Thesis for the degree of licentiate of technology

Statistical Properties of Point Process Learning for Gibbs Processes

Julia Jansson

Department of Mathematical Sciences

Division of Applied Mathematics and Statistics

Chalmers University of Technology and the University of Gothenburg

Göteborg, Sweden 2024

Statistical Properties of Point Process Learning for Gibbs Processes
Julia Jansson

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology and the University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

Author e-mail: juljans@chalmers.se

Cover: A thinning of a point pattern, where the black points are kept and the red crosses are deleted (see Fig. 2.6). This illustrates the notion of thinning, or alternatively, a cross-validation split of a point pattern, which is central to the topic of the thesis (Point Process Learning).

Typeset with LaTeX
Printed by Chalmers Reproservice
Göteborg, Sweden 2024

Statistical Properties of Point Process Learning for Gibbs Processes

Julia Jansson

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis studies Point Process Learning (PPL), which is a novel statistical learning framework that uses point process cross-validation and point process prediction errors, and includes different hyperparameters. Specifically, statistical properties of PPL are explored, in the context of Gibbs point processes. Paper 1 demonstrates PPL's advantages over pseudolikelihood, which is a state-of-the-art parameter estimation method and a special case of Takacs-Fiksel estimation (TF), with particular focus on Gibbs hard-core processes. Paper 2 compares PPL to TF, and shows that TF is a special case of PPL, when the cross-validation scheme tends to leave-one-out cross-validation. In addition, Paper 2 shows that for four common Gibbs models, namely Poisson, hard-core, Strauss and Geyer saturation processes, one can choose hyperparameters so that PPL outperforms TF in terms of mean square error.

In Paper 1 and 2, parameter estimation with PPL is done by minimizing loss functions, while Paper 3 explores an alternative approach to PPL, namely estimating equations. Further, statistical properties of the parameter estimator are derived in Paper 3, such as consistency and asymptotic normality for large samples, as well as bias and variance for small samples. It is concluded that the estimating equation approach is not feasible for PPL, whereby the original loss function-based approach is preferred. Moving on, Paper 3 then provides a theoretical foundation for the loss functions through an empirical risk formulation.

To conclude, PPL is shown to be a flexible and robust competitor to state-of-the-art methods for parameter estimation.

Keywords: cross-validation, estimating equation, Gibbs processes, loss function, Papangelou conditional intensity, point processes, prediction error, pseudolikelihood, Takacs-Fiksel, thinning

Acknowledgements

First, I would like to thank my PhD supervisor Ottmar Cronie. You have given so much of your time and energy into supervision and I really appreciate your passion for the topic. I am also extremely grateful to Aila Särkkä for being a wonderful co-supervisor. You have also helped me immensely by guiding me on the right path when needed. Thank you also to Christophe Biscio for your mathematical guidance and for inviting me for a research visit to Aalborg. It has also been really valuable to have Konstantinos Konstantinou and Mathis Rost as PhD colleagues in the same group, thank you for all the fruitful discussions.

Thank you to the Mathematics department at Chalmers and GU for being a great workplace. I would like to acknowledge some special people from the department who have helped me in different ways during my PhD journey (as examiner, director of research school, head of unit, deputy head of department): Philip Gerlee, Johan Tykesson, Umberto Picchini, Annika Lang, Irina Petterson and Marija Cvijovic. I would also like to thank all the administrative staff for doing an excellent job and helping me with all non-mathematical issues that come up. Further, I really appreciate the community of PhD students at the department for good company. Since the list is far too long, I will focus on highlighting a few whose support have been invaluable: Robin van Haastrecht, Rickard Cullman, Ruben Seyer and Ioanna Motschan-Armen. I also want to acknowledge the visiting PhD students Verena Schwarz, Nathan Gilliot and Martin Voigt Veiling; your time here was short, but sweet!

Next, I want to thank the spatial statistics community at large for being so welcoming and for introducing me to the wonderful world of spatial statistics and point processes. Thank you especially to Giada Adelfio and Nicoletta D'Angelo for inviting me to visit Palermo and for your warm hospitality. Further, I am grateful to the developers of the `spatstat` package, especially Adrian Baddeley and Ege Rubak for kindly answering implementation-related questions.

I would like to extend my gratitude to the Åforsk foundation and SVeFUM ("Stiftelsen för Vetenskaplig Forskning och Utbildning i Matematik") for financing some of my conference related expenses.

Finally I would like to express my gratitude to all my friends and family, and I really appreciate everyone who has listened to me and supported me through tough times. Above all, I owe gratitude to my husband Henrik Jansson Valter and my parents Tünde Fülöp and Patrik Jansson. This thesis would not have been possible without your love, help and encouragement.

Julia Jansson
Göteborg, November 2024

Contents

List of Publications

This thesis consists of an extended summary and the following appended manuscripts.

Paper 1: Jansson, J., Cronie, O., Moradi M., Biscio C.A.N. *Point Process Learning: a cross-validation-based statistical framework for point processes.* in A. Pollice and P. Mariani (eds), Methodological and Applied Statistics and Demography I: SIS 2024, Short Papers, Plenary and Specialized Sessions, Vol. 52 of Proceedings of the Scientific Meeting of the Italian Statistical Society, the Italian Statistical Society, Springer, Bari, Italy. (Reproduced with permission from Springer Nature)

Paper 2: Jansson, J., Cronie, O. *Comparison of Point Process Learning and its special case Takacs-Fiksel estimation.* (2024) arXiv:2405.19523. Submitted.

Paper 3: Jansson, J., Cronie, O., Biscio C.A.N. *Two perspectives on Point Process Learning: estimating equations and empirical risk minimisation.* (2024) Manuscript.

Author contribution:

Paper 1: I did the simulations and plots and contributed in the writing.

Paper 2: I proved parts of the main results (Theorem 2 and 3), and proved Theorem 4 and all the lemmas. I was solely responsible for the simulation study and plots. Further, I did most of the writing of the paper.

Paper 3: I proved most of the results and did most of the writing of the paper.

Other publications not included in this thesis:

Paper 4: Jansson, J., Jansson, P. *Level-p-complexity of Boolean functions using thinning, memoization, and polynomials.* Journal of Functional Programming, (2023).

Paper 5: Cronie, O., **Jansson, J.**, Konstantinou, K. *Discussion of the Paper "Marked Spatial Point Processes: Current State and Extensions to Point Processes on Linear Networks".* Journal of Agricultural, Biological and Environmental Statistics (2024).

List of Notation

Point Processes

$\mathcal{S}$	Complete separable metric space
$b(u, r)$	Closed ball with radius r around location $u \in \mathcal{S}$
$d(\cdot, \cdot)$	Distance metric
$(\Omega, \mathcal{F}, \mathbb{P})$	Probability space
$\mathbf{N}$	Space of point patterns/configurations
$\mathcal{N}$	Borel σ-algebra generated by a Prohorov-type metric on $\mathbf{N}$
X	A point process $\{x_i\}_{i=1}^N$ where N is random
P_X	The distribution induced by the point process X on $(\mathbf{N}, \mathcal{N})$
$\#$	Cardinality mapping
$\lambda_X(u\|\mathbf{x})$	Papangelou conditional intensity
$\lambda_X^{(n)}(u_1, \ldots, u_n\|\mathbf{x})$	nth order Papangelou conditional intensity
$j_X^{(n)}(\cdot\|A)$	Local Janossy density with respect to a bounded $A \subseteq \mathcal{S}$
$\rho_X(u)$	Intensity function
$\rho_X^{(n)}(u_1, \ldots, u_n)$	nth order product density/intensity function
f_X	Density of P_X w.r.t. the distribution of a finite Poisson process
$\breve{X}$	A marked point process $\{(x_i, m_i)\}_{i=1}^N = \{(x_i, m(x_i))\}_{i=1}^N$
$\mathcal{M}$	Mark space, should also be a complete separable metric space
$\breve{\lambda}$	Conditional intensity of the marked point process $\breve{X}$
$p(u)$	Retention probability function in thinning
$\mathbf{1}\{\cdot\}$	Indicator function

Gibbs Processes

R	Distance parameter (hard-core, Strauss and Geyer models)
β	Intensity related parameter (hard-core, Strauss and Geyer models)
γ	Interaction parameter (Strauss and Geyer models)
s	Saturation threshold, used as a parameter in the Geyer model
$D_R(u; \mathbf{x})$	Number of points which lie within a distance R of the location u

Inference

θ_0	True parameter generating the point process X
$\hat{\theta}$	Estimator of the true parameter θ_0
h	Test function
$e_h(\theta)$	Estimating equation for TF, or h-innovation
$S_h(\theta)$	Sensitivity matrix for the estimating equation $e_h(\theta)$
$\Sigma_h(\theta)$	Covariance matrix for the estimating equation $e_h(\theta)$
G_h	Godambe information
Σ_{hg}	Covariance matrix $Cov(e_h(\theta), e_g(\theta))$

Point Process Learning

X^T	Training set
X^V	Validation set
$\mathcal{I}_\theta^h(A; X^T, X^V)$	Point process prediction error
$V_\theta(u, X^T, X^V)$	PPL-weight
k	Number of cross-validation splits
$\mathcal{T}_k$	Index set of non-empty cross-validation splits
$\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$	Loss functions
$I_{h,k}(\theta, \mathbf{x})$	Mean of prediction errors over cross-validation splits
$\widehat{\theta}_{h,k}(\mathbf{x})$	Estimator solving (4.5)
$\mathcal{D}I_{h,k}(\theta, \mathbf{x})$	Jacobian matrix of $I_{h,k}(\theta, \mathbf{x})$, w.r.t. θ
$S_{h,k}(\theta)$	Sensitivity matrix for $I_{h,k}(\theta, \mathbf{x})$
$\Sigma_{h,k}(\theta)$	Covariance matrix for $I_{h,k}(\theta, \mathbf{x})$
$G_{h,k}$	Godambe information
$\Sigma_{(h,k),(g,k)}(\theta_0)$	Covariance matrix $Cov(I_{h,k}(\theta_0), I_{g,k}(\theta_0))$
Φ_{hg}	Difference between $\Sigma_{(h,k),(g,k)}(\theta_0)$ and $\Sigma_{h,k}(\theta_0)$

List of Abbreviations

i.i.d.	Independent and identically distributed
a.s.	Almost sure(ly)
a.e.	Almost everywhere
GNZ formula	Formula named after Georgii, Nguyen and Zessin, see (2.1)
CV	Cross-validation
MCMC	Markov Chain Monte Carlo
MLE	Maximum likelihood estimation
MSE	Mean square error
PPL	Point Process Learning
TF	Takacs-Fiksel estimation
PL	Pseudolikelihood estimation
`R`	R programming language [R Core Team, 2020]
`spatstat`	Package in `R` for spatial statistics [Baddeley et al., 2015]
`ppm`	Function for parameter estimation in `spatstat`
2D	Two-dimensional

List of Figures

Chapter 1

Introduction

In classical statistical inference and statistical learning, there is an underlying assumption that the data come from independent and identically distributed (i.i.d.) random variables. In the field of point pattern analysis, this assumption is often violated, making it difficult to apply classical statistical inference/learning methods directly. Examples of spatio-temporal data where the i.i.d. assumption is violated include ambulance calls [Bayisa et al., 2023], nerve fibers [Konstantinou et al., 2023] and seismic waveforms for earthquakes [D'Angelo et al., 2023]. Here, it is more suitable to consider a *generalised random sample*, which allows dependency among the variables and the sample size to be random. Such a generalised random sample is precisely what constitutes a point process. See Figure 1.1 for two examples of point patterns which can be modelled with point processes. The aim of this licentiate thesis is to explore a recent approach to statistical learning in the context of point processes, based on work by Cronie et al. [2024b].

Gibbs point processes are flexible and natural for modelling point patterns with dependence between the points. Such processes are defined either by a density with respect to a Poisson point process or by the so-called Papangelou conditional intensity.

While maximum likelihood estimation is an established method for parameter estimation, it requires computationally intensive estimation of an unknown normalising constant, the so-called partition function which appears in the density functions of Gibbs point processes [Møller and Waagepetersen, 2004, Section 8.4]. Therefore, alternative estimation methods based on the Papangelou conditional intensity are more suitable. The state-of-the art is Takacs-Fiksel estimation (TF) [Takacs and Fiksel, 1986], which has pseudolikelihood estimation (PL) as a special case [Coupier, 2019]. For practical applications of Gibbs point processes, PL, or its approximation logistic regression likelihood estimation [Baddeley et al., 2014], is

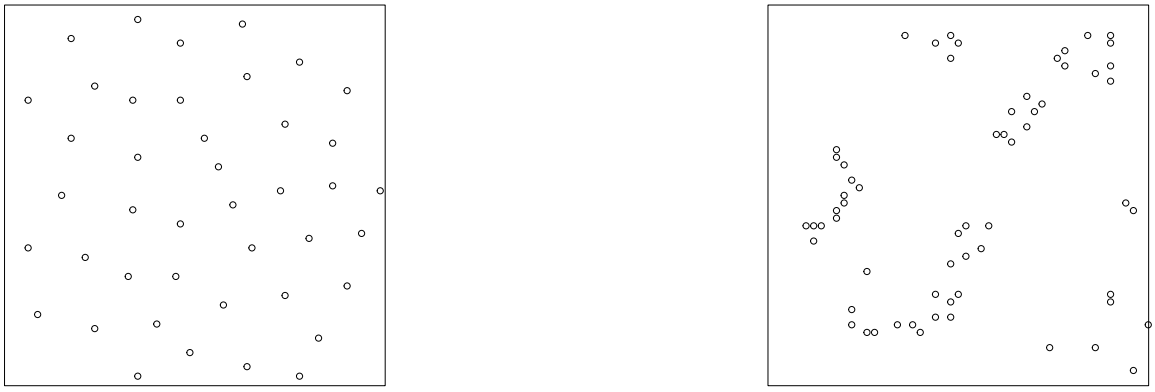

Figure 1.1: Examples of point patterns. Left: Locations of the centres of 42 biological cells observed under optical microscopy [Ripley, 1977]. Right: Locations of 62 seedlings and saplings of California Giant Redwood (Sequoiadendron giganteum) [Ripley, 1977]

the most commonly used method, partly because it is the default in the R package spatstat [Baddeley et al., 2015]. Even though PL has been widely used, it has its disadvantages. First of all, it is not necessarily the optimal special case of TF [Coeurjolly et al., 2016]. Moreover, it has poor performance when there are strong interactions present [Baddeley et al., 2015]. In addition, as shown in Paper 1, it suffers from identifiability issues, even in the context of rather basic models.

