Supplementary material for "A cross-validation-based statistical theory for point processes"

Ottmar Cronie

Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Gothenburg, Sweden; ottmar@chalmers.se, ottmar.cronie@gu.se

Mehdi Moradi

Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden.

Christophe A.N. Biscio

Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark.

SUMMARY

The following supplementary material document contains a discussion on alternative cross-validation approaches in Section S1, and it covers additional material on kernel intensity estimation, including additional results from the simulation study in the main text, in Section S2. Section S3 deals with hyperparameters, including a discussion on test functions and the definition of a data-driven hyperparameter selection algorithm. Section S4 presents kernel intensity estimation for two datasets while Section S5 presents higher-order statements and proofs of the results in the main text. Finally, in Section S6 we study large sample properties of point process learning.

S1. ALTERNATIVE CROSS-VALIDATION APPROACHES

The application of cross-validation procedures to point process statistics is not new; k-fold cross-validation, in particular leave-one-out cross-validation, has been applied in various statistical settings (Loader, 1999; Guan, 2007a,b; Hessellund et al., 2022). As indicated in Section 3.2, though, this is not an instance of independent thinning, whereby it is not immediate how such procedures could be appropriately incorporated into the proposed framework.

As a sort of mix between the cross-validation methods in Definition 3, one could consider letting $\mathbb{x}_i^V$ be a p-thinning of $\mathbb{x}$ with retention probability $p_i = i/k$, and letting $\mathbb{x}_i^T = \mathbb{x} \setminus \mathbb{x}_i^V$, $i = 1, \dots, k \geq 1$, whereby the validation sets would obtain an increasing expected number of points. On the one hand, the possible advantage over Monte-Carlo cross-validation is that we only have to choose the parameter k. On the other hand, it is not likely that it would perform better than an "optimally" chosen pair (k, p) for Monte-Carlo cross-validation. A further variant of Definition 3 is to consider an empirical Bayes-type cross-validation approach, where the considered retention probability would be estimated non-parametrically, e.g. by means of a scaled intensity estimate, using (a part of) the data. The issue here is that we do not actually employ independent thinning to generate the training and validation sets. It is further not clear whether there are any actual benefits of doing this, but this may be worth exploring further.

S2. SUPPLEMENTARY MATERIAL ON KERNEL INTENSITY ESTIMATION

S2.1. *State of the art in bandwidth selection*

We next give an account of two of the best-performing approaches for point processes in $\mathbb{R}^d$ in the literature. The Poisson process likelihood leave-one-out cross-validation approach (Loader, 1999; Baddeley et al., 2015) maximizes $\theta \mapsto \sum_{x \in \mathbb{x} \cap W} \log \widehat{\rho}_\theta(x, \mathbb{x} \setminus \{x\}) - \int_W \widehat{\rho}_\theta(u, \mathbb{x}) \mathrm{d}u$ in order to obtain an optimal bandwidth. To express this as a loss function $\mathcal{L}(\theta)$, we may either multiply it by -1 or consider the square of its derivative with respect to θ, assuming sufficient differentiability. Cronie & van Lieshout (2018) noted that this approach is particularly suited when data come from a Poisson process. In particular, the derivative of the Poisson process likelihood leave-one-out cross-validation function with respect to θ results in the univariate prediction error $\mathcal{I}^{h_\theta}_{\xi_\theta}(W; \mathbb{x})$, where $\xi_\theta = \widehat{\rho}_\theta$ and $h_\theta(x; \mathbb{x} \setminus \{x\}) = \widehat{\rho}_\theta(x, \mathbb{x} \setminus \{x\})^{-1} \partial \widehat{\rho}_\theta(x, \mathbb{x} \setminus \{x\}) / \partial \theta$. Motivated by the Campbell formula, Cronie & van Lieshout (2018) instead proposed to select the bandwidth by minimizing

$$\mathcal{L}(\theta) = \left[\sum_{x \in \mathbb{x} \cap W} h_\theta(x, \mathbb{x}) - \int_W f\{\widehat{\rho}_\theta(u, \mathbb{x})\} \widehat{\rho}_\theta(u, \mathbb{x}) \mathrm{d}u \right]^2, \tag{S1}$$

where $h_\theta(x, \mathbb{x}) = f\{\widehat{\rho}_\theta(x, \mathbb{x} \setminus \{x\})\}$ for some $f : \mathbb{R} \to \mathbb{R}$. We see that (S1) is the square of the univariate prediction error $\mathcal{I}^{h_\theta}_{\xi_\theta}(W; \mathbb{x})$, where $\xi_\theta = \widehat{\rho}_\theta$ and $h_\theta(x; \mathbb{x} \setminus \{x\}) = f\{\widehat{\rho}_\theta(x, \mathbb{x} \setminus \{x\})\}$. Hereby, Cronie & van Lieshout (2018), in fact, implicitly carried out Takacs–Fiksel estimation, where the non-parametric intensity estimator $\widehat{\rho}_\theta$, $\theta \in \Theta$, is treated as a parametrized conditional intensity. They further found that the choice $h_\theta(x, \mathbb{x}) = f\{\widehat{\rho}_\theta(x, \mathbb{x})\}$ with $f(x) = 1/x$ gives rise to (S1) being the square of a (conjectured) monotonic function of $\theta \geq 0$ when using $e_\theta(u, x) \equiv 1$ and a Gaussian kernel. They showed that this outperforms e.g. the Poisson process likelihood leave-one-out cross-validation approach and Moradi et al. (2019) further indicated that the choice $f(x) = 1/x$ promotes a low variance, in contrast to a low bias, which makes it particularly suited for aggregated point processes. Here, $f(x) = 1/x$ sets the integral in (S1) to $|W|$ when $\widehat{\rho}_\theta(u, \mathbb{x}) > 0$, $u \in W$, whereby the bandwidth is selected by estimating the (known) size of the study region with a sum of reciprocal intensity estimates.

S2.2. *Simulation study: additional results*

Consider the simulation study on bandwidth selection in Section 5.2. In Figure S1 below we illustrate the performance of point process learning when combining the loss function $\mathcal{L}_2$ and the prediction errors in (12) with $f(x) = x^{-\gamma}$, $\gamma = 1/2$, using MCCV with $p = 0.1, 0.3, 0.5, 0.7, 0.9$ and $k = 400$.

S3. HYPERPARAMETERS

Section 5.1 illustrates that point process learning involves a few choices to be made before the estimation can take place, e.g. how to combine the prediction errors (into a loss function), which test functions to employ and which cross-validation setup to use. These may all be viewed as hyperparameters to be chosen. Since our point process learning framework in general cannot be expressed through unbiased estimating equations, finding optimal hyperparameters similarly to Guan et al. (2015); Coeurjolly et al. (2016) seems unfeasible. A further idea is to apply calculus of variations to find a minimizer of the variances in Theorem 2. Aside from the potential associated intractability with such mathematically motivated approaches, the obtained optimality may be model specific. Hence, it seems one has to resort to either rules of thumb or some data-driven way to choose the hyperparameters.

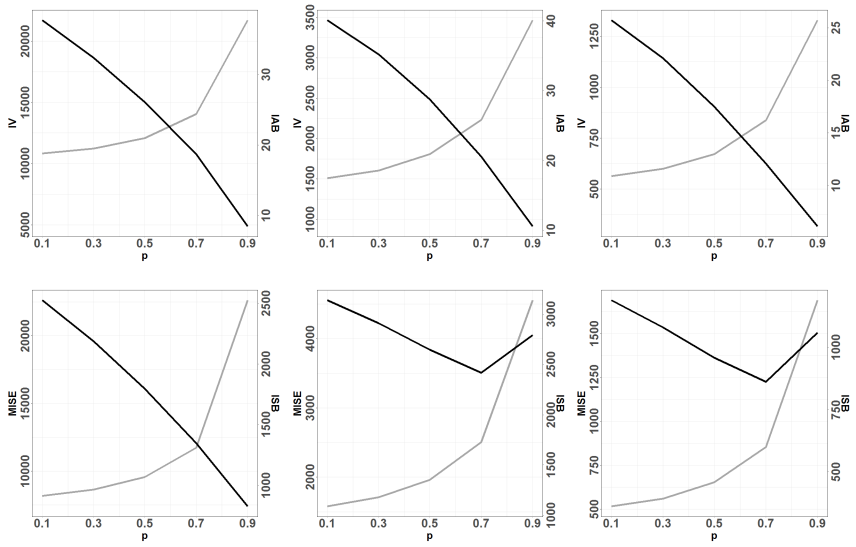

Fig. S1: Performance of the loss function $\mathcal{L}_2$, using Monte-Carlo cross-validation with $p = 0.1, 0.3, 0.4, 0.7, 0.9$ and $k = 400$ together with the test function $f(x) = x^{-0.5}$. Columns: log-Gaussian Cox process (left), Poisson process (middle) and determinantal point process (right). 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).

For the kernel intensity estimation, in Section 5.2 we explore different rules of thumb, showing that there exist specific hyperparameter choices which yield (substantially) better estimates than both the state of the art and other potential hyperparameter choices. We see e.g. that the chosen test function and cross-validation setup impact the quality of the obtained estimates. Although we believe these rules of thumb to be generally applicable, there is of course a risk that they do not perform equally well beyond the scope of the models we consider. In Section S3.2 we introduce an algorithm for data-driven hyperparameter selection and numerically evaluate it in the context of kernel intensity estimation, where it successfully manages to select the cross-validation parameters. This should be seen in light of mathematically derived, and likely model-dependent, optimal hyperparameter choices.

S3.1. *Test functions*

The literature offers a few suggestions on suitable test functions (recall Section 4.2). Most notably, when ξ_θ is differentiable in θ, in the univariate setting of (9), the test function $h_\theta^{ij}(\cdot) = \partial\xi_\theta(\cdot)/\partial\theta = \nabla\xi_\theta(\cdot)$ turns $\theta \mapsto \mathcal{I}_{ij}(\theta)$ into a Poisson process likelihood score-type function. A further group of test function candidates encountered in the literature may be summarized as $h_\theta^{ij}(\cdot) = f(\xi_\theta(\cdot))$, where $f(x) = x^{-\gamma}$, $\gamma \in \mathbb{R}$ (Baddeley et al., 2005; Cronie & van Lieshout, 2018). In the innovations setting, $\gamma = 0$ corresponds to so-called raw innovations, $\gamma = 1/2$ to Pearson innovations and $\gamma = 1$ to Stoyan–Grabarnik/inverse innovations. In particular, when $\gamma = 1$, (8) becomes $\mathcal{I}_{ij}(\theta) = \sum_{x \in \mathbb{x}_i^V \cap W} \xi_\theta(x; \mathbb{x}_i^T)^{-1} - |\bar{W}|$. The size of the support $\bar{W} = \{u \in W : \xi_\theta(u; \mathbb{x}_i^T) > 0\} \subseteq W$, which may vary depending on $\mathbb{x}_i^T$, is (approximately) $|W|$ if $\xi_\theta(\cdot; \mathbb{x}_i^T)$ is strictly positive on (most of) W. This is convenient from a computational point of view, since we do not have to compute the integral in (8). Further, setting $h_\theta^{ij}(\cdot) = f\{\xi_\theta^n(\cdot)\}$, where $f(x) = -x\log(x)$, results in entropy-type prediction errors; this is partly motivated by the connection

between entropy and Kullback-Leibler divergence (Daley & Vere-Jones, 2008). Finally, if we use h_θ^{ij} divided by the integral of $h_\theta^{ij}(\cdot)\xi_\theta^n(\cdot)$ as test function, we have e.g. that (8) becomes $\mathcal{I}_{ij}(\theta) = \sum_{x\in\aleph_i^V\cap W}\xi_\theta^n(x;\aleph_i^T)^{-1}\bar{f}_\theta^{ij}(x;\aleph_i^T) - 1$, where $\bar{f}_\theta^{ij}(\cdot)$ is a density function on W. Viewing the first part of the summand as a quadrature weight for the quadrature point x, this sum may be viewed as an approximation of the integral of $\bar{f}_\theta^{ij}$, and minimization of $\theta \mapsto \mathcal{I}_{ij}(\theta)$ as a density estimation problem.

S3.2. *Data-driven hyperparameter selection*

Our proposed approach to hyperparameter selection, which is found in Algorithm 1 below, is motivated by a commonly encountered cross-validation-based algorithm in the statistical learning literature (James et al., 2013). We emphasize that Algorithm 1 only deals with the setting where the estimation is carried out by minimizing a loss function; recall Section 5.1.

