CENTRE FOR **STOCHASTIC GEOMETRY**
AND ADVANCED **BIOIMAGING**

www.csgb.dk

RESEARCH REPORT 2020

Achmad Choiruddin, Jean-François Coeurjolly and Rasmus Waagepetersen

Information criteria for inhomogeneous spatial point processes

No. 04, March 2020

Information criteria for inhomogeneous spatial point processes

Achmad Choiruddin[1], Jean-François Coeurjolly[2]
and Rasmus Waagepetersen[3]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS)
[2]Department of Mathematics, Université du Québec à Montréal (UQAM)
[3]Department of Mathematical Sciences, Aalborg University

Abstract

The theoretical foundation for a number of model selection criteria is established in the context of inhomogeneous point processes and under various asymptotic settings: infill, increasing domain, and combinations of these. For inhomogeneous Poisson processes we consider Akaike's information criterion and the Bayesian information criterion, and in particular we identify the point process analogue of 'sample size' needed for the Bayesian information criterion. Considering general inhomogeneous point processes we derive new composite likelihood and composite Bayesian information criteria for selecting a regression model for the intensity function. The proposed model selection criteria are evaluated using simulations of Poisson processes and cluster point processes.

Keywords: Akaike's information criterion, Bayesian information criterion, composite information criterion, composite likelihood, inhomogeneous point process, intensity function, model selection.

1 Introduction

Fitting a regression model to the intensity function of a point process is one of the most fundamental tasks in statistical analysis of point pattern data , see e.g. Møller and Waagepetersen (2017) or Coeurjolly and Lavancier (2019) for a recent review of this problem. If the data in question can be viewed as a realization of a Poisson process, regression parameters are usually estimated by maximum likelihood. If the point process is not Poisson, the likelihood function is often computationally intractable. In such cases the Poisson likelihood function can still be used as a composite likelihood function for estimating regression parameters. This is e.g. the approach underlying the popular `spatstat` `R` package (Baddeley et al., 2015) procedure `kppm`.

Considering regression models, model selection is often a pertinent task. In case of a Poisson process, the Akaike information criterion (AIC) (Akaike, 1973) seems an obvious approach (and implemented in the function `logLik.ppm` of `spatstat`) since in this case the likelihood function is available. If the assumption of a Poisson process is not tenable, generalization of the AIC to a composite likelihood information criterion (CIC) (Varin and Vidoni, 2005) is relevant. Yet another alternative is the Bayesian information criterion (BIC) (Schwarz, 1978). The classical framework for the information criteria mentioned typically involves a sample of independent observations where the sample size plays a crucial role for the BIC. However, in the case of a unique realization of a point process, it is not obvious how to define sample size. In some sense, it is one, but this choice is obviously not useful for asymptotic justifications of information criteria. Instead, sample size must be linked to properties of the observation window or the point process intensity function. Several proposals for defining 'sample size' have been considered in the literature. Choiruddin et al. (2018) use the size of the observation window while Thurman et al. (2015) consider the number of observed points. Jeffrey et al. (2018) use the sum of the number of data points and the number of dummy points used in a numerical approximation of the likelihood (Berman and Turner, 1992).

In this paper we first establish the theoretical foundation for AIC and CIC in the context of intensity function model selection for a point process. This includes asymptotic results for estimates of the 'least false parameter value', see e.g. Claeskens and Hjort (2008, Section 2.2). Next we derive the BIC in case of a Poisson process and we thus identify what is the meaning of sample size in this context. We also consider the generalization of BIC to composite likelihood BIC (CBIC) (Gao and Song, 2010) using a concept of effective degrees of freedom derived for the CIC. Our asymptotic developments are established under an original setting which embraces both infill asymptotic (the number of points in a fixed domain increases) and increasing domain asymptotics (the volume of the observation window tends to infinity) which are often considered in the literature.

The rest of the article is organized as follows. The problem of selecting a model for the intensity function is specified in Section 2. In Section 3 we discuss asymptotic results for intensity function regression parameter estimators under a 'double' asymptotic framework. We derive the AIC and CIC for spatial point processes in Section 4 and develop the BIC and CBIC in Section 5. The different model selection criteria are compared in a simulation study in Section 6. Section 7 gives some concluding remarks. Proofs are given in the Appendices A–E.

2 Intensity model selection

A spatial point process $\mathbf{X}$ defined on $\mathbb{R}^d$ is a locally finite random subset of $\mathbb{R}^d$. If for bounded $B \subset S$ we denote by $N(B)$ the cardinality of $\mathbf{X} \cap B$, locally finite means that $N(B)$ is a finite integer almost surely. For a bounded domain $W \subset S$, $|W|$ denotes the volume of W. The intensity function λ and the pair correlation function

g of $\mathbf{X}$ are defined (if they exist) by the equations

$$\mathbb{E}N(A) = \int_A \lambda(u)\mathrm{d}u$$
$$\mathbb{E}\{N(A)N(B)\} = \int_A \lambda(u)\mathrm{d}u + \int_A \int_B \lambda(u)\lambda(v)g(u,v)\mathrm{d}u\mathrm{d}v$$

for any bounded $A, B \subset \mathbb{R}^d$.

If the counts $N(A)$ are Poisson distributed, $\mathbf{X}$ is said to be a Poisson process. In this case counts $N(B_1), \dots, N(B_m)$ are independent whenever the subsets $B_1, \dots, B_m$ are disjoint and the pair correlation function is identically equal to one. For our asymptotic considerations, we assume that a sequence of spatial point processes $\mathbf{X}_n$ is observed within a sequence of bounded windows $W_n \subset \mathbb{R}^d$, $n = 1, 2, \dots$ We denote by λ_n and g_n the intensity and pair correlation function of $\mathbf{X}_n$. With an abuse of notation, we denote for any $n \geq 1$ expectation and variance under the sampling distribution of $\mathbf{X}_n$ by $\mathbb{E}$ and $\mathbb{V}\mathrm{ar}$.

For modelling the intensity function we assume that $p \geq 1$ covariates $z_1, \dots, z_p$ are available where for each $i = 1, \dots, p$, z_i is a locally integrable function on $\mathbb{R}^d$. Let I_l, $l = 1, \dots, 2^p$, denote the subsets of $\{1, \dots, p\}$ and let $p_l = |I_l| + 1$ where $|I_l|$ is the cardinality of I_l. We consider models for λ_n specified in terms of the 2^p subsets of the covariates. For each $l = 1, \dots, 2^p$ and $n \geq 1$ we define the log-linear model

$$\rho_{l,n}\{u; \mathbf{z}_l(u); \boldsymbol{\beta}_l\} = \theta_n \exp\{\boldsymbol{\beta}_l^\top \mathbf{z}_l(u)\} \tag{2.1}$$

where $\boldsymbol{\beta}_l = \{\beta_0, (\beta_j)_{j \in I_l}\} \in \mathbb{R}^{p_l}$ and $\mathbf{z}_l(u) = [1, \{z_j(u)\}_{j \in I_l}]$. The quantity θ_n should not be regarded as a parameter to be estimated. For $n \geq 1$, θ_n could e.g. represent a timespan over which $\mathbf{X}_n$ is observed. In the following, with an abuse of notation, we just write $\rho_n(u; \boldsymbol{\beta}_l)$ for $\rho_{l,n}\{u; \mathbf{z}_l(u); \boldsymbol{\beta}_l\}$ and similarly for related quantities.

The problem we consider is to select among the 2^p intensity models $\mathcal{M}_l$, $l = 1, \dots, 2^p$, given by

$$\mathcal{M}_l = \{\rho_n(\,\cdot\,; \boldsymbol{\beta}_l) \mid \boldsymbol{\beta}_l = \{\beta_0, (\beta_j)_{j \in I_l}\} \in \mathbb{R}^{p_l}\}, \qquad l = 1, \dots, 2^p.$$

The distinction between β_0 and the other parameters is necessary because our objective is to select among 2^p different models which all contain an intercept term. Note that the true intensity function λ_n does not necessarily correspond to any of the suggested models $\mathcal{M}_l$, $l = 1, \dots, 2^p$.

3 Estimation of the intensity function

In this section we discuss estimation of the intensity function using a Poisson likelihood function and associated asymptotic results.

3.1 The Poisson likelihood function

The density of a Poisson point process with intensity $\rho_n(\,\cdot\,; \boldsymbol{\beta}_l)$ and observed in W_n is given by (see e.g. Møller and Waagepetersen (2004))

$$p_n(\mathbf{x}; \boldsymbol{\beta}_l) = \Big\{\prod_{u \in \mathbf{x}} \rho_n(u; \boldsymbol{\beta}_l)\Big\} \exp\Big\{|W_n| - \int_{W_n} \rho_n(u; \boldsymbol{\beta}_l)\mathrm{d}u\Big\} \tag{3.1}$$

for locally finite point configurations $\mathbf{x} \subset W_n$. We emphasize that in the following we assume neither that $\mathbf{X}_n$ introduced in the previous section is a Poisson process nor that $\rho_n(\,\cdot\,;\boldsymbol{\beta}_l)$ coincides with the intensity function λ_n of $\mathbf{X}_n$.

Combining (2.1) and (3.1), up to a constant, the log of (3.1) evaluated at $\mathbf{X}_n$ becomes

$$\ell_n(\boldsymbol{\beta}_l) = \sum_{u\in\mathbf{X}_n} \boldsymbol{\beta}_l^\top \mathbf{z}_l(u) - \theta_n \int_{W_n} \exp\{\boldsymbol{\beta}_l^\top \mathbf{z}_l(u)\}\mathrm{d}u. \tag{3.2}$$

Let $e_n(\boldsymbol{\beta}_l)$ be the corresponding estimating function given by

$$e_n(\boldsymbol{\beta}_l) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_l}\ell_n(\boldsymbol{\beta}_l) = \sum_{u\in\mathbf{X}_n} \mathbf{z}_l(u) - \theta_n \int_{W_n} \mathbf{z}_l(u)\rho_n(u;\boldsymbol{\beta}_l)\mathrm{d}u. \tag{3.3}$$

For any $\boldsymbol{\beta}_l \in \mathbb{R}^{p_l}$, the sensitivity (or Fisher information) matrix is

$$\mathbf{S}_n(\boldsymbol{\beta}_l) = -\mathbb{E}\left\{\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_l^\top} e_n(\boldsymbol{\beta}_l)\right\} = \theta_n \int_{W_n} \mathbf{z}_l(u)\mathbf{z}_l(u)^\top \rho_n(u;\boldsymbol{\beta}_l)\mathrm{d}u. \tag{3.4}$$

We assume $\mathbf{S}_n(\boldsymbol{\beta}_l)$ is positive definite for all $\boldsymbol{\beta}_l$ (see also condition C5 in the next section). We can then define the estimator of $\boldsymbol{\beta}_l$ as

$$\hat{\boldsymbol{\beta}}_{l,n} = \underset{\boldsymbol{\beta}_l\in\mathbb{R}^{p_l}}{\arg\max}\; p_n(\mathbf{X}_n;\boldsymbol{\beta}_l) = \underset{\boldsymbol{\beta}_l\in\mathbb{R}^{p_l}}{\arg\max}\; \ell_n(\boldsymbol{\beta}_l). \tag{3.5}$$

If $\mathbf{X}_n$ is indeed a Poisson process with intensity function $\rho_n(\cdot;\boldsymbol{\beta}_l)$, then $\hat{\boldsymbol{\beta}}_l$ is the maximum likelihood estimator and the sensitivity (3.4) equals the observed information matrix $-\mathrm{d}e_n(\boldsymbol{\beta}_l)/\mathrm{d}\boldsymbol{\beta}_l^\top$.