In this licentiate thesis, we study a new method for parameter estimation for Gibbs processes called Point Process Learning (PPL), which was recently introduced by Cronie et al. [2024b]. Specifically, PPL is a prediction-based statistical theory for point processes which utilises a parametrised Papangelou conditional intensity. Inspired by statistical learning, PPL is based on the combination of two novel concepts for point processes: cross-validation and prediction errors. The cross-validation approach uses thinning to split a point process/pattern into pairs of training and validation sets, while the prediction errors measure discrepancy between two point processes.

According to Cronie et al. [2024b], PPL outperforms the state-of-the-art in kernel intensity estimation, i.e. the Cronie and van Lieshout [2018] approach. Further, Paper 1 shows that PPL outperforms PL for the Gibbs hard-core process, and Paper 2 shows that it is possible to choose hyperparameters so that PPL outperforms TF for four common Gibbs models: Poisson, hard-core, Strauss and Geyer saturation processes. Notably, TF is shown to be a special case of PPL in Paper 2.

Cronie et al. [2024b] show that the expectation of the prediction errors is zero if and only if the so-called PPL-weight is of a certain form. In Paper 2, we provide general expressions for the PPL-weight for Gibbs models. Specifically, we investigate the

weight expressions for Poisson, hard-core, Strauss and Geyer saturation processes. For the Poisson process, the weight takes a simple form but for the other models, the weight is intractable. Therefore we discuss different practical choices for the weight.

In Cronie et al. [2024b] and Paper 1 and 2, parameter estimation with PPL includes minimising loss functions. This is done since the loss functions approximate the expectation of a prediction error, which we know is zero in the true parameter. Hence, by minimising the loss functions, we aim to get a good parameter estimate. Solving estimating equations is another type of parameter estimation which is well-established for other parameter estimation methods, such as TF [Coeurjolly et al., 2016]. Following the lines of Guan et al. [2015] and Coeurjolly et al. [2016], Paper 3 explores the estimating equation approach for PPL. Furthermore, Paper 3 presents an empirical risk formulation of PPL, and some additional statistical properties of PPL are derived.

The thesis is structured as follows. In Chapter 2, we recall necessary background about point processes. In Chapter 3, we focus on Gibbs point processes, introducing the models mentioned earlier, and recalling methods used for simulation and inference for Gibbs models. Chapter 4 describes the new statistical methodology, PPL. Lastly, in Chapter 5, a summary of the papers included in this licentiate thesis is given, and in Chapter 6, conclusions and future work are discussed.

1.1 Aims

The overarching aim of this thesis is to study PPL, by exploring properties of the method and comparing it to other state-of-the-art methods in the field. More specifically, the aims of Papers 1-3 are presented in the following list:

1. Compare PPL to PL for the Gibbs hard-core process.
2. Study the relationship between PPL and TF, by exploring asymptotic properties of PPL and comparing PPL to TF for some Gibbs models.
3. Explore two different perspectives on PPL, by attempting to set up PPL within an estimating equation approach and providing an empirical risk minimisation formulation of PPL.

Chapter 2

Point Processes

This chapter provides an introduction to theoretical concepts for point processes used throughout the thesis.

2.1 Definition of a Point Process

A point process generates point patterns according to some distribution. Recall Figure 1.1 where we have seen two examples of point patterns. Note that a point pattern can be spatial (e.g. points in 2D), temporal (points at different times), or spatio-temporal (e.g. ambulance calls recorded with both location and time, see Moradi [2018]). To be able to define what a point process is, we therefore start with the underlying space $\mathcal{S}$ where the points are located.

Let $\mathcal{S}$ be a general space, like a compact subset of $\mathbb{R}^d$, a sphere or a linear network (see Figure 2.1). In applications, it is common to consider a bounded subset W of $\mathcal{S}$, called the observation window; see Figure 1.1 for two examples of point patterns observed in a square window. In the simulation studies of Paper 1 and Paper 2, $\mathcal{S}$ is given by a bounded subset of $\mathbb{R}^2$. Formally, we require $\mathcal{S}$ to be a complete separable metric space with a distance metric $d(\cdot, \cdot)$, which is equipped with a suitable reference measure

$$A \mapsto |A| = \int_A \mathrm{d}x, \qquad A \subseteq \mathcal{S}.$$

Here, we require our reference measure $|\cdot|$ to be σ-finite and locally finite. Further, all sets considered are Borel sets. Throughout the thesis, a closed ball with radius r around $u \in \mathcal{S}$ will be denoted by $b(u, r) = \{v \in \mathcal{S} : d(u, v) \leq r\}$.

When rolling a die, a random variable can describe the possible outcomes, such as rolling a six. If the possible outcomes of a random variable live on a more general

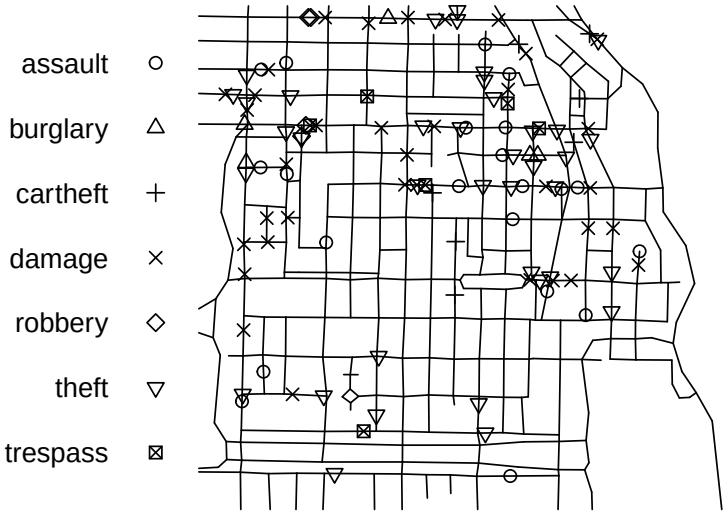

Figure 2.1: An example of a marked point pattern observed on a linear network. Spatial locations of crimes reported in the period April 25 to May 8 2002, in an area of Chicago (Illinois, USA) close to the University of Chicago [Ang et al., 2012].

space than the natural numbers, such as for example $\mathcal{S}$, we call the random variable a random element. A point process $X = \{x_i\}_{i=1}^N$ in $\mathcal{S}$ is a random element whose outcomes are collections of points, so-called point patterns. Recall that a random sample is a sequence of independent and identically distributed (i.i.d.) random variables $x_1, \ldots, x_n$, where the sample size n is fixed. Hence, we may view a point process X as a generalisation of a classical random sample, where we allow the sample size N to be random and the sample points x_i to be dependent random variables.

Formally, $X = \{x_i\}_{i=1}^N$, $0 \leq N \leq \infty$, is defined as a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(\mathbf{N}, \mathcal{N})$ [Chiu et al., 2013, Section 4.1.1]. Here, $\mathbf{N}$ is the collection of point patterns/configurations $\mathbf{x} = \{x\}_{i=1}^n \subseteq \mathcal{S}$, $0 \leq n \leq \infty$. We assume that they are locally finite, i.e. satisfying that the cardinality $\#(\mathbf{x} \cap A)$ is almost surely (a.s.) finite for any bounded $A \subseteq \mathcal{S}$. Here $\mathcal{N}$ is the smallest σ-algebra such that for all bounded Borel sets $A \subseteq \mathcal{S}$ the mapping $\mathbf{x} \to \#(\mathbf{x} \cap A)$ is measurable [van Lieshout, 2000]. Throughout, it is also assumed that the point process X is simple which means that a.s. no two points of X have the same location. Moreover, we refer to X as a finite point process if $\#(X \cap \mathcal{S}) < \infty$ a.s.. Hence, if $\mathcal{S}$ is a bounded set then X is automatically a finite point process, due to the local finiteness.

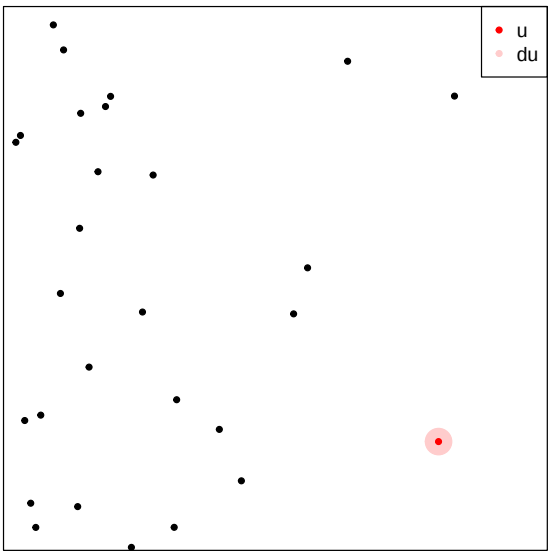

Figure 2.2: A point pattern $\mathbf{x}$ with black points, and a red point called u. Here du is an infinitesimal region around the point u, shaded in red. This image illustrates how the conditional intensity $\lambda_X(u|\mathbf{x})$ is defined. Here $\lambda_X(u|\mathbf{x})\mathrm{d}u$ is interpreted as the probability of finding a point in an infinitesimal region du around $u \in \mathcal{S}$, given that the point process agrees with the configuration $\mathbf{x}$ outside du.

2.2 Distributional Properties

Now we introduce the distribution of a point process, first in terms of the conditional intensity, then we define Janossy densities, as well as the intensity and classical density function. The point process X induces a distribution P_X on $(\mathbf{N}, \mathcal{N})$, which is governed by its finite dimensional distributions, and the distribution of a point process X is completely characterised by its (Papangelou) conditional intensity λ_X. It can be interpreted as follows: $\lambda_X(u|\mathbf{x})\mathrm{d}u$ is the probability of finding a point in an infinitesimal region du around $u \in \mathcal{S}$, given that the point process agrees with the configuration $\mathbf{x}$ outside du [van Lieshout, 2000]; see Figure 2.2 for an illustration. Conditional intensities have a central role in the study of point processes, and they are very useful for describing local interactions in point processes, most notably Gibbs point processes; see Chapter 3.

One way to define the conditional intensity $\lambda : \mathcal{S} \times \mathbf{N} \to [0, \infty)$ is through the Georgii-Nguyen-Zessin (GNZ) [Georgii, 1976, Nguyen and Zessin, 1979] formula, which states that

$$\mathbb{E}\left[\sum_{x \in X} h(x, X \setminus \{x\})\right] = \int_{\mathcal{S}} \mathbb{E}[h(u, X)\lambda_X(u|X)]\mathrm{d}u \tag{2.1}$$

for any non-negative (potentially infinite) measurable function h on $\mathcal{S} \times \mathbf{N}$. Note that this means that (2.1) also holds for any integrable function h. Here, it should

be emphasised that the expectation is defined as

$$\mathbb{E}[g(X)] = \int_{\mathbf{N}} g(\mathbf{x}) P_X(d\mathbf{x})$$

for $g : \mathbf{N} \to \mathbb{R}$. Note that if we replace $h(x, X \setminus \{x\})$ by $h(x, X)$ on the left-hand side of (2.1), we have to replace $h(u, X)$ by $h(u, X \cup \{u\})$ on its right-hand side [Betsch, 2023].

A model, or the point process X it generates, is called *attractive* if $\lambda_X(u|\mathbf{x}) \leq \lambda_X(u|\mathbf{y})$ and *repulsive* if $\lambda_X(u|\mathbf{x}) \geq \lambda_X(u|\mathbf{y})$, whenever $\mathbf{x} \subseteq \mathbf{y}$ [Møller and Waagepetersen, 2004, Section 6.1.1].

It is also possible to define higher-order conditional intensities. For any $n \geq 1$ and $u_1 \ldots, u_n \in \mathcal{S}$, we consider

$$\lambda_X^{(n)}(u_1, \ldots, u_n|\mathbf{x}) = \lambda_X(u_1|\mathbf{x})\lambda_X(u_2|\mathbf{x} \cup \{u_1\}) \cdots \lambda_X(u_n|\mathbf{x} \cup \{u_1, \ldots, u_{n-1}\}),$$

which is the nth order conditional intensity function of X. Let $X^n_{\neq} = \{(x_1, \ldots, x_n) : x_1, \ldots, x_n \in X, x_i \neq x_j \text{ if } i \neq j\}$ be the point process on $\mathcal{S}^n$ consisting of all distinct n-tuples of elements of X. Then, the nth order conditional intensity function also satisfies the GNZ formula

$$\begin{aligned}&\mathbb{E}\left[\sum_{(x_1,\ldots,x_n)\in X^n_{\neq}} h(x_1, \ldots, x_n, X \setminus \{x_1, \ldots, x_n\})\right] \\ &= \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} \mathbb{E}[h(u_1 \ldots, u_n, X)\lambda_X^{(n)}(u_1 \ldots, u_n|X)]\mathrm{d}u_1 \ldots \mathrm{d}u_n,\end{aligned}$$

where $h : \mathcal{S}^n \times \mathbf{N} \to \mathbb{R}$ [Coeurjolly et al., 2017]. Note that $\lambda_X^{(1)}(u|\mathbf{x}) = \lambda_X(u|\mathbf{x})$.

