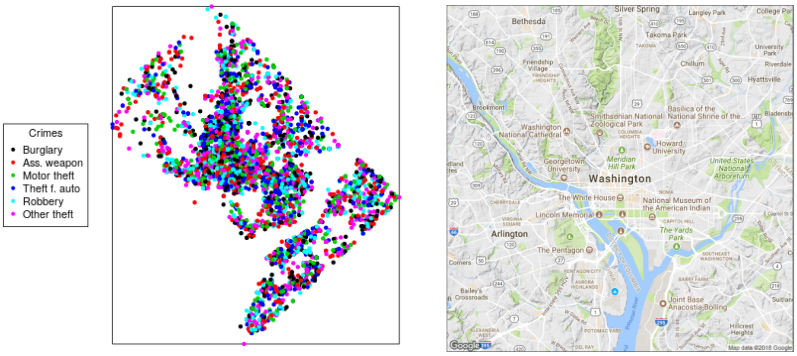


**Fig. B.10:** Left: street crimes locations ($n = 5378$). Right: a map of Washington DC.

six types of crimes with numbers in parantheses are 1) Burglary (259), 2)

[1] `http://opendata.dc.gov/datasets/`

Assault with weapon (332), 3) Motor vehicle theft (335), 4) Theft from automobile (1832), 5) Robbery (366), and 6) Other theft (2254). This data set has previously been considered by Hessellund et al. (2019) who focused on the dependency of the street crime intensity functions on spatial covariates using the model (B.3). In the following we focus on the second order properties as described by the PCFs and cross PCFs. We refer to Hessellund et al. (2019) for more details regarding the covariates and the fitted intensity functions.

We apply the regularized estimation approach described in Sections 3.2-3.3 where we first determine $q$ using cross validation without regularization and next, for the chosen $q$, use another cross validation to select the regularization parameter $\lambda$ to potentially obtain a sparse submodel for the chosen $q$. For the cross validation we use $K = 5$ and $L = 10$ and choose $q$ among $\{0, 1, \ldots, 5\}$ and $\lambda$ in the set $\{100, 80, 60, 50, 40, 30, 20, 10, 5, 2.5, 1, 0.25, 0\}$. For the second order composite likelihood we use $R = 1000$ meters.

The left panel in Figure B.11 shows cross validation scores for each $q$, where the 1-SE criterion leads to choosing $q = 0$ while MIN chooses $q = 1$. The middle panel show CV-scores for each $\lambda$ with $q = 1$ where the minimum is obtained for $\lambda = 0$. The right panel in Figure B.11 shows a non-parametric estimate of $\rho_0$ using the estimator described in Hessellund et al. (2019) with bandwidth chosen as described in Section E in the supplementary material. In the following we focus on the results with $q = 1$ and $\lambda = 0$. Hence we obtain an estimate of $\boldsymbol{\alpha}$ without any regularization.

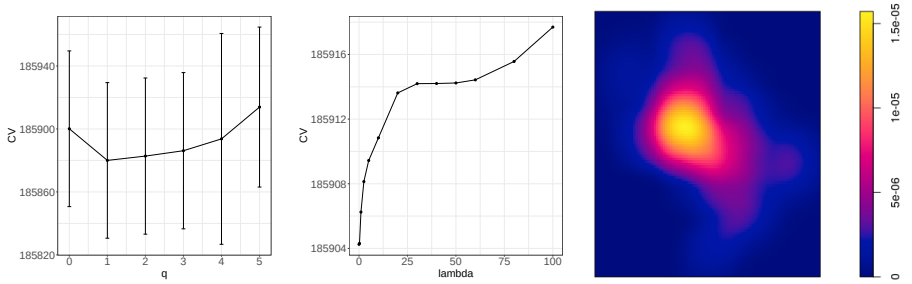


**Fig. B.11:** Left: CV-score for $q$ with one standard error bars. Middle: CV-score for $\lambda$ given $q = 1$. Right: non-parametric estimate of $\rho_0$ with bandwidth 2654 meters.

The parameter estimates for each street crime are given in Table B.6 except for the common latent field correlation scale parameter estimate which is $\hat{\xi} = 102.5$. The $\sigma_i$ estimates are small to moderate for the first five crimes while the estimate $\hat{\sigma}_6$ for Other theft is about two times larger than the other $\sigma_i$ estimates. Regarding the latent field $Y_1$ the $\alpha_{i1}$ estimates are pretty small for Assault, Vehicle theft, Theft from auto, and Robbery while $\alpha_{11}$ for Burglary and $\alpha_{61}$ for Other theft have fairly large estimates 0.78 and $-0.93$. The resulting estimated PCFs and cross PCFs are shown in the left and middle panels

**Table B.6:** Table of parameter estimates for each street crime for $(q, \lambda) = (1, 0)$. Last two columns show estimates of $\boldsymbol{\alpha}_{\cdot l}$, $l = 1, 2$ with $(q, \lambda) = (2, 2.5)$.

| Crime type | $\hat{\boldsymbol{\alpha}}$ | $\hat{\sigma}$ | $\hat{\varphi}$ | $\hat{\boldsymbol{\alpha}}_{\cdot 1}$ | $\hat{\boldsymbol{\alpha}}_{\cdot 2}$ |
|---|---|---|---|---|---|
| Burglary | 0.78 | 0.50 | 245.8 | 0 | 0.76 |
| Assault | -0.12 | 0.51 | 457.5 | 0 | -0.09 |
| Vehicle Theft | 0.49 | 0.14 | 20.5 | 0 | 0.47 |
| Theft F. Auto | 0.09 | 0.58 | 2483.1 | 0 | 0.08 |
| Robbery | -0.30 | 0.53 | 485.2 | 0 | -0.26 |
| Other theft | -0.93 | 0.96 | 20.5 | 0 | -0.97 |

of Figure B.12. The overall conclusion is that most crimes are moderately clustered except for Burglary and Other theft with strongest clustering for Other theft. Also the cross dependencies seem fairly weak except for pairs involving the crimes Burglary, Vehicle theft and Other theft (crimes 1,3,6) with Burglary and Vehicle theft being positively correlated and Burglary and Other theft being negatively correlated. The interpretation of these results is that except for moderate random fluctuations, the spatial patterns of Assault, Vehicle theft, Theft from auto and Robbery are quite well described by their intensity functions depending on the common factor $\rho_0$ as well as covariate effects. On the other hand, the random intensity functions for Burglary and Other theft seem subject to more pronounced deviations from the intensity functions and these deviations are negatively correlated. In other words if a cluster of Burglaries not explained by the intensity function is present in a certain area then there tends to be less Other theft committed in the same area and vice versa.

