ESAIM: PROCEEDINGS AND SURVEYS, February 2025, Vol. 80, p. 2-16
Editors: Pierre Calka, Clément Marteau - Journées MAS 2022, Université de Rouen

REGULARIZATION TECHNIQUES FOR INHOMOGENEOUS (SPATIAL) POINT PROCESSES INTENSITY AND CONDITIONAL INTENSITY ESTIMATION *, **, ***

Jean-Francois Coeurjolly[1], Ismaïla Ba[2] and Achmad Choiruddin[3]

Abstract. Point processes are stochastic models generating interacting points or events in time and/or space. Among characteristics of these models, first-order intensity and conditional intensity functions are often considered. We focus on inhomogeneous parametric forms of these functions assumed to depend on a certain number of spatial covariates. When this number of covariates is large, we are faced with a high-dimensional problem. This paper provides an overview of these questions and existing solutions based on regularizations.

Résumé. Les processus ponctuels constituent une classe de modèles stochastiques permettant de modéliser des évènements dans le temps et/ou l'espace en interaction. Parmi les caractéristiques d'un processus ponctuel, l'intensité et l'intensité conditionnelle d'ordre un sont souvent considérées. Nous nous concentrons ici sur des formes paramétriques inhomogènes de ces fonctions que nous supposons dépendre d'un certain nombre de covariables spatiales. Lorsque ce nombre est élevé, nous faisons face à un problème de grande dimension. Ce papier a pour objectif de présenter un aperçu de ces problèmes et solutions existantes.

1. Introduction

Spatial point processes are stochastic processes which model point patterns distributed in a space say S (usually a subset of $\mathbb{R}^d$), such as locations of crime events, species of trees, earthquake occurrences, disease cases (see e.g. [3, 29]). Modeling and inferring the intensity or conditional intensity of a spatial point process often constitutes the first and important task in the description and analysis of a spatial point pattern [12]. Roughly speaking, the intensity function measures the probability of an event to occur at a specific location, say $u \in S$, while the conditional intensity measures the probability to observe a point at u given $\mathbf{x}$ the observed set of events (i.e. points). This paper is focused on inhomogeneous models and in particular on parametric (conditional) intensity models for which intensities can be explained by covariates.

To illustrate this, let us consider the motivating example in dimension $d = 2$. This example, also described in Section 2.3, is depicted in Figure 1. The point pattern corresponds to the locations of 3605 trees in a tropical

* *JF Coeurjolly is supported by Labex PERSYVAL-lab ANR-11-LABX-0025*

** *A Choiruddin is supported by Grant 1961/PKS/ITS/2023 from Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia*

*** *JF Coeurjolly thanks all the organizers of the "Journées MAS 2022" and the members of the group MAS from SMAI for the opportunity offered to us to prepare this short review/tutorial paper.*

[1] Univ. Grenoble Alpes, LJK, 38000 Grenoble, France

[2] Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada

[3] Department of Statistics, Institut Teknologi Sepuluh Nopember, 60111 Surabaya, Indonesia

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202580002

forest (see [3]). It appears to be highly inhomogeneous in space and probably also clustered. In addition to the locations of trees, we also observe spatial information available on the whole observation domain. These covariates are for instance the altitude map, the slope of elevation map, and level maps of some soil nutrients. It makes sense to consider a parametric model of the intensity or conditional intensity (see Section 2.3 and in particular Equation (3) for a typical exponential family model) that depends of this set of covariates with cardinality close to 100 in this application. Among these covariates, many of them are correlated and probably only a few of them are informative. The objective is therefore to estimate (in particular) a parameter vector of length p (≈ 100 for this example) where probably only some of its components are non-zero. This number, say s, is (of course) unknown.

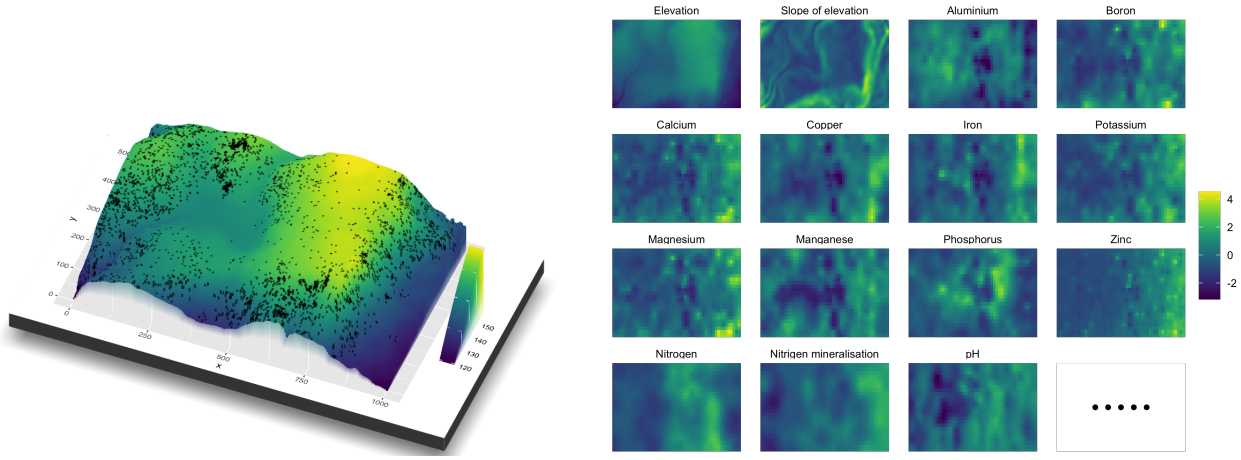

FIGURE 1. Left: locations of Beilschmiedia pendula trees in Barro Colorado Island (dataset obtained from the Center of Tropical Forest) represented with the ground topography; Right: some of available spatial covariates (observed on the same observation domain) such as the altitude, slope of elevation and values of soil nutrients (e.g. Aluminium, Boron).

The above phenomenon and problem arise in many various fields in statistics (see e.g. [24,27]). When it comes to spatial point patterns, methodology based on regularization has been developed for intensity/conditional intensity estimation. The idea is to consider a standard composite likelihood (e.g. Poisson likelihood or pseudo-likelihood, see Section 3.1 for more details) and to add a penalty or regularization term with the objective to estimate parameters and perform features selection at the same time. The literature on this large topic is quite recent and covers methodological, theoretical and computational aspects [1, 6, 7, 9, 16, 36, 37].

Most often, in the literature (e.g. [3, 29, 33]), estimating the intensity or conditional intensity function are two questions which are treated separately. This was mainly justified by the fact that the object to model, mathematical tools (Campbell or Georgii-Nguyen-Zessin (GNZ) equations, see Equations (1)-(2)), statistical methodologies and proofs appear, at first glance, really different. The contribution of this paper is to make a short overview of regularization techniques applied to point processes by trying, as far as possible, to present the two problems (intensity or conditional intensity estimation) in a similar way in order to shed the light on their similarities. Doing this allows us to consider a very large class of models like Poisson, Cox point processes, Determinantal point processes (DPP), Gibbs processes (see Table 2.2), that is models for which the modelling of either the intensity or conditional intensity function is very natural.

In order to present methodologies and setting in a similar way and with the objective to summarize the existing literature, some choices are necessary. For instance, even if more general regularization techniques and asymptotic regimes were considered in [1, 6, 7], we present only a part of asymptotic results and for one type of penalty, namely for the adaptive lasso. However, the theoretical contribution (see Theorem 2) of this paper is to propose an extension of a consistency result (consistency of the estimator and oracle properties) under a

setting which extends previous ones: we assume that the mean number of points is a sequence indexed by, say n, which tends to infinity as $n \to \infty$. This includes two standard settings in spatial statistics, namely increasing domain asymptotics or infill asymptotics. Moreover, we assume that both p and s can tend to infinity with n and we allow the regularization parameters to be stochastic. Letting p and s increase with n is quite standard when we deal with high-dimensional statistics (see e.g. [21]). Finally, the fact that regularization parameters can be stochastic is of significant practical interest as it is often the case to select these tuning parameters using the same data used to estimate the parameters [27] (see the end of Section 4 for a discussion on this point). The rest of the paper is organized as follows. Background on spatial point process is described in Section 2. We detail the statistical inference, theoretical results, and numerical aspects in Sections 3-4.

2. Spatial point processes

2.1. Notation and intensity functions

We consider spatial point processes in $\mathbb{R}^d$. For ease of exposition, we view a point process as a random locally finite subset $\mathbf{X}$ of a Borel set $S \subseteq \mathbb{R}^d$, $d \geq 1$. For readers interested in measure theoretical details, we refer to e.g. [15, 33] or [17]. This setting implies the following facts. First, we consider simple point processes (two points cannot occur at the same location). Second, we exclude manifold-valued point processes (like circular or spherical point processes), and marked point processes, even if most of the concepts and methodologies presented hereafter exist or can be straightforwardly adapted to such contexts.