2.2.1 Janossy Densities

Next, we introduce the so-called local Janossy densities $j_X^{(n)}(\cdot|A)$ of X, $n \geq 0$, with respect to a bounded $A \subseteq \mathcal{S}$ [Daley and Vere-Jones, 2008]. The first-order local Janossy density $j_X^{(1)}(\{u\}|A)\mathrm{d}u$ gives the probability that $X \cap A$ has exactly one point located in the infinitesimal region du of $u \in \mathcal{S}$. The Janossy densities satisfy [Betsch, 2023, Lemma 3.1]

$$\begin{aligned}&j_X^{(n)}(\mathbf{x}|A) = j_X^{(n)}(\{x_1, \ldots, x_n\}|A) = \\ &\frac{1}{n!}\mathbf{1}\{x_1, \ldots, x_n \in A\}\mathbb{E}[\mathbf{1}\{\#(X \cap A) = 0\}\lambda_X^{(n)}(x_1, \ldots, x_n|X)] \qquad (2.2)\end{aligned}$$

for $\mathbf{x} \in \mathbf{N}$. Note that $j_X^{(0)}$ represents the probability that $X = \emptyset$ and that $j_X^{(n)}(\{u_1, \ldots, u_n\}|A)\mathrm{d}u_1 \cdots \mathrm{d}u_n$ gives the probability that $X \cap A$ has exactly n points which are located in the infinitesimal neighbourhoods $du_1, \ldots, du_n$ of $u_1, \ldots, u_n \in \mathcal{S}$.

We remark that obtaining Janossy densities from conditional intensities, as we do here, is the reverse of the classical setup [Daley and Vere-Jones, 2008], where one starts with the Janossy densities and shows that they yield the conditional intensity as a ratio. More specifically, for finite point processes it holds that [van Lieshout, 2000, Møller and Waagepetersen, 2004, Daley and Vere-Jones, 2008]

$$\lambda_X(u|\mathbf{x}) = \frac{j_X((\mathbf{x} \setminus \{u\}) \cup \{u\})}{j_X(\mathbf{x} \setminus \{u\})}, \qquad u \in \mathcal{S}. \tag{2.3}$$

Note here the shortened notation

$$j_X(\mathbf{y}) = j_X(\mathbf{y}|\mathcal{S}) = \sum_{n=0}^{\infty} \mathbf{1}\{\#(\mathbf{y} \cap \mathcal{S}) = n\} j_X^{(n)}(\mathbf{y}), \quad \mathbf{y} \in \mathbf{N}. \tag{2.4}$$

2.2.2 Intensity and Density Function

We now introduce the intensity function $\rho_X(u)$ of X, which reflects the infinitesimal probability of finding a point of X in the distinct location $u \in \mathcal{S}$. For two-dimensional point patterns, the intensity can be visualised as a colormap which shows how likely it is to find a point in different parts of the window, see Figure 2.3. The nth order intensity function of X, which is the intensity function of $X^n_{\neq}$, is given by

$$\rho_X^{(n)}(u_1, \ldots, u_n) = \mathbb{E}[\lambda_X^{(n)}(u_1, \ldots, u_n|X)],$$

where $\rho_X^{(1)}(u) = \rho_X(u)$.

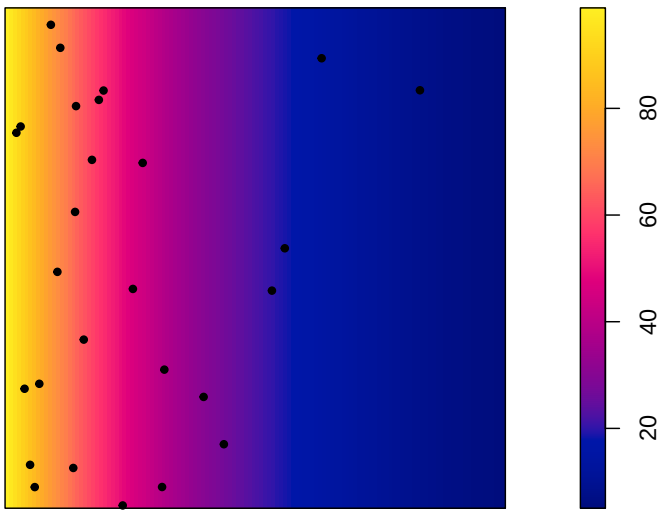

Figure 2.3: A realisation of an inhomogeneous Poisson process with intensity $\rho(u) = 100\exp(-3u_x)$ where $u = (u_x, u_y) \in \mathbb{R}^2$. The intensity is shown in the background with a colour gradient.

Lastly, we introduce the classical density f_X. Let us consider a finite point process X on a bounded domain $\mathcal{S}$, with Janossy densities $j_X^{(n)}(\cdot) = j_X^{(n)}(\cdot|\mathcal{S})$, $n \geq 0$. Then the density function f_X is defined as [Daley and Vere-Jones, 2008, Exercise 10.4.1]

$$f_X(\mathbf{x}) = \frac{\mathrm{e}^{\int_A \rho_0(u)\mathrm{d}u}}{\prod_{x\in\mathbf{x}} \rho_0(x)} j_X(\mathbf{x}), \qquad \mathbf{x} \in \mathbf{N}, \tag{2.5}$$

where $j_X(\mathbf{x})$ is as in (2.4). This is the density of P_X with respect to the distribution of a finite Poisson process with intensity function ρ_0 on A; see Section 2.3 for a definition of the Poisson process. As can be seen in (2.5), f_X is simply a scaling of j_X. Similarly to (2.3) it holds that

$$\lambda_X(u|\mathbf{x}) = \frac{f_X((\mathbf{x} \setminus \{u\}) \cup \{u\})}{f_X(\mathbf{x} \setminus \{u\})}, \qquad u \in \mathcal{S}. \tag{2.6}$$

2.3 Poisson Processes

Now we can give a first example of a point process; the Poisson process.

Definition 2.3.1 (Poisson distribution)**.** *A discrete random variable Y is said to have a Poisson distribution, with parameter $\lambda > 0$ if its probability mass function is given by*

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, \ldots.$$

Then we write $Y \sim Poi(\lambda)$.

Definition 2.3.2 (Poisson process)**.** *Let $\mu(A) = \int_A \rho(u)\mathrm{d}u$, $A \subseteq \mathcal{S}$. A Poisson process X on $\mathcal{S}$ with intensity function ρ is a point process satisfying [Møller and Waagepetersen, 2004, van Lieshout, 2000]:*

1. *$\#(X \cap A) \sim Poi(\mu(A))$ for every bounded set $A \subset \mathcal{S}$,*
2. *for any k disjoint bounded sets $A_1, \ldots, A_k \subset \mathcal{S}$, the random variables $\#(X \cap A_1), \ldots, \#(X \cap A_k)$ are independent.*

If the intensity function ρ is constant, the process is called a *homogeneous* Poisson process with rate or intensity ρ, otherwise it is said to be an *inhomogeneous* Poisson process [Møller and Waagepetersen, 2004]. In Figure 2.3 a realisation of an inhomogeneous Poisson process with intensity $\rho(u) = 100\exp(-3u_x)$ where $u = (u_x, u_y) \in \mathbb{R}^2$, is illustrated. See Figure 2.4 for a realisation of a homogeneous Poisson process with intensity 100. The homogeneous Poisson process describes the property of Complete Spatial Randomness (CSR) which means that the points in a point pattern are uniformly and independently distributed in space. Note that when we consider a Poisson process with unit rate, that is a homogeneous point process with $\rho = 1$, (2.5) turns into $f_X(\mathbf{x}) = \mathrm{e}^{|A|}\, j_X(\mathbf{x})$.

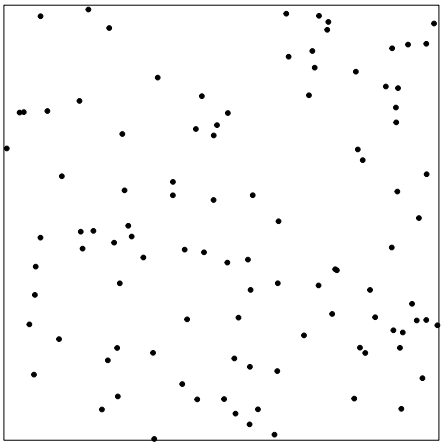

Figure 2.4: A realisation of a homogeneous Poisson process with intensity 100.

For a Poisson process it holds that [Mecke, 1967]

$$\mathbb{E}\left[\sum_{u\in X} h(u, X\setminus\{u\})\right] = \int_{\mathcal{S}} \mathbb{E}[h(u,X)\rho(u)]\mathrm{d}u, \tag{2.7}$$

where $h : \mathcal{S}\times\mathbf{N}\to[0,\infty)$. Note the similarity to the GNZ formula (2.1), where we have the conditional intensity instead of the intensity. We observe that for the Poisson process, the conditional intensity is the same as the intensity. We see this by using (2.1) in the first equality and (2.7) in the second equality:

$$\int_{\mathcal{S}} \mathbb{E}[h(u,X)\lambda(u|X)]\mathrm{d}u = \mathbb{E}\left[\sum_{u\in X} h(u,X\setminus\{u\})\right] = \int_{\mathcal{S}} \mathbb{E}[h(u,X)\rho(u)]\mathrm{d}u.$$

Setting $h(u,\cdot) = \mathbf{1}\{u\in A\}$ for any bounded $A\subseteq\mathcal{S}$, we now get $\int_A \mathbb{E}[\lambda(u|X)]\mathrm{d}u = \int_A \mathbb{E}[\rho(u)]\mathrm{d}u$ which means that $\rho_X(u) = \lambda_X(u|X)$ a.s. for almost every u.

2.4 Operations on Point Processes

In this section we will introduce operations on point processes, like marking, thinning and cross-validation. Marking means that we start with a point process and attach additional components, called marks, to all points. Two examples of marked point patterns can be seen in Figure 2.1 and 2.5.

Formally, a marked point process is a point process $\breve{X} = \{(x_i, m_i)\}_{i=1}^N$ on $\mathcal{S}\times\mathcal{M}$, i.e. a random element in the space of point patterns in $\mathcal{S}\times\mathcal{M}$, with the additional property that its projection onto $\mathcal{S}$, its so-called ground process $X = \{x_i\}_{i=1}^N$, is a well-defined point process in $\mathcal{S}$ [van Lieshout, 2000]. Here, $\mathcal{M}$ is called the mark space, and is required to be a complete separable metric space, like $\mathcal{S}$. One example is a binary mark space as in Figure 2.5 where the cells have an "on"

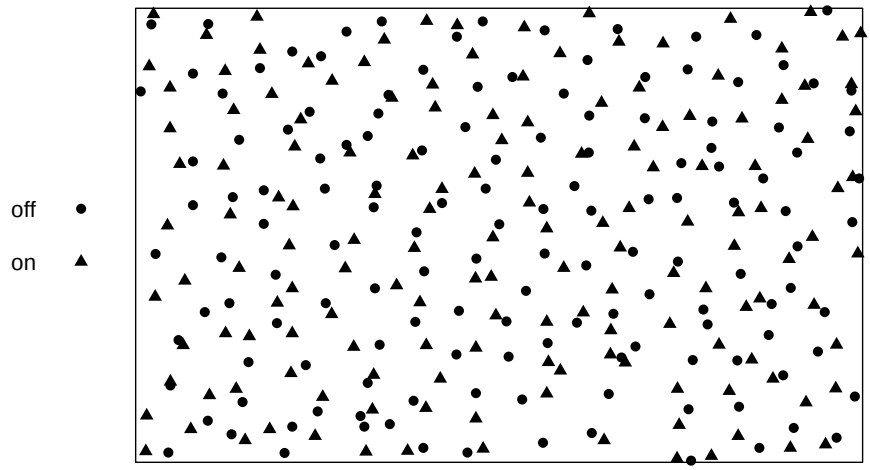

Figure 2.5: Displaced amacrine cells in the retina of a rabbit, with 152 "on" cells and 142 "off" cells in a rectangular sampling frame [Diggle, 1986].

and "off" property. In Figure 2.1, a discrete mark space can be seen, with seven types of crimes. Other examples include continuous marks, function-valued marks, graph-valued marks and random sets [Cronie et al., 2024a, Eckardt et al., 2024, Eckardt and Moradi, 2024]. See van Lieshout [2000], Daley and Vere-Jones [2008] for theory for general mark spaces.

A realisation of a marked point process is called a marked point pattern, given by an unordered set $\breve{\mathbf{x}} = \{(x_i, m_i)\}_{i=1}^n$, where $\mathbf{x} = \{x\}_{i=1}^n$ is a point pattern on $\mathcal{S}$ and m_i, $i = 1, \ldots, n$, the corresponding marks. The marked process $\breve{X}$ has the conditional intensity $\breve{\lambda} : (\mathcal{S} \times \mathcal{M}) \times \breve{\mathbf{N}} \to [0, \infty)$, where $\breve{\mathbf{N}}$ denotes the collection of marked point patterns $\breve{\mathbf{x}}$. Note that $\breve{\lambda}$ satisfies the GNZ formula, which is defined as in the unmarked case, since $\breve{X}$ is also a point process.

2.4.1 Thinning

Thinning is achieved by applying some mechanism to a point process/pattern so that each point is retained or deleted with a certain probability. Thinning can also be seen as binary marking where 0 corresponds to deleting the point and 1 corresponds to keeping the point [Cronie et al., 2024b]. Formally, we set the mark space to $\mathcal{M} = \{0, 1\}$, let the reference measure on $\mathcal{M}$ be the counting measure and let $\breve{X}$ be a binary marking of X. Then the thinning of X is defined as the point process $X^V = \{x : (x, m) \in \breve{X}, m = 1\} \subseteq X$. We let the point process with deleted points be denoted by $X^T = X \setminus X^V$. The dual operation of thinning is superpositioning; from X^V and X^T we can construct the superposition X.