We also tried out $q = 2$ for which the cross validation score is quite close to the one for $q = 1$. For $q = 2$ the cross validation selected $\lambda = 2.5$. The last columns in Table B.6 show the estimate of $\boldsymbol{\alpha}$ obtained with $(q, \lambda) = (2, 2.5)$. The lasso regularization has shrunk $\hat{\boldsymbol{\alpha}}_{\cdot 1}$ to $\mathbf{0}$, while the estimate of $\boldsymbol{\alpha}_{\cdot 2}$ is quite similar to the estimate of $\boldsymbol{\alpha}_{\cdot 1}$ for $q = 1$. In view of this, one may argue that the lasso regularization makes our estimation approach more robust, since a too large selected $q$ can be counterbalanced by regularization on $\boldsymbol{\alpha}$ with a $\lambda > 0$.

Quite different conclusions are obtained with the simple non-parametric analysis. Figure B.12 shows that the non-parametric estimates of the PCFs and cross PCFs are all considerably above the reference value 1 which would imply strong clustering within and between the different types of crime. These results may well be explained by bias of the non-parametric estimates.

For model assessment we consider global envelope tests based on differences between semi-parametric and consistent non-parametric estimates for all 20 ratios $g_{ij}/g_{66}$, $1 \leq i \leq j \leq 6$, $(i, j) \neq (6, 6)$. The $p$-values obtained are between 0.089 and 0.624 and hence do not provide evidence against our

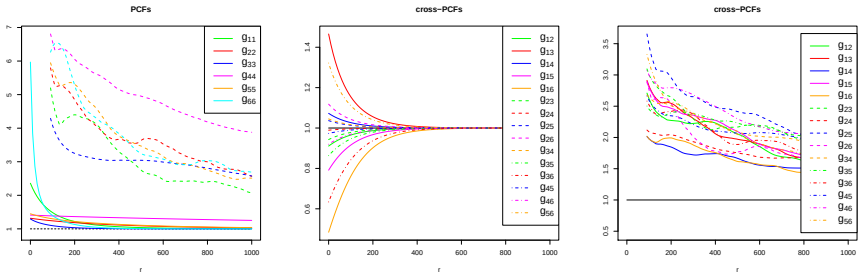


**Fig. B.12:** Left: semi-parametric (solid) and simple non-parametric (dashed) estimates of PCFs for $(q, \lambda) = (1, 0)$. Middle: semi-parametric estimates of cross PCFs for $(q, \lambda) = (1, 0)$. Right: simple non-parametric estimates of cross PCFs.

model. Some representative global envelope plots for the differences are shown in Section I in the supplementary material.

We finally consider an explorative analysis focusing on patterns in the common latent process $Y_1$. We define 'residuals' $\Delta \log \Lambda_i(\mathbf{u})$ by $\log \Lambda_i(\mathbf{u}) - \mu_i - \boldsymbol{\beta}_i^\mathsf{T} \mathbf{z}(\mathbf{u}) - \frac{1}{p} \sum_{l=1}^{p} [\log \Lambda_l(\mathbf{u}) - \mu_l - \boldsymbol{\beta}_l^\mathsf{T} \mathbf{z}(\mathbf{u})]$. Due to the sum-to-zero constraint on $\boldsymbol{\alpha}$ we obtain:

$$\Delta \log \Lambda_i(\mathbf{u}) = \alpha_{i1} Y_1(u) + \sigma_i U_i - \frac{1}{p} \sum_{l=1}^{p} \sigma_l U_l(\mathbf{u}).$$

Estimating $\Delta \log \Lambda_i$ by replacing $\Lambda_i$ by a kernel estimate and the parameters by their conditional likelihood estimates, we obtain:

$$\widehat{Y}_1(u) = (\hat{\boldsymbol{\alpha}}^\mathsf{T} \hat{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\alpha}} [\widehat{\Delta} \log \Lambda_1(u), \ldots, \widehat{\Delta} \log \Lambda_p(u)]^\mathsf{T}.$$

The left plot in Figure B.13 shows $\widehat{Y}_1$ where the $\Lambda_i$ are estimated by kernel smoothing using a bandwidth of 3 km. There is some resemblance between $\hat{Y}_1(\mathbf{u}; h)$ and the spatial distribution of median income shown in the middle plot of Figure B.13. Log median income is included as a covariate in the regression model for the log intensity so it may be the case that $\hat{Y}_1(\mathbf{u})$ reflects nonlinear effects of the financial status of a neighborhood, cf. the right plot in Figure B.13.

# 6 Conclusion

The methodology introduced in this paper provides a major step forward regarding second order analysis of multivariate point processes with complex intensity functions. Existing approaches (such as simple non-parametric estimation or the approach in Diggle et al., 2007) rely on estimating the intensity functions using kernel estimators and this tends to result in strong bias

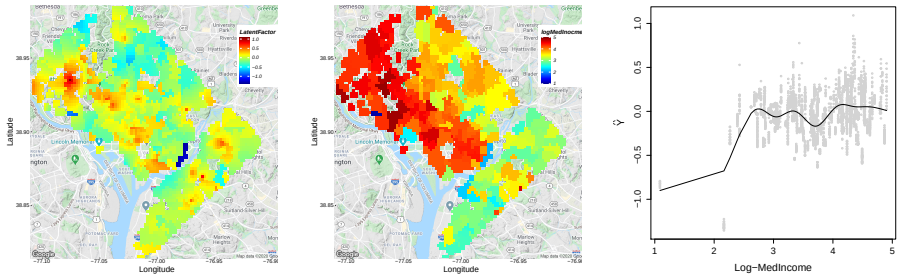


**Fig. B.13:** Left: latent factor $\widehat{Y}_1$. Middle: log median income within census tracts. Right: $\hat{Y}_1$ versus log median income.

and/or large variance for subsequent estimation of PCFs and cross PCFs. In contrast, in the context of the model (B.3), our approach circumvents the need to estimate the complex unknown intensity function factor $\rho_0$ and the resulting PCF and cross PCFs appear to be close to unbiased according to our simulation studies. For the data examples considered we obtain simple and interpretable models that may result in better understanding of the interplay between respectively cells in tumors and different types of crimes.