Thus, $\mathbf{X} \cap B$ stands for the restriction of $\mathbf{X}$ to a set $B \subseteq S$ and we let $|B|$ denote the volume of any bounded $B \subset S$. Local finiteness of $\mathbf{X}$ means that $\mathbf{X} \cap B$ is finite almost surely (a.s.), that is the number of points $N(B)$ of $\mathbf{X} \cap B$ is finite a.s., whenever B is bounded. We let $\mathcal{N}$ stand for the state space consisting of the locally finite subsets (or point configurations) of S.

The distribution of $\mathbf{X}$ can be characterized by the finite-dimensional distributions of counting variables, or by the void probability, i.e. the probability to have no point in any compact set. However, these are usually not accessible and it is easier to summarize (and estimate) interpretable statistical measures such as intensity functions and conditional intensity functions. A more rigorous introduction on intensities, Palm intensities and conditional intensities and their links with reduced moment measures, Palm measures and reduced Campbell measures can be found in [12]. To get quicker to the core of the paper, we introduce them through Campbell theorem and Georgii-Nguyen-Zessin (GNZ) formula which may be viewed as integrals characterizations.

Theorem 1 (Campbell theorem and GNZ formula)**.** *The k-th order intensity function $\rho^{(k)}$ and the k-th order Papangelou conditional intensity function $\lambda^{(k)}$ are defined such that for any measurable function $h^{(k)} : (\mathbb{R}^d)^k \to \mathbb{R}^+$ and $\tilde{h}^{(k)} : (\mathbb{R}^d)^k \times \mathcal{N} \to \mathbb{R}^+$, we have respectively*

$$\mathbb{E}\{ \sum_{u_1,\ldots,u_k}^{\neq} h^{(k)}(u_1,\ldots,u_k)\} = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} h^{(k)}(u_1,\ldots,u_k)\rho^{(k)}(u_1,\ldots,u_k)\mathrm{d}u_1 \ldots \mathrm{d}u_k. \tag{1}$$

$$\mathbb{E}\{ \sum_{u_1,\ldots,u_k}^{\neq} \tilde{h}^{(k)}(\{u_1,\ldots,u_k\}, \mathbf{X} \setminus \{u_1,\ldots,u_k\})\} = \mathbb{E}\left\{ \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \tilde{h}^{(k)}(u_1,\ldots,u_k)\lambda^{(k)}(\{u_1,\ldots,u_k\}, \mathbf{X})\mathrm{d}u_1 \ldots \mathrm{d}u_k \right\}. \tag{2}$$

When $k = 1$, we more simply speak of the intensity function or the Papangelou conditional intensity function. It is relevant to have the following interpretation of such functions: $\rho(u)$ (resp. $\lambda(u, \mathbf{X})$) can be interpreted as the probability to observe a point in $B(u, \mathrm{d}u)$ an infinitesimal ball centered at u (resp. one point in $B(u, \mathrm{d}u)$ given the rest of the configuration of points outside the ball is $\mathbf{X}$). Similar interpretations are available when

$k > 1$. Equations (1)-(2) can be combined to show that $\rho(u) = \mathbb{E}\{\lambda(u, \mathbf{X})\}$ (also valid when $k > 1$). Figure 2 illustrates briefly the functions ρ and λ.

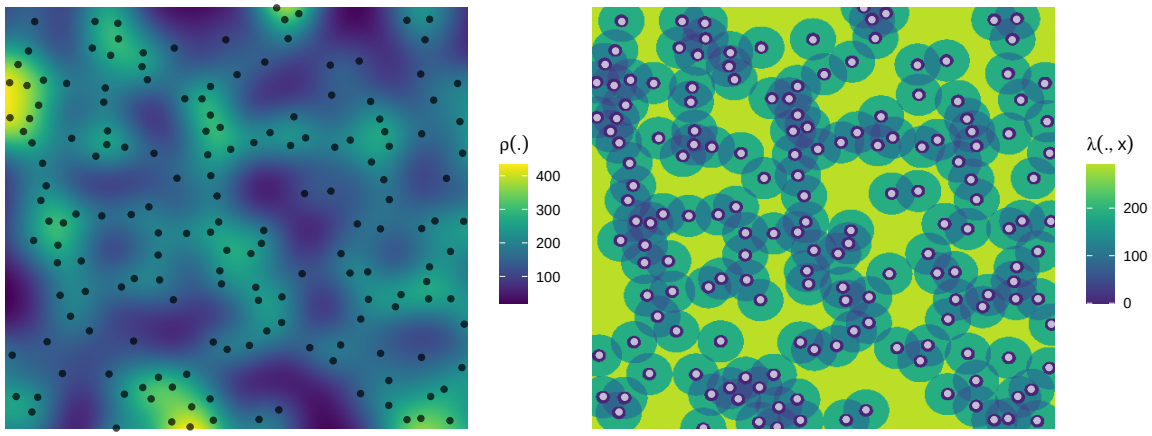

FIGURE 2. Left: Illustration of the concept of intensity function. The latent image is the true intensity function of the model (not specified) generating the pattern. The higher the intensity function the more likely a point can occur at this location; Right: Illustration of the conditional intensity function. The latent image depicts Papangelou conditional intensity function for a Gibbs model (not specified) given the configuration of points. The higher the intensity the more likely a point can be added at this location (and given this set of observed points). We can observe that this model produces repulsive patterns as the conditional intensity is zero around points and quite small in balls around observed points.

2.2. **Classes of models**

The reference model is the Poisson point process often defined as follows.

Definition 1. *Let ρ be a locally integrable function on $\mathbb{R}^d$. A point process $\mathbf{X}$ satisfying the following statements is called the Poisson point process on $\mathbb{R}^d$ with intensity function ρ:*

- *for any $m \geq 1$, and for any disjoint and bounded $B_1, \ldots, B_m \subset S$, the random variables $\mathbf{X} \cap B_1, \ldots, \mathbf{X} \cap B_m$ are independent;*
- *$N(B)$ follows a Poisson distribution with parameter $\int_B \rho(u)\mathrm{d}u$ for any bounded $B \subset S$.*

Poisson point processes model (eventually inhomogeneous) patterns with no interaction between points. As a consequence, it can be easily proved that for such processes, $\rho^{(k)}(u_1, \ldots, u_k) = \lambda^{(k)}(\{u_1, \ldots, u_k\}, \mathbf{X}) = \prod_i \rho(u_i)$. Large classes of models exist to introduce dependence between points. A survey can be found in [12] and the references therein. This is debatable but, to our point of view, the main classes are: Cox processes (which includes Neymann-Scott, shot noise Cox or log-Gaussian Cox processes, see e.g. [33]) defined as Poisson point processes with random driven intensity obtained from a random field; Gibbs point processes (see [17]), which are (in a bounded domain) defined via a density with respect to a Poisson point process with intensity 1; Determinantal point processes (DPP), e.g. [32], for which intensities are defined through the determinant of a kernel function. We do not intend to define rigorously these models (see e.g. [12] and references therein), however it is worth pointing out that these models are very different by the kind of interaction they model, their flexibility and, in particular as regards the concern of this paper, the fact that the intensity function and/or the Papangelou conditional intensity function is explicit or not. Table 2.2 is an attempt to present the diversity of these models, their richness which also makes this research area attractive and fruitful.

Model	Type of interaction	Is $\rho(\cdot)$ explicit?	Is $\lambda(u,\mathbf{x})$ explicit?
Poisson	no interaction	yes	yes
Cox	attraction	yes	no
Gibbs	attraction/repulsion	no	yes
DPP	repulsion	yes	yes and no

TABLE 1. Attempt to classify most of spatial point process models. Column type of interaction refers to the type of patterns the corresponding model can produce. Last two columns answer the question of tractability of the intensity and conditional intensity function. A "no" means that for most of models there is no explicit expression. "yes" definitely means that the expression is explicit and can be used from a statistical point of view. The "yes and no" is more in between "yes" and "no". An explicit expression can be obtained but is quite complex to exploit.

2.3. **Inhomogeneous parametric models**

Let us consider once more Figure 1 to motivate this section and paper. It is often the case in spatial statistics, that we observe a point pattern, here the locations of 3605 trees of Beilschmiedia pendula observed within a tropical forest in the Barro Colorado Island (see [3] for more details on this dataset) together with spatial covariates which are information available on the whole observation domain. A quick look at Figure 1 is enough to be convinced of the inhomogeneity (and maybe non independence) characteristic of the point pattern and that it makes completely sense to relate the distribution of trees with covariates such as the elevation, the slope of elevation or levels of soil nutrients.

