Feature selection for spatial point processes

Achmad Choiruddin

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministérial : 7 août 2006

Présentée par

Achmad Choiruddin

Thèse dirigée par **Jean-François Coeurjolly**
et codirigée par **Frédérique Letué**

préparée au sein du **Laboratoire Jean Kuntzmann**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Sélection de variables pour des processus ponctuels spatiaux

Thèse soutenue publiquement le **15 Septembre 2017**,
devant le jury composé de :

M. Stéphane Girard
Directeur de recherche, INRIA Grenoble Rhône-Alpes, Président
M. Jorge Mateu
Professeur, Universitat Jaume I, Rapporteur
M. Vivian Viallon
Maître de Conférences, Université Claude Bernard de Lyon, Rapporteur
Mme. Hermine Biermé
Professeur, Université de Poitiers, Examinatrice
M. Frédéric Lavancier
Maître de Conférences, Université de Nantes, Examinateur
M. Jean-François Coeurjolly
Professeur, Université du Québec à Montréal, Directeur de thèse
Mme. Frédérique Letué
Maître de Conférences, Université Grenoble Alpes, Directrice de thèse

Sélection de variables pour des processus ponctuels spatiaux

*Feature selection for spatial point processes

Abstract

Abstract. Recent applications such as forestry datasets involve the observations of spatial point pattern data combined with the observation of many spatial covariates. We consider in this thesis the problem of estimating a parametric form of the intensity function in such a context. This thesis develops feature selection procedures and gives some guarantees on their validity. In particular, we propose two different feature selection approaches: the lasso-type methods and the Dantzig selector-type procedures. For the methods considering lasso-type techniques, we derive asymptotic properties of the estimates obtained from estimating functions derived from Poisson and logistic regression likelihoods penalized by a large class of penalties. We prove that the estimates obtained from such procedures satisfy consistency, sparsity, and asymptotic normality. For the Dantzig selector part, we develop a modified version of the Dantzig selector, which we call by the adaptive linearized Dantzig selector (ALDS), to obtain the intensity estimates. More precisely, the ALDS estimates are defined as the solution to an optimization problem which minimizes the sum of coefficients of the estimates subject to a linear approximation of the score vector as a constraint. We find that the estimates obtained from such methods have asymptotic properties similar to the ones proposed under lasso-type techniques using an adaptive lasso regularization term. We investigate the computational aspects of the methods developed using either lasso-type procedures or the Dantzig selector-type approaches. We make links between spatial point processes intensity estimation and generalized linear models (GLMs), so we only have to deal with feature selection procedures for GLMs. Thus, easier computational procedures are implemented and computationally fast algorithm are proposed. Simulation experiments are conducted to highlight the finite sample performances of the

estimates from each of two proposed approaches. Finally, our methods are applied to model the spatial locations a species of tree in the forest observed with a large number of environmental factors.

Keywords: Campbell theorem, Dantzig selector, lasso, logistic regression likelihood, Poisson likelihood

Résumé. Les applications récentes telles que les bases de données forestières impliquent des observations de données spatiales associées à l'observation de nombreuses covariables spatiales. Nous considérons dans cette thèse le problème de l'estimation d'une forme paramétrique de la fonction d'intensité dans un tel contexte. Cette thèse développe les procédures de sélection des variables et donne des garanties quant à leur validité. En particulier, nous proposons deux approches différentes pour la sélection de variables: les méthodes de type lasso et les procédures de type sélecteur de Dantzig. Pour les méthodes envisageant les techniques de type lasso, nous dérivons les propriétés asymptotiques des estimations obtenues par les équations estimantes dérivées des vraisemblances de Poisson et de la régression logistique pénalisées par une grande classe de pénalités. Nous prouvons que les estimations obtenues par de ces procédures satisfont la consistance, sparsité et la normalité asymptotique. Pour la partie sélecteur de Dantzig, nous développons une version modifiée du sélecteur de Dantzig, que nous appelons le sélecteur Dantzig linéarisé adaptatif (ALDS), pour obtenir les estimations d'intensité. Plus précisément, les estimations ALDS sont définies comme la solution à un problème d'optimisation qui minimise la somme des coefficients des estimations sous contrainte de la norme d'une approximation linéaire du vecteur score. Nous constatons que les estimations obtenues par de ces méthodes ont des propriétés asymptotiques semblables à celles proposées précédemment à l'aide de méthode régularisation du lasso adaptatif. Nous étudions les aspects computationnels des méthodes développées en utilisant les procédures de type lasso et de type Sélector Dantzig. Nous établissons des liens entre l'estimation de l'intensité des processus ponctuels spatiaux

et les modèles linéaires généralisés (GLM). Ainsi, des procédures de calcul plus faciles sont implémentées et un algorithme rapide est proposé. Des études de simulation sont menées pour évaluer les performances des estimations de chacune des deux approches proposées. Enfin, nos méthodes sont appliquées pour modéliser les positions d'arbres observées avec un grand nombre de facteurs environnementaux.

Mots clés: Théorème de Campbell, sélecteur de Dantzig, lasso, vraisemblance de la régression logistique, vraisemblance de Poisson

Acknowledgments

First and foremost, I would like to thank my supervisors Jean-François Coeurjolly and Frédérique Letué, without whom this work quite simply would not exist, for their her helpful advice on matters academic, professional, and personal. Their guidance has shaped me into a more rigorous thinker, critical writer, skilled researcher, effective presenter and amiable colleague. I cannot thank them enough for an enormous amount of time they have dedicated in reading various drafts of journal articles, presentation slides, and this thesis. I am always grateful to them for involving me into this research project and for introducing me to the spatial statistics community which open a lot of opportunities, including winning the Mahar Schützenberger prize and continuing to my postdoc at Aalborg University right after my Ph.D. defense. My experience with them during my 3-years Ph.D. has generally augmented my love of statistics.

I want to express my gratitude to my thesis committee members. I wish to thank Stéphane Girard, Hermine Biermé and Frédéric Lavancier for a fruitful discussion during the defense. To Jorge Mateu and Vivian Viallon, thank you for a keen discussion which opens some ideas for extension and improvement.

Thanks to my colleagues at Laboratory Jean Kuntzmann (LJK) for providing a pleasant and inspiring working environment. Special thanks go to Hélène, Laurence, Bruno, and Frédéric for their help and assistance regarding administration and IT services during my stay at LJK. Thanks to Adeline, Rémy, Vincent, and Olivier for informal but helpful discussions. For my good friends and office mates Adrien, Modibo, and Djihad: I was enjoying so much spending my time with you discussing mathematics, statistics, R and many other kinds of stuff outside the "Ph.D. world". I hope to see you again someday :)

I would like to thank the people I met during the conferences, workshops, and summer schools for nice questions and discussions which broaden for more ideas to improve and extend this thesis. A special thank goes to Christophe Biscio, who calls me a mathematical cousin, for nice chats during the conferences we met. I never know that you will be becoming my office mate at Aalborg University. Thanks for reading and advice which improves the Chapter 2 of this thesis. Thanks to Adrian Baddeley, Rasmus Waagepetersen and Ege Rubak for fruitful discussion which gives new perspectives on the research topic I am concerning on.

I would like to extend my gratitude towards all my teachers (and later colleagues) at University Aix-Marseille, especially Nicolas Pech and Marie-Christine Roubaud, for their the motivation, aid, and support from the very first of my day arriving in France. I wish to deliver my special thanks to my high school teacher Anah Herawati and my bachelor teacher Sri Mumpuni Retnaningsih for their motivation and encouragement to be brave chasing whatever our dreams are. Growing up from the modest family, continuing the study at university is an issue. I hope this reminds me, and all people reading my manuscript, that any dream is possible with hard work, trust in ourselves and support from people around.

My family and friends, especially my parents and my parents in law, are thanked for their continuous support and for keeping me in their life despite the long stays abroad. Thanks to all Indonesians in Grenoble for all nice memory. In particular, thank you for those who helped me preparing "the pôt" and documentation for my Ph.D. defense.

Finally, to my wife Saldhyna: Thank you for moving to the other side of the world with me and for standing by me through both good and bad times. Thank you for all your love, friendship, and support. I could not imagine doing this without you, and I look forward to diving into our future adventures.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction (English)

Spatial point pattern data occur in a wide range of scientific areas, e.g., ecology, epidemiology, biology, geosciences, criminology, and astronomy. The statistical and probabilistic framework to treat spatial point pattern data is spatial point process. Some recent spatial point process textbooks include Møller and Waagepetersen (2004), Illian et al. (2008), Diggle (2013), and Baddeley et al. (2015). In this thesis, the main focus is on the estimation of spatial point processes intensity. The intensity serves as the first-order characteristics of a spatial point process and often becomes the main interest in many studies, either in theoretical aspect or in application. Examples include the study of spatial variation of specific disease risk (e.g., Diggle, 1990, 2013), crime rate analysis in a city (e.g., Baddeley et al., 2015; Shirota et al., 2017), and modelling the spatial distribution of trees species in a forest related to some environmental factors (e.g., Waagepetersen, 2007; Renner and Warton, 2013).

In this study, we consider inhomogeneous spatial point processes described by an intensity function which depends on spatial covariates. Although it will be stated in every chapter of this manuscript, we write here that we focus on intensity functions $\rho(\cdot;\boldsymbol{\beta})$ with log-linear form

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

where $\mathbf{z}(u) = \{z_1(u), \ldots, z_p(u)\}^\top$ are the p spatial covariates measured at location u and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^\top$ is a real p-dimensional parameter. An example in epidemiology is the study of spatial variation of cancer risk in a city given the locations of patients

residence and the location of an industrial incinerator (see e.g., Diggle, 1990). In this study, the main question is whether there is an evidence of increased cancer risk for the resident close to the incinerator, so the distance between the home address of the patients to the incinerator is treated as a potential covariate. In criminology, one example is the analysis of car thefts rate in a city considering some demographical information as covariates (see e.g., Shirota et al., 2017). Besides that, another application also occurs in ecology by the interest of modeling the spatial distribution of trees species in a forest related to some environmental factors such as topological attributes and soil properties. Therefore, the main concern to these studies is to find the relationship between the intensity and the spatial covariates by assessing the magnitudes of the components of the vector $\boldsymbol{\beta}$. For Poisson point process models which serve as a tractable class for complete spatial randomness, maximum likelihood estimation (e.g., Berman and Turner, 1992; Rathbun and Cressie, 1994) is straightforward to implement since the likelihood function is easy to evaluate. However, for more general spatial point processes whose likelihoods are often intractable, computationally expensive Markov chain Monte Carlo methods are required (Møller and Waagepetersen, 2004). To overcome this computational issue, many estimating equation-based methods which are computationally competitive and also have nice theoretical properties are developed, for example by Waagepetersen (2007), Guan and Shen (2010), Baddeley et al. (2014), and Guan et al. (2015). It is to be noticed that the estimating equation-based methods are not restricted to the class of Poisson point processes, thus, it can be applied to attractive or repulsive point pattern data.

In recent decades, with the advancement of technology and huge investment in data collection, many applications for estimating the intensity function which involves a large number of covariates are rapidly available. An example which will be used throughout this thesis is an ecological study. In particular, we are interested in studying the spatial distribution of *Beilschmiedia pendula Lauraceae* trees locations surveyed in a 50-hectare region ($D = 1000m \times 500m$) in a tropical rain forest in Barro Corolado Island, Panama. More complete data that we have from censuses conducted in the same observation region contain locations of 297 species of trees (see Condit, 1998;

Hubbell et al., 1999, 2005). Furthermore, information regarding environmental covariates such as topological attributes and soil nutrients have been also collected. Figure 1.1 depicts the spatial distribution of 3604 locations of *Beilschmiedia pendula Lauraceae* trees superimposed on the map of slope field (left) and on the map of concentration of Phosphorus in the soil (middle). In addition, we present in Figure 1.1 (right) the 3604 locations of *Beilschmiedia pendula Lauraceae* trees (•) along with the 1928 locations of *Ocotea whitei* trees (▲). Some research questions regarding this study are: (a) which areas *Beilschmiedia pendula Lauraceae* does and does not prefer to live on? (b) which environmental factors which have significant effect on the appearance of *Beilschmiedia pendula Lauraceae* and how to choose them? (c) how large those factors influence the intensity of *Beilschmiedia pendula Lauraceae*? (d) is there any competition between *Beilschmiedia pendula Lauraceae* and other species of trees in the forest? The study which relates the distribution of a species to the environment is also commonly known as species distribution modeling (e.g., Elith and Leathwick, 2009; Franklin, 2010; Renner, 2013). Species distribution modeling becomes one of the main interests in ecology and biology since it aims to answer many important questions such as questions (a)-(d). These are useful, for example, to have information regarding the conservation efforts and studies of the impact of activities on habitats (e.g., Franklin, 2010). Species distribution modeling is also able to give the prediction of species distribution to discover unstudied regions that may be preferable for a species (e.g., Elith and Leathwick, 2009).

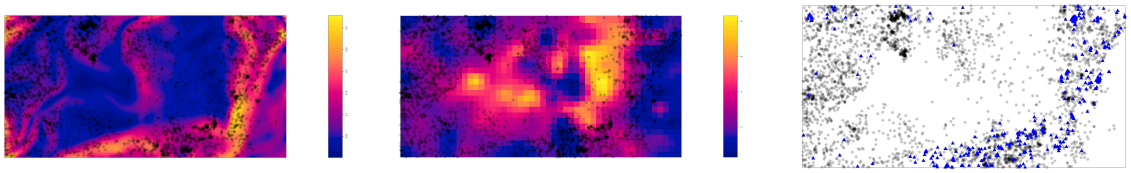

Figure 1.1: Maps of 3604 locations of *Beilschmiedia pendula Lauraceae* trees superimposed on the map of elevation field (left), on the map of concentration of Phosporus in the soil (middle), and on the map of 1928 locations of *Ocotea whitei* trees (right).

For such an application, note that modeling the intensity of *Beilschmiedia pendula Lauraceae* as a function of any possible spatial covariates consisting of environmental

factors will involve a large number of covariates, so maximum likelihood estimation or estimating equation-based methods become undesirable. First, these methods cannot perform variable selection which leads to a hard interpretation of the model. Second, as the number of covariates is large, employing these methods will yield large variance for parameter estimates.

The main goal of this research is to study and develop feature selection procedures for spatial point processes intensity estimation. In particular, we consider two different feature selection procedures: the lasso-type approaches and the Dantzig selector-type methods. We investigate both theoretical and computational aspects. For theoretical aspects, we study the asymptotic properties of our estimates and evaluate whether or not the estimates obtained from such feature selection procedures satisfy consistency, sparsity, and asymptotic normality. In this thesis, we do not investigate the non-asymptotic properties of our estimates, for example, by studying the oracle inequalities as studied by Bühlmann and Van De Geer (2011) for instance. Even if it seems feasible for Poisson point process using for instance concentration inequalities obtained by Reynaud-Bouret (2003), it is not straightforwardly applicable for more general spatial point processes due to the lack of such concentration inequalities for general spatial point processes. By focusing on asymptotic properties, we are able to make our results available for very large classes of spatial point processes which exhibit strong dependence (i.e., very clustered or repulsive point processes). If we had stayed with Poisson case, we would probably have studied the problem differently. Furthermore, in the application we consider in this thesis, it is not realistic to model *Beilschmiedia pendula Lauraceae* by a Poisson point process model as these data exhibit clustering mainly due to seed dispersal (see e.g., Waagepetersen and Guan, 2009; Thurman et al., 2015). From a computational point of view, as we make links between spatial point processes intensity estimation and generalized linear models (GLMs), we only have to deal with feature selection procedures for GLMs which are easy to implement and computationally fast. It is worth emphasizing that our proposed methods are not only limited to the application in ecology and can be applied in different contexts such as the ones studied by Yue and Loh (2015); Renner et al. (2015); Shirota et al. (2017).