A limitation of our approach, shared with existing methods, is that we have not provided confidence intervals for parameter estimates or confidence bands for estimated PCFs or cross PCFs. One topic for further research would be to establish asymptotic results for parameter estimates within the framework of estimating function inference. This was done by Hessellund et al. (2019) regarding inference for the intensity function but the current problem of inferring cross PCFs entail considerable additional theoretical difficulties.

The impact of using regularization was not very strong in our simulation studies when moderate values of $q$ were considered. However, the crimes data example indicates that the use of regularization may add robustness to the estimation procedure if a too large $q$ is selected.

**Acknowledgements**

We thank Arnulf Mayer, Dept. of Radiation Oncology, University Medical Center, Mainz, Germany, for providing the fluorescence image and the point pattern data. Kristian B. Hessellund and Rasmus Waagepetersen were supported by The Danish Council for Independent Research | Natural Sciences, grant DFF - 7014-00074 'Statistics for point processes in space and beyond', and by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by grant 8721 from the Villum Foundation.

## SUPPLEMENTARY MATERIAL

The supplementary material for this paper contains further plots, an algorithm for updating regularized $\boldsymbol{\alpha}$ and auxiliary results.

# A Conditional probability and likelihood

Define $X^{\text{pooled}} = \cup_{i=1}^{p} X_i$ with intensity function $\sum_i \rho_i$ and second order joint intensity $\sum_{l,k} \rho_{lk}$. Define further the measure

$$C(A \times B \times \{i\} \times \{j\}) = \mathrm{E} \sum_{\mathbf{u},\mathbf{v} \in X^{\text{pooled}}}^{\neq} 1[\mathbf{u} \in A, \mathbf{v} \in B, \mathbf{u} \in X_i, \mathbf{v} \in X_j]$$

By Radon-Nikodym's theorem,

$$C(A \times B \times \{i\} \times \{j\}) = \int_{A \times B} p_{ij}(\mathbf{u}, \mathbf{v}) \left[ \sum_{l,k} \rho_{lk}(\mathbf{u}, \mathbf{v}) \right] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v},$$

where for almost all $(\mathbf{u}, \mathbf{v})$, $p_{ij}(\mathbf{u}, \mathbf{v})$ is a probability function on $\{1, \ldots, p\} \times \{1, \ldots, p\}$. This follows because $C(\cdot \times \cdot \times \{i\} \times \{j\})$ is absolutely continuous with respect to the second order factorial measure

$$\alpha(A \times B) = \mathrm{E} \sum_{\mathbf{u},\mathbf{v} \in X^{\text{pooled}}}^{\neq} 1[\mathbf{u} \in A, \mathbf{v} \in B]$$

of $X^{\text{pool}}$ which has density $\sum_{l,k} \rho_{lk}$. It is then natural to interpret $p_{ij}(\mathbf{u}, \mathbf{v})$ as the conditional (Palm) probability that $\mathbf{u} \in X_i, \mathbf{v} \in X_j$ given that $\mathbf{u}, \mathbf{v} \in X^{\text{pool}}$ since $\sum_{l,k} \rho_{lk}(\mathbf{u}, \mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$ is the 'probability' that $X^{\text{pool}}$ 'has points at $\mathbf{u}$ and $\mathbf{v}$'. On the other hand,

$$C(A \times B \times \{i\} \times \{j\}) = \int_{A \times B} \rho_{ij}(\mathbf{u}, \mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.$$

Thus we obtain (B.6).

Another way to arrive at (B.6) is to define point processes $X_{ij} = \{(\mathbf{u}, \mathbf{v}) | \mathbf{u} \in X_i, \mathbf{v} \in X_j, u \neq v\}$ with intensity functions $\rho_{ij}(\mathbf{u}, \mathbf{v})$. We can further define the union $\tilde{X}^{\text{pool}} = \cup_{i,j} X_{ij}$ with intensity function $\sum_{k,l} \rho_{kl}$. If we now condition on $\tilde{X}^{\text{pool}}$ and consider a point $(\mathbf{u}, \mathbf{v}) \in \tilde{X}^{\text{pool}}$, then (B.6) is the conditional probability that this point comes from $X_{ij}$. We could also define $\tilde{X}^{\text{pool}}$ as $\cup_{i \leq j} X_{ij}$ in which case we would get the conditional probabilities

$$\mathrm{q}_{ij}(\mathbf{u}, \mathbf{v}) = \frac{\rho_{ij}(\mathbf{u}, \mathbf{v})}{\sum_{l \leq k} \rho_{lk}(\mathbf{u}, \mathbf{v})}. \tag{B.13}$$

**Remark A.1**
An important property of (B.7) is that the score function is unbiased, see Lemma B.1. For this to hold it is crucial that when we sum over all $l, k$ in the denominator of (B.6) we also use product over all $i, j$ in (B.7). Alternatively, using (B.13) we could define

$$L(\boldsymbol{\theta}) = \prod_{i \leq j} \overset{\neq}{\prod_{\substack{\mathbf{u} \in X_i \cap W \\ \mathbf{v} \in X_j \cap W \\ ||\mathbf{u}-\mathbf{v}|| \leq R}}} \mathrm{q}_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}). \tag{B.14}$$

In this case a pair $\{\mathbf{u}, \mathbf{v}\}$ with $\mathbf{u} \in X_i$ and $\mathbf{v} \in X_j$ only appears once for $i \neq j$ since the sum is now only over $i \leq j$. However, a pair $\mathbf{u} \neq \mathbf{v} \in X_i$ will contribute twice to the likelihood. We tried out the two alternatives (B.7) and (B.14) on a number of data sets and got very similar estimates.