In this application, we could be interested to model either the intensity and/or the Papangelou conditional intensity. We focus in this paper on exponential family models, where for any $u \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{N}$

$$\rho(u) = \exp\left\{\boldsymbol{\beta}^\top \mathbf{z}(u)\right\} \qquad \text{and} \qquad \lambda(u,\mathbf{x}) = \exp\left\{\boldsymbol{\beta}^\top \mathbf{z}(u) + \boldsymbol{\psi}^\top \mathbf{s}(u,\mathbf{x})\right\}. \tag{3}$$

In both definitions, $\boldsymbol{\beta} = \{\beta_1(u),\dots,\beta_p(u)\}^\top \in \mathbb{R}^p$ represents the main parameter vector of interest, $\mathbf{z}(u) = \{z_1(u),\dots,z_p(u)\}^\top$, $z_i : \mathbb{R}^d \to \mathbb{R}$ corresponds to the spatial covariates. We let $\boldsymbol{\psi} \in \mathbb{R}^l$ and $\mathbf{s}(u,\mathbf{x}) = \{s_1(u,\mathbf{x}),\cdots,s_l(u,\mathbf{x})\}^\top$ denote the parameter vector and the sufficient statistics defining the interaction term (more details are given below). This parameter vector $\boldsymbol{\psi}$ is also estimated using data. To rewrite (3) into the same formalism we suggest the reformulation

$$\rho(u;\boldsymbol{\beta}_\rho) = \exp\left\{\boldsymbol{\beta}_\rho^\top \mathbf{z}_\rho(u)\right\} \qquad \text{and} \qquad \lambda(u,\mathbf{x};\boldsymbol{\beta}_\lambda) = \exp\left\{\boldsymbol{\beta}_\lambda^\top \mathbf{z}_\lambda(u,\mathbf{x})\right\} \tag{4}$$

where $\boldsymbol{\beta}_\rho \in \mathbb{R}^p$ (resp. $\boldsymbol{\beta}_\lambda \in \mathbb{R}^{p+l}$), $\mathbf{z}_\rho(u) = \mathbf{z}(u)$ and $\mathbf{z}_\lambda(u,\mathbf{x}) = \{\mathbf{z}(u)^\top, \mathbf{s}(u,\mathbf{x})^\top\}^\top$. In the rest of the paper, each time we index a vector, matrix, random quantity by ρ (resp. λ) means that we refer to the estimation of ρ (resp. λ). And when a comment applies to the two problems, we write $\bullet$. Hence, for instance $\boldsymbol{\beta}_\bullet$ is the parameter of interest to be estimated and stands either for $\boldsymbol{\beta}_\rho$ or $\boldsymbol{\beta}_\lambda$. We point out that non-exponential family models can be considered but, as seen in Section 3.1, exponential models can be fitted very quickly using a tricky analogy with generalized linear models.

As a direct consequence of (4), the distribution of $\mathbf{X}$ is necessarily non-stationary. It is therefore highly relevant to ask the following questions: (A) given $\boldsymbol{\beta}_\rho$, $\mathbf{z}_\rho$, are there models with intensity ρ? (B) given $\boldsymbol{\beta}_\lambda$

and $\mathbf{z}_\lambda$, are there models with Papangelou conditional intensity λ? Answer to (A) is easy, as the Poisson point process already answers to this question. It is also quite simple to design inhomogeneous Cox point process or DPP to achieve this task (see e.g. [6, 32]). The question for (B) is much more complex (at least if $l \geq 1$, otherwise we are back to the Poisson case). As seen from Table 2.2, the question is essentially related to the existence of non-stationary Gibbs models. [18] (and the references therein) is one of the most popular existence result in the stationary case. In the non-stationary case, [38] provides sufficient conditions (which already cover a large class of examples): there exists at least one Gibbs measure with Papangelou conditional intensity λ if it satisfies for any $u \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{N}$

$$\lambda(u, \mathbf{x} : \boldsymbol{\beta}_\lambda) = \lambda\{u, \mathbf{x} \cap B(u, R); \boldsymbol{\beta}_\lambda\} \quad \text{and} \quad \lambda(u, \mathbf{x}; \boldsymbol{\beta}_\lambda) \leq \bar{\lambda} \tag{5}$$

where $R, \bar{\lambda} < \infty$. The first part (finite range property) means that the Papangelou conditional intensity at u depends only on points of $\mathbf{x}$ close to u. The second one, called local stability property, tells that the process is stochastically dominated by a Poisson point process. To set the ideas, the inhomogeneous Strauss model with $l = 1$ and $s_1(u, \mathbf{x}) = \sum_{v \in \mathbf{x}} \mathbf{1}(\|v - u\| \leq R)$ (number of R-closed neighbors of u in $\mathbf{x}$) satisfies (5) for any $\psi \in [0, 1]$ and $R < \infty$. We refer to [1, 3] for more complex examples.

Hence, the problem of inferring ρ or λ given by (4) is a well-posed one. The aim of next sections is to estimate $\boldsymbol{\beta}_\bullet$ based on a single observation $\mathbf{x}$ in an observation domain say D of $\mathbf{X}$, a spatial point process defined on $\mathbb{R}^d$ with intensity ρ (or conditional intensity λ).

3. Statistical inference for low-dimensional parametric ρ, λ

3.1. Poisson and pseudo-likelihoods

We consider composite likelihood-based techniques to estimate $\boldsymbol{\beta}_\bullet$. In particular we define the (log-)Poisson likelihood and (log-)pseudolikelihood respectively given by

$$\ell_\rho(\boldsymbol{\beta}_\rho; \mathbf{X}) = \sum_{u \in \mathbf{X} \cap D} \log \rho(u; \boldsymbol{\beta}_\rho) - \int_D \rho(u; \boldsymbol{\beta}_\rho) \mathrm{d}u \tag{6}$$

$$\ell_\lambda(\boldsymbol{\beta}_\lambda; \mathbf{X}) = \sum_{u \in \mathbf{X} \cap D \ominus R} \log \lambda(u, \mathbf{X} \setminus u; \boldsymbol{\beta}_\lambda) - \int_{D \ominus R} \lambda(u, \mathbf{X}; \boldsymbol{\beta}_\lambda) \mathrm{d}u, \tag{7}$$

where $D \ominus R$ in (7) stands for the domain eroded by the interaction range R. We use the notation $\ell_\bullet^{(1)}$ and $\ell_\bullet^{(2)}$ to denote the gradient vector and the Hessian matrix with respect to $\boldsymbol{\beta}_\bullet$. In addition, our main result Theorem 2 requires the definition of $\mathbf{H}_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X}) = -\ell_\bullet^{(2)}(\mathbf{X}; \boldsymbol{\beta}_\bullet)$ which are given by

$$\mathbf{H}_\rho(\boldsymbol{\beta}_\rho; \mathbf{X}) = \mathbf{H}_\rho(\boldsymbol{\beta}_\rho) = \int_D \mathbf{z}_\rho(u) \mathbf{z}_\rho(u)^\top \rho(u; \boldsymbol{\beta}_\rho) \mathrm{d}u \tag{8}$$

$$\mathbf{H}_\lambda(\boldsymbol{\beta}_\lambda; \mathbf{X}) = \int_{D \ominus R} \mathbf{z}_\lambda(u, \mathbf{X}) \mathbf{z}_\lambda(u, \mathbf{X})^\top \lambda(u, \mathbf{X}; \boldsymbol{\beta}_\lambda) \mathrm{d}u \tag{9}$$

and we also define the matrix $\mathbf{H}_\lambda(\boldsymbol{\beta}_\lambda) = \mathbb{E}\{\mathbf{H}_\lambda(\boldsymbol{\beta}_\lambda; \mathbf{X})\}$.

Let us comment on these methodologies. We focus first on (6) to estimate ρ given by (4). When $\mathbf{X}$ comes from a Poisson point process, as suggested by its name, (6) corresponds to the (log-)Poisson likelihood and its maximum was shown to converge by [34]. The method remains unbiased for general models, as the gradient of (6) can be easily interpreted as an estimating equation according to Campbell Theorem (1). [40] and then [39] proved asymptotic properties for general point processes under increasing domain asymptotic for large classes of mixing point processes. The efficiency of the method has been improved by [26] and then by [25] using quasi-likelihood. Infill asymptotic results have been obtained very recently by [8].

Equation (7) has also an intuitive origin: [31] obtained this criterion as the limit of a certain product of conditional densities. This paper together with [30] were also the first ones to establish consistency and asymptotic normality for stationary exponential family Gibbs models satisfying (5) under the increasing domain framework. These results were extended by [4, 10, 19] for more general models including infinite range pairwise interaction point processes or non-hereditary Gibbs models. Variant of the pseudo-likelihood include the Takacs-Fiksel method, the logistic regression likelihood and a form of quasi-likelihood respectively studied by [2, 11, 14]. The first results for inhomogeneous models were proved very recently by [1], still in the increasing domain framework. The popularity of (6)-(7) lies without doubt by their simple implementation. This is discussed in the next section.

