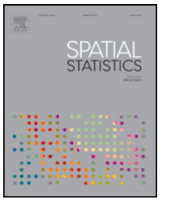
Takacs-Fiksel estimation as a special case of Point Process Learning

Spatial Statistics 74 (2026) 100991

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Takacs-Fiksel estimation as a special case of Point Process Learning☆

Julia Jansson Valter, Ottmar Cronie *

Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Empirical risk minimisation
Gibbs point process model
Innovations
Loss function
Papangelou conditional intensity
Thinning

ABSTRACT

In the context of parameter estimation for Gibbs point processes, the state-of-the-art method is Takacs-Fiksel estimation, of which pseudolikelihood estimation is a special case. An alternative method is the recently proposed Point Process Learning approach, based on point process cross-validation and point process prediction errors. Since both Takacs-Fiksel estimation and Point Process Learning are motivated by the Georgii–Nguyen–Zessin formula, which defines Gibbs point processes, in this paper we study Point Process Learning in relation to Takacs-Fiksel estimation. We show that, upon applying appropriate scaling and letting the cross-validation regime tend to leave-one-out cross-validation in Point Process Learning, averages of prediction errors converge to the innovation-based loss function in Takacs-Fiksel estimation. We further provide an empirical risk formulation of Point Process Learning, which highlights the nature of our asymptotic results, and show that the underlying convergence mechanism can be partially understood through a conditional law of large numbers for statistics of conditionally independent thinnings. We finally illustrate our theoretical findings through simulations for a Strauss process, focusing on both convergence diagnostics and comparison of parameter estimation performance between the two approaches.

1. Introduction

Gibbs point processes are flexible and natural for modelling point patterns with dependence between the points. Such processes can be defined in different but equivalent ways, most notably by locally having densities with respect to Poisson processes or by satisfying the Georgii–Nguyen–Zessin formula, which in turn yields the Papangelou conditional intensity function of the process (Betsch, 2023; Coeurjolly and Lavancier, 2019). For Gibbs point processes, estimation methods based on the Papangelou conditional intensity are both tractable and computationally the most convenient. Here, the state-of-the art is the Takacs-Fiksel estimation method (Takacs, 1986; Fiksel, 1984), which has the pseudolikelihood estimation method as a special case (Coeurjolly and Lavancier, 2019).

An alternative estimation method for Gibbs point processes, also based on the Papangelou conditional intensity, is the recent Point Process Learning approach of Cronie et al. (2024b). Point Process Learning is a prediction-based statistical theory for point processes, which is inspired by previous work by Moradi et al. (2019), Cronie and van Lieshout (2018) and Takacs-Fiksel estimation (Takacs, 1986; Fiksel, 1984). More specifically, it is based on the combination of two concepts that are novel for general 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 via a parametrised Papangelou conditional intensity.

☆ This article is part of a Special issue entitled: 'SPASTA_dawn of AI' published in Spatial Statistics.

* Corresponding author.
E-mail address: ottmar@chalmers.se (O. Cronie).

https://doi.org/10.1016/j.spasta.2026.100991
Received 14 November 2025; Received in revised form 26 April 2026; Accepted 27 April 2026
Available online 23 May 2026

Both Takacs-Fiksel estimation and the prediction errors of Point Process Learning are closely linked to the innovations of Baddeley et al. (2005), which in turn are based on the Georgii–Nguyen–Zessin formula (Georgii, 1976; Nguyen and Zessin, 1979). Consequently, it is natural to suspect a close relationship between Takacs-Fiksel estimation and Point Process Learning. In this paper, we study Point Process Learning in relation to Takacs-Fiksel estimation, specifically by showing that Point Process Learning produces Takacs-Fiksel estimation as a limiting case, when the cross-validation regime tends to leave-one-out cross-validation. More specifically, in Theorems 1 and 2 we establish this relationship when the underlying cross-validation in Point Process Learning is given by Monte-Carlo cross-validation or block cross-validation, respectively. This strengthens the position of Point Process Learning as a powerful method for parameter estimation for Gibbs processes.

To illustrate the theoretical results in practice, we compare Point Process Learning and Takacs-Fiksel estimation, for a Strauss process, using Monte-Carlo cross-validation with increasing values of the number of cross-validation splits. The general Point Process Learning formulation contains several hyperparameters to be specified, namely the cross-validation regime/parameters and a certain test function which determines how much a validation point contributes to the total prediction error. For Takacs-Fiksel estimation, a specific test function choice yields pseudolikelihood estimation, but one might also choose a different test function. Motivated by e.g. Kresin and Schoenberg (2023), in our simulations the test function is fixed to be the so-called Stoyan–Grabarnik test function (Stoyan and Grabarnik, 1991), for both Point Process Learning and Takacs-Fiksel estimation.

The remainder of the paper is organised as follows. In Section 2 we review Gibbs processes, Takacs-Fiksel estimation and Point Process Learning. In Section 3 we develop an empirical risk formulation of Point Process Learning and relate the associated loss functions to theoretical risk quantities. In Section 4 we establish our main results, showing that appropriately scaled averages of prediction errors converge to Takacs-Fiksel innovations, and we provide additional remarks to place these results in a broader context. In Section 5 we present a simulation study which both illustrates the convergence behaviour and compares parameter estimation performance. The Appendix contains additional simulation results and all proofs; throughout, items with the prefix 'A' refer to the Appendix.

2. Preliminaries

Let S be a general (complete separable metric) space with metric $d(\cdot,\cdot)$, which is equipped with a suitable (non-atomic and locally finite) reference measure $A \mapsto |A| = \int_A \mathrm{d}x$, $A \subseteq S$. All sets considered in this paper are Borel sets and a closed ball around $u \in S$, with radius $r > 0$, will be denoted by $b(u,r) = \{v \in S : d(u,v) \leq r\}$. Examples of general spaces include (compact subsets of) the Euclidean space $\mathbb{R}^d$, $d \geq 1$, e.g. $S = [0,1]^2 \subseteq \mathbb{R}^2$ as in our simulation study, spheres and linear networks (Cronie et al., 2020; Baddeley et al., 2015). We consider the assumed generality for the space S to ensure that the theory in this paper is valid beyond the usual setting, where S is a compact subset of some Euclidean space, but the reader may simply think of S as representing the usual setting. Moreover, all functions considered in this paper are taken to be measurable.

A (simple) point process $X = \{x_i\}_{i=1}^N$ in S is a random mechanism whose outcomes are collections of points, so-called point patterns. Hence, we may view 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 $(\mathbf{N}, \mathcal{N})$ (van Lieshout, 2000; Møller and Waagepetersen, 2004). Here, $\mathbf{N}$ is the collection of point patterns/configurations $\mathbf{x} = \{x\}_{i=1}^n \subseteq S$, $0 \leq n \leq \infty$, which are locally finite, i.e. those satisfying that the cardinality $\#(\mathbf{x} \cap A)$ is finite for any bounded $A \subseteq S$. Further, the σ-algebra $\mathcal{N}$ is the Borel σ-algebra generated by a Prohorov-type metric on $\mathbf{N}$ (Daley and Vere-Jones, 2003, Section A2.5–A2.6). Note that, by construction, X is simple, which means that almost surely (a.s.) no two points of X have the same location. The point process X induces a distribution P_X on $(\mathbf{N}, \mathcal{N})$, which is governed by its finite dimensional distributions. Note that X is commonly referred to as a finite point process if $\#(X \cap S) < \infty$ a.s., which e.g. holds whenever S is a bounded set, due to the local finiteness.

2.1. Gibbs point processes

A point process X is called a Gibbs process if its distribution satisfies the Georgii–Nguyen–Zessin (GNZ) formula (see e.g. van Lieshout (2000, Section 1.8.2) or Betsch (2023)), which states that

$$\mathbb{E}\left[\sum_{u \in X} h(u, X \setminus \{u\})\right] = \int_S \mathbb{E}[h(u,X)\lambda_X(u|X)]\mathrm{d}u \tag{1}$$

for any non-negative (potentially infinite) measurable function h on $S \times \mathbf{N}$, whereby it also holds for any integrable function h. The distribution of a Gibbs process X is completely characterised by its (Papangelou) conditional intensity λ_X, which can be interpreted as follows: $\lambda_X(u|\mathbf{x})\mathrm{d}u$ is the probability of finding a point of the point process in an infinitesimal neighbourhood du of $u \in S$, with measure $|du| = \mathrm{d}u$, given that the point process agrees with the configuration $\mathbf{x}$ outside du. 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). In addition, it is called locally stable if there exists a $|\cdot|$-locally integrable function ϕ such that $\lambda_X(u|\cdot) \leq \phi(u) < \infty$, for any $u \in S$ (Betsch, 2023). Conditional intensities have a central role in the study of point processes, e.g. X has the intensity function $\rho_X(u) = \mathbb{E}[\lambda_X(u|X)]$, $u \in S$.

It is often convenient to express the conditional intensity of a parametrised Gibbs model P_θ, $\theta \in \Theta$, as

$$\lambda_\theta(u|\mathbf{x}) = \mathrm{e}^{\Phi_1(u;\theta)+\Phi_2(u,\mathbf{x};\theta)} = \widetilde{\rho}_\theta(u)\,\mathrm{e}^{\Phi_2(u,\mathbf{x};\theta)}, \tag{2}$$

for specific functions Φ_1 and Φ_2, where $P_X = P_{\theta_0}$ and $\lambda_X = \lambda_{\theta_0}$ for some $\theta_0 \in \Theta$. Here, $\widetilde{\rho}_{\theta_0}(u)$ controls the propensity of X to place a point at location $u \in S$, in absence of interaction with other points, while $e^{\Phi_2(u,\mathbf{x};\theta_0)}$ scales up/down $\widetilde{\rho}_{\theta_0}(u)$, depending on the strength and type of interaction between points that the distribution of X exhibits. Note in particular that we here have $\rho_X(u) = \widetilde{\rho}_{\theta_0}(u)\mathbb{E}[e^{\Phi_2(u,X;\theta_0)}]$, which is typically not known explicitly. One exception is a Poisson process, where $\Phi_2(\cdot) = 0$, which implies that $\rho_X(u) = \lambda_X(u|\mathbf{x}) = \widetilde{\rho}_{\theta_0}(u)$.

A Strauss process (Strauss, 1975) is a Gibbs process with parameter vector $\theta = (\beta, R, \gamma)$, where $R > 0$ is the interaction radius, $\beta > 0$ is an intensity-related parameter and $\gamma \in [0, 1]$ is the interaction parameter. Here, $\Phi_1(u;\theta) = \log\beta$ and $\Phi_2(u,\mathbf{x};\theta) = D_R(u,\mathbf{x})\log\gamma$, where $D_R(u;\mathbf{x}) = \sum_{x\in\mathbf{x}\setminus\{u\}} \mathbf{1}\{d(u,x) \le R\}$ counts the R-close neighbours of u in the point pattern $\mathbf{x}$. It is a repulsive point process, with behaviour depending on the interaction parameter γ. When $\gamma = 1$, we obtain a Poisson process with intensity $\beta > 0$ and when $\gamma = 0$, using the convention that $0/0 = 1$, we obtain a Gibbs hard-core process. This process is specified with $\theta = (\beta, R) \in \Theta = (0,\infty)^2$, $\Phi_1(u;\theta) = \log\beta$ and $\Phi_2(u,\mathbf{x};\theta) = \log \mathbf{1}\{u \notin \bigcup_{x\in\mathbf{x}} b(x,R)\}$.

2.2. Statistical setting

Consider the typical setting where we observe a point pattern $\mathbf{x} \in \mathbf{N}$, which is a realisation of a point process X on S, with unknown conditional intensity λ_X. As is commonly the case in parametric Gibbs process modelling, we do not deal with the setting where we observe X restricted to some bounded sub-domain $W \subseteq S$, forcing us to take edge effects into account. Given a model family $\Lambda_\Theta = \{\lambda_\theta : \theta \in \Theta\}$ with Euclidean parameter space Θ, we here assume that the conditional intensity λ_X is given by λ_{θ_0}, for some unknown $\theta_0 \in \Theta$. Our objective is now to find the member λ_θ in Λ_Θ which is "closest to" λ_{θ_0} in some suitable sense. To this end, we need some criterion $\mathcal{L}(\theta;\mathbf{x})$, $\theta \in \Theta$, to optimise in order to obtain an estimate $\widehat{\theta} = \widehat{\theta}(\mathbf{x}) \in \Theta$, which in turn yields our estimate $\lambda_{\widehat{\theta}}$ of λ_{θ_0}.

2.3. Takacs-Fiksel estimation

In Baddeley et al. (2005), h-innovations were introduced as

$$\mathcal{I}^{h_\theta}_{\lambda_\theta}(A;X,X) = \sum_{x\in X\cap A} h_\theta(x; X\setminus\{x\}) - \int_A h_\theta(u;X)\lambda_\theta(u;X)\mathrm{d}u, \tag{3}$$

for a test function $h_\theta : S \times \mathbf{N} \to \mathbb{R}$, which is non-negative or integrable. These innovations form the basis of both point process residuals and Takacs-Fiksel estimation; see e.g. Coeurjolly and Lavancier (2019) for details. In Takacs-Fiksel estimation, the estimate $\widehat{\theta}(\mathbf{x})$ is obtained as the $\theta \in \Theta$ which either (i) minimises the norm/absolute value of $\mathcal{I}^{h_\theta}_{\lambda_\theta}(S;\mathbf{x},\mathbf{x})$ or (ii) yields that $\mathcal{I}^{h_\theta}_{\lambda_\theta}(S;\mathbf{x},\mathbf{x}) = 0$; Coeurjolly et al. (2016) consider the latter. The motivation here is that by the GNZ formula we have that $\mathbb{E}[\mathcal{I}^{h_\theta}_{\lambda_\theta}(S;X,X)] = 0$ when $\theta = \theta_0$. When the aim is to solve $\mathcal{I}^{h_\theta}_{\lambda_\theta}(S;\mathbf{x},\mathbf{x}) = 0$, in order to have a solvable system of equations, one considers a vector-valued test function $h_\theta(\cdot) = (h^1_\theta(\cdot),\dots,h^d_\theta(\cdot))^T \in \mathbb{R}^d$, where $d \ge 1$ is equal to the dimension of Θ. Depending on the model Λ_Θ and the test function h_θ, this mathematically driven approach is, however, not always simple to carry out (Coeurjolly et al., 2016). Hereby, numerically minimising the norm $\theta \mapsto \|\mathcal{I}^{h_\theta}_{\lambda_\theta}(S;\mathbf{x},\mathbf{x})\|$, with respect to $\theta \in \Theta$, may be a more appealing alternative, in particular if the model has many parameters or a complex interaction structure.

A prominent special case of Takacs-Fiksel estimation is pseudolikelihood estimation. Here, the vectorial test function used in the innovation is given by the normalised gradient $h_\theta(\cdot) = \nabla_\theta\lambda_\theta(\cdot)/\lambda_\theta(\cdot) \in \mathbb{R}^d$, where d is the dimension of Θ. Setting the resulting innovation to 0 is equivalent to maximising the log-pseudolikelihood function

$$\theta \mapsto \sum_{x\in\mathbf{x}} \log\lambda_\theta(x|\mathbf{x}\setminus\{x\}) - \int_S \lambda_\theta(u|\mathbf{x})\mathrm{d}u, \qquad \theta \in \Theta.$$

In particular, for a Poisson process model with intensity $\rho_\theta(u) = \lambda_\theta(u|\cdot)$, $u \in S$, the above is in fact the actual log-likelihood function and the innovation is the associated score function to be minimised. It further holds that the popular maximum logistic regression likelihood method of Baddeley et al. (2014) is a numerically stable approximation of pseudolikelihood estimation (Coeurjolly et al., 2016).

2.4. Point Process Learning

We here briefly recall the statistical methodology called Point Process Learning. As previously mentioned, it is based on the combination of point process cross-validation and point process prediction errors.