If $\mathbf{X}_n$ is not Poisson, $\hat{\boldsymbol{\beta}}_{l,n}$ may be viewed as a composite likelihood estimator (Schoenberg, 2005; Waagepetersen, 2007). In the situation where the intensity function $\rho_n(\,\cdot\,;\boldsymbol{\beta}_l)$ coincides with the true intensity function λ_n, asymptotic properties of maximum likelihood or composite likelihood estimators obtained as maximizers of Poisson likelihood functions have been established in various settings by Rathbun and Cressie (1994), Waagepetersen (2007), Guan and Loh (2007) and Waagepetersen and Guan (2009). In the next section we investigate the more intriguing situation where the intensity model is misspecified.

3.2 Framework and asymptotic results for misspecified intensity functions

To handle the situation where $\rho_n(\,\cdot\,;\boldsymbol{\beta}_l)$ does not coincide with the intensity function of $\mathbf{X}_n$, we follow Varin and Vidoni (2005) and define a (composite) Kullback-Leibler divergence between the model $\mathcal{M}_l$ with parameter $\boldsymbol{\beta}_l$ and the true sampling distribution. That is,

$$\mathrm{KL}_n(\boldsymbol{\beta}_l) = \mathbb{E}\left\{\ell_n - \ell_n(\boldsymbol{\beta}_l)\right\} \tag{3.6}$$

where ℓ_n is the Poisson log-likelihood obtained with the true intensity λ_n. For a window W_n and model $\mathcal{M}_l$ we let

$$\boldsymbol{\beta}^*_{l,n} = \underset{\boldsymbol{\beta}_l\in\mathbb{R}^{p_l}}{\arg\max}\, \mathbb{E}\ell_n(\boldsymbol{\beta}_l)$$

denote the 'least wrong parameter value' under model $\mathcal{M}_l$, provided the maximum exists. It is easy to see by explicit evaluation of the right hand side that

$$-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_l{}^\top}\left\{-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_l}\mathbb{E}\ell_n(\boldsymbol{\beta}_l)\right\} = \mathbf{S}_n(\boldsymbol{\beta}_l)$$

and so condition C5 stated below implies that $\boldsymbol{\beta}^*_{l,n}$ is well-defined as a unique maximum when n is large enough. Also it is easy to see that

$$\mathbb{E}e_n(\boldsymbol{\beta}^*_{l,n}) = 0, \tag{3.7}$$

which means that $\hat{\boldsymbol{\beta}}_{l,n}$ given by (3.5) is a candidate to estimate $\boldsymbol{\beta}^*_{l,n}$.

The remainder of this section is devoted to asymptotic results for $\hat{\boldsymbol{\beta}}_{l,n}$ within the above framework of a misspecified intensity function. We thereby extend the results in the references mentioned in Section 3.1. In contrast to these references which used either increasing domain or infill asymptotics, we moreover consider a 'double asymptotic' framework as formalized by condition C3 presented below.

Two matrices are crucial for the asymptotic results. The sensitivity matrix $\mathbf{S}_n(\boldsymbol{\beta}_l)$ is given by (3.4) regardless of whether the model is misspecified or not. The variance-covariance matrix $\boldsymbol{\Sigma}_{l,n}$ of (3.3) is, using the Campbell theorem, given by

$$\begin{aligned}\boldsymbol{\Sigma}_{l,n} = &\int_{W_n} \mathbf{z}_l(u)\mathbf{z}_l(u)^\top \lambda_n(u)\mathrm{d}u \\ &+ \int_{W_n}\int_{W_n} \mathbf{z}_l(u)\mathbf{z}_l(u)^\top \lambda_n(u)\lambda_n(v)\{g_n(u,v)-1\}\mathrm{d}u\mathrm{d}v.\end{aligned} \tag{3.8}$$

Observe that $\boldsymbol{\Sigma}_{l,n}$ does not depend on $\boldsymbol{\beta}_l$ (whence its notation). Our results will be based on the following assumptions where for a square matrix $\mathbf{M}$, $\nu_{\min}(\mathbf{M})$ (resp. $\nu_{\max}(\mathbf{M})$) stands for the smallest (resp. largest) eigenvalue. We use $a_n \asymp b_n$ to denote that $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$.

[C1] As $n \to \infty$, $\sup_{n\geq 1}\|\boldsymbol{\beta}^*_{l,n}\| = \mathcal{O}(1)$.

[C2] $\mathbf{z}_l : \mathbb{R}^d \to \mathbb{R}$ is continuous, $0 < \inf_{u\in\mathbb{R}^d}\|\mathbf{z}_l(u)\| < \sup_{u\in\mathbb{R}^d}\|\mathbf{z}_l(u)\| < \infty$.

[C3] The sequence $(\tau_n := \theta_n|W_n|)_{n\geq 1}$ is an increasing sequence, such that

$$\liminf_{n\to\infty}\theta_n > 0 \qquad \text{and} \qquad \lim_{n\to\infty}\tau_n = \infty.$$

The sets W_n are convex and compact.

[C4] As $n \to \infty$, $\tau_n^{-1}e_n(\boldsymbol{\beta}^*_{l,n}) \to 0$ almost surely.

[C5] For any $\boldsymbol{\beta}_l \in \mathbb{R}^{p_l}$ and $n \geq 1$, $\mathbf{S}_n(\boldsymbol{\beta}_l)$ is positive definite. In addition, for any $n \geq 1$, there exists a set $B_n \subseteq W_n$ such that $|B_n| \asymp |W_n|$ and a $c > 0$ such that $\inf_n \inf_{\boldsymbol{\phi}\in\mathbb{R}^{p_l},\|\boldsymbol{\phi}\|=1}\inf_{u\in B_n}|\boldsymbol{\phi}^\top\mathbf{z}_l(u)| \geq c$. Finally, we assume that $\liminf_{n\to\infty}\nu_{\min}\left(\tau_n^{-1}\boldsymbol{\Sigma}_{l,n}\right) > 0$.

[C6] As $n \to \infty$, $\|\boldsymbol{\Sigma}_{l,n}\| = \mathcal{O}(\tau_n)$.

[C7] As $n \to \infty$, $\boldsymbol{\Sigma}_{l,n}^{-1/2}e_n(\boldsymbol{\beta}^*_{l,n}) \to N(0, \mathbf{I}_{p_l})$ in distribution.

We can then state the following asymptotic result which is verified in Appendix A.

Theorem 1.

(i) Assume conditions C1–C4 hold, then there exists $v = v_n(\hat{\boldsymbol{\beta}}_{l,n}, \boldsymbol{\beta}^*_{l,n})$*, such that almost surely*

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n} \\ = \frac{\mathbf{z}_l(v)}{\|\mathbf{z}_l(v)\|^2} \log\Big[1 + \tau_n^{-1} e_n(\boldsymbol{\beta}^*_{l,n})^\top \frac{\mathbf{z}_l(v)}{\|\mathbf{z}_l(v)\|^2} \exp\{-\boldsymbol{\beta}^{*\top}_{l,n} \mathbf{z}_l(v)\}\Big].\end{aligned} \tag{3.9}$$

and $\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n} \to 0$ *almost surely as* $n \to \infty$*.*

(ii) Assume conditions C1–C3 and C5–C6 hold. Then, $\hat{\boldsymbol{\beta}}_{l,n}$ *is a root-*τ_n *consistent estimator of* $\boldsymbol{\beta}^*_{l,n}$*, i.e.*

$$\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n} = \mathcal{O}_P(\tau_n^{-1/2}). \tag{3.10}$$

(iii) If in addition, C7 holds, then as $n \to \infty$*,*

$$\boldsymbol{\Sigma}_{l,n}^{-1/2} \mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n}) \to N(0, \mathbf{I}_{p_l}) \tag{3.11}$$

in distribution.

We stress that Theorem 1 (ii)–(iii) do not require the strong consistency of $\tau_n^{-1} e_n(\boldsymbol{\beta}^*_{l,n})$ to 0, i.e. condition C4. We conclude by some remarks regarding the assumptions C1–C7. Condition C1 ensures that the sequence of 'least wrong parameter value' does not diverge with n.

Condition C3 is different from existing conditions as it embraces both of the standard asymptotic frameworks considered in the literature:

- infill asymptotics: $W_n = W$ with W a bounded set of $\mathbb{R}^d$ and $\theta_n \to \infty$ as $n \to \infty$.
- increasing domain asymptotics: $\theta_n = \theta > 0$ and $(W_n)_{n\geq 1}$ is a sequence of bounded domains of $\mathbb{R}^d$ such that $|W_n| \to \infty$.

It is also valid if both asymptotics are considered at the same time. The assumed convexity in C3 enables the use of the mean value theorem in the proofs of our theoretical results.

In condition C2, assuming an upper bound for $\|\mathbf{z}_l(u)\|$ is quite standard. The upper bound further implies that assuming a lower bound is not really restrictive since for each covariate z_j we can always find some k so that $|z_j(u) + k| \neq 0$ for any $u \in \mathbb{R}^d$. Replacing z_j by $z_j + k$ while changing the intercept from β_0 to $\beta_0 - \beta_j k$ leaves the model unchanged. The continuity assumption is used to prove the strong consistency of $\hat{\boldsymbol{\beta}}_{l,n}$.

Condition C4 is also used to ensure the strong consistency of $\hat{\boldsymbol{\beta}}_{l,n}$. This condition can be seen as a law of large numbers. In Example 2, we present a class of models where such an assumption is valid under the generalized asymptotic condition C3. We can observe that if there exists $p > 0$ such that $\mathbb{E}\{\|\tau_n^{-1/2} e_n(\boldsymbol{\beta}^*_{l,n}\|^p\} = \mathcal{O}(1)$ and

$\sum_n \tau_n^{-p/2} < \infty$, then C4 ensues from an application of Borel-Cantelli's lemma. At least in the increasing domain framework, assuming such a moment assumption is quite standard to derive a central limit theorem.

Condition C5 is very similar to the assumption required by Rathbun and Cressie (1994) under the Poisson case or by Waagepetersen and Guan (2009) for more general point processes, within the increasing domain asymptotic framework. Note in particular that C5 combined with C1–C2 ensures that

$$\liminf_{n\to\infty} \nu_{\min}\{\tau_n^{-1}\mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})\} > 0.$$

Under the Poisson case, $g_n = 1$ and so C6 is obviously satisfied if $\sup_{u\in W_n} \lambda_n(u) = \mathcal{O}(\theta_n)$. For more general point processes, assume further that $g_n = g$ does not depend on n and is invariant under translations. Then, thanks to C2, the assumption

$$\int_{\mathbb{R}^d} \{g(o,w) - 1\}\mathrm{d}w < \infty.$$

will imply C6. This assumption is quite standard and satisfied by a large class of models, see e.g. Waagepetersen and Guan (2009).

Continuing within the increasing domain framework, C7 was established by Rathbun and Cressie (1994) in the Poisson case, by Guan and Loh (2007) and Waagepetersen and Guan (2009) for α-mixing point processes, and by Lavancier et al. (2020) for determinantal point processes. The infill asymptotic framework was used to establish C7 in case of Poisson cluster processes in Waagepetersen (2007). However, the 'double' asymptotic framework has never been considered. Below, we provide an example of a model which satisfies C4, C6 and C7 under the new general asymptotic setting C3.