**Remark A.2**
Note that if an ordered pair $(\mathbf{u}, \mathbf{v})$ appears in the product in (B.7) then so does $(\mathbf{v}, \mathbf{u})$. Hence in a practical implementation we may restrict the product to $i \leq j$ and if $i = j$ only include unordered pairs $\{\mathbf{u}, \mathbf{v}\}$ with $\mathbf{u} \neq \mathbf{v} \in X_i$. We can finally square to get (B.7).

# B Theoretical results concerning conditional composite likelihood score and Hessian

In this section $\boldsymbol{\theta}^*$ denotes the parameter vector for which the data is generated.

**Lemma B.1**
The score function $\mathrm{e}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\tau}} l_{-}(\boldsymbol{\theta})$ is unbiased meaning $\mathrm{E}[\mathrm{e}(\boldsymbol{\theta}^*)] = 0$.

*Proof.*

$$\begin{aligned}
\mathrm{E}[\mathrm{e}(\boldsymbol{\theta}^*)] &= \mathrm{E}\left[ \sum_{i,j} \overset{\neq}{\sum_{\substack{\mathbf{u} \in X_i \cap W \\ \mathbf{v} \in X_j \cap W}}} \left( \frac{\sum_{k,l} \nabla_{\boldsymbol{\tau}} \rho_{kl}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{kl})}{\sum_{k,l} \rho_{kl}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{kl})} - \frac{\nabla_{\boldsymbol{\tau}} \rho_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{ij})}{\rho_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{ij})} \right) \right] \\
&= \sum_{i,j} \int_{W^2} \left( \frac{\sum_{k,l} \nabla_{\boldsymbol{\tau}} \rho_{kl}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{kl})}{\sum_{k,l} \rho_{kl}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{kl})} - \frac{\nabla_{\boldsymbol{\tau}} \rho_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{ij})}{\rho_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{ij})} \right) \rho_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{ij}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\
&= \int_{W^2} \sum_{k,l} \nabla_{\boldsymbol{\tau}} \rho_{kl}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{kl}) - \sum_{i,j} \nabla_{\boldsymbol{\tau}} \rho_{ij}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*_{ij}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} = \mathbf{0}. \qquad \square
\end{aligned}$$

**Lemma B.2**
The expected Hessian matrix of $l_{-}(\boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\tau}$ is given by:

$$\mathrm{H}(\boldsymbol{\theta}^*) = \int_{W^2} 1_R(\mathbf{u},\mathbf{v})\mathrm{Cov}(Z(\mathbf{u},\mathbf{v},\boldsymbol{\theta}^*)) \sum_{i,j} \rho_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\theta}^*)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v},$$

where for each $\mathbf{u}$, $\mathbf{v}$, $Z(\mathbf{u},\mathbf{v},\boldsymbol{\theta}^*)$ denotes a random vector which takes values $\nabla_{\boldsymbol{\tau}}\log\rho_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\theta}^*)$ with probabilities $\mathrm{p}_{ij}(\mathbf{u},\mathbf{v};\boldsymbol{\theta}^*)$, $i,j = 1,\ldots,p$.

*Proof.* We suppress the arguments $\mathbf{u}$, $\mathbf{v}$ and $\boldsymbol{\theta}^*$ in order to save space. The Hessian matrix for $l_{-}(\boldsymbol{\theta})$ is:

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\tau}^{\mathsf{T}}} e(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\tau}} l_{-}(\boldsymbol{\theta}) = \\
&\sum_{i,j} \sum_{\substack{\mathbf{u}\in X_i\cap W \\ \mathbf{v}\in X_j\cap W}}^{\neq} \frac{(\sum_{k,l}\nabla^2_{\boldsymbol{\tau}}\rho_{kl})(\sum_{k,l}\rho_{kl}) - (\sum_{k,l}\nabla_{\boldsymbol{\tau}}\rho_{kl})(\sum_{k,l}\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\rho_{kl})}{(\sum_{k,l}\rho_{kl})^2} \quad - \\
&\sum_{i,j} \sum_{\substack{\mathbf{u}\in X_i\cap W \\ \mathbf{v}\in X_j\cap W}}^{\neq} \frac{(\nabla^2_{\boldsymbol{\tau}}\rho_{ij})\rho_{ij} - (\nabla_{\boldsymbol{\tau}}\rho_{ij})(\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\rho_{ij})}{\rho^2_{ij}}
\end{aligned}$$

The expected Hessian is then given by:

$$\begin{aligned}
&\mathrm{H}(\boldsymbol{\theta}^*) \\
&= \sum_{i,j}\int_{W^2} 1_R \frac{(\sum_{k,l}\nabla^2_{\boldsymbol{\tau}}\rho_{kl})(\sum_{k,l}\rho_{kl}) - (\sum_{k,l}\nabla_{\boldsymbol{\tau}}\rho_{kl})(\sum_{k,l}\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\rho_{kl})}{(\sum_{k,l}\rho_{kl})^2}\rho_{ij}dudv \quad - \\
&\sum_{i,j}\int_{W^2} 1_R \frac{(\nabla^2_{\boldsymbol{\tau}}\rho_{ij})\rho_{ij} - (\nabla_{\boldsymbol{\tau}}\rho_{ij})(\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\rho_{ij})}{\rho^2_{ij}}\rho_{ij}dudv \\
&= \int_{W^2} 1_R \sum_{i,j} \frac{(\nabla_{\boldsymbol{\tau}}\rho_{ij})(\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\rho_{ij})}{\rho_{ij}} - \frac{(\sum_{k,l}\nabla_{\boldsymbol{\tau}}\rho_{kl})(\sum_{k,l}\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\rho_{kl})}{(\sum_{k,l}\rho_{kl})} dudv \\
&= \int_{W^2} 1_R \left(\sum_{k,l}\rho_{kl}\right)\left(\sum_{i,j}\nabla_{\boldsymbol{\tau}}\log(\rho_{ij})\nabla^{\mathsf{T}}_{\boldsymbol{\tau}}\log(\rho_{ij})\mathrm{p}_{ij}\right) dudv \\
&\quad - \int_{W^2} 1_R \left(\sum_{k,l}\rho_{kl}\right)\left((\sum_{k,l}\nabla_{\boldsymbol{\tau}}\log(\rho_{kl})\mathrm{p}_{kl})(\sum_{k,l}\nabla_{\boldsymbol{\tau}}\log(\rho_{kl})\mathrm{p}_{kl})^{\mathsf{T}}\right) dudv \\
&= \int_{W^2} 1_R \left(\sum_{i,j}\rho_{ij}\right)\mathrm{Cov}(Z(\boldsymbol{\theta}^*))dudv. \qquad \square
\end{aligned}$$