2.4.1. Point process cross-validation

The general idea of cross-validation is to repeatedly split a dataset into a training set and a validation set (see e.g. Arlot and Celisse (2010)); common examples include leave-one-out cross-validation and k-fold cross-validation. Cronie et al. (2024b) introduced

thinning as an instance of binary marking (Cronie et al., 2024a; D'Angelo et al., 2023) and cross-validation splitting/partitioning as generating a collection of pairs (X_i^T, X_i^V), $X_i^T = X \setminus X_i^V$, $i = 1, \dots, k \geq 1$, where $X_1^V, \dots, X_k^V$ are thinnings of X. For mathematical tractability reasons, independent thinning-based cross-validation is, in general, particularly appealing. The most straightforward method here is Monte-Carlo cross-validation, where all $\mathbf{x}_i^V$ are obtained by attaching independent and identically distributed (iid) marks with a Bernoulli distribution with parameter p to the points of $\mathbf{x}$. Note that we here may have that $\mathbf{x}_i^V \cap \mathbf{x}_j^V \neq \emptyset$, $i \neq j$, when $k \geq 2$. As an alternative, where $\mathbf{x}_i^V \cap \mathbf{x}_j^V = \emptyset$, we have multinomial k-fold cross-validation, which has a hierarchical structure. Here we independently attach marks $m(x)$ to all $x \in \mathbf{x}$, where the mark distribution is given by a multinomial distribution with $p_i(x) = \mathbb{P}(m(x) = i) = 1/k$, $i = 1, \dots, k$. We then let $\mathbf{x}_i^V = \{x \in \mathbf{x} : m(x) = i\}$ and $\mathbf{x}_i^T = \mathbf{x} \setminus \mathbf{x}_i^V$, $i = 1, \dots, k$. A further alternative mentioned by Cronie et al. (2024b) is block cross-validation, where we instead let $p_i(x) = \mathbb{P}(m(x) = i) = \mathbf{1}\{x \in \mathcal{S}_i\}$, $i = 1, \dots, k$, for a fixed partition $\{\mathcal{S}_i\}_{i=1}^k$ of $\mathcal{S}$. In this paper, we will only focus on independent thinning-based cross-validation, so that X^V is an independent thinning of X, based on the retention probability function $p(u) \in (0, 1)$, $u \in \mathcal{S}$.

2.4.2. Point process prediction errors

For a given training–validation pair (X^T, X^V) and test function $h_\theta : \mathcal{S} \times \mathbf{N} \to \mathbb{R}$, the associated point process prediction error is given by

$$\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V) = \sum_{x \in X^V \cap A} h_\theta(x; X^T) - \int_A h_\theta(u; X^T) \xi_\theta(u; X^T) \mathrm{d}u \tag{4}$$

for any $A \subseteq \mathcal{S}$, where $\xi_\theta(u; X^T) = p(u)\mathbb{E}[\lambda_\theta(u|X)|X^T]$ since we consider independent thinning-based cross-validation by means of a retention probability function $p(\cdot)$. Prediction errors for a point pattern $\mathbf{x}$ are obtained by replacing (X^T, X^V) by $(\mathbf{x}^T, \mathbf{x}^V)$. See Cronie et al. (2024b, Definition 4) for a general definition for two arbitrary point processes, Z and Y.

Cronie et al. (2024b, Theorem 2) state that, for any $A \subseteq \mathcal{S}$ and any test function choice, the prediction error $\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)$ has expectation 0 if and only if the parameter θ is set to the true parameter. Motivated by this, the idea of Point Process Learning is to find a choice $\theta = \widehat{\theta} \in \Theta$ such that $\mathbb{E}[\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)]$ is as close to 0 as possible.

As pointed out in Cronie et al. (2024b), there is a link between point process prediction errors and the innovations of Baddeley et al. (2005). Namely, by letting $X^T = X^V = X$ in a prediction error and setting $\xi_\theta(u; X) = \lambda_\theta(u|X)$, the prediction errors in (4) reduce to the h-innovations found in (3). This motivates exploring how Point Process Learning is related to Takacs-Fiksel estimation and, in Section 4, we show that Takacs-Fiksel estimation is a limiting case of Point Process Learning.

2.5. Remarks on test functions

That the test function is allowed to depend on the model parameters highlights the fact that it is natural to let $h_\theta(\cdot) = f(\lambda_\theta(\cdot))$ for some suitable function $f : \mathbb{R} \to \mathbb{R}$. One interesting such choice is

$$h_\theta(\cdot) = 1/\xi_\theta(\cdot)^\alpha, \qquad \alpha \geq 0, \tag{5}$$

which becomes $h_\theta(\cdot) = 1/\lambda_\theta(\cdot)^\alpha$ in the case of innovations and Takacs-Fiksel estimation (Cronie et al., 2024b). The test function in (5) with $\alpha = 0$ results in the raw test function, while $\alpha = 1/2$ renders the Pearson test function (Baddeley et al., 2005). The choice $\alpha = 1$ corresponds to the Stoyan and Grabarnik (1991) test function, also called the inverse test function (Baddeley et al., 2005), which has been a common component in various statistical approaches found in the literature (Cronie and van Lieshout, 2018; Cronie et al., 2024b; Cronie and van Lieshout, 2016; Kresin and Schoenberg, 2023; Stoyan and Stoyan, 2000). A particular appeal of the Stoyan–Grabarnik test function is that it sets the integral in the prediction error (recall (4)) to $|A|$, if either ξ_θ is strictly positive a.e. or under the convention that $0/0 = 1$. Moreover, given the omnipresent variance–bias trade-off in estimation, it mainly seems to target the variance of the estimator (Moradi et al., 2019).

3. Empirical risk formulation and loss functions

In Section 2 we saw that the objectives of Point Process Learning and Takacs-Fiksel estimation are to find choices $\theta = \widehat{\theta} \in \Theta$ which minimise the magnitudes of $\mathbb{E}[\mathcal{I}_{\xi_\theta}^{h_\theta}(\mathcal{S}; X^T, X^V)]$ and $\mathbb{E}[\mathcal{I}_{\lambda_\theta}^{h_\theta}(\mathcal{S}; X, X)]$, respectively. These objectives raise the natural question whether Point Process Learning can be formulated through the lens of (empirical) risk minimisation, thus following the path of classical statistical learning (Vapnik, 1999) and further justifying the name Point Process Learning.

3.1. Risk functions

We note that we can let the risk, which we aim to minimise in order to obtain an estimator $\widehat{\theta}$ for θ_0, be given by

$$R(\theta) = \mathbb{E}[\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)]^2, \qquad \theta \in \Theta, \tag{6}$$

which satisfies the bounds

$$\begin{aligned} R(\theta)^{1/2} &= |\mathbb{E}[\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)]| \leq \mathbb{E}[|\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)|] = R_1(\theta), \\ R(\theta) &\leq \mathbb{E}[\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)^2] = \mathrm{Var}(\mathcal{I}_{\xi_\theta}^{h_\theta}(A; X^T, X^V)) + R(\theta) = R_2(\theta), \end{aligned} \tag{7}$$

by Jensen's inequality, since $x \mapsto |x|^v$, $v \geq 1$, is convex; note that by Cronie et al. (2024b, Theorem 2) the risk coincides with the squared prediction error bias, i.e. $R(\theta) = \mathbb{E}[\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)]^2 = \mathbb{E}[\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V) - \mathcal{I}^{h_{\theta_0}}_{\xi_{\theta_0}}(A; X^T, X^V)]^2$. In addition, by Hölder's inequality we have that $R_1(\theta) \leq R_2(\theta)^{1/2}$, so $R_2(\theta) = 0$ implies that $R_1(\theta) = 0$ and $R(\theta) = 0$, and $R_1(\theta) = 0$ implies that $R(\theta) = 0$. Note further that although $R(\theta)^{1/2}$ and $R(\theta)$ have the same minimisers, their respective upper bounds $R_1(\theta)$ and $R_2(\theta)$ do not have to have the same minimisers as each other.

3.2. Empirical risk estimation

Unfortunately, deriving an exact expression for the risk in (6), which depends on a range of different things, including the specific model and the cross-validation approach employed, is generally a hard task. Hence, we proceed by following the path of statistical learning and estimate (6) by means of the empirical risk.

3.2.1. Empirical risk estimation based on repeated sampling

Given iid copies $\widetilde{X}_1, \dots, \widetilde{X}_n$ of the underlying point process X, consider $\widetilde{\mathbf{X}}_n = \{(\widetilde{X}_i^T, \widetilde{X}_i^V)\}_{i=1}^n$, where $(\widetilde{X}_i^T, \widetilde{X}_i^V)$ is one training–validation pair generated from X_i, based on the chosen thinning regime. Now, classical large sample arguments suggest using

$$\widehat{R}^{\mathcal{T}}(\theta; \widetilde{\mathbf{X}}_n) = \left(\frac{1}{\#\mathcal{T}(\widetilde{\mathbf{X}}_n)} \sum_{i \in \mathcal{T}(\widetilde{\mathbf{X}})} \mathcal{I}^{h_\theta}_{\xi_\theta}(A; \widetilde{X}_i^V, \widetilde{X}_i^T) \right)^2 \approx R(\theta), \qquad \theta \in \Theta,$$

to estimate $R(\theta)$. Here, $\mathcal{T}$ generates an index set $\mathcal{T}(\widetilde{\mathbf{X}}_n) \subseteq \{1, \dots, n\}$, determining which of the training–validation pairs we should include, where natural choices are given by $\mathcal{T}(\widetilde{\mathbf{X}}_n) = \{1, \dots, n\}$ and $\mathcal{T}(\widetilde{\mathbf{X}}_n) = \{i \in \{1, \dots, n\} : \widetilde{X}_i^T \neq \emptyset, \widetilde{X}_i^V \neq \emptyset\}$. Note that the latter choice, where we exclude all pairs where either $\widetilde{X}_j^T$ or $\widetilde{X}_j^V$ is empty, often makes sense in practice, since it might not make sense to predict the empty set from all of $\widetilde{X}_j$, or all of $\widetilde{X}_j$ from the empty set. Note further that, in practice, one would consider realisations $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_n$ of $\widetilde{X}_1, \dots, \widetilde{X}_n$ and let $\widetilde{\mathbf{X}}_n = \{(\widetilde{\mathbf{x}}_i^T, \widetilde{\mathbf{x}}_i^V)\}_{i=1}^n$, resulting in the empirical risk estimate $\widehat{R}^{\mathcal{T}}(\theta; \{(\widetilde{\mathbf{x}}_i^T, \widetilde{\mathbf{x}}_i^V)\}_{i=1}^n)$, $\theta \in \Theta$, to be minimised in order to obtain an estimate $\widehat{\theta}$.

3.2.2. Empirical risk estimation based on a single realisation and loss functions

As laid out in Section 2.2, in this paper we consider the typical statistical setting, where we only have access to one single realisation $\mathbf{x}$ of the underlying point process X. The empirical risk above would here be given by $\widehat{R}^{\mathcal{T}}(\theta; \{(\widetilde{\mathbf{x}}^T, \widetilde{\mathbf{x}}^V)\})$, $\theta \in \Theta$, which e.g. in the case of exclusion of empty training or validation sets reads $\widehat{R}^{\mathcal{T}}(\theta; \{(\widetilde{\mathbf{x}}^T, \widetilde{\mathbf{x}}^V)\}) = \mathbf{1}\{\mathbf{x}^T, \mathbf{x}^V \neq \emptyset\} \mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}^T, \mathbf{x}^V)^2$, $\theta \in \Theta$. Since there is a significant risk that we obtain a training–validation pair $(\mathbf{x}^T, \mathbf{x}^V)$ for which the magnitude of $\mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}^T, \mathbf{x}^V) - \mathbb{E}[\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)]$ is large, even if $\theta = \theta_0$, we conclude that using only one training–validation pair is not ideal.

Instead of appealing to classical iid large sample theory to estimate the expectation in (6), as we did above, if the realisation $\mathbf{x} = X(\omega)$ is central in the distribution of X we note that $R(\theta) = \mathbb{E}[\mathbb{E}[\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)|X]]^2 \approx \mathbb{E}[\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)|X = \mathbf{x}]^2$. Note that the last term can be treated as an integral with respect to a probability kernel (all involved spaces are topologically Polish) and the only randomness which remains in it comes from the cross-validation/thinning procedure. Now, given a cross-validation $\mathbf{X}_k = \{(X_i^T, X_i^V)\}_{i=1}^k$ of X, based on some thinning procedure, this suggests employing

$$\widehat{R}^{\mathcal{T}}(\theta; \mathbf{X}_k) = \widehat{R}^{\mathcal{T}}(\theta; \{(X_i^T, X_i^V)\}_{i=1}^k) = \left(\frac{1}{\#\mathcal{T}(\mathbf{X}_k)} \sum_{i \in \mathcal{T}(\mathbf{X}_k)} \mathcal{I}^{h_\theta}_{\xi_\theta}(A; X_i^T, X_i^V) \right)^2$$

for (conditional iid) estimation of $\mathbb{E}[\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)|X] \approx R(\theta)$.

Similarly, we can consider the empirical counterparts of the upper bounds $R_j(\theta) = \mathbb{E}[\mathbb{E}[|\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)|^j|X]] \approx \mathbb{E}[|\mathcal{I}^{h_\theta}_{\xi_\theta}(A; X^T, X^V)|^j|X = \mathbf{x}]$, $j = 1, 2$, in (7), i.e.

$$\widehat{R}_j^{\mathcal{T}}(\theta; \mathbf{X}_k) = \widehat{R}_j^{\mathcal{T}}(\theta; \{(X_i^T, X_i^V)\}_{i=1}^k) = \frac{1}{\#\mathcal{T}(\mathbf{X}_k)} \sum_{i \in \mathcal{T}(\mathbf{X}_k)} \left| \mathcal{I}^{h_\theta}_{\xi_\theta}(A; X_i^T, X_i^V) \right|^j, \qquad j = 1, 2,$$

which aim to approximate bounds of conditional expectations involving X by averages based on the subsamples in $\mathbf{X}_k$. Note that in each of the cases above, we have to make a choice for how $\mathcal{T}$ includes/excludes training–validation pairs and, as before, natural choices include $\mathcal{T}(\mathbf{X}_k) = \{1, \dots, k\}$ and $\mathcal{T}(\mathbf{X}_k) = \{i \in \{1, \dots, k\} : \widetilde{X}_i^T \neq \emptyset, \widetilde{X}_i^V \neq \emptyset\}$.