Algorithm 1. Hyperparameter selection

1 Let $k_E \geq 1$ and choose a cross-validation method, with associated parameters (e.g. $p_E = 1/k_E$ for multinomial cross-validation), to generate k_E cross-validation splits. Refer to the corresponding validation sets $\aleph_j^E = \aleph_j^V$ as test sets and denote the corresponding training sets $\aleph_j^T = \aleph \setminus \aleph_j^E$ by $\aleph_j$, $j = 1, \ldots, k_E$.
2 Specify a space Θ_γ of permissible hyperparameter choices.
3 **if** *the estimation is based on* (8)*,* **then**
4 specify a general loss function $\mathcal{L}_\gamma : \Theta \times (\mathcal{X}\times\mathcal{X})^\infty \to [0,\infty)$, $\gamma \in \Theta_\gamma$, and a goodness of fit/prediction accuracy mapping, $\mathcal{G}$, which is defined on $\Theta \times \Theta_\gamma \times \mathcal{X} \times \mathcal{X}^\infty$, where small means better predictive performance.
5 **if** *the estimation is based on* (9)*,* **then**
6 specify a general loss function $\widetilde{\mathcal{L}}_\gamma : \Theta \times \mathcal{X}^\infty \to [0,\infty)$, $\gamma \in \Theta_\gamma$, and a goodness of fit/prediction accuracy mapping, $\mathcal{G}$, which is defined on $\Theta \times \Theta_\gamma \times \mathcal{X}$, where small means better predictive performance.
7 **for** $j = 1, \ldots, k_E$ **do**
8 **for** $\gamma \in \Theta_\gamma$ **do**
9 Generate $k = k(\gamma)$ cross-validation splittings $\{(\aleph_{ij}^T, \aleph_{ij}^V)\}_{i=1}^k$, based on $\aleph_j$ and the cross-validation method corresponding to γ; if the cross-validation parameters are not hyperparameters, this may be done only once directly after step 7.
10 **if** *the estimation is based on* (8)*,* **then**
11 Find $\widehat{\theta}_j(\gamma) \in \Theta$ by minimising $\mathcal{L}_\gamma(\theta; \{(\aleph_{ij}^T, \aleph_{ij}^V)\}_{i=1}^k)$, $\theta \in \Theta$, and let $\mathcal{G}_j(\gamma) = \mathcal{G}[\widehat{\theta}_j(\gamma), \gamma, \aleph_j^E, \{(\aleph_{ij}^T, \aleph_{ij}^V)\}_{i=1}^k]$.
12 **if** *the estimation is based on* (9)*,* **then**
13 Find $\widehat{\theta}_j(\gamma) \in \Theta$ by minimising $\widetilde{\mathcal{L}}_\gamma(\theta; \{\aleph_{ij}^T\}_{i=1}^k)$, $\theta \in \Theta$, and let $\mathcal{G}_j(\gamma) = \mathcal{G}(\widehat{\theta}_j(\gamma), \gamma, \aleph_j^E)$.
14 Given some suitable mapping M, define $\bar{M}(\gamma) = M(\mathcal{G}_1(\gamma), \ldots, \mathcal{G}_{k_E}(\gamma)) \geq 0$, $\gamma \in \Theta_\gamma$, and find a minimizer $\widehat{\gamma}$ of $\bar{M}(\gamma)$, $\gamma \in \Theta_\gamma$.
15 Carry out the final estimation based on the full dataset, i.e. without holding out any test sets, employing the loss function $\mathcal{L}_{\widehat{\gamma}}(\cdot)$ when using (8) and $\widetilde{\mathcal{L}}_{\widehat{\gamma}}$ when using (9).

The subjective choices in Algorithm 1, which have to be fixed a priori (according to some best-practice principle/rule of thumb), are k_E (and p_E), i.e. parameters related to the test set

generation, as well as the forms of the mappings $\mathcal{G}$ and M. The former measures how well a particular hyperparameter choice results in a good prediction of a particular test set, and the latter specifies how all such test set-based measures should be combined. Natural choices for M include $M(x_1,\dots,x_{k_E}) = -(x_1+\cdots+x_{k_E})/k_E$, $M(x_1,\dots,x_{k_E}) = -\operatorname{med}\{x_1,\dots,x_{k_E}\}$ and $M(x_1,\dots,x_{k_E}) = -\min\{x_1,\dots,x_{k_E}\}$. The choice of $\mathcal{G}$, however, is a bit more delicate since it is specifically $\mathcal{G}$ which quantifies how well a given hyperparameter choice actually performs.

It should be noted that penalization, e.g. regularization, may be achieved by adding the penalty in question to the general loss function, $\mathcal{L}_\gamma$ or $\widetilde{\mathcal{L}}_\gamma$, and the penalization parameter, $\widetilde{\gamma} \geq 0$, would be included in the hyperparameter vector. In classical statistical learning, the standard cross-validation algorithm (James et al., 2013) is commonly used to select $\widetilde{\gamma}$ as well as any parameters included in the actual penalty, and we believe that Algorithm 1 may fulfil the same purpose in the current context. For instance, in the case of elastic net regularization (Zou & Hastie, 2005), where the penalty $R(\theta;\alpha)$ itself has a parameter $\alpha \in [0,1]$, which governs how much ridge penalization versus lasso penalization is imposed, we add $\widetilde{\gamma}R(\theta;\alpha)$ to the general loss function and include the pair $(\widetilde{\gamma},\alpha) \in \Theta_{\widetilde{\gamma}} \times \Theta_\alpha \subseteq [0,\infty) \times [0,1]$ in the hyperparameter vector γ. In particular, in intensity estimation also other types of penalization, e.g. smoothness, may be of interest.

To make Algorithm 1 and its associated choices a bit more concrete, we next illustrate how it may be used in kernel estimation; recall that the estimation is based on (8) here. Some of the choices we can make here include:

- In the case of Monte-Carlo cross-validation, we have the sets Θ_p and Θ_k of permissible choices for p and k, respectively. E.g., we may want to evaluate $p \in \Theta_p = \{0.1,\dots,0.9\}$ and $k \in \Theta_k = \{100,\dots,400\}$. In the case of multinomial cross-validation, we would instead only consider Θ_k, e.g. $\Theta_k = \{2,5,10,20,30\}$.
- Choices for the parametrization of the different test functions may be considered, e.g. $h_\theta(\cdot) = f\{\xi^1_\theta(\cdot)\} = f\{\widehat{\rho}_\theta(\cdot)p/(1-p)\}$, where $f(x) = x^{-\beta}$ and $\beta \in \Theta_\beta = \{0,1/4,1/2,3/4,1\}$.
- Choices for the forms of the general loss functions in steps 3 and 5 may be considered. E.g., we may want to evaluate which of the loss functions $\mathcal{L}_i$, $i = 1,2,3$, in (10)-(11) performs the best, which we parameterize according to $i \in \Theta_{\mathcal{L}} = \{1,2,3\}$. This pertains to Step 3 in Algorithm 1.
- The choice of kernel may be treated as a hyperparameter. To exemplify, consider the family of beta kernels κ_ϕ, $\phi \geq 0$, which are also known as multi-weight kernels (Hall et al., 2004). The box kernel is obtained by setting $\phi = 0$, the Epanechnikov kernel by setting $\phi = 1$ and the Gaussian kernel may be viewed as a degenerate limit case (having applied proper scaling), which we represent by $\phi = \infty$; see e.g. Cronie & van Lieshout (2018) for details. Comparing these three special cases may then be represented by the hyperparameter choice $\phi \in \Theta_\phi = \{0,1,\infty\}$.
- Choices of edge correction methods may be treated as a hyperparameter. The three most common edge correction methods are $e_\theta(u,x) \equiv 1$ (no edge correction), $e_\theta(u,x) = \int_W \kappa_\theta(v-x)\mathrm{d}v$ (local edge correction), $e_\theta(u,x) = \int_W \kappa_\theta(v-u)\mathrm{d}v$ (global edge correction), which we could parameterize by $e=0$, $e=1$ and $e=2$, respectively, i.e. $e \in \Theta_e = \{0,1,2\}$.

Depending on what we would like to include in our hyperparameter vector, we would thus let Θ_γ be the product space generated by a combination of the spaces above. We could thus in principle let $\gamma = (p,k,\beta,i) \in \Theta_\gamma = \Theta_p \times \Theta_k \times \Theta_\beta \times \Theta_{\mathcal{L}}$, using a Gaussian kernel with no edge correction in the bandwidth selection (but local edge correction in the final intensity estimate).

We emphasize, however, that we cannot guarantee that this would work well in practice. Turning to the choice of $\mathcal{G}$ and $\bar{M}$, we believe that

$$\bar{M}(\gamma) = \sum_{j=1}^{k_E} \mathcal{G}_j(\gamma)/k_E = \sum_{j=1}^{k_E} \mathcal{G}[\widehat{\theta}_j(\gamma), \gamma, \aleph_j^E, \{(\aleph_{ij}^T, \aleph_{ij}^V)\}_{i=1}^k]/k_E,$$

$$\mathcal{G}_j(\gamma) = \mathcal{I}^{h_{\widehat{\theta}_j(\gamma)}}_{\xi_{\widehat{\theta}_j(\gamma)}}(W; \aleph_j^E) = -\sum_{x \in \aleph_j^E} \widehat{\rho}_{\widehat{\theta}_j(\gamma)}(x; \aleph_j^E) \log\{\widehat{\rho}_{\widehat{\theta}_j(\gamma)}(x; \aleph_j^E)\} + \int_W \widehat{\rho}_{\widehat{\theta}_j(\gamma)}(u; \aleph_j^E)^2 \log\{\widehat{\rho}_{\widehat{\theta}_j(\gamma)}(u; \aleph_j^E)\} \mathrm{d}u, \tag{S2}$$

should make sense here. This means that we let $\mathcal{G}$ be a prediction error based on the entropy-motivated test function mentioned in Section S3.1, whereby we essentially would consider residuals in the sense of Baddeley et al. (2005).

We next evaluate Algorithm 1 numerically in the context of kernel intensity estimation. More precisely, we repeat the experiment in Section 5.2, for the exact same realizations for each model, but for each realization we run Algorithm 1 to choose i) the cross-validation parameter $k \in \Theta_k = \{2, 5, 10, 20, 30\}$ for multinomial cross-validation, and ii) the cross-validation parameter $p \in \Theta_p = \{0.1, \ldots, 0.9\}$ for Monte-Carlo cross-validation; in the Monte-Carlo cross-validation case, we fix $k = 100$ (recall the discussion in Section 5.2 about $k = 100$ being sufficiently large). To create the test sets, we consider multinomial cross-validation and let $k_E = 2, 5, 10, 20$ to see if some general recommendation can be given for k_E. We further restrict ourselves to the loss function $\mathcal{L}_2$ and the test function $h_\theta(\cdot) = f\{\xi_\theta^1(\cdot)\} = \{\widehat{\rho}_\theta(\cdot)p/(1-p)\}^{-1}$, and let $\mathcal{G}$ and $\bar{M}$ be as in (S2). Based on the results, which can be found in Table S1, we conclude that when multinomial cross-validation is used, in terms of keeping MISE low for all three models, the recommendation is to set k_E to 5, 10 or 20; arguably, $k_E = 20$ performs slightly better. The results are essentially equivalent to what we obtained when we fixed $k = 2$; recall Figure 3. In the case of Monte-Carlo cross-validation, we recommend to set $k_E = 2$, which results in a performance slightly worse than fixing p to our rule of thumb in Section 5.2, i.e. $p \in [0.5, 0.7]$. Although, as expected, we do not obtain as good results as when adhering to the rules of thumb, we here have the benefit that the algorithm tends to adapt to the kind of process that has generated the data, which of course is practically useful since we rarely/never know the true data-generating process. Moreover, these observations indicate that common recommendations about the number of folds to use in classical k-fold cross-validation, typically $k = 5$ or $k = 10$, do not necessarily apply in the current context. Here, the computation times scale with a factor k_E with respect to the computation times provided in Section 5.2 of the main text, without parallelization.

S4. DATA ANALYSIS

Point pattern data arise in various applications and fields. A few common examples of point patterns include collections of astronomical objects (Babu & Feigelson, 1996; Kerscher, 2000), climatic events (Toreti et al., 2019), crimes (Ang et al., 2012; Moradi et al., 2018; Chaudhuri et al., 2021), disease cases (Meyer et al., 2012; Diggle, 2014), earthquakes (Ogata, 1998; Marsan & Lengline, 2008; Iftimi et al., 2019), farms (Chaiban et al., 2019), queuing events (Brémaud, 1981; Baccelli & Brémaud, 2013), traffic accidents (Rakshit et al., 2019; Moradi & Mateu, 2020; Moradi et al., 2020), trees (forestry) (Stoyan & Penttinen, 2000; Cronie et al., 2013) and geology (Dehghani & Vahidi-Asl, 2019).

