Spatial point processes intensity estimation with a diverging number of covariates

Achmad Choiruddin, Jean-François Coeurjolly, Frédérique Letué

Spatial point processes intensity estimation with a diverging number of covariates

Achmad Choiruddin[1], Jean-François Coeurjolly[2, 3], and Frédérique Letué[3]

[1]Department of Mathematical Sciences, Aalborg University, Denmark
[2]Department of Mathematics, Université du Québec à Montréal (UQAM), Canada
[3]Laboratory Jean Kuntzmann, Department of Probability and Statistics, Université Grenoble Alpes, France

December 27, 2017

Abstract: Feature selection procedures for spatial point processes parametric intensity estimation have been recently developed since more and more applications involve a large number of covariates. In this paper, we investigate the setting where the number of covariates diverges as the domain of observation increases. In particular, we consider estimating equations based on Campbell theorems derived from Poisson and logistic regression likelihoods regularized by a general penalty function. We prove that, under some conditions, the consistency, the sparsity, and the asymptotic normality are valid for such a setting. We support the theoretical results by numerical ones obtained from simulation experiments and an application to forestry datasets.

Key words and phrases: Campbell formula, estimating equation, high dimensional regression, regularization method, variable selection.

1 Introduction

1.1 Background

Spatial point pattern data arise in many contexts, e.g. ecology (see e.g. Møller and Waagepetersen, 2004; Renner et al., 2015), epidemiology (e.g. Diggle, 1990, 2013), criminology (e.g. Baddeley et al., 2015; Shirota et al., 2017), biology (e.g. Illian et al., 2008) and astronomy (e.g. Baddeley et al., 2015), where interest lies in describing the distribution of an event in space. Stochastic models generating spatial

point patterns are called spatial point processes (see e.g. Møller and Waagepetersen, 2004; Illian et al., 2008; Diggle, 2013; Baddeley et al., 2015).

Usually, the first step to analyze spatial point pattern data is to investigate the intensity. The intensity serves as the first-order characteristics of a spatial point process and often becomes the main interest in many studies, especially when the intensity is suspected to depend on spatial covariates. Examples include the study of spatial variation of specific disease risk related to pollution sources (e.g. Diggle, 1990, 2013), crime rate analysis in a city related to some demographical information (e.g. Shirota et al., 2017), and modeling of the spatial distribution of trees species in a forest related to some environmental factors (e.g. Waagepetersen, 2007; Thurman et al., 2015; Renner et al., 2015).

We focus in this study on the log-linear model for the intensity function of an inhomogeneous spatial point process defined by

$$\rho(u;\boldsymbol{\beta}) = \exp(\mathbf{z}(u)^\top \boldsymbol{\beta}), u \in D \subset \mathbb{R}^d, \tag{1.1}$$

where $\mathbf{z}(u) = \{z_1(u), \ldots, z_p(u)\}^\top$ are the p spatial covariates measured at location u, d represents the state space of the spatial point processes (usually $d = 2, 3$) and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^\top$ is a real p-dimensional parameter. Hence, our main concern is to assess the magnitudes of the vector $\boldsymbol{\beta}$. For parametric estimation, while maximum likelihood estimation (e.g. Berman and Turner, 1992; Rathbun and Cressie, 1994) has been widely implemented for Poisson point processes models, estimating equation-based methods (e.g. Waagepetersen, 2007, 2008; Guan and Shen, 2010; Baddeley et al., 2014) are simpler to implement for more general spatial point processes models, overcoming the possible drawback of MCMC methods which are usually computational expensive (Møller and Waagepetersen, 2004). However, when the number of covariates is relatively large, maximum likelihood estimation and estimating equation-based methods become undesirable: all covariates are selected yielding an increasing standard error for parameter estimates.

1.2 Feature selection techniques

To select significant covariates, one may consider a traditional procedure such as a stepwise method. This technique starts with an initial set of covariates, then considers adding or deleting a covariate from the current set at each iteration using a criterion such as an F-statistic or AIC. However, such procedure has a number of limitations: it can be numerically unstable and exhibits high variance due to its discrete procedure (e.g. Breiman, 1996; Fan and Li, 2001; Friedman et al., 2008). It is even computationally unfeasible especially when the number of covariates is too large (e.g. Breiman, 1996; Zou, 2006).

To overcome this drawback, regularization techniques have recently been developed for spatial point processes intensity estimation. Such methods are able to perform variable selection while keeping interesting properties in terms of prediction. For Poisson point process models, the idea is to penalize the Poisson likelihood by a penalty function such as l_1 penalty (see Renner and Warton, 2013; Thurman and Zhu, 2014). For more general point process models, instead of employing the likelihood of the processes which often requires computational intensive MCMC methods

(Møller and Waagepetersen, 2004), penalized versions of estimating equations based on Campbell theorem derived both from Poisson and logistic regression likelihoods have been developed (see Thurman et al., 2015; Choiruddin et al., 2017). Furthermore, Thurman et al. (2015) and Choiruddin et al. (2017) show that, under some conditions, the estimates obtained from such procedures are consistent, sparse, and asymptotically normal.

1.3 Issues in high dimensional data

The motivation of our paper comes from a study of biodiversity in a 50-hectare region ($D = 1,000\text{m} \times 500\text{m}$) of the tropical moist forest of Barro Colorado Island (BCI) in central Panama, where censuses have been carried out such that all free-standing woody stems at least 10 mm diameter at breast height were identified, tagged, and mapped, resulting in maps of over 350,000 individual trees with around 300 species (see Condit, 1998; Hubbell et al., 1999, 2005). In the same region, many environmental covariates such as topographical attributes and soil properties have been also collected. In particular, we are interested to study the spatial distribution of 3,604 locations of *Beilschmiedia pendula Lauraceae* (BPL) trees and to model its intensity as a parametric function of 93 covariates consisting of 2 topological attributes, 13 soil properties and 78 interactions between two soil nutrients.

Although it seems that the number of covariates is not very large with respect to the number of data points, two hours are required to estimate the parameters and select covariates using a standard stepwise procedure. To do this, we use the `step` function in `R` which intrinsically assumes that $\mathbf{X}$ is a Poisson point process since we use the AIC in the stepwise procedure. For a general point process, other criteria could be investigated such as the one based on the F statistic, but they require to estimate the asymptotic covariance matrix of the estimates at each step: even if we know the right covariates, it is known as a difficult task (see e.g. Coeurjolly and Guan (2014)) especially when the number of parameters is large. That would easily triple the time of this estimation/selection procedure. To evaluate the performance of such a selection/estimation procedure, a simulation would be required which is unrealistic (1000 replications of a single model would take 250 days). This motivates us to consider regularization methods.

Thurman et al. (2015) and Choiruddin et al. (2017) are the first two theoretical works. Both these works have the important limitation that the number of covariates p is finite. We extend this in the present paper. Asymptotic properties which consider a diverging number of parameters for M-estimators have a long story (e.g Huber, 1973; Portnoy, 1984) but have recently been investigated for penalized regression estimators by Fan and Peng (2004); Zou and Zhang (2009). In particular, as argued by Fan and Peng (2004), even though the asymptotic properties (i.e., consistency, sparsity, and asymptotic normality) proposed by Fan and Li (2001) for penalized generalized linear models under the assumption that p is finite, are encouraging, there are many naive and simple model selection procedures which possess those properties. Establishing the validity of these asymptotic properties in a diverging number of parameters setting is, therefore, a major importance. We study this type of asymptotic properties in the spatial point processes framework. Hence,

our work can be regarded as an extension of the study conducted by Choiruddin et al. (2017).

A standard way of measuring asymptotic for spatial point process is the increasing domain asymptotic. Therefore, we investigate the problem where $p = p_n$ grows with $|D_n|$ the volume of the observation domain. In our setting, $|D_n|$ plays the same role as n, the number of observations, in standard problems such as in linear models or generalized linear models. We obtain consistency, sparsity, and asymptotic normality for our estimator. One of our main assumptions is that $p_n^3/|D_n| \to 0$ as $n \to \infty$, which is similar to the one required by Fan and Peng (2004) when $|D_n|$ is simply replaced by n (the sample size in their context).

Our results are general: (1) a large choice of penalty functions (either convex or non-convex function) and methods (e.g. ridge, lasso, elastic net, SCAD, and MC+) are available; (2) we include a large class of mixing spatial point processes. The implementation is done by combining the `spatstat` (Baddeley et al., 2015) `R` package with the two `R` packages implementing penalized methods for generalized linear models: `glmnet` (Friedman et al., 2010) and `ncvreg` (Breheny and Huang, 2011).

1.4 Outline of the paper

In Section 2, we introduce brief background on spatial point processes as well as regularization methods for spatial point processes intensity estimation. Section 3 presents our asymptotic results. We investigate in Section 4 the finite sample performance of our estimates in a simulation study and in an application to tropical forestry datasets. Conclusion and discussion are presented in Section 5. Proofs of the main results are postponed to Appendices A-C.

2 Regularization methods for spatial point processes

This section gives brief introduction on spatial point processes and reviews regularization methods for spatial point processes intensity estimation previously studied by Choiruddin et al. (2017) when the number of parameters is finite.

Let $\mathbf{X}$ be a spatial point process on $\mathbb{R}^d$. We view $\mathbf{X}$ as a locally finite random subset of $\mathbb{R}^d$. Let $D \subset \mathbb{R}^d$ be a compact set of Lebesgue measure $|D|$ which will play the role of the observation domain. A realization of $\mathbf{X}$ in D is thus a set $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, where $x \in D$ and m is the observed number of points in D. Suppose $\mathbf{X}$ has intensity function ρ and second-order product density $\rho^{(2)}$. Campbell theorem (see e.g. Møller and Waagepetersen, 2004) states that, for any function $k : \mathbb{R}^d \to [0, \infty)$ or $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$

$$\mathbb{E}\Big(\sum_{u \in \mathbf{X}} k(u)\Big) = \int_{\mathbb{R}^d} k(u)\rho(u)\mathrm{d}u \tag{2.2}$$

$$\mathbb{E}\Big(\sum_{u,v \in \mathbf{X}}^{\neq} k(u,v)\Big) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(u,v)\rho^{(2)}(u,v)\mathrm{d}u\mathrm{d}v. \tag{2.3}$$

We may interpret $\rho(u)\mathrm{d}u$ as the probability of occurence of a point in an infinitesimally small ball with centre u and volume $\mathrm{d}u$. Intuitively, $\rho^{(2)}(u,v)\mathrm{d}u\mathrm{d}v$ is the probability for observing a pair of distinct points from $\mathbf{X}$ occuring jointly in each of two infinitesimally small balls with centres u, v and volume $\mathrm{d}u, \mathrm{d}v$. For further background materials on spatial point processes, see for example Møller and Waagepetersen (2004); Illian et al. (2008).

In our study, we assume that the intensity function depends on parameter $\boldsymbol{\beta}$, $\rho = \rho(\cdot;\boldsymbol{\beta})$. The standard parametric methods for estimating $\boldsymbol{\beta}$ are by maximizing the weighted Poisson likelihood (e.g. Guan and Shen, 2010) or the weighted logistic regression likelihood (e.g. Baddeley et al., 2014; Choiruddin et al., 2017) given respectively by

$$\ell_{\mathrm{PL}}(w;\boldsymbol{\beta}) = \sum_{u\in\mathbf{X}\cap D} w(u)\log\rho(u;\boldsymbol{\beta}) - \int_D w(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u, \tag{2.4}$$

$$\begin{aligned}\ell_{\mathrm{LRL}}(w;\boldsymbol{\beta}) = &\sum_{u\in\mathbf{X}\cap D} w(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})}{\delta(u)+\rho(u;\boldsymbol{\beta})}\right)\\ &- \int_D w(u)\delta(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})+\delta(u)}{\delta(u)}\right)\mathrm{d}u,\end{aligned} \tag{2.5}$$

where $w(\cdot)$ is a weight non-negative function depending on the first and the second-order characterictics of $\mathbf{X}$ and $\delta(\cdot)$ is a non-negative real-valued function. The solution of maximizing (2.4) (resp. (2.5)) is called Poisson estimator (resp. the logistic regression estimator). We refer readers to Guan and Shen (2010) for further details on the weight function $w(\cdot)$ and to Baddeley et al. (2014) for the role of function $\delta(\cdot)$.