3.2. **Implementation using the Berman-Turner's approximation**

The main challenge to maximize (6)-(7) is the second term which involves an integral over the observation domain and needs to be approximated numerically. The Berman-Turner scheme [3] is a popular approach that involves discretizing the integral using both quadrature points and data points,

$$\int_D \rho(u;\boldsymbol{\beta}_\rho)\mathrm{d}u \approx \sum_{i=1}^{N+M} w(u_i)\rho(u_i;\boldsymbol{\beta}_\rho) \quad \text{and} \quad \int_{D\ominus R} \lambda(u,\mathbf{x};\boldsymbol{\beta}_\lambda)\mathrm{d}u \approx \sum_{i=1}^{N+M} w(u_i)\lambda(u_i,\mathbf{x};\boldsymbol{\beta}_\lambda),$$

where $u_i, i = 1, \ldots, N+M$ are quadrature points in D or $D \ominus R$ (depending on the problem) involving N data points and M dummy points and where the $w(u_i) > 0$ are quadrature weights such that $\sum_i w(u_i) = |D|$ (or $|D \ominus R|$). Using this technique, (6)-(7) are then approximated by

$$\ell_\rho(\boldsymbol{\beta}_\rho;\mathbf{x}) \approx \tilde{\ell}_\rho(\boldsymbol{\beta}_\rho;\mathbf{x}) = \sum_{i=1}^{N+M} w_i\{y_i \log \rho_i(\boldsymbol{\beta}_\rho) - \rho_i(\boldsymbol{\beta}_\rho)\} \tag{10}$$

$$\ell_\lambda(\boldsymbol{\beta}_\lambda;\mathbf{x}) \approx \tilde{\ell}_\lambda(\boldsymbol{\beta}_\lambda;\mathbf{x}) = \sum_{i=1}^{N+M} w_i\{y_i \log \lambda_i(\mathbf{x};\boldsymbol{\beta}_\lambda) - \lambda_i(\mathbf{x};\boldsymbol{\beta}_\lambda)\} \tag{11}$$

where $w_i = w(u_i), y_i = w_i^{-1}\mathbf{1}(u_i \in \mathbf{x} \cap D)$ (or $D \ominus R$), $\rho_i(\boldsymbol{\beta}_\rho) = \rho(u_i;\boldsymbol{\beta}_\rho)$ and $\lambda_i(\mathbf{x};\boldsymbol{\beta}_\lambda) = \lambda(u_i,\mathbf{x};\boldsymbol{\beta}_\rho)$. It is now relevant to remark that Equations (10)-(11) are equivalent to a weighted quasi-likelihood function of independent Poisson variables y_i with weights w_i. Therefore, the implementation can take advantage of any software implementing generalized linear models. These methods are in particular implemented in the `spatstat` `R` package [3]. The accuracy of the approximation of $\ell_\bullet$ by $\tilde{\ell}_\bullet$ increases when N is small with respect to M. If N is too large or if increasing M leads to numerical instabilities, the induced bias can be non negligible. In these situations, alternatives based on the use of an approximation of (6)-(7) by a logistic regression likelihood are available (see e.g. [2, 40]).

To sum up Section 3: when p is moderate, we have at our disposal a bunch of statistical methodologies to estimate either the intensity or conditional intensity function. These methodologies are well-studied from a mathematical point of view and efficiently implemented.

4. Inference for sparse (conditional) intensity

4.1. **Setting and additional notation**

The application summarized by Figure 1 suggests that more refined methods are necessary. The number of spatial covariates is large (close to 100 if one considers topographic, levels of soil nutrients and levels of combinations of soil nutrients) which leads to numerical problems if one considers methods described in Section 3.1. Known as the curse of dimensionality, these problems can be alleviated if one assumes sparsity in the (conditional) intensity model and makes use of regularized versions of (6)-(7).

Let us turn to the setting of the present paper. We assume that we observe one realization from $\mathbf{X}_n$ where $(\mathbf{X}_n)_{n\geq 1}$ is a sequence of point processes defined in $\mathbb{R}^d$ and observed in D_n. We assume that either the intensity ρ or the Papangelou conditional intensity λ is modelled by (4). The true parameter (to be estimated) is denoted by $\boldsymbol{\beta}_{0,\bullet}$ and we assume it can be decomposed as $\boldsymbol{\beta}_{0,\rho} = (\boldsymbol{\beta}_{01,\rho}^\top, \boldsymbol{\beta}_{02,\rho}^\top)^\top$ and $\boldsymbol{\beta}_{0,\lambda} = (\boldsymbol{\beta}_{01,\lambda}^\top, \boldsymbol{\beta}_{02,\lambda}^\top, \boldsymbol{\psi})^\top$ where $\boldsymbol{\beta}_{01,\bullet} = 0$ and where all components of $\boldsymbol{\beta}_{02,\bullet}$ are non zero. We index any (random) vector, matrix in the same way. Thus $\mathbf{z}_{01,\bullet}$ corresponds to the set of non-informative covariates while the set $\mathbf{z}_{02,\bullet}$ represents the set of active features. We assume that $\boldsymbol{\beta}_{02,\rho}$ or $(\boldsymbol{\beta}_{02,\lambda}^\top, \boldsymbol{\psi}^\top)^\top$ has length s_n. Thus $\boldsymbol{\beta}_{01,\bullet}$ has length $p_n - s_n$. The sequences s_n and p_n may increase with n. Finally, to quantify the amount of increase of data with n, we assume for both problems that ρ, λ, $\boldsymbol{\beta}_{0,\bullet}$ and D_n are such that $\mu_n \to \infty$ as $n \to \infty$ where

$$\mu_n = \mathbb{E}\{N(D_n)\} = \int_{D_n} \rho(u; \boldsymbol{\beta}_{0,\rho}) \mathrm{d}u = \int_{D_n} \mathbb{E}\left\{\lambda(u, \mathbf{X}_n; \boldsymbol{\beta}_{0,\lambda})\right\} \mathrm{d}u. \tag{12}$$

We assume that for any $n \geq 1$, the model is well-defined. In particular for λ, this means the sequence of Papangelou conditional intensity functions satisfies (5). Note that μ_n is a function of $D_n, \boldsymbol{\beta}_{02,\bullet}, \mathbf{z}_{02,\bullet}(u)$ and s_n. We believe this kind of framework is original and quite general. It embraces the well-known frameworks called increasing domain asymptotics and infill asymptotics. For the increasing domain context, $D_n \to \mathbb{R}^d$ and usually $\boldsymbol{\beta}_{02,\bullet}$ depends only on n through s_n. For the infill asymptotics, $D_n = D$ is assumed to be a bounded domain of $\mathbb{R}^d$ and usually $(\mathbf{z}_2)_1(u) = 1$, $(\boldsymbol{\beta}_{02,\bullet})_1 = \theta_n \to \infty$ as $n \to \infty$. In some sense, the parameter μ_n plays the role of the sample size in standard inference. To reduce notation in the following, unless it is ambiguous, we do not index $\mathbf{X}$, ρ, λ, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$, $\mathbf{z}_\bullet(u), \ell_\bullet$ with n.

When the number of parameters is large, regularization methods allow one to perform both estimation and variable selection simultaneously. When $p_n = p$, [6] consider several regularization procedures which consist in adding a convex or non-convex penalty term to (6)-(7). A quite similar approach was considered in [1] for Gibbs point processes. To ease the presentation and focus more on the similarities between the two problems of inferring (4), we only consider the ℓ^1 regularization which gives rise to the adaptive lasso procedure. The ℓ^1-regularized versions of (6)-(7) are given by

$$Q_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X}) = \frac{1}{\mu_n} \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X}) - \sum_{j=1}^{p_n} \tau_{n,j} |\beta_{j,\bullet}| \tag{13}$$

where the real numbers $\tau_{n,j}$ are non-negative tuning parameters, also called regularization parameters. The adaptive lasso estimator is then defined by

$$\hat{\boldsymbol{\beta}}_\bullet = \arg \max_{\boldsymbol{\beta}_\bullet \in \mathbb{R}^{p_n}} Q_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X}). \tag{14}$$

When $\tau_{n,j} = 0$ for $j = 1, \ldots, p_n$, the method reduces to the maximum Poisson likelihood or pseudo-likelihood estimator and when $\tau_{n,j} = \tau_n$ to the standard lasso estimator. Note that in the formulation (13), if we want the model to necessarily have an intercept term and/or if one does not want to regularize the parameter vector corresponding to the interaction term for λ, we can simply set the corresponding tuning parameters to 0. Finally, the choice of μ_n as a normalization factor in (13) follows the implementation of the adaptive lasso procedure for generalized linear models in standard softwares (e.g. `R` package `glmnet` [22]).

4.2. **Asymptotic results for the adaptive lasso**

Our result relies upon the following conditions:

- ($\mathcal{C}$.1) For any $n \geq 1$, the intensity or the conditional intensity functions has the log-linear specification given by (4) where $\boldsymbol{\beta}_\bullet \in \mathbb{R}^{p_n}$. For λ, we assume that it satisfies (5).
- ($\mathcal{C}$.2) $(\mu_n)_{n\geq 1}$ is an increasing sequence of real numbers, such that $\mu_n \to \infty$ as $n \to \infty$.