Table S1: Algorithm 1 for selecting the cross-validation parameters in the simulation study in Section 5.2. For multinomial cross-validation we consider $k \in \Theta_k = \{2, 5, 10, 20, 30\}$ and for Monte-Carlo cross-validation we consider $p \in \Theta_p = \{0.1, \dots, 0.9\}$ with $k = 100$ fixed. Throughout, we consider the loss function $\mathcal{L}_2$ in (10) and the test function $h_\theta(\cdot) = f\{\xi^1_\theta(\cdot)\} = \{\widehat{\rho}_\theta(\cdot)p/(1-p)\}^{-1}$, and let $\mathcal{G}$ and $\bar{M}$ be as in (S2). To generate the test sets, we consider multinomial cross-validation with k_E folds. The table also includes the Cronie & van Lieshout (2018) approach (CvL) results from Section 5.2.

	k_E	2	5	10	20	CvL
Log-Gaussian Cox	IAB	24.76	25.85	25.86	26.08	19.48
Multinomial	ISB	1357.32	1429.57	1428.04	1455.58	963.47
	IV	12294.92	11228.12	11288.70	11214.43	17597.99
	MISE	13652.25	12657.69	12716.74	12670.01	18561.47
Log-Gaussian Cox	IAB	42.42	48.97	48.97	48.97	
Monte-Carlo	ISB	3121.58	4008.46	4008.46	4008.46	
	IV	5110.25	3603.55	3603.55	3603.55	
	MISE	8231.83	7612.01	7612.01	7612.01	
Poisson	IAB	26.46	27.67	28.23	28.43	15.80
Multinomial	ISB	1762.95	1860.62	1906.46	1927.86	921.82
	IV	1985.08	1958.83	1801.30	1779.76	4408.21
	MISE	3748.03	3819.45	3707.77	3707.62	5330.04
Poisson	IAB	38.45	51.00	51.00	51.00	
Monte-Carlo	ISB	2936.40	4588.65	4588.65	4588.65	
	IV	1374.09	677.67	677.67	677.67	
	MISE	4310.48	5266.32	5266.32	5266.32	
Determinantal	IAB	14.87	15.27	15.57	15.81	9.14
Multinomial	ISB	534.52	550.51	567.12	575.92	276.75
	IV	807.81	765.58	757.18	741.90	2002.55
	MISE	1342.34	1316.10	1324.30	1317.81	2279.31
Determinantal	IAB	24.70	32.35	32.35	32.35	
Monte-Carlo	ISB	1109.54	1709.93	1709.93	1709.93	
	IV	584.93	253.21	253.21	253.21	
	MISE	1694.47	1963.14	1963.14	1963.14	

To illustrate the applicability of our methodology on data in different general spaces, we next carry out kernel intensity estimation for the two datasets in Figure 1, where the first one lives in a Euclidean domain $W \subseteq \mathbb{R}^2$ and the second one lives on a linear network $S = W = L = \bigcup_{i=1}^k l_i$. Both datasets can be downloaded through the R package spatstat (Baddeley et al., 2015). In both cases, we employ Monte-Carlo cross-validation with $(k, p) = (400, 0.7)$ and $\mathcal{L}_2$. Moreover, in analogy with Section 5.2, we let the test function be given by $h_\theta(u, \mathbb{x}_i^T) = \{p\widehat{\rho}_\theta(u; \mathbb{x}_i^T)/(1-p)\}^{-1}$, which yields the prediction errors

$$\mathcal{I}_i(\theta) = \frac{1-p}{p} \sum_{x \in \mathbb{x}_i^V} \frac{1}{\widehat{\rho}_\theta(x, \mathbb{x}_i^T)} - |W|.$$

In analogy with Section 5.2, we use no edge correction in $\widehat{\rho}_\theta$ when we select the bandwidth, but use it for the generation of the final intensity estimate $\widehat{\rho}_{\widehat{\theta}}(u, \mathbb{x})$, $u \in W$.

The first dataset (see the left panel of Figure 1) is a point pattern of 3605 Beilschmiedia pendula tree locations on Barro Colorado Island, Panama. Figure S2 shows the dataset $\aleph \subseteq W \subseteq \mathbb{R}^2$ together with a kernel intensity estimate (3), obtained through our bandwidth selection approach. Visibly, the obtained bandwidth, 56.65 (metres), leads to an estimate which adapts well to the inhomogeneity of the events.

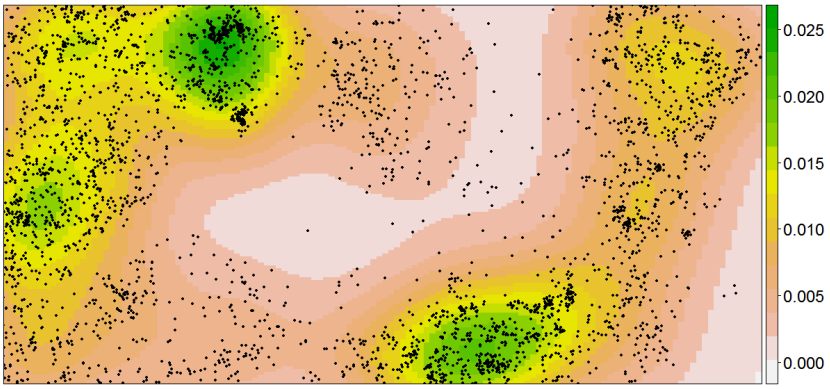

Fig. S2: Locations of tropical rain forest trees on Barro Colorado Island, Panama, with an obtained kernel intensity estimate overlaid.

Our second dataset $\aleph \subseteq S = W = L$ (see the right panel of Figure 1) consists of the locations of 566 spines on one branch of the dendritic tree of a rat neuron, i.e. nerve cell (Jammalamadaka et al., 2013; Baddeley et al., 2014); courtesy of the Kosik Lab, UC Santa Barbara. Here, spines refer to small protrusions on the dendrites, which are branching filaments which extend from the main body of a neuron. The linear network $L = \bigcup_{i=1}^{n} l_i$, which consists of 640 vertices (maximum vertex degree 4) and 639 line segments $l_i = [u_i, v_i] = \{tu_i + (1-t)v_i : 0 \leq t \leq 1\} \subseteq \mathbb{R}^2$, has a total length of 1933.653 microns. To obtain an intensity estimate, we consider the kernel intensity estimator proposed by Rakshit et al. (2019). More specifically, we employ the following variant of (3) in the linear network setting:

$$\widehat{\rho}_\theta(u, \aleph) = \sum_{x \in \aleph} \frac{\kappa_\theta(u-x)}{e_\theta(u,x)}, \qquad u \in L, \tag{S3}$$

where the edge correction term is given by e.g.

$$e_\theta(u,x) = e_\theta(x) = \int_L \kappa_\theta(x-v)\mathrm{d}v$$

and we recall that the reference measure $|A| = \int_A \mathrm{d}u$, $A \subseteq L$, corresponds to integration with respect to arc length, i.e. the 1-dimensional Hausdorff measure on L. Given that Moradi et al. (2018); Moradi (2018) obtained good results with the bandwidth selection approach of Cronie & van Lieshout (2018) in the linear network setting, and given the results in Section 5.2, we have indications that our new approach should also do a good job here. We obtain the bandwidth 24.83 (microns) and in Figure S3 we see that the lower sub-branch has an intensity which is almost double of what on average is seen on the remaining network. Visually, this corresponds well with the data.

Fig. S3: A point pattern of locations of spines on one branch of the dendritic tree of a rat neuron, with an obtained kernel intensity estimate overlaid.

S5. Higher-order statements and proofs of the results in the main text

S5.1. *Preliminaries*

The main text focused on first-order characteristics. However, in the sections below, we state and prove the results in the main text in the nth-order setting. Consequently, we first provide an overview of the nth-order setting.

Given a suitable probability space $(\Omega, \mathcal{F}, \mathrm{pr})$, we recall from Section 2.1 that a point process $X = \{x_i\}_{i=1}^{N}$, $0 \leq N \leq \infty$, in a general space 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 $\aleph = \{x_1, \ldots, x_n\} \subseteq S$, $0 \leq n \leq \infty$, such that $\#(\aleph \cap A) = \sum_{i=1}^{n} \mathbb{1}(x_i \in A) < \infty$ for any bounded (Borel set) $A \subseteq S$. The σ-algebra $\mathcal{N}$, which is generated by the cardinality mappings $\aleph \mapsto \#(\aleph \cap A) \in \{0, 1, \ldots, \infty\}$, $A \subseteq S$, $\aleph \in \mathcal{X}$, coincides with the Borel σ-algebra generated by a certain metric for measures on S, and X can be identified with the (discrete) random measure $X(A) = \#(X \cap A)$, $A \subseteq S$ (Daley & Vere-Jones, 2003, 2008). The distribution $P(E) = P_X(E) = \mathrm{pr}(X \in E)$, $E \in \mathcal{N}$, is governed by the finite dimensional distributions of X (van Lieshout, 2000), which for a finite point process ($N < \infty$ a.s.) are determined by its so-called Janossy densities; see the discussion around (S11) for details.

Most of the relevant distributional characteristics considered in the literature can be obtained through (combinations of) expectations of the form

$$E\left\{ \sum_{x_1,\ldots,x_n \in X}^{\neq} h(x_1, \ldots, x_n, X \setminus \{x_1, \ldots, x_n\}) \right\} = E\left\{ \sum_{x \in X_{\neq}^n} h(x, X \setminus \{x\}) \right\}, \tag{S4}$$

for $h : S^n \times \mathcal{X} \to \mathbb{R}$ which are permutation invariant over S^n, $n \geq 1$; unless h is non-negative (and possibly infinite), it is assumed to be integrable. Here, $X_{\neq}^n = \{(x_1, \ldots, x_n) \in X^n : x_i \neq x_j \text{ if } i \neq j\} \subseteq S^n$ is the point process consisting of all distinct n-tuples of elements of X (Daley & Vere-Jones, 2008); e.g., $(x_1, x_2), (x_2, x_1) \in X_{\neq}^2$ if $x_1, x_2 \in X$. Below we consider different

subclasses of functions h, which yield different integral identities for (S4) and, in turn, define different point process characteristics of interest.

The subclass of functions h in (S4) which are constant over $\mathcal{X}$, i.e. of the form $h(x_1, \dots, x_n)$, defines the nth-order product density $\rho^{(n)}$ of X through the Campbell formula/theorem (Daley & Vere-Jones, 2008, Section 9.5), which states that (S4) equals

$$\int_{S^n} h(u_1, \dots, u_n)\rho^{(n)}(u_1, \dots, u_n)\mathrm{d}u_1 \cdots \mathrm{d}u_n. \tag{S5}$$

We have that $\rho^{(n)}(\cdot)$, which is the Radon-Nikodym derivative of the nth-order factorial moment measure $(A_1 \times \cdots \times A_n) \mapsto E\{\sum^{\neq}_{x_1,\dots,x_n \in X} \prod_{i=1}^{n} \mathbb{1}(u_i \in A_i)\}$, $A_i \subseteq S$, with respect to the product measure $|\cdot|^n$, coincides with the first-order product density of the point process $X^n_{\neq}$ (Daley & Vere-Jones, 2008). Heuristically, since X is simple, for disjoint infinitesimal neighbourhoods $A_i = du_i$, $\mathrm{d}u_i = |du_i|$, of the points $u_i \in S$, $i = 1, \dots, n$, we obtain that $\mathrm{pr}\{X(du_1) = 1, \dots, X(du_n) = 1\} = E\{X(du_1) \cdots X(du_n)\} = \rho^{(n)}(u_1, \dots, u_n)\mathrm{d}u_1 \cdots \mathrm{d}u_n$. Here, the particular case $n = 1$ gives us the intensity function $\rho = \rho^{(1)}$ of X; see Section 2.1 for details. It is worth noting that the correlation functions, $g^{(n)}(u_1, \dots, u_n) = \rho^{(n)}(u_1, \dots, u_n)/(\rho(u_1) \cdots \rho(u_n))$, $n \geq 1$, quantify interaction (van Lieshout, 2011): when $g^{(n)}(u_1, \dots, u_n)$ is larger than 1 we speak of clustering/aggregation between points of X in the (infinitesimal) vicinity of $u_1, \dots, u_n \in S$, while we speak of inhibition/regularity/repulsion when it is less than 1; for a Poisson process, which represents complete spatial randomness (Diggle, 2014), we have $g^{(n)}(\cdot) \equiv 1$ for any $n \geq 1$.