These standard methods cannot perform variable selection. To do so, Thurman et al. (2015) and Choiruddin et al. (2017) suggest to maximize a penalized version of (2.4)-(2.5)

$$Q(w;\boldsymbol{\beta}) = \ell(w;\boldsymbol{\beta}) - |D|\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{2.6}$$

where $\ell(w;\boldsymbol{\beta})$ is either the Poisson likelihood (2.4) or the logistic regression likelihood (2.5). We refer the second term of (2.6) to a penalization term. In this term, we have mainly two parts: (1) a penalty function p_λ parameterized by $\lambda \geq 0$ and (2) the volume of the observation domain $|D|$ which plays the same role as the sample size in the spatial point process framework.

For any non-negative λ, we say that $p_\lambda(\cdot)$ is a penalty function if p_λ is a non-negative function with $p_\lambda(0) = 0$. Some examples, described in Table 1, include l_2 penalty (Hoerl and Kennard, 1988), l_1 penalty (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), and MC+ (Zhang, 2010). Note that, as indicated by (2.6), we allow each direction to have different tuning parameters $\lambda_j, j = 1,\ldots,p$. Such a method is called an adaptive method (e.g. adaptive lasso (Zou, 2006) and adaptive elastic net (Zou and Zhang, 2009)). For further backgrounds about penalty function and regularization methods, see, for example, Friedman et al. (2008).

Table 1: Examples of penalty function.

Penalty	$p_\lambda(\theta)$
l_2 penalty	$\frac{1}{2}\lambda\theta^2$
l_1 penalty	$\lambda\|\theta\|$
Enet	$\lambda\{\gamma\|\theta\| + \frac{1}{2}(1-\gamma)\theta^2\}$, for any $0 < \gamma < 1$
SCAD	$\lambda\theta\mathbb{I}(\theta \leq \lambda) + \frac{\gamma\lambda\theta - \frac{1}{2}(\theta^2+\lambda^2)}{\gamma-1}\mathbb{I}(\lambda \leq \theta \leq \gamma\lambda) + \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}\mathbb{I}(\theta \geq \gamma\lambda)$, for any $\gamma > 2$
MC+	$\left(\lambda\theta - \frac{\theta^2}{2\gamma}\right)\mathbb{I}(\theta \leq \gamma\lambda) + \frac{1}{2}\gamma\lambda^2\mathbb{I}(\theta \geq \gamma\lambda)$, for any $\gamma > 1$

3 Asymptotic properties

In this section, we present asymptotic properties of the regularized Poisson estimator when both $|D_n| \to \infty$ and $p_n \to \infty$ as $n \to \infty$. In particular, we consider $\mathbf{X}$ as a d-dimensional point process observed over a sequence of observation domain $D = D_n, n = 1, 2, \ldots$ which expands to $\mathbb{R}^d$ as $n \to \infty$. We assume that $\mathbf{X}$ has a log-linear form given by (1.1) for which the dimension of parameter $\boldsymbol{\beta}$, denoted now by p_n, diverges to ∞ as $n \to \infty$. In Section 3.1, we provide notation and conditions, and discuss the differences from the setting where p is fixed. Our main results are presented in Section 3.2. For sake of conciseness, we do not present the asymptotic results for the regularized logistic regression estimator. The results are very similar. The main difference is lying in the conditions ($\mathcal{C}$.6) and ($\mathcal{C}$.7) for which the matrices $\mathbf{A}_n, \mathbf{B}_n$, and $\mathbf{C}_n$ have a different expression (see Remark 2).

3.1 Notation and conditions

Throughout this section and Appendices A-C, let

$$\begin{aligned}\ell_n(w;\boldsymbol{\beta}) =& \ell_{n,\mathrm{PL}}(w;\boldsymbol{\beta}) \\ =& \sum_{u\in\mathbf{X}\cap D_n} w(u)\log\rho(u;\boldsymbol{\beta}) - \int_{D_n} w(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u, \qquad (3.7)\\ Q_n(w;\boldsymbol{\beta}) =& \ell_n(w;\boldsymbol{\beta}) - |D_n|\sum_{j=1}^{p_n} p_{\lambda_{n,j}}(|\beta_j|), \qquad (3.8)\end{aligned}$$

be respectively the weighted Poisson likelihood and its penalized version.

Let $\boldsymbol{\beta}_0 = \{\beta_{01}, \ldots, \beta_{0s}, \beta_{0(s+1)}, \ldots, \beta_{0p_n}\}^\top = \{\boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{02}^\top\}^\top = (\boldsymbol{\beta}_{01}^\top, \mathbf{0}^\top)^\top$ denote the p_n-dimensional vector to estimate, where $\boldsymbol{\beta}_{01}$ is the s-dimensional vector of non-zero coefficients and $\boldsymbol{\beta}_{02}$ is the $(p_n - s)$-dimensional vector of zero coefficients. We assume that the number of non-zero coefficients, s, does not depend on n. Let $\mathbf{z}_{01}$ and $\mathbf{z}_{02}$ denote the corresponding s-dimensional and $(p_n - s)$-dimensional vectors of spatial covariates. We denote the regularized Poisson estimator by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$.

We recall the classical definition of strong mixing coefficients adapted to spatial point processes (e.g. Politis et al., 1998): for $k, l \in \mathbb{N} \cup \{\infty\}$ and $q \geq 1$, define

$$\alpha_{k,l}(q) = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathscr{F}(\Lambda_1), B \in \mathscr{F}(\Lambda_2), \\ \Lambda_1 \in \mathscr{B}(\mathbb{R}^d), \Lambda_2 \in \mathscr{B}(\mathbb{R}^d), |\Lambda_1| \leq k, |\Lambda_2| \leq l, d(\Lambda_1, \Lambda_2) \geq q\}, \quad (3.9)$$

where $\mathscr{F}$ is the σ-algebra generated by $\mathbf{X} \cap \Lambda_i, i = 1, 2, d(\Lambda_1, \Lambda_2)$ is the minimal distance between sets Λ_1 and Λ_2, and $\mathscr{B}(\mathbb{R}^d)$ denotes the class of Borel sets in $\mathbb{R}^d$.

We define the $p_n \times p_n$ matrices $\mathbf{A}_n(w; \boldsymbol{\beta}_0), \mathbf{B}_n(w; \boldsymbol{\beta}_0)$ and $\mathbf{C}_n(w; \boldsymbol{\beta}_0)$ by

$$\mathbf{A}_n(w; \boldsymbol{\beta}_0) = \int_{D_n} w(u)\mathbf{z}(u)\mathbf{z}(u)^\top \rho(u; \boldsymbol{\beta}_0)\mathrm{d}u,$$
$$\mathbf{B}_n(w; \boldsymbol{\beta}_0) = \int_{D_n} w(u)^2\mathbf{z}(u)\mathbf{z}(u)^\top \rho(u; \boldsymbol{\beta}_0)\mathrm{d}u,$$
$$\mathbf{C}_n(w; \boldsymbol{\beta}_0) = \int_{D_n}\int_{D_n} w(u)w(v)\mathbf{z}(u)\mathbf{z}(v)^\top \{g(u, v) - 1\}\rho(u; \boldsymbol{\beta}_0)\rho(v; \boldsymbol{\beta}_0)\mathrm{d}u\mathrm{d}v,$$

where $g(u, v)$ is the classical pair correlation function (Møller and Waagepetersen, 2004) given by

$$g(u, v) = \frac{\rho^{(2)}(u, v)}{\rho(u)\rho(v)},$$

when both ρ and $\rho^{(2)}$ exist with the convention $0/0 = 0$. For a Poisson point process, we have $g(u, v) = 1$ since $\rho^{(2)}(u, v) = \rho(u)\rho(v)$. If, for example, $g(u, v) > 1$ (resp. $g(u, v) < 1$), this indicates that pair of points are more likely (resp. less likely) to occur at locations u, v than for a Poisson point process.

We denote the $s \times s$ top-left corner of $\mathbf{A}_n(w; \boldsymbol{\beta}_0)$ (resp. $\mathbf{B}_n(w; \boldsymbol{\beta}_0)$, $\mathbf{C}_n(w; \boldsymbol{\beta}_0)$) by $\mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)$ (resp. $\mathbf{B}_{n,11}(w; \boldsymbol{\beta}_0), \mathbf{C}_{n,11}(w; \boldsymbol{\beta}_0)$). It is worth noticing that $\mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)$, $\mathbf{B}_{n,11}(w; \boldsymbol{\beta}_0)$ and $\mathbf{C}_{n,11}(w; \boldsymbol{\beta}_0)$ depend on n only through D_n and not through p_n. In what follows, for a squared symmetric matrix $\mathbf{M}_n$, $\nu_{\min}(\mathbf{M}_n)$ and $\nu_{\max}(\mathbf{M}_n)$ denote respectively the smallest and largest eigenvalue of $\mathbf{M}_n$.

Under the conditions ($\mathcal{C}$.8)-($\mathcal{C}$.9), we define the sequences a_n, b_n and c_n by

$$a_n = \max_{j=1,\ldots,s} |p'_{\lambda_{n,j}}(|\beta_{0j}|)|, \quad (3.10)$$

$$b_n = \inf_{j=s+1,\ldots,p_n} \inf_{\substack{|\theta| \leq \epsilon_n \\ \theta \neq 0}} p'_{\lambda_{n,j}}(\theta), \text{ for } \epsilon_n = K_1\sqrt{\frac{p_n}{|D_n|}}, \quad (3.11)$$

$$c_n = \max_{j=1,\ldots,s} |p''_{\lambda_{n,j}}(|\beta_{0j}|)|, \quad (3.12)$$

where K_1 is any positive constant.

Consider the following conditions ($\mathcal{C}$.1)-($\mathcal{C}$.9) which are required to derive our asymptotic results:

($\mathcal{C}$.1) For every $n \geq 1, D_n = nE = \{ne : e \in E\}$, where $E \subset \mathbb{R}^d$ is convex, compact, and contains the origin of $\mathbb{R}^d$ in its interior.

($\mathcal{C}$.2) The intensity function has the log-linear specification given by (1.1) where $\boldsymbol{\beta} \in \Theta$ and Θ is an open convex bounded set of $\mathbb{R}^{p_n}$. Furthermore, we assume that there exists a neighborhood $\Xi(\boldsymbol{\beta}_0)$ of $\boldsymbol{\beta}_0$ such that

$$\sup_{n\geq 1} \sup_{\boldsymbol{\beta}\in\Xi(\boldsymbol{\beta}_0)} \sup_{u\in\mathbb{R}^d} \rho(u;\boldsymbol{\beta}) < \infty.$$

($\mathcal{C}$.3) The covariates $\mathbf{z}$ and the weight function w satisfy

$$\sup_{n\geq 1} \sup_{i=1,\ldots,p_n} \sup_{u\in\mathbb{R}^d} |z_i(u)| < \infty, \quad \text{and} \quad \sup_{u\in\mathbb{R}^d} w(u) < \infty.$$

($\mathcal{C}$.4) There exists an integer $t \geq 1$ such that for $k = 2, \ldots, 2+t$, the product density $\rho^{(k)}$ exists and satisfies $\rho^{(k)} < \infty$.

