Biometrika (2023), **...**, ..., *pp*. 1–25
Printed in Great Britain
Advance Access publication on ...

A cross-validation-based statistical theory for point processes

BY OTTMAR CRONIE
Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, 412 96 Gothenburg, Sweden
ottmar@chalmers.se, ottmar.cronie@gu.se

MEHDI MORADI
Department of Mathematics and Mathematical Statistics, Umeå University, 901 87 Umeå, Sweden
mehdi.moradi@umu.se

CHRISTOPHE A.N. BISCIO
Department of Mathematical Sciences, Aalborg University, 9220 Aalborg, Denmark
christophe@math.aau.dk

SUMMARY

Motivated by cross-validation's general ability to reduce overfitting and mean square error, we develop a cross-validation-based statistical theory for general point processes. It is based on the combination of two novel concepts for general point processes: cross-validation and prediction errors. Our cross-validation approach uses thinning to split a point process/pattern into pairs of training and validation sets, while our prediction errors measure discrepancy between two point processes. The new statistical approach, which may be used to model different distributional characteristics, exploits the prediction errors to measure how well a given model predicts validation sets using associated training sets. Having indicated that our new framework generalizes many existing statistical approaches, we then establish different theoretical properties for it, including large sample properties. We further recognize that non-parametric intensity estimation is an instance of Papangelou conditional intensity estimation, which we exploit to apply our new statistical theory to kernel intensity estimation. Using independent thinning-based cross-validation, we numerically show that the new approach substantially outperforms the state of the art in bandwidth selection. Finally, we carry out intensity estimation for a dataset in forestry (Euclidean domain) and a dataset in neurology (linear network).

Some key words: kernel intensity estimation; Papangelou conditional intensity; prediction; spatial statistics; thinning

1. INTRODUCTION

A point process in a general space (Daley & Vere-Jones, 2003, 2008; Kallenberg, 2017) can be viewed as a generalized random sample, where we allow the sample size to be random and/or the sample points to be dependent random variables; an independent and identically distributed (iid) sample is called a Binomial point process (van Lieshout, 2000). Consequently, they have been extensively applied to analyse and model various event data sources, e.g. in forestry and epidemiology (Diggle, 2014). As one typically observes only one point pattern, i.e. point process

realization, classical iid sample statistics is infeasible and, in addition, likelihood estimation is generally intractable (van Lieshout, 2000). This has led to the development of a range of innovative statistical approaches (Coeurjolly & Lavancier, 2019), where the associated estimation criteria to be optimized, which take the full observed pattern as input, are not (explicitly) based on predictive estimation ideas. Generally speaking, models with small cross-validation errors yield good out-of-sample prediction performances and tend to result in little overfitting and small mean square errors (Hastie et al., 2009; Arlot & Celisse, 2010). Hence, a general cross-validation-based statistical theory for point processes could reduce mean (integrated) square errors when fitting different distributional characteristics, but such a theory does not currently exist.

This paper addresses the development of a general cross-validation approach as well as a cross-validation-based predictive statistical theory for general point processes. Our cross-validation approach, which is inspired by our previous work on point process subsampling (Moradi et al., 2019), is defined through thinning and allows us to consider a form of conditional iid sampling of a point process. Besides cross-validation, our statistical approach is further based on a new notion of point process prediction errors, which is inspired by our previous work on non-parametric intensity estimation (Cronie & van Lieshout, 2018). Our prediction errors, which allow us to measure the quality of a proposed estimate, can be thought of as measures of discrepancy between two point processes. More specifically, any prediction error is given by the difference between two parametrized terms, a random one and a deterministic one. The random term is a sum over the first point process, where each summand depends on i) the second point process, ii) a point of the first point process, and iii) a candidate parameter. The deterministic term is equal to the expectation of the random term if and only if the candidate parameter is set to the true one, for a well specified model. When the two point processes coincide, i.e. under auto-prediction, in a certain setting our prediction errors reduce to so-called innovations, originally introduced by Baddeley et al. (2005, 2008) to define residuals for (Papangelou) conditional intensity models. Our prediction errors further reduce to various loss functions for existing estimation approaches, e.g. i) the approach of Takacs (1986) and Fiksel (1984) for conditional intensity modelling (Coeurjolly et al., 2016), with pseudo-likelihood estimation as a special case, ii) the quasi-likelihood approach of Guan et al. (2015) for parametric intensity estimation, which has composite-/Poisson- and Palm-likelihood estimation (Coeurjolly & Lavancier, 2019) as special cases, and iii) the non-parametric intensity estimation approach of Cronie & van Lieshout (2018). By combining our two new concepts, we arrive at the definition of our new statistical theory, referred to as point process learning due to its similarities with risk minimization in statistical learning (Vapnik, 1999).

We establish different properties and variations of point process learning, in particular how it can be applied to parametric product density/intensity estimation and conditional intensity estimation. We then focus on non-parametric intensity estimation, which we indicate is an instance of conditional intensity estimation. In particular, we apply our new approach to bandwidth selection in kernel intensity estimation, when the cross-validation is achieved through independent thinning. We find that point process learning numerically outperforms the state of the art, i.e. the Cronie & van Lieshout (2018) approach, in terms of mean integrated square error, regardless of the degree of spatial interaction in the underlying point process.

For readability, in the main text we state all theory for first-order/univariate statistics and defer higher-order statements and proofs to the supplementary material of the paper. Throughout, labels starting with the prefix 'S' refer to items in the supplementary material.

2. PRELIMINARIES

2.1. *Point processes and distributional characteristics*

Consider a general (complete separable metric) space S, which is endowed with a notion of size in the form of a (locally- and σ-finite Borel) reference measure $A \mapsto |A| = \int_A \mathrm{d}u$, $A \subseteq S$; we here reserve the notation "$\subseteq$" for Borel sets of S. For convenience, one may think of $S = \mathbb{R}^d$, $d \geq 1$, equipped with the d-dimensional Euclidean metric $d(u, v) = \|u - v\|$ and Lebesgue measure $|\cdot|$, or a linear network $S = L = \bigcup_{i=1}^{k} l_i$, i.e. a union of connected line segment $l_i \subseteq \mathbb{R}^2$, where $d(\cdot, \cdot)$ is the shortest-path metric and $|\cdot|$ represents arc length integration on L (Baddeley et al., 2015; Cronie et al., 2020); see Figure 1 for illustrations. Throughout, we will implicitly assume that functions are sufficiently measurable/integrable, and we abbreviate the terms almost sure(ly) and almost everywhere by a.s. and a.e..

Given a suitable probability space $(\Omega, \mathcal{F}, \mathrm{pr})$, a point process $X = \{x_i\}_{i=1}^{N}$, $0 \leq N \leq \infty$, in S may be defined as a random element in the measurable space $(\mathcal{X}, \mathcal{N}) = (\mathcal{X}_S, \mathcal{N})$ of point patterns/configurations $\mathbb{x} = \{x_1, \ldots, x_n\} \subseteq S$, $0 \leq n \leq \infty$, which are locally finite, i.e. where the cardinality $\#(\mathbb{x} \cap A) = \sum_{i=1}^{n} \mathbb{1}(x_i \in A)$ is finite for bounded $A \subseteq S$ (Daley & Vere-Jones, 2003, 2008; Møller & Waagepetersen, 2004). A member x_i of a point pattern $\mathbb{x}$ or a point process X is commonly called an event. We identify X with the random measure $X(A) = \#(X \cap A)$, $A \subseteq S$, which is simple, meaning that a.s. $X(\{u\}) \in \{0, 1\}$, $u \in S$, i.e. X has at most one event at any location.