# C Quadratic approximation and least squares

Omitting the first term not depending on $\boldsymbol{\tau}$ and letting $K = -\mathrm{H}^{-1}(\tilde{\boldsymbol{\tau}})\mathrm{e}(\tilde{\boldsymbol{\tau}})$, the quadratic approximation (B.8) can be rewritten as follows:

$$\begin{aligned}
&(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})^{\mathsf{T}}\mathrm{e}(\tilde{\boldsymbol{\tau}}) + \frac{1}{2}(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})^{\mathsf{T}}\mathrm{H}(\tilde{\boldsymbol{\tau}})(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}}) \\
=&-(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})^{\mathsf{T}}\mathrm{H}(\tilde{\boldsymbol{\tau}})K + \frac{1}{2}(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})^{\mathsf{T}}\mathrm{H}(\tilde{\boldsymbol{\tau}})(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}}) \\
=&\frac{1}{2}(K-(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}}))^{\mathsf{T}}\mathrm{H}(\tilde{\boldsymbol{\tau}})(K-(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}})) - \frac{1}{2}K^{\mathsf{T}}\mathrm{H}(\tilde{\boldsymbol{\tau}})K.
\end{aligned}$$

Hence, minimizing (B.8) is a least squares problem:

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}}\left(\|\mathrm{H}(\tilde{\boldsymbol{\tau}})^{1/2}(K-(\boldsymbol{\tau}-\tilde{\boldsymbol{\tau}}))\|^2\right) = \arg\min_{\boldsymbol{\tau}}\left(\|Y-X\boldsymbol{\tau}\|^2\right),$$

where

$$Y = \mathrm{H}(\tilde{\boldsymbol{\tau}})^{1/2}\left(-\mathrm{H}(\tilde{\boldsymbol{\tau}})^{-1}\mathrm{e}(\tilde{\boldsymbol{\tau}}) + \tilde{\boldsymbol{\tau}}\right) \quad \text{and} \quad X = \mathrm{H}(\tilde{\boldsymbol{\tau}})^{1/2}.$$

# D Update of regularized $\boldsymbol{\alpha}$

Following Shi et al. (2016) we update $\boldsymbol{\alpha}$ by minimizing the augmented Lagrangian object function

$$Q_{\lambda,\mu}(\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{1}{2}||Y-X\boldsymbol{\alpha}||^2 + \lambda\sum_{i=1}^{p}\sum_{j=1}^{q}|\alpha_{ij}| + \boldsymbol{\eta}\mathbf{C}\boldsymbol{\alpha} + \frac{\mu}{2}||\mathbf{C}\boldsymbol{\alpha}||^2,$$

where $\boldsymbol{\eta} \in \mathbb{R}^q$ is the Lagrange multiplier and $\mu > 0$ is a penalty parameter that we set to 1 as in Shi et al. (2016). Letting $\boldsymbol{\alpha}^{\text{current}}$, $\boldsymbol{\eta}^{\text{current}}$, $\boldsymbol{\alpha}^{\text{new}}$, and $\boldsymbol{\eta}^{\text{new}}$ denote temporary vectors used in the iterative algorithm, we initialize $\boldsymbol{\alpha}^{\text{current}} = \boldsymbol{\alpha}^{(n)}$ and $\boldsymbol{\eta}^{\text{current}} = \mathbf{0}$. We then iterate updates

$$\boldsymbol{\alpha}^{\text{new}} \leftarrow \arg\min_{\boldsymbol{\alpha}} Q_{\lambda,\mu}(\boldsymbol{\alpha},\boldsymbol{\eta}^{\text{current}},\boldsymbol{\pi}^{(n)}) \quad \boldsymbol{\eta}^{\text{new}} \leftarrow \boldsymbol{\eta}^{\text{current}} + \mu\mathbf{C}\boldsymbol{\alpha}^{\text{new}}.$$

The updating is terminated if for some $\tilde{\varepsilon}$ and $\tilde{\tilde{\varepsilon}}$,

$$||\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{current}}|| < \tilde{\varepsilon} \quad \text{and} \quad ||\boldsymbol{\eta}^{\text{new}} - \boldsymbol{\eta}^{\text{current}}|| < \tilde{\tilde{\varepsilon}}, \tag{B.15}$$

in which case $\hat{\boldsymbol{\alpha}} := \boldsymbol{\alpha}^{\text{new}}$. Otherwise $\boldsymbol{\alpha}^{\text{current}} \leftarrow \boldsymbol{\alpha}^{\text{new}}$ and $\boldsymbol{\eta}^{\text{current}} \leftarrow \boldsymbol{\eta}^{\text{new}}$ and a new iteration takes place.