($\mathcal{C}$.5) For the strong mixing coefficients (3.9), we assume that there exists some $\tilde{t} > d(2+t)/t$ such that $\alpha_{2,\infty}(q) = O(q^{-\tilde{t}})$.

($\mathcal{C}$.6) $\liminf_n \nu_{\min}\big(|D_n|^{-1}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}\big) > 0$.

($\mathcal{C}$.7) $\liminf_n \nu_{\min}\big(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0)\big) > 0$.

($\mathcal{C}$.8) The penalty function $p_\lambda(\cdot)$ is non-negative on $\mathbb{R}^+$, continuously differentiable on $\mathbb{R}^+ \setminus \{0\}$ with derivative p'_λ assumed to be a Lipschitz function on $\mathbb{R}^+ \setminus \{0\}$. Furthermore, given $(\lambda_{n,j})_{n\geq 1}$, for $j = 1, \ldots, s$, we assume that there exists $(\tilde{r}_{n,j})_{n\geq 1}$, where $\tilde{r}_{n,j}\sqrt{|D_n|/p_n} \to \infty$ as $n \to \infty$, such that, for n sufficiently large, $p_{\lambda_{n,j}}$ is thrice continuously differentiable in the ball centered at $|\beta_{0j}|$ with radius $\tilde{r}_{n,j}$ and we assume that the third derivative is uniformly bounded.

($\mathcal{C}$.9) $p_n^3/|D_n| \to 0$ as $n \to \infty$.

Conditions ($\mathcal{C}$.1)-($\mathcal{C}$.8) are quite similar to the ones required by Choiruddin et al. (2017) in the setting when the number of parameters to estimate is fixed. Condition ($\mathcal{C}$.2) is slightly stronger since we have to ensure that $\rho(u;\boldsymbol{\beta})$ is finite for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}_0$. Note that $\sup_{u\in\mathbb{R}^d} \rho(u;\boldsymbol{\beta}_0) < \infty$ follows directly from condition ($\mathcal{C}$.3). We derive asymptotic properties when both $|D_n|$ and p_n tend to infinity with n. However, to obtain an estimator which is consistent and has two other properties: sparsity and asymptotic normality, we need that the number of covariates does not grow too fast with respect to the volume of the observation domain. This condition is stated by condition ($\mathcal{C}$.9) which is similar to the one required by Fan and Peng (2004) when $|D_n|$ is simply replaced by n (the sample size in their context).

3.2 Main results

We state our main results here. Proofs are relegated to Appendices A-C.

We first show in Theorem 1 that the regularized Poisson estimator converges in probability and exhibits its rate of convergence.

Theorem 1. *Assume the conditions (C.1)-(C.5) and (C.7)-(C.9) hold. Let a_n and c_n be given respectively by (3.10) and (3.12). If $a_n = O(|D_n|^{-1/2})$ and $c_n = o(1)$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q_n(w;\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}\| = O_{\mathrm{P}}\big(\sqrt{p_n}(|D_n|^{-1/2} + a_n)\big)$.*

This implies that, the regularized Poisson estimator is root-$(|D_n|/p_n)$ consistent. Note that, as expected, the convergence rate is $\sqrt{p_n}$ times the convergence rate of the estimator obtained when p is fixed (see Theorem 1 Choiruddin et al., 2017). In addition, when we compare our results with the ones obtained by Fan and Peng (2004), who also considered a diverging number of parameters setting, our estimator has the same rate of convergence when we replace $|D_n|$ by n to their context. This rate of convergence also appears in other contexts considering diverging number of parameters setting (see e.g. Lam and Fan, 2008; Zou and Zhang, 2009; Li et al., 2011; Cho and Qu, 2013; Wang and Zhu, 2017).

Now, we demonstrate in Theorem 2 that such a root-$(|D_n|/p_n)$ consistent estimator ensures the sparsity of $\hat{\boldsymbol{\beta}}$; that is, the estimate will correctly set $\boldsymbol{\beta}_2$ to zero with probability tending to 1 as $n \to \infty$, and $\hat{\boldsymbol{\beta}}_1$ is asymptotically normal.

Theorem 2. *Assume the conditions (C.1)-(C.9) are satisfied. If $a_n\sqrt{|D_n|} \to 0$, $b_n\sqrt{|D_n|/p_n^2} \to \infty$ and $c_n\sqrt{p_n} \to 0$ as $n \to \infty$, the root-$(|D_n|/p_n)$ consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$ in Theorem 1 satisfies:*

(i) Sparsity: $\mathrm{P}(\hat{\boldsymbol{\beta}}_2 = 0) \to 1$ *as* $n \to \infty$,

(ii) Asymptotic Normality: $|D_n|^{1/2}\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_s)$,

where

$$\begin{aligned}\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0) =& |D_n|\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\boldsymbol{\Pi}_n\}^{-1}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\} \\ & \{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\boldsymbol{\Pi}_n\}^{-1}, \qquad (3.13)\\ \boldsymbol{\Pi}_n =& \mathrm{diag}\{p''_{\lambda_{n,1}}(|\beta_{01}|), \ldots, p''_{\lambda_{n,s}}(|\beta_{0s}|)\}. \qquad (3.14)\end{aligned}$$

As a consequence, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$. Here, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}$ is the inverse of $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}$, where $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}$ is any square matrix with $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\big(\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\big)^\top = \boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$.

Remark 1. For lasso and adaptive lasso, $\boldsymbol{\Pi}_n = \mathbf{0}$. For other penalties, since $c_n = o(1)$, then $\|\boldsymbol{\Pi}_n\| = o(1)$. Since $\|\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\| = O(|D_n|)$ from conditions (C.1)-(C.3), $|D_n|\,\|\boldsymbol{\Pi}_n\|$ is asymptotically negligible with respect to $\|\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\|$.

Remark 2. Theorems 1 and 2 remain true for the regularized logistic regression estimator if we replace in the expression of the matrices $\mathbf{A}_n, \mathbf{B}_n$, and $\mathbf{C}_n$, $w(u)$ by $w(u)\delta(u)/(\rho(u;\boldsymbol{\beta}_0)+\delta(u)), u \in D_n$ and extend the condition ($\mathcal{C}$.3) by adding $\sup_{u\in\mathbb{R}^d}\delta(u) < \infty$.

The proofs of Theorems 1 and 2 for this estimator are slightly different mainly because unlike the Poisson likelihood for which we have $\ell_n^2(w;\boldsymbol{\beta}) = -\mathbf{A}_n(w;\boldsymbol{\beta})$, for the regularized logistic regre $\ell_n^2(w;\boldsymbol{\beta})$ is now stochastic and we only have $\mathbb{E}(\ell_n^2(w;\boldsymbol{\beta})) = -\mathbf{A}_n(w;\boldsymbol{\beta})$. Despite the additional difficulty, we maintain that no additional assumption is required.

We show in Theorem 2 that the sparsity and asymptotic normality are still valid when the number of parameters diverges. By Remark 1, when n is large enough, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ in (3.13) becomes approximately

$$|D_n|\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0)+\mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1},$$

which is precisely the asymptotic covariance matrix of the estimator of $\boldsymbol{\beta}_{01}$ obtained by maximizing the likelihood function or solving estimating equations based on the submodel knowing that $\boldsymbol{\beta}_{02} = \mathbf{0}$. This shows that when n is sufficiently large, our estimator is as efficient as the oracle one.

To satisfy Theorem 2, we require that $a_n\sqrt{|D_n|} \to 0$, $b_n\sqrt{|D_n|/p_n^2} \to \infty$ and $c_n\sqrt{p_n} \to 0$ as $n \to \infty$ simultaneously. In particular, conditions on a_n and c_n ensure the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$ while condition on b_n is used to prove the sparsity. Conditions regarding a_n and c_n are similar to the ones imposed by Fan and Peng (2004) when $|D_n|$ is replaced by n in their context. However, we require a slightly stronger condition on b_n than the one required by Fan and Peng (2004) which in the present setting could be written as $b_n\sqrt{|D_n|/p_n} \to \infty$. As compensation, we do not need to impose, as Fan and Peng (2004) did, for any $0 < K_2 < \infty$, $\nu_{\max}\big(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0)\big) < K_2$. Such a condition is not straightforwardly satisfied in our setting since the other conditions only imply that $\nu_{\max}\big(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0)\big) = O(p_n)$.

Further details regarding a_n, b_n and c_n for each method are presented in Table 2. For the ridge regularization method, $b_n = 0$, preventing from applying Theorem 2 for this penalty. For lasso and elastic net, $a_n = K_3 b_n$ for some constant $K_3 > 0$ (K_3=1 for lasso). The two conditions $a_n\sqrt{|D_n|} \to 0$ and $b_n\sqrt{|D_n|/p_n^2} \to \infty$ as $n \to \infty$ cannot be satisfied simultaneously. This is different for the adaptive versions where a compromise can be found by adjusting the $\lambda_{n,j}$'s, as well as the two non-convex penalties SCAD and MC+, for which λ_n can be adjusted. For the regularization methods we consider in this study, the condition $c_n\sqrt{p_n} \to 0$ is implied by the condition $a_n\sqrt{|D_n|} \to 0$ as $n \to \infty$ and condition ($\mathcal{C}$.9).

4 Numerical results

This section is devoted to present numerical results. More precisely, we conduct simulation experiments in Section 4.1 to assess the finite sample peformance of our estimates and apply our method to an application in ecology in Section 4.2. We

Table 2: Details of the sequences a_n, b_n and c_n for a given regularization method.

Method	a_n	b_n	c_n
Ridge	$\lambda_n \max_{j=1,\dots s} \{\lvert\beta_{0j}\rvert\}$	0	λ_n
Lasso	λ_n	λ_n	0
Enet	$\lambda_n \left[(1-\gamma) \max_{j=1,\dots s} \{\lvert\beta_{0j}\rvert\} + \gamma\right]$	$\gamma\lambda_n$	$(1-\gamma)\lambda_n$
AL	$\max_{j=1,\dots s} \{\lambda_{n,j}\}$	$\min_{j=s+1,\dots p_n} \{\lambda_{n,j}\}$	0
Aenet	$\max_{j=1,\dots s} \{\lambda_{n,j}\big((1-\gamma)\lvert\beta_{0j}\rvert + \gamma\big)\}$	$\gamma \min_{j=s+1,\dots p_n} \{\lambda_{n,j}\}$	$(1-\gamma) \max_{j=1,\dots,s} \{\lambda_{n,j}\}$
SCAD	0^*	${\lambda_n}^{**}$	0^*
MC+	0^*	$\lambda_n - \frac{K_1\sqrt{p_n}}{\gamma\sqrt{\lvert D_n\rvert}}$ **	0^*

* if $\lambda_n \to 0$ for n sufficient large
** if $\lambda_n \sqrt{|D_n|/p_n^2} \to \infty$ for n sufficient large

apply the regularized Poisson likelihood (PL) and the regularized weighted Poisson likelihood (WPL) to select covariates and estimate their coefficients. Similar approach can be straightforwardly used for the regularized versions using logistic regression likelihood.