The distribution of a point process X is most conveniently described by its (Papangelou) conditional intensity, λ. It satisfies the Georgii–Nguyen–Zessin (GNZ) formula/theorem, which states that (Daley & Vere-Jones, 2008)

$$E\left\{\sum_{x \in X} h(x, X \setminus \{x\})\right\} = \int_S E\left\{h(u, X)\lambda(u; X)\right\} \mathrm{d}u$$

for non-negative (possibly infinite) and integrable $h : S \times \mathcal{X} \to \mathbb{R}$. It has the interpretation that the conditional probability of finding a point of X in an infinitesimal neighbourhood du of $u \in S$, given that X agrees with $\mathbb{x}$ outside du, satisfies $\mathrm{pr}\{X(du) = 1 \mid X \cap S \setminus du = \mathbb{x} \cap S \setminus du)\} = \lambda(u; \mathbb{x})\mathrm{d}u$ (Coeurjolly et al., 2017). This interpretation is motivated by the fact that, for a finite point process, i.e. if $N = X(S) < \infty$ a.s., we can express λ as a ratio of Janossy densities, which thus implies that $\lambda(\cdot)$ can be readily derived when the Janossy densities are known in closed form (Daley & Vere-Jones, 2008). Unfortunately, the Janossy densities, which yield the likelihood function, are generally not tractable (van Lieshout, 2000), but luckily there exist many models with explicit forms for λ, e.g. Poisson, Cox (Møller & Waagepetersen, 2004), Hawkes (Yang et al., 2019), Markov (van Lieshout, 2000) and hybrid Gibbs (Baddeley et al., 2015) point processes. Regarding dependencies in X, when $\mathbb{y} \subseteq \mathbb{x}$, if $\lambda(\cdot; \mathbb{y})$ is smaller/larger than or equal to $\lambda(\cdot; \mathbb{x})$, we call X attractive/repulsive.

By letting h in the GNZ formula be constant over its second argument, we obtain the Campbell formula, which yields that $E\{\lambda(u; X)\} = \rho(u)$, the intensity function of X, where $E\{X(A)\} = \int_A \rho(u)\mathrm{d}u$, $A \subseteq S$. Heuristically, $\mathrm{pr}\{X(du) = 1\} = E\{X(du)\} = \rho(u)\mathrm{d}u$ and whenever $\rho(\cdot) \equiv \rho > 0$ is (non-)constant, we say that X is (in)homogeneous.

By replacing X by the point process consisting of all distinct n-tuples of elements of X, $X^n_{\neq} = \{(x_1, \ldots, x_n) \in X^n : x_i \neq x_j \text{ if } i \neq j\} \subseteq S^n$, we obtain nth-order conditional intensities $\lambda^{(n)}$ and product densities/intensities $\rho^{(n)}$ (Coeurjolly et al., 2017); see Section S5.1 for details.

2.2. *Point process statistics*

Assume that we observe/sample a point pattern $\mathbb{x} = \{x_1, \dots, x_n\}$ within some, potentially bounded, study region/domain $W \subseteq S$, $|W| > 0$, which has been generated by $X \cap W$, for some unknown point process X. Here, in contrast to the classical iid setting, we have only one realization of the random element of interest.

Statistical settings typically deal with estimation of some particular characteristic of X and it turns out that the associated estimators can be characterized by what we will refer to as a general parametrized estimator family $\Xi_\Theta = \{\xi_\theta : \theta \in \Theta\}$, where

$$\xi_\theta(u; \mathbb{x}), \quad u \in S, \quad \mathbb{x} \in \mathcal{X}, \quad \theta \in \Theta, \tag{1}$$

are real-valued and $\xi_\theta(\cdot; \mathbb{x})$ is either non-negative or integrable for any $\mathbb{x}$. Typically, $\Theta \subseteq \mathbb{R}^l$, $l \geq 1$, but one could imagine other forms of parametrization; cf. Vapnik (1999). When each ξ_θ is constant over $\mathbb{x} \in \mathcal{X}$, i.e. it does not depend on $\mathbb{x}$, we set

$$\xi_\theta(u; \mathbb{x}) \equiv \xi_\theta(u), \quad u \in S, \quad \mathbb{x} \in \mathcal{X}, \quad \theta \in \Theta. \tag{2}$$

This definition naturally extends to the nth-order setting; see Section S5.3.

To carry out estimation, one typically finds a minimizer, an estimate $\widehat{\theta} = \widehat{\theta}_W(\mathbb{x}) \in \Theta$, through some loss function $\mathcal{L}(\theta) = \mathcal{L}(\xi_\theta, \mathbb{x}, W)$, $\theta \in \Theta$. Ideally, $\mathcal{L}(\theta)$ is constructed such that the estimator, $\widehat{\theta}_W(X)$, properly describes the characteristic of interest, in some suitable distributional sense, e.g. a mean square error sense. When we do not work under model miss-specification, we assume that the true characteristic of interest is parametrized by some $\theta_0 \in \Theta$. Many common statistical frameworks (Møller & Waagepetersen, 2017; Coeurjolly & Lavancier, 2019) can be expressed through general parametrized estimator families with accompanying loss functions. Examples include parametric product density/intensity estimation, $\rho^{(n)}_\theta$, $\theta \in \Theta$, $n \geq 1$, encompassing also Palm likelihood estimation and K-function-based minimum contrast estimation, taking the form (2), as well as parametric conditional intensity estimation, λ_θ, $\theta \in \Theta$, and non-parametric product density/intensity estimation, which have the form (1). Mathematically speaking, a non-parametric intensity estimator $\widehat{\rho}_\theta$ (van Lieshout, 2012) has the form of a parametrized conditional intensity; it is an attractive model since the addition of a point to $\mathbb{x}$ (close to u) increases the value of $\widehat{\rho}_\theta(u, \mathbb{x})$.

We will illustrate our new theory by focusing on non-parametric intensity estimation, in particular kernel intensity estimation (van Lieshout, 2012): for a point pattern $\mathbb{x} \subseteq W \subseteq S = \mathbb{R}^d$,

$$\widehat{\rho}_\theta(u, \mathbb{x}) = \sum_{x \in \mathbb{x}} \frac{\kappa_\theta(u - x)}{e_\theta(u, x)} = \sum_{x \in \mathbb{x}} \frac{\theta^{-d} \kappa\{(u - x)/\theta\}}{e_\theta(u, x)}, \qquad u \in W, \tag{3}$$

where the kernel κ is a symmetric density function and e_θ is an edge correction term compensating for potential interactions with points outside W; examples include $e_\theta(u, x) = \int_W \kappa_\theta(v - x)\mathrm{d}v$, which ensures that $\int_W \widehat{\rho}_\theta(u, \mathbb{x})\mathrm{d}u = \#\mathbb{x}$, and $e_\theta(u, x) \equiv 1$, which represents no edge correction. The main challenge here is optimal selection of the bandwidth, i.e. the smoothing parameter $\theta \in \Theta = (0, \infty)$, as, generally speaking, the kernel choice has a much less pronounced role than the chosen θ (Silverman, 1986). For other, possibly non-Euclidean, domains, κ and thereby the kernel estimator may look somewhat different and also be quite abstract (Di Marzio et al., 2014; McSwiggan et al., 2017; Rakshit et al., 2019; Mateu et al., 2020). In certain cases, however, there are straightforward extensions of (3); see Section S4 for the case of linear networks.

Our main focus in this paper will be bandwidth selection, and we here follow Cronie & van Lieshout (2018), who, in the context of Takacs–Fiksel estimation, implicitly suggested the following for non-parametric intensity estimation: fit (3) to $\mathbb{x}$, using a suitable conditional intensity estimation method, to obtain $\widehat{\theta}$ and the final intensity estimate $\widehat{\rho}_{\widehat{\theta}}(\cdot; \mathbb{x})$. We see that if the method is

doing a good job in the sense that $\widehat{\rho}_{\widehat{\theta}} = \lambda_{\widehat{\theta}}$ is close to the conditional intensity λ of X, we have that $\widehat{\rho}_{\widehat{\theta}}(\cdot; \aleph) \approx \lambda(\cdot; \aleph) \approx E[\lambda(\cdot; X)] = \rho(\cdot)$; the last approximation follows since, on average, $\lambda(\cdot; \aleph)$ is close to the expectation of $\lambda(\cdot; X)$. Since we apply the same model $\widehat{\rho}_\theta$, $\theta \in \Theta$, regardless of the (unknown) underlying distribution, this is (in general) an instance of model misspecification.

3. Thinning-based cross-validation

3.1. *Thinning*

Heuristically, a thinning $Z \subseteq X$ is generated by applying some rule/mechanism to X which either retains or deletes each $x \in X$ (Chiu et al., 2013). At the same time, marked point processes are used when each event carries additional information, not directly connected to S, e.g. a label, a quantitative measurement, a function or a set (Chiu et al., 2013; Cronie & van Lieshout, 2016; Ghorbani et al., 2020). We next formalize thinning through bivariate markings of point processes.