When h is not necessarily constant over $\mathcal{X}$, (S4) equals (Daley & Vere-Jones, 2008)

$$\int_{S^n \times \mathcal{X}} h(u_1, \dots, u_n, \aleph)\mathcal{C}^!_n[d\{(u_1, \dots, u_n), \aleph\}], \tag{S6}$$

where $\mathcal{C}^!_n(A \times E)$, $A \subseteq S^n$, $E \in \mathcal{N}$, is the nth-order reduced Campbell measure. If $\mathcal{C}^!_n$ in (S6) is absolutely continuous with respect to the nth-order factorial moment measure, (S6) may be expressed as

$$\int_{S^n} E^!_{u_1,\dots,u_n}\{h(u_1, \dots, u_n, X)\}\, \rho^{(n)}(u_1, \dots, u_n)\mathrm{d}u_1 \cdots \mathrm{d}u_n. \tag{S7}$$

The equality between (S4) and (S7), via (S6), is referred to as the reduced Campbell–Mecke formula/theorem (Daley & Vere-Jones, 2008, Section 13). Moreover, the family of expectations in (S7) are governed by the nth-order reduced Palm distributions $P^!_{u_1,\dots,u_n}(E)$, $u_1, \dots, u_n \in S$, $E \in \mathcal{N}$. They satisfy that $P^!_{u_1,\dots,u_n}(\cdot)$ is the distribution of a point process $X^!_{u_1,\dots,u_n}$, which may be interpreted as X conditioned on having points at the locations $u_1, \dots, u_n$ which are removed upon realization. The associated product densities are given by (Coeurjolly et al., 2017, Equation (9))

$$\rho^{!(k)}(v_1, \dots, v_k|u_1, \dots, u_n) = \frac{\rho^{(k+n)}(v_1, \dots, v_k, u_1, \dots, u_n)}{\rho^{(n)}(u_1, \dots, u_n)}, \quad k \geq 1, \tag{S8}$$

when the nth-order product density of X satisfies $\rho^{(n)}(u_1, \dots, u_n) > 0$ and otherwise by 0. Further, by imposing absolute continuity of $\mathcal{C}^!_n$ with respect to the distribution of X, we obtain that (S6), and thereby (S4), equals (Daley & Vere-Jones, 2008)

$$\int_{S^n} E\left\{h(u_1, \dots, u_n, X)\lambda^{(n)}(u_1, \dots, u_n; X)\right\}\mathrm{d}u_1 \cdots \mathrm{d}u_n, \tag{S9}$$

where $\lambda^{(n)}$ in (S9) is the nth-order (Papangelou) conditional intensity, which further satisfies $E\{\lambda^{(n)}(u_1,\ldots,u_n;X)\} = \rho^{(n)}(u_1,\ldots,u_n)$. The equality between (S4) and (S9) is called the Georgii–Nguyen–Zessin (GNZ) formula/theorem, and point processes for which it is well-defined are sometimes called Gibbs processes (Coeurjolly et al., 2017). Moreover,

$$\lambda^{(n)}(u_1,\ldots,u_n;\aleph) = \lambda(u_1;\aleph)\lambda(u_2;\aleph\cup\{u_1\})\cdots\lambda(u_n;\aleph\cup\{u_1,\ldots,u_{n-1}\}), \tag{S10}$$

with $\lambda = \lambda^{(1)}$ referred to as 'the' conditional intensity, has the interpretation that the conditional probability of finding points of X in disjoint infinitesimal neighbourhoods du_i of $u_i \in S$, $i = 1,\ldots,n$, given that X agrees with $\aleph$ outside $du_1 \cup \ldots \cup du_n$, satisfies $\mathrm{pr}\{X(du_1) = 1,\ldots,X(du_n) = 1 \mid X \cap S \setminus (du_1 \cup \cdots \cup du_n) = \aleph \cap S \setminus (du_1 \cup \cdots \cup du_n)\} = \lambda^{(n)}(u_1,\ldots,u_n;\aleph)\mathrm{d}u_1\cdots\mathrm{d}u_n$ (Coeurjolly et al., 2017). This interpretation is motivated by the fact that, for a finite point process, the finite dimensional distributions are governed by Janossy measures, which may admit densities, $\{j_n\}_{n\geq 0}$, satisfying (Daley & Vere-Jones, 2008, Section 15.5)

$$\lambda(u,\aleph) = \begin{cases} j_{n+1}(\aleph\cup\{u\})/j_n(\aleph), & u \notin \aleph = \{x_1,\ldots,x_n\} \in \mathcal{X}, \\ j_n(\aleph)/j_{n-1}(\aleph\setminus\{u\}), & u \in \aleph = \{x_1,\ldots,x_n\} \in \mathcal{X}, \end{cases} \quad u \in S. \tag{S11}$$

Heuristically, $j_n(u_1,\ldots,u_n)\mathrm{d}u_1\cdots\mathrm{d}u_n$ gives the probability of X being contained in infinitesimal neighbourhoods of $u_1,\ldots,u_n \in S^n$ (Daley & Vere-Jones, 2003). Hence, $\lambda(\cdot)$ can be readily derived when the Janossy densities, which yield the (intractable) likelihood function, are known in closed form. Finally, in addition to attractiveness and repulsiveness (recall Section 2.1), there is also local stability: if $\lambda(\cdot;\aleph) \leq \phi^*(\cdot)$ for any $\aleph \in \mathcal{X}$ and some $|\cdot|$-integrable function ϕ^* on S, then X is called ϕ^*-locally stable (Møller & Waagepetersen, 2004, Section 6.1.1). It is noteworthy that both attractiveness/repulsiveness and local stability of λ transfer to $\lambda^{(n)}$.

Given two general spaces S and $\mathcal{M}$, with reference measures $|A|$, $A \subseteq S$, and $\nu_{\mathcal{M}}(B)$, $B \subseteq \mathcal{M}$, equip the (general) product space $\breve{S} = S \times \mathcal{M}$ with the product reference measure $\breve{\nu}(A \times B) = |A|\nu_{\mathcal{M}}(B)$. A point process $\breve{X} = \{(x_i,m_i)\}_{i=1}^N \subseteq \breve{S}$ is called a marked point process with marks $m_i \in \mathcal{M}$, $i = 1,\ldots,N$, if $X = \{x_i\}_{i=1}^N$ is a well-defined point process in S. The corresponding point configuration space is here denoted by $(\breve{\mathcal{X}},\breve{\mathcal{N}})$. It should be emphasized that the nth-order conditional intensity $\breve{\lambda}^{(n)}$ of $\breve{X}$ lives on $\breve{S}^n \times \breve{\mathcal{X}}$, and the intensity satisfies $\breve{\rho}^{(n)}\{(u_1,m_1),\ldots,(u_n,m_n)\} = E[\breve{\lambda}^{(n)}\{(u_1,m_1),\ldots,(u_n,m_n);\breve{X}\}]$.

S5.2. *A higher-order version of Theorem 1*

The result below is an nth-order version of Theorem 1 in the main text.

THEOREM S1. *Let Z be a p-thinning of a point process X on S, with retention probability $p(u) \in (0,1)$, $u \in S$, and $Y = X \setminus Z$. For any non-negative or integrable $h : S^n \times \mathcal{X} \to \mathbb{R}$, $n \geq 1$,*

$$E\left\{\sum_{x\in Z^n_{\neq}} h(x,Y)\right\} = E\left\{\sum_{x=(x_1,\ldots,x_n)\in Y^n_{\neq}} h(x,Y\setminus\{x\})\prod_{i=1}^n \frac{p(x_i)}{1-p(x_i)}\right\}. \tag{S12}$$

Moreover, provided that they exist, the nth-order conditional intensities, product densities and reduced Palm product densities of Z and X a.e. satisfy

$$\begin{aligned}
\lambda_Z^{(n)}(u_1,\dots,u_n,Z) &\overset{a.s.}{=} p(u_1)\cdots p(u_n)E\{\lambda_X^{(n)}(u_1,\dots,u_n;X)\mid Z\},\\
\rho_Z^{(n)}(u_1,\dots,u_n) &= p(u_1)\cdots p(u_n)\rho_X^{(n)}(u_1,\dots,u_n),\\
\rho_Z^{!(n)}(u_1,\dots,u_n|v_1,\dots,v_k) &= p(u_1)\cdots p(u_n)\rho_X^{!(n)}(u_1,\dots,u_n|v_1,\dots,v_k). \qquad \text{(S13)}
\end{aligned}$$

Given the associated marked point process representation $\breve{X}$ in Definition 1, when the nth-order conditional intensities of $\breve{X}$ and Y exist, they satisfy $E[\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\}\mid Y] = \lambda_Y^{(n)}(u_1,\dots,u_n;Y)\prod_{i=1}^n p(u_i)/[\prod_{i=1}^n\{1-p(u_i)\}]$ for almost all $u_1,\dots,u_n\in S$. In particular, for a p-thinning with retention probability $p\in(0,1)$, we set $p(\cdot)\equiv p$ in all expressions above.

Proof of Theorem S1. Starting with expression (S13), the form of the conditional intensity is a direct consequence of Decreusefond & Vasseur (2018, Theorem 4.7) and (S10). The result on the product densities follows from e.g. Baccelli et al. (2020, Proposition 2.3.24), and combining this with (S8), we obtain the result on reduced Palm product densities.

The structure of the proof of the prediction formula in (S12) follows the lines of the proof of Last & Penrose (2017, Exercise 5.9). Consider the random measure representation of X, where there are random variables $N = X(S)\in\{0,\dots,\infty\}$ and $X_1,\dots X_N\in S$ such that $X(A)=\sum_{i=1}^N\delta_{X_i}(A)=\sum_{i=1}^N\mathbb{1}(X_i\in A)$, $A\subseteq S$. An independent thinning of X has the same distribution as $Z(\cdot)=\sum_{i=1}^N B_i\delta_{X_i}(\cdot)$, where i) conditional on N and $X_1,\dots,X_N$, the random variables $B_1,\dots,B_N$ are mutually independent and, ii) for any $i=1,\dots,N$, conditional on X_i, the random variable B_i is Bernoulli distributed with parameter $p(X_i)$. Similarly, $Y=X\setminus Z$ has the random measure representation $Y(\cdot)=X(\cdot)-Z(\cdot)=\sum_{i=1}^N(1-B_i)\delta_{X_i}(\cdot)$. For any $m\geq n$, let $\mathcal{A}_m$ be the set of all n-tuples of distinct integers $i_1,\dots,i_n\in\{1,\dots,m\}$; if m is infinite, we let $i_1,\dots,i_n$ be finite. It now follows that

$$\begin{aligned}
&E\left[\sum_{x_1,\dots,x_n\in Z}^{\neq} h(x_1,\dots,x_n,Y)\prod_{i=1}^n\{1-p(x_i)\}\right]\\
=&E\left[\sum_{i_1,\dots,i_n\in\mathcal{A}_N} h(X_{i_1},\dots,X_{i_n},Y)\prod_{j=1}^n B_{i_j}\{1-p(X_{i_j})\}\right]\\
=&E\left\{\sum_{i_1,\dots,i_n\in\mathcal{A}_N} h(X_{i_1},\dots,X_{i_n},Y)\prod_{j=1}^n E(1-B_{i_j}\mid X)\prod_{j=1}^n B_{i_j}\right\}\\
=&E\left\{\sum_{i_1,\dots,i_n\in\mathcal{A}_N} h(X_{i_1},\dots,X_{i_n},Y\setminus\{X_{i_1},\dots,X_{i_n}\})\prod_{j=1}^n E(1-B_{i_j}\mid X)\prod_{j=1}^n B_{i_j}\right\},
\end{aligned}$$

where we have used that $p(X_{i_j})=E(B_{i_j}\mid X)$ and that $Y\cap\{X_{i_1},\dots,X_{i_n}\}=\emptyset$, i.e. $Y=Y\setminus\{X_{i_1},\dots,X_{i_n}\}$, when $B_{i_j}=1$ for all $j=1,\dots,n$. By the conditional independence of the B_{i_j}'s, we have that $\prod_{j=1}^n E(1-B_{i_j}\mid X)=E\{\prod_{j=1}^n(1-B_{i_j})\mid X\}$, and writing $\tilde{h}_{i_1,\dots,i_n}(X,Y)=$

$h(X_{i_1},\dots,X_{i_n},Y\setminus\{X_{i_1},\dots,X_{i_n}\})$, we obtain

$$
\begin{aligned}
&E\left\{\sum_{i_1,\dots,i_n\in\mathcal{A}_N} h(X_{i_1},\dots,X_{i_n},Y\setminus\{X_{i_1},\dots,X_{i_n}\})\prod_{j=1}^{n}E(1-B_{i_j}\mid X)\prod_{j=1}^{n}B_{i_j}\right\}\\
=&E\left[\sum_{i_1,\dots,i_n\in\mathcal{A}_N}\tilde{h}_{i_1,\dots,i_n}(X,Y)E\left\{\prod_{j=1}^{n}(1-B_{i_j})\mid X\right\}\prod_{j=1}^{n}B_{i_j}\right]\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left[\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)E\left\{\prod_{j=1}^{n}(1-B_{i_j})\mid X\right\}\prod_{j=1}^{n}B_{i_j}\right]\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left[E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\mid X\right\}\prod_{j=1}^{n}B_{i_j}\right],
\end{aligned}
$$

where the last equality follows from the "pulling out known factors" property of conditional expectations; N and $\tilde{h}_{i_1,\dots,i_n}(X,Y)$ are measurable with respect to the σ-algebra generated by X. By the law of total expectation, it follows that

$$
\begin{aligned}
&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left[E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\,\middle|\,X\right\}\prod_{j=1}^{n}B_{i_j}\right]\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left(E\left[E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\,\middle|\,X\right\}\prod_{j=1}^{n}B_{i_j}\,\middle|\,X\right]\right)\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left[E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\,\middle|\,X\right\}E\left(\prod_{j=1}^{n}B_{i_j}\,\middle|\,X\right)\right]\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left[E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\,\middle|\,X\right\}\prod_{j=1}^{n}p(X_{i_j})\right]\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left[E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\prod_{j=1}^{n}p(X_{i_j})\,\middle|\,X\right\}\right]\\
=&\sum_{i_1,\dots,i_n\in\mathcal{A}_\infty}E\left\{\mathbb{1}(N\geq\max\{i_1,\dots,i_n\})\tilde{h}_{i_1,\dots,i_n}(X,Y)\prod_{j=1}^{n}(1-B_{i_j})\prod_{j=1}^{n}p(X_{i_j})\right\}\\
=&E\left\{\sum_{i_1,\dots,i_n\in\mathcal{A}_N}h(X_{i_1},\dots,X_{i_n},Y\setminus\{X_{i_1},\dots,X_{i_n}\})\prod_{j=1}^{n}(1-B_{i_j})\prod_{j=1}^{n}p(X_{i_j})\right\},
\end{aligned}
$$

where we have used the fact that $E(\prod_{j=1}^{n}B_{i_j}\mid X)=\prod_{j=1}^{n}E(B_{i_j}|X)=\prod_{j=1}^{n}p(X_{i_j})$ by the conditional independence of the B_{i_j}'s, as well as the above-mentioned property of conditional expectations for $\prod_{j=1}^{n}p(X_{i_j})$ and the σ-algebra generated by X. Exploiting the representation $Y(\cdot)=X(\cdot)-Z(\cdot)=\sum_{i=1}^{N}(1-B_i)\delta_{X_i}(\cdot)$, we finally obtain that

$$
\begin{aligned}
&E\left\{\sum_{i_1,\dots,i_n\in\mathcal{A}_N}h(X_{i_1},\dots,X_{i_n},Y\setminus\{X_{i_1},\dots,X_{i_n}\})\prod_{j=1}^{n}(1-B_{i_j})\prod_{j=1}^{n}p(X_{i_j})\right\}\\
=&E\left\{\sum_{x_1,\dots,x_n\in Y}^{\neq}h(x_1,\dots,x_n,Y\setminus\{x_1,\dots,x_n\})\prod_{i=1}^{n}p(x_i)\right\},
\end{aligned}
$$

which proves (S12).