Example 2. Let $\mathbf{C}_n$ be a homogeneous Poisson point process on $\mathbb{R}^d$ with intensity θ_n. Given $\mathbf{C}_n$, let $\mathbf{X}_{n,c}$, $c \in \mathbf{C}_n$, be independent inhomogeneous Poisson point processses on W_n with intensity $\alpha k(u-c)\rho(u)$ where $\alpha > 0$, k is a symmetric density on $\mathbb{R}^d$, and ρ is a non-negative bounded function. Then, $\mathbf{X}_n = \bigcup_{c\in\mathbf{C}_n} \mathbf{X}_{n,c}$ is an inhomogeneous Poisson cluster point process (with inhomogeneous offspring). It can be shown that $\lambda_n(u) = \theta_n\alpha\rho(u)$ and $g_n(u,v) = 1 + (k*k)(v-u)/\theta_n$ where the notation $*$ denotes convolution. Assuming that $\sup_u \rho(u) = \mathcal{O}(1)$, there exists $K \geq 0$ such that

$$\begin{aligned}
\|\boldsymbol{\Sigma}_{l,n}\| &\leq K\big\{\tau_n + \theta_n^2 \int_{W_n}\int_{W_n} |g_n(u,v) - 1|\mathrm{d}u\mathrm{d}v\big\} \\
&\leq K\big\{\tau_n + \theta_n \int_{W_n}\int_{W_n} |(k*k)(v-u)|\mathrm{d}u\mathrm{d}v\big\} \\
&\leq K\big\{\tau_n + \tau_n \int_{\mathbb{R}^d} |(k*k)(w)|\mathrm{d}w\big\} = \mathcal{O}(\tau_n).
\end{aligned}$$

Thus C6 is satisfied. In Appendix B we show that this model also satisfies C4 and C7 within the 'double' asymptotic framework. Along the same lines one can show that C7 holds for the inhomogeneous Poisson point process with intensity function $\lambda_n(u) = \theta_n\rho(u)$.

4 Akaike and composite information criteria

One criterion for model selection would be to choose the model that minimizes the Kullback-Leibler divergence, i.e. the model for which (3.6) evaluated at $\boldsymbol{\beta}_{l,n}^*$ is smallest. Of course this criterion is not useful in practice since the true model is unknown. Following Varin and Vidoni (2005) we instead choose the model that minimizes an estimate of the expected value of (3.6) evaluated at $\hat{\boldsymbol{\beta}}_{l,n}$, or equivalently, that minimizes an estimate of $2\mathbb{E}C_n(\hat{\boldsymbol{\beta}}_{l,n})$ with $C_n(\boldsymbol{\beta}_l) = -\mathbb{E}\ell_n(\boldsymbol{\beta}_l)$. We follow Varin and Vidoni (2005, Lemmas 1–2) to derive in our context the following result.

Proposition 3. *Assume conditions C1–C3 and C5–C6 hold. Also assume that there exists $\varepsilon > 0$ such that*

$$\sup_n \mathbb{E}\big\{\|e_n(\boldsymbol{\beta}_{l,n}^*)^\top \mathbf{M}_n e_n(\boldsymbol{\beta}_{l,n}^*)\|^{1+\varepsilon}\big\} < \infty \tag{4.1}$$

where $\mathbf{M}_n = \mathbf{S}_n(\tilde{\boldsymbol{\beta}}_{l,n})^{-1} - \mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}$ and $\tilde{\boldsymbol{\beta}}_{l,n}$ is on the line segment between $\hat{\boldsymbol{\beta}}_{l,n}$ and $\boldsymbol{\beta}_{l,n}^*$. Then,

$$\mathbb{E}\{C_n(\hat{\boldsymbol{\beta}}_{l,n})\} = \mathbb{E}\{-\ell_n(\hat{\boldsymbol{\beta}}_{l,n})\} + \operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}\boldsymbol{\Sigma}_{l,n}\} + o(1).$$

In other words, $-2\ell_n(\hat{\boldsymbol{\beta}}_l) + 2\operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}\boldsymbol{\Sigma}_{l,n}\}$ is an asymptotically unbiased estimator of $2\mathbb{E}\{C_n(\hat{\boldsymbol{\beta}}_{l,n})\}$.

Remark 4. The technical condition (4.1) implies that the sequence

$$\{\tilde{e}_n := e_n(\boldsymbol{\beta}_{l,n}^*)^\top \mathbf{M}_n e_n(\boldsymbol{\beta}_{l,n}^*)\}_{n\geq 1}$$

is a uniformly integrable sequence of random variables. This ensures that convergence in probability of $\tilde{e}_n$ to 0 implies that $\mathbb{E}\tilde{e}_n = o(1)$.

To estimate the effective degrees of freedom $p_l^* = \operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}\boldsymbol{\Sigma}_{l,n}\}$ we first simply estimate $\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)$ by $\mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})$. The estimation of $\boldsymbol{\Sigma}_{l,n}$ is more difficult. Following Claeskens and Hjort (2008, p. 31), note that if $\rho_n(\,\cdot\,;\boldsymbol{\beta}_{l,n}^*)$ coincides with the true intensity function of $\mathbf{X}_n$, then $\boldsymbol{\Sigma}_{l,n}$ coincides with $\boldsymbol{\Sigma}_n(\boldsymbol{\beta}_{l,n}^*)$ given by $\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*) + T_2$ where

$$T_2 := \int_{W_n^2} \mathbf{z}_l(u)\mathbf{z}_l^\top(v)\rho_n(u;\boldsymbol{\beta}_{l,n}^*)\rho_n(v;\boldsymbol{\beta}_{l,n}^*)\{g_n(u,v)-1\}\mathrm{d}u\mathrm{d}v.$$

In this case we get

$$\begin{aligned}\operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}\boldsymbol{\Sigma}_n(\boldsymbol{\beta}_{l,n}^*)\} &= \operatorname{trace}\{\mathbf{I}_{p_l} + \mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}T_2\}\\ &= p_l + \operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}T_2\}.\end{aligned}$$

We propose to use $p_{l,\mathrm{approx}}^* = p_l + \operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)^{-1}T_2\}$ as an approximation of p_l^*. In practice we replace $\boldsymbol{\beta}_{l,n}^*$ by $\hat{\boldsymbol{\beta}}_{l,n}$ and g_n by an estimate obtained by fitting a valid parametric model for g_n and thus obtain $\hat{p}_{l,\mathrm{approx}}^*$. Our composite model selection criterion then becomes

$$\mathrm{CIC}_l = -2l_n(\hat{\boldsymbol{\beta}}_{l,n}) + 2\hat{p}_{l,\mathrm{approx}}^*. \tag{4.2}$$

For a Poisson process, $g_n = 1$ in which case $\hat{p}^*_{l,\text{approx}} = p_l$ and (4.2) reduces to the popular Akaike's information criterion. For a clustered point process, $g_n > 1$ meaning that $\hat{p}^*_{l,\text{approx}} > p_l$. Thus we penalize more the complexity of the model in the case of a clustered point process. This seems to make sense since random clustering of points (i.e. not due to covariates) may erroneously be picked up by covariates that actually had no effect in the data generating mechanism. Hence there is a greater risk of picking a too complex model for the intensity function in case of a clustered point process than for a Poisson process.

5 Bayesian information criterion

The motivation of the Bayesian information criterion (BIC) is quite different from the derivation of the AIC and CIC criteria considered in the previous section. A main difference is that there is initially no reference to a Kullback-Leibler distance or asymptotics related to a 'least false parameter value'. Instead, the true model is considered to be one of the models $\mathcal{M}_l$ and the idea is to choose the model that has maximum posterior probability within the specified Bayesian framework. However, the asymptotic concepts again play a role in order to derive asymptotic expansions of the posterior probabilities. Section 5.1 covers the Poisson process case. Section 5.2 proposes a composite likelihood BIC in the case where data is not generated from a Poisson process. Note that in this case, it is still assumed that the true intensity function corresponds to one of intensity functions for the models $\mathcal{M}_l$.

5.1 BIC in the Poisson process case

The BIC criterion (see e.g. Schwarz (1978); Lebarbier and Mary-Huard (2006)) defines the best model $\mathcal{M}_{\text{BIC}}$ as

$$\mathcal{M}_{\text{BIC}} = \underset{\mathcal{M}_l, l=1,\dots,2^p}{\arg\max} \; \mathrm{P}_n\left(\mathcal{M}_l \mid \mathbf{X}_n\right). \tag{5.1}$$

From Bayes formula,

$$\mathrm{P}_n(\mathcal{M}_l \mid \mathbf{X}_n) = \frac{p_n(\mathbf{X}_n \mid \mathcal{M}_l)\mathrm{P}(\mathcal{M}_l)}{p_n(\mathbf{X}_n)}$$

Letting $p(\boldsymbol{\beta}_l \mid \mathcal{M}_l)$ denote the prior density of $\boldsymbol{\beta}_l$ given $\mathcal{M}_l$,

$$p_n(\mathbf{X}_n \mid \mathcal{M}_l) = \int_{\mathbb{R}^{p_l}} p_n(\mathbf{X}_n; \beta_l) p(\boldsymbol{\beta}_l \mid \mathcal{M}_l) \mathrm{d}\boldsymbol{\beta}_l$$

since $p_n(\mathbf{X}_n; \beta_l)$ is the conditional density of $\mathbf{X}_n$ given $\mathcal{M}_l$ and $\boldsymbol{\beta}_l$. We assume that the prior distribution over models is non-informative, so that the BIC criterion defines the best model as

$$\mathcal{M}_{\text{BIC}} = \underset{\mathcal{M}_l, l=1,\dots,2^p}{\arg\max} \; p_n(\mathbf{X}_n \mid \mathcal{M}_l). \tag{5.2}$$

In principle, one could evaluate the $p_n(\mathbf{X}_n \mid \mathcal{M}_l)$ using numerical quadrature and then determine $\mathcal{M}_{\text{BIC}}$. However, this is computationally costly and also the need to elicit a specific prior $p(\boldsymbol{\beta}_l \mid \mathcal{M}_l)$ for each model may be a nuisance. Our next result therefore proposes an asymptotic expansion of $\log p_n(\mathbf{X}_n \mid \mathcal{M}_l)$. The methodology is standard (basically a Laplace approximation) and well-known in the literature, see Tierney and Kadane (1986) or e.g. Lebarbier and Mary-Huard (2006) and the references therein. However, due to our spatial framework and the double asymptotic point of view considered in this paper, the standard results do not apply straightforwardly. Due to the use of asymptotic results we again need to rely on the notions of a true model and the 'least false parameter value' $\boldsymbol{\beta}^*_{l,n}$.

We impose the following conditions on the prior for $\boldsymbol{\beta}_l$ and the mean $\mu_n = \mathbb{E}\mathbf{X}_n$.

[C8] The prior density $p(\boldsymbol{\beta}_l \mid \mathcal{M}_l)$ of $\boldsymbol{\beta}_l$ given $\mathcal{M}_l$ is continuously differentiable on $\mathbb{R}^{p_l}$.

[C9] $\mu_n \asymp \tau_n$.

Note that μ_n is the marginal mean of $\mathbf{X}_n$ under the true intensity model λ_n as discussed in Section 2. Condition C9 seems reasonable since it would hold if λ_n coincided with any of the specified parametric intensity models $\rho_n(\,\cdot\,;\boldsymbol{\beta}_l)$. We also need to slightly strengthen assumption C1 and replace it by

[C1′] As $n \to \infty$, there exists $\boldsymbol{\beta}^*_l \in \mathbb{R}^{p_l}$ such that $\lim_{n\to\infty} \boldsymbol{\beta}^*_{l,n} = \boldsymbol{\beta}^*_l$.

The following result is verified in Appendix D using Łapiński (2019, Theorem 2), which is a rigorous statement of a multivariate Laplace approximation.

Proposition 5. *Under the conditions C1′, C2–C5 and C8–C9, we have, almost surely with respect to the distribution of $\{\mathbf{X}_n\}_{n\geq 1}$, as $n \to \infty$,*

$$\begin{aligned}\log p_n(\mathbf{X}_n \mid \mathcal{M}_l) = \ell_n(\hat{\boldsymbol{\beta}}_{l,n}) &- \frac{p_l}{2}\log(\mu_n)\\ &+ \frac{p_l}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mu_n^{-1}\mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})\}\\ &+ \log p(\hat{\boldsymbol{\beta}}_{l,n} \mid \mathcal{M}_l) + \mathcal{O}(\mu_n^{-1/2}).\end{aligned} \tag{5.3}$$

The criterion (5.2) is defined entirely within the specified Bayesian framework. Hence no reference to a true model and no need for asymptotic results. However, this is changed when we derive the expansion (5.3) for (5.2). The expansion is around $\hat{\boldsymbol{\beta}}_{l,n}$ and for technical reasons, when applying the Laplace approximation, convergence of $\hat{\boldsymbol{\beta}}_{l,n}$ is needed. Therefore we need the assumptions that ensure strong consistency of $\hat{\boldsymbol{\beta}}_{l,n}$.