If we replace X by the observed point pattern $\mathbf{x}$ and thus let $\mathbf{X}_k = \{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^k$ be a cross-validation of $\mathbf{x}$, we obtain the loss functions

$$\mathcal{L}_j(\theta) = \widehat{R}_j^{\mathcal{T}}(\theta; \{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^k) = \frac{1}{\#\mathcal{T}(\mathbf{X}_k)} \sum_{i \in \mathcal{T}(\mathbf{X}_k)} \left| \mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}_i^T, \mathbf{x}_i^V) \right|^j, \qquad j = 1, 2, \tag{8}$$

$$\mathcal{L}_3(\theta) = \widehat{R}^{\mathcal{T}}(\theta; \{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^k) = \left(\frac{1}{\#\mathcal{T}(\mathbf{X}_k)} \sum_{i \in \mathcal{T}(\mathbf{X}_k)} \mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}_i^T, \mathbf{x}_i^V) \right)^2, \tag{9}$$

typically with $A = S$; here, $\mathcal{T}(\mathbf{X}_k) = \emptyset$ results in both (8) and (9) returning the value 0. These are exactly the loss functions originally proposed by Cronie et al. (2024b). Recalling the discussion below (7) about the relationships between $R(\theta)$, $R_1(\theta)$ and $R_2(\theta)$, similarly,

$$\mathcal{L}_3(\theta) \leq \mathcal{L}_2(\theta), \qquad \mathcal{L}_3(\theta) \leq \mathcal{L}_1(\theta)^2, \qquad \mathcal{L}_1(\theta)^2 \leq \mathcal{L}_2(\theta), \tag{10}$$

$$\mathcal{L}_2(\theta) - \mathcal{L}_3(\theta) = \frac{1}{\#\mathcal{T}(\mathbf{X}_k)} \sum_{i \in \mathcal{T}(\mathbf{X}_k)} \left(\mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}_i^T, \mathbf{x}_i^V) - \frac{1}{\#\mathcal{T}(\mathbf{X}_k)} \sum_{j \in \mathcal{T}(\mathbf{X}_k)} \mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}_j^T, \mathbf{x}_j^V) \right)^2,$$

by Jensen's inequality, the triangle inequality, the Cauchy–Schwarz inequality and properties of squared sums. Since minimisation of $R_2(\theta)$ targets both the prediction error variance and the squared bias, we anticipate that $\mathcal{L}_2$ tends to emphasise a lower estimator variance than $\mathcal{L}_3$, at the cost of a higher estimator bias. Similarly, $\mathcal{L}_1$ can be interpreted as targeting a robust measure of central tendency (related to the median), whereby we expect it to emphasise bias less than $\mathcal{L}_3$. Hereby, $\mathcal{L}_1$ and $\mathcal{L}_2$ should yield similar results when the distribution of the iid random variables $\mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}_i^T, \mathbf{x}_i^V)$, $i \in \mathcal{T}(\mathbf{X})$, is not too asymmetric. Moreover, when considering test functions which are allowed to take negative values, cancellation between different $\mathcal{I}^{h_\theta}_{\xi_\theta}(A; \mathbf{x}_i^T, \mathbf{x}_i^V)$, $i \in \mathcal{T}(\mathbf{X})$, may take place in $\mathcal{L}_3$, which indicates that it might not always be the most appropriate choice in practice.

4. Takacs-Fiksel estimation as a limiting case of Point Process Learning

We now arrive at the main results of the paper, which essentially state that an innovation constitutes a limit of averages of scaled prediction errors, under specific assumptions on the cross-validation regime employed. Specifically, the cross-validation regime should tend to leave-one-out cross-validation. In other words, the results tell us that an innovation (Takacs-Fiksel estimation) is a limiting special case of an average of prediction errors, which is reflected by the $\mathcal{L}_3$ loss function. Hence, comparison of Takacs-Fiksel estimation and the general Point Process Learning setup can be expressed as a comparison of Point Process Learning with a specific hyperparameter choice and Point Process Learning with a free choice of hyperparameters.

Theorem 1, whose proof can be found in Appendix A.2.1, addresses the convergence when we apply Monte-Carlo cross-validation, let $\mathcal{T}(\mathbf{X}_k) = \{1, \dots, k\}$ and $k \to \infty$. Heuristically, it tells us that by employing the $\mathcal{L}_3$ loss function we approximately obtain h-innovations, when k is large and $p = 1/\sqrt{k}$. Hence, to perform Takacs-Fiksel estimation one could, in theory, carry out Point Process Learning with such a setup for the cross-validation regime.

Theorem 1. *Assume that $\lambda_\theta(u|\mathbf{x})$ and $h_\theta(u; \mathbf{x})$ are bounded for any $\theta \in \Theta$, $u \in S$ and $\mathbf{x} \in \mathbf{N}$. Moreover, for any $k \geq 2$, 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, for all $\epsilon > 0$ and for any bounded $A \subseteq S$,*

$$\lim_{k \to \infty} \mathbb{P}\left(\left| p_k \sum_{i=1}^k \mathcal{I}^{h_\theta}_{\xi_\theta}(A; X_i^V(p_k), X_i^T(p_k)) - \mathcal{I}^{h_\theta}_{\lambda_\theta}(A; X, X) \right| > \epsilon \right) = 0.$$

We next provide a similar result, where $\mathcal{T}(\mathbf{X}_k) = \{1, \dots, k\}$ and $k \to \infty$, but this time under block-cross-validation; its proof can be found in Appendix A.2.2. Here we consider partitions which become finer and finer, with the size of any member of the partition rendering a fold-retention probability tending to 0.

Theorem 2. *Assume that $\lambda_\theta(u|\mathbf{x})$ and $h_\theta(u; \mathbf{x})$ are bounded for any $\theta \in \Theta$, $u \in S$ and $\mathbf{x} \in \mathbf{N}$. Given a bounded $A \subseteq S$, let $\{(X_{ik}^T, X_{ik}^V)\}_{i=1}^k$, $k \geq 2$, be block cross-validations of $X \cap A$, based on partitions $\{A_{ik}\}_{i=1}^k$ of A, with associated retention probabilities $p_{ik}(u) = \mathbf{1}\{u \in A_{ik}\}$, $i = 1, \dots, k$. Assume further that the partition sizes satisfy $\max_{i=1,\dots,k} |A_{ik}| \to 0$ as $k \to \infty$ and that for any $i = 1, \dots, k$ there exists only one $j = 1, \dots, k+1$ such that $A_{ik} \subseteq A_{j(k+1)}$, i.e. we have a refinement. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left| \sum_{i=1}^k \mathcal{I}^{h_\theta}_{\xi_\theta}(A; X_{ik}^V, X_{ik}^T) - \mathcal{I}^{h_\theta}_{\lambda_\theta}(A; X \cap A, X \cap A) \right| > \varepsilon \right) = 0.$$

It should be emphasised that the boundedness conditions imposed in Theorems 1 and 2 indicate that the convergence might not hold for any model and test function combination. We note, however, that the imposed conditional intensity boundedness implies that the results hold at least for all locally stable models. Turning to the test function, in Corollary 1 below, which we prove in Appendix A.2.3, we show that the boundedness conditions hold for a large class of Gibbs models in combination with the test function (5).

Corollary 1. *Let $\lambda_\theta(u|\mathbf{x}) = \mathrm{e}^{\Phi_1(u;\theta) + \Phi_2(u,\mathbf{x};\theta)}$, $\theta \in \Theta$, be a Gibbs model where Θ is bounded and $|\Phi_1(u;\theta)|, |\Phi_2(u, \mathbf{x}; \theta)|$ are bounded for any $u \in A$ and any $\mathbf{x} \in \mathbf{N}$ such that $\mathbf{x} \subseteq A$. Assume further that the test function is given by (5) and that the conditional intensity of X is given by λ_{θ_0}, $\theta_0 \in \Theta$. It then follows that the conditions of Theorems 1 and 2 are satisfied.*

Note that there is an immediate example where Corollary 1 is not satisfied when using the test function in (5), with $\alpha > 0$, and that is the Gibbs hard-core process, presented in Section 2.1. The support of $\lambda_\theta(\cdot|\mathbf{x})$ is here given by $S \setminus \bigcup_{x \in \mathbf{x}} b(x, R)$, whereby $h_\theta(u; \mathbf{x})$ is infinite for $u \in \bigcup_{x \in \mathbf{x}} b(x, R)$. A way to resolve this issue, in general, is to replace the test function by a new truncated version of the test function, given by $\min\{h_\theta(\cdot), C\}$, for some large constant $C > 0$.

4.1. Additional remarks

Although the results and observations above are satisfactory for our purposes, a few remarks are in place. To begin with, we conjecture that both Theorems 1 and 2 can be proved under less restrictive boundedness conditions.

4.1.1. Generalised multinomial cross-validation

We further conjecture that a result of a similar spirit could also be obtained for a generalisation of multinomial k-fold and block cross-validation, which we refer to as generalised multinomial cross-validation and present in Definition 1.

Definition 1. Given $k \geq 2$, consider a collection of $k \geq 2$ functions $p_i(u) \in [0,1]$, $u \in S$, $i = 1, \dots, k$, satisfying $\sum_{i=1}^{k} p_i(u)\mathrm{d}u = 1$ for any $u \in S$. Then, attach iid marks $m(x)$ to all $x \in \mathbf{x}$, according to $\mathbb{P}(m(x) = i) = p_i(x)$, $i = 1, \dots, k$. We define *generalised multinomial (k-fold) cross-validation* by letting $\mathbf{x}_i^V = \{x \in \mathbf{x} : m(x) = i\}$ and $\mathbf{x}_i^T = \mathbf{x} \setminus \mathbf{x}_i^V$, $i = 1, \dots, k$.

4.1.2. Conditional weak law of large numbers

In Theorems 1 and 2 we have carried out the rather involved task of showing that under certain specific scalings and structures of the prediction errors, the attained limit coincides with the loss underlying Takacs-Fiksel estimation, i.e. an innovation. Now, it turns out that the convergence mechanism underlying Theorems 1 and 2 can be partially understood as a conditional law of large numbers principle for statistics of conditionally independent thinnings. Its statement can be found in Lemma 1 below while its proof can be found in Appendix A.2.4.

Lemma 1. *Given a point process X on S and any bounded $A \subseteq S$, for any $k \geq 1$, let $X_{i,k}^V$, $i = 1, \dots, k$, be conditionally iid given X and let $T_k : \mathbf{N} \to \mathbb{R}$ be such that the common conditional expectation $\mu_k(X) = \mathbb{E}[T_k(X_{i,k}^V \cap A)|X]$ is well-defined. If $\mathbb{E}[k^{-1}\operatorname{Var}(T_k(X_{1,k}^V \cap A)|X)] \to 0$ when $k \to \infty$, then*

$$\frac{1}{k}\sum_{i=1}^{k} T_k(X_{i,k}^V \cap A) - \mu_k(X) \to 0$$

in probability as $k \to \infty$. If we additionally have that $\mu_k(X)$ tends to $T(X \cap A)$ in probability when $k \to \infty$, for some $T : \mathbf{N} \to \mathbb{R}$, it also follows that

$$\frac{1}{k}\sum_{i=1}^{k} T_k(X_{i,k}^V \cap A) - T(X \cap A) \to 0$$

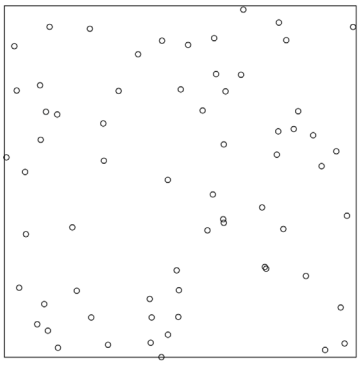
in probability, as $k \to \infty$.

A part of the convergence in Theorems 1 and 2 is a manifestation of the general conditional weak law in Lemma 1, which shows that averaging statistics over conditionally independent validation sets yields convergence towards their conditional expectation. However, Lemma 1 does not explicitly identify the limiting object. To do so, Theorems 1 and 2 evaluate how thinning mechanisms interact with conditional intensities and test functions, in order to show that the limiting conditional expectation coincides with an innovation, for specific choices of prediction errors and appropriate scalings with respect to the chosen cross-validation regime. Hence, the substantive point in Theorems 1 and 2 is not only the conditionally independent averaging represented by Lemma 1, but also the identification of appropriate scalings and limiting conditional means to turn averaged prediction errors into innovations.

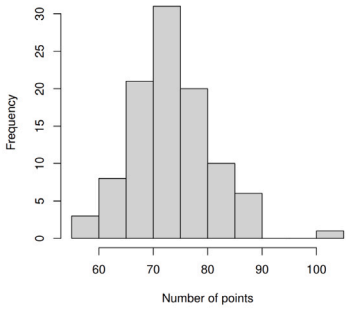
Remark 4.1. Note first that in Lemma 1 we may let all the statistics T_k coincide, i.e. let $T_k(\cdot) = T(\cdot)$, $k \geq 1$. Note, in addition, that one in fact can relax Lemma 1 by requiring that X_{ik}^V, $i = 1, \dots, k$, $k \geq 1$, are only conditionally independent (without identical distribution). Given $\mu_{ik}(X) = \mathbb{E}[T_{ik}(X_{ik}^V \cap A)|X]$ and $T_{ik} : \mathbf{N} \to \mathbb{R}$, $i = 1, \dots, k$, $k \geq 1$, and imposing that $\lim_{k\to\infty} \mathbb{E}[k^{-2}\sum_{i=1}^{k} \operatorname{Var}(T_{ik}(X_{ik}^V \cap A)|X)] = 0$, analogous arguments yield that $k^{-1}\sum_{i=1}^{k} T_{ik}(X_{ik}^V \cap A) - \mu_{ik}(X) \to 0$ in probability as $k \to \infty$. If, in addition, there is a function $T : \mathbf{N} \to \mathbb{R}$ such that $\frac{1}{k}\sum_{i=1}^{k} \mu_{ik}(X) \to T(X \cap A)$ in probability, as $k \to \infty$, we also have $\frac{1}{k}\sum_{i=1}^{k} T_{ik}(X_{ik}^V \cap A) - T(X \cap A) \to 0$ in probability as $k \to \infty$.

4.1.3. The central role of $\mathcal{L}_3$

Recalling (10), we note that existence of the empirical risk limit $\lim_{k\to\infty} \widehat{R}(\theta; \mathbf{X}_k)$, i.e. convergence of $\mathcal{L}_3(\theta)$, does not guarantee convergence of $\mathcal{L}_1(\theta) = \widehat{R}_1(\theta; \mathbf{X}_k)$ or $\mathcal{L}_2(\theta) = \widehat{R}_2(\theta; \mathbf{X}_k)$. The inequalities in (10) show that $\mathcal{L}_3$ is the smallest of the three loss functions, $\mathcal{L}_2$ is the largest and $\mathcal{L}_1$ is intermediate; by (7), we have complete analogy for the theoretical counterparts $R(\theta)$, $R_1(\theta)$ and $R_2(\theta)$. Moreover, the identity for $\mathcal{L}_2(\theta) - \mathcal{L}_3(\theta)$ in (10) reveals that $\mathcal{L}_2$ decomposes into the squared mean prediction error $\mathcal{L}_3$ and the empirical variance of the prediction errors across cross-validation splits, so $\mathcal{L}_3$ isolates the squared average (signed) prediction error, while $\mathcal{L}_2$, as well as $\mathcal{L}_1$, additionally incorporates variability across validation sets. Since innovations are defined as signed sums of contributions, there may be cancellation between positive and negative terms. $\mathcal{L}_3$, which is based on the squared average of the signed prediction errors, preserves this cancellation structure while $\mathcal{L}_1$ and $\mathcal{L}_2$ attenuate/remove such cancellation by taking absolute values or squaring before averaging. Since Theorems 1 and 2 and Lemma 1 involve averages of signed sums, the asymptotic results are tied specifically to the only loss function which reflects signed sums, namely $\mathcal{L}_3$. Put differently, only $\mathcal{L}_3$ targets the same limiting object as these results do. Note, in particular, that Lemma 1 yields convergence to a conditional expectation, whereas Theorems 1 and 2 show that this expectation indeed coincides with an innovation.



(a) Example of a point pattern used in the simulation study.



(b) Histogram for simulated Strauss process point counts.

Fig. 1. Simulated realisations from the Strauss process on $S = [0,1]^2$ with parameters $R = 0.05, \beta = 100$ and $\gamma = 0.5$. Left: Illustration of a simulated realisation. Right: Histogram for the total point counts.