Next, let $\psi_0(\breve{x}) = \{x : (x,m) \in \breve{x} \cap S \times \{0\}\}$, $\breve{x} \in \breve{\mathcal{X}}$, and consider any non-negative or integrable $h : S^n \times \mathcal{X} \to \mathbb{R}$. Applying the GNZ formula to the left-hand side of (S12) yields

$$
\begin{aligned}
&E\left[\sum_{x_1,\dots,x_n \in Z}^{\neq} h(x_1,\dots,x_n,Y)\prod_{i=1}^{n}\{1-p(x_i)\}\right]\\
=&E\left[\sum_{(x_1,m_1),\dots,(x_n,m_n)\in \breve{X}}^{\neq} h\{x_1,\dots,x_n,\psi_0(\breve{X})\}\prod_{i=1}^{n} m_i\{1-p(x_i)\}\right]\\
=&\int_{S^n} E\left[\prod_{i=1}^{n}\{1-p(u_i)\}h\{u_1,\dots,u_n;\psi_0(\breve{X})\}\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\}\right]\mathrm{d}u_1\cdots\mathrm{d}u_n\\
=&\int_{S^n} E\left[\prod_{i=1}^{n}\{1-p(u_i)\}h(u_1,\dots,u_n;Y)\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\}\right]\mathrm{d}u_1\cdots\mathrm{d}u_n,
\end{aligned}
$$

since the reference measure on the mark space is the counting measure on $\mathcal{M} = \{0,1\}$. On the other hand, applying the GNZ formula to the right-hand side of (S12) yields

$$
\begin{aligned}
&E\left\{\sum_{x_1,\dots,x_n\in Y}^{\neq} h(x_1,\dots,x_n,Y\setminus\{x_1,\dots,x_n\})\prod_{i=1}^{n}p(x_i)\right\}\\
&=\int_{S^n} E\left\{\prod_{i=1}^{n}p(u_i)h(u_1,\dots,u_n;Y)\lambda_Y^{(n)}(u_1,\dots,u_n;Y)\right\}\mathrm{d}u_1\cdots\mathrm{d}u_n.
\end{aligned}
$$

The equality of these two expressions for arbitrary $h : S^n \times \mathcal{X} \to \mathbb{R}$ yields that, for almost every $u_1,\dots,u_n \in S^n$ and every non-negative or integrable $h^* : \mathcal{X} \to \mathbb{R}$,

$$
E\left(h^*(Y)\left[\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\} - \frac{\prod_{i=1}^{n}p(u_i)}{\prod_{i=1}^{n}\{1-p(u_i)\}}\lambda_Y^{(n)}(u_1,\dots,u_n;Y)\right]\right) = 0,
$$

which concludes the proof. □

S5.3. *A higher-order version of Theorem 2*

Theorem S2 is an nth-order version of Theorem 2 and Corollary 1 in the main text. We state and prove the result in terms of nth-order general parametrized estimator families $\Xi_\Theta^n = \{\xi_\theta^n : \theta \in \Theta\}$, $n \geq 1$, where

$$
\xi_\theta^n(u_1,\dots,u_n;\aleph), \quad u_1,\dots,u_n \in S, \quad \aleph \in \mathcal{X}, \quad \theta \in \Theta, \tag{S14}
$$

are real-valued and $\xi_\theta^n(\cdot;\aleph)$ is either non-negative or integrable for any $\aleph$. When each ξ_θ^n is constant over $\aleph \in \mathcal{X}$,

$$
\xi_\theta^n(u_1,\dots,u_n;\aleph) \equiv \xi_\theta^n(u_1,\dots,u_n), \quad u_1,\dots,u_n \in S, \quad \aleph \in \mathcal{X}, \quad \theta \in \Theta. \tag{S15}
$$

In Theorem S2 we show that the weight function appearing in the independent thinning setting has an additional bound when the point process is locally stable. In particular, for an attractive and locally stable point process, e.g. an area-interaction process (Møller & Waagepetersen, 2004), we have both upper and lower bounds for the weight function $w(\cdot)$ in (S20).

THEOREM S2. *Given a point process X in S, let Z be an arbitrary thinning of X, $Y = X \setminus Z$, and $\breve{X}$ the associated bivariate point process representation in Definition 1. Consider further some fixed $n \geq 1$, and let $\Xi_\Theta^n = \{\xi^n\}$ and $\mathcal{H}_\Theta = \{h\}$ consist of one element each.*

When $\xi^n, h : S^n \to \mathbb{R}$ are of the form (S15), we have that $\mathcal{I}_{\xi^n}^h(\cdot;Z,Y) = \mathcal{I}_{\xi^n}^h(\cdot;Z)$ satisfies

$$
E\{\mathcal{I}_{\xi^n}^h(A;Z)\} = \int_A h(u_1,\dots,u_n)\left\{\rho_Z^{(n)}(u_1,\dots,u_n) - \xi^n(u_1,\dots,u_n)\right\}\mathrm{d}u_1\cdots\mathrm{d}u_n \tag{S16}
$$

for any $A \subseteq S^n$, where $\rho_Z^{(n)}(\cdot)$ denotes the nth-order product density of Z; the variance of $\mathcal{I}^h_{\xi^n}(A; Z)$ can be found in expression (S21). Moreover, the expectation in (S16) is 0 for any $A \subseteq S^n$ and any test function h of the form (S15) if and only if

$$\xi^n(u_1, \ldots, u_n) \overset{a.e.}{=} \rho_Z^{(n)}(u_1, \ldots, u_n). \tag{S17}$$

If, instead, $\xi^n, h : S^n \times \mathcal{X} \to \mathbb{R}$ are of the form (S14), when $\breve{X}$ admits an nth-order conditional intensity $\breve{\lambda}^{(n)}(\cdot; \breve{X})$, for any $A \subseteq S^n$ we have

$$E\{\mathcal{I}^h_{\xi^n}(A; Z, Y)\} = \int_A E\left[h(u; Y)\left\{\breve{\lambda}_1^{(n)}(u; \breve{X}) - \xi^n(u; Y)\right\}\right] \mathrm{d}u, \tag{S18}$$

where $\breve{\lambda}_1^{(n)}(u; \breve{X}) = \breve{\lambda}^{(n)}\{(u_1, 1), \ldots, (u_n, 1); \breve{X}\}$, $u = (u_1, \ldots, u_n) \in S^n$, $n \geq 1$; the variance of $\mathcal{I}^h_{\xi^n}(A; Z, Y)$ can be found in expression (S26). Assume further $E\{\breve{\lambda}_1^{(n)}(u_1, \ldots, u_n; \breve{X})^2\} < \infty$ for $|\cdot|^n$-almost any $(u_1, \ldots, u_n) \in S^n$. Then, for any $A \subseteq S^n$ and any test function h such that $E\{h(u_1, \ldots, u_n; Y)^2\} < \infty$, we have that $E\{\mathcal{I}^h_{\xi^n}(A; Z, Y)\} = 0$ if and only if

$$\xi^n(u_1, \ldots, u_n; Y) \overset{a.e.}{=} E\left\{\breve{\lambda}_1^{(n)}(u_1, \ldots, u_n; \breve{X}) \mid Y\right\}. \tag{S19}$$

In particular, when Z is an independent thinning of X, based on a retention probability function $p(u) \in (0, 1)$, $u \in S$, then (S17) reads $\xi(u_1, \ldots, u_n) \overset{a.e.}{=} p(u_1) \cdots p(u_n)\rho_X^{(n)}(u_1, \ldots, u_n)$. Moreover, the right-hand side of (S19) is given by

$$p(u_1) \cdots p(u_n) E\{\lambda_X^{(n)}(u_1, \ldots, u_n; X) \mid Y\} = w(u_1, \ldots, u_n, Z, Y)\lambda_X^{(n)}(u_1, \ldots, u_n; Y), \tag{S20}$$

with $w(u_1, \ldots, u_n, Z, Y) = p(u_1) \cdots p(u_n)\lambda_X^{(n)}(u_1, \ldots, u_n; Y)^{-1} E\{\lambda_X^{(n)}(u_1, \ldots, u_n; X) \mid Y\}$. In particular, $w(u_1, \ldots, u_n, Z, Y) \leq p(u_1) \cdots p(u_n)$ if X is repulsive, $w(u_1, \ldots, u_n, Z, Y) \geq p(u_1) \cdots p(u_n)$ if X is attractive and $w(u_1, \ldots, u_n, Z, Y) = p(u_1) \cdots p(u_n)$ if X is a Poisson process. In addition, (S19) is smaller than or equal to $\prod_{i=1}^n p(u_i)\phi^*(u_i)$ if X is ϕ^*-locally stable.

Proof of Theorem S2. Recall that $\Xi_\Theta^n = \{\xi^n\} = \{\xi\}$ and $\mathcal{H}_\Theta = \{h\}$ here consist of one element each. Moreover, for ease of notation, we sometimes write $\mathrm{d}u$ for $\mathrm{d}u_1 \cdots \mathrm{d}u_n$.

When h and ξ are of the form (S15), by the Campbell formula we have that for any $A \subseteq S^n$,

$$\begin{aligned} E\{\mathcal{I}^h_\xi(A; Z, Y)\} =& E\left\{\sum_{(x_1, \ldots, x_n) \in Z^n_{\neq} \cap A} h(x_1, \ldots, x_n)\right\} - \int_A h(u_1, \ldots, u_n)\xi(u_1, \ldots, u_n)\mathrm{d}u \\ =& \int_A h(u_1, \ldots, u_n)\left\{\rho_Z^{(n)}(u_1, \ldots, u_n)\mathrm{d}u - \xi(u_1, \ldots, u_n)\right\}\mathrm{d}u. \end{aligned}$$

Hence, $E\{\mathcal{I}^h_\xi(A; Z, Y)\} = 0$ for any (bounded) $A \subseteq S^n$ and function h if and only if $\xi(u_1, \ldots, u_n) = \rho_Z^{(n)}(u_1, \ldots, u_n)$ for $|\cdot|^n$-almost every $(u_1, \ldots, u_n) \in S^n$; see e.g. Møller & Waagepetersen (2004, Section 2.3.3). We further have that

$$\begin{aligned} \mathrm{var}\{\mathcal{I}^h_\xi(A; Z, Y)\} =& \mathrm{var}\left\{\sum_{(x_1, \ldots, x_n) \in Z^n_{\neq} \cap A} h(x_1, \ldots, x_n)\right\} \\ =& E\left[\left\{\sum_{(x_1, \ldots, x_n) \in Z^n_{\neq} \cap A} h(x_1, \ldots, x_n)\right\}^2\right] - \left\{\int_A h(u_1, \ldots, u_n)\rho_Z^{(n)}(u_1, \ldots, u_n)\mathrm{d}u\right\}^2, \end{aligned}$$

where, by Poinas et al. (2019, Equation (B.3)),

$$
\begin{aligned}
&E\left[\left\{\sum_{(x_1,\dots,x_n)\in Z^n_{\neq}\cap A} h(x_1,\dots,x_n)\right\}^2\right] \\
=&E\left(\left[\sum_{x_1,\dots,x_n\in Z}^{\neq} \mathbb{1}\{(x_1,\dots,x_n)\in A\}h(x_1,\dots,x_n)\right]^2\right) \\
=&E\left[\left\{\sum_{\mathtt{y}\subseteq Z} n!\mathbb{1}(\#\mathtt{y}=n)\mathbb{1}(\mathtt{y}\in A)h(\mathtt{y})\right\}^2\right] \\
=&\sum_{j=0}^{n}\frac{(n!)^2}{(2n-j)!}\binom{n}{j}\binom{2n-j}{n}\int_{S^{2n-j}} h(u_1,\dots,u_n)h(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j}) \\
&\times \mathbb{1}\{(u_1,\dots,u_n)\in A\}\mathbb{1}\{(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j})\in A\} \\
&\times \rho_Z^{(2n-j)}(u_1,\dots,u_{2n-j})\mathrm{d}u_1\cdots\mathrm{d}u_{2n-j}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\operatorname{var}\{\mathcal{I}^h_\xi(A;Z)\} =&\sum_{j=0}^{n} j!\binom{n}{j}^2\int_{S^{2n-j}} h(u_1,\dots,u_n)h(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j}) \\
&\times \mathbb{1}\{(u_1,\dots,u_n)\in A\}\mathbb{1}\{(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j})\in A\}S \\
&\times \rho_Z^{(2n-j)}(u_1,\dots,u_{2n-j})\mathrm{d}u_1\cdots\mathrm{d}u_{2n-j} \\
&-\left\{\int_A h(u_1,\dots,u_n)\rho_Z^{(n)}(u_1,\dots,u_n)\mathrm{d}u_1\cdots\mathrm{d}u_n\right\}^2, \qquad \text{(S21)}
\end{aligned}
$$

where $j=0$ yields that $\{u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j}\}=\{u_{n+1},\dots,u_{2n}\}$ and $j=n$ yields that $\{u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j}\}=\{u_1,\dots,u_n\}$.