($\mathcal{C}.3$) As $n \to \infty$, $\ell_\bullet^{(1)}(\boldsymbol{\beta}_{0,\bullet}; \mathbf{X}) = O_{\mathrm{P}}\left(\sqrt{p_n \mu_n}\right)$.
($\mathcal{C}.4$) The matrix $\mathbf{H}_\rho(\boldsymbol{\beta}_{0,\rho})$ or the matrices $\mathbf{H}_\lambda(\boldsymbol{\beta}_{0,\lambda}; \mathbf{X})$ and $\mathbf{H}_\lambda(\boldsymbol{\beta}_{0,\lambda})$ satisfy

$$\inf_{n\geq 1} \inf_{\boldsymbol{\phi}\in\mathbb{R}^{p_n}, \|\boldsymbol{\phi}\|=1} \mu_n^{-1} \boldsymbol{\phi}^\top \mathbf{H}_\bullet(\boldsymbol{\beta}_{0,\bullet}) \boldsymbol{\phi} > 0 \quad \text{and} \quad \sup_{\boldsymbol{\phi}\in\mathbb{R}^{p_n}, \|\boldsymbol{\phi}\|=1} \boldsymbol{\phi}^\top \left\{ \mathbf{H}_\lambda(\boldsymbol{\beta}_{0,\lambda}; \mathbf{X}) - \mathbf{H}_\lambda(\boldsymbol{\beta}_{0,\lambda}) \right\} \boldsymbol{\phi} = o_{\mathrm{P}}(\mu_n).$$

($\mathcal{C}.5$) As $n \to \infty$, $p_n^4/\mu_n \to 0$.
($\mathcal{C}.6$) For any $c \in \mathbb{R}$, $\tilde{\boldsymbol{\beta}}_\bullet = \boldsymbol{\beta}_{0,\bullet} + c\sqrt{p_n/\mu_n}$ and $j = 1, \ldots, p_n - s_n$

$$\int_{D_n} \|\mathbf{z}_\rho(u)\|^3 \rho(u; \tilde{\boldsymbol{\beta}}_\rho) \mathrm{d}u = O(p_n^{3/2}) \quad \text{and} \quad \int_{D_n \ominus R} \|\mathbf{z}_\lambda(u, \mathbf{X})\|^3 \lambda(u, \mathbf{X}; \tilde{\boldsymbol{\beta}}_\lambda) \mathrm{d}u = O_{\mathrm{P}}(p_n^{3/2})$$
$$\int_{D_n} |(\mathbf{z}_\rho)_j(u)| \|\mathbf{z}_\rho(u)\| \rho(u; \tilde{\boldsymbol{\beta}}_\rho) \mathrm{d}u = O(\sqrt{p_n}) \quad \text{and} \quad \int_{D_n \ominus R} |(\mathbf{z}_\lambda)_j(u, \mathbf{X})| \|\mathbf{z}_\lambda(u, \mathbf{X})\| \lambda(u, \mathbf{X}; \tilde{\boldsymbol{\beta}}_\lambda) \mathrm{d}u = O_{\mathrm{P}}(\sqrt{p_n}).$$

($\mathcal{C}.7$) Let $a_n = \max_{j=p_n - s_n + 1, \ldots, p_n} \tau_{n,j}$ and $b_n = \min_{j=1,\ldots,p_n - s_n} \tau_{n,j}$. The $\tau_{n,j}$ are allowed to be stochastic and we assume that, as $n \to \infty$

$$a_n \sqrt{\frac{s_n \mu_n}{p_n}} = o_{\mathrm{P}}(1) \qquad \text{and} \qquad \frac{1}{b_n} \sqrt{\frac{p_n^2}{\mu_n}} = o_{\mathrm{P}}(1).$$

Let us discuss these conditions. Conditions ($\mathcal{C}.1$)-($\mathcal{C}.2$) have already been explained. It is worth saying that in an attempt to embrace both problems of estimating ρ or λ (and ease the presentation) some conditions may appear useless or too vague. This is the case for ($\mathcal{C}.3$), ($\mathcal{C}.4$) and ($\mathcal{C}.6$). We invite the reader to refer to [1,7] where conditions on second-order moments (for ρ) or second-order Papangelou conditional intensity (for λ and in the increasing domain framework) are presented. Essentially, condition ($\mathcal{C}.3$) is obtained by proving that the variance of the score behaves as $p_n \mu_n$ (which corresponds to np_n for standard GLMs for instance). Condition ($\mathcal{C}.4$) shows that the smallest eigenvalue of $\mu_n^{-1} \mathbf{H}_\bullet(\boldsymbol{\beta}_{0,\bullet})$ is positive for any n and could be compared to an assumption of the form $n^{-1}\mathbf{X}\mathbf{X}^\top$ tends to a positive definite matrix for standard linear models with fixed number of covariates (where $\mathbf{X}$ would stand for the design matrix). Condition ($\mathcal{C}.6$) looks meaningless however, for instance for ρ and $c = 0$, these are fulfilled as soon as $\sup_n \sup_i \sup_u |(\mathbf{z})_i(u)| < \infty$ (again see [1, 7] for more details). Condition ($\mathcal{C}.5$) reveals one limitation of this asymptotic result. The number of covariates is allowed to diverge with n but cannot increase too quickly. Such an assumption is also standard if one is interested in asymptotic properties of estimators for high-dimensional GLMs, see [21].

Condition ($\mathcal{C}.7$) expresses the compromise to be considered on s_n, p_n, μ_n and the regularization parameters to ensure consistency and oracle properties. These are quite similar to the corresponding ones considered by [1, 7] and that kind of compromise is also present in the pioneering work [20] dealing with linear models. The basic situation is the following one: let $s_n = s$, $p_n = p$ and $\tau_{n,j} = \tau_n$ a deterministic sequence (standard lasso), then ($\mathcal{C}.5$) cannot be fullfilled since $\tau_n \sqrt{\mu_n}$ and $\frac{1}{\tau_n \sqrt{\mu_n}}$ cannot simultaneously tend to 0 as $n \to \infty$. Allowing the regularization parameters sequences to evolve differently enables to satisfy the above constraint and also enables to consider the situation where s_n and p_n diverge with n. Finally, we include here the possibility to have stochastic regularization parameters. This does not imply strong modifications in the proofs (compared to proofs of similar results in [1, 6, 7]) but its interest is significant for the choice of regularization parameters, see the discussion after the presentation of Theorem 2.

Theorem 2. *Let $\hat{\boldsymbol{\beta}}_\bullet$ be given by* (14). *Assume that the conditions ($\mathcal{C}.1$)-($\mathcal{C}.7$) hold, then the following properties hold.*

(i) Consistency: $\hat{\boldsymbol{\beta}}_\bullet$ satisfies $\hat{\boldsymbol{\beta}}_\bullet - \boldsymbol{\beta}_{0,\bullet} = O_{\mathrm{P}}(\sqrt{p_n/\mu_n})$.
(ii) Sparsity: $\mathrm{P}(\hat{\boldsymbol{\beta}}_{1,\bullet} = 0) \to 1$ as $n \to \infty$.

A look at the proof of Theorem 2(i) shows in particular that the consistency remains true under the first part of condition ($\mathcal{C}.7$) and so remains valid for the standard lasso and actually even if no regularization is considered. The combination of Theorem 2 (i)-(ii) justifies the interest of a regularization technique. With the same rate of convergence than the one of the unregularized estimator, we ensure that the estimator of $\boldsymbol{\beta}_{1,\bullet}$ (corresponding to non-informative covariates) can be set to 0 with probability tending to 1.

We decided not to include more results to keep the paper readable and short. [1, 7] explore for example the asymptotic normality of $\hat{\boldsymbol{\beta}}_{2,\rho}$ and $(\hat{\boldsymbol{\beta}}_{2,\lambda}^\top, \hat{\boldsymbol{\psi}}^\top)^\top$ respectively, and provide estimates of asymptotic covariance matrices. In the same vein, we only focus on the adaptive lasso penalty. More convex penalties (e.g. adaptive elastic net) as well as non-convex penalties (such as MC+, SCAD penalties) are considered in aforementioned references.