4.1.4. Relation to other asymptotic regimes

It should finally be emphasised that the asymptotic regime considered here is different than the increasing domain regime commonly encountered in the context of a stationary point process X in a Euclidean domain. For instance, given a Gibbs process X in $S = \mathbb{R}^d$, $d \geq 1$, generated by a distribution belonging to a wide family of models, Schreiber and Yukich (2013, Theorem 2.1) consider sub-domains $A_n = [-n^{1/d}/2, n^{1/d}/2]^d \subseteq \mathbb{R}^d$, $n \geq 1$, and show that, for any bounded $f : A_1 \to \mathbb{R}$ and any non-negative so-called stabilising test function h, the random integrals/sums $Y_n^h[f] = \frac{1}{n}\int_{\mathbb{R}^d} f(u) X_n^h(du)$, $n \geq 1$, generated by the random measures $X_n^h(A) = \sum_{x \in X \cap A_n} \mathbf{1}\{n^{-1/d} x \in A\} h(x, X \cap A_n \setminus \{x\})$, $A \subseteq \mathbb{R}^d$, converge in quadratic mean, and thereby in probability, to $\tau_X \mathbb{E}[h(o, X)\lambda_X(o|X)] \int_{A_1} f(u)\mathrm{d}u$, for specific $\tau_X > 0$; here o denotes the origin in $\mathbb{R}^d$. As an equivalent result also holds for marked Gibbs processes (Schreiber and Yukich, 2013, Remark Section 2), one could potentially exploit the results in Schreiber and Yukich (2013) to prove a similar increasing domain result for the prediction errors in (4), under corresponding regularity on X. Given the different (and difficult) nature of this problem, compared to the thinning-based one considered here, the required assumptions in the increasing domain regime need to be more restrictive than the ones considered here. Our assumed regime should also be contrasted to the infill asymptotic regime, considered by e.g. van Lieshout (2021), where one studies the convergence of functionals whose inputs are given by superpositions $\bigcup_{i=1}^n X_i$, $n \geq 1$, based on a sequence $X_1, X_2, \ldots$ of iid point processes.

5. Simulation study

To validate our findings, we carried out a simulation study to illustrate Theorem 1. Concerning the setup for Point Process Learning, we considered Monte-Carlo cross-validation, in order to reflect Theorem 1. We then compared this with Takacs-Fiksel estimation, which was effectively implemented as Point Process Learning using leave-one-out cross-validation. We computed the absolute difference, i.e.

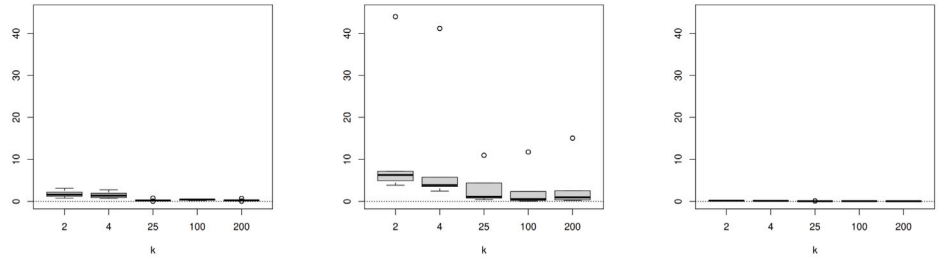
$$\left| p_k \sum_{i=1}^{k} \mathcal{I}_{\xi_\theta}^{h_\theta}(A; \mathbf{x}_i^V(p_k), \mathbf{x}_i^T(p_k)) - \mathcal{I}_{\lambda_\theta}^{h_\theta}(A; \mathbf{x}, \mathbf{x}) \right|, \tag{11}$$

for a simulated point pattern $\mathbf{x}$, to verify that this quantity decreases as k increases. Given a simulated point pattern $\mathbf{x}$ and a fixed value of k, to calculate the first term, i.e. $p_k \sum_{i=1}^{k} \mathcal{I}_{\xi_\theta}^{h_\theta}(A; \mathbf{x}_i^V(p_k), \mathbf{x}_i^T(p_k))$, we carried out Monte-Carlo cross-validation on the point pattern $\mathbf{x}$, obtaining $\{\mathbf{x}_i^V(p_k), \mathbf{x}_i^T(p_k)\}_{i=1}^k$, where the retention probability is $p_k = 1/\sqrt{k}$. We then computed the prediction error for each $i = 1, \ldots, k$, summed over these terms and weighted by $p_k = 1/\sqrt{k}$. For the second term, i.e. $\mathcal{I}_{\lambda_\theta}^{h_\theta}(A; \mathbf{x}, \mathbf{x})$, our implementation carried out leave-one-out cross-validation, which means going through every point x in the point pattern $\mathbf{x}$ and leaving out the point exactly one time, to obtain the training set $\mathbf{x} \setminus \{x\}$. Recalling that $\xi_\theta(u; X^T) = p(u)\mathbb{E}[\lambda_\theta(u|X)|X^T]$, we here approximated $\mathbb{E}[\lambda_\theta(u|X)|X^T = \mathbf{x}_i^T(p_k)]$ by $\lambda_\theta(x|\mathbf{x}_i^T(p_k))$, since these two terms tend to each other as $k \to \infty$. To limit the scope of the simulation study, in both the Point Process Learning estimation and the Takacs-Fiksel estimation we employed the Stoyan–Grabarnik test function.

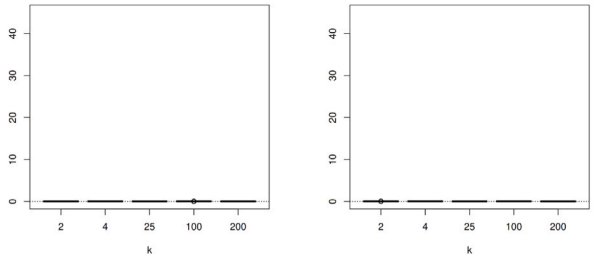
Given a grid of parameter values θ, we calculated the absolute difference in (11) for a two-dimensional conditional intensity model $\lambda_\theta(u|\mathbf{x})$, $u = (u_x, u_y) \in S = [0,1]^2$, $\mathbf{x} \in \mathbf{N}$, given by the Strauss process; recall Section 2.1. More specifically, we considered 100 simulated realisations from the Strauss process, with parameters $R = 0.05$, $\beta = 100$ and $\gamma = 0.5$; see Fig. 1(a) for an example of a realisation from this specific model. On average, this yielded around 74 points per point pattern, and a histogram for the number

Table 1
Values of $q_k = \mathbb{P}(Y_k = 1)$, where $Y_k \sim Bin(k, 1/\sqrt{k})$, for selected k.

k	2	4	15	20	25	30	35	40	50	60	70
q_k	0.414	0.250	0.059	0.037	0.024	0.016	0.011	0.008	0.004	0.002	0.001



(a) $(\beta, R, \gamma) = (50, 0.035, 0.1)$. (b) $(\beta, R, \gamma) = (50, 0.05, 0.1)$. (c) $(\beta, R, \gamma) = (100, 0.05, 0.5)$.



(d) $(\beta, R, \gamma) = (150, 0.05, 0.9)$. (e) $(\beta, R, \gamma) = (150, 0.065, 0.9)$.

Fig. 2. Box plots for the absolute differences between the weighted prediction error sums and the innovations, when k increases, and $p_k = 1/\sqrt{k}$, for five different parameter settings.

of points in the simulated realisations from this Strauss model can be found in Fig. 1(b). The grids for the parameters were set to $50, 55, \dots, 150$ for β, $0.0350, 0.0365, \dots, 0.0650$ for R and $0.10, 0.14, \dots, 0.90$ for γ.

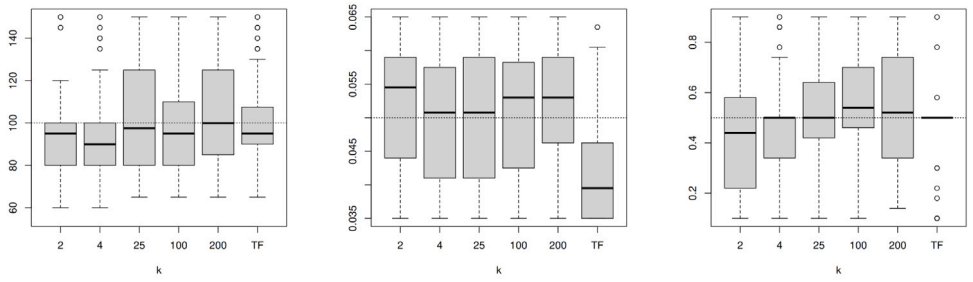
Since we wanted to study how the absolute differences are affected by increasing the value of k, we considered five different values for k, namely 2, 4, 25, 100 and 200, which correspond to the p_k values $1/\sqrt{2} \approx 0.71$, $1/\sqrt{4} = 0.5$, $1/\sqrt{25} = 0.2$, $1/\sqrt{100} = 0.1$ and $1/\sqrt{200} \approx 0.07$.

The probability of having a point $x \in \mathbf{x}$ included in exactly one validation set $\mathbf{x}_i^V$, $i = 1, \dots, k$, is given by $q_k = \mathbb{P}(Y_k = 1)$, where $Y_k \sim Bin(k, 1/\sqrt{k})$; see Table 1 for (approximate) values of q_k for selected values of k. Now, for a specific pattern $\mathbf{x}$, we have that the probability of including each of its points $x \in \mathbf{x}$ in exactly one validation set is given by $q_k^{\#\mathbf{x}}$, where $\#\mathbf{x}$ is the total point count in $\mathbf{x}$. As we can tell from the point count histogram in Fig. 1(b), $q_k^{\#\mathbf{x}}$ is very small for any of the realisations we have generated and, as k increases, the chance that we in practice would encounter the leave-one-out scenario, which represents innovations/Takacs-Fiksel estimation, rapidly tends to 0.

5.1. Absolute differences

Our parameter grid corresponds to a total of 9261 parameter settings, but here we only present the absolute differences for five different parameter settings; see Fig. 2. For each value of k, we illustrate the absolute differences through a boxplot, which captures the variability for the 100 different realisations.

First, we look at parameter values in the lower end of the grid. In Fig. 2(a) we see that the absolute differences decrease as k increases, which indicates what we found in Theorem 1. In Fig. 2(b) we see that the absolute differences are generally higher than in Fig. 2(a), and that there are outliers. The absolute differences decrease for $k = 2, 4, 25$ and then increase slightly for $k = 100$ and 200, which likely indicates that the absolute differences would fluctuate around 0 if we were to increase k further. In Fig. 2(c) we present the absolute differences for the true parameter and here we see that the absolute differences are small for all values of k. Lastly, focusing on parameter values in the higher end of the grid, in both Figs. 2(d) and 2(e) we see that the absolute differences are small for all values of k.



(a) Boxplots for β. (b) Boxplots for R. (c) Boxplots for γ.

Fig. 3. In each graph, six boxplots are seen showing the β, R and γ parameter estimates. First the Point Process Learning estimates are shown for the $\mathcal{L}_3$ loss function with increasing values of k and $p_k = 1/\sqrt{k}$. The last boxplot in each graph shows the parameter estimate for Takacs-Fiksel estimation. The dotted lines show the true parameter value, that is $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

5.2. Parameter estimates

We have validated our theoretical results by means of absolute differences. We next aim to compare Point Process Learning and Takacs-Fiksel estimation in terms of parameter estimation. To this end, we use the same setup as we previously did, considering 100 realisations of the same Strauss process and the same parameter grid. In the case of Point Process Learning, we use Monte-Carlo cross-validation with the same five values for k (and the corresponding p_k values) as before, as well as the same procedure for calculating prediction errors as before. In the case of Takacs-Fiksel estimation, we use leave-one-out cross-validation and the same procedure for calculating innovations. However, instead of calculating the absolute difference between the weighted sum of prediction errors and the innovations, we here use norm-based minimisation over the parameter grid to obtain our parameter estimates. For Point Process Learning, we numerically minimise the $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ loss functions, thus obtaining three different parameter estimates per scenario. For Takacs-Fiksel estimation, we minimise the squared innovations in order to obtain parameter estimates.

In Fig. 3 we provide boxplots for the parameter estimates, to illustrate the variability over the 100 Strauss model realisations. Specifically, we compare the parameter estimates from Point Process Learning, using the $\mathcal{L}_3$ loss function, to the parameter estimates obtained using Takacs-Fiksel estimation. In Tables 2 and 3 we find the corresponding mean and variance values for the parameter estimates. Here we focus on the $\mathcal{L}_3$ since this is the most relevant loss function in connection to our theoretical results.

For the parameter β, it seems that going from the smallest to the largest values of k yields an increase in the estimator variance in combination with a mean which is closer to the true parameter value of 100. Hence, the choice of k seems to influence the trade-off between variance and bias. Fig. 3(a) also illustrates similar trends for the median and the variability of the estimates of β. Hence, as k increases, the Point Process Learning estimates seem to improve in terms of a bias, but at the cost of an increased estimator variance. In the case of Takacs-Fiksel estimation, the mean is very close to the true value, while the median is slightly below the true value, and the variance is on a similar scale as the Point Process Learning estimates.

Turning to the parameter R, Tables 2 and 3 suggest that all approaches and setups are essentially unbiased with small standard errors; note that the bias is slightly smaller for Point Process Learning with larger values of k while Takacs-Fiksel produces a comparatively smaller empirical standard error. At the same time, Fig. 3(b) shows that, in contrast to the Point Process Learning setups, Takacs-Fiksel systematically underestimates the actual value of R, i.e. 0.05; both of the associated quartiles and the median fall below the true value. The fact that we for Takacs-Fiksel estimation have an average estimate of R which is very close to 0.05 is likely an effect of an extreme among the estimates pulling up their overall mean. Note, on the other hand, that systematic under-/overestimation is not encountered in any of the Point Process Learning settings. In contrast to Takacs-Fiksel estimation, here the empirical estimator distributions are closer to being centred around the true parameter value; the medians are closer to the true parameter value, and the upper and lower quartiles are always on opposite sides of the true parameter value.

Turning to γ, in Fig. 3(c) we see that Takacs-Fiksel estimation has a median given precisely by the true parameter value, i.e. 0.5, while the mean value falls slightly below it (see Table 3). At the same time, there are values of k for Point Process Learning such that the produced estimate median is very close to the true parameter value and such that the estimated bias is slightly smaller than for Takacs-Fiksel estimation. On the other hand, Takacs-Fiksel produces a slightly smaller empirical standard error than Point Process Learning.

The results for the loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ can be found in Appendix A.1; see Figs. A.1 and A.2 and Tables A1 and A2. The results for these two are very similar, but there are some differences with respect to $\mathcal{L}_3$. We see that the β estimates produced by Point Process Learning in combination with these loss functions have more negative biases but lower standard error than the ones obtained when using $\mathcal{L}_3$. When it comes to the parameter R, the results are very similar to what they were under $\mathcal{L}_3$. For γ, both the mean and median values increase as k increases, while the variance becomes smaller, so large values of k lead to parameter estimates deviating significantly from the true value of 0.5. Note also the variance–bias trade-off for β and γ: both $\mathcal{L}_1$ and $\mathcal{L}_2$ yield lower standard errors than $\mathcal{L}_3$, at the cost of a higher bias.

We finally note that an extensive simulation study, which more exhaustively compares the estimator performance of Point Process Learning to that of Takacs-Fiksel estimation, can be found in the preprint of this paper, Jansson and Cronie (2024).