In independent thinning each point $x \in \mathbf{x}$ is deleted (attached mark 0) with prob-

ability $1 - p(x)$, independently of all other deletions. Here $p(\cdot)$ is the associated retention probability function defined as $p(u) = \mathbb{P}(m = 1)$, $(u, m) \in \mathcal{S} \times \mathcal{M}$. In particular, in p-thinning, we have constant $p(u) = p \in (0, 1)$, $u \in \mathcal{S}$. This corresponds to an independent binary marking according to a Bernoulli distribution with parameter p. A p-thinning of a point pattern is illustrated in Figure 2.6.

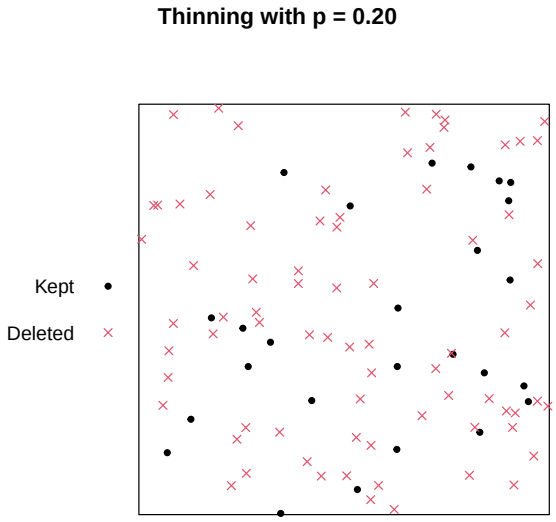

Figure 2.6: The point pattern in Figure 2.4 thinned with p-thinning using $p = 0.2$.

When X^V is an independent thinning of a Poisson process X, X^V is also a Poisson process, but with intensity $\rho_{X^V}(\cdot) = p(\cdot)\rho_X(\cdot)$ [Chiu et al., 2013, p. 161]. Returning to Figure 2.6 we see that the thinned process is also a Poisson process, but with intensity $0.2 \cdot 100 = 20$.

2.4.2 Cross-validation

Cross-validation (CV) generally involves partitioning a dataset into a training set and a validation set [Arlot and Celisse, 2010]. Cronie et al. [2024b] define CV for point processes as follows.

Definition 2.4.1 (Point process cross-validation)**.** *Given a point process $X \subseteq S$ and $k \geq 1$ thinnings $X_1^V, \ldots, X_k^V$ of X, we refer to the collection of pairs $(X_i^T, X_i^V) = (Y_i, Z_i)$, $X_i^T = X \setminus X_i^V$, $i = 1, \ldots, k$, as a cross-validation splitting/partitioning. Analogously, for a point pattern $\mathbf{x}$, we may generate thinnings $(\mathbf{x}_i^T, \mathbf{x}_i^V)$, $i = 1, \ldots, k$.*

A common CV method in classical statistics is k-fold CV, of which leave-one-out CV, where k is given by the sample size n, is a special case [Arlot and Celisse, 2010]. For point processes, k-fold CV is mathematically intractable since it is a dependent thinning-generated splitting scheme [Cronie et al., 2024b]. An example

of an independent thinning-based CV method is Monte-Carlo CV, where all points $x \in \mathbf{x}$ are assigned to $\mathbf{x}_i^V$ with a fixed common probability $p \in (0,1)$, independently of the other assignments, and the remaining points are assigned to $\mathbf{x}_i^T$. Note that each split $(\mathbf{x}_i^T, \mathbf{x}_i^V)$ is generated from the point pattern $\mathbf{x}$ according to this procedure, which means that the splits need not be disjoint, i.e. we may have that $\mathbf{x}_i^V \cap \mathbf{x}_j^V \neq \emptyset$ if $i \neq j$. An example of independent thinning-based CV with disjoint splits is block CV, where we partition the space $\mathcal{S}$ into disjoint regions $\mathcal{S}_i$. We independently attach marks m to all $x \in \mathbf{x}$, where the mark distribution is given by a multinomial distribution with $p(x) = P(m = i) = \mathbf{1}\{x \in \mathcal{S}_i\}$ for $(x, m) \in \mathcal{S} \times \mathcal{M}$, $i \in \mathcal{M}$ and $\mathcal{M} = \{1, \ldots, k\}$. This means that each point in $\mathcal{S}_i$ gets the mark i. We then let $\mathbf{x}_i^V = \{x \in \mathbf{x} : m = i\}$ for $(x, m) \in \mathcal{S} \times \mathcal{M}$ and $\mathbf{x}_i^T = \mathbf{x} \setminus \mathbf{x}_i^V$, $i = 1, \ldots, k$. For further point process CV methods, see [Cronie et al., 2024b] and Paper 2.

Chapter 3

Gibbs Processes

The theory of Gibbs measures originates from statistical physics, where one aims to describe the probability of the different states in an interacting particle system, for example the Ising model which is visualised in Figure 3.1. Heuristically, the probability of being in state x for an interacting particle system is described as

$$P(X = x) = \frac{1}{Z} e^{-\beta H(x)}, \tag{3.1}$$

where $H(x)$ denotes the so-called energy function of the system, Z is a normalising constant, and β is a parameter describing the "temperature" of the system.

Carrying this concept over to point processes we consider the points to interact with each other, similarly to the particles in statistical physics. However, instead of being on a discrete space like a lattice, as seen in Figure 3.1, the points are now in continuous space. Gibbs point processes constitute a large class of point processes, and can be applied in forestry [Ripley, 1977, Särkkä, 1995, Eckel et al., 2009], medicine [Ripley, 1977, Iftimi et al., 2018], and materials science [Häbel et al., 2019], just to name a few examples.

The first part of this chapter defines a Gibbs process and gives some examples of Gibbs processes: the hard-core process, the Strauss process and the Geyer saturation process. These three processes will be used later in the simulation study in Paper 1 and 2. Next, simulation and inference for Gibbs processes is described.

3.1 Gibbs Point Process Models

We can define the density of a Gibbs process similarly as in (3.1), but the Janossy density has an intractable normalising constant, whereby exact likelihood inference

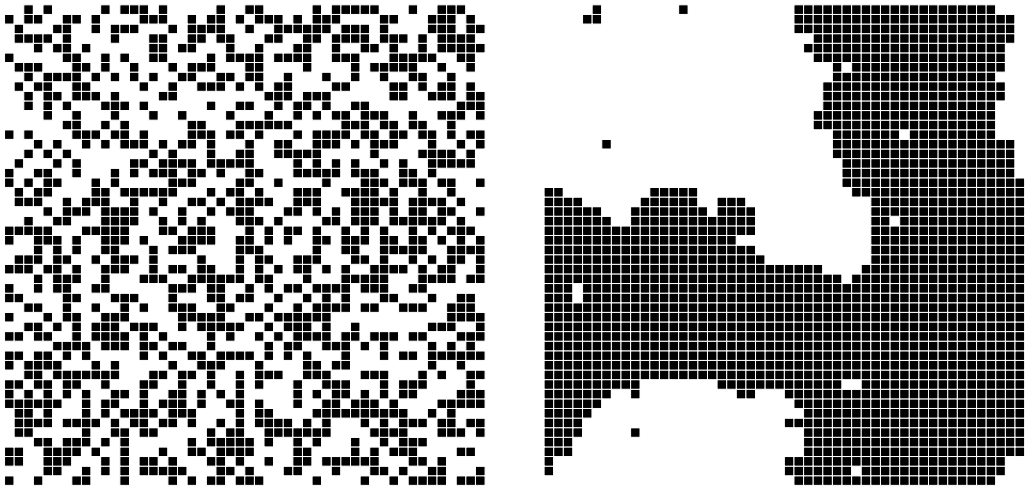

Figure 3.1: The Ising model on a 50×50 square lattice, where up and down spins are visualised with white and black squares, respectively. Left: An initial configuration, where the up and down spins are assigned at random and independent of each other. Right: The model after 100 steps of the Metropolis-Hastings algorithm, simulating the Ising model dynamics with $\beta = 0.67$. As we can see the system moves towards an equilibrium.

is typically not feasible, which makes it more practical to consider Gibbs processes in terms of their conditional intensity. We thus follow Betsch [2023] and define a Gibbs process as any point process with existing conditional intensity. Nguyen and Zessin [1979] showed that a Gibbs process may be characterised in this way, by the GNZ formula. Often it is convenient to express the conditional intensity of a Gibbs process in exponential form

$$\lambda_X(u; \mathbf{x}) = \mathrm{e}^{\Phi_1(u)+\Phi_2(u,\mathbf{x})} = \widetilde{\rho}(u)\,\mathrm{e}^{\Phi_2(u,\mathbf{x})}, \tag{3.2}$$

for functions Φ_1 and Φ_2. Note that this is possible since the conditional intensity λ takes non-negative values. Therefore, it is common to refer to a Gibbs process as a point process with conditional intensity in exponential form, see e.g. Dereudre [2019, Theorem 2]. In (3.2), the term $\widetilde{\rho}(u) = e^{\Phi_1(u)}$ only depends on the spatial location, $u \in \mathcal{S}$, but not on the other points in the point pattern $\mathbf{x}$. The Poisson process is an example of a Gibbs process with no inter-point interaction, i.e. $\Phi_2(\cdot) \equiv 0$ so then $\rho_X(u) = \lambda_X(u|\mathbf{x}) = \widetilde{\rho}(u)$.

3.1.1 Hard-core Models

The first example of a Gibbs process is the hard-core process, where each point is surrounded by a circular area called the "hard core", where no other points are

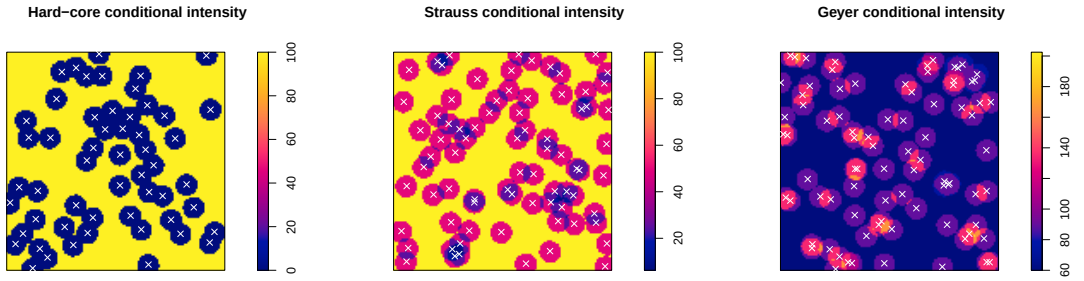

Figure 3.2: Examples how the conditional intensity can look for a point pattern for the hard-core ($R = 0.05, \beta = 100$), Strauss ($R = 0.05, \beta = 100, \gamma = 0.5$) and Geyer saturation processes ($R = 0.05, \beta = 60, \gamma = \sqrt{1.5}, s = 2$) using the same parameters as in the simulation study of Paper 2.

allowed to enter. The conditional intensity is given by

$$\lambda(u|\mathbf{x}) = \beta \mathbf{1}\left\{u \notin \bigcup_{x \in \mathbf{x}} b(x, R)\right\} = \begin{cases} \beta & \text{if } u \notin \bigcup_{x \in \mathbf{x}} b(x, R), \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta > 0$ and $R > 0$ is referred to as the hard-core distance. Hard-core models are repulsive, which is shown for completeness in Lemma 1 of Paper 2. The result is intuitive, since each point "pushes away" points that are "too close". Note however, that points which are further away from each other do not interact. It is possible to simulate realisations of the hard-core process by simulating realisations of the Poisson process, and removing those instances where the hard-core constraint is violated, i.e. when some points are too close to each other, but this approach is slow.

See the first plot in Figure 3.2 for a realisation of a hard-core process with parameters $R = 0.05$ and $\beta = 100$. We also see the conditional intensity illustrated with colours in the plot, and in this case it is either 100 or 0 depending on the distance to the points.

3.1.2 Strauss Models

The next example was introduced by Strauss [1975] as "a model for clustering" but it is in fact not, which was pointed out by Kelly and Ripley [1976]. For completeness, it is shown that the Strauss process is repulsive in Lemma 2 in Paper 2. The Strauss process is similar to the hard-core process, but with an additional parameter γ, determining how much the points interact with each other. A point u interacts with a point x if u is closer than R to x, where $R > 0$ is called the

interaction radius. Its conditional intensity is given by

$$\begin{aligned}\lambda(u|\mathbf{x}) &= \beta\gamma^{D_R(u,\mathbf{x})},\\ D_R(u;\mathbf{x}) &= \sum_{x\in\mathbf{x}\setminus\{u\}} \mathbf{1}\{d(u,x)\le R\} = \sum_{x\in\mathbf{x}\setminus\{u\}} \mathbf{1}\{u\in b(x,R)\},\end{aligned} \tag{3.3}$$

where $\gamma \in [0,1]$ is called the interaction parameter. Here $\gamma = 1$ corresponds to the Poisson process with intensity $\beta > 0$ and $\gamma = 0$ gives us the hard-core process, with the convention that $0^0 = 1$.

The second plot in Figure 3.2 illustrates a realisation of a Strauss process with parameters $R = 0.05, \beta = 100$ and $\gamma = 0.5$. We also see the conditional intensity illustrated with colours in the plot.

3.1.3 Geyer-saturation Models

Geyer [1999] suggested an extension of the Strauss process, called the Geyer saturation process. In the Strauss process, if $\gamma > 1$, this can lead to infinite values of the conditional intensity, but in the Geyer saturation process, this is resolved with the help of a threshold parameter s. Here, the overall contribution from each point is trimmed to never exceed the maximum value of the threshold. The density for the Geyer saturation process is

$$f(\mathbf{x}) = \frac{1}{Z}\beta^n \prod_{y\in\mathbf{x}} \gamma^{\min(s,D_R(y,\mathbf{x}))},$$

where Z is a normalising constant, β, γ, R, s are parameters, and $n = \#(\mathbf{x}\cap\mathcal{S})$. The parameter $s \ge 0$ is called the saturation threshold, and puts an upper limit on how much any single point contributes to the density. For $s = 0$ the process reduces to a Poisson process and for $s = \infty$ the process is a Strauss process with interaction parameter γ^2.