Since $\mu_n \asymp \tau_n$ and from the strong consistency of $\hat{\boldsymbol{\beta}}_{l,n}$, we have that

$$\log\det\{\mu_n^{-1}\mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})\} = \mathcal{O}_{\mathrm{P}}(1)$$

while C8 ensures that $\log p(\hat{\boldsymbol{\beta}}_{l,n} \mid \mathcal{M}_l)$ is $\mathcal{O}_{\mathrm{P}}(1)$. So, if we neglect terms which are $\mathcal{O}_{\mathrm{P}}(1)$ in (5.3), we follow the standard heuristic and suggest to define a first version of the BIC criterion as

$$-2\ell_n(\hat{\boldsymbol{\beta}}_{l,n}) + p_l\log(\mu_n) \tag{5.4}$$

where we remind that p_l is the length of $\boldsymbol{\beta}_l$.

In practice μ_n is not known. However, since $\mathbb{V}\mathrm{ar}\, N(W_n) = \mu_n$ it follows that $N(W_n)/\mu_n - 1 = \mathcal{O}_{\mathrm{P}}(\mu_n^{-1/2}) = o_{\mathrm{P}}(1)$. This justifies to define the BIC criterion in the following natural way

$$\mathrm{BIC}_l = -2\ell_n(\hat{\boldsymbol{\beta}}_{l,n}) + p_l \log\{N(W_n)\}. \tag{5.5}$$

5.2 Composite likelihood BIC

Suppose $\mathbf{X}_n$ has an intensity function of the form (2.1) but is not a Poisson process. Then as mentioned in Section 3.1, (3.2) may be viewed as a composite likelihood score for estimating $\boldsymbol{\beta}_l$. In this case, following Gao and Song (2010), (5.2) may be viewed as a composite likelihood BIC. Again we obtain (5.5) from (5.2) by Laplace approximation. However, Gao and Song (2010) suggest to replace p_l by the 'effective degrees of freedom' p_l^* considered in Section 4. Thus our proposed composite likelihood BIC is

$$\mathrm{CBIC}_l = -2\ell_n(\hat{\boldsymbol{\beta}}_{l,n}) + p^*_{l,\mathrm{approx}} \log\left\{N(W_n)\right\}$$

which becomes equal to the ordinary BIC in (5.5) for a Poisson process. From a practical point of view, we simply estimate $p^*_{l,\mathrm{approx}}$ as in (4.2).

6 Simulation study

To evaluate the proposed model selection criteria we conduct two simulation studies with $p = 6$ spatial covariates, of which four have zero effect. The first study in Section 6.1 considers AIC and BIC in the Poisson point process case. The second study considers the clustered Thomas point process where we employ the CIC and CBIC criteria to select the best model and compare with results obtained using AIC and BIC (assuming wrongly that the simulated data are from a Poisson point process).

The covariates are obtained from the BCI dataset (Hubbell and Foster, 1983; Condit et al., 1996; Condit, 1998) which in addition to locations of around 300 species of trees observed in $W = [0, 1000] \times [0, 500]$ (m^2) contains a number of spatial covariates. In particular, we center and scale the two topological covariates (elevation and slope of elevation) and four soil nutrients (aluminium, boron, calcium, and copper). The six covariates are depicted in Figure 1.

Skipping the dependence on n in the notation, we model the intensity function as

$$\rho(u, \boldsymbol{\beta}) = \omega \exp\{\beta_1 z_1(u) + \cdots + \beta_6 z_6(u)\}, \tag{6.1}$$

where $z_1, \ldots, z_6$ are the centered and scaled covariates as in Figure 1 and $\beta_1, \ldots, \beta_6$ are the regression coefficients. We consider different settings for ω and W. Note that ω plays the role of $\theta \exp(\beta_0)$ and will be adjusted to obtain desired expected numbers of points μ in W. When the simulation involves an observation window W different from $[0, 1000] \times [0, 500]$, the covariates are simply rescaled to fit W.

The model selection criteria are compared in terms of the true positive rate (TPR), false positive rate (FPR), expected Kullback-Leibler divergence (MKL), and

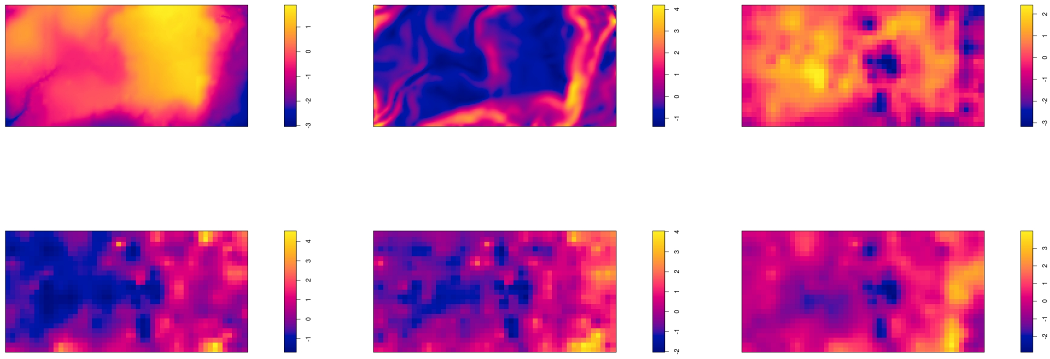

Figure 1: Maps of covariates used in the simulation study. From left to right: First row: elevation (z_1), slope (z_2) and aluminium (z_3); Second row: boron (z_4), calcium (z_5) and copper (z_6).

mean integrated squared error of the intensity function (MISE). The TPR (resp. FPR) are the expected fractions of informative (resp. non-informative) covariates included in the selected model. For a point process observed on W, with intensity $\rho(\,\cdot\,;\boldsymbol{\beta}^*)$ where $\boldsymbol{\beta}^*$ stands for the true parameter estimated by $\hat{\boldsymbol{\beta}}$, MKL and MISE are estimated by averaging the following KL and ISE across simulations:

$$\mathrm{KL} = \int_W \big[\rho(u;\boldsymbol{\beta}^*)\{\log \rho(u;\boldsymbol{\beta}^*) - 1\} - \rho(u;\hat{\boldsymbol{\beta}})\{\log \rho(u;\hat{\boldsymbol{\beta}}) - 1\}\big]\mathrm{d}u$$

$$\mathrm{ISE} = \int_W \{\rho(u;\boldsymbol{\beta}^*) - \rho(u;\hat{\boldsymbol{\beta}})\}^2 \mathrm{d}u.$$

6.1 Poisson point process model

We consider two different scenarios to illustrate both types of asymptotics.

- **Scenario 1 (infill asymptotics).** $W = [0,1] \times [0,0.5]$. We adjust $\omega = \theta \exp(\beta_0)$ such that μ equals either 50 or 200.
- **Scenario 2 (increasing domain asymptotics).** $W = [0,500] \times [0,250]$ or $[0,1000] \times [0,500]$ and ω is chosen so that $\mu = 200$ or $\mu = 800$.

For each scenario and the choices of W and ω, 500 simulations are generated from inhomogeneous Poisson point processes with intensity function given by (6.1) using the function `rpoispp` from the `spatstat` `R` package. We set $\beta_1 = 0.5$ and $\beta_2 = -0.25$ to represent moderate effects of elevation and slope, and set $\beta_3 = \cdots = \beta_6 = 0$. For each simulation, parameters are estimated by maximizing the Berman-Turner approximation (see e.g. Baddeley and Turner, 2000) of the Poisson log-likelihood (3.2) using a number m of quadrature dummy points to approximate the integral in (3.2). The estimation is done using the `ppm` function. Then, the model is selected according to the AIC and BIC-type criteria. We consider different variants of the BIC criterion, namely

$$\mathrm{BIC}_l(\pi) = -2\ell(\hat{\boldsymbol{\beta}}_l) + p^*_{l,\mathrm{approx}} \log(\pi), \qquad l = 1,\ldots,64,$$

where π represents a penalty. Note that (omitting dependence on l)

$$\text{AIC} = \text{BIC}\{\exp(2)\}$$

and that BIC(N) ($N = N(W)$) corresponds to the criterion used by Thurman et al. (2015) and is also the criterion suggested by the present paper. We also consider BIC($|W|$) used by Choiruddin et al. (2018) and BIC($N + m$) considered by Jeffrey et al. (2018).

Table 1: True positive (TPR) and false positive (FPR) rates in percent, MISE, and MKL, estimated from 500 simulations of inhomogeneous Poisson point processes on different observation domains. Model selections are based on AIC or BIC criteria of the form $\text{BIC}_l = -2\ell(\hat{\boldsymbol{\beta}}_l) + p_l \log \pi$ ($l = 1, \ldots, 64$) for different penalty terms $\pi = N, N + 400\mu, |W|$. For convenience, the four rows with MISE are multiplied by respectively .001, .01, 100, and 1000.

		AIC	BIC(π), $\pi =$		
			N	$N + 400\mu$	$\|W\|$
$W = [0, 1] \times [0, 0.5]$	TPR	70	59	49	100
$\mu = 50$	FPR	16	5	1	100
	MISE	6.5	6.0	6.0	7.9
	MKL	3.0	2.9	2.9	3.6
$W = [0, 1] \times [0, 0.5]$	TPR	94	83	63	100
$\mu = 200$	FPR	17	2	0	100
	MISE	2.6	2.3	3.3	3.2
	MKL	2.9	2.8	4.2	3.5
$W = [0, 500] \times [0, 250]$	TPR	95	83	63	62
$\mu = 200$	FPR	16	2	0	0
	MISE	1.0	0.9	1.3	1.3
	MKL	2.8	2.7	4.1	4.2
$W = [0, 1000] \times [0, 500]$	TPR	100	99	99	98
$\mu = 800$	FPR	18	1	0	0
	MISE	10.4	6.5	6.7	7.0
	MKL	2.9	1.8	1.9	2.0

For both scenarios, we perform estimation with $m = 4\mu$ (the rule of thumb suggested by `spatstat`) and model selection with the criteria AIC, BIC(N), BIC(N+4μ) and BIC($|W|$). Similar results are obtained with BIC(N) and BIC($N + 4\mu$) and we omit the results for the latter. We also perform estimation and selection with $m = 400\mu$ and only report results for BIC($N + 400\mu$) since the results with the criteria AIC, BIC(N) and BIC($|W|$) are similar to those obtained when the estimation is performed with $m = 4\mu$.

When $|W|$ is small, especially when $\log|W| < 0$, the criterion BIC($|W|$) obviously fails as it selects the most complex model regardless of the value of μ. Thereby the FP rate becomes 100%. In addition, as indicated in the second and third rows of

Table 1 where the point patterns have the same average number of points, it is worth noticing that BIC($|W|$) selects pretty different models. In particular this criterion has an undesirable strong dependence on the choice of length unit.

The criterion BIC($N + m$) with a large m also fails since the TPR with this criterion and $m = 400\mu$ is very small compared to the other criteria, especially when the expected number of points is small or moderate. The AIC criterion achieves a high TPR in all situations but fails since it suffers from a high FPR, even in the scenario 2 where $\mu = 800$.

In all cases, BIC(N) provides the best trade-off between TPR and FPR and the results improve when ω or $|W|$ is increased. The minimal values of MKL and MISE are further always obtained with BIC(N). In case of a Poisson process, we therefore recommend BIC(N) (simply denoted BIC in the following).