Table 2
Mean and variance (over 100 realisations) for the β, R and γ parameter estimates. The Point Process Learning values are shown for the $\mathcal{L}_3$ loss function with increasing values of k and $p_k = 1/\sqrt{k}$. The true parameter values are $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

		2	4	25	100	200
β	Mean	92.8	92.7	101.6	97.0	102.9
	Variance	332.5	316.9	524.7	501.0	478.9
R	Mean	0.0414	0.0521	0.0497	0.0503	0.0507
	Variance	8.47×10^{-5}	9.17×10^{-5}	9.46×10^{-5}	8.35×10^{-5}	7.19×10^{-5}
γ	Mean	0.439	0.443	0.522	0.556	0.535
	Variance	0.0522	0.0319	0.0338	0.0359	0.0432

Table 3
Mean and variance (over 100 realisations) for the β, R and γ parameter estimates for Takacs-Fiksel estimation. The true parameter values are $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

		TF
β	Mean	100.3
	Variance	401.2
R	Mean	0.0520
	Variance	5.39×10^{-5}
γ	Mean	0.446
	Variance	0.0231

6. Discussion

Point Process Learning is a cross-validation-based statistical methodology for point processes, which was recently introduced by Cronie et al. (2024b) for the purpose of estimating parameters in Gibbs point processes models, via their conditional intensity functions. Intuitively, since both Takacs-Fiksel estimation and the prediction errors of Point Process Learning are based on the Georgii–Nguyen–Zessin formula, it seems like Takacs-Fiksel estimation, which has pseudolikelihood estimation as a special case (Coeurjolly and Lavancier, 2019), should be a special (limiting) case of Point Process Learning. In the main results of this paper, we show that this is indeed the case. More specifically, in Section 4 we show that by letting the cross-validation scheme on which Point Process Learning is based tend to leave-one-out cross-validation, it follows that Point Process Learning, when combined with appropriate scaling and a specific loss function, referred to as $\mathcal{L}_3(\theta)$, converges in probability to the loss function used in Takacs-Fiksel estimation. We also discuss how our results can be partially understood as a conditional law of large numbers principle for statistics of conditionally independent thinnings and how our asymptotic regimes relate to different asymptotic regimes found in the literature. In Section 3, we additionally provide a risk minimisation formulation of Point Process Learning, along with bounds $R_1(\theta)$ and $R_2(\theta)$ for the risk $R(\theta) = R_3(\theta)$, and we show that empirical versions of these functions correspond, respectively, to the loss functions $\mathcal{L}_1(\theta)$, $\mathcal{L}_2(\theta)$ and $\mathcal{L}_3(\theta)$ originally proposed for Point Process Learning; in Section 4 we further discuss how $\mathcal{L}_3(\theta)$ is inherently tied to the limiting objects in our asymptotic results, i.e. to the loss function used in Takacs-Fiksel estimation.

In Section 5, we illustrate the derived convergence through a simulation study for a Strauss process. For most parameter values, we observe that the absolute difference appearing in one of our asymptotic results, specifically (11), decreases as the number of cross-validation splits, k, increases. The decrease is, however, not uniform over the parameter space. E.g., although we observe a decrease in the difference (11) as k increases when the parameter vector is given by $(\beta, R, \gamma) = (50, 0.05, 0.1)$, the difference does not vanish completely for any of the choices of k we consider. To study this further and avoid numerical instabilities, the number of points in the simulated point patterns could be increased.

Previous work has compared Point Process Learning to Takacs-Fiksel estimation, by looking specifically at pseudolikelihood estimation. Specifically, through a simulation study, Jansson et al. (2024) showed that Point Process Learning with Monte-Carlo cross-validation and the Stoyan–Grabarnik test function outperforms pseudolikelihood estimation in terms of MSE in the case of a Gibbs hard-core process; see Section 2.1 for the definition of Gibbs hard-core family.

Pseudolikelihood estimation is a special case of Takacs-Fiksel estimation, but it does not necessarily represent the optimal case (Coeurjolly et al., 2016). As appealing as pseudolikelihood estimation may be, it has its issues, most notably poor performance when there are strong interactions present (Diggle et al., 1994). In addition, it suffers from identifiability issues, even in the context of rather basic models. It was shown in Jansson et al. (2024) that a range of parameter choices for the hard-core distance in a Gibbs hard-core process give the same value of the log-pseudolikelihood function. This stems from the fact that the conditional intensity of a Gibbs hard-core process is not differentiable with respect to the hard-core distance. In practice, estimation can be carried out with the function `ppm` in the `R` package `spatstat` (Baddeley et al., 2015), which uses pseudolikelihood estimation for the intensity-related parameter and a plug-in approach for the hard-core distance parameter.

Given the results in Jansson et al. (2024), one may hope to obtain an understanding of whether general Point Process Learning renders lower estimator variances than Takacs-Fiksel estimation, by looking closer at (limits of) the prediction error variance expression in Cronie et al. (2024b). Using the limits in Theorems 1 and 2, in combination with Cronie et al. (2024b, Theorem

2), one may hope to shed some light on how the bias, the variance and the MSE of the general Point Process Learning setup relates to its limiting case Takacs-Fiksel estimation. Considering this approach, we have, however, not succeeded to theoretically show that either of the two methods results in a lower bias, variance or MSE for the estimators than the other method. In future work, we intend to compare Point Process Learning and Takacs-Fiksel estimation further, with simulation studies and data analysis.

Funding

All simulations were run on the Vera cluster, provided by Chalmers e-Commons at Chalmers University of Technology.[1] Ottmar Cronie has been supported by the Swedish Research Council (2023-03320).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are grateful to Aila Särkkä for valuable discussions and helpful insights. They are also very grateful to an anonymous referee for many constructive comments and suggestions.

Appendix

A.1. Additional simulation results

In Fig. A.1, Table A1, Fig. A.2 and Table A2 we provide the results for the parameter estimates for the $\mathcal{L}_1$ and $\mathcal{L}_2$ loss functions.

A.2. Proofs

Below we provide the proofs of the results in the main text.

A.2.1. Proof of Theorem 1

To prove Theorem 1 we need a dominated convergence theorem for sequences of random variables converging in probability. This result can be found stated in different places (e.g. Exercise 2(ix) on https://terrytao.wordpress.com/2015/10/23/275a-notes-3-the-weak-and-strong-law-of-large-numbers/) and its proof exploits that such sequences have subsequences which converge a.s. to the same limit.

Lemma A.1 (*Dominated Convergence in Probability*). *Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying a non-negative random variable Z, with $\mathbb{E}[Z] < \infty$, and a sequence of random variables which satisfy $Y_n \xrightarrow{p} Y$, as $n \to \infty$, and $|Y_n| \leq Z$ a.s. for all n. We then have $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$ and $\mathbb{E}[|Y_n - Y|] \to 0$ as $n \to \infty$.*

We are now ready to prove Theorem 1.

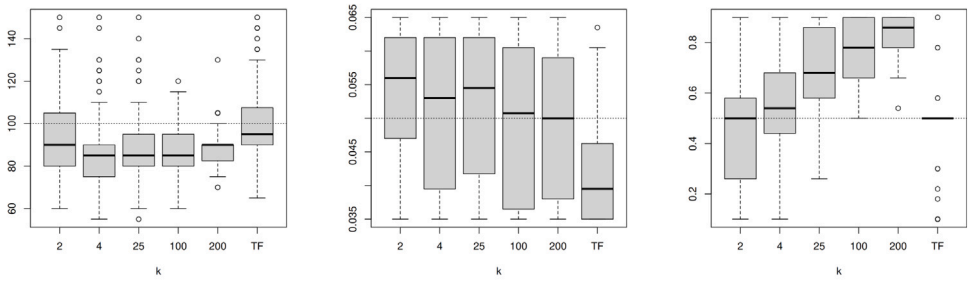
Proof of Theorem 1. Let $A \subseteq S$ be arbitrary and bounded. Consider a sequence a_k, $k \geq 2$, and a sequence $p_k \in (0,1)$, $k \geq 1$, which decreases to 0 as $k \to \infty$. We want to make choices for these sequences such that

$$a_k \sum_{i=1}^{k} \mathcal{I}_{\xi_\theta}^{h_\theta}(A; X_i^V(p_k), X_i^T(p_k)) \xrightarrow{p} \mathcal{I}_{\lambda_\theta}^{h_\theta}(A; X, X)$$

as $k \to \infty$. As in the statement of the theorem, for all $k \geq 2$, we let $\{(X_i^T(p_k), X_i^V(p_k))\}_{i=1}^k$ and $\{(\mathbf{x}_i^T(p_k), \mathbf{x}_i^V(p_k))\}_{i=1}^k$ be Monte-Carlo cross-validation rounds for X and $\mathbf{x}$, respectively, which have been generated by the retention probability p_k. Now, consider

$$\begin{aligned}
\Delta(k; \mathbf{x}) =& a_k \sum_{i=1}^{k} \mathcal{I}_{\xi_\theta}^{h_\theta}(A; \mathbf{x}_i^V(p_k), \mathbf{x}_i^T(p_k)) - \mathcal{I}_{\lambda_\theta}^{h_\theta}(A; \mathbf{x}, \mathbf{x}) \\
=& a_k \sum_{i=1}^{k} \left(\sum_{x \in \mathbf{x}_i^V(p_k) \cap A} h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - \int_A h_\theta(u; \mathbf{x}_i^T(p_k)) \xi_\theta(u; \mathbf{x}_i^T(p_k)) \mathrm{d}u \right) \\
& - \left(\sum_{x \in \mathbf{x} \cap A} h_\theta(x; \mathbf{x} \setminus \{x\}) - \int_A h_\theta(u; \mathbf{x}) \lambda_\theta(u|\mathbf{x}) \mathrm{d}u \right)
\end{aligned}$$

[1] https://www.c3se.chalmers.se/about/Vera/

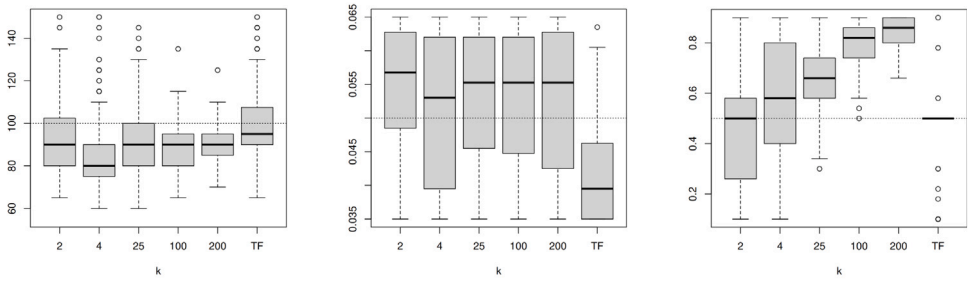


(a) Boxplots for β. (b) Boxplots for R. (c) Boxplots for γ.

Fig. A.1. In each graph, six boxplots are seen showing the β, R and γ parameter estimates. First the Point Process Learning estimates are shown for the $\mathcal{L}_1$ loss function with increasing values of k and $p_k = 1/\sqrt{k}$. The last boxplot in each graph shows the parameter estimate for Takacs-Fiksel estimation. The dotted lines show the true parameter value, that is $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

Table A1
Mean and variance (over 100 realisations) for the β, R and γ parameter estimates. The Point Process Learning values are shown for the $\mathcal{L}_1$ loss function with increasing values of k and $p_k = 1/\sqrt{k}$. The true parameter values are $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

		2	4	25	100	200
β	Mean	92.4	87.8	87.9	86.5	87.8
	Variance	336.1	354.2	306.2	146.1	103.2
R	Mean	0.0535	0.0512	0.0519	0.0498	0.0492
	Variance	9.97×10^{-5}	1.20×10^{-4}	1.16×10^{-4}	1.29×10^{-4}	1.18×10^{-4}
γ	Mean	0.455	0.557	0.695	0.769	0.835
	Variance	0.0533	0.0428	0.0254	0.0160	0.00597



(a) Boxplots for β. (b) Boxplots for R. (c) Boxplots for γ.

Fig. A.2. In each graph, six boxplots are seen showing the β, R and γ parameter estimates. First the Point Process Learning estimates are shown for the $\mathcal{L}_2$ loss function with increasing values of k and $p_k = 1/\sqrt{k}$. The last boxplot in each graph shows the parameter estimate for Takacs-Fiksel estimation. The dotted lines show the true parameter value, that is $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

$$
\begin{aligned}
&= \sum_{x \in \mathbf{x} \cap A} a_k \sum_{i=1}^{k} \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\} h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\}) \\
&\quad - \left(a_k \sum_{i=1}^{k} \int_A h_\theta(u; \mathbf{x}_i^T(p_k)) \xi_\theta(u; \mathbf{x}_i^T(p_k)) \mathrm{d}u - \int_A h_\theta(u; \mathbf{x}) \lambda_\theta(u|\mathbf{x}) \mathrm{d}u \right) \\
&= \Delta_1(k; \mathbf{x}) - \Delta_2(k; \mathbf{x}).
\end{aligned}
$$

Convergence of $\Delta_1(k; X)$

Starting with $\Delta_1(k; \mathbf{x})$, given any $x \in \mathbf{x}$, we note that

$$
\min_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\}) \sum_{i=1}^{k} a_k \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}
$$

Table A2
Mean and variance (over 100 realisations) for the β, R and γ parameter estimates. The Point Process Learning values are shown for the $\mathcal{L}_2$ loss function with increasing values of k and $p_k = 1/\sqrt{k}$. The true parameter values are $\beta_0 = 100$, $R_0 = 0.05$ and $\gamma_0 = 0.5$, respectively.

		2	4	25	100	200
β	Mean	91.8	87.8	92.2	89.2	90.0
	Variance	328.0	395.6	289.5	135.2	93.7
R	Mean	0.0543	0.0506	0.0530	0.0529	0.0525
	Variance	9.51×10^{-5}	1.31×10^{-4}	9.61×10^{-5}	1.03×10^{-4}	1.15×10^{-4}
γ	Mean	0.457	0.569	0.664	0.786	0.837
	Variance	0.0496	0.0490	0.0200	0.00925	0.00373

$$\leq \sum_{i=1}^{k} a_k \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\} h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\})$$
$$\leq \max_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\}) \sum_{i=1}^{k} a_k \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}.$$

Next, we will show that $\min_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\})$ and $\max_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\})$ tend to $h_\theta(x; \mathbf{x} \setminus \{x\})$ in probability and that $\sum_{i=1}^{k} a_k \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\} \to 1$ in probability, as $k \to \infty$, provided that we make the choice $a_k = p_k = 1/\sqrt{k}$. Having shown this, by appealing to Slutsky's lemma (Ferguson, 1996, Theorem 6'), as $k \to \infty$, we obtain that both the upper and the lower bound tend to $h_\theta(x; \mathbf{x} \setminus \{x\})$ in probability, which in turn implies that $\sum_{i=1}^{k} a_k \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\} h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) \to h_\theta(x; \mathbf{x} \setminus \{x\})$ in probability, whereby $\Delta_1(k; \mathbf{x}) \to 0$ in probability when $k \to \infty$. Hence, as this holds for any realisation $\mathbf{x}$ of X, we have that $\Delta_1(k; X) \to 0$ in probability as $k \to \infty$.

For a fixed $x \in \mathbf{x}$, let $S_k = a_k \sum_{i=1}^{k} \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}$ and note that $\mathbb{E}[S_k] = a_k k \mathbb{E}[\mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}] = a_k k p_k$. We want to have that $\lim_{k\to\infty} a_k k p_k = 1$ and we want to show that

$$S_k - \mathbb{E}[S_k] = a_k \sum_{i=1}^{k} \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\} - a_k k p_k = a_k k p_k \left(\frac{1}{k} \sum_{i=1}^{k} \frac{\mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}}{p_k} - 1 \right)$$

tends to 0 in probability, as $k \to \infty$. Note that all $\mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}$ are independent Bernoulli random variables with mean $\mathbb{E}[\mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}] = p_k$ and variance $\mathrm{Var}(\mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}) = p_k(1 - p_k)$. By Markov's inequality,

$$\begin{aligned}
\varepsilon \mathbb{P}(|S_k - \mathbb{E}[S_k]| > \varepsilon) &\leq \mathbb{E}[(S_k - \mathbb{E}[S_k])^2] = \mathrm{Var}(S_k) \\
&= \mathrm{Var}\left(a_k \sum_{i=1}^{k} \mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\} \right) \\
&= a_k^2 k \,\mathrm{Var}\left(\mathbf{1}\{x \in \mathbf{x}_i^V(p_k)\}\right) = a_k^2 k p_k (1 - p_k)
\end{aligned}$$

for any $\varepsilon > 0$, which we want to show tends to 0. We thus want to have both $\lim_{k\to\infty} a_k k p_k = 1$ and $\lim_{k\to\infty} a_k^2 k p_k (1 - p_k) = (\lim_{k\to\infty} a_k k p_k)(\lim_{k\to\infty} a_k (1 - p_k)) = 0$, which may be achieved by letting

$$a_k = p_k = 1/\sqrt{k},$$

yielding that the upper bound above is given by $k^{-1/2}(1 - k^{-1/2})$. Hence, since $\varepsilon > 0$ was arbitrary, with $a_k = p_k = 1/\sqrt{k}$ we obtain that $\lim_{k\to\infty} S_k = \lim_{k\to\infty} \mathbb{E}[S_k] = 1$ in probability.