The rest of this manuscript is organized as follows. In Chapter 3, we develop lasso-type procedures based on convex and non-convex regularization techniques. We consider estimating equations based on the Campbell formulas derived from Poisson likelihood (Waagepetersen, 2007; Guan and Shen, 2010) and logistic regression likelihood (Baddeley et al., 2014) penalized by a penalty function, written by

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

where $\ell(w;\boldsymbol{\beta})$ is either the Poisson or the logistic regression likelihoods, $|D|$ is the volume of the observation domain and $p_\lambda(|\theta|)$ is a penalty function parameterized by nonnegative λ. Note that if $p_\lambda(|\theta|)$ is a l_1 norm penalty, it corresponds to lasso regularization method (Tibshirani, 1996). We consider a general form of penalty function $p_\lambda(|\theta|)$, which can be a convex or a non-convex function, to make our results available in a more general setting. As representatives, we consider seven regularization methods including convex and non-convex penalties, i.e., ridge (Hoerl and Kennard, 1988), lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006), adaptive elastic net (Zou and Zhang, 2009), SCAD (Fan and Li, 2001), and MC+ (Zhang, 2010). We evaluate their theoretical properties and computational implementation. From theoretical point of view, we find that the regularization methods employing two adaptive methods (i.e., adaptive lasso and adaptive elastic net) and two non-convex penalties (i.e., SCAD and MC+) satisfy consistency, sparsity, and asymptotic normality. It is worth mentioning that the asymptotic considered in this thesis is the increasing domain asymptotic, namely we consider spatial point processes observed over a sequence of bounded domains D_n such that $|D_n| \to \infty$ as $n \to \infty$. In our setting, $|D_n|$ plays the same role as n, the number of observations, in standard problems such as lasso-type methods for linear models or generalized linear models. From computational point of view, our procedure is straightforward to implement in `R` since we combine the existing `R` package `spatstat` (Baddeley et al., 2015) devoted to the analysis of spatial point pattern data with two `R` packages `glmnet` (Friedman et al., 2010) and `ncvreg` (Breheny and Huang, 2011). We also assess the finite properties of

our estimates obtained from such procedures in a simulation experiment and apply our methods to model the intensity of *Beilschmiedia pendula Lauraceae* as a function of 15 spatial covariates consisting of 2 topological properties and 13 soil nutrients.

In Chapter 3, we are restricted to the assumption where the number of covariates is fixed. This leads to two issues: application and theoretical study. In the application considered in this study, modeling the intensity of *Beilschmiedia pendula Lauraceae* as a function of environmental covariates and their possible interactions can increase the number of covariates considerably, so the setting for a diverging number of covariates should be considered. In the theoretical study for a finite number of parameters setting, it has been proved by Fan and Peng (2004) that, in general, penalization regression setting, there are many naive and simple model selection procedures which possess the consistency, sparsity, and asymptotic normality. Therefore, the validity of such asymptotic properties for spatial point processes considering the situation when the number of covariates diverges becomes important to study. We relax this assumption in Chapter 4 by allowing the number of covariates to grow to infinity as the observation domain increases. We investigate the theoretical properties considered in Chapter 3 but extend to the situation when the number of parameters diverges. More precisely, we consider the asymptotic study which allows that both $|D_n|$ (the sequence of observed domains volume) and p_n (the sequence of covariates number) tend to infinity as n goes to infinity. We prove that the results obtained in Chapter 3 are still valid with a few restrictions on the sequence of parameters p_n, by the main argument requiring $p_n^3/|D_n| \to 0$ as $n \to \infty$.

Apart from regularization techniques studied in Chapters 3 and 4, we develop in Chapter 5 the Dantzig selector-type methods for spatial point processes intensity estimation. In particular, we propose a modified version of the Dantzig selector based on linear approximation in the constraint vector which we call by the adaptive linearized Dantzig selector (ALDS). The Dantzig selector (Candes and Tao, 2007) was initially designed for linear regression models and attracted a lot of attention because of its two significant contributions: computational and theoretical aspects. From a computational point of view, an efficient algorithm has been proposed as the implementation

of the Dantzig selector results in a linear programming. For theoretical aspects, Candes and Tao (2007) provided sharp non-asymptotic bounds on the l_2 norm of estimated coefficients error and showed that the error is within a factor of $\log p$ of the error that would be achieved if the locations of the non-zero coefficients were known. As $\log p$ grows very slowly, the Dantzig selector only pays a small price for adaptively choosing the significant variables and is then very suitable for a very large dataset. Some extended studies have been conducted, for example, by James and Radchenko (2009) who studied the computational implementation of the Dantzig selector for generalized linear models, Antoniadis et al. (2010) who extended the theoretical results and implementation of the Dantzig selector for the class of Cox's proportional hazards model, and Li et al. (2014) who developed the Dantzig selector for censored linear regression models and evaluated its asymptotic properties. The general idea of the Dantzig selector is to minimize the l_1 norm of the parameters subject to a constraint on the score vector given by

$$\min |D| \sum_{j=1}^{p} \lambda_j |\beta_j| \text{ subject to } |\mathbf{U}_j(\boldsymbol{\beta})| \leq |D| \lambda_j \text{ for } j = 1, \ldots, p. \tag{1.1}$$

More precisely, $\mathbf{U}_j(\boldsymbol{\beta})$ is the jth component of either the score vector of a likelihood function or an estimating function and $\lambda_j \geq 0, j = 1, 2, \ldots, p$, are the tuning parameters which can be different for every j. Our focus in this chapter is to evaluate the asymptotic properties of the Dantzig selector-type estimator and compare them to the ones obtained from regularization methods developed in Chapters 3 and 4, especially with lasso since the similarities between lasso and Dantzig selector have been discovered in linear models (see e.g., Meinshausen et al., 2007; Bickel et al., 2009; James et al., 2009; Asif and Romberg, 2010). We show under some conditions that the estimates from ALDS are sparse and asymptotically normal. In addition, by proposing the linear approximation on the constraint vector, we show that the complex optimization problem (1.1) can be reduced to a linear programming problem, so a computationally efficient algorithm can be introduced.

CHAPTER 2

Introduction (Français)

Les données spatiales apparaissent dans un large éventail de domaines scientifiques, par exemple, l'écologie, l'épidémiologie, la biologie, les géosciences, la criminologie et l'astronomie. Le cadre statistique et probabiliste pour le traitement des données ponctuelles spatiaux est celui des processus ponctuels spatiaux. Certains livres récents sur les processus ponctuels spatiaux incluent Møller and Waagepetersen (2004), Illian et al. (2008), Diggle (2013), and Baddeley et al. (2015). Dans cette thèse, l'accent principal est mis sur l'estimation de l'intensité des processus ponctuels spatiaux. L'intensité sert de caractéristique de premier ordre d'un processus ponctuel spatial et devient souvent l'intérêt principal pour de nombreuses études, tant dans l'aspect théorique que dans l'application. Des exemples sont l'étude de la variation spatiale du risque spécifique de maladie (e.g., Diggle, 1990, 2013), l'analyse du taux de criminalité dans une ville (e.g., Baddeley et al., 2015; Shirota et al., 2017), et la modélisation de la répartition spatiale des espèces d'arbres dans une forêt en fonction de certains facteurs environnementaux (e.g., Waagepetersen, 2007; Renner and Warton, 2013).

Dans cette étude, nous considérons des processus ponctuels spatiaux inhomogènes décrits par une fonction d'intensité qui dépend de covariables spatiales. Comme indiqué dans chaque chapitre de ce manuscrit, nous nous concentrons sur les fonctions d'intensité $\rho(\cdot;\boldsymbol{\beta})$ avec la forme log-linéaire

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

où $\mathbf{z}(u) = \{z_1(u), \ldots, z_p(u)\}^\top$ sont les p covariables spatiales mesurées à l'emplacement

u et $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^\top$ est un paramètre réel de dimension p. Un exemple d'épidémiologie est l'étude de la variation spatiale du risque de cancer dans une ville compte tenu de l'emplacement de la résidence des patients et de la localisation d'un incinérateur industriel (e.g., Diggle, 1990). Dans cette étude, la question principale est de savoir s'il existe une preuve du risque croissant de cancer pour le résident près de l'incinérateur, de sorte que la distance entre l'adresse résidentielle des patients à l'incinérateur est considérée comme une covariable potentielle. En criminologie, un exemple est l'analyse du taux de vols de voitures dans une ville en considérant certaines informations démographiques comme des covariables (e.g., Shirota et al., 2017). Par ailleurs, une autre application apparaît également en écologie par l'intérêt de la modélisation de la répartition spatiale des espèces d'arbres liée à certains facteurs environnementaux tels que les attributs topologiques et les propriétés du sol. Par conséquent, le but principal de ces études est de trouver une relation entre l'intensité et les covariables spatiales en évaluant les grandeurs des composantes du vecteur $\boldsymbol{\beta}$. Pour les modèles de processus ponctuels de Poisson, l'estimation par maximum de vraisemblance (e.g., Berman and Turner, 1992; Rathbun and Cressie, 1994) est simple à mettre en œuvre puisque la fonction de vraisemblance est facile à évaluer. Cependant, pour des processus ponctuels spatiaux plus généraux dont les vraisemblances sont souvent compliqués, des méthodes de Monte-Carlo par chaînes de Markov coûteuses sont requises (Møller and Waagepetersen, 2004). Pour surmonter ce problème de calcul, de nombreuses méthodes basées sur les équations estimantes, qui sont avantageuses en termes de calcul et ont également de bonnes propriétés théoriques, sont développées, par exemple par Waagepetersen (2007), Guan and Shen (2010), Baddeley et al. (2014), et Guan et al. (2015). Il faut remarquer que les méthodes basées sur les équations estimantes ne sont pas limitées à la classe des processus ponctuels de Poisson, et peuvent être appliquées à des données attractives ou répulsives.

Au cours des dernières décennies, avec l'avancée de la technologie et l'investissement énorme dans la collecte de données, de nombreuses applications qui impliquent un grand nombre de covariables sont disponibles rapidement. Un exemple qui sera utilisé tout au long de cette thèse est une étude écologique. En particulier, nous nous

intéressons à étudier la répartition spatiale des sites d'arbres de *Beilschmiedia pendula Lauraceae* dans une région de 50 hectares ($D = 1000m \times 500m$) dans une forêt tropicale humide à l'île Barro Corolado, au Panama. Des données plus complètes issues des recensements menés dans la même région d'observation contiennent les emplacements de 297 espèces d'arbres (e.g., Condit, 1998; Hubbell et al., 1999, 2005). En outre, des informations concernant les covariables environnementales, telles que les attributs topologiques et les éléments nutritifs des sols ont également été collectées. La Figure 2.1 représente la répartition spatiale des 3604 emplacements d'arbres de *Beilschmiedia pendula Lauraceae* superposés sur la carte du champ de pente (gauche) et sur la carte de concentration de Phosporus dans le sol (milieu). De plus, nous présentons dans la Figure 2.1 (droite) les 3604 emplacements des arbres *Beilschmiedia pendula Lauraceae* (•) ainsi que les 1928 emplacements des arbres *Ocotea whitei* (▲). Les questions de recherche concernant cette étude sont: (a) Dans quelles regions *Beilschmiedia pendula Lauraceae* préfèrent-elles et ne préfèrent-elles pas vivre? (b) Quels sont les facteurs environnementaux qui ont un effet significatif sur l'apparition de *Beilschmiedia pendula Lauraceae*? (c) Comment ces facteurs influencent-ils l'intensité de *Beilschmiedia pendula Lauraceae*? (d) Existe-t-il une concurrence entre *Beilschmiedia pendula Lauraceae* et d'autres espèces d'arbres dans la forêt? L'étude qui relie la répartition d'une espèce à l'environnement est également connue sous le nom de modélisation de la distribution d'espèces (e.g., Elith and Leathwick, 2009; Franklin, 2010; Renner, 2013). La modélisation de la distribution d'espèces devient l'un des intérêts principaux en 'écologie et en biologie car elle vise à répondre à de nombreuses questions importantes telles que les questions (a) - (d). Celles-ci sont utiles, par exemple, pour avoir des informations concernant les efforts de conservation et les études d'impact des activités sur les habitats (e.g., Franklin, 2010). La modélisation de la distribution d'espèces est également capable de prédire la répartition des espèces pour découvrir des régions non étudiées qui peuvent être préférables pour une espèce (e.g., Elith and Leathwick, 2009).

Pour une telle application, notez que la modélisation de l'intensité de *Beilschmiedia pendula Lauraceae* en fonction de toutes les covariables spatiales possibles composées de facteurs environnementaux impliquera un grand nombre de covariables, les méthodes

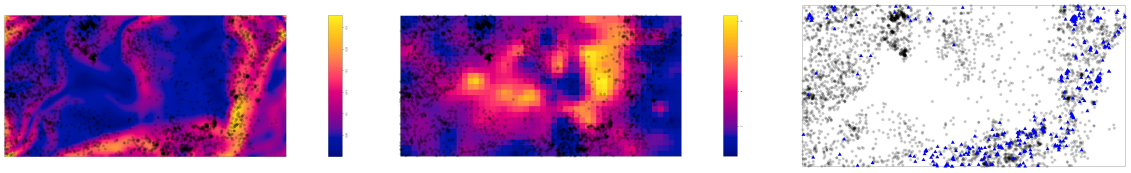

Figure 2.1: Cartes de 3604 emplacements des arbres de *Beilschmiedia pendula Lauraceae* superposés sur la carte du champ d'élévation (gauche), sur la carte de concentration de Phosporus dans le sol (milieu), et sur la carte de 1928 emplacements des arbres de *Ocotea whitei* (droite).

du maximum de vraisemblance ou les méthodes basées sur les équations estimantes deviennent alors insuffisantes. Premièrement, ces méthodes ne peuvent pas effectuer une sélection de variables qui entraîne une interprétation rigoureuse du modèle. Deuxièmement, comme le nombre de covariables est élevé, l'utilisation de ces méthodes entraînera une grande variance pour les estimations de paramètres.

L'objectif principal de cette étude est d'étudier et de développer des procédures de sélection des variables pour l'estimation de l'intensité des processus ponctuels spatiaux. En particulier, nous proposons deux approches différentes: les méthodes de type lasso et de type sélecteur de Dantzig. Nous examinons les aspects théoriques et computationneles. Du point de vue théorique, nous étudions les propriétés asymptotiques des estimateurs et évaluons si les estimateurs obtenues par ces procédures satisfont les propriétés de consistance, sparsité et normalité asymptotique. Dans cette thèse, nous n'étudions pas les propriétés non-asymptotiques de nos estimateurs, par exemple, en étudiant les inégalités d'oracle développées par Bühlmann and Van De Geer (2011). Même si cela semble faisable pour un processus ponctuel de Poisson en utilisant par exemple des inégalités de concentration obtenues par Reynaud-Bouret (2003), ce n'est pas directement applicable pour des processus ponctuels spatiaux plus généraux en raison de l'absence de telles inégalités de concentration pour les processus ponctuels spatiaux dans le cas général. En mettant l'accent sur les propriétés asymptotiques, nous sommes en mesure de rendre nos résultats qui sont applicable pour de très grandes classes de processus ponctuels spatiaux qui présentent une forte dépendance (c'est-à-dire des processus ponctuels très cluster ou répulsifs). Si nous étions restés dans le cas

de Poisson, nous aurions probablement étudié le problème différemment. En plus, dans l'application que nous considérons dans cette thèse, ce n'est pas réaliste de modéliser *Beilschmiedia pendula Lauraceae* par un modèle de processus ponctuel de Poisson car ces données montrent un regroupement qui sont principalement dû à la dispersion des graines (e.g., Waagepetersen and Guan, 2009; Thurman et al., 2015). Du point de vue computationel, comme nous établissons les liens entre l'estimation de l'intensité des processus ponctuels spatiaux et les modèles linéaires généralisés (GLM), nous devons seulement traiter les procédures de sélection de variables pour les GLM qui sont faciles à implémenter et rapides. Il convient de souligner que nos méthodes ne se limitent pas à l'application en écologie et donc peuvent être appliquées dans différents contextes, tels que ceux étudiés par Yue and Loh (2015); Renner et al. (2015); Shirota et al. (2017).