When h and ξ are of the form (S14), we start by defining

$$
\begin{aligned}
H_1(A) =& \sum_{(x_1,\dots,x_n)\in Z^n_{\neq}\cap A} h(x_1,\dots,x_n;Y\setminus\{x_1,\dots,x_n\}), \\
H_2(A) =& \int_A h(u_1,\dots,u_n;Y)\xi(u_1,\dots,u_n;Y)\mathrm{d}u, \\
\mu_1(A) =& E\left\{H_1(A)\right\}, \\
\mu_2(A) =& E\left\{H_2(A)\right\}, \quad A\subseteq S^n,
\end{aligned}
$$

where

$$
\begin{aligned}
E\{\mathcal{I}^h_\xi(A;Z,Y)\} =&\mu_1(A)-\mu_2(A), \qquad \text{(S22)} \\
E\{\mathcal{I}^h_\xi(A;Z,Y)^2\} =&E\{H_1(A)^2\}+E\{H_2(A)^2\}-2E\{H_1(A)H_2(A)\}, \\
\operatorname{var}\{\mathcal{I}^h_\xi(A;Z,Y)\} =&E\{H_1(A)^2\}+E\{H_2(A)^2\}-2E\{H_1(A)H_2(A)\}-\{\mu_1(A)-\mu_2(A)\}^2.
\end{aligned}
$$

Next, recall the associated marked point process $\breve{X}$ in Definition 1, with conditional intensity $\breve{\lambda}^{(n)}(\cdot)$. Given $\psi_0(\breve{\mathbb{x}}) = \{x : (x, m) \in \breve{\mathbb{x}} \cap S \times \{0\}\}$, $\breve{\mathbb{x}} \in \breve{\mathcal{X}}$, by the GNZ formula,

$$\begin{aligned}
\mu_1(A) =& E\left\{ \sum_{(x_1,\ldots,x_n)\in Z^n_{\neq}\cap A} h(x_1,\ldots,x_n; Y\setminus\{x_1,\ldots,x_n\}) \right\} \\
=& E\left\{ \sum_{((x_1,m_1),\ldots,(x_n,m_n))\in \breve{X}^n_{\neq}\cap(A\times\mathcal{M}^n)} \prod_{i=1}^n m_i h(x_1,\ldots,x_n;\psi_0[\breve{X}\setminus\{(x_1,m_1),\ldots,(x_n,m_n)\}]) \right\} \\
=& \int_A \sum_{m_1,\ldots,m_n\in\{0,1\}} E\left[\prod_{i=1}^n m_i h\{u_1,\ldots,u_n;\psi_0(\breve{X})\}\breve{\lambda}^{(n)}\{(u_1,m_1),\ldots,(u_n,m_n);\breve{X}\} \right] \mathrm{d}u \\
=& \int_A E\left[h\{u_1,\ldots,u_n;\psi_0(\breve{X})\}\breve{\lambda}^{(n)}\{(u_1,1),\ldots,(u_n,1);\breve{X}\} \right] \mathrm{d}u \\
=& \int_A E\left[h(u_1,\ldots,u_n;Y)\breve{\lambda}^{(n)}\{(u_1,1),\ldots,(u_n,1);\breve{X}\} \right] \mathrm{d}u,
\end{aligned}$$

since the reference measure on the mark space is the counting measure on the mark space $\mathcal{M} = \{0, 1\}$. On the other hand, by the Fubini-Tonelli theorem,

$$\mu_2(A) = E\left\{ \int_A h(u_1,\ldots,u_n;Y)\xi(u_1,\ldots,u_n;Y)\mathrm{d}u \right\} = \int_A E\Big\{ h(u_1,\ldots,u_n;Y)\xi(u_1,\ldots,u_n;Y) \Big\}\mathrm{d}u.$$

Hence, $E\{\mathcal{I}^h_\xi(A; Z, Y)\} = 0$ for any (bounded) $A \subseteq S^n$ if and only if

$$E\left(h(u_1,\ldots,u_n;Y)\left[\breve{\lambda}^{(n)}\{(u_1,1),\ldots,(u_n,1);\breve{X}\} - \xi(u_1,\ldots,u_n;Y)\right]\right) = 0,$$

for $|\cdot|^n$-almost every $(u_1, \ldots, u_n) \in S^n$; see e.g. Møller & Waagepetersen (2004, Section 2.3.3). Moreover, under the assumption that $E[\breve{\lambda}^{(n)}\{(u_1, 1), \ldots, (u_n, 1); \breve{X}\}^2] < \infty$ and $E\{h(u_1, \ldots, u_n; Y)^2\} < \infty$, L_2-projection yields that

$$\xi(u_1,\ldots,u_n;Y) = E[\breve{\lambda}^{(n)}\{(u_1,1),\ldots,(u_n,1);\breve{X}\}|Y].$$

We next turn to the variance. Similarly to Poinas et al. (2019, Equation (B.3)), we find that

$$\begin{aligned}
E\{H_1(A)^2\} =& E\Bigg[\sum_{(x_1,\ldots,x_n)\in Z^n_{\neq}} \sum_{(y_1,\ldots,y_n)\in Z^n_{\neq}} \mathbb{1}\{(x_1,\ldots,x_n),(y_1,\ldots,y_n)\in A\} \\
& \times h(x_1,\ldots,x_n;Y\setminus\{x_1,\ldots,x_n\})h(y_1,\ldots,y_n;Y\setminus\{y_1,\ldots,y_n\}) \Bigg] \\
=& n!^2 E\left\{ \sum_{\mathbb{x}=\{x_1,\ldots,x_n\}\subseteq Z} \sum_{\mathbb{y}=\{y_1,\ldots,y_n\}\subseteq Z} \mathbb{1}(\mathbb{x},\mathbb{y}\in A)h(\mathbb{x};Y\setminus\mathbb{x})h(\mathbb{y};Y\setminus\mathbb{y}) \right\} \\
=& n!^2 \sum_{j=0}^n E\left[\sum_{\mathbb{x}=\{x_1,\ldots,x_n\}\subseteq Z} \sum_{\mathbb{y}=\{y_1,\ldots,y_n\}\subseteq Z} \mathbb{1}\{\#(\mathbb{x}\cap\mathbb{y})=j\}\mathbb{1}(\mathbb{x},\mathbb{y}\in A)h(\mathbb{x};Y\setminus\mathbb{x})h(\mathbb{y};Y\setminus\mathbb{y}) \right],
\end{aligned}$$

where the factor $n!^2$ comes from the fact that when we go from n-subsets to n-tuples we count the same thing $n!$ times; we can rearrange $(x_1, \ldots, x_n)$ in $n!$ different ways. In the last sum, assuming that $\#(\mathbb{x} \cap \mathbb{y}) = j$, i.e. that $\mathbb{x}$ and $\mathbb{y}$ have j elements $x_i = y_{i'} \in Z$ in common, there are $\binom{n}{j}$ ways in which the elements in $\mathbb{x} \cap \mathbb{y}$ can be chosen from $\mathbb{x}$ and $\mathbb{y}$. The remaining $2n - j$ elements now need to be assigned to $\mathbb{x} \setminus (\mathbb{x} \cap \mathbb{y})$ and $\mathbb{y} \setminus (\mathbb{x} \cap \mathbb{y})$. There are $\binom{2n-j}{n-j} = \binom{2n-j}{n}$ ways to assign elements of $(\mathbb{x} \cap \mathbb{y})^c$ to $\mathbb{x} \setminus (\mathbb{x} \cap \mathbb{y})$ so that $\#\mathbb{x} = n$; the remaining elements will

automatically be assigned to $\mathbb{y} \setminus (\mathbb{x} \cap \mathbb{y})$. In the following, we let $z_1, \dots, z_j$ denote the elements in $\mathbb{x} \cap \mathbb{y}$, $z_{j+1}, \dots, z_n$ the elements only in $\mathbb{x}$, and $z_{n+1}, \dots, z_{2n-j}$ the ones only in $\mathbb{y}$. Consequently,

$$\begin{aligned} E\{H_1(A)^2\} =& n!^2 \sum_{j=0}^{n} \tbinom{n}{j} \tbinom{2n-j}{n} E\Bigg[\sum_{\{z_1,\dots,z_{2n-j}\} \subseteq Z} \mathbb{1}\{(z_1,\dots,z_n) \in A\} \\ & \times \mathbb{1}\{(z_1,\dots,z_j,z_{n+1},\dots,z_{2n-j}\} \in A) h(z_1,\dots,z_n; Y \setminus \{z_1,\dots,z_n\}) \\ & \times h(z_1,\dots,z_j,z_{n+1},\dots,z_{2n-j}; Y \setminus \{z_1,\dots,z_j,z_{n+1},\dots,z_{2n-j}\}) \Bigg] \\ =& \sum_{j=0}^{n} \tbinom{n}{j} \tbinom{2n-j}{n} \frac{n!^2}{(2n-j)!} E\Bigg[\sum_{(z_1,\dots,z_{2n-j}) \subseteq Z_{\neq}^{2n-j}} \mathbb{1}\{(z_1,\dots,z_n) \in A\} \\ & \times \mathbb{1}\{(z_1,\dots,z_j,z_{n+1},\dots,z_{2n-j}) \in A\} h(z_1,\dots,z_n; Y \setminus \{z_1,\dots,z_n\}) \\ & \times h(z_1,\dots,z_j,z_{n+1},\dots,z_{2n-j}; Y \setminus \{z_1,\dots,z_j,z_{n+1},\dots,z_{2n-j}\}) \Bigg], \end{aligned}$$

where

$$\frac{(n!)^2}{(2n-j)!} \tbinom{n}{j} \tbinom{2n-j}{n} = \tbinom{n}{j} \frac{(n!)^2}{(2n-j)!} \frac{(2n-j)!}{n!(n-j)!} = j! \tbinom{n}{j}^2.$$

By applying the GNZ formula to each term in the sum in the last equation, it follows that

$$\begin{aligned} E\{H_1(A)^2\} =& \sum_{j=0}^{n} j! \tbinom{n}{j}^2 \int_{S^{2n-j}} \mathbb{1}\{(u_1,\dots,u_n),(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j}) \in A\} \\ & \times E\Big[h(u_1,\dots,u_n;Y) h(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j};Y) \\ & \times \breve{\lambda}^{(2n-j)}\{(u_1,1),\dots,(u_{2n-j},1);\breve{X}\} \Big] \mathrm{d}u_1 \cdots \mathrm{d}u_{2n-j}. \end{aligned} \tag{S23}$$

We further have that

$$E\{H_2(A)^2\} = \int_A \int_A E\{h(u_1,\dots,u_n;Y) h(v_1,\dots,v_n;Y) \xi(u_1,\dots,u_n;Y) \xi(v_1,\dots,v_n;Y)\} \mathrm{d}u \mathrm{d}v \tag{S24}$$

and

$$\begin{aligned} E\{H_1(A)H_2(A)\} =& E\left\{ \sum_{(x_1,\dots,x_n) \in Z_{\neq}^n \cap A} h(x_1,\dots,x_n; Y \setminus \{x_1,\dots,x_n\}) \int_A h(v_1,\dots,v_n;Y) \xi(v_1,\dots,v_n;Y) \mathrm{d}v \right\} \\ =& E\Bigg\{ \sum_{(x_1,\dots,x_n) \in Z_{\neq}^n \cap A} h(x_1,\dots,x_n; Y \setminus \{x_1,\dots,x_n\}) \\ & \times \Bigg[\int_A h\{v_1,\dots,v_n; (Y \setminus \{x_1,\dots,x_n\}) \cup \{x_1,\dots,x_n\}\} \\ & \times \xi(v_1,\dots,v_n; (Y \setminus \{x_1,\dots,x_n\}) \cup \{x_1,\dots,x_n\}) \mathrm{d}v \Bigg] \Bigg\} \\ =& E\left\{ \sum_{(x_1,\dots,x_n) \in Z_{\neq}^n \cap A} \widetilde{h}(x_1,\dots,x_n; Y \setminus \{x_1,\dots,x_n\}) \right\}, \end{aligned}$$

where

$$\begin{aligned} \widetilde{h}(x_1,\dots,x_n; Y \setminus \{x_1,\dots,x_n\}) =& h(x_1,\dots,x_n; Y \setminus \{x_1,\dots,x_n\}) \int_A h\{v; (Y \setminus \{x_1,\dots,x_n\}) \cup \{x_1,\dots,x_n\}\} \\ & \times \xi\{v; (Y \setminus \{x_1,\dots,x_n\}) \cup \{x_1,\dots,x_n\}\} \mathrm{d}v. \end{aligned}$$