We next turn to the convergence of the minima $\min_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\})$ and $\max_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\})$. For any $k \geq 2$ and any $i \in \{1, \dots, k\}$, let $m_{i,k}(y) \in \{0, 1\}$, $y \in \mathbf{x}$, be the corresponding marking, which yields a sequence of iid random variables with $\mathbb{P}(m_{i,k}(y) = 0) = 1 - p_k$ and $\mathbb{P}(m_{i,k}(y) = 1) = p_k$. We have that $\mathbf{x}_i^T(p_k) = \{y \in \mathbf{x} : m_{i,k}(y) = 0\}$. Now, for any $\varepsilon > 0$, by the law of total probability,

$$\begin{aligned}
&\mathbb{P}(|h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\})| > \varepsilon) \\
=&\mathbb{P}(|h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\})| > \varepsilon | \mathbf{x}_i^T(p_k) = \mathbf{x}) \mathbb{P}(\mathbf{x}_i^T(p_k) = \mathbf{x}) \\
&+ \mathbb{P}(|h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\})| > \varepsilon | \mathbf{x}_i^T(p_k) \neq \mathbf{x})(1 - \mathbb{P}(\mathbf{x}_i^T(p_k) = \mathbf{x})) \\
=&\mathbb{P}(|h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\})| > \varepsilon | \mathbf{x}_i^T(p_k) \neq \mathbf{x})(1 - \mathbb{P}(\mathbf{x}_i^T(p_k) = \mathbf{x})) \\
=&\mathbb{P}(|h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\})| > \varepsilon | \mathbf{x}_i^T(p_k) \neq \mathbf{x}) \\
&\times \left(1 - \binom{\#\mathbf{x}}{\#\mathbf{x}} p_k^0 (1 - p_k)^{\#\mathbf{x}}\right) \\
=&\mathbb{P}(|h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\}) - h_\theta(x; \mathbf{x} \setminus \{x\})| > \varepsilon | \mathbf{x}_i^T(p_k) \neq \mathbf{x})(1 - (1 - p_k)^{\#\mathbf{x}}).
\end{aligned}$$

Since $p_k \downarrow 0$ as $k \to \infty$, we obtain $1 - (1 - p_k)^{\#\mathbf{x}} \to 0$, whereby the expression above tends to 0. Since $\varepsilon > 0$ was arbitrary, $h_\theta(x; \mathbf{x}_i^T(p_k) \setminus \{x\})$ tends to $h_\theta(x; \mathbf{x} \setminus \{x\})$ in probability, as $k \to \infty$. Since we have shown this for any $i \in \{1, \dots, k\}$, it also holds for $\min_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\})$ and $\max_{j=1,\dots,k} h_\theta(x; \mathbf{x}_j^T(p_k) \setminus \{x\})$.

One may hope to strengthen the convergence above to hold a.s., which could be achieved by applying a combination of the Borel–Cantelli lemma and Ferguson (1996, Lemma 1), provided that the right-hand side of $\sum_{k\geq 2}\mathbb{P}(|h_\theta(x;\mathbf{x}_i^T(p_k)\setminus\{x\})-h_\theta(x;\mathbf{x}\setminus\{x\})|>\varepsilon)\leq\sum_{k\geq 2}(1-(1-p_k)^{\#\mathbf{x}})$ is finite. However, $1+\sum_{k\geq 2}(1-(1-k^{-b}))=\zeta(b)$, the Riemann zeta function evaluated in b, is finite only if $b>1$. In other words, we would have had to have $p_k=k^{-b}$ for some $b>1$, as opposed to $p_k=k^{-1/2}$.

Convergence of $\Delta_2(k;X)$

We next want to show that $\lim_{k\to\infty}\Delta_2(k;X)\overset{p}{=}0$. Having fixed $a_k=p_k=1/\sqrt{k}$,

$$\begin{aligned}
&\Delta_2(k;X)\\
=&a_k\sum_{i=1}^{k}\int_A h_\theta(u;X_i^T(p_k))\xi_\theta(u;X_i^T(p_k))\mathrm{d}u-\int_A h_\theta(u;X)\lambda_\theta(u|X)\mathrm{d}u\\
=&a_k\sum_{i=1}^{k}\int_A h_\theta(u;X_i^T(p_k))p_k\,\mathbb{E}\left[\lambda_\theta(u|X)\middle|X_i^T(p_k)\right]\mathrm{d}u-\int_A h_\theta(u;X)\lambda_\theta(u|X)\mathrm{d}u\\
=&\frac{1}{k}\sum_{i=1}^{k}\int_A\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)\middle|X_i^T(k^{-1/2})\right]\mathrm{d}u-\int_A h_\theta(u;X)\lambda_\theta(u|X)\mathrm{d}u\\
=&\mathbb{E}\left[\int_A\frac{1}{k}\sum_{i=1}^{k}h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)\mathrm{d}u\middle|X_i^T(k^{-1/2})\right]-\int_A h_\theta(u;X)\lambda_\theta(u|X)\mathrm{d}u,
\end{aligned}$$

where the last equality follows from the Fubini–Tonelli theorem for conditional expectations, which requires that the Fubini–Tonelli theorem holds for the unconditional version of the statement. If the variance of $\Delta_2(k;X)$ tends to 0 then we obtain the required result as a consequence of Markov's inequality.

We first show that $\lim_{k\to\infty}\mathbb{E}[\Delta_2(k;X)]=0$. Note that

$$\begin{aligned}
\mathbb{E}[\Delta_2(k;X)]=&\int_A\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\mathbb{E}[h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)|X_i^T(k^{-1/2})]]\mathrm{d}u\\
&-\int_A\mathbb{E}[h_\theta(u;X)\lambda_\theta(u|X)]\mathrm{d}u\\
=&\int_A\mathbb{E}[h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)]-\mathbb{E}[h_\theta(u;X)\lambda_\theta(u|X)]\mathrm{d}u\\
=&\mathbb{E}\left[\sum_{x\in X\cap A}h_\theta(x;X_i^T(k^{-1/2})\setminus\{x\})-h_\theta(x;X\setminus\{x\})\right]
\end{aligned}$$

by the Fubini–Tonelli theorem, the law of total expectation and the GNZ formula. We already know that for any $\mathbf{x}$ and $x\in\mathbf{x}$, the deviation $h_\theta(x;\mathbf{x}_i^T(p_k)\setminus\{x\})-h_\theta(x;\mathbf{x}\setminus\{x\})$ tends to 0 in probability as $k\to\infty$. Therefore, we obtain $\lim_{k\to\infty}\sum_{x\in X\cap A}h_\theta(x;X_i^T(k^{-1/2})\setminus\{x\})-h_\theta(x;X\setminus\{x\})\overset{p}{=}0$; note that, by definition, $\#(X\cap A)<\infty$ a.s. for all bounded $A\subseteq S$. Next, we want to apply Lemma A.1 here to obtain that $\lim_{k\to\infty}\mathbb{E}[\Delta_2(k;X)]=\mathbb{E}[\lim_{k\to\infty}\Delta_2(k;X)]=0$. To do so, we need to ensure that $|\sum_{x\in X\cap A}h_\theta(x;X_i^T(k^{-1/2})\setminus\{x\})-h_\theta(x;X\setminus\{x\})|$ is bounded by an integrable random variable for each $k\geq 2$. Since h_θ is bounded, $|h_\theta(x;\mathbf{x})-h_\theta(x;\mathbf{y})|\leq 2\max\{|h_\theta(x;\mathbf{x})|,|h_\theta(x;\mathbf{y})|\}<\infty$ for all $\mathbf{x},\mathbf{y}$. We thus obtain the bounding random variable $2\sum_{x\in X\cap A}\max\{|h_\theta(x;X_i^T(k^{-1/2})\setminus\{x\})|,|h_\theta(x;X\setminus\{x\})|\}$, which has finite expectation since $\#(X\cap A)<\infty$ a.s. by the local finiteness of X. Hence, $\lim_{k\to\infty}\mathrm{Var}(\Delta_2(k;X))=\lim_{k\to\infty}\mathbb{E}[\Delta_2(k;X)^2]$.

Turning to the second moment, we have

$$\begin{aligned}
&\mathbb{E}[\Delta_2(k;X)^2]\\
=&\mathbb{E}\left[\mathbb{E}\left[\int_A\frac{1}{k}\sum_{i=1}^{k}h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)\mathrm{d}u\middle|X_i^T(k^{-1/2})\right]^2\right]\\
&-2\mathbb{E}\left[\int_{A^2}\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)\middle|X_i^T(k^{-1/2})\right]h_\theta(v;X)\lambda_\theta(v|X)\mathrm{d}u\mathrm{d}v\right]\\
&+\mathbb{E}\left[\int_{A^2}h_\theta(u;X)\lambda_\theta(u|X)h_\theta(v;X)\lambda_\theta(v|X)\mathrm{d}u\mathrm{d}v\right]\\
\leq&\mathbb{E}\left[\left(\int_A\frac{1}{k}\sum_{i=1}^{k}h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)\mathrm{d}u\right)^2\right]\\
&-2\int_{A^2}\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))\lambda_\theta(u|X)\middle|X_i^T(k^{-1/2})\right]h_\theta(v;X)\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v\\
&+\int_{A^2}\mathbb{E}\left[h_\theta(u;X)h_\theta(v;X)\lambda_\theta(u|X)\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v\\
=&E_1(k)-2E_2(k)+E_3,
\end{aligned}$$

where the inequality is a consequence of conditioning being a contractive projection of L^p spaces.

Now, we first want to show that $E_1(k) \to E_3$ when $k \to \infty$. We have

$$\begin{aligned}
E_1(k) =& \mathbb{E}\left[\left(\int_A \frac{1}{k}\sum_{i=1}^k h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)\mathrm{d}u\right)^2\right] \\
=& \mathbb{E}\left[\int_{A^2} \frac{1}{k}\sum_{i=1}^k h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)\frac{1}{k}\sum_{j=1}^k h_\theta(v; X_j^T(k^{-1/2}))\lambda_\theta(v|X)\mathrm{d}u\mathrm{d}v\right] \\
=& \int_{A^2} \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^k h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)\frac{1}{k}\sum_{j=1}^k h_\theta(v; X_j^T(k^{-1/2}))\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v \\
=& \int_{A^2} \frac{1}{k^2}\sum_{i=1}^k\sum_{j=1}^k \mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)h_\theta(v; X_j^T(k^{-1/2}))\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v \\
=& \int_{A^2} \mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)h_\theta(v; X_j^T(k^{-1/2}))\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v,
\end{aligned}$$

for any $i, j \in \{1, \dots, k\}$. Now, we already know that $h_\theta(u; \mathbf{x}_i^T(k^{-1/2})) - h_\theta(u; \mathbf{x})$ and $h_\theta(v; \mathbf{x}_j^T(k^{-1/2})) - h_\theta(v; \mathbf{x})$ both tend to 0 in probability as $k \to \infty$. Since this holds for all $\mathbf{x}$, it also holds for X in probability. Further, by Lemma A.1 and Slutsky's lemma (Ferguson, 1996, Theorem 6') we get

$$\begin{aligned}
&\lim_{k\to\infty} E_1(k) \\
=& \int_{A^2} \mathbb{E}\left[\lim_{k\to\infty}(h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)h_\theta(v; X_j^T(k^{-1/2}))\lambda_\theta(v|X))\right]\mathrm{d}u\mathrm{d}v \\
=& \int_{A^2} \mathbb{E}\left[\left(\lim_{k\to\infty} h_\theta(u; X_i^T(k^{-1/2}))\lambda_\theta(u|X)\right)\left(\lim_{k\to\infty} h_\theta(v; X_j^T(k^{-1/2}))\lambda_\theta(v|X)\right)\right]\mathrm{d}u\mathrm{d}v \\
=& \int_{A^2} \mathbb{E}\left[h_\theta(u; X)\lambda_\theta(u|X)h_\theta(v; X)\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v = E_3.
\end{aligned}$$

Note that Lemma A.1 is applicable here since λ_θ and h_θ are bounded by assumption.

Focusing on $E_2(k)$, by the self-adjointness of conditional expectations we have

$$\begin{aligned}
&E_2(k) \\
=& \int_{A^2} \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^k \mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\middle| X_i^T(k^{-1/2})\right] h_\theta(v; X)\lambda_\theta(u|X)\lambda_\theta(v|X)\right]\mathrm{d}u\mathrm{d}v,
\end{aligned}$$

which we want to show tends to E_3 in probability.

We start by showing that $\mathbb{E}[h_\theta(u; X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})]$ tends to $h_\theta(u; X)$ in probability. We do so using Markov's inequality:

$$\begin{aligned}
&\epsilon^2\mathbb{P}\left(\left|\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\middle| X_i^T(k^{-1/2})\right] - h_\theta(u; X)\right| > \epsilon\right) \\
\leq& \mathbb{E}\left[\left(\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\middle| X_i^T(k^{-1/2})\right] - h_\theta(u; X)\right)^2\right] \\
=& \mathbb{E}\Big[\Big(\mathbb{E}\Big[h_\theta(u; X_i^T(k^{-1/2}))\Big| X_i^T(k^{-1/2})\Big] - h_\theta(u; X_i^T(k^{-1/2})) \\
&+ h_\theta(u; X_i^T(k^{-1/2})) - h_\theta(u; X)\Big)^2\Big] \\
\leq& \mathbb{E}\left[\left(\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\middle| X_i^T(k^{-1/2})\right] - h_\theta(u; X_i^T(k^{-1/2}))\right)^2\right] \\
&+ \mathbb{E}\left[\left(h_\theta(u; X_i^T(k^{-1/2})) - h_\theta(u; X)\right)^2\right].
\end{aligned}$$

We already know from before that $\lim_{k\to\infty} h_\theta(x; X_i^T(k^{-1/2})) - h_\theta(x; X) \stackrel{p}{=} 0$. Thus, by applying Lemma A.1 the second term goes to 0. We continue with the first term:

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))\middle| X_i^T(k^{-1/2})\right] - h_\theta(u; X_i^T(k^{-1/2}))\right)^2\right] \\
=& \mathbb{E}\Big[\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right]^2 + h_\theta(u; X_i^T(k^{-1/2}))^2 \\
&- 2\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right] h_\theta(u; X_i^T(k^{-1/2}))\Big] \\
=& \mathbb{E}\Big[\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right]^2\Big] + \mathbb{E}\Big[h_\theta(u; X_i^T(k^{-1/2}))^2\Big] \\
&- 2\mathbb{E}\Big[\mathbb{E}\left[h_\theta(u; X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right] h_\theta(u; X_i^T(k^{-1/2}))\Big].
\end{aligned}$$

Now we know that $\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right]^2\right] \leq \mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2\right]$ due to conditioning being a contractive projection of L^p spaces. Further, by the 'taking out what is known' property and the law of total expectation,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right]h_\theta(u;X_i^T(k^{-1/2}))\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2|X_i^T(k^{-1/2})\right]\right]\\ =&\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2\right].\end{aligned}$$

By putting all of this together we obtain

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right]^2\right]+\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2\right]\\ &-2\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})\right]h_\theta(u;X_i^T(k^{-1/2}))\right]\\ &\leq \mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2\right]+\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2\right]-2\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))^2\right]=0.\end{aligned}$$