Le reste de ce manuscrit est organisé comme suit. Au Chapitre 3, nous développons des procédures de type lasso basées sur des techniques de régularisation convexes et non-convexes. Nous considérons les équations estimantes basées sur le théorème de Champbell dérivées des vraisemblances de Poisson (Waagepetersen, 2007; Guan and Shen, 2010) et de la régression logistique (Baddeley et al., 2014) pénalisé par une fonction de pénalité

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

où $\ell(w;\boldsymbol{\beta})$ est la vraisemblance de Poisson ou la vraisemblance de la régression logistique, $|D|$ est le volume du domaine d'observation et $p_\lambda(|\theta|)$ est une fonction de pénalité paramétrisé par λ (positif). Notez que si $p_\lambda(|\theta|)$ est une pénalité de norme l_1, cela correspond à la méthode de régularisation lasso (Tibshirani, 1996). Nous considérons une forme générale de fonction de pénalité $p_\lambda(|\theta|)$, qui peut être une fonction convexe ou non-convexe, pour rendre nos résultats applicables dans un cadre plus général. Dans cette thèse, nous considérons sept méthodes de régularisation incluant les fonctions de pénalité convexe et non-convexe, c.-à-d., ridge (Hoerl and Kennard, 1988), lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), lasso adaptatif (Zou, 2006), elastic

net adaptatif (Zou and Zhang, 2009), SCAD (Fan and Li, 2001), et MC+ (Zhang, 2010). Nous évaluons leurs propriétés asymptotiques et leur implémentation computationnelle. Du point de vue théorique, c'est à noter que l'asymptotique considéré dans cette thèse est le domaine croissant asymptotique, à savoir nous considérons les processus ponctuels spatiaux observés sur une séquence de domaines bornés D_n tels que $|D_n| \to \infty$ à mesure que $n \to \infty$. Dans notre contexte, $|D_n|$ joue le même rôle que n, le nombre d'observations, dans les problèmes standard tels que les méthodes de type lasso pour les modèles linéaires ou les modèles linéaires généralisés. Du point de vue computationel, notre procédure est simple à implémenter dans `R` puisque nous combinons le paquet `R spatstat` (Baddeley et al., 2015) avec `glmnet` (Friedman et al., 2010) et `ncvreg` (Breheny and Huang, 2011). Nous évaluons également les propriétés finies de nos estimateurs obtenus par de telles procédures dans les études de simulation et appliquons nos méthodes pour modéliser l'intensité de *Beilschmiedia pendula Lauraceae*.

Au Chapitre 3, nous nous limitons à l'hypothèse où le nombre de covariables est fixé. Cela conduit à deux problèmes: l'application et la théorie. Dans l'application considérée dans cette étude, la modélisation de l'intensité de *Beilschmiedia pendula Lauraceae* en fonction des covariables environnementales et de leurs interactions peut augmenter considérablement le nombre de covariables, donc le contexte où le nombre de covariables diverge doit être considéré. Pour les propriétés asymptotiques considérées dans cette étude, cela a été prouvé par Fan and Peng (2004), dans le cadre de la régression pénalisée générale, qu'il existe de nombreuses procédures de sélection de variables simples qui satisfont aussi les propriétés de consistance, sparsité et normalité asymptotique dans le cadre de paramètres finis. Nous relaxons cette hypothèse au Chapitre 4 en permettant au nombre de covariables de croître à l'infini à mesure que le domaine d'observation augmente. Nous étudions les propriétés asymptotiques considérées au Chapitre 3 mais étendons à la situation lorsque le nombre de paramètres diverge. Nous considérons l'étude asymptotique qui permet à la fois que $|D_n|$ (la séquence des domaines du volume d'observation) et p_n (la suite des nombres de covariables) tend vers l'infini quand n tend vers l'infini. Nous prouvons que les résultats

obtenus au Chapitre 3 sont toujours valables avec quelques restrictions sur la suite des paramètres p_n, par l'argument principal exigeant que $p_n^3/|D_n| \to 0$ quand $n \to \infty$.

Outre les techniques de régularisation étudiées aux Chapitres 3 et 4, nous développons au Chapitre 5, les méthodes de type sélecteur de Dantzig pour l'estimation de l'intensité des processus ponctuels spatiaux. En particulier, nous proposons une version modifiée du sélecteur de Dantzig basé sur une approximation linéaire du vecteur de contrainte que nous appelons le sélecteur Dantzig linéarisé adaptatif (ALDS). Le sélecteur de Dantzig (Candes and Tao, 2007) a été initialement conçu pour les modèles de régression linéaire et a attiré beaucoup d'attention en raison de ses deux contributions significatives: les aspects computationnel et théoriques. Du point de vue computationel, un algorithme efficace a été proposé comme l'implementation du sélecteur Dantzig qui résulte en une programmation linéaire. Du point de vue théorique, Candes and Tao (2007) ont fourni des bornes non asymptotiques optimales sur la norme l_2 de l'erreur d'estimation des coefficients et ont montré que cette erreur est un facteur $\log p$ de l'erreur qui serait atteinte si les emplacements des coefficients non nuls étaient connus. Comme $\log p$ croit très lentement, le sélecteur de Dantzig ne paie qu'un petit prix pour le choix adaptatif des variables significatives et est donc très approprié pour un très grand jeu de données. Des études approfondies ont été menées, par exemple, par James and Radchenko (2009) qui ont étudié la mise en oeuvre du sélecteur de Dantzig pour les modèles linéaires généralisés, Antoniadis et al. (2010) qui ont étendu les résultats théoriques et la mise en oeuvre du sélecteur Dantzig pour la classe du modèle à risques proportionnels de Cox, et Li et al. (2014), qui ont développé le sélecteur de Dantzig pour les modèles de régression linéaire censurée et évalué ses propriétés asymptotiques. L'idée générale du sélecteur de Dantzig est de minimiser la norme l_1 des paramètres soumis à une contrainte sur le vecteur de score donné par

$$\min |D| \sum_{j=1}^{p} \lambda_j |\beta_j| \text{ avec des contraintes } |\mathbf{U}_j(\boldsymbol{\beta})| \leq |D| \lambda_j \text{ for } j = 1, \ldots, p. \tag{2.1}$$

Plus précisément, $\mathbf{U}_j(\boldsymbol{\beta})$ est la $j$ème composante du vecteur de score d'une fonction de vraisemblance ou d'une fonction estimante et $\lambda_j \geq 0, j = 1, 2, \ldots, p$, sont les paramètres

de régularisation qui peuvent être différents pour chaque j. Nos objectifs dans ce chapitre sont d'évaluer les propriétés asymptotiques des estimateurs obtenus par des méthodes de type sélecteur de Dantzig et de les comparer à celles obtenues par des méthodes de régularisation développées dans les Chapitres 3 et 4, surtout avec lasso puisque les similarités entre le lasso et le sélecteur de Dantzig ont été découvertes dans les modèles linéaires (e.g., Meinshausen et al., 2007; Bickel et al., 2009; James et al., 2009; Asif and Romberg, 2010). Nous montrons dans certaines conditions que les estimations de l'ALDS sont sparse et asymptotiquement normales. De plus, en proposant l'approximation linéaire sur le vecteur de contraintes, nous montrons que le problème d'optimisation complexe (2.1) peut être réduit à un problème de programmation linéaire, ainsi un algorithme efficace peut être introduit.

CHAPTER 3

Convex and non-convex regularization methods for spatial point processes intensity estimation

3.1 Introduction

Spatial point pattern data arise in many contexts where interest lies in describing the distribution of an event in space. Some examples include the locations of trees in a forest, gold deposits mapped in a geological survey, stars in a cluster star, animal sightings, locations of some specific cells in retina, or road accidents (see e.g. Møller and Waagepetersen, 2004; Illian et al., 2008; Baddeley et al., 2015). Interest in methods for analyzing spatial point pattern data is rapidly expanding accross many fields of science, notably in ecology, epidemiology, biology, geosciences, astronomy, and econometrics.

One of the main interests when analyzing spatial point pattern data is to estimate the intensity which characterizes the probability that a point (or an event) occurs in an infinitesimal ball around a given location. In practice, the intensity is often assumed to be a parametric function of some measured covariates (e.g. Waagepetersen, 2007; Guan and Loh, 2007; Møller and Waagepetersen, 2007; Waagepetersen, 2008; Waagepetersen and Guan, 2009; Guan and Shen, 2010; Coeurjolly and Møller, 2014). In this study, we assume that the intensity function ρ is parameterized by a vector $\boldsymbol{\beta}$ and has a log-linear specification

$$\rho(u;\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)), \tag{3.1}$$

where $\mathbf{z}(u) = \{z_1(u), \ldots, z_p(u)\}^\top$ are the p spatial covariates measured at location u

and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^\top$ is a real p-dimensional parameter. When the intensity is a function of many variables, covariates selection becomes inevitable.

Variable selection in regression has a number of purposes: provide regularization for good estimation, obtain good prediction, and identify clearly the important variables (e.g. Fan and Lv, 2010; Mazumder et al., 2011). Identifying a set of relevant features from a list of many features is in general combinatorially hard and computationally intensive. In this context, convex relaxation techniques such as lasso (Tibshirani, 1996) have been effectively used for variable selection and parameter estimation simultaneously. The lasso procedure aims at minimizing:

$$-\log L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

where $L(\boldsymbol{\beta})$ is the likelihood function for some model of interest. The ℓ_1 penalty shrinks coefficients towards zero, and can also set coefficients to be exactly zero. In the context of variable selection, the lasso is often thought of as a convex surrogate for the best-subset selection problem:

$$-\log L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_0.$$

The ℓ_0 penalty $\|\boldsymbol{\beta}\|_0 = \sum_{i=1}^{p} \mathbb{I}(|\beta_i| > 0)$ penalizes the number of nonzero coefficients in the model.

Since lasso can be suboptimal in model selection for some cases (e.g. Fan and Li, 2001; Zou, 2006; Zhang and Huang, 2008), many regularization methods then have been developped, motivating to go beyond ℓ_1 regime to more aggressive non-convex penalties which bridges the gap between ℓ_1 and ℓ_0 such as SCAD (Fan and Li, 2001) and MC+ (Zhang, 2010).

More recently, there were several works on implementing variable selection for spatial point processes in order to reduce variance inflation from overfitting and bias from underfitting. Thurman and Zhu (2014) focused on using adaptive lasso to select variables for inhomogeneous Poisson point processes. This study then later was extended to the clustered spatial point processes by Thurman et al. (2015) who established the

asymptotic properties of the estimates in terms of consistency, sparsity, and normality distribution. They also compared their results employing adaptive lasso to SCAD and adaptive elastic net in the simulation study and application, using both regularized weighted and unweighted estimating equations derived from the Poisson likelihood. Yue and Loh (2015) considered modelling spatial point data with Poisson, pairwise interaction point processes, and Neyman-Scott cluster models, incorporated lasso, adaptive lasso, and elastic net regularization methods into generalized linear model framework for fitting these point models. Note that the study by Yue and Loh (2015) also used an estimating equation derived from the Poisson likelihood. However, Yue and Loh (2015) did not provide the theoretical study in detail. Although, in application, many penalty functions have been employed to regularization methods for spatial point processes intensity estimation, the theoretical study is still restricted to some specific penalty functions.

In this chapter, we propose regularized versions of estimating equations based on Campbell formula derived from the Poisson and the logistic regression likelihoods to estimate the intensity of the spatial point processes. We consider both convex and non-convex penalty functions. We provide general conditions on the penalty function to ensure an oracle property and a central limit theorem. Thus, we extend the work by Thurman et al. (2015) and obtain the theoretical results for more general penalty functions and under less restrictive assumptions on the asymptotic covariance matrix (see Remark 3.6.3). The logistic regression method proposed by Baddeley et al. (2014) is as easy to implement as the Poisson likelihood method, but is less biased since it does not require deterministic numerical approximation. We prove that the estimates obtained by regularizing the logistic regression likelihood can also satisfy asymptotic properties (see Remark 3.6.2). Our procedure is straightforward to implement since we only need to combine the `spatstat` `R` package with the two `R` packages `glmnet` and `ncvreg`.

The remainder of this chapter is organized as follows. Section 3.2 gives backgrounds on spatial point processes. Section 3.3 describes standard parameter estimation methods when there is no regularization, while regularization methods are developed in

Section 3.4. Section 3.5 develops numerical details induced by the methods introduced in Sections 3.3-3.4. Asymptotic properties following the work by Fan and Li (2001) for generalized linear models are presented in Section 3.6. Section 3.7 investigates the finite-sample properties of the proposed method in a simulation study, followed by an application to tropical forestry datasets in Section 3.8, and finished by conclusion and discussion in Section 3.9. Proofs of the main results are postponed to Sections 3.11.1-3.11.3.

3.2 Spatial point processes

Let $\mathbf{X}$ be a spatial point process on $\mathbb{R}^d$. Let $D \subset \mathbb{R}^d$ be a compact set of Lebesgue measure $|D|$ which will play the role of the observation domain. We view $\mathbf{X}$ as a locally finite random subset of $\mathbb{R}^d$, i.e. the random number of points of $\mathbf{X}$ in B, $N(B)$, is almost surely finite whenever $B \subset \mathbb{R}^d$ is a bounded region. Suppose $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$ denotes a realization of $\mathbf{X}$ observed within a bounded region D, where $x_i, i = 1, \ldots, m$ represent the locations of the observed points, and m is the number of points. Note that m is random and $0 \leq m < \infty$. If $m = 0$ then $\mathbf{x} = \emptyset$ is the empty point pattern in D. For further background material on spatial point processes, see for example Møller and Waagepetersen (2004).

3.2.1 Moments

The first and second-order properties of a point process are described by intensity measure and second-order factorial moment measure. First-order properties of a point process indicate the spatial distribution events in domain of interest. The intensity measure μ on $\mathbb{R}^d$ is given by

$$\mu(B) = \mathbb{E}N(B), \ B \subseteq \mathbb{R}^d.$$

If the intensity measure μ can be written as

$$\mu(B) = \int_B \rho(u)\mathrm{d}u, \ B \subseteq \mathbb{R}^d,$$

where ρ is a nonnegative function, then ρ is called the intensity function. If ρ is constant, then $\mathbf{X}$ is said to be homogeneous or first-order stationary with intensity ρ. Otherwise, it is said to be inhomogeneous. 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 $\mathrm{d}u$.

Second-order properties of a point process indicate the spatial coincidence of events in the domain of interest. The second-order factorial moment measure $\alpha^{(2)}$ on $\mathbb{R}^d \times \mathbb{R}^d$ is given by

$$\alpha^{(2)}(C) = \mathbb{E} \sum_{u,v \in \mathbf{X}}^{\neq} \mathbb{I}[(u,v) \in C], \ C \subseteq \mathbb{R}^d \times \mathbb{R}^d.$$

where the $\neq$ over the summation sign means that the sum runs over all pairwise different points u, v in $\mathbf{X}$, and $\mathbb{I}[.]$ is the indicator function. If the second-order factorial moment measure $\alpha^{(2)}$ can be written as

$$\alpha^{(2)}(C) = \int \int \mathbb{I}[(u,v) \in C] \rho^{(2)}(u,v) \mathrm{d}u \mathrm{d}v, \ C \subseteq \mathbb{R}^d \times \mathbb{R}^d,$$

where $\rho^{(2)}$ is a nonnegative function, then $\rho^{(2)}$ is called the second-order product density. Intuitively, $\rho^{(2)}(u,v)\mathrm{d}u\mathrm{d}v$ is the probability for observing a pair of points from $\mathbf{X}$ occuring jointly in each of two infinitesimally small balls with centres u, v and volume $\mathrm{d}u, \mathrm{d}v$. Fore more detail description of moment measures of any order, see appendix C in Møller and Waagepetersen (2004).

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 $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$

$$\mathbb{E} \sum_{u \in \mathbf{X}} k(u) = \int k(u) \rho(u) \mathrm{d}u \tag{3.2}$$

$$\mathbb{E} \sum_{u,v \in \mathbf{X}}^{\neq} k(u,v) = \int \int k(u,v) \rho^{(2)}(u,v) \mathrm{d}u \mathrm{d}v. \tag{3.3}$$

In order to study whether a point process deviates from independence (i.e., Poisson

point process), we often consider the pair correlation function 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 (Section 3.2.2.1), we have $\rho^{(2)}(u,v) = \rho(u)\rho(v)$ so that $g(u,v) = 1$. 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 with the same intensity function as $\mathbf{X}$. In the same spirit, we can define $\rho^{(k)}$ the k-th order intensity function (see Møller and Waagepetersen, 2004, for more details). If for any u, v, $g(u,v)$ depends only on $u - v$, the point process $\mathbf{X}$ is said to be second-order reweighted stationary.