Definition 1. *Given a point process $X = \{x_i\}_{i=1}^N \subseteq S$, a thinning Z of X with retention probability $p : S \times \mathcal{X} \to [0, 1]$ may be defined as the marginal point process $Z = \{x : (x, m) \in \breve{X}, m = 1\}$ of a bivariate marking $\breve{X} = \{(x_i, m_i)\}_{i=1}^N \subseteq S \times \mathcal{M}$, $\mathcal{M} = \{0, 1\}$, of X. Here, $m_i = m(x_i) \in \mathcal{M}$, $i = 1, \dots, N$, for some (possibly random) marking function $m(\cdot)$, governing the retention probability. We let the reference measure on $\mathcal{M}$ be the counting measure.*

When $\breve{X}$ is independently marked, i.e. the marks are independent conditional on X, whereby the retention probability $p(u)$, $u \in S$, does not depend on X, we say that Z is an independent thinning. If, in addition, $p(\cdot) \equiv p \in [0, 1]$, we say that Z is a p-thinning.

Independent thinnings are particularly tractable and in Theorem 1 we provide important results on such thinnings, which will be used to establish certain properties of our statistical theory. Theorem 1 is stated and proved in the general nth-order setting in Section S5.

Theorem 1. *Let Z be a p-thinning of a point process X on S, with retention probability $p(u) \in (0, 1)$, $u \in S$. Given $Y = X \setminus Z$, for any non-negative or integrable $h : S \times \mathcal{X} \to \mathbb{R}$,*

$$E\left\{\sum_{x \in Z} h(x, Y)\right\} = E\left\{\sum_{x \in Y} h(x, Y \setminus \{x\}) \frac{p(x)}{1 - p(x)}\right\}.$$

Moreover, provided that they exist, the conditional intensity and the intensity of Z a.e. satisfy

$$\begin{aligned} \lambda_Z(u, Z) &\overset{a.s.}{=} p(u) E\{\lambda_X(u; X) \mid Z\}, \\ \rho_Z(u) &= p(u)\rho_X(u), \end{aligned}$$

where λ_X and ρ_X are the conditional intensity and the intensity of X. Moreover, the associated Palm intensities (see Section S5) satisfy $\rho_Z^!(u|v) = p(u)\rho_X^!(u|v)$. Given the associated marked point process representation $\breve{X}$ in Definition 1, when the conditional intensities of $\breve{X}$ and Y exist, they satisfy $E[\breve{\lambda}\{(u, 1); \breve{X}\} \mid Y] = \lambda_Y(u; Y)p(u)/\{1 - p(u)\}$ for almost all $u \in S$. In particular, for a p-thinning with retention probability $p \in (0, 1)$, we set $p(\cdot) \equiv p$ above.

3.2. *Cross-validation for point processes*

Broadly speaking, cross-validation refers to a family of techniques which, in different ways, repeatedly split the dataset $\aleph$ into a training set $\aleph_i^T \subseteq \aleph$ and a validation set $\aleph_i^V = \aleph \setminus \aleph_i^T$, $i = 1, \dots, k$, with the aim of assessing a model's generalizability (Arlot & Celisse, 2010); we here use the terms training, validation and test sets in a rather loose fashion. In the classical random sample setting, generalizability is often described in terms of out-of-sample prediction performance with

respect to "new" data, i.e. additional draws from the underlying distribution. The notion of "new" data makes little sense in the context of point processes, where one is typically dealing with just one realization of a sample which (potentially) possesses both dependence and a random sample size. However, the equally common terms "unseen" and "hold-out" make more sense here since they have a clear meaning for point processes, namely splitting through thinning.

DEFINITION 2 (POINT PROCESS CROSS-VALIDATION). *Given $k \geq 1$ thinnings $Z_1, \ldots, Z_k$ of a point process $X \subseteq S$, we refer to the collection of pairs $(X_i^T, X_i^V) = (Y_i, Z_i)$, $Y_i = X \setminus Z_i$, $i = 1, \ldots, k$, as a cross-validation splitting/partitioning. For a point pattern $\mathbb{x}$, we write $(\mathbb{x}_i^T, \mathbb{x}_i^V)$.*

If we have access to $k' \geq 1$ independent copies $X_1, \ldots, X_{k'}$ of X, we consider k splits for each X_i, giving $k \times k'$ training-validation pairs. Clearly, we may carry out the partitioning in an infinite number of ways by using different thinning strategies. However, dependent thinnings are hard to deal with since, for arbitrary point processes, it is generally hard to derive distributional properties for them; we essentially have no control over the dependence structures between the training and validation sets. In classical k-fold cross-validation, $\mathbb{x}$ is split into k folds, having (approximately) the same fixed cardinality, and in each round, $i = 1, \ldots, k$, the ith fold plays the role of $\mathbb{x}_i^V$, while the union of the remaining $k-1$ folds plays the role of $\mathbb{x}_i^T$. By setting $k = \#\mathbb{x}$, so that $\mathbb{x}_i^V = \{x_i\}$ and $\mathbb{x}_i^T = \mathbb{x} \setminus \{x_i\}$, $i = 1, \ldots, \#\mathbb{x}$, we obtain leave-one-out cross-validation. These sequential algorithms do not result in independent thinning, since the assignment of a point of $\mathbb{x}$ to a given fold depends on the assignments of other points; a fold runs full once it contains a given number of points of $\mathbb{x}$. Thus, due to the related intractability, we will not look closer at these. Instead, we argue that cross-validation procedures for point processes should be based on independent thinning. The main argument is that then, as we saw in Theorem 1, we have control over distributional properties of most characteristic of interest, e.g. intensity functions. There are, however, infinitely many ways to carry out independent thinning.

For simplicity reasons, we argue that cross-validation procedures for point processes should be based on p-thinning; see Figure 1 for illustrations in two different spatial domains. Next, we provide two p-thinning-based procedures: Monte-Carlo cross-validation which, in the literature, is also referred to as repeated random sub-sampling validation, and multinomial cross-validation, which is our variant of classical k-fold cross-validation. The main difference between the two is that the latter, which is computationally more efficient than the former, involves only one parameter, k, and does not allow for $\mathbb{x}_i^V$ and $\mathbb{x}_j^V$, $i \neq j$, to overlap.

DEFINITION 3. *Given $k \geq 1$ p-thinnings $\mathbb{z}_1, \ldots, \mathbb{z}_k$, $p \in (0,1)$, of a point pattern $\mathbb{x}$, we define Monte-Carlo cross-validation as setting $\mathbb{x}_i^V = \mathbb{z}_i$ and $\mathbb{x}_i^T = \mathbb{x} \setminus \mathbb{z}_i$, $i = 1, \ldots, k$.*

Given some $k \geq 2$, randomly label the point pattern $\mathbb{x}$ with iid marks $m(x) \in \{1, \ldots, k\}$, $x \in \mathbb{x}$, from a multinomial distribution with parameters k and $p_1 = \ldots = p_k = 1/k$. We define (k-fold) multinomial cross-validation by $\mathbb{x}_i^V = \{x \in \mathbb{x} : m(x) = i\}$ and $\mathbb{x}_i^T = \mathbb{x} \setminus \mathbb{x}_i^V$, $i = 1, \ldots, k$.

In Monte-Carlo cross-validation, for any point pattern $\mathbb{x}$ and a suitable function f on $\mathcal{X}$, by the law of large numbers and the central limit theorem, conditionally on $X = \mathbb{x}$, the mean $k^{-1} \sum_{i=1}^k f(X_i^V)$ converges a.s. to $E[f(X_1^V)]$ and weakly to a Gaussian random variable; in practice, associated statistical procedures may be stopped when we see indications of convergence. Further, when $p \approx 0$ we obtain something similar to the classical leave-one-out approach, where the advantage of the former over the latter is that by Theorem 1 we have theoretical control over distributional properties of X_i^T and X_i^V. In multinomial cross-validation, each validation set is a p-thinning with retention probability $1/k$ and each training set is a p-thinning with retention probability $1 - 1/k = (k-1)/k$, whereby various distributional properties are known, e.g. the intensity of X_i^V is $\rho_p(\cdot) = \rho(\cdot)/k$; recall Theorem 1. Moreover, for large datasets, multinomial

and classical k-fold cross-validation should yield very similar results since, proportionally, the point counts of the folds become approximately the same in the two. Additionally, common rules of thumb for k-fold cross-validation in the iid setting (Arlot & Celisse, 2010), e.g. $k = 5$ or $k = 10$, do not necessarily apply in the current context. Sometimes, there are optimal choices for p and k such that our new statistical approach performs better with Monte-Carlo cross-validation than with multinomial cross-validation, but such choices seem related to e.g. the sample size or the dependence structure of X (Section 5.2).