The update leading to $\boldsymbol{\alpha}^{\text{new}}$ is conducted using cyclical updating of the entries $\alpha_{ij}^{\text{new}}$ in $\boldsymbol{\alpha}^{\text{new}}$. The update of the $ij$th entry is

$$\begin{aligned} c_1 &\leftarrow X_{\cdot ij}^{\mathsf{T}} \left( Y - \sum_{lk \neq ij} X_{\cdot lk} \alpha_{lk}^{\text{new}} \right) \\ c_2 &\leftarrow \mu \left( \sum_{lk \neq ij} \alpha_{lk}^{\text{new}} \mathbf{C}_{\cdot lk}^{\mathsf{T}} \mathbf{C}_{\cdot ij} + \mathbf{C}_{\cdot ij}^{\mathsf{T}} \boldsymbol{\eta} / \mu \right) \\ \alpha_{ij}^{\text{new}} &\leftarrow \frac{S\left(c_1 - c_2, \lambda\right)}{X_{\cdot ij}^{\mathsf{T}} X_{\cdot ij} + \mu \mathbf{C}_{\cdot ij}^{\mathsf{T}} \mathbf{C}_{\cdot ij}}, \end{aligned} \tag{B.16}$$

where $X_{\cdot ij}$ and $C_{\cdot ij}$ are the columns of $X$ and $C$ corresponding to $\alpha_{ij}$ (i.e. when $\boldsymbol{\alpha}$ is laid out columnwise) and $S(\cdot, \cdot)$ is a soft-thresholding operator given by:

$$S(c, \lambda) = \begin{cases} c - \lambda & \text{if } c > 0 \text{ and } \lambda < |c| \\ c + \lambda & \text{if } c < 0 \text{ and } \lambda < |c| \\ 0 & \text{if } \lambda > |c|. \end{cases}$$

Algorithm 2 gives an overview of the regularized cyclical block descent algorithm.

**Algorithm 2** Regularized cyclical block descent algorithm

1: Simulate initial parameters $\hat{\boldsymbol{\alpha}}^{(0)}$, $\hat{\boldsymbol{\xi}}^{(0)}$, $\hat{\sigma}^{2(0)}$ and $\hat{\boldsymbol{\varphi}}^{(0)}$
2: $n := 0$
3: ***repeat***
4: $\quad \tilde{\boldsymbol{\alpha}} := \boldsymbol{\alpha}^{(n)}$, $\tilde{\boldsymbol{\xi}} := \boldsymbol{\xi}^{(n)}$, $\sigma^2 := \sigma^{2(n)}$ and $\boldsymbol{\varphi} := \boldsymbol{\varphi}^{(n)}$
5: $\quad$ update $\tilde{\boldsymbol{\alpha}}$ using augmented Lagrangian method followed by line search
6: $\quad$ update $\tilde{\boldsymbol{\xi}}$, $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\varphi}}$ in turn using (B.9) followed by line search
7: $\quad \boldsymbol{\alpha}^{(n+1)} := \tilde{\boldsymbol{\alpha}}$, $\boldsymbol{\xi}^{(n+1)} := \tilde{\boldsymbol{\xi}}$, $\sigma^{2(n+1)} := \tilde{\sigma}^2$, and $\boldsymbol{\varphi}^{(n+1)} := \tilde{\boldsymbol{\varphi}}$
8: $\quad n := n + 1$
9: ***until*** relative convergence for object function (B.11)
10: ***return*** $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(n)}$

# E Bandwidth selection for $\rho_0$

Following Hessellund et al. (2019) we can estimate $\rho_0$ using the semi-parametric kernel estimator:

$$\hat{\rho}_0(\mathbf{u}) = \frac{1}{p} \sum_{i=1}^{p} \sum_{\mathbf{v} \in X_i} \exp(-\hat{\boldsymbol{\beta}}_i^{\mathsf{T}} \mathbf{z}(\mathbf{v})) \frac{k((\mathbf{u} - \mathbf{v})/b)}{b^d c_b(\mathbf{v})}, \tag{B.17}$$

where $k$ is a $d$-dimensional kernel, $b$ is the bandwidth, and $c_b(v) = \int_W k(\mathbf{u} - \mathbf{v})\mathrm{d}\mathbf{u}$ is an edge correction factor.

We suggest to choose the bandwidth according to a criterion inspired by Cronie and van Lieshout (2016) who consider the squared difference between the observation window area and an estimate (depending on the bandwidth) of this area. However, exact knowledge of the observation window area may not always be available (this is e.g. the case for the crimes data where the observation window depends on the complex urban structure of Washington DC). To handle this we take advantage of our multivariate setup. Define $X^{\text{pooled}} = \cup_{i=1}^{p} X_i$ with intensity function $\rho^{\text{pooled}}(\mathbf{u}) = \sum_{i=1}^{p} \rho_i(\mathbf{u})$. An estimator of $\rho^{\text{pooled}}$ is simply given by $\hat{\rho}^{\text{pooled}}(\mathbf{u}) = \hat{\rho}_0(\mathbf{u}) \sum_{i=1}^{p} \exp(\hat{\boldsymbol{\beta}}_i^{\mathsf{T}} \mathbf{z}(\mathbf{u}))$. We can then define two different estimators, $\hat{\omega}$ and $\hat{w}$, for the area of the observation window:

$$\hat{\omega}(b) = \frac{1}{p} \sum_{i=1}^{p} \sum_{\mathbf{u} \in X_i} \frac{1}{\hat{\rho}_0(\mathbf{u}) \exp(\hat{\boldsymbol{\beta}}_i^{\mathsf{T}} \mathbf{z}(\mathbf{u}))} \quad \text{and} \quad \hat{w}(b) = \sum_{\mathbf{u} \in X^{\text{pooled}}} \frac{1}{\hat{\rho}^{\text{pooled}}(\mathbf{u})}$$

where the dependence on $b$ is through $\hat{\rho}_0(\cdot)$. We then select the bandwidth $b$ that minimizes $(\hat{\omega}(b) - \hat{w}(b))^2$. Hence, the bandwidth can be selected without specifying the observation window.