3.2.2 Modelling the intensity function

We discuss spatial point process models specified by deterministic or random intensity function. Particularly, we consider two important model classes, namely Poisson and Cox processes. Poisson point processes serve as a tractable model class for no interaction or complete spatial randomness. Cox processes form major classes for clustering or aggregation. For conciseness, we focus on the two later classes of models. We could also have presented determinantal point processes (e.g. Lavancier et al., 2015) which constitute an interesting class of repulsive point patterns with explicit moments. This has not been further investigated for sake of brevity. In this study, we focus on log-linear models of the intensity function given by (3.1).

3.2.2.1 Poisson point process

A point process $\mathbf{X}$ on D is a Poisson point process with intensity function ρ, assumed to be locally integrable, if the following conditions are satisfied:

1. for any $B \subseteq D$ with $0 \leq \mu(B) < \infty$, $N(B) \sim Poisson(\mu(B))$,
2. conditionally on $N(B)$, the points in $\mathbf{X} \cap B$ are i.i.d. with joint density proportional to $\rho(u)$, $u \in B$.

A Poisson point process with a log-linear intensity function is also called a modulated Poisson point process (e.g. Møller and Waagepetersen, 2007; Waagepetersen, 2008). In particular, for Poisson point processes, $\rho^{(2)}(u,v) = \rho(u)\rho(v)$, and $g(u,v) = 1, \forall u, v \in D$.

3.2.2.2 Cox processes

A Cox process is a natural extension of a Poisson point process, obtained by considering the intensity function of the Poisson point process as a realization of a random field. Suppose that $\mathbf{\Lambda} = \{\mathbf{\Lambda}(u) : u \in D\}$ is a nonnegative random field. If the conditional distribution of $\mathbf{X}$ given $\mathbf{\Lambda}$ is a Poisson point process on D with intensity function $\mathbf{\Lambda}$, then $\mathbf{X}$ is said to be a Cox process driven by $\mathbf{\Lambda}$ (see e.g. Møller and Waagepetersen, 2004). There are several types of Cox processes. Here, we consider two types of Cox processes: a Neyman-Scott point process and a log Gaussian Cox process.

Neyman-Scott point processes. Let $\mathbf{C}$ be a stationary Poisson process (mother process) with intensity $\kappa > 0$. Given $\mathbf{C}$, let $\mathbf{X}_c, c \in \mathbf{C}$, be independent Poisson processes (offspring processes) with intensity function

$$\rho_c(u;\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)) k(u-c;\omega)/\kappa,$$

where k is a probability density function determining the distribution of offspring points around the mother points parameterized by ω. Then $\mathbf{X} = \cup_{c\in\mathbf{C}}\mathbf{X}_c$ is a special case of an *inhomogeneous Neyman-Scott point process* with mothers $\mathbf{C}$ and offspring $\mathbf{X}_c, c \in \mathbf{C}$. The point process $\mathbf{X}$ is a Cox process driven by $\mathbf{\Lambda}(u) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)) \sum_{c\in\mathbf{C}} k(u-c,\omega)/\kappa$ (e.g. Waagepetersen, 2007; Coeurjolly and Møller, 2014) and we can verify that the intensity function of $\mathbf{X}$ is indeed

$$\rho(u;\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)).$$

One example of *Neyman-Scott point process* is the *Thomas process* where

$$k(u) = (2\pi\omega^2)^{-d/2} \exp(-\|u\|^2/(2\omega^2))$$

is the density for $N_d(0, \omega^2\mathbf{I}_d)$. Conditionally on a parent event at location c, children

events are normally distributed around c. Smaller values of ω correspond to tighter clusters, and smaller values of κ correspond to fewer number of parents. The parameter vector $\boldsymbol{\psi} = (\kappa, \omega)^\top$ is referred to as the interaction parameter as it modulates the spatial interaction (or, dependence) among events.

Log Gaussian Cox process. Suppose that $\log \mathbf{\Lambda}$ is a Gaussian random field. Given $\mathbf{\Lambda}$, the point process $\mathbf{X}$ follows Poisson process. Then $\mathbf{X}$ is said to be a log Gaussian Cox process driven by $\mathbf{\Lambda}$ (Møller and Waagepetersen, 2004). If the random intensity function can be written as

$$\log \mathbf{\Lambda}(u) = \boldsymbol{\beta}^\top \mathbf{z}(u) + \boldsymbol{\phi}(u) - \sigma^2/2,$$

where $\boldsymbol{\phi}$ is a zero-mean stationary Gaussian random field with covariance function $c(u, v; \boldsymbol{\psi}) = \sigma^2 R(v - u; \zeta)$ which depends on parameter $\boldsymbol{\psi} = (\sigma^2, \zeta)^\top$ (Møller and Waagepetersen, 2007; Coeurjolly and Møller, 2014). The intensity function of this log Gaussian Cox process is indeed given by

$$\rho(u; \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)).$$

One example of correlation function is the exponential form (e.g. Waagepetersen and Guan, 2009)

$$R(v - u; \zeta) = \exp(-\|u - v\|/\zeta), \text{ for } \zeta > 0.$$

Here, $\boldsymbol{\psi} = (\sigma^2, \zeta)^\top$ constitutes the interaction parameter vector, where σ^2 is the variance and ζ is the correlation scale parameter.

3.3 Parametric intensity estimation

One of the standard ways to fit models to data is by maximizing the likelihood of the model for the data. While maximum likelihood method is feasible for parametric Poisson point process models (Section 3.3.1), computationally intensive Markov chain Monte Carlo (MCMC) methods are needed otherwise (Møller and Waagepetersen,

2004). As MCMC methods are not yet straightforward to implement, estimating equations based on Campbell theorem have been developed (see e.g. Waagepetersen, 2007; Møller and Waagepetersen, 2007; Waagepetersen, 2008; Guan and Shen, 2010; Baddeley et al., 2014). We review the estimating equations derived from the Poisson likelihood in Section 3.3.2-3.3.3 and from the logistic regression likelihood in Section 3.3.4.

3.3.1 Maximum likelihood estimation

For an inhomogeneous Poisson point process with intensity function ρ parameterized by $\boldsymbol{\beta}$, the likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{u \in \mathbf{X} \cap D} \rho(u; \boldsymbol{\beta}) \exp\left(\int_D \big(1 - \rho(u; \boldsymbol{\beta})\big) \mathrm{d}u \right),$$

and the log-likelihood function of $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}) = \sum_{u \in \mathbf{X} \cap D} \log \rho(u; \boldsymbol{\beta}) - \int_D \rho(u; \boldsymbol{\beta}) \mathrm{d}u, \tag{3.4}$$

where we have omitted the constant term $\int_D 1 \mathrm{d}u = |D|$. As the intensity function has log-linear form (3.1), (3.4) reduces to

$$\ell(\boldsymbol{\beta}) = \sum_{u \in \mathbf{X} \cap D} \boldsymbol{\beta}^\top \mathbf{z}(u) - \int_D \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)) \mathrm{d}u.$$

Rathbun and Cressie (1994) showed that the maximum likelihood estimator is consistent, asymptotically normal and asymptotically efficient as the sample region goes to $\mathbb{R}^d$.

3.3.2 Poisson likelihood

Let $\boldsymbol{\beta}_0$ be the true parameter vector. By applying Campbell theorem (5.2) to the score function, i.e. the gradient vector of $\ell(\boldsymbol{\beta})$ denoted by $\ell^{(1)}(\boldsymbol{\beta})$, we have

$$\begin{aligned}
\mathbb{E}\ell^{(1)}(\boldsymbol{\beta}) &= \mathbb{E} \sum_{u \in \mathbf{X} \cap D} \mathbf{z}(u) - \int_D \mathbf{z}(u) \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)) \mathrm{d}u \\
&= \int_D \mathbf{z}(u) \exp(\boldsymbol{\beta}_0^\top \mathbf{z}(u)) \mathrm{d}u - \int_D \mathbf{z}(u) \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)) \mathrm{d}u \\
&= \int_D \mathbf{z}(u) (\exp(\boldsymbol{\beta}_0^\top \mathbf{z}(u)) - \exp(\boldsymbol{\beta}^\top \mathbf{z}(u))) \mathrm{d}u = 0
\end{aligned}$$

when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. So, the score function of the Poisson log-likelihood appears to be an unbiased estimating equation, even though $\mathbf{X}$ is not a Poisson point process. The estimator maximizing (3.4) is referred to as the Poisson estimator. The properties of the Poisson estimator have been carefully studied. Schoenberg (2005) showed that the Poisson estimator is still consistent for a class of spatio-temporal point process models. The asymptotic normality for a fixed observation domain was obtained by Waagepetersen (2007) while Guan and Loh (2007) established asymptotic normality under an increasing domain assumption and for suitable mixing point processes.

Regarding the parameter $\boldsymbol{\psi}$ (see Section 3.2.2.2), Waagepetersen and Guan (2009) studied a two-step procedure to estimate both $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, and they proved that, under certain mixing conditions, the parameter estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}})$ enjoy the properties of consistency and asymptotic normality.

3.3.3 Weighted Poisson likelihood

Although the estimating equation approach derived from the Poisson likelihood is simpler and faster to implement than maximum likelihood estimation, it potentially produces a less efficient estimate than that of maximum likelihood (Waagepetersen, 2007; Guan and Shen, 2010) because information about interaction of events is ignored. To regain some lack of efficiency, Guan and Shen (2010) proposed a weighted Poisson log-likelihood function given by

$$\ell(w; \boldsymbol{\beta}) = \sum_{u \in \mathbf{X} \cap D} w(u) \log \rho(u; \boldsymbol{\beta}) - \int_D w(u) \rho(u; \boldsymbol{\beta}) \mathrm{d}u, \tag{3.5}$$

where $w(\cdot)$ is a weight surface. By regarding (3.5), we see that a larger weight $w(u)$ makes the observations in the infinitesimal region $\mathrm{d}u$ more influent. By Campbell theorem, $\ell^{(1)}(w; \boldsymbol{\beta})$ is still an unbiased estimating equation. In addition, Guan and Shen (2010) proved that, under some conditions, the parameter estimates are consistent and asymptotically normal.

Guan and Shen (2010) showed that a weight surface $w(\cdot)$ that minimizes the trace of the asymptotic variance-covariance matrix of the estimates maximizing (3.5) can

result in more efficient estimates than Poisson estimator. In particular, the proposed weight surface is

$$w(u) = \{1 + \rho(u)f(u)\}^{-1},$$

where $f(u) = \int_D \{g(\|v-u\|;\boldsymbol{\psi}) - 1\}\mathrm{d}u$ and $g(\cdot)$ is the pair correlation function. For a Poisson point process, note that $f(u) = 0$ and hence $w(u) = 1$, which reduces to maximum likelihood estimation. For general point processes, the weight surface depends on both the intensity function and the pair correlation function, thus incorporates information on both inhomogeneity and dependence of the spatial point processes. When clustering is present so that $g(v-u) > 1$, then $f(u) > 0$ and hence the weight decreases with $\rho(u)$. The weight surface can be achieved by setting $\hat{w}(u) = \{1 + \hat{\rho}(u)\hat{f}(u)\}^{-1}$. To get the estimate $\hat{\rho}(u)$, one uses parametric estimation considering $\rho(u;\boldsymbol{\beta})$ with $\boldsymbol{\beta}$ is substituted by $\tilde{\boldsymbol{\beta}}$ given by Poisson estimates, that is, $\hat{\rho}(u;\boldsymbol{\beta}) = \rho(u;\tilde{\boldsymbol{\beta}})$. Alternatively, $\hat{\rho}(u)$ can also be computed nonparametrically by kernel method. Furthermore, Guan and Shen (2010) suggessted to approximate $f(u)$ by $K(r) - \pi r^2$, where $K(\cdot)$ is the Ripley's $K-$function estimated by

$$\hat{K}(r) = \sum_{u,v\in\mathbf{X}\cap D}^{\neq} \frac{\mathbb{I}[\|u-v\| \le r]}{\hat{\rho}(u)\hat{\rho}(v)|D \cap D_{u-v}|}.$$

Guan et al. (2015) extended the study by Guan and Shen (2010) and considered more complex estimating equations. Specifically, $w(u)\mathbf{z}(u)$ is replaced by a function $h(u;\boldsymbol{\beta})$ in the derivative of (3.5) with respect to $\boldsymbol{\beta}$. The procedure results in a slightly more efficient estimate than the one obtained from (3.5). However, the computational cost is more important and since we combine estimating equations and penalization methods (see Section 3.4.1), we have not considered this extension.

3.3.4 Logistic regression likelihood

Although the estimating equations discussed in Section 3.3.2 and 3.3.3 are unbiased, these methods do not, in general, produce unbiased estimator in practical implementations. Waagepetersen (2008) and Baddeley et al. (2014) proposed another estimating

function which is indeed close to the score of the Poisson log-likelihood but is able to obtain less biased estimator than Poisson estimates. In addition, their proposed estimating equation is in fact the derivative of the logistic regression likelihood.

Following Baddeley et al. (2014), we define the weighted logistic regression log-likelihood function by

$$\ell(w;\boldsymbol{\beta}) = \sum_{u\in\mathbf{X}\cap D} w(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})}{\delta(u)+\rho(u;\boldsymbol{\beta})}\right) - \int_D w(u)\delta(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})+\delta(u)}{\delta(u)}\right)\mathrm{d}u, \tag{3.6}$$

where $\delta(u)$ is a nonnegative real-valued function. Its role as well as an explanation of the name 'logistic method' will be explained further in Section 3.5.2. Note that the score of (3.6) is an unbiased estimating equation. Waagepetersen (2008) showed asymptotic normality for Poisson and certain clustered point processes for the estimator obtained from a similar procedure. Furthermore, the methodology and results were studied by Baddeley et al. (2014) considering spatial Gibbs point processes.

To determine the optimal weight surface $w(\cdot)$ for logistic method, we follow Guan and Shen (2010) who minimized the trace of the asymptotic covariance matrix of the estimates. We obtain the weight surface defined by

$$w(u) = \frac{\rho(u)+\delta(u)}{\delta(u)\{1+\rho(u)f(u)\}},$$

where $\rho(u)$ and $f(u)$ can be estimated as in Section 3.3.3.

3.4 Regularization techniques

This section discusses convex and non-convex regularization methods for spatial point process intensity estimation.

3.4.1 Methodology

Regularization techniques were introduced as alternatives to stepwise selection for variable selection and parameter estimation. In general, a regularization method attempts to maximize the penalized log-likelihood function $\ell(\boldsymbol{\theta}) - \eta\sum_{j=1}^{p} p_{\lambda_j}(|\theta_j|)$, where $\ell(\boldsymbol{\theta})$

is the log-likelihood function of $\boldsymbol{\theta}$, η is the number of observations, and $p_\lambda(\theta)$ is a nonnegative penalty function parameterized by a real number $\lambda \geq 0$.

Let $\ell(w;\boldsymbol{\beta})$ be either the weighted Poisson log-likelihood function (3.5) or the weighted logistic regression log-likelihood function (3.6). In a similar way, we define the penalized weighted log-likelihood function given by

$$Q(w;\boldsymbol{\beta}) = \ell(w;\boldsymbol{\beta}) - |D|\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{3.7}$$

where $|D|$ is the volume of the observation domain, which plays the same role as the number of observations η in our setting, λ_j is a nonnegative tuning parameter corresponding to β_j for $j = 1, \ldots, p$, and p_λ is a penalty function described in details in the next section.