Theorem 2 extends the results obtained by [1,7]. In particular for the estimation problem of λ, we consider the possibility to include an infill asymptotics and we allow the number of parameters for the interaction parameter, that is the length of $\boldsymbol{\psi}$, to increase with n. But the main improvement is that for both problems, we allow the tuning parameters $\tau_{n,j}$ to be stochastic, see Section 4.3 and this of signigicant practical interest. Indeed, [41] and many authors after this work suggest to scale the regularization parameters as $\tau_{n,j} = \tau_n/|(\check{\boldsymbol{\beta}}_\bullet)_j|^\gamma$ where τ_n is a non-negative sequence, $\gamma > 0$ and where $\check{\boldsymbol{\beta}}_\bullet$ is the unregularized estimator (or actually any estimate producing no zero). This popular idea is indeed very natural as it is expected that the $\tau_{n,j}$ for the non-informative covariates will be much larger than the ones for the informative ones and so estimates of parameters corresponding to non-informative covariates will be more likely set to zero thanks to the ℓ^1 penalty. We claim that τ_n can be adjusted to fulfill condition ($\mathcal{C}.7$). Indeed, since $(\check{\boldsymbol{\beta}}_\bullet)_j - (\boldsymbol{\beta}_{0,\bullet})_j = O_{\mathrm{P}}(\sqrt{p_n/\mu_n})$, it is easily deduced that $a_n = O_{\mathrm{P}}(\tau_n)$ and $b_n^{-1} = \max_{j \leq p_n - s_n} \tau_{n,j}^{-1} = O_{\mathrm{P}}\{\tau_n^{-1}(p_n/\mu_n)^{\gamma/2}\}$. Now, since $s_n \leq p_n$ and $p_n^4/\mu_n = o(1)$ by condition ($\mathcal{C}.5$),

$$a_n \sqrt{\frac{s_n \mu_n}{p_n}} = O_{\mathrm{P}}(\tau_n \sqrt{\mu_n}) \quad \text{and} \quad \frac{1}{b_n}\sqrt{\frac{p_n^2}{\mu_n}} = o_{\mathrm{P}}\left(\frac{1}{\tau_n}\frac{1}{\mu_n^{-1/4-3\gamma/8}}\right).$$

And so condition ($\mathcal{C}.7$) is in particular satisfied if $\tau_n\sqrt{\mu_n} \to 0$ and $\tau_n \mu_n^{1/4+3\gamma/8} \to \infty$ as $n \to \infty$. For instance, if $\tau_n = \mu_n^{-\alpha}$ with $\alpha > 0$, this imposes the non-empty condition $1/2 < \alpha < 1/4 + 3\gamma/8$ (if $\gamma > 2/3$). A discussion on how to select τ_n and/or γ and other numerical considerations are presented in the next section.

4.3. **Numerical considerations, algorithms and implementation**

To implement regularization methods for spatial point processes in particular within `R`, we combine the existing `R` package `spatstat` [3] (devoted to the analysis of spatial point pattern data) with two `R` packages `glmnet` [22] (for convex penalties) and `ncvreg` [5] (for non-convex penalties) which use coordinate descent procedures/algorithms [1, 5, 6, 16, 22]. Those methods rely on the tuning parameters $\tau_{n,j}$. Following [41], we suggest to use $\tau_{n,j} = \tau_n/|(\check{\boldsymbol{\beta}}_\bullet)_j|^\gamma$ where $\tau = \tau_n$ is to be chosen (the parameter γ has less influence and is often set to 1). Large values of τ yield estimates with high biases and low variances, whereas small values of τ produce estimates with low biases and high variances. Therefore, an optimal choice of the tuning parameter τ is necessary to control the trade-off between the bias and the variance. To select τ, it is reasonable first to identify a decreasing sequence of τ ranging from a maximum value of τ for which all penalized coefficients are zero to $\tau = 0$ (which corresponds to the unregularized parameter estimator); and second to define a criterion to select τ by an optimization (minimization) procedure. Following [1, 8], we suggest the use of an information criteria such as the (composite) Bayesian information criterion (BIC) [23, 35] or the (composite) extended regularized information criterion (ERIC) [28] to select τ. As explained in [6], the computational cost of this approach is cheaper than cross-validation techniques. These techniques are well-known for standard models but are less appropriate for spatial models which are by essence dependent models. Let us define first the (composite) BIC, which we denote by cBIC,

$$\mathrm{cBIC}_\bullet(\tau) = -2\ell_\bullet(\hat{\boldsymbol{\beta}}_\bullet; \mathbf{x}) + \log(N)\, d_\bullet(\tau) \tag{15}$$

where N is the observed number of points and

$$d_\bullet(\tau) = \mathrm{trace}(\hat{\mathbf{H}}_\bullet(\hat{\boldsymbol{\beta}}_\bullet)\hat{\boldsymbol{\Sigma}}_\bullet(\hat{\boldsymbol{\beta}}_\bullet)) \tag{16}$$

with $\boldsymbol{\Sigma}_\bullet(\boldsymbol{\beta}_\bullet) = \mathbf{H}_\bullet(\boldsymbol{\beta}_\bullet)^{-1}\mathbf{V}_\bullet(\boldsymbol{\beta}_\bullet)\mathbf{H}_\bullet(\boldsymbol{\beta}_\bullet)^{-1}$ and $\mathbf{V}_\bullet(\boldsymbol{\beta}_\bullet) = \mathrm{Var}(\ell_\bullet^{(1)}(\boldsymbol{\beta}_\bullet;\mathbf{X}))$. It is worth pointing out that $d(\tau)$ is called the effective number of parameters in the model with tuning parameter τ and for models with tractable likelihood functions like the inhomogeneous Poisson point process, $d(\tau)$ corresponds to the number of non-zero coefficients in $\hat{\boldsymbol{\beta}}_\bullet$ and the criterion reduces to BIC. For Gibbs models, estimates of $\mathbf{H}_\lambda$ and $\boldsymbol{\Sigma}_\lambda$ can be efficiently computed using the `vcov` function of the `spatstat R` package [13]. Let us now define the (composite) ERIC, which we denote by cERIC and which is designed for the purpose of taking into account the effects of the tuning parameter τ,

$$\mathrm{cERIC}_\bullet(\tau) = -2\ell_\bullet(\hat{\boldsymbol{\beta}}_\bullet;\mathbf{x}) + \log\left(\frac{N}{|D|\tau}\right) d_\bullet(\tau). \tag{17}$$

To sum up, we choose the tuning parameter $\tau \geq 0$ which minimizes either cBIC or cERIC. For more numerical details as well as implementation of the methodology, we refer the reader to [1, 6, 8, 9]. A simulation study conducted in [1] for Gibbs models shows that the criterion cERIC tends to produce better results in terms of selection and prediction. We end this section by mentioning that a recent version of the `spatstat R` package allows to include elastic net regularization in the `ppm` function through the option `improve.type='enet'` [6,9].

A. Proof of Theorem 2

Proof. (i) Let $\mathbf{k} \in \mathbb{R}^{p_n}$. We remind the reader that the estimator of $\boldsymbol{\beta}_{0,\bullet}$ defined as the maximum of $Q_\bullet$ given by (13). We aim at proving that for any given $\varepsilon > 0$, there exists sufficiently large K such that for n sufficiently large

$$\mathrm{P}\left\{\sup_{\|\mathbf{k}\|=K} \Delta_\bullet(\mathbf{k};\mathbf{X}) > 0\right\} \leq \varepsilon \quad \text{where} \quad \Delta_\bullet(\mathbf{k};\mathbf{X}) = Q_\bullet(\boldsymbol{\beta}_{0,\bullet} + \sqrt{p_n/\mu_n}\mathbf{k};\mathbf{X}) - Q_\bullet(\boldsymbol{\beta}_{0,\bullet};\mathbf{X}). \tag{18}$$

Equation (18) will imply that with probability at least $1-\varepsilon$, there exists a local maximum in the ball $\{\boldsymbol{\beta}_{0,\bullet} + \sqrt{p_n/\mu_n}\mathbf{k} :, \|\mathbf{k}\| \leq K\}$. We decompose $\Delta_\bullet(\mathbf{k};\mathbf{X}) = T_{1,\bullet} + T_{2,\bullet}$ with

$$T_{1,\bullet} = \mu_n^{-1}\left\{\ell_\bullet(\boldsymbol{\beta}_{0,\bullet} + \sqrt{p_n/\mu_n}\mathbf{k}) - \ell_\bullet(\boldsymbol{\beta}_{0,\bullet};\mathbf{X})\right\} \tag{19}$$

$$T_{2,\bullet} = \sum_{j=1}^{p_n} \tau_{n,j}\left(|(\boldsymbol{\beta}_{0,\bullet})_j| - |(\boldsymbol{\beta}_{0,\bullet})_j + \sqrt{p_n/\mu_n}k_j|\right). \tag{20}$$

Since $\rho(u;\cdot)$ and $\lambda(u,\mathbf{x};\cdot)$ are infinitely continuously differentiable for any $u \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{N}$, $\ell_\bullet(\cdot;\mathbf{X})$ is in particular twice continuously differentiable. Using a second-order Taylor expansion there exists $t \in (0,1)$ such that

$$\mu_n T_{1,\bullet} = \sqrt{\frac{p_n}{\mu_n}}\mathbf{k}^\top \ell_\bullet^{(1)}(\boldsymbol{\beta}_{0,\bullet};\mathbf{X}) + T_{11,\bullet} + T_{12,\bullet}$$

where (remind that by definition $\ell_\bullet^{(2)} = -\mathbf{H}_\bullet$)

$$T_{11,\bullet} = -\frac{1}{2}\frac{p_n}{\mu_n}\mathbf{k}^\top \mathbf{H}_\bullet(\boldsymbol{\beta}_{0,\bullet};\mathbf{X})\mathbf{k} \tag{21}$$

$$T_{12,\bullet} = \frac{1}{2}\frac{p_n}{\mu_n}\mathbf{k}^\top\left\{\mathbf{H}_\bullet(\boldsymbol{\beta}_{0,\bullet};\mathbf{X}) - \mathbf{H}_\bullet(\boldsymbol{\beta}_{0,\bullet} + t\sqrt{p_n/\mu_n};\mathbf{X})\right\}\mathbf{k}. \tag{22}$$