6.2 Thomas point process model

To generate a simulation from a Thomas point process with intensity (6.1), we first generate a parent point pattern from a stationary Poisson point process $\mathbf{C}$ with intensity $\kappa > 0$. Given $\mathbf{C}$, clusters $\mathbf{X}_c$, $c \in \mathbf{C}$, are generated from inhomogeneous Poisson point processes with intensity functions

$$\rho_c(u; \boldsymbol{\beta}) = \omega \exp\{\beta_1 z_1(u) + \cdots + \beta_6 z_6(u)\} k(u - c; \gamma)/\kappa,$$

where $k(u-c;\gamma) = (2\pi\gamma^2)^{-1}\exp(-\|u-c\|^2/(2\gamma^2))$ is the density for $\mathcal{N}(0, \gamma^2 \mathbf{I}_2)$. Finally, $\mathbf{X} = \bigcup_{c \in \mathbf{C}} \mathbf{X}_c$ is an inhomogeneous Thomas point process with intensity (6.1). The regression parameters are set as follows: $\beta_1 = 2$, $\beta_2 = -1$, $\beta_3 = \cdots = \beta_6 = 0$. We consider $\kappa = 4 \times 10^{-4}$ and two scale parameters $\gamma = 5$ and $\gamma = 15$. A lower value for γ tends to produce more clustered patterns. We consider the observation domains $W = [0, 500] \times [0, 250]$ and $W = [0, 1000] \times [0, 500]$ with ω adjusted to give expected numbers of points $\mu = 400$ and $\mu = 1600$ for the two windows. The chosen value of κ implies on average 50 parent points on $W = [0, 500] \times [0, 250]$ and 200 parent points on $W = [0, 1000] \times [0, 500]$.

The `R` function `rThomas` is used to simulate the point patterns and the function `kppm` to estimate the regression parameters. We use $m = 16\mu$ dummy points for the different integral approximations. Models are then selected using four criteria: AIC, BIC, CIC and CBIC. When using AIC or BIC we implicitly assume wrongly that the simulated patterns come from a Poisson point process and thus we set $p^*_{l,\text{approx}} = p_l$. For the composite likelihood type criteria CIC and CBIC, we compute $\hat{p}^*_{l,\text{approx}}$ using the `R` function `vcov.kppm`. The parameters κ and γ are estimated using minimum contrast estimation with the tuning parameter $r_{\max}$ (Waagepetersen and Guan, 2009) set to 20 when $\gamma = 5$ and to 50 when $\gamma = 15$.

Results are reported in Table 2. The AIC and BIC criteria ignore the second-order structure of the simulated point patterns and we observe that overall AIC and BIC produce high TPR but also very high FPR. The CIC and CBIC give much more reasonable trade-offs between TPR and FPR and also (with one exception) give smaller MKL and MISE than AIC and BIC.

Focusing on CIC and BIC, we first notice that the means of the estimates of $p^*_{l,\text{approx}}$ are high. This is because the considered clustered point processes are far from

Table 2: True positive (TPR) and false positive (FPR) rates in percent, MISE, MKL (divided by 10), and mean and standard deviations of estimates of $p^*_{l,\text{approx}}$, based on 500 simulations from inhomogeneous Thomas point processes with $\kappa = 4 \times 10^{-4}$ and scale parameter $\gamma = 5$ or 15, observed on different observation domains. Parameters are adjusted to have on average $\mu = 400$ points in the small window and 1600 points in the larger one. Model selection is based on AIC, BIC, CIC, and CBIC.

		AIC	BIC	CIC	CBIC
$W = [0, 500] \times [0, 250]$	TPR	98	96	88	81
$\mu = 400$	FPR	81	65	24	17
$\gamma = 5$	MISE	4.5	4.4	2.7	2.5
	MKL	9.7	9.6	7.5	9.1
	Mean($\hat{p}^*_l$)	—	—	103.6	76.6
	SD($\hat{p}^*_l$)	—	—	122.7	70.1
$W = [0, 1000] \times [0, 500]$	TPR	100	100	97	94
$\mu = 1600$	FPR	81	65	20	11
$\gamma = 5$	MISE	3.9	3.8	2.6	2.4
	MKL	10.3	10.2	7.9	8.2
	Mean($\hat{p}^*_l$)	—	—	101.1	81.3
	SD($\hat{p}^*_l$)	—	—	104.6	70.2
$W = [0, 500] \times [0, 250]$	TPR	100	99	95	93
$\mu = 400$	FPR	70	49	43	32
$\gamma = 15$	MISE	2.1	2.0	1.8	1.8
	MKL	5.2	5.0	4.9	5.4
	Mean($\hat{p}^*_l$)	—	—	48.9	40.9
	SD($\hat{p}^*_l$)	—	—	77.3	61.3
$W = [0, 1000] \times [0, 500]$	TPR	100	100	96	94
$\mu = 1600$	FPR	78	57	36	24
$\gamma = 15$	MISE	2.8	2.8	2.3	2.5
	MKL	7.7	7.5	7.3	9.3
	Mean($\hat{p}^*_l$)	—	—	96.1	78.1
	SD($\hat{p}^*_l$)	—	—	172.1	124.6

Poisson models. We also notice that even when the number of points is quite large, the estimation of $p^*_{l,\text{approx}}$ is inaccurate with standard deviations of the estimates of the same order as the means. Nevertheless, the resulting CIC and CBIC give reasonable results. Comparing CIC and CBIC, CIC in general has a higher FPR than CBIC but on the other hand always gives the smallest MKL. The TPR and MISE are quite similar for CIC and BIC. Hence in the case of the clustered point process considered here, CIC and CBIC clearly outperform AIC and BIC but there is not a clear winner between CIC and CBIC.

7 Discussion

In this paper we establish a theoretical foundation for various model selection criteria for inhomogeneous point processes under various asymptotic settings. In case of a Poisson process a main contribution is to identify in relation to BIC, the correct interpretation of 'sample size' which based on our theoretical derivation is the expected number of points which in practice is estimated by the observed number of points. This interpretation is supported by our simulation study which also supports the common understanding that BIC may be preferable to AIC which tends to pick too complex models.

More generally for selecting a regression model for the intensity function of a general point process we develop composite model selection criteria, CIC and CBIC that clearly outperform AIC and BIC in the simulation study for a clustered point process. One issue regarding CIC and BIC is to estimate the bias correction for the estimate of the composite Kullback-Leibler divergence which depends on the unknown true intensity function and pair correlation function. Here, inspired by the approach underlying AIC, for a given model we simply plug in the fitted intensity function and pair correlation function for the model in question. This is computationally convenient and we leave it as an open problem to develop more precise estimates.

Further interesting topics for future research would be to study the theoretical foundation of criteria for selecting models for the conditional intensity of a Gibbs process fitted by pseudo-likelihood. Another interesting problem is selection of the penalization parameter when the intensity function is estimated using regularization methods like the lasso (see e.g Thurman et al., 2015; Choiruddin et al., 2018). This is not covered by our theoretical results which rely on asymptotic results for unbiased estimating functions or Bayesian considerations.

Acknowledgments

The research of J.-F. Coeurjolly is supported by the Natural Sciences and Engineering Research Council. Rasmus Waagepetersen is supported by The Danish Council for Independent Research | Natural Sciences, grant DFF – 7014-00074 "Statistics for point processes in space and beyond", and by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by grant 8721 from the Villum Foundation.

A Proof of Theorem 1

Proof. (i) Using a zero order Taylor expansion of $e_n(\hat{\boldsymbol{\beta}}_{l,n})$ (which equals zero by definition) around $\boldsymbol{\beta}^*_{l,n}$ expressed with integral remainder term, we have

$$e_n(\boldsymbol{\beta}^*_{l,n}) = \left(\int_0^1 \int_{W_n} \theta_n \mathbf{z}_l(u) \mathbf{z}_l(u)^\top \exp[\{\boldsymbol{\beta}^*_{l,n} + t(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n})\}^\top \mathbf{z}_l(u)] \mathrm{d}u \mathrm{d}t \right) \\ \times (\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n}).$$

By rearranging the right-hand side of the latter equation and using Fubini's theorem, we have

$$
\begin{aligned}
e_n(\boldsymbol{\beta}_{l,n}^*) &= \theta_n \Big[\int_{W_n} \mathbf{z}_l(u) \exp\{{\boldsymbol{\beta}_{l,n}^*}^\top \mathbf{z}_l(u)\} \\
&\qquad \times \int_0^1 \{(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_{l,n}^*)^\top \mathbf{z}_l(u)\} \exp\{t(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_{l,n}^*)^\top \mathbf{z}_l(u)\} \mathrm{d}t \Big] \mathrm{d}u \\
&= \theta_n \int_{W_n} \mathbf{z}_l(u) \exp\{{\boldsymbol{\beta}_{l,n}^*}^\top \mathbf{z}_l(u)\} \big[\exp\{(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_{l,n}^*)^\top \mathbf{z}_l(u)\} - 1\big] \mathrm{d}u.
\end{aligned}
$$

Now, using C2–C3, we can apply a mean value theorem for multiple integrals: there exists $v = v_n(\hat{\boldsymbol{\beta}}_{l,n}, \boldsymbol{\beta}_{l,n}^*) \in W_n$ such that

$$
e_n(\boldsymbol{\beta}_{l,n}^*) = \tau_n \mathbf{z}_l(v) \exp\{{\boldsymbol{\beta}_{l,n}^*}^\top \mathbf{z}_l(v)\}[\exp\{(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_{l,n}^*)^\top \mathbf{z}_l(v)\} - 1]. \tag{A.1}
$$

Since $\|\mathbf{z}_l(v)\| > 0$ almost surely by condition C2, we deduce (3.9) from (A.1) using a little algebra. Let

$$
A_n = \tau_n^{-1} e_n(\boldsymbol{\beta}_{l,n}^*)^\top \frac{\mathbf{z}_l(v)}{\|\mathbf{z}_l(v)\|^2} \exp\{-{\boldsymbol{\beta}_{l,n}^*}^\top \mathbf{z}_l(v)\}.
$$

By conditions C1–C4, it is clear that $A_n \to 0$ almost surely as $n \to \infty$. Using the continuity of $t \mapsto \log(1+t)$ and again condition C2, we deduce that $\log(1 + A_n)\mathbf{z}_l(v)/\|\mathbf{z}_l(v)\|^2$ tends to 0 almost surely.

(ii) The proof consists in applying a modified version of Waagepetersen and Guan (2009, Theorem 2) to the sequence of estimating functions $e_n(\boldsymbol{\beta}_l)$. The modification is given in Appendix E and is needed to handle a sequence of 'least false' parameter values $\boldsymbol{\theta}_n^*$ instead of a unique 'true' value $\boldsymbol{\theta}^*$. Thus, we need to prove the assumptions G1–G4 in Appendix E. We leave the reader to check that conditions C3 and C5 imply conditions G1 and G2 with $\mathbf{V}_n = \sqrt{\tau_n}\mathbf{I}_p$ and $\mathbf{J}_n(\boldsymbol{\theta}_n^*) = \mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)$.

To prove assumption G3, we have to prove that as $n \to \infty$,

$$
\sup_{\sqrt{\tau_n}\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{l,n}^*\| \le d} \frac{\|\mathbf{S}_n(\boldsymbol{\beta}_l) - \mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)\|_M}{\tau_n} \to 0,
$$

where $\|\mathbf{M}\|_M = \max_{ij}|M_{ij}|$. Consider a $\boldsymbol{\beta}_l$ such that $\sqrt{\tau_n}\|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{l,n}^*\| \le d$. Under conditions C1–C3, there exist $K_1, K_2, K < \infty$ such that

$$
\begin{aligned}
\|\mathbf{S}_n(\boldsymbol{\beta}_l) - \mathbf{S}_n(\boldsymbol{\beta}_{l,n}^*)\|_M &= \left\| \theta_n \int_{W_n} \mathbf{z}_l(u)\mathbf{z}_l(u)^\top \left\{\rho(u;\boldsymbol{\beta}_l) - \rho(u;\boldsymbol{\beta}_{l,n}^*)\right\} \mathrm{d}u \right\|_M \\
&\le K_1\, \tau_n \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{l,n}^*\| \exp\{K_2(\|\boldsymbol{\beta}_{l,n}^*\| + d/\sqrt{\tau_n})\} \\
&\le K\tau_n \|\boldsymbol{\beta}_l - \boldsymbol{\beta}_{l,n}^*\| = \mathcal{O}(\sqrt{\tau_n}) = \mathcal{O}_{\mathrm{P}}(\sqrt{\tau_n}).
\end{aligned} \tag{A.2}
$$

To verify condition G4, it is sufficient to note that condition C6 implies that the variance of $\tau_n^{-1/2} e_n(\boldsymbol{\beta}_{l,n}^*)$ is bounded. Letting $\hat{\boldsymbol{\beta}}_{l,n}$ denote the (unique for n large enough) solution of $e_n(\boldsymbol{\beta}_l) = 0$ or, equivalently,

$$
\hat{\boldsymbol{\beta}}_{l,n} = \arg\max_{\boldsymbol{\beta}_l \in \mathbb{R}^{p_l}} \ell_n(\boldsymbol{\beta}_l)
$$

we can then conclude that $\hat{\boldsymbol{\beta}}_{l,n}$ is a root-τ_n consistent estimator of $\boldsymbol{\beta}^*_{l,n}$, that is, $\sqrt{\tau_n}(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n})$ is bounded in probability, which proves (3.10).