3.4.2 Penalty functions and regularization methods

For any $\lambda \geq 0$, we say that $p_\lambda(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is a penalty function if p_λ is a nonnegative function with $p_\lambda(0) = 0$. Examples of penalty function are the

- ℓ_2 norm: $p_\lambda(\theta) = \frac{1}{2}\lambda\theta^2$,
- ℓ_1 norm: $p_\lambda(\theta) = \lambda\theta$,
- Elastic net: for $0 < \gamma < 1, p_\lambda(\theta) = \lambda\{\gamma\theta + \frac{1}{2}(1-\gamma)\theta^2\}$,
- SCAD: for any $\gamma > 2, p_\lambda(\theta) = \begin{cases} \lambda\theta & \text{if } \theta \leq \lambda \\ \frac{\gamma\lambda\theta - \frac{1}{2}(\theta^2+\lambda^2)}{\gamma-1} & \text{if } \lambda \leq \theta \leq \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if } \theta \geq \gamma\lambda, \end{cases}$
- MC+: for any $\gamma > 1, p_\lambda(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma} & \text{if } \theta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if } \theta \geq \gamma\lambda. \end{cases}$

The first and second derivatives of the above functions are given by Table 3.1. It is to be noticed that p'_λ is not differentiable at $\theta = \lambda, \gamma\lambda$ (resp. $\theta = \gamma\lambda$) for SCAD (resp. for MC+) penalty.

Table 3.1: The first and the second derivatives of several penalty functions.

Penalty	$p'_\lambda(\theta)$	$p''_\lambda(\theta)$
ℓ_2	$\lambda\theta$	λ
ℓ_1	λ	0
Elastic net	$\lambda\{(1-\gamma)\theta+\gamma\}$	$\lambda(1-\gamma)$
SCAD	$\begin{cases}\lambda & \text{if } \theta \le \lambda \\ \frac{\gamma\lambda-\theta}{\gamma-1} & \text{if } \lambda \le \theta \le \gamma\lambda \\ 0 & \text{if } \theta \ge \gamma\lambda\end{cases}$	$\begin{cases}0 & \text{if } \theta < \lambda \\ \frac{-1}{\gamma-1} & \text{if } \lambda < \theta < \gamma\lambda \\ 0 & \text{if } \theta > \gamma\lambda\end{cases}$
MC+	$\begin{cases}\lambda - \frac{\theta}{\gamma} & \text{if } \theta \le \gamma\lambda \\ 0 & \text{if } \theta \ge \gamma\lambda\end{cases}$	$\begin{cases}\frac{-1}{\gamma} & \text{if } \theta < \gamma\lambda \\ 0 & \text{if } \theta > \gamma\lambda\end{cases}$

As a first penalization technique to improve ordinary least squares, ridge regression (e.g. Hoerl and Kennard, 1988) works by minimizing the residual sum of squares subject to a bound on the ℓ_2 norm of the coefficients. As a continuous shrinkage method, ridge regression achieves its better prediction through a bias-variance trade-off. Ridge can also be extended to fit generalized linear models. However, the ridge cannot reduce model complexity since it always keeps all the predictors in the model. Then, it was introduced a method called lasso (Tibshirani, 1996), where it employs ℓ_1 penalty to obtain variable selection and parameter estimation simultaneously. Despite lasso enjoys some attractive statistical properties, it has some limitations in some senses (Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006; Zhang and Huang, 2008; Zhang, 2010), making huge possibilities to develop other methods. In the scenario where there are high correlations among predictors, Zou and Hastie (2005) proposed an elastic net technique which is a convex combination between ℓ_1 and ℓ_2 penalties. This method is particularly useful when the number of predictors is much larger than the number of observations since it can select or eliminate the strongly correlated predictors together.

The lasso procedure suffers from nonnegligible bias and does not satisfy an oracle property asymptotically (Fan and Li, 2001). Fan and Li (2001) and Zhang (2010), among others, introduced non-convex penalties to get around these drawbacks. The idea is to bridge the gap between ℓ_0 and ℓ_1, by trying to keep unbiased the estimates of nonzero coefficients and by shrinking the less important variables to be exactly

zero. The rationale behind the non-convex penalties such as SCAD and MC+ can also be understood by considering its first derivative (see Table 3.1). They start by applying the similar rate of penalization as the lasso, and then continuously relax that penalization until the rate of penalization drops to zero. However, employing non-convex penalties in regression analysis, the main challenge is often in the minimization of the possible non-convex objective function when the non-convexity of the penalty is no longer dominated by the convexity of the likelihood function. This issue has been carefully studied. Fan and Li (2001) proposed the local quadratic approximation (LQA). Zou and Li (2008) proposed a local linear approximation (LLA) which yields an objective function that can be optimized using least angle regression (LARS) algorithm (Efron et al., 2004). Finally, Breheny and Huang (2011) and Mazumder et al. (2011) investigated the application of coordinate descent algorithm to non-convex penalties.

Table 3.2: Details of some regularization methods.

Method	$\sum_{j=1}^{p} p_{\lambda_j}(\lvert\beta_j\rvert)$
Ridge	$\sum_{j=1}^{p} \frac{1}{2}\lambda\beta_j^2$
Lasso	$\sum_{j=1}^{p} \lambda\lvert\beta_j\rvert$
Enet*	$\sum_{j=1}^{p} \lambda\{\gamma\lvert\beta_j\rvert + \frac{1}{2}(1-\gamma)\beta_j^2\}$
AL*	$\sum_{j=1}^{p} \lambda_j\lvert\beta_j\rvert$
Aenet*	$\sum_{j=1}^{p} \lambda_j\{\gamma\lvert\beta_j\rvert + \frac{1}{2}(1-\gamma)\beta_j^2\}$
SCAD	$\sum_{j=1}^{p} p_\lambda(\lvert\beta_j\rvert)$, with $p_\lambda(\theta) = \begin{cases} \lambda\theta & \text{if } (\theta \le \lambda) \\ \frac{\gamma\lambda\theta - \frac{1}{2}(\theta^2+\lambda^2)}{\gamma-1} & \text{if } (\lambda \le \theta \le \gamma\lambda) \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if } (\theta \ge \gamma\lambda) \end{cases}$
MC+	$\sum_{j=1}^{p} \left\{ \left(\lambda\lvert\beta_j\rvert - \frac{\beta_j^2}{2\gamma}\right)\mathbb{I}(\lvert\beta_j\rvert \le \gamma\lambda) + \frac{1}{2}\gamma\lambda^2\mathbb{I}(\lvert\beta_j\rvert \ge \gamma\lambda) \right\}$

* Enet, AL and Aenet, respectively, stand for elastic net, adaptive lasso and adaptive elastic net

In (3.7), it is worth emphasizing that we allow each direction to have a different regularization parameter. By doing this, the ℓ_1 and elastic net penalty functions are extended to the adaptive lasso (e.g. Zou, 2006) and adaptive elastic net (e.g. Zou and Zhang, 2009). Table 3.2 details the regularization methods considered in this study.

3.5 Numerical methods

We present numerical aspects in this section. For nonregularized estimation, there are two approaches that we consider. Weighted Poisson regression is explained in Section 3.5.1, while logistic regression is reviewed in Section 3.5.2. Penalized estimation procedure is done by employing coordinate descent algorithm (Section 3.5.3). We separate the use of the convex and non-convex penalties in Section 3.5.3.1 and 3.5.3.2.

3.5.1 Weighted Poisson regression

Berman and Turner (1992) developed a numerical quadrature method to approximate maximum likelihood estimation for an inhomogeneous Poisson point process. They approximated the likelihood by a finite sum that had the same analytical form as the weighted likelihood of generalized linear model with Poisson response. This method was then extended to Gibbs point processes by Baddeley and Turner (2000). Suppose we approximate the integral term in (3.4) by Riemann sum approximation

$$\int_D \rho(u;\boldsymbol{\beta})\mathrm{d}u \approx \sum_i^M v_i \rho(u_i;\boldsymbol{\beta})$$

where $u_i, i = 1,\ldots,M$ are points in D consisting of the m data points and $M-m$ dummy points. The quadrature weights $v_i > 0$ are such that $\sum_i v_i = |D|$. To implement this method, the domain is firstly partitioned into M rectangular pixels of equal area, denoted by a. Then one dummy point is placed in the center of the pixel. Let Δ_i be an indicator whether the point is an event of point process ($\Delta_i = 1$) or a dummy point ($\Delta_i = 0$). Without loss of generality, let $u_i,\ldots,u_m$ be the observed events and $u_{m+1},\ldots,u_M$ be the dummy points. Thus, the Poisson log-likelihood function (3.4) can be approximated and rewritten as

$$\ell(\boldsymbol{\beta}) \approx \sum_i^M v_i\{y_i \log \rho(u_i;\boldsymbol{\beta}) - \rho(u_i;\boldsymbol{\beta})\}, \text{ where } y_i = v_i^{-1}\Delta_i. \tag{3.8}$$

Equation (3.8) corresponds to a quasi Poisson log-likelihood function. Maximizing (3.8) is equivalent to fitting a weighted Poisson generalized linear model, which can

be performed using standard statistical software. Similarly, we can approximate the weighted Poisson log-likelihood function (3.5) using numerical quadrature method by

$$\ell(w;\boldsymbol{\beta}) \approx \sum_{i}^{M} w_i v_i \{y_i \log \rho(u_i;\boldsymbol{\beta}) - \rho(u_i;\boldsymbol{\beta})\}. \tag{3.9}$$

where w_i is the value of the weight surface at point i. The estimate $\hat{w}_i$ is obtained as suggested by Guan and Shen (2010). The similarity beetween (3.8) and (3.9) allows us to compute the estimates using software for generalized linear model as well. This fact is in particular exploited in the `ppm` function in the `spatstat` `R` package (Baddeley and Turner, 2005; Baddeley et al., 2015) with option `method="mpl"`. To make the presentation becomes more general, the number of dummy points is denoted by $\texttt{nd}^2$ for the next sections.

3.5.2 Logistic regression

To perform well, the Berman-Turner approximation often requires a quite large number of dummy points. Hence, fitting such generalized linear models can be computationally intensive, especially when dealing with a quite large number of points. When the unbiased estimating equations are approximated using deterministic numerical approximation as in Section 3.5.1, it does not always produce unbiased estimator. To achieve unbiased estimator, we estimate (3.6) by

$$\ell(w;\boldsymbol{\beta}) \approx \sum_{u\in\mathbf{X}\cap D} w(u) \log\left(\frac{\rho(u;\boldsymbol{\beta})}{\delta(u)+\rho(u;\boldsymbol{\beta})}\right) + \sum_{u\in\mathcal{D}\cap D} w(u)\log\left(\frac{\delta(u)}{\rho(u;\boldsymbol{\beta})+\delta(u)}\right), \tag{3.10}$$

where $\mathcal{D}$ is dummy point process independent of $\mathbf{X}$ and with intensity function δ. The form (3.10) is related to the estimating equation defined by Baddeley et al. (2014, eq. 7). Besides that, we consider this form since if we apply Campbell theorem to the last term of (3.10), we obtain

$$\mathbb{E}\sum_{u\in\mathcal{D}\cap D} w(u)\log\left(\frac{\delta(u)}{\rho(u;\boldsymbol{\beta})+\delta(u)}\right) = \int_D w(u)\delta(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})+\delta(u)}{\delta(u)}\right)\mathrm{d}u,$$

which is exactly what we have in the last term of (3.6). In addition, conditional on $\mathbf{X}\cup\mathcal{D}$, (3.10) is the weighted likelihood function for Bernoulli trials, $y(u) = 1\{u\in\mathbf{X}\}$

for $u \in \mathbf{X} \cup \mathcal{D}$, with

$$\mathrm{P}\{y(u) = 1\} = \frac{\rho(u;\boldsymbol{\beta})}{\delta(u) + \rho(u;\boldsymbol{\beta})} = \frac{\exp\Big(-\log\delta(u) + \boldsymbol{\beta}^\top \mathbf{z}(u)\Big)}{1 + \exp\Big(-\log\delta(u) + \boldsymbol{\beta}^\top \mathbf{z}(u)\Big)}.$$

Precisely, (3.10) is a weighted logistic regression with offset term $-\log\delta$. Thus, parameter estimates can be straightforwardly obtained using standard software for generalized linear models. This approach is in fact provided in the `spatstat` package in `R` by calling the `ppm` function with option `method="logi"` (Baddeley et al., 2014, 2015).

In `spatstat`, the dummy point process $\mathcal{D}$ generates $\mathtt{nd}^2$ points in average in D from a Poisson, binomial, or stratified binomial point process. Baddeley et al. (2014) suggested to choose $\delta(u) = 4m/|D|$, where m is the number of points (so, $\mathtt{nd}^2 = 4m$). Furthermore, to determine δ, this option can be considered as a starting point for a data-driven approach (see Baddeley et al., 2014, for further details).

3.5.3 Coordinate descent algorithm

LARS algorithm (Efron et al., 2004) is a remarkably efficient method for computing an entire path of lasso solutions. For linear models, the computational cost is of order $O(Mp^2)$, which is the same order as a least squares fit. Coordinate descent algorithm (Friedman et al., 2007, 2010) appears to be a more competitive algorithm for computing the regularization paths by costs $O(Mp)$ operations. Therefore we adopt cyclical coordinate descent methods, which can work really fast on large datasets and can take advantage of sparsity. Coordinate descent algorithms optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence criterion is reached. We detail this for some convex and non-convex penalty functions in the next two sections. Here, we only present the coordinate descent algorithm for fitting generalized weighted Poisson regression. A similar approach is used to fit penalized weighted logistic regression.

3.5.3.1 Convex penalty functions

Since $\ell(w;\boldsymbol{\beta})$ given by (3.9) is a concave function of the parameters, the Newton-Raphson algorithm used to maximize the penalized log-likelihood function can be done using the iteratively reweighted least squares (IRLS) method. If the current estimate of the parameters is $\tilde{\boldsymbol{\beta}}$, we construct a quadratic approximation of the weighted Poisson log-likelihood function using Taylor's expansion:

$$\ell(w;\boldsymbol{\beta}) \approx \ell_Q(w;\boldsymbol{\beta}) = -\frac{1}{2M}\sum_{i}^{M}\nu_i(y_i^* - \mathbf{z}_i^\top\boldsymbol{\beta})^2 + C(\tilde{\boldsymbol{\beta}}), \tag{3.11}$$

where $C(\tilde{\boldsymbol{\beta}})$ is a constant, y_i^* are the working response values and ν_i are the weights,

$$\nu_i = w_i v_i \exp(\mathbf{z}_i^\top\tilde{\boldsymbol{\beta}})$$
$$y_i^* = \mathbf{z}_i^\top\tilde{\boldsymbol{\beta}} + \frac{y_i - \exp(\mathbf{z}_i^\top\tilde{\boldsymbol{\beta}})}{\exp(\mathbf{z}_i^\top\tilde{\boldsymbol{\beta}})}.$$

Regularized Poisson linear model works by firstly identifying a decreasing sequence of $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, for which starting with minimum value of $\lambda_{\max}$ such that the entire vector $\hat{\boldsymbol{\beta}} = 0$. For each value of λ, an outer loop is created to compute $\ell_Q(w;\boldsymbol{\beta})$ at $\tilde{\boldsymbol{\beta}}$. Secondly, a coordinate descent method is applied to solve a penalized weighted least squares problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Omega(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{-\ell_Q(w;\boldsymbol{\beta}) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|)\}. \tag{3.12}$$

The coordinate descent method is explained as follows. Suppose we have the estimate $\tilde{\beta}_l$ for $l \neq j$, $l, j = 1, \ldots, p$. The method consists in partially optimizing (3.12) with respect to β_j, that is

$$\min_{\beta_j}\Omega(\tilde{\beta}_1, \ldots, \tilde{\beta}_{j-1}, \beta_j, \tilde{\beta}_{j+1}, \ldots, \tilde{\beta}_p).$$

Friedman et al. (2007) have provided the form of the coordinate-wise update for penalized regression using several penalties such as nonnegative garrote (Breiman, 1995), lasso, elastic net, fused lasso (Tibshirani et al., 2005), group lasso (Yuan and Lin, 2006), Berhu penalty (Owen, 2007), and LAD-lasso (Wang et al., 2007a). For instance,

the coordinate-wise update for the elastic net, which embraces the ridge and lasso regularization by setting respectively γ to 0 or 1, is

$$\tilde{\beta}_j \leftarrow \frac{S\left(\sum_{i=1}^{M} \nu_j z_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\gamma\right)}{\sum_{i=1}^{M} \nu_j z_{ij}^2 + \lambda(1-\gamma)}, \tag{3.13}$$

where $\tilde{y}_i^{(j)} = \tilde{\beta}_0 + \sum_{l \neq j} z_{il}\tilde{\beta}_l$ is the fitted value excluding the contribution from covariate z_{ij}, and $S(z, \lambda)$ is the soft-thresholding operator with value

$$S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+ = \begin{cases} z - \lambda & \text{if } z > 0 \text{ and } \lambda < |z| \\ z + \lambda & \text{if } z < 0 \text{ and } \lambda < |z| \\ 0 & \text{if } \lambda \geq |z|. \end{cases} \tag{3.14}$$

The update (3.13) is repeated for $j = 1, \ldots, p$ until convergence. Coordinate descent algorithm for several convex penalties is implemented in the `R` package `glmnet` (Friedman et al., 2010). For (3.13), we can set $\gamma = 0$ to implement ridge and $\gamma = 1$ to lasso, while we set $0 < \gamma < 1$ to apply elastic net regularization. For adaptive lasso, we follow Zou (2006), take $\gamma = 1$ and replace λ by $\lambda_j = \lambda/|\tilde{\beta}_j|^\tau$, where $\tilde{\boldsymbol{\beta}}$ is an initial estimate, say $\tilde{\boldsymbol{\beta}}(ols)$ or $\tilde{\boldsymbol{\beta}}(ridge)$, and τ is a positive tuning parameter. To avoid the computational evaluation for choosing τ, we follow Zou (2006, Section 3.4) and Wasserman and Roeder (2009) who also considered $\tau = 1$, so we choose $\lambda_j = \lambda/|\tilde{\beta}_j(ridge)|$, where $\tilde{\boldsymbol{\beta}}(ridge)$ is the estimates obtained from ridge regression. Implementing adaptive elastic net follows along similar lines.