To numerically evaluate the parameters estimates, we apply Berman-Turner method (Berman and Turner, 1992) combined with coordinate descent algorithm (Friedman et al., 2007) to perform variable selection and parameter estimation. Berman-Turner device allows to show that maximizing (2.4) is equivalent to fitting a weighted Poisson generalized linear model, so the standard software for generalized linear models (GLMs) can be used. This has been exploited by the `spatstat` `R` package (Baddeley et al., 2015). As we make links between spatial point processes intensity estimation and GLMs, we only have to deal with feature selection procedures for GLMs. Hence, we clearly have many advantages: the various computational strategies are carefully studied, and, in particular, efficiently implemented in `R`. In this study, to compute the regularization path solutions, we employ coordinate descent algorithm (Friedman et al., 2007). This is implemented in the `glmnet` (Friedman et al., 2010) for regularization methods for GLMs using some convex penalties (i.e., ridge, lasso, elastic net, adaptive lasso and adaptive elastic net) and in the `ncvreg` (Breheny and Huang, 2011) for regularization methods for GLMs using some non-convex penalties (i.e., SCAD and MC+). More details for computational strategies are discussed in detail by Choiruddin et al. (2017).

Our methods rely on the tuning parameter λ. Some previous studies (see e.g. Zou et al., 2007; Wang et al., 2007, 2009) suggest to use a modified BIC criterion to select the tuning parameter. We follow the literature and choose λ by minimizing

WQBIC(λ), a modified version of the BIC criterion, defined by

$$\text{WQBIC}(\lambda) = -2\ell(w; \hat{\boldsymbol{\beta}}(\lambda)) + s(\lambda) \log |D|,$$

where $s(\lambda) = \sum_{j=1}^{p} \mathbb{I}\{\hat{\beta}_j(\lambda) \neq 0\}$ is the number of selected covariates with non-zero regression coefficients and $|D|$ is the volume of observation domain. To implement the adaptive methods (i.e., adaptive lasso and adaptive elastic net), we follow Zou (2006) and define $\lambda_j = \lambda/|\tilde{\beta}_j(ridge)|, j = 1, \cdots, p$, where $\tilde{\boldsymbol{\beta}}(ridge)$ is the estimates obtained from ridge regression and λ is a tuning parameter chosen by WQBIC(λ) criterion as described above. Following Choiruddin et al. (2017), we fix $\gamma = 0.5$ for elastic net and its adaptive version, $\gamma = 3.7$ for SCAD, and $\gamma = 3$ for MC+. For further discussion regarding the selection of γ for SCAD and MC+, see e.g. Fan and Li (2001) and Breheny and Huang (2011).

4.1 Simulation study

In this section, we investigate the behavior of our estimators in a simulation experiment in different situations when a large number of covariates for fitting spatial point process intensity estimation is involved. We intend to extend the setting considered by Choiruddin et al. (2017). We start with relatively complex situation where strong multicollinearity is present (Scenarios 1a and 2a) and we then consider a more complex setting using real datasets (Scenarios 1b and 2b). We have two different scenarios (Scenarios 1 and 2) for which the number of true covariates as well as their coefficients are different.

The spatial domain we consider is $D = [0, 1000] \times [0, 500]$. The true intensity function has the form $\rho(u; \boldsymbol{\beta}_0) = \exp(\mathbf{z}(u)^\top \boldsymbol{\beta}_0)$, where $\mathbf{z}(u) = \{1, z_1(u), \ldots, z_{50}(u)\}^\top$ and $\boldsymbol{\beta}_0 = \{\beta_0, \beta_{01}, \cdots, \beta_{050}\}$. We set β_0 such that the mean number of points over D is equal to 1600. We consider two different scenarios described as follows.

Scenario 1. We define the true vector $\boldsymbol{\beta}_0 = \{\beta_0, 2, 0.75, 0, \cdots, 0\}$. To define the covariates, we center and scale the 201×101 pixel images of elevation (x_1) and gradient of elevation (x_2) contained in the `bei` datasets of `spatstat` library in `R` and use them as two true covariates. In addition, we create two settings to define extra covariates:

a. First, we generate 48 201×101 pixel images of covariates as a standard Gaussian white noise and denote them by $x_3, \ldots, x_{50}$. Second, we transform them, together with x_1 and x_2, to have multicollinearity. In particular, we define $\tilde{\mathbf{z}}(u) = \mathbf{V}^\top \mathbf{x}(u)$, where $\mathbf{x}(u) = \{x_1(u), \ldots, x_{50}(u)\}^\top$. More precisely, $\mathbf{V}$ is such that $\boldsymbol{\Omega} = \mathbf{V}^\top \mathbf{V}$, and $(\Omega)_{ij} = (\Omega)_{ji} = 0.7^{|i-j|}$ for $i, j = 1, \ldots, 50$, except $(\Omega)_{12} = (\Omega)_{21} = 0$, to preserve the correlation between x_1 and x_2. In this setting, $\mathbf{z}(u) = \{1, \tilde{\mathbf{z}}(u)\}$.

b. We center and scale the 13 50×25 pixel images of soil nutrients obtained from the study in tropical forest of Barro Colorado Island (BCI) in central Panama (see Condit, 1998; Hubbell et al., 1999, 2005) and convert them to be 201×101 pixel images as x_1 and x_2.

In addition, we consider the interaction between two soil nutrients such that we have 50 covariates in total. We use 48 covariates (13 soil nutrients and 35 interactions between them) as the extra covariates. Together with x_1 and x_2, we keep the structure of the covariance matrix to preserve the complexity of the situation. In this setting, we have $\mathbf{z}(u) = \mathbf{x}(u) = \{1, x_1(u), \ldots, x_{50}(u)\}^\top$.

Scenario 2. In this setting, we consider five true covariates out of 50 covariates. In addition of elevation (x_1) and gradient of elevation (x_2), we convert 50×25 pixel images of concentration of Aluminium (x_3), Boron (x_4) and Calcium (x_5) in the soil to be 201×101 pixel images as x_1 and x_2 and set them to be other three true covariates. All five covariates are centered and scaled. We define the true coefficient vector $\boldsymbol{\beta}_0 = \{\beta_0, 5, 4, 3, 2, 1, 0, \cdots, 0\}$. As in Scenario 1, we make two settings to define 45 extra covariates:

a. This setting is similar to that of Scenario 1a. We generate 45 201×101 pixel images of covariates as standard Gaussian white noise, denote them by $x_6, \ldots, x_{50}$, and define $\tilde{\mathbf{z}}(u) = \mathbf{V}^\top \mathbf{x}(u)$, where $\mathbf{V}$ is such that $\boldsymbol{\Omega} = \mathbf{V}^\top \mathbf{V}$, and $(\Omega)_{ij} = (\Omega)_{ji} = 0.7^{|i-j|}$ for $i, j = 1, \ldots, 50$, except $(\Omega)_{kl} = (\Omega)_{lk} = 0$, for $k, l = 1, \cdots, 5, k \neq l$, to preserve the correlation among x_1 - x_5. We still define $\mathbf{z}(u) = \{1, \tilde{\mathbf{z}}(u)\}$.

b. We use the real dataset as in Scenario 1b and consider similar setting. In this setting, we define 5 true covariates which have different regression coefficients as in Scenario 1b.

With these scenarios, we simulate 2000 spatial point patterns from a Thomas point process using the `rThomas` function in the `spatstat` package. We set the interaction parameter κ to be $\kappa = 5 \times 10^{-4}, \kappa = 5 \times 10^{-5}$ and let $\omega = 20$. Briefly, smaller values of ω correspond to tighter clusters, and smaller values of κ correspond to a fewer number of parents (see e.g. Møller and Waagepetersen, 2004, for further details regarding the Thomas point process). For each scenario with different κ, we fit the intensity to the simulated point pattern realizations.

We report the performances of our estimates in terms of two characteristics: selection and prediction properties. We present the selection properties in Table 3 and the prediction properties in Table 4

To evaluate the selection properties of the estimates, we consider the true positive rate (TPR), the false positive rate (FPR), and the positive predictive value (PPV). We want to find the methods which have a TPR close to 100% meaning that it can select correctly all the true covariates, a FPR close to 0 showing that it can remove all the extra covariates from the model, and a PPV close to 100% indicating that, for Scenario 1 (resp. Scenario 2), it can keep exactly the two (resp. five) true covariates and remove all the 48 (resp. 45) extra covariates. In general, for both regularized PL and regularized WPL, the best selection properties are obtained from larger κ (5×10^{-4}) which indicates weaker spatial dependence. To compare the regularization methods, we emphasize here that the main difference between regularization methods which satisfy (adaptive lasso, adaptive elastic net, SCAD,

Table 3: Empirical selection properties (TPR, FPR, and PPV in %) based on 2000 replications of Thomas processes on the domain D for two different values of κ and for the two different scenarios. Different penalty functions are considered as well as two estimating equations, the regularized Poisson likelihood (PL) and the regularized weighted Poisson likelihood (WPL).

Method	Regularized PL			Regularized WPL			Regularized PL			Regularized WPL		
	$\kappa = 5 \times 10^{-4}$						$\kappa = 5 \times 10^{-5}$					
	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV
	Scenario 1a											
Lasso	100[1]	13	28	96	4	62	97	23	20	64	1	76
Enet	100[1]	34	12	93	8	48	97	48	10	59	2	58
AL	100[1]	1	92	97	0[1]	96	95	3	68	70	0[1]	98
Aenet	100[1]	2	76	97	1	85	95	6	52	67	0[1]	95
SCAD	100[1]	7	41	97	1	87	96	4	61	56	0[1]	79
MC+	100[1]	8	37	96	1	85	96	5	58	52	1	74
	Scenario 1b											
Lasso	100[1]	45	10	91	11	52	100[1]	96	4	20	6	22
Enet	100[1]	63	7	87	18	31	100[1]	98	4	15	6	14
AL	100[1]	26	19	95	5	81	99	85	5	26	5	35
Aenet	100[1]	30	15	95	6	74	100[1]	87	5	24	5	30
SCAD	100[1]	26	18	93	5	76	100[1]	76	5	23	4	28
MC+	100[1]	26	17	93	5	76	99	76	5	22	5	27
	Scenario 2a											
Lasso	98	93	10	84	73	14	98	96	10	47	35	16
Enet	99	98	10	85	80	11	99	98	10	46	38	12
AL	95	49	18	83	35	27	95	64	15	50	23	28
Aenet	96	52	17	84	40	21	96	68	14	48	26	20
SCAD	86	74	13	65	45	36	75	60	21	39	26	30
MC+	87	78	13	65	47	35	73	60	22	39	26	30
	Scenario 2b											
Lasso	80	64	13	75	60	12	78	69	11	64	57	9
Enet	85	73	12	82	69	11	84	79	11	68	64	8
AL	56	26	19	54	25	20	59	35	17	48	30	13
Aenet	59	30	18	57	29	18	64	43	15	52	36	11
SCAD	43	21	20	42	20	23	46	24	27	41	25	16
MC+	44	21	20	43	20	23	46	24	26	41	26	16

[1] Approximate value

Table 4: Empirical prediction properties (Bias, SD, and RMSE) based on 2000 replications of Thomas processes on the domain D for two different values of κ and for the two different scenarios. Different penalty functions are considered as well as two estimating equations, the regularized Poisson likelihood (PL) and the regularized weighted Poisson likelihood (WPL).