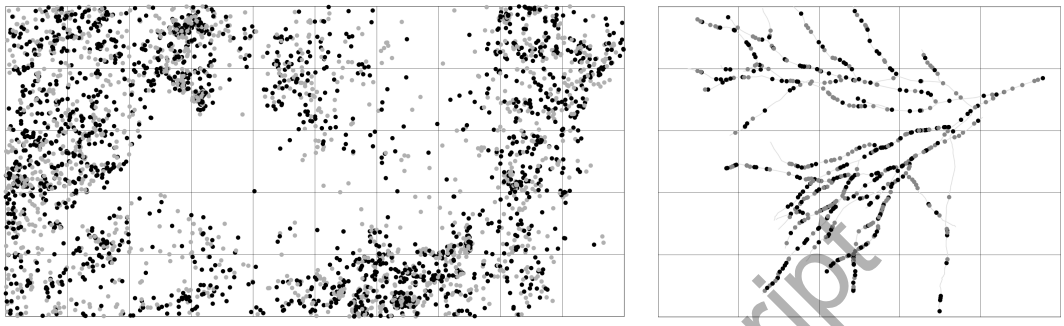

Fig. 1: A point pattern $\mathbb{x} \subseteq W \subseteq S$ together with a partition $\{W_i\}_{i=1}^k$, yielding validation sets $\mathbb{x}_i^V = \mathbb{x} \cap W_i$, $i = 1, \ldots, k \geq 2$, as well as a p-thinning-based training-validation pair (black dots indicate validation points). Left: Euclidean domain, $S = \mathbb{R}^2$. Right: Linear network, $S = L$.

Another natural and appealing approach here is block cross-validation (Roberts et al., 2017): choose $k \geq 1$ subsets $W_i \subseteq W$, $i = 1, \ldots, k$, potentially a partition of W, and let $\mathbb{x}_i^V = \mathbb{x} \cap W_i$; see Figure 1 for an illustration. This is a form of independent thinning-based k-fold cross-validation, where the ith fold is obtained through the retention probability $p_i(u) = \mathbb{1}(u \in W_i)$, $u \in W$, $i = 1, \ldots, k$. Note the philosophical difference between this, where we sample "from left to right/top to bottom", and the p-thinning-based approach, where we sample "from above". As our general statistical theory in Section 5 accommodates any (independent) thinning-based cross-validation approach, this extrapolation/interpolation kind of approach can also be combined with our new theory. However, in contrast to Definition 3, where the procedures are governed by the hyperparameters k and p, we here see a few challenges. E.g., it is not evident how to best choose the sets W_i for an unknown model (cf. Mattfeldt et al., 2013), the counts $\#\mathbb{x}_i^V$ may be highly varying, and we might introduce edge effects to be corrected for (Cronie & Särkkä, 2011; Chiu et al., 2013; Baddeley et al., 2015).

In Section S1 we further discuss alternatives to and merits of Definition 3. Moreover, Section S3.2 illustrates that, in addition to training and validation sets, we may generate test sets by treating each training set as an original point pattern to which cross-validation is applied.

4. Point process prediction

4.1. *Prediction errors*

Our statistical approach heavily relies on a new notion of prediction errors for point processes, which e.g. may be used to predict properties of one point process from another point process.

Definition 4 (Point process prediction errors). *Let* $\Xi_\Theta = \{\xi_\theta : \theta \in \Theta\}$ *and* $\mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}$ *be two general parametrized estimator families, both either of the form* (1) *or* (2)*. We refer to the members of* $\mathcal{H}_\Theta$ *as test functions. The associated families of* ($\mathcal{H}_\Theta$*-weighted)*

bivariate prediction errors $\{\mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{z},\mathbb{y}) : A \subseteq S, \mathbb{y},\mathbb{z} \in \mathcal{X}\}_{\theta\in\Theta}$ *and univariate prediction errors* $\{\mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{y}) : A \subseteq S, \mathbb{y} \in \mathcal{X}\}_{\theta\in\Theta}$ *are defined as the (signed) Borel measures*

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

and $\mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{y}) = \mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{y},\mathbb{y})$. *In particular, when* Ξ_Θ *and* $\mathcal{H}_\Theta$ *are of the form* (2) *then* $\mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{z},\mathbb{y}) = \mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{z}) = \int_A h_\theta(u)\mathcal{I}^1_{\xi_\theta}(du;\mathbb{z})$ *for any* $\mathbb{y},\mathbb{z} \in \mathcal{X}$ *and* $A \subseteq S$.

Regarding the interpretation of (4), which will be motivated in Section 4.3, for two point processes Z and Y, the random (signed) measure $\mathcal{I}^{h_\theta}_{\xi_\theta}(A;Z,Y)$, $A \subseteq S$, represents an empirical measure of how well Y predicts points of Z in A via ξ_θ and h_θ; univariate prediction errors thus correspond to auto-prediction. The test function h_θ weights the associated contributions of distinct points, ξ_θ is intended to describe the distributional properties of the superposition $Z \cup Y$, and $\mathcal{I}^{h_\theta}_{\xi_\theta}(A;Z,Y)$ estimates how well the specific choice $\theta \in \Theta$ does in predicting points of Z from Y. In some sense, this is a coupling idea. Further, by replacing the test function in (4) by either $h_\theta(u; W \cap \mathbb{y}\setminus\{u\})$ or $\mathbb{1}(\mathbb{y}\subseteq W)h_\theta(u;\mathbb{y}\setminus\{u\})$, for some (possibly) bounded $W \subseteq S$, we ensure that $\mathbb{y} \in \mathcal{X}$ is contained in W. Indeed, a prediction error family forms a (signed) transition kernel (Kallenberg, 2017) and it is straightforward to extend (4) to the nth-order case (see Section S5), where we sum over $\mathbb{z}^n_{\neq} \cap A \subseteq S^n$ and $\xi_\theta, h_\theta : S^n \times \mathcal{X} \to \mathbb{R}$.

To compute prediction errors in practice, we need to numerically approximate the integral in (4). Given any quadrature rule with quadrature points $\mathbb{v} \in \mathcal{X}$, $\mathbb{z} \subseteq \mathbb{v}$, and quadrature weights $\{w_v : v \in \mathbb{v}\}$, $\sum_{v\in\mathbb{v}} w_v = |S|$, we may exploit the Berman–Turner device (Berman & Turner, 1992) to approximate the integral in (4), i.e. $\mathcal{I}^{h_\theta}_{\xi_\theta}(A;\mathbb{z},\mathbb{y}) \approx \sum_{v\in\mathbb{v}\cap A} h_\theta(v;\mathbb{y})\{\mathbb{1}(v\in\mathbb{z}) - \xi_\theta(v;\mathbb{y})w_v\}$.

4.2. *Auto-prediction: Innovations and classical statistical approaches*

Before we turn to studying different properties of our prediction errors, we point to a connection between them and the so-called innovations of Baddeley et al. (2005, 2008), as well as the related estimating equation approaches considered in the literature; see Møller & Waagepetersen (2017); Coeurjolly & Lavancier (2019) and the references therein. An innovation coincides with a univariate prediction error $\mathcal{I}^{h_\theta}_{\lambda_\theta}(A;X)$, where X is some point process and $\xi_\theta(\cdot;\cdot) = \lambda_\theta(\cdot;\cdot)$ belongs to a parametric family of conditional intensity functions. It should be emphasized that Baddeley et al. (2005, 2008) used innovations to define point process residuals, obtained by replacing θ by an estimate $\widehat{\theta}$ in a univariate prediction error.