3.5.3.2 Non-convex penalty functions

Breheny and Huang (2011) have investigated the application of coordinate descent algorithm to fit penalized generalized linear model using SCAD and MC+, for which the penalty is non-convex. Mazumder et al. (2011) also studied the coordinate-wise optimization algorithm in linear models considering more general non-convex penalties.

Mazumder et al. (2011) concluded that, for a known current estimate $\tilde{\theta}$, the uni-

variate penalized least squares function $Q_u(\theta) = \frac{1}{2}(\theta - \tilde{\theta})^2 + p_\lambda(|\theta|)$ should be convex to ensure that the coordinate-wise procedure converges to a stationary point. Mazumder et al. (2011) found that this turns out to be the case for SCAD and MC+ penalty, but it cannot be satisfied by bridge (or power) penalty and some cases of log-penalty.

Breheny and Huang (2011) derived the solution of coordinate descent algorithm for SCAD and MC+ in generalized linear models cases, and it is implemented in the `ncvreg` package of `R`. Let $\tilde{\boldsymbol{\beta}}_l$ be a vector containing estimates $\tilde{\beta}_l$ for $l \neq j$, $l, j = 1, \ldots, p$, and we wish to partially optimize (3.12) with respect to β_j. If we define $\tilde{g}_j = \sum_{i=1}^{M} \nu_j z_{ij}(y_i - \tilde{y}_i^{(j)})$ and $\tilde{\eta}_j = \sum_{i=1}^{M} \nu_j z_{ij}^2$, the coordinate-wise update for SCAD is

$$\tilde{\beta}_j \leftarrow \begin{cases} \frac{S(\tilde{g}_j, \lambda)}{\tilde{\eta}_j} & \text{if } |\tilde{g}_j| \leq \lambda(\tilde{\eta}_j + 1) \\ \frac{S(\tilde{g}_j, \gamma\lambda/(\gamma-1))}{\tilde{\eta}_j - 1/(\gamma-1)} & \text{if } \lambda(\tilde{\eta}_j + 1) \leq |\tilde{g}_j| \leq \tilde{\eta}_j\lambda\gamma \\ \frac{\tilde{g}_j}{\tilde{\eta}_j} & \text{if } |\tilde{g}_j| \geq \tilde{\eta}_j\lambda\gamma, \end{cases}$$

for any $\gamma > \max_j(1 + 1/\tilde{\eta}_j)$. Then, for $\gamma > \max_j(1/\tilde{\eta}_j)$ and the same definition of $\tilde{g}_j$ and $\tilde{\eta}_j$, the coordinate-wise update for MC+ is

$$\tilde{\beta}_j \leftarrow \begin{cases} \frac{S(\tilde{g}_j, \lambda)}{\tilde{\eta}_j - 1/\gamma} & \text{if } |\tilde{g}_j| \leq \tilde{\eta}_j\lambda\gamma \\ \frac{\tilde{g}_j}{\tilde{\eta}_j} & \text{if } |\tilde{g}_j| \geq \tilde{\eta}_j\lambda\gamma, \end{cases}$$

where $S(z, \lambda)$ is the soft-thresholding operator given by (3.14).

3.5.4 Selection of regularization or tuning parameter

It is worth noticing that coordinate descent procedures (and other computation procedures computing the penalized likelihood estimates) rely on the tuning parameter λ so that the choice of λ is also becoming an important task. The estimation using a large value of λ tends to have smaller variance but larger biases, whereas the estimation using a small value of λ leads to have zero biases but larger variance. The trade-off between the biases and the variances yields an optimal choice of λ (Fan and Lv, 2010).

To select λ, it is reasonable to identify a range of λ values extending from a maximum value of λ for which all penalized coefficients are zero to $\lambda = 0$ (e.g. Friedman

et al., 2010; Breheny and Huang, 2011). After that, we select a λ value which optimizes some criterion. By fixing a path of $\lambda \geq 0$, we select the tuning parameter λ which minimizes WQBIC(λ), a weighted 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 nonzero regression coefficients and $|D|$ is the observation volume which represents the sample size. For linear regression models, $\mathbf{Y} = \mathbf{X}^\top \hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$, Wang et al. (2007b) proposed BIC-type criterion for choosing λ by

$$\text{BIC}(\lambda) = \log \frac{\|\mathbf{Y} - \mathbf{X}^\top \hat{\boldsymbol{\beta}}(\lambda)\|^2}{\eta} + \frac{1}{\eta} \log(\eta) \text{DF}(\lambda),$$

where η is the number of observations and DF(λ) is the degree of freedom. This criterion is consistent, meaning that, it selects the correct model with probability approaching 1 in large samples when a set of candidate models contains the true model. Their findings is in line with the study of Zhang et al. (2010) for which the criterion was presented in more general way, called generalized information criterion (GIC). The criterion WQBIC is the specific form of GIC proposed by Zhang et al. (2010).

The selection of γ for SCAD and MC+ is another task, but we fix $\gamma = 3.7$ for SCAD and $\gamma = 3$ for MC+, following Fan and Li (2001) and Breheny and Huang (2011) respectively, to avoid more complexities.

3.6 Asymptotic theory

In this section, we present the asymptotic results for the regularized weighted Poisson likelihood estimator when considering $\mathbf{X}$ 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 $n \to \infty$. The regularization parameters $\lambda_j = \lambda_{n,j}$ for $j = 1, \ldots, p$ are now indexed by n. These asymptotic results also hold for the regularized unweighted Poisson likelihood estimator. For sake of conciseness, we do not present the asymptotic results for the regularized logistic regression estimate. 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 3.6.2).

3.6.1 Notation and conditions

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

$$\alpha_{k,l}(q) = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathscr{F}(\Lambda_1), B \in \mathscr{F}(\Lambda_2), \\ \Lambda_1 \in \mathscr{B}(\mathbb{R}^d), \Lambda_2 \in \mathscr{B}(\mathbb{R}^d), |\Lambda_1| \leq k, |\Lambda_2| \leq l, d(\Lambda_1, \Lambda_2) \geq q\}, \quad (3.15)$$

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 $\mathbb{R}^d$.

Let $\boldsymbol{\beta}_0 = \{\beta_{01}, \ldots, \beta_{0s}, \beta_{0(s+1)}, \ldots, \beta_{0p}\}^\top = \{\boldsymbol{\beta}_{01}^\top, \boldsymbol{\beta}_{02}^\top\}^\top = (\boldsymbol{\beta}_{01}^\top, \mathbf{0}^\top)^\top$ denote the p-dimensional vector of true coefficient values, where $\boldsymbol{\beta}_{01}$ is the s-dimensional vector of nonzero coefficients and $\boldsymbol{\beta}_{02}$ is the $(p\text{-}s)$-dimensional vector of zero coefficients.

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

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

Consider the following conditions ($\mathcal{C}.1$)-($\mathcal{C}.8$) which are required to derive our asymptotic results, where o denotes the origin of $\mathbb{R}^d$:

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

($\mathcal{C}.2$) We assume that the intensity function has the log-linear specification given by (3.1) where $\beta \in \Theta$ and Θ is an open convex bounded set of $\mathbb{R}^p$.

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

$$\sup_{u \in \mathbb{R}^d} ||\mathbf{z}(u)|| < \infty \quad \text{and} \quad \sup_{u \in \mathbb{R}^d} |w(u)| < \infty.$$

($\mathcal{C}$.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 (3.15), 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) There exists a $p \times p$ positive definite matrix $\mathbf{I}_0$ such that for all sufficiently large n, $|D_n|^{-1}\{\mathbf{B}_n(w;\boldsymbol{\beta}_0) + \mathbf{C}_n(w;\boldsymbol{\beta}_0)\} \geq \mathbf{I}_0$.

($\mathcal{C}$.7) There exists a $p \times p$ positive definite matrix $\mathbf{I}'_0$ such that for all sufficiently large n, we have $|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0) \geq \mathbf{I}'_0$.

($\mathcal{C}$.8) The penalty function $p_\lambda(\cdot)$ is nonnegative 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 $|D_n|^{1/2}\tilde{r}_{n,j} \to \infty$ 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.

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

$$a_n = \max_{j=1,\ldots s} |p'_{\lambda_{n,j}}(|\beta_{0j}|)|, \tag{3.16}$$

$$b_n = \inf_{j=s+1,\ldots,p} \inf_{\substack{|\theta|\leq \epsilon_n \\ \theta \neq 0}} p'_{\lambda_{n,j}}(\theta), \text{ for } \epsilon_n = K_1|D_n|^{-1/2}, \tag{3.17}$$

$$c_n = \max_{j=1,\ldots s} |p''_{\lambda_{n,j}}(|\beta_{0j}|)|. \tag{3.18}$$

These sequences a_n, b_n and c_n, detailed in Table 3.3 for the different methods considered in this chapter, play a central role in our results. Even if this will be discussed later in Section 3.6.3, we specify right now that we require that $a_n|D_n|^{1/2} \to 0$, $b_n|D_n|^{1/2} \to \infty$ and $c_n \to 0$.

Table 3.3: Details of the sequences a_n, b_n and c_n for a given regularization method.

Method	a_n	b_n	c_n
Ridge	$\lambda_n \max_{j=1,\ldots s}\{\lvert\beta_{0j}\rvert\}$	0	λ_n
Lasso	λ_n	λ_n	0
Enet	$\lambda_n\left[(1-\gamma)\max_{j=1,\ldots s}\{\lvert\beta_{0j}\rvert\}+\gamma\right]$	$\gamma\lambda_n$	$(1-\gamma)\lambda_n$
AL	$\max_{j=1,\ldots s}\{\lambda_{n,j}\}$	$\min_{j=s+1,\ldots p}\{\lambda_{n,j}\}$	0
Aenet	$\max_{j=1,\ldots s}\{\lambda_{n,j}\big((1-\gamma)\lvert\beta_{0j}\rvert+\gamma\big)\}$	$\gamma\min_{j=s+1,\ldots p}\{\lambda_{n,j}\}$	$(1-\gamma)\max_{j=1,\ldots,s}\{\lambda_{n,j}\}$
SCAD	0^*	${\lambda_n}^{**}$	0^*
MC+	0^*	${\lambda_n - \frac{K_1}{\gamma\lvert D_n\rvert^{1/2}}}^{**}$	0^*

* if $\lambda_n \to 0$ as $n \to \infty$
** if $|D_n|^{1/2}\lambda_n \to \infty$ as $n \to \infty$

3.6.2 Main results

We state our main results here. Proofs are relegated to Sections 3.11.1-3.11.3.

We first show in Theorem 3.6.1 that the penalized weighted Poisson likelihood estimator converges in probability and exhibits its rate of convergence.

Theorem 3.6.1. *Assume the conditions (C.1)-(C.8) hold and let a_n and c_n be given by (3.16) and (3.18). If $a_n = O(|D_n|^{-1/2})$ and $c_n = o(1)$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(w;\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}\| = O_{\mathrm{P}}(|D_n|^{-1/2} + a_n)$.*

This implies that, if $a_n = O(|D_n|^{-1/2})$ and $c_n = o(1)$, the penalized weighted Poisson likelihood estimator is root-$|D_n|$ consistent. Furthermore, we demonstrate in Theorem 3.6.2 that such a root-$|D_n|$ 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 3.6.2. *Assume the conditions (C.1)-(C.8) hold. If $a_n|D_n|^{1/2} \to 0$, $b_n|D_n|^{1/2} \to \infty$ and $c_n \to 0$ as $n \to \infty$, the root-$|D_n|$ consistent local maximizers $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$ in Theorem 1 satisfy:*

(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

$$\begin{aligned}\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0) =& |D_n|\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\boldsymbol{\Pi}_n\}^{-1}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\} \\ & \{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\boldsymbol{\Pi}_n\}^{-1}, \qquad (3.19)\\ \boldsymbol{\Pi}_n =& \text{diag}\{p''_{\lambda_{n,1}}(|\beta_{01}|), \ldots, p''_{\lambda_{n,s}}(|\beta_{0s}|)\}, \qquad (3.20)\end{aligned}$$

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

As a consequence, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$. Note that $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}$ 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 $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\left(\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\right)^\top = \boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$.

Remark 3.6.1. *For lasso and adaptive lasso,* $\boldsymbol{\Pi}_n = \mathbf{0}$*. For other penalties, since* $c_n = o(1)$*, then* $\|\Pi_n\| = o(1)$*. Since* $\|\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\| = O(|D_n|)$ *from conditions (*$\mathcal{C}$*.2) and (*$\mathcal{C}$*.3),* $|D_n|\,\|\boldsymbol{\Pi}_n\|$ *is asymptotically negligible with respect to* $\|\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\|$*.*

Remark 3.6.2. *Theorems 3.6.1 and 3.6.2 remain true for the regularized weighted logistic regression likelihood estimates if we extend the condition (*$\mathcal{C}$*.3) by replacing 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 by adding* $\sup_{u\in\mathbb{R}^d}\delta(u) < \infty$*.*

Remark 3.6.3. *We want to highlight here the main theoretical differences with the work by Thurman et al. (2015). First, the methodology and results are available for the logistic regression likelihood. Second, we consider very general penalty function while Thurman et al. (2015) only considered the adaptive lasso method. Third, we do not assume, as in Thurman et al. (2015), that* $|D_n|^{-1}\mathbf{M}_n \to \mathbf{M}$ *as* $n \to \infty$ *(where* $\mathbf{M}_n$ *is* $\mathbf{A}_n, \mathbf{B}_n$*, or* $\mathbf{C}_n$*), when* $\mathbf{M}$ *is a positive definite matrix. Instead we assume sharper condition assuming* $\lim_{n\to\infty}\nu_{\min}(|D_n|^{-1}\mathbf{M}_n) > 0$*, where* $\mathbf{M}_n$ *is either* $\mathbf{A}_n$ *or* $\mathbf{B}_n + \mathbf{C}_n$ *and* $\nu_{\min}(\mathbf{M}')$ *is the smallest eigenvalue of a positive definite matrix* $\mathbf{M}'$*. This makes the proofs a little bit more technical.*

3.6.3 Discussion of the conditions

We adopt the conditions ($\mathcal{C}$.1)-($\mathcal{C}$.6) based on the paper from Coeurjolly and Møller (2014). In condition ($\mathcal{C}$.1), the assumption that E contains o in its interior can be made without loss of generality. If instead u is an interior point of E, then condition ($\mathcal{C}$.1) could be modified to that any ball with centre u and radius $r > 0$ is contained in $D_n = nE$ for all sufficiently large n. Condition ($\mathcal{C}$.3) is quite standard. From conditions ($\mathcal{C}$.2)-($\mathcal{C}$.5), the matrices $\mathbf{A}_n(w;\boldsymbol{\beta}_0)$, $\mathbf{B}_n(w;\boldsymbol{\beta}_0)$ and $\mathbf{C}_n(w;\boldsymbol{\beta}_0)$ are bounded by $|D_n|$ (see e.g. Coeurjolly and Møller, 2014).