Method	Regularized PL			Regularized WPL			Regularized PL			Regularized WPL		
	$\kappa = 5 \times 10^{-4}$						$\kappa = 5 \times 10^{-5}$					
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
	Scenario 1a											
Lasso	0.19	0.19	0.27	0.43	0.29	0.52	0.29	0.60	0.67	0.94	0.53	1.08
Enet	0.27	0.22	0.35	0.72	0.32	0.79	0.34	0.66	0.74	1.21	0.40	1.27
AL	0.05	0.18	0.19	0.14	0.24	0.28	0.19	0.60	0.63	0.57	0.57	0.81
Aenet	0.07	0.19	0.20	0.20	0.27	0.33	0.22	0.60	0.64	0.69	0.55	0.88
SCAD	0.19	0.19	0.27	0.29	0.32	0.43	0.14	0.55	0.57	1.10	0.71	1.31
MC+	0.20	0.19	0.28	0.32	0.37	0.49	0.15	0.55	0.57	1.15	0.72	1.35
	Scenario 1b											
Lasso	0.18	1.03	1.05	0.57	0.58	0.81	1.97	8.00	8.23	1.85	2.11	2.81
Enet	0.27	1.32	1.34	0.81	0.73	1.09	1.87	7.73	7.96	1.94	2.02	2.80
AL	0.18	0.73	0.76	0.28	0.43	0.51	1.26	6.23	6.36	1.68	1.70	2.39
Aenet	0.21	0.72	0.75	0.36	0.44	0.57	1.05	5.45	5.55	1.76	1.49	2.31
SCAD	0.26	0.99	1.02	0.39	0.63	0.74	1.20	5.55	5.68	1.71	1.59	2.34
MC+	0.26	0.99	1.03	0.40	0.64	0.76	1.21	5.53	5.66	1.71	1.59	2.33
	Scenario 2a											
Lasso	1.45	1.89	2.38	2.24	2.47	3.34	0.94	8.86	8.91	4.53	5.79	7.35
Enet	1.54	1.89	2.44	2.38	2.62	3.54	1.27	6.54	6.66	4.95	4.85	6.93
AL	1.57	1.80	2.39	2.20	2.16	3.09	1.33	6.38	6.52	4.31	4.50	6.23
Aenet	2.05	1.60	2.59	2.64	2.11	3.38	1.95	4.75	5.13	4.89	3.73	6.14
SCAD	2.26	1.75	2.86	3.84	2.43	4.54	3.74	3.45	5.09	5.79	2.73	6.40
MC+	2.45	1.77	3.02	3.95	2.39	4.61	3.81	3.41	5.12	5.82	2.71	6.42
	Scenario 2b											
Lasso	3.28	2.87	4.36	3.36	3.20	4.64	3.85	13.41	13.95	4.61	11.20	12.11
Enet	3.39	2.45	4.18	3.48	2.75	4.44	3.76	7.86	8.71	4.66	6.96	8.37
AL	3.64	1.59	3.97	3.69	1.78	4.10	3.89	8.99	9.80	4.70	6.95	8.39
Aenet	3.71	1.34	3.95	3.79	1.58	4.10	4.03	4.89	6.34	4.88	4.38	6.55
SCAD	4.56	2.22	5.07	4.67	2.27	5.19	5.22	3.27	6.16	5.65	3.18	6.48
MC+	4.53	2.24	5.05	4.64	2.29	5.18	5.23	3.25	6.15	5.66	3.21	6.51

and MC+) and which cannot satisfy (lasso, elastic net) our theorems is that the methods which cannot satisfy our theorems tend to over-select covariates, leading to suffering from larger FPR and smaller PPV in general. Among all regularization methods considered in this study, adaptive lasso and adaptive elastic net seem to outperform the other methods in most cases. Although adaptive lasso and adaptive elastic net perform quite similarly, the adaptive lasso is slightly better.

In this simulation study, we are still able to show that even when the strong multicollinearity exists such as in Scenario 1a, our proposed methods work well for the penalization methods satisfying our theorems. However, as probably expected, our methods are getting difficult to distinguish between the important and the noisy covariates as the setting becomes more and more complex. In the experiments we conduct, we find that the regularized PL and WPL (with adaptive lasso) perform quite similar for the easiest (Scenario 1a) and the toughest (Scenario 2b) setting. For Scenarios 1b and 2a, the regularized WPL with adaptive lasso seems to be more favorable. From Table 3, we would recommend in general to combine the regularized WPL with the adaptive lasso to perform variable selection.

Table 4 gives the prediction properties of the estimates (except for β_0 which is excluded) in terms of biases, standard deviations (SD), and square root of mean squared errors (RMSE), some criteria we define by

$$\text{Bias} = \left[\sum_{j=1}^{50} \{\hat{\mathbb{E}}(\hat{\beta}_j) - \beta_{0j}\}^2\right]^{\frac{1}{2}}, \text{SD} = \left[\sum_{j=1}^{50} \hat{\sigma}_j^2\right]^{\frac{1}{2}}, \text{RMSE} = \left[\sum_{j=1}^{50} \hat{\mathbb{E}}(\hat{\beta}_j - \beta_{0j})^2\right]^{\frac{1}{2}},$$

where $\hat{\mathbb{E}}(\hat{\beta}_j)$ and $\hat{\sigma}_j^2$ are respectively the empirical mean and variance of the estimates $\hat{\beta}_j$, for $j = 1, \ldots, 50$.

In general, the properties improve with larger κ due to weaker spatial dependence. Regarding the regularization methods considered in this study, adaptive lasso and adaptive elastic net perform best. Adaptive elastic net becomes more preferable than adaptive lasso for a clustered process ($\kappa = 5 \times 10^{-5}$) and for a structured spatial data (Scenarios 1b and 2b). The adaptive elastic net is more efficient than the adaptive lasso especially in the complex situation: large number of covariates, strong multicollinearity, clustered processes, and complex spatial structure due to the advantage of combining l_1 and l_2 penalties.

By employing regularized WPL, we have potentially more efficient estimates than that of the regularized PL, especially for the more clustered process. However, this does not mean that the regularized WPL is able to improve the RMSE since it usually introduces extra biases. Regularized WPL seems more appropriate for the case having covariates with complex spatial structure (Scenarios 1b and 2b). Otherwise, regularized PL is more favorable. From Table 4, when the focus is on prediction, we would recommend to apply adaptive elastic net as a general advice, and we would combine with regularized WPL if the covariates have complex spatial structure (e.g. Scenarios 1b and 2b) or combine with regularized PL if there is no evidence of complex spatial structure in the covariates (e.g. Scenarios 1a and 2a).

Note that, from Table 3, the adaptive lasso is more preferable if the focus is on variable selection while, from Table 4, the adaptive elastic net is more favorable if the focus is for prediction. To have a more general recommendation, we would

recommend applying adaptive elastic net when we are faced with a complex situation: a large number of covariates, strong multicollinearity, clustered processes and complex spatial structure. By combining l_1 and l_2 penalties, the adaptive elastic net provides a nice balance between selection and prediction properties. This is why in most complex cases (Scenario 2 with $\kappa = 5 \times 10^{-5}$), adaptive elastic net decides to choose more covariates than adaptive lasso (which includes true and noisy covariates) to suffer from slightly less appropriate properties for the selection performance but to be able to improve significantly the prediction properties.

4.2 Application to forestry datasets

We now consider the study of ecology in a tropical rainforest in Barro Corrolado Island (BCI), Panama, described previously in Section 1. In particular, we are interested in studying the spatial distribution of 3,604 locations of *Beilschmiedia pendula Lauraceae* (BPL) trees by modeling its intensity as a log-linear function of 93 covariates consisting of 2 topological attributes, 13 soil properties, and 78 interactions between two soil nutrients.Regarding the relatively large number of covariates, we apply our proposed methods to select few covariates among them and estimate their coefficients. In particular, we use the regularized Poisson methods with the lasso, adaptive lasso, and adaptive elastic net. Note that we center and scale all the covariates to observe which covariates owing relatively large effect on the intensity.

Table 5: Number of selected and non-selected covariates among 93 covariates by regularized Poisson likelihood with lasso, adaptive lasso and adaptive elastic net regularization.

Method	Regularized PL		Regularized WPL	
	#Selected	#Non-selected	#Selected	#Non-selected
LASSO	77	16	20	73
AL	50	43	9	84
AENET	69	24	9	84

We present in Table 5 the number of selected and non-selected covariates by each method. Out of 93 covariates, more than 50% from the total number of covariates are selected by regularized PL while much fewer covariates are selected by regularized WPL. The regularized PL seems to overfit the model.

Regarding lasso method, 77 covariates are selected by regularized PL method while 20 covariates are selected by regularized WPL. Compared to the two adaptive methods (i.e., adaptive lasso and adaptive elastic net), lasso tends to keep less important covariates. This may explain why lasso cannot satisfy our Theorem 2. In terms of selection properties, adaptive lasso and adaptive elastic net perform similarly when regularized WPL is applied.

Table 6: Nine common covariates selected

Covariates	Regularized PL			Regularized WPL		
	LASSO	AL	AENET	LASSO	AL	AENET
Elev	0.33	0.37	0.34	0.23	0.14	0.14
Slope	0.37	0.37	0.37	0.45	0.44	0.46
Cu	0.45	0.30	0.30	0.16	0.22	0.19
Mn	0.11	0.10	0.11	0.18	0.14	0.14
P	-0.49	-0.45	-0.48	-0.50	-0.43	-0.39
Zn	-0.69	-0.54	-0.70	-0.21	-0.31	-0.25
Al:P	-0.28	-0.24	-0.28	-0.13	-0.14	-0.13
Mg:P	0.49	0.26	0.30	0.38	0.38	0.34
N.Min:pH	0.42	0.39	0.39	0.22	0.17	0.17

Table 6 gives the information regarding nine covariates commonly selected among the six methods. Although the magnitudes of the estimates can be slightly different, the signs all agree with each other.

These results suggest that BPL trees favor to live in the areas of higher elevation and slope with a high concentration of Copper and Manganese in the soil. Furthermore, BPL trees prefer to live in the areas with lower concentration levels of Phosphorus and Zinc in the soil. The interaction between Aluminum and Phosphorus gives a negative association with the appearance of BPL trees while the interaction between Magnesium and Phosphorus and the interaction between Nitrogen mineralization and pH show a positive association with the occurrence of BPL trees. The maps of 3,604 locations of BPL trees, as well as the nine commonly selected covariates, are depicted in Figure 1.

5 Conclusion and discussion

We consider feature selection techniques for spatial point processes intensity estimation by regularizing estimating equations derived from Poisson and logistic regression likelihoods in a setting where the number of parameters diverges as the volume of observation domain increases. Under some conditions, we prove that the estimates obtained from such setting satisfy consistency, sparsity, and asymptotic normality. Our results are available for large classes of spatial point processes and for many penalty functions.

We conduct simulation experiments to evaluate the finite sample properties of the regularized Poisson estimator and regularized weighted Poisson estimator. From the results, we would recommend in general the combination between regularized WPL and adaptive lasso if the concern is on variable selection. Furthermore, when

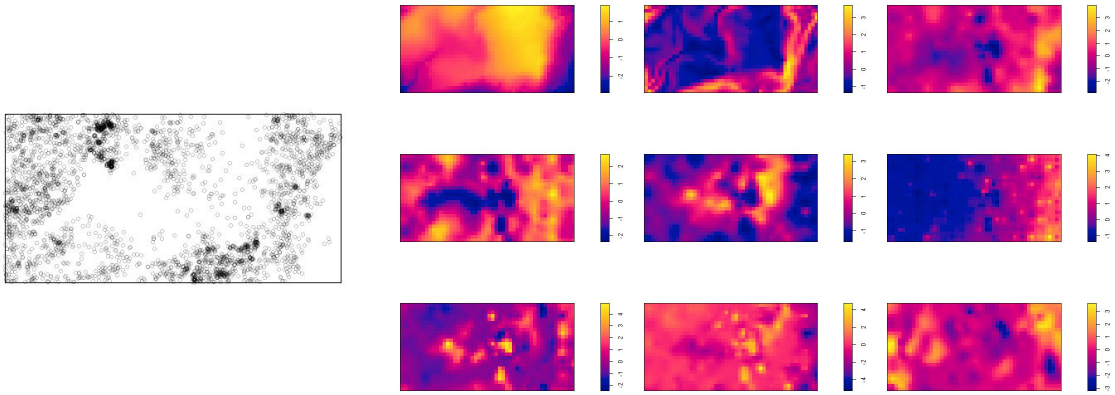

Figure 1: Maps of 3,604 locations of BPL trees and the nine common selected covariates, from left to right, row 1: elevation, slope and Copper, row 2: Manganese, Phosphorus and Zinc, row 3: the interaction between Aluminum and Phosphorus, between Magnesium and Phosphorus, and between Nitrogen mineralisation and pH.

the focus is on prediction, the regularized WPL combined with the adaptive elastic net is more preferable for the situation where there is a complex spatial structure in the covariates. For more general advice, we would recommend using the adaptive elastic net rather than the adaptive lasso since the adaptive elastic net is able to balance the selection and the prediction properties by combining the l_1 and the l_2 penalties.