To summarise, we then have that

$$\begin{aligned}&\varepsilon^2\mathbb{P}\left(\left|\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))\middle|X_i^T(k^{-1/2})\right]-h_\theta(u;X)\right|>\varepsilon\right)\\ \leq&\mathbb{E}\left[\left(\mathbb{E}\left[h_\theta(u;X_i^T(k^{-1/2}))\middle|X_i^T(k^{-1/2})\right]-h_\theta(u;X_i^T(k^{-1/2}))\right)^2\right]\\ &+\mathbb{E}\left[\left(h_\theta(u;X_i^T(k^{-1/2}))-h_\theta(u;X)\right)^2\right]\\ &\leq 0+\mathbb{E}\left[\left(h_\theta(u;X_i^T(k^{-1/2}))-h_\theta(u;X)\right)^2\right]\longrightarrow 0\end{aligned}$$

when $k\to\infty$. Since ε is arbitrary, this means that $\mathbb{E}[h_\theta(u;X_i^T(k^{-1/2}))|X_i^T(k^{-1/2})]$ tends to $h_\theta(u;X)$ in probability. Then, we can see that $\lim_{k\to\infty}E_2(k)\overset{p}{=}E_3$ by using Lemma A.1, which gives us that

$$\lim_{k\to\infty}\mathbb{E}[\Delta_2(k;X)^2]\leq\lim_{k\to\infty}E_1(k)-2\lim_{k\to\infty}E_2(k)+E_3\overset{p}{=}E_3-2E_3+E_3=0$$

which in turn gives us that $\lim_{k\to\infty}\mathrm{Var}(\Delta_2(k;X))=\lim_{k\to\infty}\mathbb{E}[\Delta_2(k;X)^2]\overset{p}{=}0$. Then we obtain $\lim_{k\to\infty}\Delta_2(k;X)\overset{p}{=}0$ due to Markov's inequality. □

A.2.2. *Proof of Theorem 2*

Proof of Theorem 2. We want to show that

$$\begin{aligned}\sum_{i=1}^k\mathcal{I}_{\xi_\theta}^{h_\theta}(A;X_{ik}^V,X_{ik}^T)&=\sum_{i=1}^k\sum_{x\in X_{ik}^V\cap A}h_\theta(x;X_{ik}^T)-\sum_{i=1}^k\int_A h_\theta(u;X_{ik}^T)\xi_\theta(u;X_{ik}^T)\mathrm{d}u\\ &=\sum_{i=1}^k\sum_{x\in X\cap A}\mathbf{1}\{x\in A_{ik}\}h_\theta(x;X\cap(A\setminus A_{ik}))\\ &\quad-\sum_{i=1}^k\int_A h_\theta(u;X\cap(A\setminus A_{ik}))\xi_\theta(u;X\cap(A\setminus A_{ik}))\mathrm{d}u\end{aligned}$$

converges in probability to

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

In order to do so, we show that both

$$\Delta_1(k;X)=\sum_{i=1}^k\sum_{x\in X\cap A}\mathbf{1}\{x\in A_{ik}\}h_\theta(x;X\cap(A\setminus A_{ik}))-\sum_{x\in X\cap A}h_\theta(x;X\cap A\setminus\{x\})$$

and

$$\Delta_2(k;X)=\int_A h_\theta(u;X\cap(A\setminus A_{ik}))\xi_\theta(u;X\cap(A\setminus A_{ik}))\mathrm{d}u-\int_A h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)\mathrm{d}u$$

tend to 0 in probability, as $k\to\infty$.

Since S is a metric space (and thereby a Hausdorff space), for any distinct points $x,y\in S$ we can find radii $r_x,r_y>0$ such that $b(x,r_x)\cap b(y,r_y)=\emptyset$. This holds in particular for distinct members $x,y\in\mathbf{x}$ of a point pattern $\mathbf{x}\in\mathbf{N}$. By the local finiteness of $\mathbf{N}$, for each $\mathbf{x}\in\mathbf{N}$ there is a universal $r_{\mathbf{x},A}>0$ such that $b(x,r_{\mathbf{x},A})\cap b(y,r_{\mathbf{x},A})=\emptyset$ for any $x,y\in\mathbf{x}\cap A$. Hence, as $\max_{i=1,\ldots,k}|A_{ik}|$ decreases, we can find some $k_{\mathbf{x},A}\geq 2$ such that when $k\geq k_{\mathbf{x},A}$ we have that $\#(A_{ik}\cap\mathbf{x})\in\{0,1\}$, i.e. each A_{ik} contains at most one element of $\mathbf{x}\cap A$.

Now, for any $u \in A$, let $A_k(u)$ be the unique A_{ik} that contains u. Then

$$\sum_{i=1}^{k} \mathbf{1}\{u \in A_{ik}\} h_\theta(u; \mathbf{x} \cap (A \setminus A_{ik})) = h_\theta(u; \mathbf{x} \cap (A \setminus A_k(u))). \tag{A.1}$$

For any $\mathbf{x} \in \mathbf{N}$, if $u \in \mathbf{x}$ then $\mathbf{x} \cap (A \setminus A_{ik}(u)) \to \mathbf{x} \setminus \{u\}$ as $k \to \infty$ and if $u \notin \mathbf{x}$ then $\mathbf{x} \cap (A \setminus A_{ik}(u)) \to \mathbf{x}$ as $k \to \infty$. We thus have

$$\begin{aligned} &\lim_{k\to\infty} h_\theta(u; \mathbf{x} \cap (A \setminus A_k(u))) \\ =&\mathbf{1}\{u \in \mathbf{x} \cap A\} h_\theta(u; \mathbf{x} \cap A \setminus \{u\}) + \mathbf{1}\{u \notin \mathbf{x} \cap A\} h_\theta(u; \mathbf{x} \cap A), \end{aligned} \tag{A.2}$$

and since this holds for any $\mathbf{x} \in \mathbf{N}$, we also have

$$\begin{aligned} &\lim_{k\to\infty} h_\theta(u; \mathbf{x} \cap (A \setminus A_k(u))) \\ \overset{a.s.}{=}&\mathbf{1}\{u \in X \cap A\} h_\theta(u; X \cap A \setminus \{u\}) + \mathbf{1}\{u \notin X \cap A\} h_\theta(u; X \cap A). \end{aligned} \tag{A.3}$$

Convergence of $\Delta_1(k; X)$

By (A.3), we obtain that

$$\begin{aligned} &\lim_{k\to\infty} \sum_{i=1}^{k} \sum_{x \in X \cap A} \mathbf{1}\{x \in A_{ik}\} h_\theta(x; X \cap (A \setminus A_{ik})) \\ =& \sum_{x \in X \cap A} \lim_{k\to\infty} h_\theta(x; \mathbf{x} \cap (A \setminus A_k(u))) \\ \overset{a.s.}{=}& \sum_{x \in X \cap A} (\mathbf{1}\{x \in X \cap A\} h_\theta(x; X \cap A \setminus \{x\}) + \mathbf{1}\{x \notin X \cap A\} h_\theta(x; X \cap A)) \\ =& \sum_{x \in X \cap A} h_\theta(x; X \cap A \setminus \{x\}), \end{aligned}$$

whereby $\Delta_1(k; X)$ tends to 0 a.s., and thereby in probability, as $k \to \infty$.

Convergence of $\Delta_2(k; X)$

We next want to show that $\lim_{k\to\infty} \Delta_2(k; X) \overset{p}{=} 0$. This means that for all $\varepsilon > 0$,

$$\lim_{k\to\infty} \mathbb{P}(|\Delta_2(k; X) - 0| > \varepsilon) = 0.$$

By using Markov's inequality, we get

$$\mathbb{P}(|\Delta_2(k; X)| \geq \varepsilon) = \mathbb{P}(\Delta_2(k; X)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[\Delta_2(k; X)^2]}{\varepsilon^2}.$$

If we can show that $\lim_{k\to\infty} \mathbb{E}[\Delta_2(k; X)^2] = 0$ then we have that

$$\lim_{k\to\infty} \varepsilon^2 \mathbb{P}(|\Delta_2(k; X)| \geq \varepsilon) \leq \lim_{k\to\infty} \mathbb{E}[\Delta_2(k; X)^2] = 0.$$

Since this then holds for any $\varepsilon > 0$, we obtain that $\lim_{k\to\infty} \Delta_2(k; X) \overset{p}{=} 0$.

We first rewrite the first integral term in $\Delta_2(k; X)$:

$$\begin{aligned} &\sum_{i=1}^{k} \int_A h_\theta(u; X \cap (A \setminus A_{ik})) \xi_\theta(u; X \cap (A \setminus A_{ik})) \mathrm{d}u \\ =&\sum_{i=1}^{k} \int_A h_\theta(u; X \cap (A \setminus A_{ik})) p_{ik}(u) \, \mathbb{E}\left[\lambda_\theta(u|X \cap A) \middle| X \cap (A \setminus A_{ik})\right] \mathrm{d}u \\ =&\int_A \sum_{i=1}^{k} h_\theta(u; X \cap (A \setminus A_{ik})) \mathbf{1}\{u \in A_{ik}\} \, \mathbb{E}\left[\lambda_\theta(u|X \cap A) \middle| X \cap (A \setminus A_{ik})\right] \mathrm{d}u \\ =&\sum_{i=1}^{k} \int_{A_{ik}} \mathbb{E}\left[h_\theta(u; X \cap (A \setminus A_{ik})) \lambda_\theta(u|X \cap A) \middle| X \cap (A \setminus A_{ik})\right] \mathrm{d}u. \end{aligned}$$

Using this, we obtain that the second moment satisfies

$$\begin{aligned} &\mathbb{E}[\Delta_2(k; X)^2] \\ =&\mathbb{E}\left[\left(\mathbb{E}\left[\sum_{i=1}^{k} \int_{A_{ik}} h_\theta(u; X \cap (A \setminus A_{ik})) \lambda_\theta(u|X \cap A) \mathrm{d}u \middle| X \cap (A \setminus A_{ik})\right]\right. \right. \\ &\left.\left. - \int_A h_\theta(u; X \cap A) \lambda_\theta(u|X \cap A) \mathrm{d}u\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{k}\int_{A_{ik}} h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\mathrm{d}u\middle|\, X\cap(A\setminus A_{ik})\right]^2\right.\\
&+\left(\int_A h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)\mathrm{d}u\right)^2\\
&-2\mathbb{E}\left[\sum_{i=1}^{k}\int_{A_{ik}} h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\mathrm{d}u\middle|\, X\cap(A\setminus A_{ik})\right]\\
&\left.\times\int_A h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)\mathrm{d}u\right]\\
\leq&\mathbb{E}\left[\left(\sum_{i=1}^{k}\int_{A_{ik}} h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\mathrm{d}u\right)^2\right]\\
&+\int_{A^2}\mathbb{E}[h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)]\mathrm{d}v\mathrm{d}u\\
&-2\sum_{i=1}^{k}\int_{A^2}\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{u\in A_{ik}\}h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\middle|\, X\cap(A\setminus A_{ik})\right]\right.\\
&\qquad\left.\times h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\right]\mathrm{d}u\mathrm{d}v\\
=&E_1(k)+E_3-2E_2(k),
\end{aligned}
$$

where the inequality is a consequence of conditioning being a contractive projection of L^p spaces. By showing that $\lim_{k\to\infty}E_1(k)=E_3$ and $\lim_{k\to\infty}E_2(k)=E_3$ we are done, since then

$$
\lim_{k\to\infty}(E_1(k)-2E_2(k)+E_3)=\lim_{k\to\infty}E_1(k)-2\lim_{k\to\infty}E_2(k)+E_3=E_3-2E_3+E_3=0.
$$

We start with $E_1(k)$. Recalling (A.1), we have that

$$
\begin{aligned}
&E_1(k)=\mathbb{E}\left[\left(\sum_{i=1}^{k}\int_{A_{ik}} h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\mathrm{d}u\right)^2\right]\\
=&\mathbb{E}\left[\left(\int_A\sum_{i=1}^{k}\mathbf{1}\{u\in A_{ik}\}h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\mathrm{d}u\right)^2\right]\\
=&\mathbb{E}\left[\int_{A^2}\sum_{i=1}^{k}\mathbf{1}\{u\in A_{ik}\}h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\sum_{j=1}^{k}\mathbf{1}\{v\in A_{jk}\}\right.\\
&\qquad\left.\times h_\theta(v;X\cap(A\setminus A_{jk}))\lambda_\theta(v|X\cap A)\mathrm{d}v\mathrm{d}u\right]\\
=&\mathbb{E}\left[\int_{A^2} h_\theta(u;\mathbf{x}\cap(A\setminus A_k(u)))\lambda_\theta(u|X\cap A)h_\theta(v;\mathbf{x}\cap(A\setminus A_k(v)))\lambda_\theta(v|X\cap A)\mathrm{d}v\mathrm{d}u\right]\\
=&\int_{A^2}\mathbb{E}\left[h_\theta(u;\mathbf{x}\cap(A\setminus A_k(u)))\lambda_\theta(u|X\cap A)h_\theta(v;\mathbf{x}\cap(A\setminus A_k(v)))\lambda_\theta(v|X\cap A)\right]\mathrm{d}v\mathrm{d}u.
\end{aligned}
$$

By Lemma A.1, Slutsky's lemma (Ferguson, 1996, Theorem 6'), the observation following (A.2) and the Fubini–Tonelli theorem, we now get that

$$
\begin{aligned}
&\lim_{k\to\infty}E_1(k)\\
=&\lim_{k\to\infty}\int_{A^2}\mathbb{E}\left[h_\theta(u;\mathbf{x}\cap(A\setminus A_k(u)))\lambda_\theta(u|X\cap A)h_\theta(v;\mathbf{x}\cap(A\setminus A_k(v)))\lambda_\theta(v|X\cap A)\right]\mathrm{d}v\mathrm{d}u\\
=&\int_{A^2}\lim_{k\to\infty}\mathbb{E}\left[h_\theta(u;\mathbf{x}\cap(A\setminus A_k(u)))\lambda_\theta(u|X\cap A)h_\theta(v;\mathbf{x}\cap(A\setminus A_k(v)))\lambda_\theta(v|X\cap A)\right]\mathrm{d}v\mathrm{d}u\\
=&\int_{A^2}\mathbb{E}\left[\lim_{k\to\infty}(h_\theta(u;\mathbf{x}\cap(A\setminus A_k(u)))\lambda_\theta(u|X\cap A)h_\theta(v;\mathbf{x}\cap(A\setminus A_k(v)))\lambda_\theta(v|X\cap A))\right]\mathrm{d}v\mathrm{d}u\\
=&\int_{A^2}\mathbb{E}\left[\left(\lim_{k\to\infty}h_\theta(u;\mathbf{x}\cap(A\setminus A_k(u)))\right)\lambda_\theta(u|X\cap A)\right.\\
&\left.\times\left(\lim_{k\to\infty}h_\theta(v;\mathbf{x}\cap(A\setminus A_k(v)))\right)\lambda_\theta(v|X\cap A)\right]\mathrm{d}v\mathrm{d}u\\
=&\int_{A^2}\mathbb{E}[(\mathbf{1}\{u\in X\cap A\}h_\theta(u;X\cap A\setminus\{u\})+\mathbf{1}\{u\notin X\cap A\}h_\theta(u;X\cap A))\lambda_\theta(u|X\cap A)\\
&\times(\mathbf{1}\{v\in X\cap A\}h_\theta(v;X\cap A\setminus\{v\})+\mathbf{1}\{v\notin X\cap A\}h_\theta(v;X\cap A))\lambda_\theta(v|X\cap A)]\mathrm{d}v\mathrm{d}u
\end{aligned}
$$

$$=\mathbb{E}\Bigg[\int_{A^2}(\mathbf{1}\{u\in X\cap A\}h_\theta(u;X\cap A\setminus\{u\})+\mathbf{1}\{u\notin X\}h_\theta(u;X\cap A))\lambda_\theta(u|X\cap A)$$
$$\times(\mathbf{1}\{v\in X\cap A\}h_\theta(v;X\cap A\setminus\{v\})+\mathbf{1}\{v\notin X\cap A\}h_\theta(v;X\cap A))\lambda_\theta(v|X\cap A)\mathrm{d}v\mathrm{d}u\Bigg].$$