(iii) We use a Taylor expansion around $\boldsymbol{\beta}^*_{l,n}$: there exists $t \in (0,1)$ and $\tilde{\boldsymbol{\beta}}_{l,n} = \hat{\boldsymbol{\beta}}_{l,n} + t(\boldsymbol{\beta}^*_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})$ such that

$$\begin{aligned} e_n(\boldsymbol{\beta}^*_{l,n}) &= e_n(\boldsymbol{\beta}^*_{l,n}) - e_n(\hat{\boldsymbol{\beta}}_{l,n}) = \mathbf{S}_n(\tilde{\boldsymbol{\beta}}_{l,n})(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n}) \\ &= \mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n}) + \mathbf{B}_n \mathbf{S}'_n \end{aligned}$$

where $\mathbf{B}_n = \sqrt{\tau_n}(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n})$ and where $\mathbf{S}'_n = \tau_n^{-1/2}\{\mathbf{S}_n(\tilde{\boldsymbol{\beta}}_{l,n}) - \mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})\}$. We now show that $\mathbf{B}_n\mathbf{S}'_n = \mathcal{O}_{\mathrm{P}}(1)$. Let d and K be given as above and let $K_3 \geq d^2 K$. Using (A.2) we obtain

$$\begin{aligned} \mathrm{P}(\|\mathbf{B}_n\mathbf{S}'_n\|_M \geq K_3) &\leq \mathrm{P}(\|\mathbf{B}_n\mathbf{S}'_n\|_M \geq K_3, \|\mathbf{B}_n\| < d) + \mathrm{P}(\|\mathbf{B}_n\| \geq d) \\ &\leq \mathrm{P}(dK\|\mathbf{B}_n\| \geq K_3) + \mathrm{P}(\|\mathbf{B}_n\| \geq d) \leq 2P(\|\mathbf{B}_n\| \geq d) \end{aligned}$$

Now from (ii), for any $\varepsilon > 0$, by choosing d large enough, $P(\|\mathbf{B}_n\| \geq d) \leq \varepsilon/2$ for n sufficiently large whereby $\mathrm{P}(\|\mathbf{B}_n\mathbf{S}'_n\|_M \geq K_3) \leq \varepsilon$ for n sufficiently large.

Finally, since by condition C5, $\|\boldsymbol{\Sigma}_{l,n}^{-1/2}\| = \mathcal{O}(\tau_n^{-1/2})$, we conclude that

$$\begin{aligned} \boldsymbol{\Sigma}_{l,n}^{-1/2}\mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n}) &= \boldsymbol{\Sigma}_{l,n}^{-1/2} e_n(\boldsymbol{\beta}^*_{l,n}) + \boldsymbol{\Sigma}_{l,n}^{-1/2}\mathcal{O}_P(1) \\ &= \boldsymbol{\Sigma}_{l,n}^{-1/2} e_n(\boldsymbol{\beta}^*_{l,n}) + o_{\mathrm{P}}(1) \end{aligned}$$

which yields the result using condition C7 and Slutsky's lemma. □

B Conditions C4 and C7 for the inhomogeneous Poisson cluster point process

In this section, we show that the inhomogeneous Poisson cluster point process presented in the end of Section 3.2 satisfies C4 and C7. Recall $\lambda_n(u) = \theta_n \alpha \rho(u)$ where $\sup_u \rho(u) = \mathcal{O}(1)$.

Proof. For $k = 1, \dots, \lfloor \tau_n \rfloor$, let $\mathbf{C}_{n,k}$ be independent inhomogeneous Poisson point processes with intensity $\theta_n / \lfloor \tau_n \rfloor$. By the property of any Poisson point process, $\mathbf{C}_n$ has the same distribution as $\bigcup_k \mathbf{C}_{n,k}$. Define $\mathbf{A}_n = \sum_{u \in \mathbf{X}_n} \mathbf{z}_l(u)$. Then,

$$\mathbf{A}_n = \sum_{k=1}^{\lfloor \tau_n \rfloor} \mathbf{Z}_{n,k} \qquad \text{where } \mathbf{Z}_{n,k} = \sum_{c \in \mathbf{C}_{n,k}} \sum_{u \in \mathbf{X}_c} \mathbf{z}_l(u).$$

Using twice the Slivnyak-Mecke Theorem (see e.g. Theorem 3.1 in Møller and Waagepetersen, 2004),

$$\begin{aligned} \mathbb{E}\mathbf{Z}_{n,k} &= \lfloor \tau_n \rfloor^{-1} \int_{W_n} \mathbf{z}_l(u)\lambda_n(u)\mathrm{d}u \\ \mathbb{V}\mathrm{ar}\,\mathbf{Z}_{n,k} &= \lfloor \tau_n \rfloor^{-1} \bigg\{ \int_{W_n} \mathbf{z}_l(u)\mathbf{z}_l(u)^\top \lambda_n(u)\mathrm{d}u \\ &\quad + \int_{W_n}\int_{W_n} \mathbf{z}_l(u)\mathbf{z}_l(v)^\top \lambda_n(u)\lambda_n(v)\big(g_n(u,v) - 1\big)\mathrm{d}u\mathrm{d}v \bigg\} \end{aligned}$$

where $g_n(u,v) = 1 + \theta_n^{-1}(k*k)(v-u)$. Hence,

$$\mathbb{E}\mathbf{A}_n = \int_{W_n} \mathbf{z}_l(u)\lambda_n(u)\mathrm{d}u \quad \text{and} \quad \mathbb{V}\mathrm{ar}\,\mathbf{A}_n = \mathbf{\Sigma}_{l,n}.$$

Moreover, by (3.7),

$$\int_{W_n} \mathbf{z}_l(u)\lambda_n(u)\mathrm{d}u = \theta_n \int_{W_n} \mathbf{z}_l(u)\rho_n(u;\boldsymbol{\beta}_{l,n}^*)\mathrm{d}u$$

so that $e_n(\boldsymbol{\beta}_{l,n}^*) = \mathbf{A}_n - \mathbb{E}\mathbf{A}_n = \sum_{k=1}^{\lfloor \tau_n \rfloor} \mathring{\mathbf{Z}}_{n,k}$ with $\mathring{\mathbf{Z}}_{n,k} = \mathbf{Z}_{n,k} - \mathbb{E}\mathbf{Z}_{n,k}$. Since C6 is satisfied as shown in Example 2 it follows that the elements of $\mathbb{V}\mathrm{ar}(\mathring{\mathbf{Z}}_{n,k})$ are $\mathcal{O}(1)$ whereby $\sum_{k\ge 1} \mathbb{V}\mathrm{ar}(\mathring{\mathbf{Z}}_{n,k})/k^2 < \infty$ (elementwise). Therefore, we can apply Kolmogorov's strong law of large numbers to establish that $\lfloor \tau_n \rfloor^{-1} e_n(\boldsymbol{\beta}_{l,n}^*)$ tends almost surely to 0. Using the same conditions and C5, we can also apply the Lindeberg-Feller theorem and obtain that as $n \to \infty$, $\mathbf{\Sigma}_{l,n}^{-1/2} e_n(\boldsymbol{\beta}_{l,n}^*) \to N(0, \mathbf{I}_{p_l})$ in distribution. □

C Proof of Proposition 3

For a multi-index $\alpha \in \mathbb{N}^p$ with cardinality $|\alpha| = \alpha_1 + \cdots + \alpha_p$ and a $|\alpha|$ times differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ we define for $u \in \mathbb{R}^p$,

$$\partial^\alpha f(u) = \frac{\partial^{|\alpha|} f(u)}{\partial u_1^{\alpha_1} \cdots \partial u_p^{\alpha_p}}.$$

and we also use the notation $u^\alpha = u_1^{\alpha_1} \cdots u_p^{\alpha_p}$.

Proof. We follow the sketch of the proof of Varin and Vidoni (2005, Lemmas 1–2). Let $\mathbf{Z}_n$ be an independent copy of $\mathbf{X}_n$. Let $\ell_n(\boldsymbol{\beta}_l, \mathbf{Z}_n)$ and $e_n(\boldsymbol{\beta}_l; \mathbf{Z}_n)$ be the composite likelihood and its corresponding estimating equation evaluated at $\mathbf{Z}_n$. And we remind the notation $\ell_n(\boldsymbol{\beta}_l) = \ell_n(\boldsymbol{\beta}_l; \mathbf{X}_n)$, $C_n(\boldsymbol{\beta}_l) = -\mathbb{E}\ell_n(\boldsymbol{\beta}_l)$, $e_n(\boldsymbol{\beta}_l) = e_n(\boldsymbol{\beta}_l; \mathbf{X}_n)$ and $\hat{\boldsymbol{\beta}}_{l,n} = \mathrm{argmax}_{\boldsymbol{\beta}_l} \ell_n(\boldsymbol{\beta}_l)$. We have

$$\mathbb{E}C_n(\hat{\boldsymbol{\beta}}_{l,n}) = -\mathbb{E}\big[\mathbb{E}\{\ell_n(\hat{\boldsymbol{\beta}}_{l,n}; \mathbf{Z}_n)\} \bigm| \mathbf{X}_n\big].$$

Using a first order Taylor expansion of $\ell_n(\,\cdot\,; \mathbf{Z}_n)$ around $\boldsymbol{\beta}_{l,n}^*$ with integral remainder term, we have

$$\ell_n(\hat{\boldsymbol{\beta}}_{l,n}; \mathbf{Z}_n) = \ell_n(\boldsymbol{\beta}_{l,n}^*; \mathbf{Z}_n) + e_n(\boldsymbol{\beta}_{l,n}^*; \mathbf{Z}_n)^\top \mathbf{\Delta}_n + \mathbf{R}_n(\boldsymbol{\beta}_{l,n}^*, \hat{\boldsymbol{\beta}}_{l,n}; \mathbf{Z}_n)$$

where $\mathbf{\Delta}_n = \hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_{l,n}^*$ and where the integral remainder term can be expressed as

$$\begin{aligned}\mathbf{R}_n(\boldsymbol{\beta}_{l,n}^*, \hat{\boldsymbol{\beta}}_{l,n}; \mathbf{Z}_n) =& 2 \sum_{\alpha \in \mathbb{N}^{p_l}, |\alpha|=2} \Bigg[\frac{(\boldsymbol{\beta}_{l,n}^* - \hat{\boldsymbol{\beta}}_{l,n})^\alpha}{\alpha!} \\ & \times \int_0^1 (1-t)\partial^\alpha \ell_n \left\{ \hat{\boldsymbol{\beta}}_{l,n} + t(\boldsymbol{\beta}_{l,n}^* - \hat{\boldsymbol{\beta}}_{l,n}); \mathbf{Z}_n \right\} \mathrm{d}t \Bigg].\end{aligned}$$