As an immediate consequence of the GNZ formula and the Campbell formula, we obtain the following. If X has conditional intensity $\lambda = \lambda_{\theta_0} \in \Xi_\Theta$, where Ξ_Θ is of the form (1), then $E\{\mathcal{I}^h_\lambda(W;X)\} = 0$ for any test function $h : S \times \mathcal{X} \to \mathbb{R}$. Also, If X has intensity $\rho = \rho_{\theta_0} \in \Xi_\Theta$, where Ξ_Θ is of the form (2), then $E\{\mathcal{I}^h_\rho(W;X)\} = 0$ for any test function satisfying $h(\cdot;\mathbb{x}) = h(\cdot)$ for any $\mathbb{x} \in \mathcal{X}$. In the case of conditional intensities, variance and covariance expressions can be found in Baddeley et al. (2008); Daley & Vere-Jones (2008); Coeurjolly & Rubak (2013), while extension to the nth-order case is straightforward. These observations indicate that univariate prediction errors may sensibly be exploited as loss functions for parameter estimation. In particular, we may use $\mathcal{L}(\theta;\mathbb{x}) = \mathcal{I}^{h_\theta}_{\xi^n_\theta}(W;\mathbb{x})^2$, $\theta \in \Theta$, to obtain an estimate. As we shall see, a particularly interesting choice for $\mathcal{H}_\Theta$ is $h_\theta(\cdot;\mathbb{y}) = f\{\xi^n_\theta(\cdot;\mathbb{y})\}$, $\theta \in \Theta$, for some suitable $f : \mathbb{R} \to \mathbb{R}$. In fact, by using different test functions, univariate prediction errors yield many existing statistical approaches as particular cases, e.g. quasi-likelihood, Poisson process likelihood, Palm-likelihood, Takacs–Fiksel, pseudo-likelihood, non-parametric product density/intensity and K-function-based

minimum contrast estimation; see e.g. Diggle (2014); Møller & Waagepetersen (2017); Coeurjolly & Lavancier (2019) and the references therein.

4.3. *Properties of point process prediction errors*

Below, in Theorem 2, which is proved and stated in nth-order form in Section S5, we derive expectations for our prediction errors, together with necessary and sufficient conditions for the prediction errors to have mean zero.

THEOREM 2. *Given a point process X in S, let Z be an arbitrary thinning of X, with $Y = X \setminus Z$, and let $\breve{X}$ be the associated bivariate point process representation in Definition 1. Let further $\Xi_\Theta = \{\xi\}$ and $\mathcal{H}_\Theta = \{h\}$ consist of one element each.*

When $\xi, h : S \to \mathbb{R}$ are of the form (2), we have that $\mathcal{I}^h_\xi(\cdot; Z, Y) = \mathcal{I}^h_\xi(\cdot; Z)$ satisfies

$$E\{\mathcal{I}^h_\xi(A; Z)\} = \int_A h(u)\,\{\rho_Z(u) - \xi(u)\}\,\mathrm{d}u$$

for any $A \subseteq S$, where $\rho_Z(\cdot)$ denotes the intensity of Z. Moreover, this expectation is 0 for any $A \subseteq S$ and any test function h of the form (2) if and only if

$$\xi(u) \overset{a.e.}{=} \rho_Z(u). \tag{5}$$

If, instead, $\xi, h : S \times \mathcal{X} \to \mathbb{R}$ are of the form (1), when $\breve{X}$ admits a conditional intensity $\breve{\lambda}(\cdot; \breve{X})$, for any $A \subseteq S$ we have

$$E\{\mathcal{I}^h_\xi(A; Z, Y)\} = \int_A E\left[h(u; Y)\left\{\breve{\lambda}\{(u, 1); \breve{X}\} - \xi(u; Y)\right\}\right]\mathrm{d}u.$$

Assume further that $E[\breve{\lambda}\{(u, 1); \breve{X}\}^2] < \infty$ for $|\cdot|$-almost any $u \in S$. Then, for any $A \subseteq S$ and test function h such that $E\{h(u; Y)^2\} < \infty$, we have that $E\{\mathcal{I}^h_\xi(A; Z, Y)\} = 0$ if and only if

$$\xi(u; Y) \overset{a.e.}{=} E\left[\breve{\lambda}\{(u, 1); \breve{X}\} \mid Y\right]. \tag{6}$$

While Theorem 2 provides expressions for expecations of prediction errors, variance expressions can be found in (S21) and (S26). These indicate that the variances are governed by the dependence structure of $(Y, Z) = (X \setminus Z, Z)$ and the test function. Moreover, the general expressions in Theorem 2 are of limited practical use, but Corollary 1, which is proved and stated in nth-order form in Section S5, shows that they become explicit when Z is an independent thinning.

COROLLARY 1. *Assume the setting of Theorem 2. When Z is an independent thinning of X, based on a retention probability function $p(u) \in (0, 1)$, $u \in S$, then (5) reads $\xi(u) \overset{a.e.}{=} p(u)\rho_X(u)$ and the right-hand side of (6) is given by*

$$p(u)E\{\lambda_X(u; X) \mid Y\} = w(u, Z, Y)\lambda_X(u; Y), \tag{7}$$

where $w(u, Z, Y) = p(u)\lambda_X(u; Y)^{-1}E\{\lambda_X(u; X) \mid Y\}$; $w(u, Z, Y) \leq p(u)$ if X is repulsive, $w(u, Z, Y) \geq p(u)$ if X is attractive and $w(u, Z, Y) = p(u)$ if X is a Poisson process, a.s..

5. THE NEW STATISTICAL APPROACH

5.1. *Point process learning*

Given the definitions in Section 3.2 and 4.1, we can now specify our new statistical approach. The philosophical argument here is that a good approach should result in a model which, given the "current" (training) data, does well in predicting "unseen" (validation) data.

DEFINITION 5. *Generate training-validation pairs* $(\mathbb{x}_i^T, \mathbb{x}_i^V) \subseteq W^2 \subseteq S^2$, $i = 1, \ldots, k$, *from at least one realization of a point process* $X \subseteq S$, *in accordance with Section 3.2. Given either the form* (1) *or* (2), *consider further a general parametrized estimator family* $\Xi_\Theta = \{\xi_\theta : \theta \in \Theta\}$ *and* $k_i \geq 1$ *test function families* $\mathcal{H}_\Theta^{ij} = \{h_\theta^{ij} : \theta \in \Theta\}$, $j = 1, \ldots, k_i$, *for each* $i = 1, \ldots, k$. *Let*

$$\mathcal{I}_{ij}(\theta) = \mathcal{I}_{\xi_\theta}^{h_\theta^{ij}}(W; \mathbb{x}_i^V, \mathbb{x}_i^T) = \sum_{x \in \mathbb{x}_i^V} h_\theta^{ij}(x; \mathbb{x}_i^T) - \int_W h_\theta^{ij}(u; \mathbb{x}_i^T)\xi_\theta(u; \mathbb{x}_i^T)\mathrm{d}u \tag{8}$$

when Ξ_Θ *and* $\mathcal{H}_\Theta^{ij}$ *are of the form* (1), *or let*

$$\mathcal{I}_{ij}(\theta) = \mathcal{I}_{\xi_\theta}^{h_\theta^{ij}}(W; \mathbb{x}_i^T) = \sum_{x \in \mathbb{x}_i^T} h_\theta^{ij}(x) - \int_W h_\theta^{ij}(u)\xi_\theta(u)\mathrm{d}u \tag{9}$$

when Ξ_Θ *and* $\mathcal{H}_\Theta^{ij}$ *are of the form* (2). *We say that any method which generates estimates in* Θ *by exploiting* (8) *or* (9) *belongs to the field of point process learning.*

The use of (9) in fact results in a point process subsampling approach, very much akin to the one proposed in Moradi et al. (2019), in the sense that it does not make explicit use of $\mathbb{x}_i^V$, $i = 1, \ldots, k$, which is different from actual cross-validation-based approaches.

Once we have made a choice for $\Xi_\Theta = \{\xi_\theta : \theta \in \Theta\}$, which governs what we are interested in fitting, some further choices remain to be made: the cross-validation approach with associated parameters, how to combine the prediction errors to carry out the estimation and the test function families employed. These, as well as other choices, may be viewed as hyperparameter choices.

When we do not want a training-validation pair with $\mathbb{x}_i^V$ and/or $\mathbb{x}_i^T$ being empty to influence the estimation, we must require that $1 \leq \#\mathbb{x}_i^V$ and/or $1 \leq \#\mathbb{x}_i^T$. This may be achieved by replacing $\mathcal{I}_{ij}(\theta)$ with $\widetilde{\mathcal{I}}_{ij}(\theta) = I_i\mathcal{I}_{ij}(\theta)$, for a suitable indicator function I_i, which in the case of (8) may be absorbed into the test function. For (8), having $\mathbb{x}_i^T$ empty makes no sense, as this results in using $\mathbb{x}_i^T = \emptyset$ to predict $\mathbb{x}_i^V = \mathbb{x}$. Hence, in most cases, one would use $I_i = \mathbb{1}(1 \leq \#\mathbb{x}_i^T \leq \#\mathbb{x} - 1)$ if we consider (8) and $I_i = \mathbb{1}(1 \leq \#\mathbb{x}_i^T \leq \#\mathbb{x})$ if we consider (9), but sometimes the preferred choice may be to set $I_i = 1$ for all $i = 1, \ldots, k$. We write $\mathcal{T}_k = \{i \in \{1, \ldots, k\} : I_i = 1\}$. As usual, nth-order extensions are straightforward.