To implement our methods, we combine the `spatstat` `R` package and the two `R` packages `glmnet` and `ncvreg` dealing with penalized generalized linear models. This results in a computationally fast procedure even when the number of covariates is large. It is worth noticing that, as other regularization methods, our methods also rely on the selection of the tuning parameter. As the study in a classical regression analysis, the BIC-type methods are proposed to obtain selection consistent estimator (see e.g. Zou et al., 2007; Wang et al., 2007, 2009). We have numerical evidence from simulation studies that this criterion can satisfy the selection consistency. Theoretical justification in this spatial point process framework is the purpose of a future research.

We apply our methods to the Barro Corrolado Island study to estimate the intensity of *Beilschmiedia pendula Lauraceae* (BPL) tree as a log-linear function of 93 environmental covariates. Regularized weighted Poisson likelihood combined with adaptive elastic net performs similarly to adaptive lasso. Among 93 covariates, we find nine spatial covariates which may have a high influence to the appearance of BPL trees, including two topological attributes: elevation and slope, four soil nutrients: Copper, Manganese, Phosphorus and Zinc, and three interaction between two soil properties: the interaction between Aluminum and Phosphorus, between Magnesium and Phosphorus, and between Nitrogen mineralisation and pH.

A further work would consider to include the 296 other species of trees, which were surveyed in the same observation region, to study the existence of any competition between BPL and other species of trees in the forest. In such a situation, the

methods used in this study may face some computational issues. The Dantzig selector (Candes and Tao, 2007) might be a good alternative since the implementation for linear models (and generalized linear models) results in a linear programming. Thus, more competitive algorithms are available. It would be interesting to bring this approach to spatial point process framework.

Acknowledgements

We thank A. L. Thurman who kindly shared the `R` code used for the simulation study in Thurman et al. (2015) and P. Breheny who kindly provided his code used in `ncvreg` `R` package. We also thank R. Drouilhet for technical help. The research of A. Choiruddin 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. The research of J.-F. Coeurjolly is supported by the Natural Sciences and Engineering Research Council of Canada. The research of F. Letué is supported by ANR-11-LABX-0025 Persyval-lab (project Persyvact2).

The BCI soils data sets were collected and analyzed by J. Dalling, R. John, K. Harms, R. Stallard and J. Yavitt with support from NSF DEB021104,021115, 0212284,0212818 and OISE 0314581, and STRI Soils Initiative and CTFS and assistance from P. Segre and J. Trani. Datasets are available at the CTFS website `http://ctfs.si.edu/webatlas/datasets/bci/soilmaps/BCIsoil.html`.

References

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

Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. CRC Press, 2015.

Mark Berman and Rolf Turner. Approximating point process likelihoods with glim. *Applied Statistics*, 41(1):31–38, 1992.

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):232–253, 2011.

Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Hyunkeun Cho and Annie Qu. Model selection for correlated data with diverging number of parameters. *Statistica Sinica*, 23(2):901–927, 2013.

Achmad Choiruddin, Jean-François Coeurjolly, and Frédérique Letué. Convex and non-convex regularization methods for spatial point processes intensity estimation. *arXiv preprint arXiv:1703.02462*, 2017.

Jean-François Coeurjolly and Yongtao Guan. Covariance of empirical functionals for inhomogeneous spatial point processes when the intensity has a parametric form. *Journal of Statistical Planning and Inference*, 155:79–92, 2014.

Jean-François Coeurjolly and Jesper Møller. Variational approach to estimate the intensity of spatial point processes. *Bernoulli*, 20(3):1097–1125, 2014.

Richard Condit. Tropical forest census plots. *Springer-Verlag and R. G. Landes Company, Berlin, Germany, and Georgetown, Texas*, 1998.

Peter J Diggle. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):349–362, 1990.

Peter J Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press, 2013.

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.

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

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning (2nd Edition)*. Springer series in statistics Springer, Berlin, 2008.

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

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

Arthur E Hoerl and Robert W Kennard. Ridge regression. *Encyclopedia of statistical sciences*, 1988.

Stephen P Hubbell, Robin B Foster, Sean T O'Brien, KE Harms, Richard Condit, B Wechsler, S Joseph Wright, and S Loo De Lao. Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, 283(5401): 554–557, 1999.

Stephen P Hubbell, Richard Condit, and Robin B Foster. Barro Colorado forest census plot data. 2005. URL http://ctfs.si.edu/datasets/bci.

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

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

Zsolt Karácsony. A central limit theorem for mixing random fields. *Miskolc Mathematical Notes*, 7:147–160, 2006.

Clifford Lam and Jiangqing Fan. Profile-kernel likelihood inference with diverging number of parameters. *The Annals of statistics*, 36(5):2232, 2008.

Gaorong Li, Heng Peng, and Lixing Zhu. Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica*, 21(1):391–419, 2011.

Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical inference and simulation for spatial point processes*. CRC Press, 2004.

Dimitris N Politis, Efstathios Paparoditis, and Joseph P Romano. Large sample inference for irregularly spaced dependent observations based on subsampling. *Sankhyā: The Indian Journal of Statistics, Series A*, 60(2):274–292, 1998.

Stephen Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. consistency. *The Annals of Statistics*, 12(4):1298–1309, 1984.

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.

Ian W Renner and David I Warton. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, 69(1): 274–281, 2013.

Ian W Renner, Jane Elith, Adrian Baddeley, William Fithian, Trevor Hastie, Steven J Phillips, Gordana Popovic, and David I Warton. Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379, 2015.

Shinichiro Shirota, Jorge Mateu, and Alan E Gelfand. Statistical analysis of origin-destination point patterns: Modeling car thefts and recoveries. *arXiv preprint arXiv:1701.05863*, 2017.

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

Andrew L Thurman, Rao Fu, Yongtao Guan, and Jun Zhu. Regularized estimating equations for model selection of clustered spatial point processes. *Statistica Sinica*, 25(1):173–188, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

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

Rasmus Plenge Waagepetersen. Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, 95(2):351–363, 2008.

Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

Yanxin Wang and Li Zhu. Variable selection and parameter estimation via WLAD–SCAD with a diverging number of parameters. *Journal of the Korean Statistical Society*, 46(3):390–403, 2017.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 2009.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

A Auxiliary Lemma

The following lemma is used in the proof of Theorem 1 and Lemma 2 (which includes Lemma 3 and Theorem 2). Throughout the proofs, the notation $\mathbf{X}_n = O_{\mathrm{P}}(x_n)$ or $\mathbf{X}_n = o_{\mathrm{P}}(x_n)$ for a random vector $\mathbf{X}_n$ and a sequence of real numbers x_n means that $\|\mathbf{X}_n\| = O_{\mathrm{P}}(x_n)$ and $\|\mathbf{X}_n\| = o_{\mathrm{P}}(x_n)$. In the same way for a vector $\mathbf{V}_n$ or a squared matrix $\mathbf{M}_n$, the notation $\mathbf{V}_n = O(x_n)$ and $\mathbf{M}_n = O(x_n)$ mean that $\|\mathbf{V}_n\| = O(x_n)$ and $\|\mathbf{M}_n\| = O(x_n)$.

Lemma 1. *Under conditions (C.1)-(C.5), the following result holds as* $n \to \infty$

$$\ell_n^{(1)}(w; \boldsymbol{\beta}_0) = O_{\mathrm{P}}\left(\sqrt{p_n |D_n|}\right). \tag{A.15}$$

Proof. Using Campbell Theorems (2.2)-(2.3), the score vector $\ell_n^{(1)}(w;\boldsymbol{\beta}_0)$ has variance

$$\mathrm{Var}[\ell_n^{(1)}(w;\boldsymbol{\beta}_0)] = \mathbf{B}_n(w;\boldsymbol{\beta}_0) + \mathbf{C}_n(w;\boldsymbol{\beta}_0).$$

Conditions ($\mathcal{C}$.4)-($\mathcal{C}$.5) allow us to obtain that $\sup_{u\in\mathbb{R}^d}\int_{\mathbb{R}^d}\{g(u,v)-1\}\mathrm{d}v < \infty$. We then deduce using conditions ($\mathcal{C}$.1)-($\mathcal{C}$.3) that

$$\mathbf{B}_n(w;\boldsymbol{\beta}_0) + \mathbf{C}_n(w;\boldsymbol{\beta}_0) = O(p_n|D_n|).$$

The result is proved since for any centered real-valued stochastic process Y_n with finite variance $\mathrm{Var}[Y_n]$, $Y_n = O_{\mathrm{P}}(\sqrt{\mathrm{Var}[Y_n]})$. □

B Proof of Theorem 1

In the proof of this result and the following ones, the notation κ stands for a generic constant which may vary from line to line. In particular this constant is independent of n, $\boldsymbol{\beta}_0$ and $\mathbf{k}$.

Proof. Let $d_n = \sqrt{p}_n(|D_n|^{-1/2} + a_n)$, and $\mathbf{k} = \{k_1, k_2, \ldots, k_{p_n}\}^\top$. We remind the reader that the estimate of $\boldsymbol{\beta}_0$ is defined as the maximum of the function Q_n (given by (3.8)) over Θ, an open convex bounded set of $\mathbb{R}^{p_n}$ for any $n \geq 1$. For any $\mathbf{k}$ such that $\|\mathbf{k}\| \leq K < \infty$, $\boldsymbol{\beta}_0 + d_n\mathbf{k} \in \Theta$ for n sufficiently large. Assume this is valid in the following. To prove Theorem 1, we aim at proving that for any given $\epsilon > 0$, there exists sufficiently large $K > 0$ such that for n sufficiently large

$$\mathrm{P}\left(\sup_{\|\mathbf{k}\|=K} \Delta_n(\mathbf{k}) > 0\right) \leq \epsilon, \quad \text{where } \Delta_n(\mathbf{k}) = Q_n(w;\boldsymbol{\beta}_0 + d_n\mathbf{k}) - Q_n(w;\boldsymbol{\beta}_0). \tag{B.16}$$

Equation (B.16) will imply that with probability at least $1-\epsilon$, there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + d_n\mathbf{k} : \|\mathbf{k}\| \leq K\}$, and therefore a local maximizer $\hat{\boldsymbol{\beta}}$ is such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_{\mathrm{P}}(d_n)$. We decompose $\Delta_n(\mathbf{k})$ as $\Delta_n(\mathbf{k}) = T_1 + T_2$ where

$$\begin{aligned} T_1 &= \ell_n(w;\boldsymbol{\beta}_0 + d_n\mathbf{k}) - \ell_n(w;\boldsymbol{\beta}_0) \\ T_2 &= |D_n|\sum_{j=1}^{p_n}\left(p_{\lambda_{n,j}}(|\beta_{0j}|) - p_{\lambda_{n,j}}(|\beta_{0j} + d_n k_j|)\right). \end{aligned}$$