Hence,

$$E\{H_1(A)H_2(A)\} = \int_A E\left[\widetilde{h}(u_1,\dots,u_n;Y)\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\}\right]\mathrm{d}u$$
$$= \int_A\int_A E\Big[h(u_1,\dots,u_n;Y)h(v_1,\dots,v_n;Y\cup\{u_1,\dots,u_n\})$$
$$\times\,\xi(v_1,\dots,v_n;Y\cup\{u_1,\dots,u_n\})\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\}\Big]\mathrm{d}u\mathrm{d}v \tag{S25}$$

and, consequently, by combining (S22) with (S23)-(S25), the variance is

$$\mathrm{var}\{\mathcal{I}^h_\xi(A;Z,Y)\} = E\{\mathcal{I}^h_\xi(A;Z,Y)^2\} - E\{\mathcal{I}^h_\xi(A;Z,Y)\}^2 \tag{S26}$$
$$= \sum_{j=0}^{n} j!\binom{n}{j}^2 \int_{S^{2n-j}} \mathbb{1}\{(u_1,\dots,u_n),(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j})\in A\}E\Big[h(u_1,\dots,u_n;Y)$$
$$\times\, h(u_1,\dots,u_j,u_{n+1},\dots,u_{2n-j};Y)\breve{\lambda}^{(2n-j)}\{(u_1,1),\dots,(u_{2n-j},1);\breve{X}\}\Big]\mathrm{d}u_1\cdots\mathrm{d}u_{2n-j}$$
$$+ \int_A\int_A E\{h(u_1,\dots,u_n;Y)h(v_1,\dots,v_n;Y)$$
$$\times\,\xi^n(u_1,\dots,u_n;Y)\xi^n(v_1,\dots,v_n;Y)\}\mathrm{d}u_1\cdots\mathrm{d}u_n\mathrm{d}v_1\cdots\mathrm{d}v_n$$
$$- 2\left(\int_A\int_A E\Big[h(u_1,\dots,u_n;Y)h(v_1,\dots,v_n;Y\cup\{u_1,\dots,u_n\})\right.$$
$$\left.\times\,\xi(v_1,\dots,v_n;Y\cup\{u_1,\dots,u_n\})\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\}\Big]\mathrm{d}u_1\cdots\mathrm{d}u_n\mathrm{d}v_1\cdots\mathrm{d}v_n\right)$$
$$- \left[\int_A E\left\{h(u_1,\dots,u_n;Y)\left[\breve{\lambda}^{(n)}\{(u_1,1),\dots,(u_n,1);\breve{X}\} - \xi(u_1,\dots,u_n;Y)\right]\right\}\mathrm{d}u_1\cdots\mathrm{d}u_n\right]^2$$

when h and ξ are of the form (S14).

Turning to the independent thinning setting, the fact that (S17) reads $\xi(u_1,\dots,u_n)\overset{a.e.}{=} p(u_1)\cdots p(u_n)\rho^{(n)}_X(u_1,\dots,u_n)$ is an immediate consequence of (S13). Moreover, by Theorem S1, expression (S19) simplifies to

$$\frac{\lambda^{(n)}_Y(u_1,\dots,u_n;Y)}{\prod_{i=1}^n\{1-p(u_i)\}}\prod_{i=1}^n p(u_i) = \prod_{i=1}^n p(u_i)E\{\lambda^{(n)}_X(u_1,\dots,u_n;X)\mid Y\},$$

where for a locally stable point process the right-hand side is bounded from above by $\prod_{i=1}^n p(u_i)\phi^*(u_i)$. When X is attractive, since $Y\subseteq X$, the monotonicity of conditional expectations (Daley & Vere-Jones, 2003, Section A3.1) implies that the right-hand side is larger than or equal to $\prod_{i=1}^n p(u_i)E\{\lambda^{(n)}_X(u_1,\dots,u_n;Y)\mid Y\} = \prod_{i=1}^n p(u_i)\lambda^{(n)}_X(u_1,\dots,u_n;Y)$. Analogously, the inequality is reversed when X is repulsive and the equality for a Poisson process follows from the fact that its conditional intensity is deterministic and given by its intensity functions. □

S6. Asymptotic results

A fundamental step in statistical theory is to ensure that, with a sufficiently large sample, an estimator approximates the target parameter sufficiently well. This often translates into establishing consistency and asymptotic normality of the obtained estimators. Here we observe one realization $\mathbb{x}$ of a point process X, observed on $W\subseteq\mathbb{R}^d$, and whose distribution depends on a parameter

$\theta_0 \in \Theta$. In contrast to the classical iid setting, there is not a consensus on the definition of sample size here; see e.g. Choiruddin et al. (2021). In the point process learning setting, we identify three different asymptotic settings: i) a resampling regime, where $\#\mathcal{T}_k \to \infty$ (recall Section 5.1), which may also be studied conditionally on $X \cap W$, ii) an increasing-domain regime, using a sequence $W_1 \subseteq W_2 \subseteq \cdots$ of observation windows for W, and iii) an in-fill regime, where we consider a growing expected number of points over a fixed window W. Setting ii) is related to the idea that we need to observe a point process on a large enough scale in order to infer on its interaction structure. Here, we prove consistency and asymptotic normality for estimators obtained under regimes i) and ii). To achieve this, we apply general results from the theory for minimum contrast estimation, which we recall below.

S6.1. *General minimum contrast theory*

We here recall the theory on minimum contrast estimation found in e.g. Dacunha-Castelle & Duflo (1986) and Guyon (1995). Let us consider a parametric model of point process distributions P_θ, $\theta \in \Theta \subseteq \mathbb{R}^l$, where Θ is a compact set. We assume that the observed point pattern x is a realization of $X \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$.

Any non-negative function M on Θ such that $\theta_0 = \arg\min_{\theta\in\Theta} M(\theta)$ is called a contrast function. Intuitively, M measures how well a parameter θ fits the observation. For a given filtration $\{\mathcal{F}_t\}_{t>0}$, we let U_t be an F_t-measurable function from Θ to $\mathbb{R}$ for all $t > 0$. Here, we will consider e.g. the case $t = |W|$ and $\mathcal{F}_t = \sigma(X \cap W_t)$, and let U_t be a specific loss function.

Let us consider the following assumptions:

($\mathcal{M}1$) For all $\theta \in \Theta$, $U_t(\theta)$ converges in probability to $M(\theta)$.
($\mathcal{M}2$) $\Theta \subseteq \mathbb{R}^l$ is a compact set and θ_0 is the unique point such that $M(\theta_0) = \min_{\theta\in\Theta} M(\theta)$.
($\mathcal{M}3$) The functions M and U_t are continuous on Θ for $t > 0$.
($\mathcal{M}4$) There exists a real valued function ϕ on $\mathbb{R}$ such that $\lim_{\eta\to 0}\phi(\eta) = 0$ and for all $\eta > 0$,

$$\limsup_{t\to\infty} \sup_{\theta,\theta':\|\theta-\theta'\|<\eta} |U_t(\theta) - U_t(\theta')| \le \phi(\eta).$$

($\mathcal{M}5$) There exists an $\epsilon > 0$ such that, for $t > 0$, U_t is twice continuously differentiable in the interior of the closed ball $b(\theta_0, \epsilon) \subseteq \Theta$.
($\mathcal{M}6$) There exists a sequence $\{a_t\}_{t>0}$ and an invertible matrix J such that

$$\sqrt{a_t}\nabla U_t(\theta_0) \to \mathcal{N}(0, J).$$

($\mathcal{M}7$) There exists an invertible matrix V such that $U_t^{(2)}(\theta_0)$ converges in probability to V.

THEOREM S3 (DACUNHA-CASTELLE & DUFLO (1986); GUYON (1995)). *For all $t > 0$, let*

$$\hat{\theta}_t = \arg\min_{\theta\in\Theta} U_t(\theta).$$

If ($\mathcal{M}1$)-($\mathcal{M}4$) hold, then $\hat{\theta}_t$ converges in probability to θ_0 as t grows to infinity. If ($\mathcal{M}1$)-($\mathcal{M}7$) hold, then, in the sense of the convergence in distribution,

$$\lim_{t\to\infty} \sqrt{a_t}(\hat{\theta}_t - \theta_0) = \mathcal{N}(0, V^{-1}JV^{-1}).$$

The proof of Theorem S3 follows the same lines as the proofs of the results in (Dacunha-Castelle & Duflo, 1986; Guyon, 1995), but with slightly modified conditions.

S6.2. *Asymptotics in point process learning*

We next exploit the minimum contrast setting in Section S6.1 to study asymptotics for the point process learning approach in Section 5.1, for the regimes i) and ii) presented in the beginning of Section S6. We focus on minimization of the loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ in (10) to obtain an estimator for θ_0, and to emphasize the dependence on k and W in the notation, we write

$$\widehat{\theta}_{k,W} = \arg\min_{\theta\in\Theta} \mathcal{L}_i(\theta) \quad (i = 1, 2). \tag{S27}$$

In contrast, since the loss function (11) cannot be expressed as a suitable sum for minimum contrast estimation, it needs a different treatment, and therefore it will be excluded from consideration here. Let further $\mathcal{P} = \{P^W, W \subseteq \mathbb{R}^d, |W| < \infty\}$ be a family of thinning processes/operators, indexed by the bounded domains of $\mathbb{R}^d$, which generates a training-validation pair from any point pattern $\mathbb{x} \subseteq W$, i.e. $P^W(\mathbb{x}) = (\mathbb{x}^T, \mathbb{x}^V)$. For k independent and identical repetitions of $\mathcal{P}$, we write $\mathcal{J}(\theta, W, \mathcal{P}_i)$, $i = 1, \ldots, k$, for either $|\widetilde{\mathcal{I}}^{h_\theta}_{\xi^n_\theta}(W^n; X_i^V, X_i^T)|$ or $\widetilde{\mathcal{I}}^{h_\theta}_{\xi^n_\theta}(W^n; X_i^V, X_i^T)^2$, depending on if we use $\mathcal{L}_1$ or $\mathcal{L}_2$ in (10).

Regime i): W fixed

Recall that we in regime i) consider the setting where $\#\mathcal{T}_k \to \infty$, which is implied by $k \to \infty$. Let us now consider the following list of assumptions:

($\mathcal{H}_W1$) Θ is compact.

($\mathcal{H}_W2$) There exists a constant B such that a.s., for all $W \subseteq \mathbb{R}^d$,

$$\sup_{\theta\in\Theta} E[\mathrm{var}\{\mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\}] < B|W|^2.$$

($\mathcal{H}_W3$) $\mathcal{J}(\cdot, W, \mathcal{P})$ is continuous on Θ and $E\{\sup_{\theta\in\Theta} \mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\} < \infty$.

($\mathcal{H}_W4$) There exists a unique $\theta^*_W \in \Theta$ such that $\theta^*_W = \arg\min_{\theta\in\Theta} |W|^{-1} E\{\mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\}$.

($\mathcal{H}_W5$) There exists a real valued function ϕ on $\mathbb{R}$ such that $\lim_{\eta\to 0} \phi(\eta) = 0$ and for all $\eta > 0$,

$$\limsup_{k\to\infty} \frac{1}{k|W|} \sum_{i=1}^{k} \sup_{\theta,\theta':\|\theta-\theta'\|<\eta} |\mathcal{J}(\theta, W, \mathcal{P}_i) - \mathcal{J}(\theta', W, \mathcal{P}_i)| \le \phi(\eta).$$

($\mathcal{H}_W6$) There exists an $\epsilon > 0$ such that the closed ball $b(\theta^*_W, \epsilon) \subseteq \Theta$ and $\mathcal{J}(\cdot, W, \mathcal{P})$ is twice continuously differentiable in the interior of the closed ball $b(\theta^*_W, \epsilon)$.

($\mathcal{H}_W7$) There exists an invertible matrix J such that

$$\frac{1}{\sqrt{k}|W|} \sum_{i=1}^{k} \nabla\mathcal{J}(\theta^*_W, W, \mathcal{P}_i) \to \mathcal{N}(0, J).$$

($\mathcal{H}_W8$) There exists an invertible matrix V such that $(k|W|)^{-1} \sum_{i=1}^{k} \mathcal{J}^{(2)}(\theta^*_W, W, \mathcal{P}_i)$ converges in probability to V, as k tends to infinity.

We here have the following result.

THEOREM S4. *Let the framework be as described in Section 5.1 and assume that, for a given $W \subseteq \mathbb{R}^d$, ($\mathcal{H}_W1$)-($\mathcal{H}_W5$) hold. Then, for all $\epsilon > 0$,*

$$\lim_{k\to\infty} \mathrm{pr}(|\hat{\theta}_{k,W} - \theta^*_W| > \epsilon \mid X \cap W) = 0.$$

If, in addition, ($\mathcal{H}_W6$)-($\mathcal{H}_W8$) hold, then

$$\lim_{k\to\infty} \sqrt{|W|}(\widehat{\theta}_{k,W} - \theta^*_W) = \mathcal{N}(0, V^{-1}JV^{-1})$$

in the sense of the convergence in distribution.