According to Baddeley et al. [2015]:

> The task of calculating the conditional intensity of the saturation process seems to be extremely prone to human error.

Indeed, the expression in Baddeley et al. [2015, Equation (13.36)] is not correct, which is pointed out in their errata. Using the density and (2.6), we obtain the correct conditional intensity expression

$$\lambda(u|\mathbf{x}) = \beta\gamma^{\min(s,D_R(u,\mathbf{x}))+\sum_{y\in\mathbf{x}}(\min(s,D_R(y,\mathbf{x}\cup\{u\}))-\min(s,D_R(y,\mathbf{x})))}.$$

See the third plot in Figure 3.2 for a realisation of a Geyer saturation processes with parameters $R = 0.05, \beta = 60, \gamma = \sqrt{1.5}$ and $s = 2$. It seems like, intuitively, when $\gamma > 1$, a Geyer model is attractive, and when $\gamma < 1$, it is repulsive. However, this

is in fact not true, and a counterexample provided by Marie-Colette van Lieshout is given in Lemma 3 in Paper 2.

3.1.4 Markov Property

A Markov chain is a stochastic process in time defined by the Markov property which states that future observations depend only on the current state and not on the whole past history of the process.

Definition 3.1.1 (Markov chain)**.** *Let $Y_0, Y_1, \ldots$ be a sequence of random variables taking values in some countable set S. Then $Y = (Y_0, Y_1, \ldots)$ is a Markov chain if it satisfies the Markov property*

$$P(Y_n = y_n | Y_0 = y_0, \ldots, Y_{n-1} = y_{n-1}) = P(Y_n = y_n | Y_{n-1} = y_{n-1}), \tag{3.4}$$

for all $n \geq 1$, and all $y_0, y_1, \ldots, y_n \in S$.

Spatial point processes may also exhibit the Markov property, but here we consider neighbours of points in the space $\mathcal{S}$, instead of the previous time slot. Consider a symmetric and reflexive relation $\sim$ on $\mathcal{S}$. The points $u, v \in \mathcal{S}$ are defined as neighbours if $u \sim v$, and the neighbourhood of a point $u \in \mathcal{S}$ is $\delta_{\mathbf{x}}(u) = \{x \in \mathbf{x} : u \sim x\}$. A Gibbs process is a Markov point process if the conditional intensity only depends on the neighbourhood, i.e. $\lambda(u|\mathbf{x}) = \lambda(u|\delta_{\mathbf{x}}(u))$. These types of processes exhibit local interactions, and in fact, all examples given, the Poisson, hard-core, Strauss and Geyer saturation processes, are also Markov processes [van Lieshout, 2000]. In spatial statistics, the terms Gibbs and Markov processes are often used interchangeably, however this might cause some confusion. According to our definition of a Gibbs process, Markov processes are a sub-class of Gibbs processes, but in Illian et al. [2008] Gibbs process are defined as only having interactions between pairs of points, and in that case Gibbs processes are a sub-class of Markov processes. In this thesis we choose to stick to Gibbs processes, noting however that the difference between Gibbs and Markov processes is small when it comes to practical applications and simulations. See Ripley and Kelly [1977], van Lieshout [2000], Møller and Waagepetersen [2004] for more about Markov processes.

3.2 Simulation

Except for the special case of the Poisson process, it is not possible to get the density of a Gibbs process in closed form due to intractable normalising constants, which makes simulation and inference for Gibbs processes challenging. There are ways to work around this, like using Monte Carlo methods, which use random samples for computation to approximate the normalising constant [Møller and Waagepetersen, 2004]. Another way is to use Markov Chain Monte Carlo

(MCMC) methods for simulating from a point process, for which the target density $\pi(\mathbf{x})$ only has to be known up to a constant. MCMC methods work by constructing a Markov chain $Y_0, Y_1, \ldots$ which has the target distribution as its limiting distribution, which is possible if certain conditions are satisfied. Originally introduced by Metropolis et al. [1953], MCMC methods were introduced to the statistics literature by Hastings [1970], see also van Lieshout [2000], Møller and Waagepetersen [2004] for more about MCMC methods for simulation of point processes.

For simulating point processes, the MCMC method called Metropolis-Hastings is most commonly used, and we hereby explain the birth-death Metropolis-Hastings algorithm by Geyer and Møller [1994]. We start with some initial configuration $\mathbf{x}_0$. Considering a Markov chain whose states are point patterns $\mathbf{x}$ in an observation window W, a proposal is made in each time step to change the current state $\mathbf{x}$ to a new state $\mathbf{y}$. With a certain probability, namely the acceptance probability $A(\mathbf{x}, \mathbf{y})$, the proposal is accepted, so that the state is changed from $\mathbf{x}$ to $\mathbf{y}$. Otherwise the proposal is rejected and the state remains as $\mathbf{x}$.

The proposals, in turn, are of two different types, either a *birth* or a *death*, where adding a point u to the point pattern $\mathbf{x}$ corresponds to a birth, while deleting a point x_i from the point pattern $\mathbf{x}$ corresponds to a death. Here, the probability of death is q, and the point to be deleted from the pattern is chosen with equal probability $1/n$ where $n = \#(\mathbf{x} \cap W)$. For the birth, the probability of it occuring is $1 - q$, and the point u is chosen at random in the window W from the proposal density $b(u|\mathbf{x})$.

We can calculate the acceptance probabilities $A(\mathbf{x}, \mathbf{y}) = \min(1, R(\mathbf{x}, \mathbf{y}))$ from the so-called Hastings ratio $R(\mathbf{x}, \mathbf{y})$,

$$R(\mathbf{x}, \mathbf{x} \cup \{u\}) = \frac{\lambda(u|\mathbf{x})}{b(u|\mathbf{x})} \frac{q}{(1-q)(n+1)},$$
$$R(\mathbf{x}, \mathbf{x} \setminus \{x_i\}) = \frac{b(x_i|\mathbf{x} \setminus \{x_i\})}{\lambda(x_i|\mathbf{x})} \frac{n(1-q)}{q}.$$

Note here that we avoid issues with the normalising constant for Gibbs processes since we only have the conditional intensity λ in the expression. Further, we have to know the expression of the birth proposal density b. In the `spatstat` implementation, *shift* proposals are also included, which means to select one of the points $x_i \in \mathbf{x}$ at random and move it to another location u, which is chosen uniformly at random in the window. The Hastings ratio for this proposal is

$$R(\mathbf{x}, \mathbf{x} \cup \{u\} \setminus \{x_i\}) = \frac{\lambda(u|\mathbf{x} \setminus \{x_i\})}{\lambda(x_i|\mathbf{x})}.$$

One drawback of the Metropolis-Hastings algorithm, is that it is not guaranteed to have converged after a finite number of time steps, and the results are approximate.

For some Gibbs processes, like the hard-core and Strauss processes, exact simulation can be used instead, see e.g. van Lieshout [2000], Møller and Waagepetersen [2004]. However, Metropolis-Hastings works well in most cases and is applicable to a wide range of models.

In the simulation study of Paper 1 and 2 we used exact simulation for the hard-core process, and exact simulation for the Strauss process in Paper 2. When simulating from the Geyer saturation process in Paper 2, default parameters were used, namely $q = 0.9$ and $b(u|\mathbf{x}) = \beta$ where β is the intensity parameter of the Geyer saturation process.

3.3 Inference

This section contains a short review of parameter estimation methods for Gibbs processes, for more information see [Dereudre, 2019, Baddeley et al., 2015, Møller and Waagepetersen, 2004]. There are some Bayesian methods, see e.g. [Møller et al., 2006, Rajala, 2014, Shirota and Gelfand, 2017, Stoica et al., 2017], but in this thesis we focus on non-Bayesian inference. In this case, there are two main methods for parameter estimation of Gibbs processes: maximum likelihood estimation (MLE) and conditional intensity-based estimation using the GNZ formula (2.1).

3.3.1 Parameter Estimation for Point Processes

Recall that, given a data-generating model λ_{θ_0}, parameter estimation means to find an estimate $\hat{\theta}$ that is as close as possible to the true parameter θ_0. To compare different estimators, we usually study the bias, variance and mean square error (MSE), defined as

$$\begin{aligned} Bias(\hat{\theta}) &= \mathbb{E}[\hat{\theta}] - \theta_0, \\ Var(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2], \\ MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta_0)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2. \end{aligned}$$

A desirable property of estimators is unbiasedness, which is defined as having zero bias, or equivalently that $\mathbb{E}[\hat{\theta}] = \theta_0$. However, an unbiased estimator with high variance is not desirable, so therefore it is useful to consider the MSE, which takes both the bias and variance into account. Another desirable property for estimators is consistency, defined as $\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta_0| > \varepsilon) = 0$. For point process asymptotics, n is not necessarily the amount of points in the pattern, but instead, what is usually considered is increasing domain asymptotics, meaning that we consider an increasing sequence of bounded subsets $A_n \subset \mathcal{S}$ tending to $\mathcal{S}$ as $n \to \infty$. Lastly, it can be useful to show asymptotic normality of an estimator, as this can be used e.g. to construct approximate confidence regions.

3.3.2 Maximum Likelihood Estimation

In classical statistics MLE is widely used, and it is used to find the most likely parameter value, by maximising the likelihood function. For point processes, the likelihood function that is maximised is the density function $f(\mathbf{x})$. However, for Gibbs processes, we again encounter the issue of an intractable normalising constant. Therefore, approximate maximum likelihood is used, for example numerical approximation of the normalising constant [Ogata and Tanemura, 1981, 1984, 1985, 1989, Ogata, 1986, Ogata and Katsura, 1988, Ogata et al., 2003]. Other approximate MLE methods are Monte Carlo MLE [Penttinen, 1984, Geyer and Møller, 1994, Geyer, 1999], stochastic approximation [Moyeed and Baddeley, 1991] and one-step Monte Carlo MLE [Huang and Ogata, 1999]. Of these methods, one-step Monte Carlo MLE, also called Huang-Ogata approximate maximum likelihood, is implemented in `spatstat` in the `ppm` function. The Huang-Ogata method uses pseudolikelihood (see Section 3.3.4) as initialisation in the first step.

3.3.3 Takacs-Fiksel Estimation

The GNZ formula seen in (2.1), is a very useful identity, defining the conditional intensity. The h-innovations introduced by Baddeley et al. [2005], which are inspired by the GNZ formula, are defined as

$$e_h(\theta) = \sum_{x \in X} h(x, X \setminus \{x\}) - \int_A h(u, X)\lambda_\theta(u|X)\mathrm{d}u, \tag{3.5}$$

for any $A \subseteq \mathcal{S}$. Considering a sequence of test functions h_i, $i = 1, \ldots, q$, where $q \geq p$ and p is the dimension of the parameter θ, we let $e_h(\theta) = (e_{h^1}(\theta), \ldots, e_{h^q}(\theta))$ be a q-dimensional vector. The original formulation of Takacs-Fiksel estimation (TF) [Takacs, 1986, Fiksel, 1988] is to find an estimator $\hat{\theta}$ of θ_0 by minimising the sum of squares $\sum_{i=1}^{q} e_{h^i}(\theta)^2$.

Note that for the h-innovations defined in (3.5), it holds that $\mathbb{E}[e_h(\theta_0)] = 0$, and we call $e_h(\theta)$ an unbiased estimating function. Considering a sequence of test functions h_i, $i = 1, \ldots, q$ as before, and letting $q = p$, we obtain a system of estimating equations $e_{h_i}(\theta) = 0$. This is the estimating equation formulation of TF [Coeurjolly et al., 2016].

3.3.4 Pseudolikelihood

A special case of TF is pseudolikelihood estimation (PL). Here we set the test function to $h(x,\cdot) = \frac{\partial}{\partial\theta}\log\lambda_\theta(x|\cdot)$ and get

$$\begin{aligned}&\sum_{x\in X}\frac{\partial}{\partial\theta}\log\lambda_\theta(x|X\setminus\{x\}) - \int_A \frac{\partial}{\partial\theta}\log\lambda_\theta(u|X)\lambda_\theta(u|X)\mathrm{d}u\\ =&\sum_{x\in X}\frac{\partial}{\partial\theta}\log\lambda_\theta(x|X\setminus\{x\}) - \int_A \frac{\partial}{\partial\theta}\lambda_\theta(u|X)\mathrm{d}u,\end{aligned}$$

for $A \subseteq \mathcal{S}$. This is the score function (derivative with respect to θ) of the log-pseudolikelihood function, which is defined as

$$\log PL(\theta) = \sum_{x\in X}\log\lambda_\theta(x|X) - \int_A \lambda_\theta(u|X)\mathrm{d}u, \quad A\subseteq\mathcal{S}.$$

Here we want to estimate the value of the parameter θ_0 by maximising the log-pseudolikelihood function. Note that maximising the log-pseudolikelihood function is equivalent to setting the score function to zero, which is the way of solving the system of estimating equations for TF, with test function $h(x,\cdot) = \frac{\partial}{\partial\theta}\log\lambda_\theta(x|\cdot)$.

PL was first introduced by Besag [1974] for lattice systems, and was extended to point processes by Ripley [1988]. The pseudolikelihood function has similar properties as the likelihood function and is asymptotically unbiased, consistent and asymptotically normal under appropriate conditions [Jensen and Møller, 1991, Jensen and Künsch, 1994, Mase, 1995, 2000, Billiot et al., 2008]. In practice, both pseudolikelihood estimators and approximate maximum likelihood estimators exhibit bias, which might be due to software implementation [Baddeley and Turner, 2014]. Additionally, PL might perform poorly when the interaction is strong [Geyer and Thompson, 1992, Diggle et al., 1994].

PL is implemented as the default option in the `ppm` function in `spatstat`. Baddeley et al. [2014] introduced another method which exploits logistic regression, also implemented in the `ppm` function, which can be seen as a numerically stable approximation of PL. There is also a recently developed parameter estimation method in the context of intensity estimation, which uses cross-validation similarly defined as in Definition 2.4.1, together with PL, see Lin and Kang [2024].