Combination of conditions ($\mathcal{C}.1$)-($\mathcal{C}.6$) are used to establish a central limit theorem for $|D_n|^{-1/2}\ell_n^{(1)}(w;\boldsymbol{\beta}_0)$ using a general central limit theorem for triangular arrays of nonstationary random fields obtained by Karácsony (2006), which is an extension from Bolthausen (1982), then later extended to nonstationary random fields by Guyon (1995). As pointed out by Coeurjolly and Møller (2014), condition ($\mathcal{C}.6$) is a spatial average assumption like when establishing asymptotic normality of ordinary least square estimators for linear models. This condition is also useful to make sure that the matrix $|D_n|^{-1}\{\mathbf{B}_n(w;\boldsymbol{\beta}_0)+\mathbf{C}_n(w;\boldsymbol{\beta}_0)\}$ is invertible. Conditions ($\mathcal{C}.6$)-($\mathcal{C}.7$) ensure that the matrix $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ is invertible for sufficiently large n. Conditions ($\mathcal{C}.1$)-($\mathcal{C}.6$) are discussed in details for several models by Coeurjolly and Møller (2014). They are satisfied for a large class of intensity functions and a large class of models including Poisson and Cox processes discussed in Section 3.2.2.

Condition ($\mathcal{C}.8$) controls the higher order terms in Taylor expansion of the penalty function. Roughly speaking, we ask the penalty function to be at least Lipschitz and thrice differentiable in a neighborhood of the true parameter vector. As it is, the condition looks technical, however, it is obviously satisfied for ridge, lasso, elastic net (and the adaptive versions). According to the choice of λ_n, it is satisfied for SCAD and MC+ when $|\beta_{0j}|$, for $j=1,\ldots,s$, is not equal to $\gamma\lambda_n$ and/or λ_n.

Theorem 3.6.2 requires the conditions $a_n|D_n|^{1/2}\to 0$, $b_n|D_n|^{1/2}\to\infty$ and $c_n\to 0$ as $n\to\infty$ simultaneously. By requiring these assumptions, the corresponding penalized weighted Poisson likelihood estimators possess the oracle property and perform as well as weighted Poisson likelihood estimator which estimates $\boldsymbol{\beta}_1$ knowing the fact that $\boldsymbol{\beta}_2=\mathbf{0}$.

For the ridge regularization method, $b_n=0$, preventing from applying Theorem 3.6.2 for this penalty. For lasso and elastic net, $a_n=K_2b_n$ for some constant $K_2>0$ (K_2=1 for lasso). The two conditions $a_n|D_n|^{1/2}\to 0$ and $b_n|D_n|^{1/2}\to\infty$ 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 considered in this study, the condition $c_n\to 0$ is implied by the

condition $a_n|D_n|^{1/2} \to 0$ as $n \to \infty$.

3.7 Simulation study

We conduct a simulation study with three different scenarios, described in Section 3.7.1, to compare the estimates of the regularized Poisson likelihood (PL) and that of the regularized weighted Poisson likelihood (WPL). We also want to explore the behaviour of the estimates using different regularization methods. Empirical findings are presented in Section 3.7.2. Furthermore, we compare, in Section 3.7.3, the estimates of the regularized (un)weighted logistic likelihood and the ones of the regularized (un)weighted Poisson likelihood.

3.7.1 Simulation set-up

The setting is quite similar to that of Waagepetersen (2007) and Thurman et al. (2015). The spatial domain is $D = [0, 1000] \times [0, 500]$. 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` (R Core Team, 2016), and use them as two true covariates. In addition, we create three different scenarios to define extra covariates:

Scenario 1. We generate eighteen 201×101 pixel images of covariates as standard Gaussian white noise and denote them by $x_3, \ldots, x_{20}$. We define $\mathbf{z}(u) = \mathbf{x}(u) = \{x_1(u), \ldots, x_{20}(u)\}^\top$ as the covariates vector. The regression coefficients for $z_3, \ldots, z_{20}$ are set to zero.

Scenario 2. First, we generate eighteen 201×101 pixel images of covariates as in the scenario 1. Second, we transform them, together with x_1 and x_2, to have multicollinearity. Third, we define $\mathbf{z}(u) = \mathbf{V}^\top \mathbf{x}(u)$, where $\mathbf{x}(u) = \{x_1(u), \ldots, x_{20}(u)\}^\top$. More precisely, $\mathbf{V}$ is such that $\mathbf{\Omega} = \mathbf{V}^\top \mathbf{V}$, and $(\Omega)_{ij} = (\Omega)_{ji} = 0.7^{|i-j|}$ for $i, j = 1, \ldots, 20$, except $(\Omega)_{12} = (\Omega)_{21} = 0$, to preserve the correlation between x_1 and x_2. The regression coefficients for $z_3, \ldots, z_{20}$ are set to zero.

Scenario 3. We consider a more complex situation. We center and scale the 13 soil nutrients covariates obtained from the study in tropical forest of Barro Colorado Island (BCI) in central Panama (see Condit, 1998; Hubbell et al., 1999, 2005), and use 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 $\mathbf{z}(u) = \mathbf{x}(u) = \{x_1(u), \ldots, x_{15}(u)\}^\top$. The regression coefficients for $z_3, \ldots, z_{15}$ are set to zero.

The different maps of the covariates obtained from scenarios 2 and 3 are depicted in Section 3.10. Except for z_3 which has high correlation with z_2, the extra covariates obtained from scenario 2 tend to have a constant value (Figure 3.3). This is completely different from the ones obtained from scenario 3 (Figure 3.4).

The mean number of points over the domain D, μ, is chosen to be 1600. We set the true intensity function to be $\log \rho(u; \boldsymbol{\beta}_0) = \{\beta_0 + \beta_1 z_1(u) + \beta_2 z_2(u)\}$, where $\beta_1 = 2$ represents a relatively large effect of elevation, $\beta_2 = 0.75$ reflects a relatively small effect of gradient, and β_0 is selected such that each realization has 1600 points in average. Furthermore, we erode regularly the domain D such that, with the same intensity function, the mean number of points over the new domain $D \ominus R$ becomes 400. The erosion is used to observe the convergence of the procedure as the observation domain expands. We consider the default number of dummy points for the Poisson likelihood, denoted by $\texttt{nd}^2$, as suggested in the `spatstat` `R` package, i.e. $\texttt{nd}^2 \approx 4m$, where m is the number of points. With these scenarios, we simulate 2000 spatial point patterns from a Thomas point process using the `rThomas` function in the `spatstat` package. We also consider two different κ parameters ($\kappa = 5 \times 10^{-4}, \kappa = 5 \times 10^{-5}$) as different levels of spatial interaction and let $\omega = 20$. For each of the four combinations of κ and μ, we fit the intensity to the simulated point pattern realizations. We also fit the oracle model which only uses the two true covariates.

All models are fitted using modified internal function in `spatstat` (Baddeley et al., 2015), `glmnet` (Friedman et al., 2010), and `ncvreg` (Breheny and Huang, 2011). A modification of the `ncvreg` `R` package is required to include the penalized weighted

Poisson and logistic likelihood methods.

3.7.2 Simulation results

To better understand the behaviour of Thomas processes designed in this study, Figure 3.1 shows the plot of the four realizations using different κ and μ. The smaller value of κ, the tighter the clusters since there are fewer parents. When $\mu = 400$, i.e. by considering the realizations observed on $D \ominus R$, the mean number of points over the 2000 replications and standard deviation are 396 and 47 (resp. 400 and 137) when $\kappa = 5 \times 10^{-4}$ (resp. $\kappa = 5 \times 10^{-5}$). When $\mu = 1600$, the mean number of points and standard deviation are 1604 and 174 (resp. 1589 and 529) when $\kappa = 5 \times 10^{-4}$ (resp. $\kappa = 5 \times 10^{-5}$).

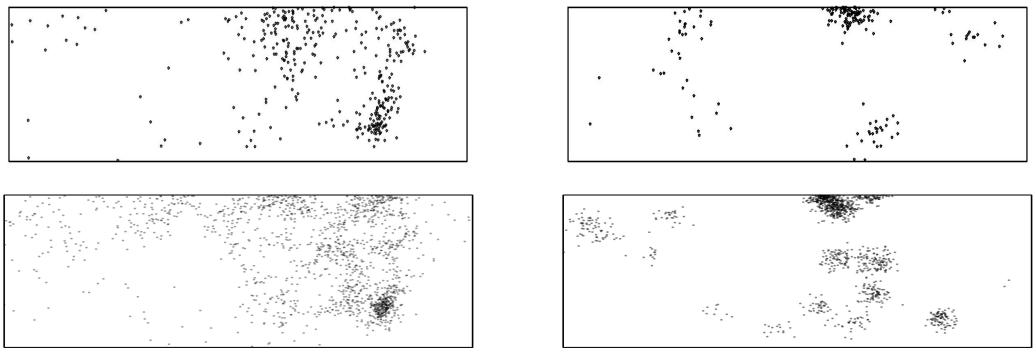

Figure 3.1: Realizations of a Thomas process for $\mu = 400$ (row 1), $\mu = 1600$ (row 2), $\kappa = 5 \times 10^{-4}$ (column 1), and $\kappa = 5 \times 10^{-5}$ (column 2).

Tables 3.4 and 3.5 present the selection properties of the estimates using the penalized PL and the penalized WPL methods. Similarly to Bühlmann and Van De Geer (2011), the indices we consider are the true positive rate (TPR), the false positive rate (FPR), and the positive predictive value (PPV). TPR corresponds to the ratio of the selected true covariates over the number of true covariates, while FPR corresponds to the ratio of the selected noisy covariates over the number of noisy covariates. TPR explains how the model can correctly select both z_1 and z_2. Finally, FPR investigates how the model uncorrectly select among z_3 to z_p ($p = 20$ for scenarios 1 and 2 and $p = 15$ for scenario 3). PPV corresponds to the ratio of the selected true covariates over

Table 3.4: Empirical selection properties (TPR, FPR, and PPV in %) based on 2000 replications of Thomas processes on the domain $D \ominus R$ ($\mu = 400$) for different values of κ and for the three 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	$\kappa = 5 \times 10^{-4}$						$\kappa = 5 \times 10^{-5}$					
	Regularized PL			Regularized WPL			Regularized PL			Regularized WPL		
	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV
Scenario 1												
Ridge	100	100	10	100	100	10	100	100	10	100	100	10
Lasso	100*	27	35	56	0*	98	89	35	34	33	0*	62
Enet	100*	59	18	39	4	36	91	60	21	31	0*	57
AL	100*	1	93	58	0*	100*	88	7	72	35	0*	67
Aenet	100*	6	72	59	0*	99	89	12	61	34	0*	64
SCAD	100*	18	41	66	0*	98	90	17	46	31	0*	56
MC+	100*	21	36	68	0*	96	90	21	42	30	0*	54
Scenario 2												
Ridge	100	100	10	100	100	10	100	100	10	100	100	10
Lasso	100*	25	35	52	1	88	90	38	29	31	0*	55
Enet	100*	52	19	49	4	62	90	60	20	24	1	38
AL	99	4	80	52	0*	100*	87	9	67	36	0*	67
Aenet	99	8	65	53	0*	99	88	14	54	35	0*	65
SCAD	100*	17	43	64	0*	92	88	17	45	28	0*	50
MC+	100*	18	41	59	1	87	88	21	41	27	0*	50
Scenario 3												
Ridge	100	100	13	100	100	13	100	100	13	100	100	13
Lasso	100*	56	24	52	2	87	98	89	15	13	2	20
Enet	100*	76	18	47	4	63	99	94	14	8	2	11
AL	100*	29	42	52	0*	100*	95	77	17	18	2	30
Aenet	100*	38	33	54	0*	99	96	82	16	15	1	25
SCAD	100*	34	33	58	0*	85	95	71	18	13	1	22
MC+	100*	35	32	56	0*	84	95	71	18	13	1	23

* Approximate value

the total number of selected covariates in the model. PPV describes how the model can approximate the oracle model in terms of selection. Therefore, we want to find the

Table 3.5: Empirical selection properties (TPR, FPR, and PPV in %) based on 2000 replications of Thomas processes on the domain D ($\mu = 1600$) for different values of κ and for the three 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	$\kappa = 5 \times 10^{-4}$						$\kappa = 5 \times 10^{-5}$					
	Regularized PL			Regularized WPL			Regularized PL			Regularized WPL		
	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV
Scenario 1												
Ridge	100	100	10	100	100	10	100	100	10	100	100	10
Lasso	100	26	35	52	0*	100*	98	48	22	56	0*	96
Enet	100	64	16	55	6	50	99	76	14	50	5	45
AL	100	0*	98	50	0	100	96	6	77	55	0*	98
Aenet	100	4	79	54	0*	100*	97	11	60	57	0*	96
SCAD	100	17	50	60	0*	100*	98	18	47	52	0*	90
MC+	100	22	47	60	0*	97	98	23	42	44	0*	79
Scenario 2												
Ridge	100	100	10	100	100	10	100	100	10	100	100	10
Lasso	100	26	33	51	0*	97	98	43	24	52	1	91
Enet	100	56	18	51	5	55	99	69	15	49	4	62
AL	100	1	92	51	0	100	96	10	67	53	0*	99
Aenet	100	4	78	51	0*	100*	97	15	52	53	0*	98
SCAD	100	21	37	53	1	85	96	16	50	45	1	77
MC+	100	24	35	47	2	76	97	19	47	42	2	72
Scenario 3												
Ridge	100	100	13	100	100	13	100	100	13	100	100	13
Lasso	100	69	19	52	1	96	100	95	14	48	4	75
Enet	100	85	16	52	5	71	100	97	14	43	5	62
AL	100	43	32	51	0*	100*	99	86	15	51	2	86
Aenet	100	49	27	52	0*	99	99	89	15	50	3	82
SCAD	100	47	27	43	2	72	99	78	17	40	2	63
MC+	100	48	26	44	2	75	99	79	17	37	2	61

* Approximate value

methods which have a TPR and a PPV close to 100%, and a FPR close to 0.

Generally, for both the penalized PL and the penalized WPL methods, the best

selection properties are obtained for a larger value of κ which shows weaker spatial dependence. For a more clustered one, indicated by a smaller value of κ, it seems more difficult to select the true covariates. As μ increases from 400 (Table 3.4) to 1600 (Table 3.5), the TPR tends to improve, so the model can select both z_1 and z_2 more frequently. Ridge, lasso, and elastic net are the regularization methods that cannot satisfy our theorems. It is firstly emphasized that all covariates are always selected by the ridge so that the rates are never changed whatever method used. For the penalized PL with lasso and elastic net regularization, it is shown that they tend to have quite large value of FPR, meaning that they wrongly keep the noisy covariates more frequently. When the penalized WPL is applied, we gain smaller FPR, but we suffer from smaller TPR at the same time. This smaller TPR actually comes from the unselection of z_2 which has smaller coefficient than that of z_1.

When we apply adaptive lasso, adaptive elastic net, SCAD, and MC+, we achieve better performance, especially for FPR which is closer to zero which automatically improves the PPV. Adaptive elastic net (resp. elastic net) has slightly larger FPR than adaptive lasso (resp. lasso). Among all regularization methods considered in this chapter, adaptive lasso seems to outperform the other ones.

Considering scenarios 1 and 2, we observe best selection properties for the penalized PL combined with adaptive lasso. As the design is getting more complex for scenario 3, applying the penalized PL suffers from much larger FPR, indicating that this method may not be able to overcome the complicated situation. However, when we use the penalized WPL, the properties seem to be more stable for the different designs of simulation study. One more advantage when considering the penalized WPL is that we can remove almost all extra covariates. It is worth noticing that we may suffer from smaller TPR when we apply the penalized WPL, but we lose the only less informative covariates. From Tables 3.4 and 3.5, when we are faced with complex situation, we would recommend the use of the penalized WPL method with adaptive lasso penalty if the focus is on selection properties. Otherwise, the use of the penalized PL combined with adaptive lasso penalty is more preferable.