Under condition ($\mathcal{C}.4$), $T_{11,\rho} \leq -(\alpha_\rho/2)p_n\|\mathbf{k}\|^2$ for some $\alpha_\rho > 0$. For the conditional intensity, again using condition ($\mathcal{C}.4$)

$$T_{11,\lambda} = -\frac{1}{2}\frac{p_n}{\mu_n}\mathbf{k}^\top \mathbf{H}_\lambda(\boldsymbol{\beta}_{0,\lambda})\mathbf{k} + \omega_{11,\lambda} \leq -\frac{\alpha_\lambda}{2}p_n\|\mathbf{k}\|^2 + \omega_{11,\lambda}$$

where $\omega_{11,\lambda} = o_{\mathrm{P}}(p_n)$. Now, for some $\tilde{\boldsymbol{\beta}}_\bullet$ on the line segment between $\boldsymbol{\beta}_{0,\bullet}$ and $\boldsymbol{\beta}_{0,\bullet} + t\sqrt{p_n/\mu_n}$, we have under condition ($\mathcal{C}.6$)

$$\begin{aligned}
T_{12,\rho} &= \frac{1}{2}\frac{p_n}{\mu_n}\mathbf{k}^\top \left\{ \int_{D_n} \mathbf{z}_{\rho(u)}\mathbf{z}_\rho(u)^\top t\sqrt{\frac{p_n}{\mu_n}}\mathbf{k}^\top \mathbf{z}_\rho(u)\rho(u;\tilde{\boldsymbol{\beta}}_\rho \mathrm{d}u \right\}\mathbf{k} \\
&= O\left(\frac{p_n}{\mu_n}\sqrt{\frac{p_n}{\mu_n}}\right)\int_{D_n}\|\mathbf{z}_\rho(u)\|^3\rho(u;\tilde{\boldsymbol{\beta}}_\rho)\mathrm{d}u = O\left(p_n\sqrt{\frac{p_n^4}{\mu_n}}\right) \\
T_{12,\lambda} &= \frac{1}{2}\frac{p_n}{\mu_n}\mathbf{k}^\top \left\{ \int_{D_n\ominus R} \mathbf{z}_{\lambda(u,\mathbf{X})}\mathbf{z}_\lambda(u,\mathbf{X})^\top t\sqrt{\frac{p_n}{\mu_n}}\mathbf{k}^\top \mathbf{z}_\lambda(u,\mathbf{X})\lambda(u,\mathbf{X};\tilde{\boldsymbol{\beta}}_\lambda \mathrm{d}u \right\}\mathbf{k} \\
&= O\left(\frac{p_n}{\mu_n}\sqrt{\frac{p_n}{\mu_n}}\right)\int_{D_n\ominus R}\|\mathbf{z}_\lambda(u,\mathbf{X})\|^3\lambda(u,\mathbf{X};\tilde{\boldsymbol{\beta}}_\lambda)\mathrm{d}u = O_{\mathrm{P}}\left(p_n\sqrt{\frac{p_n^4}{\mu_n}}\right).
\end{aligned}$$

Hence, under condition ($\mathcal{C}.5$), $T_{12,\rho} = o(p_n)$ and $T_{12,\lambda} = o_{\mathrm{P}}(p_n)$, which yields

$$T_{1,\bullet} \leq \frac{1}{\mu_n}\sqrt{\frac{p_n}{\mu_n}}\mathbf{k}^\top \ell_\bullet^{(1)}(\boldsymbol{\beta}_{0,\bullet}) - \frac{\alpha_\bullet}{2}\frac{p_n}{\mu_n}\|\mathbf{k}\|^2 + \omega_{1,\bullet}$$

where $\omega_{1,\rho} = o(p_n/\mu_n)$ and $\omega_{1,\lambda} = o_{\mathrm{P}}(p_n/\mu_n)$. Regarding the term $T_{2,\bullet}$, we have

$$\begin{aligned}
T_{2,\bullet} &= \sum_{j=1}^{p_n-s_n}\tau_{n,j}\left(|(\boldsymbol{\beta}_{0,\bullet})_j| - |(\boldsymbol{\beta}_{0,\bullet})_j + \sqrt{p_n/\mu_n}k_j|\right) + \sum_{j=p_n-s_n+1}^{p_n}\tau_{n,j}\left(|(\boldsymbol{\beta}_{0,\bullet})_j| - |(\boldsymbol{\beta}_{0,\bullet})_j + \sqrt{p_n/\mu_n}k_j|\right) \\
&\leq \sum_{j=p_n-s_n+1}^{p_n}\tau_{n,j}\left(|(\boldsymbol{\beta}_{0,\bullet})_j| - |(\boldsymbol{\beta}_{0,\bullet})_j + \sqrt{p_n/\mu_n}k_j|\right) \\
&\leq a_n\sqrt{\frac{p_n}{\mu_n}}\sum_{j=p_n-s_n+1}^{p_n}|k_j| \leq a_n\sqrt{\frac{s_np_n}{\mu_n}}\|\mathbf{k}\| =: \omega_{2,\bullet}
\end{aligned}$$

where under condition ($\mathcal{C}.7$), $\omega_{2,\bullet} = O(a_n\sqrt{s_np_n/\mu_n}p_n/\mu_n) = o_{\mathrm{P}}(p_n/\mu_n)$. We deduce that as $n \to \infty$

$$\Delta_\bullet(\mathbf{k};\mathbf{X}) \leq \frac{1}{\mu_n}\sqrt{\frac{p_n}{\mu_n}}\|\ell_\bullet^{(1)}(\boldsymbol{\beta}_{0,\bullet};\mathbf{X})\|\,\|\mathbf{k}\| - \frac{\alpha_\bullet}{2}\frac{p_n}{\mu_n}\|\mathbf{k}\|^2 + \omega_\bullet$$

where $\omega_\bullet = \omega_{1,\bullet} + \omega_{2,\bullet} = o_{\mathrm{P}}(p_n)$ whereby we continue with

$$\mathrm{P}\left\{\sup_{\|\mathbf{k}\|=K}\Delta_\bullet(\mathbf{k};\mathbf{X}) > 0\right\} \leq \mathrm{P}\left(L_n \geq \frac{\alpha_\bullet}{2}K\sqrt{p_n\mu_n}\right)$$

where $L_n = \|\ell_\bullet^{(1)}(\boldsymbol{\beta}_{0,\bullet};\mathbf{X})\| + \mu_n\sqrt{\frac{\mu_n}{p_n}}\omega_\bullet$. This leads to the result since under condition ($\mathcal{C}.3$)

$$L_n = O_P\left(\sqrt{p_n\mu_n}\right) + O\left(\mu_n\sqrt{\frac{\mu_n}{p_n}}\omega_\bullet\right) = O_P\left(\sqrt{p_n\mu_n}\right).$$

(ii) Following (i), we intend to prove that for any $\tilde{\boldsymbol{\beta}}_\bullet \in \mathbb{R}^{s_n}$ (where $\tilde{\boldsymbol{\beta}}_\rho = \boldsymbol{\beta}_{2,\rho}$ or $\tilde{\boldsymbol{\beta}}_\lambda = (\boldsymbol{\beta}_{2,\lambda}^\top, \boldsymbol{\psi}^\top)^\top$) satisfying $\|\tilde{\boldsymbol{\beta}}_\bullet - \tilde{\boldsymbol{\beta}}_{0,\bullet}\| = O_{\mathrm{P}}(\sqrt{p_n/\mu_n})$ and any $K_1 > 0$

$$Q_\bullet\left\{(\mathbf{0}^\top, \tilde{\boldsymbol{\beta}}_\bullet^\top)^\top; \mathbf{X}\right\} = \max_{\|\boldsymbol{\beta}_{1,\bullet}\| \leq K_1\sqrt{p_n/\mu_n}} Q_\bullet\left\{(\boldsymbol{\beta}_{1,\bullet}^\top, \tilde{\boldsymbol{\beta}}_\bullet^\top)^\top; \mathbf{X}\right\}.$$

To this end, it is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\tilde{\boldsymbol{\beta}}_\bullet$ such that $\|\tilde{\boldsymbol{\beta}}_\bullet - \tilde{\boldsymbol{\beta}}_{0,\bullet}\| = O_{\mathrm{P}}(\sqrt{p_n/\mu_n})$, we have for any $j = 1, \ldots, p_n - s_n$

$$\frac{\partial Q_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} < 0 \quad \text{for} \quad 0 < (\boldsymbol{\beta}_\bullet)_j < \varepsilon_n, \qquad \frac{\partial Q_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} > 0 \quad \text{for} \quad -\varepsilon_n < (\boldsymbol{\beta}_\bullet)_j < 0. \tag{23}$$