3.3.5 Optimal Test Functions for Parameter Estimation

It is important to note that for the parameter estimation methods, we deal with hyperparameters, like the test function h. For TF, one common choice is the test function which gives us PL, as described above. Another option is $h(\cdot) = 1/\lambda(\cdot)$, which is the Stoyan-Grabarnik test function [Stoyan and Grabarnik, 1991], also called the inverse test function [Baddeley et al., 2005].

An estimating equation is optimal in the sense of Godambe optimality when the test function chosen maximises the Godambe information $G_h = S_h(\theta)\Sigma_h(\theta)^{-1}S_h(\theta)$, which consists of the sensitivity matrix $S_h(\theta) = -\mathbb{E}[\frac{\partial e_h(\theta)}{\partial \theta^T}]$ and the covariance matrix $\Sigma_h(\theta) = Var(e_h(\theta))$ for the estimating function $e_h(\theta)$. The argument behind Godambe optimality is based on an asymptotic approximation, in terms of increasing domain asymptotics. It can be shown (see Guan et al. [2015]) that a sufficient condition for the test function g to be Godambe optimal is that

$$\Sigma_{hg}(\theta_0) = S_g(\theta_0), \tag{3.6}$$

for all test functions h, where $\Sigma_{hg}(\theta_0) = \mathrm{Cov}(e_h(\theta_0), e_g(\theta_0))$.

In the context of intensity-based estimation, Guan et al. [2015] found an optimal estimating equation, called quasi-likelihood, where the test function used was found by solving (3.6). Coeurjolly et al. [2016] followed the lines of Guan et al. [2015] and attempted to find an optimal estimating equation for TF, which might outperform PL. However, as conditonal-intensity based estimation is far more mathematically challenging than intensity-based estimation, due to interactions between points, the test function that Coeurjolly et al. [2016] introduced was a semi-optimal test function.

Chapter 4

Point Process Learning

This thesis studies Point Process Learning (PPL) which is a new statistical methodology introduced by Cronie et al. [2024b]. This chapter briefly recalls the main concepts, but we refer to Cronie et al. [2024b] and Paper 1, 2 and 3 for more information.

4.1 Prediction Errors

Given a point process X with conditional intensity $\lambda_{\theta_0}(u|X)$, $u \in \mathcal{S}$, consider a training-validation pair (X^T, X^V), as defined in Definition 2.4.1. Then, the point process prediction errors are defined as (see [Cronie et al., 2024b] and Paper 2)

$$\mathcal{I}_\theta^h(A; X^T, X^V) = \sum_{x \in X^V \cap A} h(x; X^T \setminus \{x\}) - \int_A h(u; X^T) V_\theta(u) \lambda_\theta(u|X^T) \mathrm{d}u, \quad (4.1)$$

where $A \subseteq \mathcal{S}$. Note the similarity to the h-innovations in (3.5). There are two main differences, the type of prediction and the weight function $V_\theta(u)$, see Section 4.2. If we use the pair (X, X) in (4.1), instead of the pair (X^T, X^V), we get so-called auto-prediction (see [Cronie et al., 2024b] and Paper 2)

$$\mathcal{I}_\theta^h(A; X, X) = \sum_{x \in X \cap A} h(x; X \setminus \{x\}) - \int_A h(u; X) V_\theta(u) \lambda_\theta(u|X) \mathrm{d}u.$$

In terms of cross-validation (CV), this approach reflects leave-one-out CV. Here we see that setting $V_\theta(u) = 1$ gives us (3.5), which means that the point process prediction errors in (4.1) are a generalization of the h-innovations in (3.5).

Similarly to the h-innovations, the prediction errors are unbiased, in the sense that

$$\mathbb{E}[\mathcal{I}_\theta^h(A; X^T, X^V)] = 0, \quad (4.2)$$

under certain conditions, see Cronie et al. [2024b, Theorem 2]. More precisely, let $\breve{\lambda}_{\theta_0}\{(u,m)|\breve{X}\}$, $(u,m) \in \mathcal{S} \times \mathcal{M}$ be the conditional intensity of the marked point process $\breve{X} = \{(x,0) : x \in X^T\} \cup \{(x,1) : x \in X^V\}$ with mark space $\mathcal{M} = \{0,1\}$. The prediction error $\mathcal{I}^h_\theta(A; X^T, X^V)$, for any $A \subseteq \mathcal{S}$, has expectation 0 if and only if the parameter θ is the true parameter θ_0. This holds under boundedness assumptions on λ_{θ_0}, $\breve{\lambda}_{\theta_0}$ and h, namely that for $|\cdot|$-almost every $u \in \mathcal{S}$ it holds that $\mathbb{E}[\lambda_{\theta_0}(u|X)^2] < \infty$, $\mathbb{E}[\breve{\lambda}_{\theta_0}((u,1)|\breve{X})^2] < \infty$ and $\mathbb{E}[h(u, X^T)^2] < \infty$ for $\theta_0 \in \Theta$.

4.2 PPL-weight

The weight function $V_\theta(u)$ is called the PPL-weight and is defined as

$$V_\theta(u) = V_\theta(u, X^T, X^V) = \mathbb{E}\left[\frac{\breve{\lambda}_\theta\{(u,1)|\breve{X}\}}{\lambda_\theta(u|X^T)}\middle| X^T\right], \quad u \in \mathcal{S}, \tag{4.3}$$

assuming that $\mathbb{E}[\lambda_\theta(u|X)^2] < \infty$ for almost all $u \in \mathcal{S}$ and all $\theta \in \Theta$ with the convention that $0/0 = 0$. It can be difficult to find the form of the PPL-weight in general, but (4.3) reduces to

$$V_\theta(u) = V_\theta(u, X^T, X^V) = p(u)\,\mathbb{E}\left[\frac{\lambda_\theta(u|X)}{\lambda_\theta(u|X^T)}\middle| X^T\right], \tag{4.4}$$

when X^V is an independent thinning of X, based on the retention probability function $p(u) \in (0,1)$, $u \in \mathcal{S}$. This is an argument for considering independent thinning-based point process CV, such as Monte-Carlo CV or block CV.

Expressions for the PPL-weight are shown in Paper 2 for general Gibbs processes, as well as for Poisson, hard-core, Strauss and Geyer saturation processes. Since the form of the weight is often mathematically intractable, approximations and estimations of the weight is further discussed in Paper 2. Common choices are $V_\theta(u) \approx p(u)$, used in Paper 1, and $V_\theta(u) \approx p(u)/(1-p(u))$, used in Cronie et al. [2024b]. Paper 2 proposes another approach, namely estimating the weight from the point pattern at hand. In Paper 2, these three choices of weights are compared in a simulation study, where $p(u)/(1-p(u))$ did not work well for any of the models. The choice between the fixed weight $p(u)$, and estimating the weight, is not as clear, and depends on the model. Usually, weight estimation yields more stable results, but it is also more computationally expensive and time consuming.

4.3 Loss Functions

Since the prediction error is zero in expectation for the true parameter, the idea of PPL is to find a good parameter estimate by minimising the expectation of the prediction error. In practice, the expectation of the prediction error has to

be approximated, as it is difficult to evaluate. To this end, Cronie et al. [2024b] introduced loss functions as follows. Given a cross-validation round $\{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^k$ of $\mathbf{x}$, let $\mathbf{X} = \{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^k$ and

$$\mathcal{T}_k = \mathcal{T}(\mathbf{X}) = \{i \in \{1, \dots, k\} : \mathbf{x}_i^T \neq \emptyset, \mathbf{x}_i^V \neq \emptyset\}.$$

This construction ensures that we exclude splits where either $\mathbf{x}^T$ or $\mathbf{x}^V$ is empty. Then the loss functions are

$$\begin{aligned} \mathcal{L}_j(\theta) &= \frac{1}{\#\mathcal{T}_k} \sum_{i \in \mathcal{T}_k} \left| \mathcal{I}_\theta^h(A; \mathbf{x}_i^T, \mathbf{x}_i^V) \right|^j, \qquad j = 1, 2, \\ \mathcal{L}_3(\theta) &= \left(\frac{1}{\#\mathcal{T}_k} \sum_{i \in \mathcal{T}_k} \mathcal{I}_\theta^h(A; \mathbf{x}_i^T, \mathbf{x}_i^V) \right)^2. \end{aligned}$$

In Paper 3, these loss functions are further motivated, by an empirical risk formulation of PPL.

In the simulation study of Paper 1, which concerned the hard-core process, the loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ yielded much better results than the $\mathcal{L}_3$ loss function. In the simulation study of Paper 2, it is seen again that the choice of loss function affects the results. For the Geyer saturation process, the $\mathcal{L}_3$ loss functions obtained the best results, while for the Poisson, hard-core and Strauss processes, the $\mathcal{L}_1$ and $\mathcal{L}_2$ loss functions performed best.

4.4 Relation to Takacs-Fiksel Estimation

Recall that h-innovations are a special case of point process prediction errors. Motivated by this fact, Paper 2 shows that Takacs-Fiksel estimation (TF) is a special case of PPL. We here present, in Theorem 1, a reformulated version of Theorem 2 and Theorem 3 in Paper 2. Theorem 1 shows that TF is a limiting case of a weighted average of prediction errors, for both Monte-Carlo CV and block CV, when the cross-validation scheme tends to leave-one-out CV.

Theorem 1 (Paper 2)**.** *Assume that $\lambda_\theta(u|\mathbf{x})$ and $h(u; \mathbf{x})$ are bounded for any $\theta \in \Theta$, $u \in \mathcal{S}$ and $\mathbf{x} \in \mathbf{N}$. Further, let $A \subseteq \mathcal{S}$ be bounded and $k \geq 2$.*

1. *Let $\{(X_i^T(p_k), X_i^V(p_k))\}_{i=1}^k$ be a Monte-Carlo cross-validation of X, based on a retention probability $p_k \in (0, 1)$. If $p_k = 1/\sqrt{k}$, then*

$$p_k \sum_{i=1}^k \mathcal{I}_\theta^h(A; X_i^V(p_k), X_i^T(p_k)) - e_h(\theta) \xrightarrow{p} 0$$

as $k \to \infty$.

2. *Let* $\{(X_{ik}^T, X_{ik}^V)\}_{i=1}^k$, *be block cross-validations of* $X \cap A$, *based on partitions* $\{A_{ik}\}_{i=1}^k$ *of* A, *with associated retention probabilities* $p_{ik}(u) = \mathbf{1}\{u \in A_{ik}\}$, $i = 1, \ldots, k$. *Assume further that the partition sizes satisfy* $\max_{i=1,\ldots,k} |A_{ik}| \to 0$ *as* $k \to \infty$ *and that for any* $i = 1, \ldots, k$ *there exists only one* $j = 1, \ldots, k+1$ *such that* $A_{ik} \subseteq A_{j(k+1)}$, *i.e. we have a refinement. Then*

$$\sum_{i=1}^{k} \mathcal{I}_\theta^h(A; X_{ik}^V, X_{ik}^T) - e_h(\theta) \xrightarrow{p} 0$$

as $k \to \infty$.

4.5 Estimating Equation Approach

We can also formulate PPL with estimating equations, similarly to how it was done with TF in Section 3.3.3. Let $h_\theta = (h_\theta^1, \ldots, h_\theta^d)^T$ be a d-dimensional test function and $\theta = (\theta_1, \ldots, \theta_d)^T \in \Theta \subseteq \mathbb{R}^d$ be a d-dimensional parameter for a given model. Further, let $\{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^k$, $k \geq 1$, be a cross-validation round for a given point pattern $\mathbf{x}$, with the associated prediction error vectors $\mathcal{I}^{h_\theta}(A; \mathbf{x}_i^T, \mathbf{x}_i^V) = (\mathcal{I}^{h_\theta^1}(A; \mathbf{x}_i^T, \mathbf{x}_i^V), \ldots, \mathcal{I}^{h_\theta^d}(A; \mathbf{x}_i^T, \mathbf{x}_i^V))^T \in \mathbb{R}^d$, $i = 1, \ldots, k$. For ease of notation we denote the mean of the prediction errors as $I_{h,k}^j(\theta; \mathbf{x}) = \frac{1}{\#\mathcal{T}_k} \sum_{i=1}^k \mathcal{I}^{h_\theta^j}(A; \mathbf{x}_i^T, \mathbf{x}_i^V)$ for $j = 1, \ldots, d$, and the vector of all the means as $I_{h,k}(\theta; \mathbf{x}) = (I_{h,k}^1(\theta; \mathbf{x}), \ldots, I_{h,k}^d(\theta; \mathbf{x}))^T$. Here, we want to solve the system of equations

$$I_{h,k}(\theta; \mathbf{x}) = (0, \ldots, 0)^T, \tag{4.5}$$

in order to find an estimate $\widehat{\theta}_{h,k}(\mathbf{x}) \in \Theta$. By (4.2), we know that $\mathbb{E}[I_{h,k}(\theta; X)] = \mathbb{E}[\mathcal{I}^{h_\theta}(A; X^T, X^V)] = (0, \ldots, 0)^T$ when $\theta = \theta_0$, which means that (4.5) is an unbiased estimating equation.

4.5.1 Large Sample Properties

In Theorem 3 of Paper 3, we provide large sample properties (consistency and asymptotic normality) of the estimator $\hat{\theta}(X, A_n)$, which hold when the observation window A_n increases. We present the results in Theorem 2 below, and refer to Paper 3 for details about the conditions and assumptions.

Theorem 2 (Paper 3)**.** *Under some regularity conditions, it holds that*

$$\hat{\theta}(X, A_n) \to \theta_0$$

in probability, as $n \to \infty$. *Under additional assumptions, it holds that*

$$|A_n|^{1/2}(\hat{\theta}(X, A_n) - \theta_0)$$

tends weakly to a Gaussian vector with mean 0 and covariance matrix $V^{-1}JV^{-1}$, *as* $n \to \infty$.