It is worth noticing that for any $\boldsymbol{\beta}_l \in \mathbb{R}^{p_l}$ and any $\alpha \in \mathbb{N}^{p_l}$ such that $|\alpha| = 2$, $\partial^\alpha \ell_n(\boldsymbol{\beta}_l; \mathbf{Z}_n)$ is a deterministic function of $\boldsymbol{\beta}_l$. Therefore

$$\mathbf{R}_n(\boldsymbol{\beta}^*_{l,n}, \hat{\boldsymbol{\beta}}_{l,n}) := \mathbf{R}_n(\boldsymbol{\beta}^*_{l,n}, \hat{\boldsymbol{\beta}}_{l,n}; \mathbf{Z}_n)$$

does not depend on $\mathbf{Z}_n$. Using this and the unbiasedness of $e_n(\boldsymbol{\beta}^*_{l,n};\,\cdot\,)$ we have

$$\mathbb{E}\{\ell_n(\hat{\boldsymbol{\beta}}_{l,n}; \mathbf{Z}_n) \mid \mathbf{X}_n\} = \mathbb{E}\{\ell_n(\boldsymbol{\beta}^*_{l,n}; \mathbf{Z}_n)\} + \mathbf{R}_n(\boldsymbol{\beta}^*_{l,n}, \hat{\boldsymbol{\beta}}_{l,n})$$

Hence,

$$\mathbb{E}\{C_n(\hat{\boldsymbol{\beta}}_{l,n})\} = -\mathbb{E}\{\ell_n(\boldsymbol{\beta}^*_{l,n})\} - \mathbb{E}\{\mathbf{R}_n(\boldsymbol{\beta}^*_{l,n}, \hat{\boldsymbol{\beta}}_{l,n})\}. \tag{C.1}$$

Now, using a first order Taylor expansion of $-\ell_n(\,\cdot\,)$ around $\boldsymbol{\beta}^*_{l,n}$, we have

$$-\ell_n(\hat{\boldsymbol{\beta}}_{l,n}) = -\ell_n(\boldsymbol{\beta}^*_{l,n}) - e_n(\boldsymbol{\beta}^*_{l,n})^\top \boldsymbol{\Delta}_n - \mathbf{R}_n(\boldsymbol{\beta}^*_{l,n}, \hat{\boldsymbol{\beta}}_{l,n}). \tag{C.2}$$

Combining (C.1) and the expectation of (C.2) we obtain

$$\mathbb{E}\{C_n(\hat{\boldsymbol{\beta}}_{l,n})\} - \mathbb{E}\{-\ell_n(\hat{\boldsymbol{\beta}}_{l,n})\} = \mathbb{E}\{e_n(\boldsymbol{\beta}^*_{l,n})^\top \boldsymbol{\Delta}_n\}.$$

Since by definition of $\hat{\boldsymbol{\beta}}_{l,n}$, $\boldsymbol{\Delta}_n = \mathbf{S}_n^{-1}(\tilde{\boldsymbol{\beta}}_{l,n}) e_n(\boldsymbol{\beta}^*_{l,n})$ where $\tilde{\boldsymbol{\beta}}_{l,n} = \hat{\boldsymbol{\beta}}_{l,n} + t(\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}^*_{l,n})$ for some $t \in (0,1)$ we continue with

$$\begin{aligned}
&\mathbb{E}\{C_n(\hat{\boldsymbol{\beta}}_{l,n})\} - \mathbb{E}\{-\ell_n(\hat{\boldsymbol{\beta}}_{l,n})\} \\
&\quad = \mathbb{E}\{e_n(\boldsymbol{\beta}^*_{l,n})^\top \mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})^{-1} e_n(\boldsymbol{\beta}^*_{l,n})\} + \mathbb{E}\{e_n(\boldsymbol{\beta}^*_{l,n})^\top \mathbf{M}_n e_n(\boldsymbol{\beta}^*_{l,n})\} \\
&\quad = \operatorname{trace}\{\mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})^{-1} \boldsymbol{\Sigma}_{l,n}\} + \mathbb{E}\{e_n(\boldsymbol{\beta}^*_{l,n})^\top \mathbf{M}_n e_n(\boldsymbol{\beta}^*_{l,n})\}
\end{aligned}$$

where $\mathbf{M}_n = \mathbf{S}_n(\tilde{\boldsymbol{\beta}}_{l,n})^{-1} - \mathbf{S}_n(\boldsymbol{\beta}^*_{l,n})^{-1}$. Since $e_n(\boldsymbol{\beta}^*_{l,n})/\sqrt{\tau_n}$ is bounded in probability by C6 and $\tau_n \mathbf{M}_n$ converges in probability since $\boldsymbol{\Delta}_n$ converges to zero in probability, we obtain $e_n(\boldsymbol{\beta}^*_{l,n})^\top \mathbf{M}_n e_n(\boldsymbol{\beta}^*_{l,n})$ converges in probability. This combined with uniform integrability (4.1) implies $\mathbb{E}\{e_n(\boldsymbol{\beta}^*_{l,n})^\top \mathbf{M}_n e_n(\boldsymbol{\beta}^*_{l,n})\} = o(1)$. ☐

D Proof of Proposition 5

Proof. We remind the notation $\tau_n = \theta_n |W_n|$.

$$\begin{aligned}
p_n(\mathbf{X}_n \mid \mathcal{M}_l) &= \int_{\mathbb{R}^{p_l}} p_n(\mathbf{X}_n, \boldsymbol{\beta}_l \mid \mathcal{M}_l) \mathrm{d}\boldsymbol{\beta}_l \\
&= \int_{\mathbb{R}^{p_l}} p_n(\mathbf{X}_n; \boldsymbol{\beta}_l) p(\boldsymbol{\beta}_l \mid \mathcal{M}_l) \mathrm{d}\beta_l \\
&= \int_{\mathbb{R}^{p_l}} p(\boldsymbol{\beta}_l \mid \mathcal{M}_l) \exp\{\tau_n\, \ell_n(\boldsymbol{\beta}_l)/\tau_n\} \,\mathrm{d}\boldsymbol{\beta}_l.
\end{aligned} \tag{D.1}$$

Now, we are in the situation where we can apply Łapiński (2019, Theorem 2), which gives rigorous conditions under which a multivariate Laplace approximation holds. First, condition C8 ensures that $p(\boldsymbol{\beta}_l \mid \mathcal{M}_l)$ is regular enough. Second, condition C5 ensures that $\tau_n \ell_n(\boldsymbol{\beta}_l)$ has a nonsingular Hessian matrix for any $\boldsymbol{\beta}_l$. Third, for $\alpha \in \mathbb{N}^{p_l}$ with $|\alpha| \leq 3$, we have for any $\boldsymbol{\beta}_l$ (since $z_{0l}^{\alpha_0} = 1^{\alpha_0} = 1$)

$$\partial^\alpha \{\tau_n^{-1} \log p_n(\mathbf{X}_n; \boldsymbol{\beta}_l)\} = -\frac{1}{|W_n|} \int_{W_n} \prod_{j \in I_l} z_j^{\alpha_j}(u) \rho(u; \boldsymbol{\beta}_l) \mathrm{d}u.$$

Therefore, under condition C2, $\partial^\alpha \{\tau_n^{-1} \log p_n(\mathbf{X}_n; \boldsymbol{\beta}_l)\}$ is uniformly bounded for $\boldsymbol{\beta}_l$ in any compact subset of $\mathbb{R}^d$.

The fact that $\int p(\boldsymbol{\beta}_l \mid \mathcal{M}_l) \mathrm{d}\boldsymbol{\beta}_l = 1 < \infty$ is sufficient to ensure Łapiński (2019, condition (5)). Finally, by the strong consistency of $\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_{l,n}^*$ to 0 and from condition C1$'$, the last condition to check is formulated as follows: we need to verify that there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, $\tau_n^{-1} \ell_n(\boldsymbol{\beta}_l)$ has a unique maximum in a closed ball $B(\boldsymbol{\beta}_l^*, \varepsilon)$ for some $\varepsilon > 0$ where $\boldsymbol{\beta}_l^*$ is given by condition C1$'$) and such that

$$\Delta = \inf_{n \geq n_0, \boldsymbol{\beta}_l \in \mathbb{R}^{p_l} \setminus B(\boldsymbol{\beta}_l^*, \varepsilon)} \tau_n^{-1} \{\ell_n(\hat{\boldsymbol{\beta}}_{l,n}) - \ell_n(\boldsymbol{\beta}_l)\} > 0. \tag{D.2}$$

By condition C5, for any $n \geq 1$, ℓ_n (and thus $\tau_n^{-1} \ell_n$) has indeed a unique maximum. Letting $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that almost surely $\hat{\boldsymbol{\beta}}_{l,n} \in B(\boldsymbol{\beta}_l^*, \varepsilon/2)$. Consider Δ with such n_0 and ε. Note that for large n and any $\boldsymbol{\beta}_l \in \mathbb{R}^{p_l} \setminus B(\boldsymbol{\beta}_l^*, \varepsilon)$, $\|\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_l\| \geq \varepsilon/2$ almost surely.

Using a first order Taylor expansion around $\hat{\boldsymbol{\beta}}_{l,n}$ and using the fact that $e_n(\hat{\boldsymbol{\beta}}_{l,n}) = 0$, we have

$$\tau_n^{-1} \{\ell_n(\boldsymbol{\beta}_l) - \ell_n(\hat{\boldsymbol{\beta}}_{l,n})\} = R_n(\hat{\boldsymbol{\beta}}_{l,n}, \boldsymbol{\beta}_{l,n})$$

where the remainder term $R_n(\hat{\boldsymbol{\beta}}_{l,n}, \boldsymbol{\beta}_{l,n})$ is

$$\begin{aligned}
&\frac{2}{\tau_n} \sum_{\alpha \in \mathbb{N}^{p_l}, |\alpha| = 2} \left[\frac{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\alpha}{\alpha!} \times \int_0^1 (1-t) \partial^\alpha \ell_n\{\hat{\boldsymbol{\beta}}_{l,n} + t(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n}); \mathbf{X}_n)\} \mathrm{d}t \right] \\
&\quad = \frac{-2}{|W_n|} \sum_{\alpha \in \mathbb{N}^{p_l}, |\alpha| = 2} \left[\frac{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\alpha}{\alpha!} \right. \\
&\qquad \left. \times \int_0^1 (1-t) \int_{W_n} \mathbf{z}^\alpha(u) \rho(u; \hat{\boldsymbol{\beta}}_{l,n} + t(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})) \mathrm{d}u \mathrm{d}t \right] \\
&\quad = \frac{-2}{|W_n|} \int_{W_n} \sum_{\alpha \in \mathbb{N}^{p_l}, |\alpha| = 2} \frac{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\alpha}{\alpha!} \mathbf{z}^\alpha(u) \\
&\qquad \times \exp\{\hat{\boldsymbol{\beta}}_{l,n} \mathbf{z}_l(u)\} \frac{\exp\{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\top \mathbf{z}_l(u)\} - (\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\top \mathbf{z}_l(u) - 1}{\{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\top \mathbf{z}_l(u)\}^2} \mathrm{d}u \\
&\quad = \frac{-1}{|W_n|} \int_{W_n} \exp\{\hat{\boldsymbol{\beta}}_{l,n}^\top \mathbf{z}_l(u)\} f\{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\top \mathbf{z}_l(u)\} \mathrm{d}u
\end{aligned}$$

where $f(t) = \exp(t) - 1 - t$. Consider the set B_n given by C5. Since for large n, $\|\hat{\boldsymbol{\beta}}_{l,n} - \boldsymbol{\beta}_l^*\| \geq \varepsilon/2$ and since f is non negative and has a unique minimum at zero, we have by condition C5 that

$$\inf_{u \in B_n} \exp\{\hat{\boldsymbol{\beta}}_{l,n}^\top \mathbf{z}_l(u)\} > \exp(-\|\hat{\boldsymbol{\beta}}_{l,n}\| c)$$

and

$$\inf_{u \in B_n} f\{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\top \mathbf{z}_l(u)\} \geq \delta > 0$$

for some $\delta > 0$. Therefore,

$$\begin{aligned} R_n(\hat{\boldsymbol{\beta}}_{l,n}, \boldsymbol{\beta}_{l,n}) &\leq \frac{-1}{|W_n|} \int_{B_n} \exp\{\hat{\boldsymbol{\beta}}_{l,n}^\top \mathbf{z}_l(u)\} f\{(\boldsymbol{\beta}_{l,n} - \hat{\boldsymbol{\beta}}_{l,n})^\top \mathbf{z}_l(u)\} \mathrm{d}u \\ &\leq -\delta \frac{|B_n|}{|W_n|} \exp(-\|\hat{\boldsymbol{\beta}}_{l,n}\| c). \end{aligned}$$

Again, by definition of B_n, the strong consistency of $\hat{\boldsymbol{\beta}}_{l,n}$ and condition C2, we conclude that for large n there exists $\delta' > 0$ such that $R_n(\hat{\boldsymbol{\beta}}_{l,n}, \boldsymbol{\beta}_{l,n}) \leq -\delta' < 0$ whereby we deduce that $\Delta > 0$.