# F Modification of Diggle et al. (2007)s second order analysis

As pointed out in Diggle et al. (2007), non-parametric estimation of both first and second order properties from the same point pattern data is an ill-posed problem due to confounding between variations in the intensity function and random clustering. In case of a bivariate case-control point process and assuming the model (B.3), Diggle et al. (2007) suggested to estimate $\rho_0$ using the control points and plug in this estimate when inferring the clustering properties of the case process. This approach can be extended to the multivariate ($p > 2$) setting as follows. For each type $i = 1, \ldots, p$ we modify (B.17) to obtain the estimator

$$\hat{\rho}_{0,-i}(\mathbf{u}) = \frac{1}{p-1} \sum_{\substack{k=1 \\ k \neq i}}^{p} \sum_{\mathbf{v} \in X_k} \frac{k((\mathbf{u} - \mathbf{v})/b)}{\exp(-\hat{\boldsymbol{\beta}}_k^{\mathsf{T}} \mathbf{z}(\mathbf{v})) b^d c_b(\mathbf{v})},$$

that does not utilize the $i$th point pattern. To ease the computation time we estimate one common bandwidth $b$ for all $i = 1, \ldots, p$ by the bandwidth selection criterion described in Section E. To estimate the PCFs and cross PCFs we use the `spatstat` procedures `pcfinhom` and `crosspcfinhom`, where we specify the intensity functions by $\hat{\rho}_i(\mathbf{u}) = \hat{\rho}_{0,-i}(\mathbf{u}) \exp(\hat{\boldsymbol{\beta}}_i^{\mathsf{T}} \mathbf{z}(\mathbf{u}))$. We manually choose the bandwidth for the PCFs and cross PCFs.

# G Analysis for tumor cells

For the tumor cells we use 80% independent thinning of the points but otherwise proceed precisely as in Section 5.1 to which we refer for details. We choose Normoxic as the baseline and estimate $\boldsymbol{\beta} = (\beta_{\text{Hyp}}, \beta_{\text{Nor}})^{\mathsf{T}}$ by $\hat{\beta}_{\text{Hyp}} = \log(2346/3693) = -0.45$ and $\hat{\beta}_{\text{Nor}} = 0$. The left panel in Figure B.14

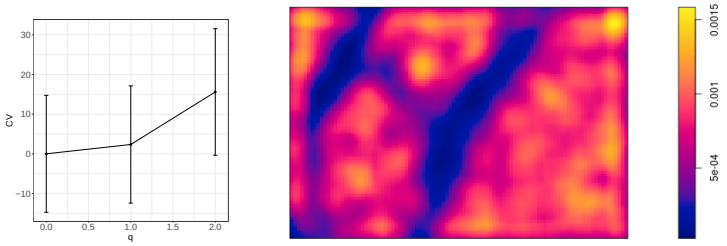


**Fig. B.14:** Left: CV-scores (minus minimum CV-score) with standard errors. Right: Non-parametric estimate of $\rho_0$ with bandwidth = 153.3.

shows that the cross validation score is minimized for $q = 0$. Hence, we model the bivariate point process as two independent LGCPs. The right panel in Figure B.14 shows the non-parametric estimate of $\rho_0$.

The parameter estimates with $q = 0$ are $\hat{\sigma}_{\text{Hyp}} = 1.45$, $\hat{\sigma}_{\text{Nor}} = 1.31$, $\hat{\varphi}_{\text{Hyp}} = 66.1$, and $\hat{\varphi}_{\text{Nor}} = 46.4$. The estimates of $\sigma_{\text{Hyp}}$ and $\sigma_{\text{Nor}}$ show that in addition to the variation caused by $\rho_0$, both the Hypoxic and Normoxic cells are highly clustered with strongest clustering for Hypoxic. This is also illustrated by the fitted PCFs in the left panel of Figure B.15.

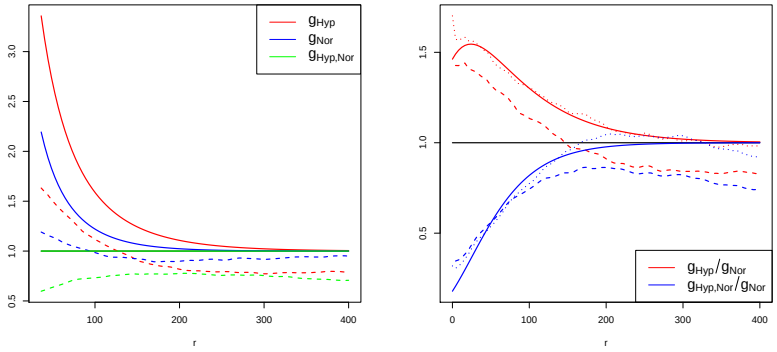


**Fig. B.15:** Left: estimated (cross) PCFs using the semi-parametric model (solid) and simple approach (dashed). Right: estimated (cross) PCF ratio using semi-parametric model (solid), simple approach (dashed) and consistent approach (dotted).

Figure B.15 shows that the agreement between the semi-parametric and consistent non-parametric estimates of cross PCF ratios $g_{\text{Hyp}}/g_{\text{Nor}}$ and $g_{\text{Hyp,Nor}}/g_{\text{Nor}}$ is very good which is confirmed by global envelope $p$-values of 0.98 in case of $g_{\text{Hyp}}/g_{\text{Nor}}$ and 0.178 for $g_{\text{Hyp,Nor}}/g_{\text{Nor}}$, see also the global envelope plots in Section H. The conclusions regarding the simple non-parametric estimates shown in Figure B.15 (left) are completely analogous to those for Stroma and CD8: the non-parametric estimates seem biased and the simple non-parametric estimates of cross PCF ratios deviate more from the consistent non-parametric estimates than the semi-parametric estimates.

# H Model assessment for lymphoma data

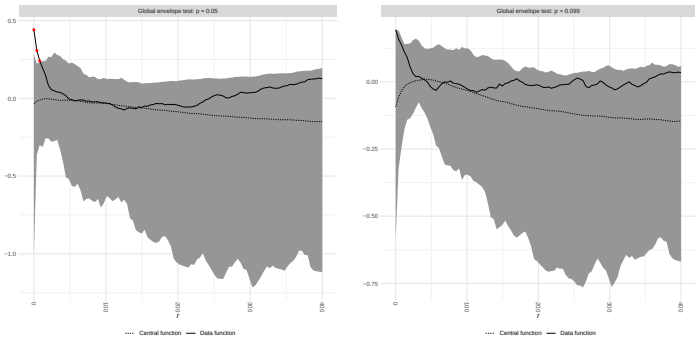


**Fig. B.16:** Differences between semi-parametric and consistent non-parametric estimates of cross PCF ratios (solid curves) with global 95% envelopes (gray shaded areas). Left: $g_{\text{Str}}/g_{\text{CD8}}$. Right: $g_{\text{Str,CD8}}/g_{\text{CD8}}$.