4.5.2 Small Sample Properties

Moving on to small sample properties we are interested in the bias and variance of the estimator $\hat{\theta}$, which we obtain expressions for using a Taylor expansion of $I_{h,k}(\theta)$ around $\hat{\theta}$. More specifically,

$$\begin{aligned} I_{h,k}(\theta;\mathbf{x}) &= I_{h,k}(\widehat{\theta}_{h,k}(\mathbf{x});\mathbf{x}) - \mathcal{D}I_{h,k}(\theta;\mathbf{x})\left(\widehat{\theta}_{h,k}(\mathbf{x}) - \theta\right) - R_{h,k}(\theta,\widehat{\theta}_{h,k}(\mathbf{x})) \quad (4.6)\\ &= -\mathcal{D}I_{h,k}(\theta;\mathbf{x})\left(\widehat{\theta}_{h,k}(\mathbf{x}) - \theta\right) - R_{h,k}(\theta,\widehat{\theta}_{h,k}(\mathbf{x})). \end{aligned}$$

where $\mathcal{D}I_{h,k}(\theta;\mathbf{x})$ is the Jacobian matrix in terms of θ, and $R_{h,k}(\theta,\widehat{\theta}_{h,k}(\mathbf{x}))$ is the remainder term. In the second equality of (4.6), $I_{h,k}(\widehat{\theta}_{h,k}(\mathbf{x});\mathbf{x})$ disappears since $\widehat{\theta}_{h,k}(\mathbf{x})$ is the estimate solving the estimating equation (4.5), i.e. it is assumed that $I_{h,k}(\widehat{\theta}_{h,k}(\mathbf{x});\mathbf{x}) = (0,\ldots,0)^T$. By reordering terms we now obtain the bias of the estimator as

$$\begin{aligned} Bias(\widehat{\theta}_{h,k}) &= -B^1_{h,k} - B^2_{h,k},\\ B^1_{h,k} &= \mathbb{E}[\mathcal{D}I_{h,k}(\theta_0)^{-1}I_{h,k}(\theta_0)],\\ B^2_{h,k} &= \mathbb{E}[\mathcal{D}I_{h,k}(\theta_0)^{-1}R_{h,k}(\theta_0,\widehat{\theta}_{h,k})], \end{aligned}$$

and the variance as

$$\begin{aligned} \mathrm{Var}(\widehat{\theta}_{h,k}) &= C^1_{h,k} + C^2_{h,k} + 2C^3_{h,k},\\ C^1_{h,k} &= \mathrm{Var}(\mathcal{D}I_{h,k}(\theta_0)^{-1}I_{h,k}(\theta_0)),\\ C^2_{h,k} &= \mathrm{Var}(\mathcal{D}I_{h,k}(\theta_0)^{-1}R_{h,k}(\theta_0,\widehat{\theta}_{h,k})),\\ C^3_{h,k} &= \mathrm{Cov}(\mathcal{D}I_{h,k}(\theta_0)^{-1}I_{h,k}(\theta_0),\mathcal{D}I_{h,k}(\theta_0)^{-1}R_{h,k}(\theta_0,\widehat{\theta}_{h,k})). \end{aligned}$$

To provide explicit expressions for the bias and variance of the estimator, we need expressions for $B^1_{h,k}$, $B^2_{h,k}$, $C^1_{h,k}$, $C^2_{h,k}$ and $C^3_{h,k}$, so to this end, some properties of the estimating function $I_{h,k}(\theta)$ are derived. For example, to obtain an expression for $\mathcal{D}I_{h,k}(\theta_0)^{-1}$ we start with partial derivatives of prediction errors. Further, a new variance expression for prediction errors is provided, since the one provided in Cronie et al. [2024b] and Paper 2 had a small calculation error. After meticulous calculations, we can conclude that the expressions for the bias and variance of the estimator are intractable.

4.5.3 Approximation with Sensitivity Matrices

For simpler computations, we follow the approach of Guan et al. [2015], Coeurjolly et al. [2016] and approximate the stochastic Jacobian matrix $\mathcal{D}I_{h,k}(\theta,X)$ with its expectation, the sensitivity matrix $S_{h,k}(\theta)$ defined as

$$S_{h,k}(\theta) = \mathbb{E}[-\mathcal{D}I_{h,k}(\theta,X)]. \quad (4.7)$$

The approximation is based on asymptotic arguments, where also the remainder term $R_{h,k}(\theta, \widehat{\theta}_{h,k}(\mathbf{x}))$ tends to 0 in probability, due to Theorem 2. This gives us an (approximately) asymptotically unbiased estimator $\widehat{\theta}_{h,k}(X)$, since

$$Bias(\widehat{\theta}_{h,k}(X)) \approx S_{h,k}^{-1}(\theta_0)\mathbb{E}[I_{h,k}(\theta_0, X)] = (0, \ldots, 0)^T.$$

Further, the covariance matrix of the estimator is also simplified, and is given by

$$\mathrm{Var}(\widehat{\theta}_{h,k}) \approx \mathrm{Var}(S_{h,k}^{-1} I_{h,k}(\theta_0)) = S_{h,k}^{-1}(\theta_0)\Sigma_{h,k}(\theta_0)S_{h,k}^{-1}(\theta_0),$$

where $\Sigma_{h,k}(\theta_0) = \mathrm{Var}(I_{h,k}(\theta_0))$. Lastly, the Godambe information $G_{h,k}$ is defined as the inverse of the covariance matrix of the estimator, namely

$$G_{h,k} = \mathrm{Var}(\widehat{\theta}_{h,k})^{-1} = S_{h,k}(\theta_0)\Sigma_{h,k}^{-1}(\theta_0)S_{h,k}(\theta_0).$$

In the context of Godambe optimality, we want to maximise the Godambe information, or equivalently, "minimise" the covariance matrix of the estimator, see Section 3.3.5.

4.5.4 Optimal Test Functions for PPL

When attempting to solve an estimating equation like (4.5), the choice of the test function h is not straightforward. In previous work for PPL ([Cronie et al., 2024b], Paper 1 and Paper 2) the Stoyan-Grabarnik test function has been used. For PPL, the (approximately) optimal test function g may be found, where the sufficient condition (3.6) becomes

$$\Sigma_{(h,k),(g,k)}(\theta_0) = S_{g,k}(\theta_0), \tag{4.8}$$

for all test functions h, where $\Sigma_{(h,k),(g,k)}(\theta_0) = \mathrm{Cov}(I_{h,k}(\theta_0), I_{g,k}(\theta_0))$. See Paper 3 for a proof that it is a sufficient condition. Paper 3 seeks to solve (4.8), similarly to Guan et al. [2015] and Coeurjolly et al. [2016], but this time, in the context of PPL. It is enough to consider the difference $\Phi_{hg} = \Sigma_{(h,k),(g,k)}(\theta_0) - S_{h,k}(\theta_0)$, and to verify condition (4.8) we aim to solve for the (approximately) optimal test function g in $\Phi_{hg} = 0$.

To this end, Paper 3 presents expressions for $\Sigma_{(h,k),(g,k)}(\theta_0)$ and $S_{h,k}(\theta_0)$, and due to the nature of these expressions, the conclusion is that solving $\Phi_{hg} = 0$ is intractable in the general case. To reduce the complexity involved in solving the equation, we assume that some parts of Φ_{hg} are negligible, along the lines of Coeurjolly et al. [2016]. Moreover, we consider the case of independent thinning-based CV, and explore the case where $k \to \infty$ in the cross-validation scheme. This leads to an equation that, for certain point processes like for example Strauss processes, could potentially be solved numerically by using the solution of a Fredholm integral, similar to the approach of Coeurjolly et al. [2016]. However, this task

is challenging, and even if successful, the resulting test function is only (approximately) semi-optimal. As an alternative, we consider the special case of a Poisson process under similar assumptions as in the general case. In this case, we derive an analytical solution of the equation and identify an (approximately) semi-optimal test function.

Chapter 5

Summary of Papers

5.1 Paper 1

Previously, Cronie et al. [2024b] showed that Point Process Learning (PPL) outperforms the state-of-the-art in kernel intensity estimation, i.e. the Cronie and van Lieshout [2018] approach. In this paper, we consider PPL in the context of parameter estimation for Gibbs processes, and apply it to the hard-core process. When using pseudolikelihood estimation (PL) for parameter estimation for the hard-core process, identifiability issues are encountered, since the test function is not differentiable with respect to the interaction distance R. In practice, estimation is done by the function `ppm` in the R package spatstat [Baddeley et al., 2015], which uses PL for the β parameter and a plug-in approach for the R parameter. In this context, PPL is more flexible, as we may let both R and β be free parameters to be estimated. The simulations in Paper 1 show that there are hyperparameter choices so that PPL outperforms PL for the hard-core process. Specifically, the choices are Monte-Carlo cross-validation (CV), loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$, and $p < 0.3$. In the case of β, the performance for the pseudolikelihood estimator is $\text{MSE}(\widehat{\beta}) = 315$. For PPL, the best performance is when $p = 0.1$, as the value of $\text{MSE}(\widehat{\beta})$ for $\mathcal{L}_1$ is given by 199 and for $\mathcal{L}_2$ it is given by 194, which are significantly lower than what PL gives rise to. For the R parameter, the `ppm` function (plug-in approach) gives the same result as PPL.

5.2 Paper 2

The main focus of this paper, is to relate PPL to Takacs-Fiksel estimation (TF). We show that TF is a special case of PPL for Monte-Carlo CV and block CV, respectively. The results hold under boundedness assumptions, and asymptotically, when the cross-validation scheme tends to leave-one-out CV. Notably, this shows the generality and flexibility of PPL.

Cronie et al. [2024b] show that the expectation of the prediction errors is zero if and only if the so-called PPL-weight is of a certain form. Here, we provide general expressions for the PPL-weight for Gibbs models. Specifically, we investigate the weight expressions for Poisson, hard-core, Strauss and Geyer saturation processes. For the Poisson process, the weight takes a simple form but for the other models, the weight is intractable. Therefore we discuss different practical choices for the weight.

We further compare PPL to TF through a simulation study for four common Gibbs models, namely Poisson, hard-core, Strauss and Geyer saturation processes. There are many different hyperparameters, for example the CV parameters, the PPL-weight and the test function. In our simulations, the test function is fixed to be the Stoyan-Grabarnik test function [Stoyan and Grabarnik, 1991], which is used previously in different contexts [Baddeley et al., 2005, Cronie and van Lieshout, 2018, Kresin and Schoenberg, 2023]. In Paper 1, PPL with the Stoyan-Grabarnik test function was compared to PL, which is a special case of TF but with another test function. The comparison here is more fair, since we compare PPL with TF, using the Stoyan-Grabarnik test function in both cases. The results of the simulations show that, for all four Gibbs models, one can choose hyperparameters so that PPL outperforms TF in terms of mean square error.

5.3 Paper 3

In this paper, we investigated whether PPL can be sensibly formulated as an estimating equation approach. The motivation comes from an earlier work by Coeurjolly et al. [2016], which introduced an estimating equation approach for TF, and from Paper 2, which highlighted connections between PPL and TF. While PPL originally relied on minimising one of the $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ loss functions, we explored the potential of using estimating equations by defining an estimating function based on the mean prediction errors across cross-validation splits.

We first derived expressions for the bias and variance of the parameter estimator, which turned out to be intractable. To combat this, we approximated the Jacobian matrix with the sensitivity matrix which yielded an approximately unbiased estimator, but still with a complicated variance expression. Efforts to determine an optimal test function g under Godambe optimality yielded only semi-optimal solutions, limited to the special case of a Poisson process and relying on numerous simplifying assumptions. Thus, we conclude that the estimating equation approach is unsuitable for PPL.

The preferred methodology for PPL remains norm minimisation with the $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ loss functions, as previously heuristically motivated in Cronie et al. [2024b]. By framing this approach within an empirical risk minimisation context [Vapnik, 1999], we provide a theoretical foundation for these loss functions.

Chapter 6

Conclusions and Future Work

In this thesis, the Point Process Learning (PPL) approach introduced in Cronie et al. [2024b] has been explored in the context of Gibbs point processes. Considering other methods for fitting parameters of Gibbs processes, like Takacs-Fiksel estimation (TF) and pseudolikelihood estimation (PL), PPL has a clear role as a competing state-of-the-art method. Further, PPL has interesting theoretical properties, as seen in Paper 3, and is a generalization of TF, as seen in Paper 2.

This thesis, together with other related work [Rost, 2024, Östling, 2024], marks the start of further development of PPL. There are several potential future paths, where the most important one is to investigate the choice of hyperparameters for PPL, e.g. the thinning probability p, the PPL-weight and the test function. There are also different types of cross-validation to consider, even though Monte-Carlo cross-validation was used in the simulation studies of both Paper 1 and 2. Weight choice is discussed in Paper 2, but remains to be investigated more.

Another choice when using PPL in practice is the type of loss function used for parameter estimation, and in Paper 1 and 2, results were presented for all three loss functions: $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$. This choice is not necessary in the estimating equation approach of PPL formulated in Paper 3, since here PPL is carried out by solving an equation system, instead of minimising loss functions. Since the estimating equation approach of PPL was not successful for practical use, we provide in Paper 3 an empirical risk formulation of PPL, where the loss functions are further motivated. So for the current state of PPL, all three loss functions are possible choices, but it might be an option to stick to one of the loss functions, e.g. the $\mathcal{L}_2$ loss function, in future studies, to narrow down the scope. One argument for choosing the $\mathcal{L}_2$ loss function, is that it generally performed best in the simulation studies of Paper 1 and 2, with exception for the Geyer saturation process. The results for the $\mathcal{L}_1$ loss function were almost the same as for the $\mathcal{L}_2$ loss function,

so it is enough to consider $\mathcal{L}_2$. Further, $\mathcal{L}_2$ was used in the original formulation of TF, and it is one of the most commonly used loss functions in machine learning.