Since $\rho(u;\cdot)$ is infinitely continuously differentiable and $\ell_n^{(2)}(w;\boldsymbol{\beta}) = -\mathbf{A}_n(w;\boldsymbol{\beta})$, then using a second-order Taylor expansion there exists $t \in (0,1)$ such that

$$\begin{aligned} T_1 = {} & d_n\mathbf{k}^\top\ell_n^{(1)}(w;\boldsymbol{\beta}_0) - \frac{1}{2}d_n^2\mathbf{k}^\top\mathbf{A}_n(w;\boldsymbol{\beta}_0)\mathbf{k} \\ & + \frac{1}{2}d_n^2\mathbf{k}^\top\left(\mathbf{A}_n(w;\boldsymbol{\beta}_0) - \mathbf{A}_n(w;\boldsymbol{\beta}_0 + td_n\mathbf{k})\right)\mathbf{k}. \end{aligned}$$

By conditions ($\mathcal{C}$.2)-($\mathcal{C}$.3), there exists a non-negative constant κ such that

$$\frac{1}{2}\|\mathbf{A}_n(w;\boldsymbol{\beta}_0) - \mathbf{A}_n(w;\boldsymbol{\beta}_0 + td_n\mathbf{k})\| \leq \kappa d_n|D_n|p_n.$$

Now, denote $\check{\nu} := \liminf_{n\to\infty} \nu_{\min}(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0))$. By condition ($\mathcal{C}$.7), we have that for any $\mathbf{k}$

$$0 < \check{\nu} \leq \frac{\mathbf{k}^\top \left(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0)\right)\mathbf{k}}{\|\mathbf{k}\|^2}.$$

Therefore, we have

$$T_1 \leq d_n \|\ell_n^{(1)}(w;\boldsymbol{\beta}_0)\| \, \|\mathbf{k}\| - \frac{\check{\nu}}{2} d_n^2 |D_n| \|\mathbf{k}\|^2 + \kappa p_n d_n^3 |D_n| \|\mathbf{k}\|^2.$$

Now by the condition ($\mathcal{C}$.9) and by assumption that $a_n = O(|D_n|^{-1/2})$, we obtain $p_n d_n = o(1)$, so $\kappa p_n d_n^3 |D_n| \|\mathbf{k}\|^2 = o(1) d_n^2 |D_n| \|\mathbf{k}\|^2$. Hence, for n sufficiently large

$$T_1 \leq d_n \|\ell_n^{(1)}(w;\boldsymbol{\beta}_0)\| \, \|\mathbf{k}\| - \frac{\check{\nu}}{4} d_n^2 |D_n| \|\mathbf{k}\|^2.$$

Regarding the term T_2,

$$T_2 \leq T_2' := |D_n| \sum_{j=1}^{s} \left(p_{\lambda_{n,j}}(|\beta_{0j}|) - p_{\lambda_{n,j}}(|\beta_{0j} + d_n k_j|) \right)$$

since for any j the penalty function $p_{\lambda_{n,j}}$ is non-negative and $p_{\lambda_{n,j}}(|\beta_{0j}|) = 0$ for $j = s+1, \ldots, p_n$.

From ($\mathcal{C}$.8), for n sufficiently large, $p_{\lambda_{n,j}}$ is twice continuously differentiable for every $\beta_j = \beta_{0j} + t d_n k_j$ with $t \in (0,1)$. Therefore using a third-order Taylor expansion, there exist $t_j \in (0,1)$, $j = 1, \ldots, s$ such that $-T_2' = T_{2,1}' + T_{2,2}' + T_{2,3}'$, where

$$T_{2,1}' = d_n |D_n| \sum_{j=1}^{s} k_j p'_{\lambda_{n,j}}(|\beta_{0j}|) \operatorname{sign}(\beta_{0,j}) \leq \sqrt{s} a_n d_n |D_n| \, \|\mathbf{k}\| \leq d_n^2 |D_n| \, \|\mathbf{k}\|,$$

$$T_{2,2}' = \frac{1}{2} d_n^2 |D_n| \sum_{j=1}^{s} k_j^2 p''_{\lambda_{n,j}}(|\beta_{0j}|) \leq c_n d_n^2 |D_n| \|\mathbf{k}\|^2,$$

$$T_{2,3}' = \frac{1}{6} d_n^3 |D_n| \sum_{j=1}^{s} k_j^3 p'''_{\lambda_{n,j}}(|\beta_{0j} + t_j d_n k_j|) \leq \kappa d_n^3 |D_n|.$$

The three inequalities above are obtained using the definitions of a_n and c_n, condition ($\mathcal{C}$.8) and Cauchy-Schwarz inequality. We deduce that for n sufficiently large

$$T_2 \leq |T_2'| \leq 2 d_n^2 |D_n| \|\mathbf{k}\|,$$

and then

$$\Delta_n(\mathbf{k}) \leq d_n \|\ell_n^{(1)}(w;\boldsymbol{\beta}_0)\| \, \|\mathbf{k}\| - \frac{\check{\nu}}{4} d_n^2 |D_n| \|\mathbf{k}\|^2 + 2 d_n^2 |D_n| \|\mathbf{k}\|.$$

We now return to (B.16): for n sufficiently large

$$\mathrm{P}\left(\sup_{\|\mathbf{k}\|=K} \Delta_n(\mathbf{k}) > 0 \right) \leq \mathrm{P}\left(\|\ell_n^{(1)}(w;\boldsymbol{\beta}_0)\| > \frac{\check{\nu}}{4} d_n |D_n| K - 2 d_n |D_n| \right).$$

Since $d_n|D_n| = O(\sqrt{p_n|D_n|})$, by choosing K large enough, there exists κ such that for n sufficiently large

$$\mathrm{P}\Big(\sup_{\|\mathbf{k}\|=K} \Delta_n(\mathbf{k}) > 0\Big) \le \mathrm{P}\Big(\|\ell_n^{(1)}(w;\boldsymbol{\beta}_0)\| > \kappa\sqrt{p_n|D_n|}\Big) \le \epsilon$$

for any given $\epsilon > 0$ from (A.15) in Lemma 1.

□

C Proof of Theorem 2

Before proving Theorem 2, we present Lemmas 2-3. Lemma 2 is used to prove Theorem 2(i) while Lemma 3 is used to derive Theorem 2(ii).

Lemma 2. *Assume the conditions (C.1)-(C.8) hold. If* $a_n = O(|D_n|^{-1/2})$ *and* $b_n\sqrt{|D_n|/p_n^2} \to \infty$ *as* $n \to \infty$, *then with probability tending to* 1, *for any* $\boldsymbol{\beta}_1$ *satisfying* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = O_{\mathrm{P}}(\sqrt{p_n/|D_n|})$, *and for any constant* $K_1 > 0$,

$$Q_n\Big(w;({\boldsymbol{\beta}_1}^\top, \mathbf{0}^\top)^\top\Big) = \max_{\|\boldsymbol{\beta}_2\| \le K_1\sqrt{p_n/|D_n|}} Q_n\Big(w;({\boldsymbol{\beta}_1}^\top, {\boldsymbol{\beta}_2}^\top)^\top\Big).$$

Proof. Let $\varepsilon_n = K_1\sqrt{p_n/|D_n|}$. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = O_{\mathrm{P}}(\sqrt{p_n/|D_n|})$, we have for any $j = s+1, \ldots, p_n$

$$\frac{\partial Q_n(w;\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for } 0 < \beta_j < \varepsilon_n, \text{ and} \tag{C.17}$$

$$\frac{\partial Q_n(w;\boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for } -\varepsilon_n < \beta_j < 0. \tag{C.18}$$

From (3.7),

$$\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + R_n,$$

where $R_n = \int_{D_n} w(u) z_j(u)\big(\rho(u;\boldsymbol{\beta}) - \rho(u;\boldsymbol{\beta}_0)\big)\mathrm{d}u$. Using similar arguments used in the proof of Lemma 1, we can prove that

$$\frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} = O_{\mathrm{P}}(\sqrt{|D_n|}).$$

Let $u \in \mathbb{R}^d$. By Taylor expansion, there exists $t \in (0,1)$, such that

$$\rho(u;\boldsymbol{\beta}) = \rho(u;\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{z}(u)\rho(u;\boldsymbol{\beta}_0 + t(\boldsymbol{\beta} - \boldsymbol{\beta}_0)).$$

For n sufficiently large, $\boldsymbol{\beta}_0+t(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \in \Xi(\boldsymbol{\beta}_0)$ defined in condition ($\mathcal{C}$.2). Therefore, for n sufficiently large, we have by Cauchy-Schwarz inequality and conditions ($\mathcal{C}$.2)-($\mathcal{C}$.3)

$$|R_n| \leq \kappa \int_{D_n} \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \|\mathbf{z}(u)\| \mathrm{d}u = O_{\mathrm{P}}(\sqrt{|D_n|p_n^2}).$$

We therefore deduce that for any $j = s+1, \ldots, p_n$

$$\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} = O_{\mathrm{P}}(\sqrt{|D_n|p_n^2}). \tag{C.19}$$

Now, we want to prove (C.17). Let $0 < \beta_j < \varepsilon_n$ and b_n be the sequence given by (3.11). By condition ($\mathcal{C}$.8), b_n is well-defined and since by the assumption $b_n\sqrt{|D_n|/p_n^2} \to \infty$, in particular, $b_n > 0$ for n sufficiently large. Therefore, for n sufficiently large,

$$\begin{aligned}
\mathrm{P}\left(\frac{\partial Q_n(w;\boldsymbol{\beta})}{\partial \beta_j} < 0\right) &= \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} - |D_n| p'_{\lambda_{n,j}}(|\beta_j|)\,\mathrm{sign}(\beta_j) < 0\right) \\
&= \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} < |D_n| p'_{\lambda_{n,j}}(|\beta_j|)\right) \\
&\geq \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} < |D_n| b_n\right) \\
&= \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} < \sqrt{|D_n|p_n^2}\sqrt{\frac{|D_n|}{p_n^2}} b_n\right).
\end{aligned}$$

The assertion (C.17) is therefore deduced from (C.19) and from the assumption that $b_n\sqrt{|D_n|/p_n^2} \to \infty$ as $n \to \infty$. We proceed similarly to prove (C.18). □

Lemma 3. *Under the conditions ($\mathcal{C}$.1)-($\mathcal{C}$.8) and the conditions required in Lemma 2, the following convergence holds in distribution as $n \to \infty$*

$$\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_{01}) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_{01})\}^{-1/2} \ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_{01}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_s), \tag{C.20}$$

where $\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0)$ is the first s components of $\ell^{(1)}_n(w;\boldsymbol{\beta}_0)$ and $\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0)$ (resp. $\mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)$) is the $s \times s$ top-left corner of $\mathbf{B}_n(w;\boldsymbol{\beta}_0)$ (resp $\mathbf{C}_n(w;\boldsymbol{\beta}_0)$).

Proof. By Lemma 2 and by using Campbell Theorems (2.2)-(2.3),

$$\mathrm{Var}[\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0)] = \mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0).$$

The remainder of the proof follows Coeurjolly and Møller (2014). Let $C_i = i + (-1/2, 1/2]^d$ be the unit box centered at $i \in \mathbb{Z}^d$ and define $\mathscr{I}_n = \{i \in \mathbb{Z}^d, C_i \cap D_n \neq \emptyset\}$. Set $D_n = \bigcup_{i \in \mathscr{I}_n} C_{i,n}$, where $C_{i,n} = C_i \cap D_n$. We have

$$\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0) = \sum_{i \in \mathscr{I}_n} Y_{i,n}$$

where

$$Y_{i,n} = \sum_{u \in \mathbf{X} \cap C_{i,n}} w(u)\mathbf{z}_{01}(u) - \int_{C_{i,n}} w(u)\mathbf{z}_{01}(u) \exp(\boldsymbol{\beta}_{01}^{\top}\mathbf{z}_{01}(u))\mathrm{d}u.$$

For any $n \geq 1$ and any $i \in \mathscr{I}_n$, $Y_{i,n}$ has zero mean, and by condition ($\mathcal{C}$.4),

$$\sup_{n \geq 1} \sup_{i \in \mathscr{I}_n} \mathbb{E}(\|Y_{i,n}\|^{2+\delta}) < \infty. \tag{C.21}$$

If we combine (C.21) with conditions ($\mathcal{C}$.1)-($\mathcal{C}$.6), we can apply Karácsony (2006, Theorem 4), a central limit theorem for triangular arrays of random fields.