How to combine the prediction errors in Definition 5 depends on the statistical analysis undertaken, but, motivated by Theorem 2, the essential idea is that all of them should be close to 0. We emphasize that a prediction error does not necessarily attain the value 0; cf. the "leave-one-out"-discussion in Cronie & van Lieshout (2018). Considering $\widetilde{\mathcal{I}}_i(\theta) = (\widetilde{\mathcal{I}}_{i1}(\theta), \ldots, \widetilde{\mathcal{I}}_{ik_i}(\theta))^\top$ and $\mathcal{I}_i(\theta) = (\mathcal{I}_{i1}(\theta), \ldots, \mathcal{I}_{ik_i}(\theta))^\top$, $i = 1, \ldots, k$, we see a couple of natural choices:

1. Assuming that $k_i = k_0 \geq 1$, $i = 1, \ldots, k$, we generate a point estimate in Θ by minimizing a loss function $\mathcal{L}(\theta) = f_\mathcal{L}(\{\mathcal{I}_i(\theta) : i \in \mathcal{T}_k\})$, for some suitable $f_\mathcal{L} : (\mathbb{R}^{k_0})^{\#\mathcal{T}_k} \to \mathbb{R}^{k_0}$.
2. Denote the estimate resulting from minimizing $\theta \mapsto \mathcal{I}_{ij}(\theta)^2$, $\theta \in \Theta$, by $\widehat{\theta}_{ij} = \widehat{\theta}\{(\mathbb{x}_i^T, \mathbb{x}_i^V), p, W, \Xi_\Theta, \mathcal{H}_\Theta^{ij}\} \in \Theta$ for $i \in \mathcal{T}_k$, $j = 1, \ldots, k_i$. The sample median and mean of these estimates may serve as point estimates while empirical quantiles may serve as confidence/uncertainty regions for the true parameter $\theta_0 \in \Theta$. In addition, the mean/median of $\xi_{\widehat{\theta}_{ij}}$, $i \in \mathcal{T}_k$, $j = 1, \ldots, k_i$, may serve as a point estimate of ξ_{θ_0} (cf. Moradi et al. (2019)).

In Section 5.2, where we focus on point process learning under p-thinning-based cross-validation, we will restrict ourselves to the case where $\mathcal{H}_\Theta^{ij} = \mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}$ for all i and j, i.e. we use one single test function family for all training-validation pairs, whereby $\widetilde{\mathcal{I}}_i(\theta) = \widetilde{\mathcal{I}}_{i1}(\theta) =$

$I_i\mathcal{I}_{i1}(\theta) = I_i\mathcal{I}_i(\theta)$ and in the case of item (2) we have $\widehat{\theta}_i = \widehat{\theta}_{i1}$, $i \in \mathcal{T}_k$. Moreover, we consider the following choices for $f_{\mathcal{L}}$ in item (1):

$$\mathcal{L}_j(\theta) = \frac{1}{k}\sum_{i=1}^{k} |\widetilde{\mathcal{I}}_i(\theta)|^j \propto \frac{1}{\#\mathcal{T}_k}\sum_{i\in\mathcal{T}_k} |\mathcal{I}_i(\theta)|^j \qquad (j = 1, 2), \tag{10}$$

$$\mathcal{L}_3(\theta) = \left\{\frac{1}{k}\sum_{i=1}^{k} \widetilde{\mathcal{I}}_i(\theta)\right\}^2 \propto \left\{\frac{1}{\#\mathcal{T}_k}\sum_{i\in\mathcal{T}_k} \mathcal{I}_i(\theta)\right\}^2, \tag{11}$$

i.e. we find a parameter θ such that all prediction error terms are close to 0 in an average sense. We have observed that even if $\theta \mapsto |\mathcal{I}_i(\theta)|^j$ is unidentifiable, i.e. is flat over a region of Θ around the minimum, when averaging such functions as in (10), we generate loss functions with seemingly smaller flat regions, which implicitly mitigates the unidentifiability. Moreover, since positively weighted sums of convex functions and $x \mapsto |x|^j$, $j = 1, 2$, are convex, the form of $\mathcal{I}_i(\theta)$, which is governed by the test function, influences the convexity of e.g. (10). We further have $\mathcal{L}_1(\theta)^2 \le \mathcal{L}_2(\theta) \le k\mathcal{L}_3(\theta)$, by Hölder's and Jensen's inequalities, with equality when $k = 1$.

Concerning test function choices, motivated by Baddeley et al. (2005); Cronie & van Lieshout (2018), a simple recommendation is $h_\theta(\cdot) = f\{\xi_\theta(\cdot)\}$, where $f(x) = x^{-\gamma}$ and $\gamma = 1/2, 1$. We have indications that $\gamma = 1$, which conveniently sets the integrals in (8) and (9) to $|W|$ if ξ_θ is positive on W, yields the estimators with the lowest variances and mean (integrated) square errors of the two. In Section S3 we provide an in-depth discussion on hyperparameter choices, in particular test function choices, and we introduce an algorithm for data-driven hyperparameter selection. Moreover, in Section S6 we present results on consistency and asymptotic normality of the estimators generated by (10) under p-thinning-based cross-validation.

5.2. *Numerical evaluation: bandwidth selection using* p*-thinning-based cross-validation*

We next apply our new theory to the problem of optimal bandwidth selection (recall Section 2.2) and numerically compare it to the state of the art, which is represented by the approach of Cronie & van Lieshout (2018); by Section S2.1, this is an instance of auto-prediction. In addition, in Section S4 we apply our new bandwidth selection approach to the two datasets illustrated in Figure 1, i.e. a point pattern of tree locations on Barro Colorado Island, Panama, and a point pattern of spines on one branch of the dendritic tree of a rat neuron.

We consider the p-thinning-based cross-validation approaches in Definition 3 and combine the loss functions $\mathcal{L}_j(\theta)$, $j = 1, 2, 3$, in (10)-(11) with the prediction errors in (8) and the indicator $I_i = \mathbb{1}(1 \le \#\mathbb{x}_i^T \le \#\mathbb{x} - 1)$. We here let $\xi_\theta(u; \mathbb{x}_i^T) = w\lambda_\theta(u; \mathbb{x}_i^T) = w\widehat{\rho}_\theta(u; \mathbb{x}_i^T)$, where $w = p/(1-p)$ is an approximation of a parametrized form of the weight function $w(\cdot)$ in (7):

$$w_\theta(u, X_i^V, X_i^T) = \frac{E\{\widehat{\rho}_\theta(u; X)|X_i^T\}}{\widehat{\rho}_\theta(u; X_i^T)/p} = \frac{\widehat{\rho}_\theta(u; X_i^T) + E\{\widehat{\rho}_\theta(u; X_i^V) \mid X_i^T\}}{\widehat{\rho}_\theta(u; X_i^T)/p} \approx p + \frac{p^2}{1-p} = w,$$

since $\widehat{\rho}_\theta(u; X) = \widehat{\rho}_\theta(u; X_i^T) + \widehat{\rho}_\theta(u; X_i^V)$. Following our general recommendation, we further let $h_\theta(u, \mathbb{x}_i^T) = f\{w\widehat{\rho}_\theta(u; \mathbb{x}_i^T)\} = \{p\widehat{\rho}_\theta(u; \mathbb{x}_i^T)/(1-p)\}^{-\gamma}$, $\gamma = 1/2, 1$, whereby

$$\mathcal{I}_i(\theta) = \sum_{x\in\mathbb{x}_i^V} \left\{\frac{p\widehat{\rho}_\theta(x, \mathbb{x}_i^T)}{1-p}\right\}^{-\gamma} - \frac{p}{1-p}\int_W \left\{\frac{p\widehat{\rho}_\theta(u, \mathbb{x}_i^T)}{1-p}\right\}^{-\gamma} \widehat{\rho}_\theta(u, \mathbb{x}_i^T)\mathrm{d}u. \tag{12}$$