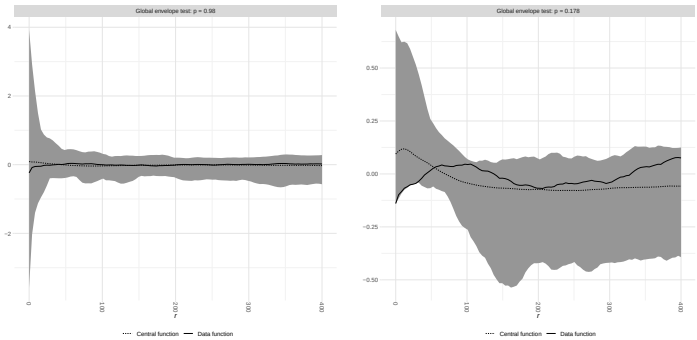


**Fig. B.17:** Differences between semi-parametric and consistent non-parametric estimates of cross PCF ratios (solid curves) with global 95% envelopes (gray shaded areas). Left: $g_{\text{Hyp}}/g_{\text{Nor}}$. Right: $g_{\text{Nor,Hyp}}/g_{\text{Nor}}$.

# I Model assessment for crime data

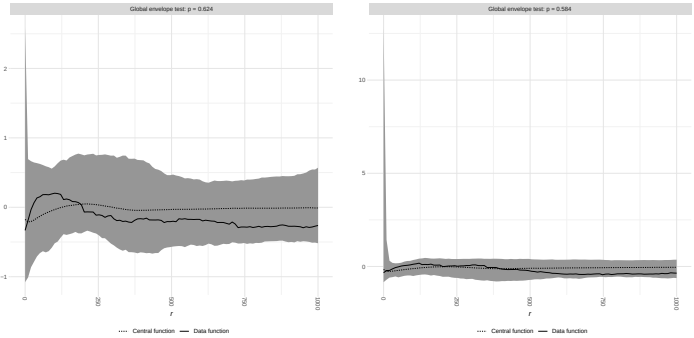


**Fig. B.18:** Differences between semi-parametric and consistent non-parametric estimates of cross PCF ratios (solid curves) with global 95% envelopes (gray shaded areas). Left: $g_1/g_6$. Right: $g_2/g_6$.

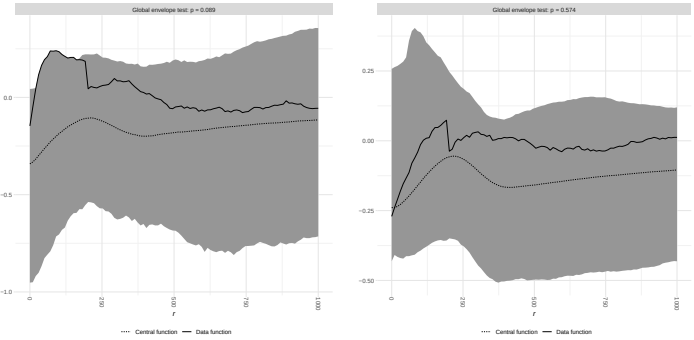


**Fig. B.19:** Differences between semi-parametric and consistent non-parametric estimates of cross PCF ratios (solid curves) with global 95% envelopes (gray shaded areas). Left: $g_{13}/g_6$. Right: $g_{16}/g_6$.

# References

Baddeley, A., Jammalamadaka, A., and Nair, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society: Series C*, 63:673–694.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54:329–350.

Choiruddin, A., Cuevas-Pacheco, F., Coeurjolly, J.-F., and Waagepetersen, R. (2019). Regularized estimation for highly multivariate log Gaussian Cox processes. *Statistics and Computing*, 30:649–662.

Cronie, O. and van Lieshout, M. N. M. (2016). Summary statistics for inhomogeneous marked point processes. *Annals of the Institute of Statistical Mathematics*, 68:905–928.

Cronie, O. and van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105:455–462.

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*, 63:550–557.

Guan, Y., Waagepetersen, R., and Beale, C. M. (2008). Second-order analysis of inhomogeneous spatial point processes with proportional intensity functions. *Journal of the American Statistical Association*, 103:769–777.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2 edition.

Henrys, P. and Brown, P. (2009). Inference for clustered inhomogeneous spatial point processes. *Biometrics*, 65:423–430.

Hessellund, K. B., Xu, G., Guan, Y., and Waagepetersen, R. (2019). Semiparametric multinomial regression for multivariate point pattern data. Under revision.

Jalilian, A., Guan, Y., Mateu, J., and Waagepetersen, R. (2015). Multivariate product-shot-noise Cox models. *Biometrics*, 71:1022–1033.

Lavancier, F., Poinas, A., and Waagepetersen, R. (2019). Adaptive estimating function inference for nonstationary determinantal point processes. *Scandinavian Journal of Statistics*. Appeared online.

Møller, J. and Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, Boca Raton.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2016). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society, Series B*, 79:381–404.

Rajala, T., Murrell, D. J., and Olhede, S. C. (2018). Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection. *Journal of the Royal Statistical Society, Series C*, 67:1237–1273.

Shaw, T., Møller, J., and Waagepetersen, R. (2020). Globally intensity-reweighted estimators for $K$- and pair correlation functions. *Australian and New Zealand Journal of Statistics*. Accepted for publication.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Annals of Applied Statistics*, 10:1019–1040.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

van Lieshout, M. N. M. (2011). A $J$-function for inhomogeneous point processes. *Statistica Neerlandica*, 65:183–201.

Waagepetersen, R. and Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Series B*, 71:685–702.

Waagepetersen, R. P., Guan, Y., Jalilian, A., and Mateu, J. (2016). Analysis of multispecies point patterns by using multivariate log-Gaussian Cox processes. *Journal of the Royal Statistical Society, Series C*, 65:77–96.

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-855-1 AALBORG UNIVERSITY PRESS