□

Proof. We now focus on the proof of Theorem 2. Since Theorem 2(i) is proved by Lemma 2, we only need to prove Theorem 2(ii), which is the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$. As shown in Theorem 1, there is a root-$(|D_n|/p_n)$ consistent local maximizer $\hat{\boldsymbol{\beta}}$ of $Q_n(w;\boldsymbol{\beta})$, and it can be shown that there exists an estimator $\hat{\boldsymbol{\beta}}_1$ in Theorem 1 that is a root-$(|D_n|/p_n)$ consistent local maximizer of $Q_n\Big(w;({\boldsymbol{\beta}_1}^{\top}, \mathbf{0}^{\top})^{\top}\Big)$, which is regarded as a function of $\boldsymbol{\beta}_1$, and that satisfies

$$\frac{\partial Q_n(w;\hat{\boldsymbol{\beta}})}{\partial \beta_j} = 0 \quad \text{for } j = 1, \ldots, s \text{ and } \hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{\top}, \mathbf{0}^{\top})^{\top}.$$

There exists $t \in (0,1)$ and $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + t(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})$ such that for $j = 1, \cdots, s$

$$\begin{aligned}
0 =& \frac{\partial \ell_n(w;\hat{\boldsymbol{\beta}})}{\partial \beta_j} - |D_n| p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\operatorname{sign}(\hat{\beta}_j) \\
=& \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \frac{\partial^2 \ell_n(w;\tilde{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) - |D_n| p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\operatorname{sign}(\hat{\beta}_j) \\
=& \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \frac{\partial^2 \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) + \sum_{l=1}^{s} \Psi_{n,jl}(\hat{\beta}_l - \beta_{0l}) \\
& - |D_n| p'_{\lambda_{n,j}}(|\beta_{0j}|)\operatorname{sign}(\beta_{0j}) - |D_n|\phi_{n,j},
\end{aligned} \tag{C.22}$$

where

$$\Psi_{n,jl} = \frac{\partial^2 \ell_n(w;\tilde{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_l} - \frac{\partial^2 \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}$$

and $\phi_{n,j} = p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\operatorname{sign}(\hat{\beta}_j) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\operatorname{sign}(\beta_{0j})$. Since p'_λ is a Lipschitz function by condition ($\mathcal{C}$.8), there exists $\kappa \geq 0$ such that by condition on a_n

$$\begin{aligned}
\phi_{n,j} &= p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\operatorname{sign}(\hat{\beta}_j) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\operatorname{sign}(\beta_{0j}) \\
&= \big(p'_{\lambda_{n,j}}(|\hat{\beta}_j|) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\big)\operatorname{sign}(\hat{\beta}_j) + p'_{\lambda_{n,j}}(|\beta_{0j}|)\big(\operatorname{sign}(\hat{\beta}_j) - \operatorname{sign}(\beta_{0j})\big) \\
&\leq \kappa\big||\hat{\beta}_j| - |\beta_{0j}|\big| + 2a_n \\
&\leq \kappa|\hat{\beta}_j - \beta_{0j}| + 2a_n.
\end{aligned} \tag{C.23}$$

We now decompose $\phi_{n,j}$ as $\phi_{n,j} = T_1 + T_2$ where

$$T_1 = \phi_{n,j}\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \quad \text{and} \quad T_2 = \phi_{n,j}\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j})$$

and where $\tilde{r}_{n,j}$ is the sequence defined in the condition ($\mathcal{C}$.8). Under this condition, the following Taylor expansion can be derived for the term T_1: there exists $t \in (0,1)$ and $\check{\beta}_j = \hat{\beta}_j + t(\beta_{0j} - \hat{\beta}_j)$ such that

$$\begin{aligned} T_1 &= p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \\ &\quad + \frac{1}{2}(\hat{\beta}_j - \beta_{0j})^2 p'''_{\lambda_{n,j}}(|\tilde{\beta}_j|)\text{sign}(\check{\beta}_j)\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \\ &= p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) + O_{\mathrm{P}}(p_n/|D_n|) \end{aligned}$$

where the latter equation ensues from Theorem 1 and condition ($\mathcal{C}$.8). Again, from Theorem 1, $\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \xrightarrow{L^1} 1$ which implies that $\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \xrightarrow{\mathrm{P}} 1$, so $T_1 = p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\big(1 + o_{\mathrm{P}}(1)\big) + O_{\mathrm{P}}(p_n/|D_n|)$.

Regarding the term T_2, we have by (C.23)

$$\begin{aligned} T_2 &\leq \{\kappa|\hat{\beta}_j - \beta_{0j}| + 2a_n\}\ \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}) \\ &= \kappa|\hat{\beta}_j - \beta_{0j}|\ \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}) + o(|D_n|^{-1/2}). \end{aligned}$$

We want to prove that $T_2 = o_{\mathrm{P}}(|D_n|^{-1/2})$. Define $S_n = |\hat{\beta}_j - \beta_{0j}|\ \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j})$ and $T_n = \mathbb{I}(S_n > \delta|D_n|^{-1/2})$ for some $\delta > 0$. We claim that the result is proved if we prove that $\mathbb{E}T_n \to 0$ for any $\delta > 0$. Condition ($\mathcal{C}$.8) implies in particular that for n large enough, $\tilde{r}_{n,j} > \sqrt{p_n/|D_n|} > \sqrt{1/|D_n|}$. Using this, it can be checked that the binary random variable T_n reduces to $T_n = \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}) \xrightarrow{L^1} 0$ as $n \to \infty$.

Then, we deduce that

$$\phi_{n,j} = p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\big(1 + o_{\mathrm{P}}(1)\big) + O_{\mathrm{P}}(p_n/|D_n|) + o_{\mathrm{P}}(|D_n|^{-1/2}). \tag{C.24}$$

Let $\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0)$ (resp. $\ell^{(2)}_{n,1}(w;\boldsymbol{\beta}_0)$) be the first s components (resp. $s \times s$ top-left corner) of $\ell^{(1)}_{n}(w;\boldsymbol{\beta}_0)$ (resp. $\ell^{(2)}_{n}(w;\boldsymbol{\beta}_0)$). Let also $\boldsymbol{\Psi}_n$ be the $s \times s$ matrix containing $\Psi_{n,jl}, j, l = 1, \ldots, s$. Finally, let the vector $\mathbf{p}'_n$, the vector $\boldsymbol{\phi}_n$ and the $s \times s$ matrix $\mathbf{M}_n$ be defined by

$$\begin{aligned} \mathbf{p}'_n &= \{p'_{\lambda_{n,1}}(|\beta_{01}|)\,\text{sign}(\beta_{01}), \ldots, p'_{\lambda_{n,s}}(|\beta_{0s}|)\,\text{sign}(\beta_{0s})\}^\top, \\ \boldsymbol{\phi}_n &= \{\phi_{n,1}, \ldots, \phi_{n,s}\}^\top, \text{ and} \\ \mathbf{M}_n &= \{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}. \end{aligned}$$

We rewrite both sides of (C.22) as

$$\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0) + \ell^{(2)}_{n,1}(w;\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) + \boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) - |D_n|\mathbf{p}'_n - |D_n|\boldsymbol{\phi}_n = 0. \tag{C.25}$$

By definition of $\mathbf{\Pi}_n$ given by (3.14) and from (C.24), we obtain $\boldsymbol{\phi}_n = \mathbf{\Pi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\big(1 + o_{\mathrm{P}}(1)\big) + O_{\mathrm{P}}(p_n/|D_n|) + o_{\mathrm{P}}(|D_n|^{-1/2})$. Using this, we deduce, by premultiplying both sides of (C.25) by $\mathbf{M}_n$, that

$$\begin{aligned}\mathbf{M}_n\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0) - \mathbf{M}_n\big(\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\mathbf{\Pi}_n\big)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \\ = O(|D_n|\,\|\mathbf{M}_n\mathbf{p}'_n\|) + o_{\mathrm{P}}(|D_n|\,\|\mathbf{M}_n\mathbf{\Pi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\|) \\ + O_{\mathrm{P}}(\|\mathbf{M}_n\|p_n) + o_{\mathrm{P}}(\|\mathbf{M}_n\||D_n|^{1/2}) \\ + O_{\mathrm{P}}(\|\mathbf{M}_n\mathbf{\Psi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\|).\end{aligned}$$

Now, $\|\mathbf{M}_n\| = O(1/\sqrt{|D_n|})$ by condition ($\mathcal{C}$.6), $\|\mathbf{\Psi}_n\| = O_{\mathrm{P}}(\sqrt{p_n|D_n|})$ by conditions ($\mathcal{C}$.2)-($\mathcal{C}$.3) and Theorem 1, and $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\| = O_{\mathrm{P}}(\sqrt{p_n/|D_n|})$ by Theorem 1 and Theorem 2(i). Finally, since by assumptions that $a_n\sqrt{|D_n|} \to 0$ and $c_n\sqrt{p_n} \to 0$ as $n \to \infty$, we deduce that

$$\begin{aligned}|D_n|\,\|\mathbf{M}_n\mathbf{p}'_n\| &= O(a_n\sqrt{D_n}) = o(1), \\ |D_n|\,\|\mathbf{M}_n\mathbf{\Pi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\| &= O_{\mathrm{P}}\left(\sqrt{|D_n|}c_n\sqrt{\frac{p_n}{|D_n|}}\right) = o_{\mathrm{P}}(1), \\ \|\mathbf{M}_n\|\,\sqrt{|D_n|} &= O(1), \\ \|\mathbf{M}_n\|\,p_n &= O\left(\sqrt{\frac{p_n^2}{|D_n|}}\right) = o(1), \\ \|\mathbf{M}_n\mathbf{\Psi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\| &= O_{\mathrm{P}}\left(\sqrt{\frac{p_n^2}{|D_n|}}\right) = o_{\mathrm{P}}(1).\end{aligned}$$

The last two lines are obtained from ($\mathcal{C}$.9). Therefore, we have that

$$\mathbf{M}_n\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0) - \mathbf{M}_n\big(\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\mathbf{\Pi}_n\big)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) = o_{\mathrm{P}}(1).$$

By (C.20) in Lemma 3 and by Slutsky's Theorem, we deduce that

$$\begin{aligned}&\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\times \\ &\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\mathbf{\Pi}_n\}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_s)\end{aligned}$$

as $n \to \infty$, which can be rewritten, in particular under ($\mathcal{C}$.7), as

$$|D_n|^{1/2}\mathbf{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_s)$$

where $\mathbf{\Sigma}_n(w, \boldsymbol{\beta}_0)$ is given by (3.13). □

Department of Mathematical Sciences, Aalborg University, Denmark
E-mail: achmad@math.aau.dk

Department of Mathematics, Université du Québec à Montréal (UQAM), Canada and
Department of Probability and Statistics, Université Grenoble Alpes, France
E-mail: coeurjolly.jean-francois@uqam.ca

Department of Probability and Statistics, Université Grenoble Alpes, France
E-mail: frederique.letue@univ-grenoble-alpes.fr