We focus only on the first part of (23) as the other one follows along similar lines. We have

$$\frac{\partial \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} = \frac{\partial \ell_\bullet(\boldsymbol{\beta}_{0,\bullet}; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} + R_\bullet$$

where

$$\begin{aligned} R_\rho &= -\int_{D_n} (\mathbf{z}_\rho)_j(u) \left\{\rho(u; \boldsymbol{\beta}_\rho) - \rho(u; \boldsymbol{\beta}_{0,\rho})\right\} \mathrm{d}u \\ R_\lambda &= -\int_{D_n} (\mathbf{z}_\lambda)_j(u, \mathbf{X}) \left\{\lambda(u, \mathbf{X}; \boldsymbol{\beta}_\lambda) - \lambda(u, \mathbf{X}; \boldsymbol{\beta}_{0,\lambda})\right\} \mathrm{d}u. \end{aligned}$$

By Taylor expansion and Cauchy-Schwarz inequality, there exists $\check{\boldsymbol{\beta}}_\bullet$ on the line segment between $\boldsymbol{\beta}_{0,\bullet}$ and $\boldsymbol{\beta}_\bullet$

$$\begin{aligned} |R_\rho| &= O\left(\|\boldsymbol{\beta}_\rho - \boldsymbol{\beta}_{0,\rho}\|\right) \int_{D_n} |(\mathbf{z}_\rho)_j(u)| \, \|\mathbf{z}_\rho(u)\| \rho(u; \check{\boldsymbol{\beta}}_\rho) \mathrm{d}u \\ |R_\lambda| &= O\left(\|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_{0,\lambda}\|\right) \int_{D_n \ominus R} |(\mathbf{z}_\lambda)_j(u, \mathbf{X})| \, \|\mathbf{z}_\lambda(u, \mathbf{X})\| \lambda(u, \mathbf{X}; \check{\boldsymbol{\beta}}_\lambda) \mathrm{d}u. \end{aligned}$$

By condition ($\mathcal{C}$.6), we deduce that $R_\bullet = O_{\mathrm{P}}(\sqrt{p_n/\mu_n}\sqrt{p_n}\mu_n) = O_{\mathrm{P}}(p_n\sqrt{\mu_n})$, which combined with condition ($\mathcal{C}$.3) yields

$$\frac{\partial \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} = O_{\mathrm{P}}(p_n\sqrt{\mu_n}). \tag{24}$$

Now, let $0 < (\boldsymbol{\beta}_\bullet)_j < \varepsilon_n$, for n sufficiently large

$$\begin{aligned} \mathrm{P}\left\{\frac{\partial Q_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} < 0\right\} &= \mathrm{P}\left\{\frac{\partial \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} - \mu_n\tau_{n,j}\mathrm{sign}(\boldsymbol{\beta}_\bullet)_j < 0\right\} \\ &= \mathrm{P}\left\{\frac{\partial \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} < \mu_n\tau_{n,j}\right\} \\ &\geq \mathrm{P}\left\{\frac{1}{b_n}\frac{\partial \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} < \mu_n\right\} \end{aligned}$$

which tends to 1 as $n \to \infty$ since by condition ($\mathcal{C}$.7) and (24)

$$\frac{1}{b_n}\frac{\partial \ell_\bullet(\boldsymbol{\beta}_\bullet; \mathbf{X})}{\partial(\boldsymbol{\beta}_\bullet)_j} = o_{\mathrm{P}}\left(\sqrt{\frac{\mu_n}{p_n}}\right) O_{\mathrm{P}}\left(p_n\sqrt{\mu_n}\right) = o_{\mathrm{P}}(\mu_n).$$

□

References

[1] Ismaïla Ba and Jean-François Coeurjolly. Inference for low-and high-dimensional inhomogeneous Gibbs point processes. *Scandinavian Journal of Statistics*, 50(3):993–1021, 2023.

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

[3] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: methodology and applications with R*. Chapman and Hall/CRC, 2015.

[4] Jean-Michel Billiot, Jean-François Coeurjolly, and Rémy Drouilhet. Maximum pseudolikelihood estimator for exponential family models of marked Gibbs point processes. *Electronic Journal of Statistics*, 2:243–264, 2008.

[5] Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232, 2011.

[6] Achmad Choiruddin, Jean-François Coeurjolly, and Frédérique Letué. Convex and non-convex regularization methods for spatial point processes intensity estimation. *Electronic Journal of Statistics*, 12(1):1210–1255, 2018.

[7] Achmad Choiruddin, Jean-François Coeurjolly, and Frédérique Letué. Adaptive lasso and Dantzig selector for spatial point processes intensity estimation. *Bernoulli*, 29(3):1849–1876, 2023.

[8] Achmad Choiruddin, Jean-François Coeurjolly, and Rasmus Waagepetersen. Information criteria for inhomogeneous spatial point processes. *Australian & New Zealand Journal of Statistics*, 63(1):119–143, 2021.

[9] Achmad Choiruddin, Tabita Yuni Susanto, Ahmad Husain, and Yuniar Mega Kartikasari. `kppmenet`: Combining the `kppm` and elastic net regularization for inhomogeneous Cox point process with correlated covariates. *Journal of Applied Statistics*, pages 1–14, 2023.

[10] Jean-François Coeurjolly and Rémy Drouilhet. Asymptotic properties of the maximum pseudo-likelihood estimator for stationary Gibbs point processes including the lennard-jones model. *Electronic Journal of Statistics*, 4:677–706, 2010.

[11] Jean-François Coeurjolly, Yongtao Guan, Mahdieh Khanmohammadi, and Rasmus Waagepetersen. Towards optimal Takacs–Fiksel estimation. *Spatial Statistics*, 18:396–411, 2016.

[12] Jean-François Coeurjolly and Frédéric Lavancier. Understanding Spatial Point Patterns Through Intensity and Conditional Intensities. In *Stochastic Geometry.*, volume 2237 of *Lecture Notes in Mathematics*, pages 45–85. Springer, 2019.

[13] Jean-François Coeurjolly and Ege Rubak. Fast covariance estimation for innovations computed from a spatial Gibbs point process. *Scandinavian Journal of Statistics*, 40(4):669–684, 2013.

[14] Jean-François Coeurjolly, David Dereudre, Rémy Drouilhet, and Frédéric Lavancier. Takacs–Fiksel method for stationary marked Gibbs point processes. *Scandinavian Journal of Statistics*, 39(3):416–443, 2012.

[15] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[16] Jeffrey Daniel, Julie Horrocks, and Gary J Umphrey. Penalized composite likelihoods for inhomogeneous Gibbs point process models. *Computational Statistics & Data Analysis*, 124:104–116, 2018.

[17] David Dereudre. Introduction to the theory of Gibbs point processes. In *Springer Lecture Notes in Stochastic Geometry*, pages 181–229. Springer, 2019.

[18] David Dereudre, Rémy Drouilhet, and Hans-Otto Georgii. Existence of Gibbsian point processes with geometry-dependent interactions. *Probability Theory and Related Fields*, 153(3-4):643–670, 2012.

[19] David Dereudre and Frédéric Lavancier. Campbell equilibrium equation and pseudo-likelihood estimation for non-hereditary Gibbs point processes. *Bernoulli*, 15(4):1368–1396, 2009.

[20] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[21] Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

[22] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[23] Xin Gao and Peter X-K Song. Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540, 2010.

[24] Christophe Giraud. *Introduction to high-dimensional statistics*. CRC Press, 2021.

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

[26] Yongtao Guan and Ye Shen. A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika*, 97(4):867–880, 2010.

[27] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[28] Francis KC Hui, David I Warton, and Scott D Foster. Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association*, 110(509):262–269, 2015.

[29] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons, 2008.

[30] Jens Ledet Jensen and Hans R Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46(3):475–486, 1994.

[31] Jens Ledet Jensen and Jesper Møller. Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability*, 1(3):445–461, 1991.

[32] Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 853–877, 2015.

[33] Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC, 2003.

[34] Stephen L Rathbun and Noel Cressie. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability*, 26(1):122–154, 1994.

[35] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[36] Andrew L Thurman, Rao Fu, Yongtao Guan, and Jun Zhu. Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica*, pages 173–188, 2015.

[37] Andrew L Thurman and Jun Zhu. Variable selection for spatial Poisson point processes via a regularization method. *Statistical Methodology*, 17:113–125, 2014.

[38] Thibault Vasseur, Jean-François Coeurjolly, and David Dereudre. Existence of inhomogeneous Gibbs point processes in the infinite volume. *Preprint*, 2020.

[39] Rasmus Waagepetersen and Yongtao Guan. Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):685–702, 2009.

[40] Rasmus Plenge Waagepetersen. An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics*, 63(1):252–258, 2007.

[41] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733, 2009.