So far, in the simulation studies in Paper 1 and 2, we have run the simulations over a range of p values and for all three loss functions, but the test function was kept fixed as the Stoyan-Grabarnik test function. In Paper 3, the choice of test function was explored in terms of attempting to find an optimal one, in the sense of Godambe optimality, with an estimating equation approach. However, due to mathematical intractability, this proved to be a difficult task, and a clear next step would be to explore different test functions in a simulation study. Here, it might be suitable to fix values of the other hyperparameters, for feasibility.

Another aspect to consider is that the simulation studies of Paper 1 and 2 have limitations. We have covered a variety of cases: an inhomogeneous Poisson process, the repulsive hard-core and Strauss processes, and the Geyer saturation process with $\gamma > 1$. However, the value of the parameter γ used for the Strauss and Geyer saturation processes, was quite close to one, which means that the processes behave similarly to a homogeneous Poisson process. This means that varying the γ parameter, and the saturation threshold s for the Geyer saturation process, would be relevant, in order to consider a wider range of point patterns for future simulation studies.

Currently, grid search is used as optimisation method for PPL, which makes the choice of grid another aspect to consider, especially when comparing results to PL which does not use grid search. Moreover, the average number of points in the simulation studies in Paper 1 and 2 were between 60 and 100 points. Testing PPL on a point pattern with a larger amount of points would be relevant, but at the moment infeasible since PPL is more computationally extensive than TF and PL. However, if the choice of hyperparameters can be simplified, much of the computational time would be reduced.

Further, it would be interesting to consider other Gibbs processes, like hybrids of Gibbs models [Baddeley et al., 2013]. Also, it would be illustrative to test PPL on a real data-set, and on point patterns on other spaces than a square 2D window. An ambition for the future is to reduce the computational time for running PPL, and to provide a package for PPL as a part of `spatstat`.

Paper 2 and 3 extend the theory around PPL; Paper 2 delves into properties of the PPL-weights and shows the relation to TF, and Paper 3 provides an empirical risk formulation of PPL, as well as an estimating equation approach. Further, Paper 3 provides large and small sample properties of the parameter estimator, and presents distributional properties of the prediction errors, like covariance matrices. Building on these results, it would be relevant to continue exploring properties of PPL, for example how the choice of the hyperparameters affects the parameter estimator.

Another possibility, would be to explore PPL in terms of moment-based models, like Cox processes [Cox, 1955], instead of conditional intensity-based models, which Gibbs processes are. It would also be interesting to look more into spatio-temporal models like the Hawkes process [Hawkes, 1971].

To summarise, two future research aims (which are also closely intertwined) are stated below:

- Investigate the choice of the hyperparameters for PPL, such as the type of cross-validation, the associated parameters p and k, the PPL-weight and the test function. Even though the PPL-weight was investigated in Paper 2, and the choice of test functions and loss functions were investigated in Paper 3, they should be explored further.
- Explore PPL more in different contexts, for example extending the simulation studies, investigate theoretical properties of PPL and applying PPL to other types of point processes beyond Gibbs processes. Aspects to consider in future simulations studies include varying the grid, the number of points, and the type of point patterns. Further, exploring PPL for other types of Gibbs processes such as hybrids of Gibbs models, as well as other types of point processes, e.g. Cox and Hawkes processes, is an important future direction including both theoretical work, as well as practical aspects.

Bibliography

Qi Wei Ang, Adrian Baddeley, and Gopalan Nair. Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4):591–617, 2012.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

Adrian Baddeley and Rolf Turner. Bias correction for parameter estimates of spatial point process models. *Journal of Statistical Computation and Simulation*, 84(8):1621–1643, 2014.

Adrian Baddeley, Rolf Turner, Jesper Møller, and Martin Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666, 2005.

Adrian Baddeley, Rolf Turner, Jorge Mateu, and Andrew Bevan. Hybrids of gibbs point process models and their implementation. 2013.

Adrian Baddeley, Jean-François Coeurjolly, Ege Rubak, and Rasmus Waagepetersen. Logistic regression for spatial gibbs point processes. *Biometrika*, 101(2):377–392, 2014.

Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R.* CRC Press, 2015.

Fekadu L Bayisa, Markus Ådahl, Patrik Rydén, and Ottmar Cronie. Regularised semi-parametric composite likelihood intensity modelling of a swedish spatial ambulance call point pattern. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–20, 2023.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Steffen Betsch. Structural properties of gibbsian point processes in abstract spaces. *Journal of Theoretical Probability*, pages 1–63, 2023.

Jean-Michel Billiot, Jean-François Coeurjolly, and Rémy Drouilhet. Maximum pseudolikelihood estimator for exponential family models of marked Gibbs point processes. *Electronic Journal of Statistics*, 2(none):234 – 264, 2008. doi: 10.1214/07-EJS160.

Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

Jean-François Coeurjolly, Jesper Møller, and Rasmus Waagepetersen. A tutorial on palm distributions for spatial point processes. *International Statistical Review*, 85(3):404–420, 2017.

Jean-François Coeurjolly, Yongtao Guan, Mahdieh Khanmohammadi, and Rasmus Waagepetersen. Towards optimal takacs–fiksel estimation. *Spatial Statistics*, 18: 396–411, 2016. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2016.08.002.

David Coupier. *Stochastic Geometry: Modern Research Frontiers*, volume 2237. Springer, 2019.

David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

Ottmar Cronie and Marie-Colette van Lieshout. A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2):455–462, 2018.

Ottmar Cronie, Julia Jansson, and Konstantinos Konstantinou. Discussion of the paper "marked spatial point processes: Current state and extensions to point processes on linear networks". *Journal of Agricultural, Biological and Environmental Statistics*, 2024a. doi: 10.1007/s13253-024-00606-0.

Ottmar Cronie, Mehdi Moradi, and Christophe AN Biscio. A cross-validation-based statistical theory for point processes. *Biometrika*, 111(2):625–641, 2024b. doi: 10.1093/biomet/asad041.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer-Verlag New York, second edition, 2008.

David Dereudre. Introduction to the theory of gibbs point processes. *Stochastic Geometry: Modern Research Frontiers*, pages 181–229, 2019.

Peter J Diggle. Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. *Journal of neuroscience methods*, 18(1-2): 115–125, 1986.

Peter J Diggle, Thomas Fiksel, Pavel Grabarnik, Yosihiko Ogata, Dietrich Stoyan, and Masaharu Tanemura. On parameter estimation for pairwise interaction point processes. *International Statistical Review/Revue Internationale de Statistique*, pages 99–117, 1994.

Nicoletta D'Angelo, Giada Adelfio, Jorge Mateu, and Ottmar Cronie. Local inhomogeneous weighted summary statistics for marked point processes. *Journal of Computational and Graphical Statistics*, pages 1–15, 2023.

Matthias Eckardt and Mehdi Moradi. Marked spatial point processes: current state and extensions to point processes on linear networks. *Journal of Agricultural, Biological and Environmental Statistics*, 29(2):346–378, 2024.

Matthias Eckardt, Farnaz Ghorbanpour, and Aila Särkkä. Second-order characteristics for spatial point processes with graph-valued marks. *arXiv preprint arXiv:2410.16903*, 2024.

Stefanie Eckel, Frank Fleischer, Pavel Grabarnik, Marian Kazda, Aila Särkkä, and Volker Schmidt. Modelling tree roots in mixed forest stands by inhomogeneous marked gibbs point processes. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(3):522–539, 2009.

Thomas Fiksel. Estimation of interaction potentials of gibbsian point processes. *Statistics*, 19(1):77–86, 1988.

Hans-Otto Georgii. Canonical and grand canonical gibbs states for continuum systems. *Communications in Mathematical Physics*, 48:31–51, 1976.

Charles J Geyer. Likelihood inference for spatial point processes. In *Stochastic geometry*, pages 79–140. Routledge, 1999.

Charles J Geyer and Jesper Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian journal of statistics*, pages 359–373, 1994.

Charles J Geyer and Elizabeth A Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):657–683, 1992.

Yongtao Guan, Abdollah Jalilian, and Rasmus Waagepetersen. Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(3):677–697, 2015.

W Keith Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Fuchun Huang and Yosihiko Ogata. Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics*, 8(3):510–530, 1999.

Henrike Häbel, Aila Särkkä, Mats Rudemo, Charlott Hamngren Blomqvist, Eva Olsson, and Matias Nordin. Colloidal particle aggregation in three dimensions. *Journal of Microscopy*, 275(3):149–158, 2019.

Adina Iftimi, Marie-Colette van Lieshout, and Francisco Montes. A multi-scale area-interaction model for spatio-temporal point patterns. *Spatial statistics*, 26: 38–55, 2018.

Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons, 2008.

Jens Ledet Jensen and Hans R Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46:475–486, 1994.

Jens Ledet Jensen and Jesper Møller. Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability*, pages 445–461, 1991.

Frank P Kelly and Brian D Ripley. A note on strauss's model for clustering. *Biometrika*, pages 357–360, 1976.

Konstantinos Konstantinou, Farnaz Ghorbanpour, Umberto Picchini, Adam Loavenbruck, and Aila Särkkä. Statistical modeling of diabetic neuropathy: Exploring the dynamics of nerve mortality. *Statistics in Medicine*, 42(23):4128–4146, 2023. doi: https://doi.org/10.1002/sim.9851.

Conor Kresin and Frederic Schoenberg. Parametric estimation of spatial–temporal point processes using the stoyan–grabarnik statistic. *Annals of the Institute of Statistical Mathematics*, pages 1–23, 2023.

Xindi Lin and Hyunseung Kang. Efficient, cross-fitting estimation of semiparametric spatial point processes. *arXiv preprint arXiv:2410.04359*, 2024.

Shigeru Mase. Consistency of the maximum pseudo-likelihood estimator of continuous state space gibbsian processes. *The Annals of Applied Probability*, 5(3): 603–612, 1995.

Shigeru Mase. Marked gibbs processes and asymptotic normality of maximum pseudo-likelihood estimators. *Mathematische Nachrichten*, 209(1):151–169, 2000.

Joseph Mecke. Stationäre zufällige maße auf lokalkompakten abelschen gruppen. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 9:36–58, 1967.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Jesper Møller and Rasmus Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. CRC Press, 2004.

Jesper Møller, Anthony N Pettitt, Robert Reeves, and Kasper K Berthelsen. An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

M Mehdi Moradi. *Spatial and Spatio-Temporal Point Patterns on Linear Networks*. PhD dissertation, University Jaume I, 2018.

Rana A Moyeed and Adrian J Baddeley. Stochastic approximation of the mle for a spatial point pattern. *Scandinavian Journal of Statistics*, pages 39–50, 1991.

Xuan Xanh Nguyen and Hans Zessin. Integral and differential characterizations of the gibbs process. *Mathematische Nachrichten*, 88(1):105–115, 1979.

Yosihiko Ogata. Likelihood estimation of interaction potentials and external fields of inhomogeneous spatial point patterns. In *Proc. Of Pacific Statist. Congress*, pages 150–154, 1986.

Yosihiko Ogata and Koichi Katsura. Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics*, 40:29–39, 1988.

Yosihiko Ogata and Masaharu Tanemura. Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals of the Institute of Statistical Mathematics*, 33(1):315–338, 1981.

Yosihiko Ogata and Masaharu Tanemura. Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3):496–518, 1984.

Yosihiko Ogata and Masaharu Tanemura. Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, pages 421–433, 1985.

Yosihiko Ogata and Masaharu Tanemura. Likelihood estimation of soft-core interaction potentials for gibbsian point patterns. *Annals of the Institute of Statistical Mathematics*, 41:583–600, 1989.

Yosihiko Ogata, Koichi Katsura, and Masaharu Tanemura. Modelling heterogeneous space–time occurrences of earthquakes and its residual analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 52(4):499–509, 2003.

Jens Östling. Point process learning estimation of bandwidth selection for the pair correlation function, 2024. MSc thesis at University of Gothenburg.

Antti Penttinen. Modelling interactions in spatial point patterns: parameter estimation by the maximum-likelihood method. *Comp. Sci., Econ. And Statist.*, 7: 1–107, 1984.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

Tuomas Rajala. A note on bayesian logistic regression for spatial exponential family gibbs point processes. *arXiv preprint arXiv:1411.0539*, 2014.

Brian D Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192, 1977.

Brian D Ripley. *Statistical inference for spatial processes*. Cambridge university press, 1988.

Brian D Ripley and Frank P Kelly. Markov point processes. *Journal of the London Mathematical Society*, 2(1):188–192, 1977.

Mathis Rost. Data augmentation for point process learning, 2024. MSc thesis at University of Gothenburg.

Aila Särkkä. Pseudo-likelihood approach for gibbs point processes in connection with field observations. *Statistics: A Journal of Theoretical and Applied Statistics,*, 26(1):89–97, 1995.

Shinichiro Shirota and Alan E Gelfand. Approximate bayesian computation and model assessment for repulsive spatial point processes. *Journal of Computational and Graphical Statistics*, 26(3):646–657, 2017.

Radu S Stoica, Anne Philippe, Pablo Gregori, and Jorge Mateu. Abc shadow algorithm: a tool for statistical analysis of spatial patterns. *Statistics and computing*, 27:1225–1238, 2017.

Dietrich Stoyan and Pavel Grabarnik. Second-order characteristics for stochastic structures connected with gibbs point processes. *Mathematische Nachrichten*, 151(1):95–100, 1991.

David J Strauss. A model for clustering. *Biometrika*, 62(2):467–475, 1975.

Roland Takacs. Estimator for the pair–potential of a gibbsian point process. *Statistics: A Journal of Theoretical and Applied Statistics*, 17(3):429–433, 1986.

Roland Takacs and Thomas Fiksel. Interaction pair-potentials for a system of ant's nests. *Biometrical Journal*, 28(8):1007–1013, 1986.

Marie-Colette van Lieshout. *Markov Point Processes and Their Applications*. Imperial College Press/World Scientific, 2000.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.