Note that Lemma A.1 is applicable here since λ_θ and h_θ are bounded by assumption. Now, for any element $\omega\in\Omega$ of the underlying probability space, by the local finiteness of $\mathbf{N}$ and the boundedness of A, the realisation $X(\omega)\cap A=\mathbf{x}$ is a discrete finite collection of points. Thus, recalling that the reference measure $|\cdot|$ is non-atomic, each $X(\omega)\cap A=\mathbf{x}$ is a $|\cdot|$-null set. This implies that the integral over the terms containing $\mathbf{1}\{u\in X\cap A\}$ and $\mathbf{1}\{v\in X\cap A\}$ are 0. Hence, $E_1(k)$ tends to

$$\mathbb{E}\left[\int_{A^2}\mathbf{1}\{u,v\notin X\cap A\}h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\mathrm{d}v\mathrm{d}u\right]$$

as $k\to\infty$. The same $|\cdot|$-null set arguments further yield that this integral is indistinguishable from the integral which we obtain by excluding $\mathbf{1}\{u,v\notin X\cap A\}$ above. Consequently,

$$\lim_{k\to\infty}E_1(k)=\int_{A^2}\mathbb{E}[h_\theta(u,X\cap A)\lambda_\theta(u|X\cap A)h_\theta(v,X\cap A)\lambda_\theta(v|X\cap A)]\mathrm{d}v\mathrm{d}u=E_3.$$

Turning to $E_2(k)$, we have that

$$\begin{aligned}E_2(k)&=\int_{A^2}\sum_{i=1}^{k}\mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}\{u\in A_{ik}\}h_\theta(u;X\cap(A\setminus A_{ik}))\lambda_\theta(u|X\cap A)\Big|X\cap(A\setminus A_{ik})\Big]\\&\quad\times h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\Big]\mathrm{d}u\mathrm{d}v\\&=\int_{A^2}\mathbb{E}\Bigg[\mathbb{E}\Bigg[\lambda_\theta(u|X\cap A)\sum_{i=1}^{k}\mathbf{1}\{u\in A_{ik}\}h_\theta(u;X\cap(A\setminus A_{ik}))\Bigg|X\cap(A\setminus A_{ik})\Bigg]\\&\quad\times h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\Bigg]\mathrm{d}u\mathrm{d}v.\end{aligned}$$

By the self-adjointness of conditional expectations we now obtain

$$\begin{aligned}E_2(k)&=\int_{A^2}\mathbb{E}\Bigg[\lambda_\theta(u|X\cap A)\sum_{i=1}^{k}\mathbf{1}\{u\in A_{ik}\}h_\theta(u;X\cap(A\setminus A_{ik}))\\&\quad\times\mathbb{E}\Big[h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\Big|X\cap(A\setminus A_{ik})\Big]\Bigg]\mathrm{d}u\mathrm{d}v\\&=\int_{A^2}\mathbb{E}\Big[\lambda_\theta(u|X\cap A)h_\theta(u;X\cap(A\setminus A_k(u)))\\&\quad\times\mathbb{E}\Big[h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\Big|X\cap(A\setminus A_k(u))\Big]\Big]\mathrm{d}u\mathrm{d}v.\end{aligned}$$

As $k\to\infty$, $A_k(u)$ is a decreasing sequence of sets, and $A\setminus A_k(u)$ is an increasing sequence of sets tending to A. This means that the σ-algebras $\sigma(X\cap(A\setminus A_k(u)))$, $k\geq 2$, are increasing (in terms of set inclusion), tending to $\sigma(X\cap A)$. Thus, we may apply Martingale convergence (Durrett, 1995, Theorem 5.7) to obtain that $\mathbb{E}[h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)|X\cap(A\setminus A_k(u))]$ a.s. tends to $\mathbb{E}[h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)|X\cap A]$ as $k\to\infty$. To see that this is indeed an increasing sequence, note that the finite dimensional distributions of $X\cap(A\setminus A_k(u))$ (which characterise its distribution) are all contained in the collection of finite dimensional distributions of $X\cap(A\setminus A_{k+1}(u))$, since $A\setminus A_k\subseteq A\setminus A_{k+1}$ by the imposed refinement property. Moreover, by the observation following (A.2) we have that $h_\theta(u;X\cap(A\setminus A_k(u)))$ a.s. tends to $\mathbf{1}\{u\in X\cap A\}h_\theta(u;X\cap A\setminus\{u\})+\mathbf{1}\{u\notin X\cap A\}h_\theta(u;X\cap A)$ as $k\to\infty$. Hence, by Lemma A.1, Slutsky's lemma (Ferguson, 1996, Theorem 6'), the $|\cdot|$-null set arguments above and the law of total expectation, we obtain

$$\begin{aligned}\lim_{k\to\infty}E_2(k)&=\int_{A^2}\mathbb{E}\Bigg[\lambda_\theta(u|X\cap A)\lim_{k\to\infty}h_\theta(u;X\cap(A\setminus A_k(u)))\\&\qquad\times\lim_{k\to\infty}\mathbb{E}\left[h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\middle|X\cap(A\setminus A_k(u))\right]\Bigg]\mathrm{d}u\mathrm{d}v\\&=\int_{A^2}\mathbb{E}\left[h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)\mathbb{E}\left[h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\middle|X\cap A\right]\right]\mathrm{d}u\mathrm{d}v\\&=\int_{A^2}\mathbb{E}\left[\mathbb{E}\left[h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\middle|X\cap A\right]\right]\mathrm{d}u\mathrm{d}v\\&=\int_{A^2}\mathbb{E}\left[h_\theta(u;X\cap A)\lambda_\theta(u|X\cap A)h_\theta(v;X\cap A)\lambda_\theta(v|X\cap A)\right]\mathrm{d}u\mathrm{d}v=E_3.\end{aligned}$$

This completes the proof. □

A.2.3. *Proof of Corollary 1*

Here follows the proof of Corollary 1, which provides sufficient conditions for Theorems 1 and 2 to be satisfied.

Proof of Corollary 1. The conditional intensity itself is bounded by assumption so we proceed with the test function. The case $\alpha = 0$ is trivial so we focus on $\alpha > 0$. Since

$$h_\theta(u; X^T) = 1/(\xi_\theta(u; X^T))^\alpha = \exp\{-\alpha \log(\xi_\theta(u; X^T))\} = \exp\{-\alpha \log(p(u)\mathbb{E}\left[\lambda_\theta(u|X)\middle| X^T\right])\}$$
$$= \exp\{\alpha\left(-\log p(u) - \log(\mathbb{E}\left[\lambda_\theta(u|X)\middle| X^T\right])\right)\},$$

the test function is bounded if $-\log p(u) - \log(\mathbb{E}\left[\lambda_\theta(u|X)\middle| X^T\right])$ is smaller than some finite positive constant. This holds because $\sup_{p(u)\in(0,1)} \log p(u) = 0$ and because $\log \mathbb{E}[\lambda_\theta(u|X)|X^T] \leq \log C < \infty$, where $\lambda_\theta(u|X) \leq C < \infty$, by the monotonicity of conditional expectations. □

A.2.4. Proof of Lemma 1

Proof of Lemma 1. Letting

$$Y_{i,k} = T_k(X^V_{i,k} \cap A) - \mu_k(X) = T_k(X^V_{i,k} \cap A) - \mathbb{E}[T_k(X^V_{i,k} \cap A)|X], \qquad i = 1, \dots, k,$$

the first statement equates to showing that $k^{-1}\sum_{i=1}^k Y_{i,k}$ tends to 0 in probability, as $k \to \infty$.

We first note that $\mathbb{E}[Y_{i,k}|X] = \mathbb{E}[T_k(X^V_{i,k} \cap A) - \mathbb{E}[T_k(X^V_{1,k} \cap A)|X]|X] = 0$ and $\mathrm{Var}(Y_{i,k}|X) = \mathbb{E}[Y^2_{i,k}|X] - \mathbb{E}[Y_{i,k}|X]^2 = \mathbb{E}[Y^2_{i,k}|X] = \mathbb{E}[T_k(X^V_{i,k} \cap A)^2|X] - 2\mathbb{E}[T_k(X^V_{i,k} \cap A)\mathbb{E}[T_k(X^V_{1,k} \cap A)|X]|X] + \mathbb{E}[\mathbb{E}[T_k(X^V_{1,k} \cap A)|X]^2|X] = \mathrm{Var}(T_k(X^V_{1,k} \cap A)|X)$, by properties of conditional expectations. Since $(X^V_{i,k}, \mathbb{E}[T_k(X^V_{1,k} \cap A)|X])$, $i = 1, \dots, k$, are conditionally iid given X, so are $Y_{i,k}$, $i = 1, \dots, k$, and hereby

$$\mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^k Y_{i,k}\middle| X\right) = \frac{1}{k^2}\sum_{i=1}^k \mathrm{Var}(Y_{i,k}|X) = \frac{1}{k}\mathrm{Var}(T_k(X^V_{1,k} \cap A)|X).$$

Combining this with the conditional version of Chebyshev's inequality, we obtain that

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^k Y_{i,k}\right| > \varepsilon\middle| X\right) \leq \frac{1}{\varepsilon^2 k}\mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^k Y_{i,k}\middle| X\right) = \frac{1}{\varepsilon^2 k}\mathrm{Var}(T_k(X^V_{1,k} \cap A)|X),$$

for any $\varepsilon > 0$, and taking expectations on both sides yields

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^k Y_{i,k}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\frac{\mathbb{E}[\mathrm{Var}(T_k(X^V_{1,k} \cap A)|X)]}{k},$$

which tends to 0 by assumption; note that $\mathbb{E}[\mathrm{Var}(T_k(X^V_{1,k} \cap A)|X)] = \mathrm{Var}(T_k(X^V_{1,k} \cap A)) - \mathrm{Var}(\mathbb{E}[T_k(X^V_{1,k} \cap A)|X])$ by the law of total variance. This completes the proof of the first statement.

Turning to the second statement, note that

$$\frac{1}{k}\sum_{i=1}^k T_k(X^V_{i,k} \cap A) - T(X \cap A) = \left(\frac{1}{k}\sum_{i=1}^k T_k(X^V_{i,k} \cap A) - \mu_k(X)\right) + \left(\mu_k(X) - T(X \cap A)\right).$$

We have just established that the first term on the right-hand side tends to 0 in probability, as $k \to \infty$, and the second term tends to 0 in probability since we have imposed that $\mu_k(X) = \mathbb{E}[T_k(X^V_{i,k} \cap A)|X] \to T(X \cap A)$ in probability, as $k \to \infty$. By an application of Slutsky's lemma (Ferguson, 1996, Theorem 6'), we now obtain that also the sum of these two terms tend 0 in probability as $k \to \infty$, and this completes the proof. □

References

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

Baddeley, A., Coeurjolly, J.-F., Rubak, E., Waagepetersen, R., 2014. Logistic regression for spatial gibbs point processes. Biometrika 101 (2), 377–392.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. CRC Press.

Baddeley, A., Turner, R., Møller, J., Hazelton, M., 2005. Residual analysis for spatial point processes (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (5), 617–666.

Betsch, S., 2023. Structural properties of gibbsian point processes in abstract spaces. J. Theoret. Probab. 36, 2501–2563.

Coeurjolly, J.-F., Guan, Y., Khanmohammadi, M., Waagepetersen, R., 2016. Towards optimal takacs–fiksel estimation. Spat. Stat. 18, 396–411.

Coeurjolly, J.-F., Lavancier, F., 2019. Understanding spatial point patterns through intensity and conditional intensities. In: Coupier, D. (Ed.), Stochastic Geometry, Lecture Notes in Mathematics, Vol 2237. Springer, pp. 45–85.

Cronie, O., Jansson, J., Konstantinou, K., 2024a. Discussion of the paper "marked spatial point processes: Current state and extensions to point processes on linear networks". J. Agric. Biol. Environ. Stat. 29, 379–388.

Cronie, O., van Lieshout, M.N.M., 2016. Summary statistics for inhomogeneous marked point processes. "Annals Stat. Math." 68, 905–928.

Cronie, O., van Lieshout, M.N.M., 2018. A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. Biometrika 105 (2), 455–462.

Cronie, O., Moradi, M., Biscio, C.A., 2024b. A cross-validation-based statistical theory for point processes. Biometrika 111 (2), 625–641.

Cronie, O., Moradi, M., Mateu, J., 2020. Inhomogeneous higher-order summary statistics for point processes on linear networks. Stat. Comput. 30 (5), 1221–1239.

Daley, D.J., Vere-Jones, D., 2003. An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods, second ed. Springer-Verlag New York.

D'Angelo, N., Adelfio, G., Mateu, J., Cronie, O., 2023. Local inhomogeneous weighted summary statistics for marked point processes. J. Comput. Graph. Statist. 1–15.

Diggle, P.J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D., Tanemura, M., 1994. On parameter estimation for pairwise interaction point processes. Int. Stat. Rev. 62, 99–117, URL https://api.semanticscholar.org/CorpusID:119917960.

Durrett, R., 1995. Probability: Theory and Examples. Duxbury Press, Second edition.

Ferguson, T.S., 1996. A Course in Large Sample Theory. Chapman & Hall.

Fiksel, T., 1984. Estimation of parameterized pair potentials of marked and non-marked gibbsian point processes. Elektron. Inf. Kybernet. 20, 270–278.

Georgii, H.-O., 1976. Canonical and grand canonical gibbs states for continuum systems. Comm. Math. Phys. 48, 31–51.

Jansson, J., Biscio, C., Moradi, M., Cronie, O., 2024. Point process learning: a cross-validation-based statistical framework for point processes. In: Pollice, A., Mariani, P. (Eds.), Methodological and Applied Statistics and Demography I: SIS 2024, Short Papers, Plenary and Specialized Sessions. In: Proceedings of the Scientific Meeting of the Italian Statistical Society, 52, Springer, the Italian Statistical Society, Bari, Italy.

Jansson, J., Cronie, O., 2024. Comparison of point process learning and its special case takacs-fiksel estimation. arXiv preprint arXiv:2405.19523.

Kresin, C., Schoenberg, F., 2023. Parametric estimation of spatial–temporal point processes using the stoyan–grabarnik statistic. Ann. Inst. Statist. Math. 1–23.

van Lieshout, M., 2000. Markov Point Processes and Their Applications. Imperial College Press/World Scientific.

van Lieshout, M., 2021. Infill asymptotics for adaptive kernel estimators of spatial intensity. Aust. N. Z. J. Stat. 63 (1), 159–181.

Møller, J., Waagepetersen, R., 2004. Statistical Inference and Simulation for Spatial Point Processes. CRC Press.

Moradi, M.M., Cronie, O., Rubak, E., Lachieze-Rey, R., Mateu, J., Baddeley, A., 2019. Resample-smoothing of voronoi intensity estimators. Stat. Comput. 29 (5), 995–1010.

Nguyen, X.X., Zessin, H., 1979. Integral and differential characterizations of the gibbs process. Math. Nachr. 88 (1), 105–115.

Schreiber, T., Yukich, J., 2013. Limit theorems for geometric functionals of gibbs point processes. Ann. L'IHP Probabilités Stat. 49 (4), 1158–1182.

Stoyan, D., Grabarnik, P., 1991. Second-order characteristics for stochastic structures connected with gibbs point processes. Math. Nachr. 151 (1), 95–100.

Stoyan, D., Stoyan, H., 2000. Improving ratio estimators of second order point process characteristics. Scand. J. Stat. 27 (4), 641–656.

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

Takacs, R., 1986. Estimator for the pair–potential of a gibbsian point process. Stat.: A J. Theor. Appl. Stat. 17 (3), 429–433.

Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer science & business media.