We first focus on $\gamma = 1$, which sets the integral in (12) to $|W|$. In both this and the state of the art approach, which is implemented in the R package spatstat (Baddeley et al., 2015), for a given realization $\mathbb{x}$ we let κ be the Gaussian kernel and $e_\theta(\cdot) \equiv 1$ in (3), and once the bandwidth

$\widehat{\theta}$ has been selected, the final intensity estimate $\widehat{\rho}_{\widehat{\theta}}$ is generated using local edge correction, i.e. $e_{\widehat{\theta}}(u,x) = \int_W \kappa_{\widehat{\theta}}(v-x)\mathrm{d}v$. We simulate 100 realizations ⁎ on $W = [0,1]^2$ and find estimates $\widehat{\rho}_{\widehat{\theta}}(u)$ of the true intensity $\rho(u)$, $u = (u_1,u_2) \in W$, for a collection of models (a subset of the models in Cronie & van Lieshout (2018)). For a given model and approach, we report IAB $= \int_W |\widehat{E}\{\widehat{\rho}_{\widehat{\theta}}(u,X)\} - \rho(u)|\mathrm{d}u$, ISB $= \int_W [\widehat{E}\{\widehat{\rho}_{\widehat{\theta}}(u,X)\} - \rho(u)]^2\mathrm{d}u$, IV $= \int_W \widehat{\mathrm{var}}\{\widehat{\rho}_{\widehat{\theta}}(u,X)\}\mathrm{d}u$ and MISE $=$ ISB $+$ IV, i.e. estimates of the integrated absolute bias, the integrated square bias, the integrated variance and the mean integrated square error.

The models we consider, which represent different kinds of spatial interaction (see Section S5.1), are the following. A log-Gaussian Cox process (aggregation) with random intensity $\Lambda(u) = \exp\{Z(u)\}$, where the Gaussian random field Z has mean function $(u_1,u_2) \mapsto 10 + 80u_1$ and covariance function $(u,v) \mapsto \sigma^2 \exp\{-r\|u-v\|_2\}$, $(\sigma^2, r) = (2\log 5, 50)$; here $\rho(u) = (10+80u_1)\exp(\sigma^2/2)$ and $E\{X(W)\} = 250$. A Poisson process (complete spatial randomness) with intensity function $\rho(u) = 10 + 480u_1$ and $E\{X(W)\} = 250$. A homogeneous determinantal point process (inhibition) with kernel $(u,v) \mapsto \sigma^2 \exp\{-r\|u-v\|_2\}$, $(\sigma^2, \beta) = (250, 50)$, independently thinned with retention probability $u \mapsto (10+80u_1)/90$ to obtain an inhomogeneous version with $\rho(u) = \sigma^2(10+80u_1)/90$ and $E\{X(W)\} \approx 138.9$.

The results for the Cronie & van Lieshout (2018) approach are as follows. For the log-Gaussian Cox process, IAB $= 19.48$, ISB $= 963.47$, IV $= 17597.99$ and MISE $= 18561.47$. For the Poisson process, IAB $= 15.80$, ISB $= 921.82$, IV $= 4408.21$ and MISE $= 5330.04$, while for the determinantal point process, IAB $= 9.14$, ISB $= 276.75$, IV $= 2002.55$ and MISE $= 2279.31$.

Turning to the new approach, with $\gamma = 1$, in Figure 2 and Figure 3 we present, respectively, the results for Monte-Carlo cross-validation, with $p = 0.1, 0.3, 0.5, 0.7, 0.9$ and $k = 400$, and multinomial cross-validation, with $k = 2, 3, \ldots, 10$, based on the loss functions $\mathcal{L}_j(\theta)$, $j = 1, 2, 3$.

Comparing Figure 2 and Figure 3 with the results for the state of the art, we see that, regardless of the choices of p, k and model, all point process learning approaches outperform the state of the art in terms of MISE. Although the state of the art performs slightly better in terms of bias, it performs comparatively poorly in terms of variance, which is consequently the reason for its higher MISE; it is worth emphasizing that it is precisely the lower variance which ensures that the Cronie & van Lieshout (2018) approach outperforms its predecessors (Moradi et al., 2019). We do, however, hypothesize that if $p \to 0$ in the Monte-Carlo cross-validation case (possibly in combination with $k \to \infty$), or $k \to \infty$ in the multinomial cross-validation case (e.g. in combination with $\mathcal{L}_3$), we would reach the same bias level as the state of the art, but still with a significantly lower MISE. We also see that the message in Figure S1, $\gamma = 1/2$, is the same in terms of performance with respect to the state of the art. In the Monte-Carlo cross-validation case, we further emphasize that increasing k beyond 100 essentially has little/no effect on the chosen performance measures, so our general suggestion is to fix $k \geq 100$. One alternative here is to sequentially increase k and stop once the loss function shows signs of convergence (guaranteed by the law of large numbers). We further conclude that multinomial cross-validation is the go-to method if computational costs are the main priority, whereas Monte-Carlo cross-validation is the go-to method if precision is prioritized; e.g., 2-fold multinomial cross-validation is roughly 400 times faster than Monte-Carlo cross-validation with $k = 400$ and $p = 0.5$. To exemplify, on a 2021 Apple MacBook with 64 gb ram and an M1 processor, selecting the bandwidth for one realization of the aggregated model in the 2-fold multinomial cross-validation case takes approximately 2.5 seconds (user/elapsed time) with our current implementation, versus 0.02 seconds for the spatstat implementation of the Cronie & van Lieshout (2018) approach; neither makes use of parallelization.

It seems that $\mathcal{L}_3$ favours a lower bias over a lower variance/MISE, whereas $\mathcal{L}_2$ favours the opposite and $\mathcal{L}_1$ seems to offer some middle-ground between the two. In Figure 2 we further see that in the case of Monte-Carlo cross-validation, $p \in [0.5, 0.7]$ tends to be a safe/good choice, which balances the trade-off between bias and variance, irrespectively of the degree of aggregation/inhibition of the underlying model. Moreover, it seems that the performance of multinomial cross-validation in terms of MISE is the best when $k = 2$ (see Figure 3), which is equivalent to Monte-Carlo cross-validation with $p = 0.5$, $k = 1$.

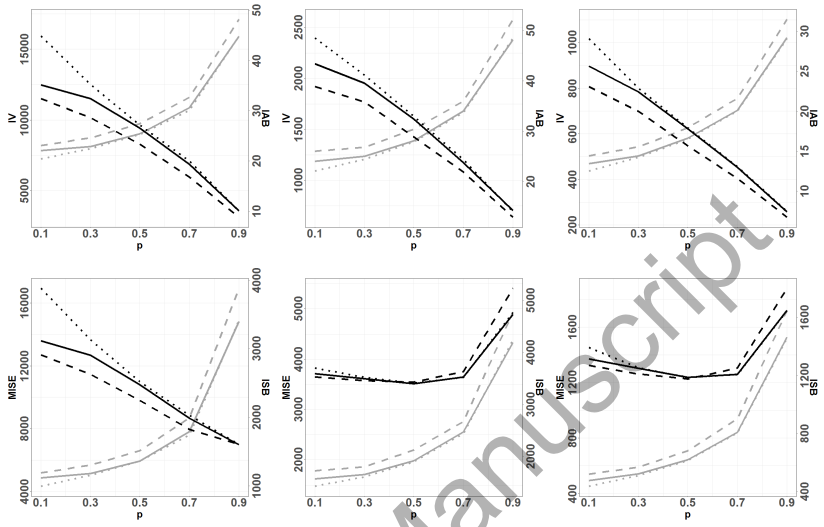

Fig. 2: Performance of the loss functions $\mathcal{L}_1$ (–), $\mathcal{L}_2$ (− − −) and $\mathcal{L}_3$ ($\cdots$), together with the test function $f(x) = 1/x$, using Monte-Carlo cross-validation with $p = 0.1, 0.3, 0.4, 0.7, 0.9$ and $k = 400$. Columns: log-Gaussian Cox process (left), Poisson process (middle) and determinantal point process (right). Top row: IAB (grey curve, right axis) and IV (black curve, left axis). Bottom row: ISB (grey curve, right axis) and MISE (black curve, left axis).