The previous statements allow us to conclude by Łapiński (2019, Theorem 2) that

$$\begin{aligned} p_n(\mathbf{X}_n \mid \mathcal{M}_l) = &\frac{p(\hat{\boldsymbol{\beta}}_{l,n} \mid \mathcal{M}_l)}{\det\{\tau_n^{-1} \mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})\}^{1/2}} \exp\{\ell_n(\hat{\boldsymbol{\beta}}_{l,n})\} \left(\frac{2\pi}{\tau_n}\right)^{p_l/2} \\ &+ \exp\{\ell_n(\hat{\boldsymbol{\beta}}_{l,n})\} \left(\frac{2\pi}{\tau_n}\right)^{p_l/2} \mathcal{O}(\tau_n^{-1/2}). \end{aligned}$$

This gives

$$\begin{aligned} &\log p_n(\mathbf{X}_n \mid \mathcal{M}_l) - \log \frac{p(\hat{\boldsymbol{\beta}}_{l,n} \mid \mathcal{M}_l)}{\det\{\tau_n^{-1} \mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})\}^{1/2}} - \ell_n(\hat{\boldsymbol{\beta}}_{l,n}) - \frac{p_l}{2} \log \frac{2\pi}{\tau_n} \\ &= \log\Big[1 + \frac{\det\{\tau_n^{-1} \mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})\}^{1/2}}{p(\hat{\boldsymbol{\beta}}_{l,n} \mid \mathcal{M}_l)} \mathcal{O}(\tau_n^{-1/2})\Big]. \end{aligned}$$

Proposition 5 now follows since

$$\log\Big[1 + \frac{\det\{\tau_n^{-1} \mathbf{S}_n(\hat{\boldsymbol{\beta}}_{l,n})\}^{1/2}}{p(\hat{\boldsymbol{\beta}}_{l,n} \mid \mathcal{M}_l)} \mathcal{O}(\tau_n^{-1/2})\Big] = \log\{1 + \mathcal{O}(\tau_n^{-1/2})\} = \mathcal{O}(\tau_n^{-1/2})$$

and using condition C9. □

E Modified version of Theorem 2 in Waagepetersen and Guan (2009)

Consider a sequence of estimating functions $u_n : \mathbb{R}^p \to \mathbb{R}^p$, $n \geq 1$ whose distribution is determined by some underlying probability measure generating the data at hand. For a matrix $\mathbf{A} = [a_{ij}]$, $\|\mathbf{A}\|_M = \max_{ij} |a_{ij}|$, and we let $\mathbf{J}_n(\boldsymbol{\theta}) = -\frac{\mathrm{d}}{\mathrm{d}\theta} u_n(\boldsymbol{\theta})$, assuming that u_n is differentiable.

Theorem 6. *Assume that there exists a sequence of invertible symmetric matrices V_n and a sequence of parameter values $\boldsymbol{\theta}_n^* \in \mathbb{R}^p$ such that*

[G1] $\|\mathbf{V}_n^{-1}\| \to 0$.

[G2] *There exists an* $l > 0$ *so that* $\mathrm{P}(l_n < l)$ *tends to zero where*

$$l_n = \inf_{\|\boldsymbol{\phi}\|=1} \boldsymbol{\phi}^\top \mathbf{V}_n^{-1} \mathbf{J}_n(\boldsymbol{\theta}_n^*) \mathbf{V}_n^{-1} \boldsymbol{\phi}.$$

[G3] *For any* $d > 0$,

$$\sup_{\|\mathbf{V}_n(\boldsymbol{\theta}-\boldsymbol{\theta}_n^*)\| \le d} \|\mathbf{V}_n^{-1}\{\mathbf{J}_n(\boldsymbol{\theta}) - \mathbf{J}_n(\boldsymbol{\theta}_n^*)\}\mathbf{V}_n^{-1}\|_M = \gamma_{nd} \to 0$$

in probability under P.

[G4] *The sequence* $u_n(\boldsymbol{\theta}_n^*)\mathbf{V}_n^{-1}$ *is bounded in probability (i.e. for each* $\epsilon > 0$ *there exists a* d *so that* $\mathrm{P}(\|\mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^*)\| > d) \le \epsilon$ *for* n *sufficiently large).*

Then for each $\epsilon > 0$, *there exists a* $d > 0$ *such that*

$$\mathrm{P}\big\{ \exists \tilde{\boldsymbol{\theta}}_n : u_n(\tilde{\boldsymbol{\theta}}_n) = 0 \text{ and } \|\mathbf{V}_n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*)\| < d \big\} > 1 - \epsilon \tag{E.1}$$

whenever n *is sufficiently large.*

Proof. The event

$$\big\{ \exists \tilde{\boldsymbol{\theta}}_n : u_n(\tilde{\boldsymbol{\theta}}_n) = 0 \text{ and } \|\mathbf{V}_n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*)\| < d \big\}$$

occurs if $\boldsymbol{\phi}^\top \mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^* + \mathbf{V}_n^{-1}\boldsymbol{\phi}) < 0$ for all $\boldsymbol{\phi}$ with $\|\boldsymbol{\phi}\| = d$ since this implies $u_n(\boldsymbol{\theta}_n^* + \mathbf{V}_n^{-1}\boldsymbol{\phi})$ for some $\|\boldsymbol{\phi}\| < d$ (Lemma 2 in Aitchison and Silvey, 1958). Hence we need to show that there is a d such that

$$\mathrm{P}\big\{ \sup_{\|\boldsymbol{\phi}\|=d} \boldsymbol{\phi}^\top \mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^* + \mathbf{V}_n^{-1}\boldsymbol{\phi}) \ge 0 \big\} \le \epsilon$$

for sufficiently large n. To this end we write

$$\boldsymbol{\phi}^\top \mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^* + \mathbf{V}_n^{-1}\boldsymbol{\phi}) = \boldsymbol{\phi}^\top \mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^*) - \boldsymbol{\phi}^\top \int_0^1 \mathbf{V}_n^{-1} \mathbf{J}_n(\boldsymbol{\theta}_n(t)) \mathbf{V}_n^{-1} \mathrm{d}t \boldsymbol{\phi}$$

where $\boldsymbol{\theta}(t) = \boldsymbol{\theta}_n^* + t\mathbf{V}_n^{-1}\boldsymbol{\phi}$. Then

$$\begin{aligned}
&\mathrm{P}\big\{ \sup_{\|\boldsymbol{\phi}\|=d} \boldsymbol{\phi}^\top \mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^* + \mathbf{V}_n^{-1}\boldsymbol{\phi}) \ge 0 \big\} \\
&\quad \le \mathrm{P}\Big\{ \sup_{\|\boldsymbol{\phi}\|=d} \boldsymbol{\phi}^\top \mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^*) \ge \inf_{\|\boldsymbol{\phi}\|=d} \boldsymbol{\phi}^\top \int_0^1 \mathbf{V}_n^{-1} \mathbf{J}_n(\boldsymbol{\theta}_n(t)) \mathbf{V}_n^{-1} \mathrm{d}t \, \boldsymbol{\phi} \Big\} \\
&\quad \le \mathrm{P}\big[\|\mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^*)\| \ge d \inf_{\|\boldsymbol{\phi}\|=1} \{\boldsymbol{\phi}^\top \mathbf{V}_n^{-1} \mathbf{J}_n(\boldsymbol{\theta}_n^*) \mathbf{V}_n^{-1} \boldsymbol{\phi}\} - dp\gamma_{nd} \big] \\
&\quad \le \mathrm{P}(\|\mathbf{V}_n^{-1} u_n(\boldsymbol{\theta}_n^*)\| \ge dl_n/2) + \mathrm{P}(p\gamma_{nd} > l_n/2).
\end{aligned}$$

The first term can be made arbitrarily small by picking a sufficiently large d and letting n tend to infinity. The second term converges to zero as n tends to infinity. $\square$

References

J. Aitchison and S. D. Silvey. Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 29:813–825, 1958.

H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

A. Baddeley and R. Turner. Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3):283–322, 2000.

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

M. Berman and R. Turner. Approximating point process likelihoods with GLIM. *Applied Statistics*, 41:31–38, 1992.

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

G. Claeskens and N.L. Hjort. *Model Selection and Model Averaging.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2008.

J.-F. Coeurjolly and F. Lavancier. Understanding spatial point patterns through intensity and conditional intensities. In *Stochastic Geometry*, number 2237 in Lecture Notes in Mathematics, chapter 2. Springer Verlag, 2019.

R. Condit. *Tropical Forest Census Plots.* Springer-Verlag and R. G. Landes Company, Berlin, Germany and Georgetown, Texas, 1998.

R. Condit, S.P. Hubbell, and R.B. Foster. Changes in tree species abundance in a neotropical forest: impact of climate change. *Journal of tropical ecology*, 12(2):231–256, 1996.

X. Gao and P. X.-K. Song. Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 1-5: 1531–1540, 2010.

Y. Guan and J. M. Loh. A thinned block bootstrap procedure for modeling inhomogeneous spatial point patterns. *Journal of the American Statistical Association*, 102:1377–1386, 2007.

S. P. Hubbell and R. B. Foster. Diversity of canopy trees in a neotropical forest and implications for conservation. In S. L. Sutton, T. C. Whitmore, and A. C. Chadwick, editors, *Tropical Rain Forest: Ecology and Management*, pages 25–41. Blackwell Scientific Publications, Oxford, 1983.

D. Jeffrey, J. Horrocks, and Umphrey G.J. Penalized composite likelihoods for inhomogeneous Gibbs point process models. *Computational Statistics & Data Analysis*, 124: 104–116, 2018.

T. M. Łapiński. Multivariate Laplace's approximation with estimated error and application to limit theorems. *Journal of Approximation Theory*, 248:105305, 2019.

F. Lavancier, A. Poinas, and R. Waagepetersen. Adaptive estimating function inference for non-stationary determinantal point processes. *Scandinavian Journal of Statistics*, 2020. Appeared online.

E. Lebarbier and T. Mary-Huard. Une introduction au critère BIC: fondements théoriques et interprétation. *Journal de la Société Française de Statistique 1 (147), 39-58*, 2006.

J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton, 2004.

J. Møller and R. P. Waagepetersen. Some recent developments in statistics for spatial point processes. *Annual review of Statistics and its Applications*, 4:317–342, 2017.

S. L. Rathbun and N. Cressie. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability*, 26: 122–154, 1994.

F.P. Schoenberg. Consistent parametric estimation of the intensity of a spatial-temporal point process. *Journal of Statistical Planning and Inference*, 128:79–93, 2005.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 03 1978.

A. L. Thurman, R. Fu, Y. Guan, and J. Zhu. Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica*, pages 173–188, 2015.

L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005.

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

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