Proof of Theorem S4. The proof is a direct application of Theorem S3 with $t = k$ and

$$U_t(\cdot) = \mathcal{L}(\cdot, k, W) = \frac{1}{k|W|}\sum_{i=1}^{k} \mathcal{J}(\cdot, W, \mathcal{P}_i). \tag{S28}$$

Hence, below we check the assumptions ($\mathcal{M}1$)-($\mathcal{M}4$) of Theorem S3.

Checking ($\mathcal{M}1$): Let $Z = \mathcal{L}(\theta, k, W)$. Since

$$\operatorname{var}(Z \mid X \cap W) = E\left[\{Z - E(Z \mid X \cap W)\}^2 \mid X \cap W\right]$$

and $E(\cdot) = E\{E(\cdot \mid X \cap W)\}$, it follows that

$$E[\{Z - E(Z \mid X \cap W)\}^2] = E\{\operatorname{var}(Z \mid X \cap W)\}.$$

Hence, by Markov's inequality, we have for all $\epsilon > 0$,

$$\operatorname{pr}\left\{|Z - E\big(Z \mid X \cap W\big)| > \epsilon\right\} \le \frac{E[\{Z - E(Z|X \cap W)\}^2]}{\epsilon^2} = \frac{E\{\operatorname{var}(Z \mid X \cap W)\}}{\epsilon^2}. \tag{S29}$$

Since the thinning processes $\mathcal{P}_i$ are identically distributed, and conditionally independent with respect to $X \cap W$, we have by (S28) and for all $\theta \in \Theta$ that

$$E\left(Z \mid X \cap W\right) = E\left\{\frac{1}{|W|}\mathcal{J}(\theta, W, \mathcal{P})|X \cap W\right\}$$

$$\operatorname{var}(Z \mid X \cap W) = \frac{1}{k|W|^2}\operatorname{var}\{\mathcal{J}(\theta, W, \mathcal{P}_1) \mid X \cap W\}.$$

Hence, by the two last equations, (S29) and ($\mathcal{H}_W2$),

$$\operatorname{pr}\left(\left|\mathcal{L}(\theta, k, W) - E\left\{\frac{1}{|W|}\mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\right\}\right| > \epsilon\right) \le \frac{B}{k\epsilon^2}$$

which implies ($\mathcal{M}1$) with $M(\theta) = E\{|W|^{-1}\mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\}$.

Checking ($\mathcal{M}2$): It follows directly from ($\mathcal{H}_W1$) and ($\mathcal{H}_W4$) with, as established above, $M(\theta) = E\{|W|^{-1}\mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\}$.

Checking ($\mathcal{M}3$): By ($\mathcal{H}_W3$), we have that $\mathcal{J}$ is continuous with respect to θ. The continuity of $M(\theta) = E\{|W|^{-1}\mathcal{J}(\theta, W, \mathcal{P}) \mid X \cap W\}$ follows from ($\mathcal{H}_W3$) and the dominated convergence theorem.

Checking ($\mathcal{M}4$)-($\mathcal{M}7$): They follow directly from ($\mathcal{H}_W5$)-($\mathcal{H}_W8$) with $a_t = k$. □

Regime ii): k fixed

We next turn to the increasing-domain regime and consider the following assumptions:

($\mathcal{H}_k1$) There exists a sequence of convex sets $\{W_j\}_{j\ge1}$ such that for all $j \ge 1$, $W_j \subseteq W_{j+1}$ and each W_j contains a ball of radius growing to infinity with j.

($\mathcal{H}_k2$) There exists a function M on Θ such that for all $\theta \in \Theta$, $\mathcal{L}(\theta, k, W_j)$ converges in probability to $M(\theta)$, as j tends to infinity.

($\mathcal{H}_k3$) $\Theta \subseteq \mathbb{R}^l$ is a compact set and θ_0 is the unique point such that $M(\theta_0) = \min_{\theta\in\Theta} M(\theta)$.

($\mathcal{H}_k4$) The functions $M(\cdot)$ and $\mathcal{L}(\cdot, k, W_j)$, $j \ge 1$, are continuous on Θ.

($\mathcal{H}_k5$) There exists a real valued function ϕ on $\mathbb{R}$ such that $\lim_{\eta\to0}\phi(\eta) = 0$ and for all $\eta > 0$,

$$\limsup_{j\to\infty} \frac{1}{k|W_j|}\sum_{i=1}^{k} \sup_{\theta,\theta':\|\theta-\theta'\|<\eta} |\mathcal{J}(\theta, W_j, \mathcal{P}_i) - \mathcal{J}(\theta', W_j, \mathcal{P}_i)| \le \phi(\eta).$$

($\mathcal{H}_k$6) There exists an $\epsilon > 0$ such that the closed ball $b(\theta_0, \epsilon) \subseteq \Theta$ and $\mathcal{L}(\cdot, k, W_j)$, $j \geq 1$, is twice continuously differentiable in the interior of the closed ball $b(\theta_0, \epsilon)$.

($\mathcal{H}_k$7) There exists an invertible matrix J such that, in the sense of convergence in distribution,

$$\lim_{j\to\infty} \sqrt{|W_j|}\frac{\partial}{\partial\theta}\mathcal{L}(\theta_0, k, W_j) = \mathcal{N}(0, J).$$

($\mathcal{H}_k$8) There exists an invertible matrix V such that $\partial^2\mathcal{L}(\theta_0, k, W_j)/\partial\theta^2$ converges in probability to V, as j tends to infinity.

Remark S1. In the case where X is ergodic, we have under weak conditions (Daley & Vere-Jones, 2008) that $M(\theta) = E[\mathcal{L}\{\theta, k, (0,1)^d\}]$ in ($\mathcal{H}_k$2). Then ($\mathcal{H}_k$4) may easily follow by imposing additional regularity assumptions on $\mathcal{J}(\cdot, k, W_j)$.

Theorem S5 below provides asymptotic results for the increasing-domain regime supplied by the assumptions above.

THEOREM S5. *Let the framework be as described in Section 5.1 and assume that ($\mathcal{H}_k$1)-($\mathcal{H}_k$5) hold for a given k. Then, for all $\epsilon > 0$,*

$$\lim_{j\to\infty} \mathrm{pr}(|\hat{\theta}_{k,W_j} - \theta_0| > \epsilon) = 0$$

If, in addition, ($\mathcal{H}_k$6)-($\mathcal{H}_k$8) hold, then, as $j \to \infty$,

$$\sqrt{|W_j|}(\widehat{\theta}_{k,W_j} - \theta_0) \to \mathcal{N}(0, V^{-1}JV^{-1})$$

in the sense of convergence in distribution.

Proof of Theorem S5. The proof is a direct application of Theorem S3 whose assumptions are immediately implied by ($\mathcal{H}_k$1)-($\mathcal{H}_k$8) with $t = l$ and $U_t(\cdot) = \mathcal{L}(\cdot, k, W_l)$. □

REFERENCES

ANG, Q. W., BADDELEY, A. & NAIR, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scand. J. Stat.* **39**, 591–617.

BABU, G. J. & FEIGELSON, E. D. (1996). Spatial point processes in astronomy. *J. Statist. Plann. Inference* **50**, 311–326.

BACCELLI, F., BLASZCZYSZYN, B. & KARRAY, M. (2020). *Random Measures, Point Processes, and Stochastic Geometry*. Inria.

BACCELLI, F. & BRÉMAUD, P. (2013). *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences*, vol. 26. Springer Science & Business Media.

BADDELEY, A., JAMMALAMADAKA, A. & NAIR, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society: Series C: Applied Statistics* , 673–694.

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.

BRÉMAUD, P. (1981). *Point processes and queues: martingale dynamics*, vol. 50. Springer.

CHAIBAN, C., BISCIO, C., THANAPONGTHARM, W., TILDESLEY, M., XIAO, X., ROBINSON, T. P., VANWAMBEKE, S. O. & GILBERT, M. (2019). Point pattern simulation modelling of extensive and intensive chicken farming in thailand: Accounting for clustering and landscape characteristics. *Agricultural Systems* **173**, 335–344.

CHAUDHURI, S., MORADI, M. & MATEU, J. (2021). On the trend detection of time-ordered intensity images of point processes on linear networks. *Comm. Statist. Simulation Comput.* .

CHOIRUDDIN, A., COEURJOLLY, J.-F. & WAAGEPETERSEN, R. (2021). Information criteria for inhomogeneous spatial point processes. *Aust. N. Z. J. Stat.* **63**, 119–143.

COEURJOLLY, J.-F., GUAN, Y., KHANMOHAMMADI, M. & WAAGEPETERSEN, R. (2016). Towards optimal takacs–fiksel estimation. *Spat. Stat.* **18**, 396–411.

COEURJOLLY, J.-F., MØLLER, J. & WAAGEPETERSEN, R. (2017). A tutorial on Palm distributions for spatial point processes. *Int. Stat. Rev.* **85**, 404–420.

CRONIE, O., NYSTRÖM, K. & YU, J. (2013). Spatiotemporal modeling of swedish scots pine stands. *Forest Science* **59**, 505–516.

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.

DACUNHA-CASTELLE, D. & DUFLO, M. (1986). *Probability and Statistics II*. Springer-Verlag New York.

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.

DECREUSEFOND, L. & VASSEUR, A. (2018). Stein's method and Papangelou intensity for Poisson or Cox process approximation. *arXiv preprint arXiv:1807.02453* .

DEHGHANI, A. & VAHIDI-ASL, M. Q. (2019). A quadrat neighborhood estimator for intensity function of point processes. *Journal of Applied Mathematics and Computing* **59**, 517–543.

DIGGLE, P. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Taylor & Francis/CRC.

GUAN, Y. (2007a). A composite likelihood cross-validation approach in selecting bandwidth for the estimation of the pair correlation function. *Scand. J. Stat.* **34**, 336–346.

GUAN, Y. (2007b). A least-squares cross-validation bandwidth selection approach in pair correlation function estimations. *Statist. Probab. Lett.* **77**, 1722–1729.

GUAN, Y., JALILIAN, A. & WAAGEPETERSEN, R. (2015). Quasi-likelihood for spatial point processes. *J. R. Statist. Soc. B* **77**, 677–697.

GUYON, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer Science & Business Media.

HALL, P., MINNOTTE, M. C. & ZHANG, C. (2004). Bump hunting with non-gaussian kernels. *Ann. Statist.* **32**, 2124–2141.

HESSELLUND, K. B., XU, G., GUAN, Y. & WAAGEPETERSEN, R. (2022). Second-order semi-parametric inference for multivariate log gaussian cox processes. *J. R. Statist. Soc. C* **71**, 244–268.

IFTIMI, A., CRONIE, O. & MONTES, F. (2019). Second-order analysis of marked inhomogeneous spatiotemporal point processes: Applications to earthquake data. *Scand. J. Stat.* **46**, 661–685.

JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning*, vol. 112. Springer.

JAMMALAMADAKA, A., BANERJEE, S., MANJUNATH, B. S. & KOSIK, K. S. (2013). Statistical analysis of dendritic spine distributions in rat hippocampal cultures. *BMC bioinformatics* **14**, 1–19.

KERSCHER, M. (2000). Statistical analysis of large-scale structure in the universe. In *Statistical physics and spatial statistics*, K. Mecke & D. Stoyan, eds., vol. 554. Springer, pp. 36–71.

LAST, G. & PENROSE, M. (2017). *Lectures on the Poisson process*, vol. 7. Cambridge University Press.

LOADER, C. (1999). *Local Regression and Likelihood*. New York: Springer.

MARSAN, D. & LENGLINE, O. (2008). Extending earthquakes' reach through cascading. *Science* **319**, 1076–1079.

MEYER, S., ELIAS, J. & HÖHLE, M. (2012). A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* **68**, 607–616.

MØLLER, J. & WAAGEPETERSEN, R. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. CRC.

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

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.

MORADI, M. & MATEU, J. (2020). First-and second-order characteristics of spatio-temporal point processes on linear networks. *J. Comput. Graph. Statist.* **29**, 432–443.

MORADI, M., MATEU, J. & COMAS, C. (2020). Directional analysis for point patterns on linear networks. *Stat* , e323.

MORADI, M., RODRIGUEZ-CORTES, F. & MATEU, J. (2018). On kernel-based intensity estimation of spatial point patterns on linear networks. *J. Comput. Graph. Statist.* **27**, 302–311.

OGATA, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50**, 379–402.

POINAS, A., DELYON, B., LAVANCIER, F. et al. (2019). Mixing properties and central limit theorem for associated point processes. *Bernoulli* **25**, 1724–1754.

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.

STOYAN, D. & PENTTINEN, A. (2000). Recent applications of point process methods in forestry statistics. *Statist. Sci.* **15**, 61–78.

TORETI, A., CRONIE, O. & ZAMPIERI, M. (2019). Concurrent climate extremes in the key wheat producing regions of the world. *Scientific reports* **9**, 1–8.

VAN LIESHOUT, M. N. M. (2000). *Markov Point Processes and Their Applications*. Imperial College Press.
VAN LIESHOUT, M. N. M. (2011). A J-function for inhomogeneous point processes. *Stat. Neerl.* **65**, 183–201.
ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301–320.

[*Received on* 1 $January$ 2022. *Editorial decision on* 1 $January$ 2022]