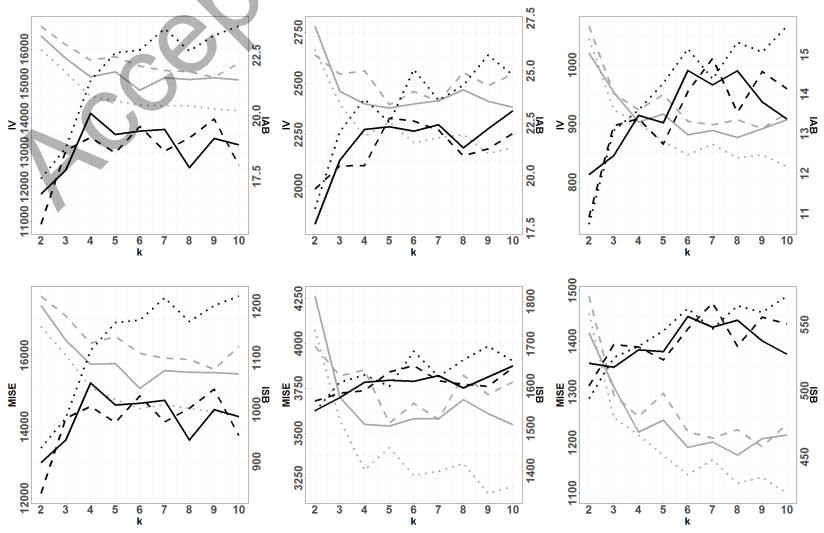

Fig. 3: The same structure as in Figure 2, using multinomial cross-validation with $k = 2, 3, \ldots, 10$.

To shed some light on the choice of test function, in Section S2.2 we also explore $\gamma = 1/2$. The overall conclusions regarding the hyperparameters p and k are the same as for $\gamma = 1$. Moreover, in Figure S1, which illustrates the same setup as in Figure 2 but with only the loss function $\mathcal{L}_2$, we see that the new approach performs much better than the state of the art in terms of MISE. Comparing the two test function choices, Figure 2 and Figure S1 reveal that $\gamma = 1/2$ reduces the bias with respect to $\gamma = 1$, but at the cost of increased variance, and thus also MISE. Moreover, in Section S3.2 we let our data-driven hyperparameter selection algorithm, Algorithm 1, select k in the case of multinomial cross-validation, which yields a performance essentially on par with our rule of thumb, $k = 2$. On the other hand, in the case of Monte-Carlo cross-validation, the performance essentially corresponds to keeping $p = 0.9$ fixed (for computational reasons we here fixed $k = 100$), which is suboptimal to our rule of thumb, i.e. $p \in [0.5, 0.7]$; it should be noted though that the gain in MISE from performing well with the log-Gaussian Cox process, which corresponds to fixing $p = 0.9$, is much bigger than the loss in MISE from performing relatively poorly with the other models. We thus see that there is merit to Algorithm 1, in particular since we in practice have no knowledge of the underlying process.

6. DISCUSSION

As our new approach outperforms the state of the art in non-parametric intensity estimation, framed as a conditional intensity estimation problem, it will likely have a central role in the future of point process statistics. We have indicated a few classical statistical approaches for parametric modelling of (conditional) intensities, which are based on univariate prediction errors/innovations. We believe that these could be improved by reframing them within our new framework.

ACKNOWLEDGEMENTS

The authors thank two referees and the editors for their constructive comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes additional background material, alternative cross-validation procedures, additional plots for the simulation study in Section S2.1, hyperparameter selection, higher-order statements and proofs of the results in the main text, examples of kernel intensity estimation for two datasets (in a Euclidean domain and on a linear network), and different asymptotic results.

REFERENCES

ARLOT, S. & CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* **4**, 40–79.

BADDELEY, A., MØLLER, J. & PAKES, A. G. (2008). Properties of residuals for spatial point processes. *Ann. Inst. Statist. Math.* **60**, 627–649.

BADDELEY, A., RUBAK, E. & TURNER, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC.

BADDELEY, A., TURNER, R., MØLLER, J. & HAZELTON, M. (2005). Residual analysis for spatial point processes. *J. R. Statist. Soc. B* **67**, 617–666.

BERMAN, M. & TURNER, T. R. (1992). Approximating point process likelihoods with glim. *J. R. Statist. Soc. C* **41**, 31–38.

CHIU, S. N., STOYAN, D., KENDALL, W. S. & MECKE, J. (2013). *Stochastic Geometry and its Applications*. John Wiley & Sons.

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 *Stochastic Geometry, Lecture Notes in Mathematics, vol 2237*, D. Coupier, ed. Springer, pp. 45–85.
COEURJOLLY, J.-F., MØLLER, J. & WAAGEPETERSEN, R. (2017). A tutorial on Palm distributions for spatial point processes. *Int. Stat. Rev.* **85**, 404–420.
COEURJOLLY, J.-F. & RUBAK, E. (2013). Fast covariance estimation for innovations computed from a spatial Gibbs point process. *Scand. J. Stat.* **40**, 669–684.
CRONIE, O., MORADI, M. & MATEU, J. (2020). Inhomogeneous higher-order summary statistics for point processes on linear networks. *Stat. Comput.* **30**, 1221–1239.
CRONIE, O. & SÄRKKÄ, A. (2011). Some edge correction methods for marked spatio-temporal point process models. *Comput. Statist. Data Anal.* **55**, 2209–2220.
CRONIE, O. & VAN LIESHOUT, M. N. M. (2016). Summary statistics for inhomogeneous marked point processes. *Ann. Inst. Statist. 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**, 455–462.
DALEY, D. J. & VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2nd ed.
DALEY, D. J. & VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer, 2nd ed.
DI MARZIO, M., PANZERA, A. & TAYLOR, C. C. (2014). Nonparametric regression for spherical data. *JASA* **109**, 748–763.
DIGGLE, P. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Taylor & Francis/CRC.
FIKSEL, T. (1984). Estimation of parameterized pair potentials of marked and non-marked gibbsian point processes. *Elektron. Inform. Kybernet.* **20**, 270–278.
GHORBANI, M., CRONIE, O., MATEU, J. & YU, J. (2020). Functional marked point processes: a natural structure to unify spatio-temporal frameworks and to analyse dependent functional data. *Test* , 1–40.
GUAN, Y., JALILIAN, A. & WAAGEPETERSEN, R. (2015). Quasi-likelihood for spatial point processes. *J. R. Statist. Soc. B* **77**, 677–697.
HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
KALLENBERG, O. (2017). *Random Measures, Theory and Applications*. Springer.
MATEU, J., MORADI, M. & CRONIE, O. (2020). Spatio-temporal point patterns on linear networks: Pseudo-separable intensity estimation. *Spat. Stat.* **37**.
MATTFELDT, T., HÄBEL, H. & FLEISCHER, F. (2013). Block bootstrap methods for the estimation of the intensity of a spatial point process with confidence bounds. *Journal of Microscopy* **251**, 84–98.
MCSWIGGAN, G., BADDELEY, A. & NAIR, G. (2017). Kernel density estimation on a linear network. *Scand. J. Stat.* **44**, 324–345.
MØLLER, J. & WAAGEPETERSEN, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. CRC.
MØLLER, J. & WAAGEPETERSEN, R. (2017). Some recent developments in statistics for spatial point patterns. *Annu. Rev. Stat. Appl.* **4**, 317–342.
MORADI, M., CRONIE, O., RUBAK, E., LACHIEZE-REY, R., MATEU, J. & BADDELEY, A. (2019). Resample-smoothing of Voronoi intensity estimators. *Stat. Comput.* **29**, 995–1010.
RAKSHIT, S., DAVIES, T., MORADI, M., MCSWIGGAN, G., NAIR, G., MATEU, J. & BADDELEY, A. (2019). Fast kernel smoothing of point patterns on a large network using two-dimensional convolution. *Int. Stat. Rev.* **87**, 531–556.
ROBERTS, D. R., BAHN, V., CIUTI, S., BOYCE, M. S., ELITH, J., GUILLERA-ARROITA, G., HAUENSTEIN, S., LAHOZ-MONFORT, J. J., SCHRÖDER, B., THUILLER, W. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929.
SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, vol. 26. Chapman & Hall/CRC.
TAKACS, R. (1986). Estimator for the pair-potential of a gibbsian point process. *Statistics* **17**, 429–433.
VAN LIESHOUT, M. N. M. (2000). *Markov Point Processes and Their Applications*. Imperial College Press.
VAN LIESHOUT, M. N. M. (2012). On estimation of the intensity function of a point process. *Methodol. Comput. Appl. Probab.* **14**, 567–578.
VAPNIK, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **10**, 988–999.
YANG, J., RAO, V. & NEVILLE, J. (2019). A Stein-Papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*.