Tables 3.6 and 3.7 give the prediction properties of the estimates in terms of bi-

Table 3.6: Empirical prediction properties (Bias, SD, and RMSE) based on 2000 replications of Thomas processes on the domain $D \ominus R$ ($\mu = 400$) for different values of κ and for the three 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	$\kappa = 5 \times 10^{-4}$						$\kappa = 5 \times 10^{-5}$					
	Regularized PL			Regularized WPL			Regularized PL			Regularized WPL		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario 1												
Oracle	0.11	0.18	0.21	0.64	0.20	0.67	0.29	0.81	0.86	0.57	0.54	0.78
Ridge	0.11	0.38	0.40	0.72	0.69	1.00	0.28	1.26	1.29	0.98	1.03	1.42
Lasso	0.28	0.32	0.42	1.06	0.32	1.11	0.47	0.99	1.10	1.40	0.73	1.58
Enet	0.24	0.38	0.44	1.28	0.28	1.31	0.45	1.04	1.13	1.59	0.58	1.70
AL	0.10	0.29	0.31	0.87	0.32	0.92	0.38	0.96	1.03	1.18	0.93	1.50
Aenet	0.14	0.30	0.33	0.93	0.39	1.01	0.40	0.96	1.04	1.29	0.82	1.53
SCAD	0.26	0.27	0.38	1.06	0.37	1.12	0.46	0.79	0.91	1.49	0.67	1.64
MC+	0.28	0.28	0.39	1.04	0.38	1.11	0.47	0.78	0.92	1.48	0.70	1.64
Scenario 2												
Oracle	0.12	0.23	0.26	0.71	0.26	0.76	0.30	0.78	0.84	0.59	0.62	0.84
Ridge	0.14	0.46	0.48	0.69	0.93	1.16	0.32	1.23	1.27	0.92	1.15	1.47
Lasso	0.34	0.33	0.48	1.20	0.37	1.26	0.45	0.96	1.06	1.50	0.69	1.65
Enet	0.38	0.40	0.55	1.40	0.35	1.44	0.44	1.03	1.12	1.78	0.49	1.85
AL	0.20	0.33	0.39	0.85	0.32	0.91	0.37	0.93	1.00	1.17	0.86	1.45
Aenet	0.25	0.33	0.42	0.96	0.34	1.02	0.40	0.94	1.02	1.29	0.78	1.51
SCAD	0.38	0.30	0.48	0.95	0.48	1.06	0.44	0.80	0.91	1.53	0.70	1.68
MC+	0.39	0.30	0.49	1.01	0.49	1.13	0.44	0.80	0.92	1.52	0.71	1.68
Scenario 3												
Oracle	0.12	0.46	0.48	0.70	0.26	0.75	0.65	1.14	1.31	0.87	0.88	1.24
Ridge	0.13	1.03	1.04	0.71	1.45	1.62	0.52	3.10	3.14	0.90	2.86	3.00
Lasso	0.20	0.69	0.71	1.26	0.40	1.32	0.51	2.91	2.95	1.93	0.68	2.04
Enet	0.21	0.83	0.86	1.53	0.40	1.58	0.52	2.94	2.99	2.03	0.60	2.12
AL	0.18	0.57	0.60	0.91	0.33	0.97	0.52	2.80	2.85	1.77	0.84	1.96
Aenet	0.22	0.61	0.65	1.04	0.36	1.10	0.52	2.80	2.85	1.86	0.73	2.00
SCAD	0.27	0.61	0.67	1.18	0.59	1.32	0.48	2.49	2.54	1.91	0.64	2.02
MC+	0.27	0.62	0.68	1.20	0.58	1.33	0.48	2.49	2.54	1.89	0.67	2.00

Table 3.7: Empirical prediction properties (Bias, SD, and RMSE) based on 2000 replications of Thomas processes on the domain D ($\mu = 1600$) for different values of κ and for the three 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	$\kappa = 5 \times 10^{-4}$						$\kappa = 5 \times 10^{-5}$					
	Regularized PL			Regularized WPL			Regularized PL			Regularized WPL		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario 1												
Oracle	0.05	0.11	0.12	0.33	0.15	0.37	0.16	0.45	0.48	0.41	0.22	0.46
Ridge	0.04	0.21	0.21	0.70	0.55	0.90	0.13	0.72	0.73	0.74	0.58	0.94
Lasso	0.14	0.19	0.24	1.03	0.20	1.05	0.23	0.60	0.64	0.99	0.43	1.08
Enet	0.11	0.22	0.24	1.14	0.29	1.17	0.20	0.62	0.65	1.12	0.43	1.20
AL	0.04	0.18	0.18	0.87	0.18	0.89	0.16	0.58	0.60	0.87	0.42	0.96
Aenet	0.05	0.18	0.18	0.96	0.22	0.99	0.17	0.58	0.60	0.90	0.48	1.02
SCAD	0.19	0.18	0.26	1.30	0.34	1.34	0.14	0.53	0.55	1.37	0.51	1.46
MC+	0.20	0.18	0.27	1.33	0.28	1.36	0.15	0.53	0.55	1.38	0.52	1.48
Scenario 2												
Oracle	0.05	0.15	0.16	0.36	0.17	0.40	0.18	0.46	0.49	0.39	0.26	0.47
Ridge	0.05	0.27	0.27	0.69	0.62	0.94	0.17	0.74	0.80	0.78	0.64	1.01
Lasso	0.16	0.20	0.25	1.16	0.24	1.18	0.23	0.60	0.64	1.14	0.43	1.22
Enet	0.17	0.23	0.29	1.24	0.24	1.26	0.23	0.63	0.67	1.33	0.42	1.40
AL	0.07	0.18	0.20	0.85	0.18	0.87	0.18	0.58	0.61	0.83	0.41	0.93
Aenet	0.09	0.19	0.21	0.94	0.20	0.96	0.20	0.59	0.62	0.92	0.41	1.01
SCAD	0.26	0.20	0.33	1.26	0.51	1.36	0.19	0.51	0.55	1.31	0.60	1.44
MC+	0.26	0.20	0.33	1.31	0.55	1.42	0.19	0.51	0.55	1.32	0.61	1.46
Scenario 3												
Oracle	0.13	0.31	0.34	0.43	0.18	0.47	0.31	0.96	1.01	0.75	0.35	0.83
Ridge	0.11	0.84	0.86	0.70	0.96	1.19	0.23	2.50	2.51	1.02	1.43	1.76
Lasso	0.12	0.64	0.65	1.14	0.29	1.17	0.22	2.41	2.42	1.40	0.61	1.52
Enet	0.13	0.71	0.73	1.35	0.30	1.39	0.23	2.42	2.43	1.63	0.56	1.73
AL	0.14	0.55	0.57	0.89	0.18	0.91	0.22	2.37	2.38	1.12	0.67	1.31
Aenet	0.15	0.56	0.58	1.00	0.22	1.03	0.22	2.36	2.37	1.26	0.64	1.41
SCAD	0.24	0.58	0.62	1.41	0.40	1.47	0.24	2.09	2.10	1.50	0.68	1.65
MC+	0.24	0.58	0.63	1.44	0.42	1.50	0.24	2.09	2.10	1.49	0.71	1.65

ases, standard deviations (SD), and square root of mean squared errors (RMSE), some criterions we define by

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

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 $j = 1, \ldots, p$, where $p = 20$ for scenarios 1 and 2, and $p = 15$ for scenario 3.

In general, the properties improve with larger value of κ and μ due to weaker spatial dependence and larger sample size. For the oracle model where the model contains only z_1 and z_2, the WPL estimates are more efficient than the PL estimates, particularly in the more clustered case, agreeing with the findings by Guan and Shen (2010).

When the regularization methods are applied, the bias increases in general, especially when we consider the penalized WPL method. The regularized WPL has a larger bias since this method does not select z_2 much more frequently. Furthermore, weighted method seems to introduce extra bias, even though the regularization is not considered as in the oracle model. For a low clustered process, the SD using the penalized WPL is similar to that of the penalized PL which may be because of the weaker dependence represented by larger κ, making weight surface $w(\cdot)$ closer to 1. However, a larger RMSE is obtained from the penalized WPL. When we observe the more clustered process, we obtain smaller SD using the penalized WPL which explains why in some cases (mainly scenario 3) the RMSE gets smaller.

For the ridge method, the bias is closest to that of the oracle model, but it has the largest SD. Among the regularization methods, the adaptive lasso method has the best performance in terms of prediction.

Considering scenarios 1 and 2, we obtain best properties when we apply the penalized PL with adaptive lasso penalty. As the design is getting much more complex for scenario 3, when we use the penalized PL with adaptive lasso, the SD is doubled and even quadrupled due to the overselection of many unimportant covariates. In particular, for the more clustered process, the better properties are even obtained by applying the regularized WPL combined with adaptive lasso. From Tables 3.6 and 3.7, when the

focus is on prediction properties, we would recommend to apply the penalized WPL combined with adaptive lasso penalty when the observed point pattern is very clustered and when covariates have a complex stucture of covariance matrix. Otherwise, the use of the penalized PL combined with adaptive lasso penalty is more favorable. Our recommendations in terms of prediction support as what we recommend in terms of selection.

3.7.3 Logistic regression

Our concern here is to compare the estimates of the penalized (un)weighted logistic likelihood to that of the penalized (un)weighted Poisson likelihood with different number of dummy points. We remind that the number of dummy points comes up when we discretize the integral terms in (3.5) and in (3.6). In the following, to ease the presentation, we use the term Poisson estimates (resp. logistic estimates) for parameter estimates obtained using the regularized Poisson likelihood (resp. the regularized logistic regression likelihood).

Table 3.8: Empirical selection properties (TPR, FPR, and PPV in %) based on 2000 replications of Thomas processes on the domain D ($\mu = 1600$) for $\kappa = 5 \times 10^{-5}$, for two different scenarios, and for three different numbers of dummy points. Different estimating equations are considered, the regularized (un)weighted Poisson and (un)weighted logistic regression likelihoods, employing adaptive lasso regularization method.

Method	nd	Scenario 2						Scenario 3					
		Unweighted			Weighted			Unweighted			Weighted		
		TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV
Poisson	20	96	35	32	53	0*	96	98	82	16	47	2	79
	40	95	6	77	52	0*	95	98	83	16	46	2	77
	80	95	4	83	50	0*	94	98	83	16	43	2	74
Logistic	20	94	11	60	49	0*	91	98	72	20	41	2	73
	40	94	8	67	50	0*	93	99	81	16	43	2	74
	80	94	5	77	50	0*	93	99	83	16	42	2	73

* Approximate value

We consider three different numbers of dummy points denoted by $\texttt{nd}^2$. By these different numbers of dummy points, we want to observe the properties with three

Table 3.9: Empirical prediction properties (Bias, SD, and RMSE) based on 2000 replications of Thomas processes on the domain D ($\mu = 1600$) for $\kappa = 5 \times 10^{-5}$, for two different scenarios, and for three different numbers of dummy points. Different estimating equations are considered, the regularized (un)weighted Poisson and (un)weighted logistic regression likelihoods, employing adaptive lasso regularization method.

Method	nd	Scenario 2						Scenario 3					
		Unweighted			Weighted			Unweighted			Weighted		
		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
No regularization													
Poisson	20	0.37	0.64	0.74	0.29	0.74	0.79	0.28	2.15	2.16	0.42	2.06	2.11
	40	0.14	0.63	0.65	0.16	0.73	0.75	0.33	2.47	2.50	0.42	2.32	2.35
	80	0.17	0.64	0.66	0.11	0.75	0.76	0.26	2.57	2.58	0.43	2.40	2.43
Logistic	20	0.03	0.69	0.69	0.32	1.34	1.37	0.20	2.31	2.32	0.36	2.95	2.97
	40	0.07	0.60	0.61	0.12	0.96	0.97	0.23	2.31	2.32	0.37	2.56	2.58
	80	0.10	0.60	0.61	0.14	0.81	0.82	0.25	2.36	2.38	0.42	2.38	2.42
Adaptive lasso													
Poisson	20	0.30	0.59	0.67	0.86	0.47	0.98	0.30	2.00	2.03	1.14	0.68	1.33
	40	0.20	0.58	0.61	0.86	0.49	0.99	0.33	2.33	2.35	1.18	0.70	1.37
	80	0.18	0.59	0.62	0.88	0.51	1.02	0.28	2.41	2.43	1.22	0.71	1.41
Logistic	20	0.19	0.50	0.53	0.95	0.55	1.09	0.23	2.06	2.07	1.26	0.73	1.45
	40	0.18	0.52	0.55	0.89	0.52	1.03	0.23	2.15	2.16	1.22	0.72	1.42
	80	0.18	0.55	0.58	0.89	0.52	1.03	0.25	2.21	2.22	1.24	0.71	1.43

different situations: (a) $\texttt{nd}^2 < m$, (b) $\texttt{nd}^2 \approx m$, and (c) $\texttt{nd}^2 > m$, where m is the number of points. In the following, $m \approx 1600$ and $\texttt{nd}^2 = 400$, 1600, and 6400. Note that the choice by default from the Poisson likelihood in `spatstat` corresponds to case (c). Baddeley et al. (2014) showed that for datasets with very large number of points and for very structured point processes, the logistic likelihood method is clearly preferable as it requires a smaller number of dummy points to perform quickly and efficiently. We want to investigate a similar comparison when these methods are regularized.

We only repeat the results for $\kappa = 5 \times 10^{-5}$ and $\mu = 1600$, and for scenarios 2 and 3. We use the same selection and prediction indices examined in Section 3.7.2 and consider only the adaptive lasso method.

Table 3.8 presents selection properties for the Poisson and logistic likelihoods with adaptive lasso regularization. For unweighted versions of the procedure, the regularized logistic method outperforms the regularized Poisson method when $\texttt{nd} = 20$, i.e. when the number of dummy points is much smaller than the number of points. When $\texttt{nd}^2 \approx m$ or $\texttt{nd}^2 > m$, the methods tend to have similar performances. When we consider weighted versions of the regularized logistic and Poisson likelihoods, the results do not change that much with nd and the regularized Poisson likelihood method slightly outperforms the regularized logistic likelihood method. In addition, for scenario 3 which considers a more complex situation, the methods tend to select the noisy covariates much more frequently.

Empirical biases, standard deviation and square root of mean squared errors are presented in Table 3.9. We include all empirical results for the standard Poisson and logistic estimates (i.e. no regularization is considered). Let us first consider the unweighted methods with no regularization. The logistic method clearly has smaller bias, especially when $\texttt{nd} = 20$, which explains why in most situations the RMSE is smaller. However, for the weighted methods, although the logistic method has smaller bias in general, it produces much larger SD, leading to larger RMSE for all cases. When we compare the weighted and the unweighted methods for logistic estimates, in general, not only do we fail to reduce the SD, but we also have larger bias. When the adaptive lasso regularization is considered, combined with the unweighted methods, we can preserve the bias in general and simultaneously improve the SD, and hence improve the RMSE. The logistic likelihood method slightly outperforms the Poisson likelihood method. When the weighted methods are considered, we obtain smaller SD, but we have larger bias. For weighted versions of the Poisson and logistic likelihoods, the results do not change that much with nd and the weighted Poisson method slightly outperforms the weighted logistic method. From Tables 3.8 and 3.9, when the number of dummy points can be chosen as $\texttt{nd}^2 \approx m$ or $\texttt{nd}^2 > m$, we would recommend to apply the Poisson likelihood method. When the number of dummy points should be chosen as $\texttt{nd}^2 < m$, the logistic likelihood method is more favorable. Our recommendations regarding whether weighted or unweighted methods follow the ones as in Section 3.7.2.

3.8 Application to forestry datasets

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, censuses have been carried out where 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 more than 300 species (see Condit, 1998; Hubbell et al., 1999, 2005). It is of interest to know how the very high number of different tree species continues to coexist, profiting from different habitats determined by e.g. topography or soil properties (see e.g. Waagepetersen, 2007; Waagepetersen and Guan, 2009). In particular, the selection of covariates among topological attributes and soil minerals as well as the estimation of their coefficients are becoming our most concern.

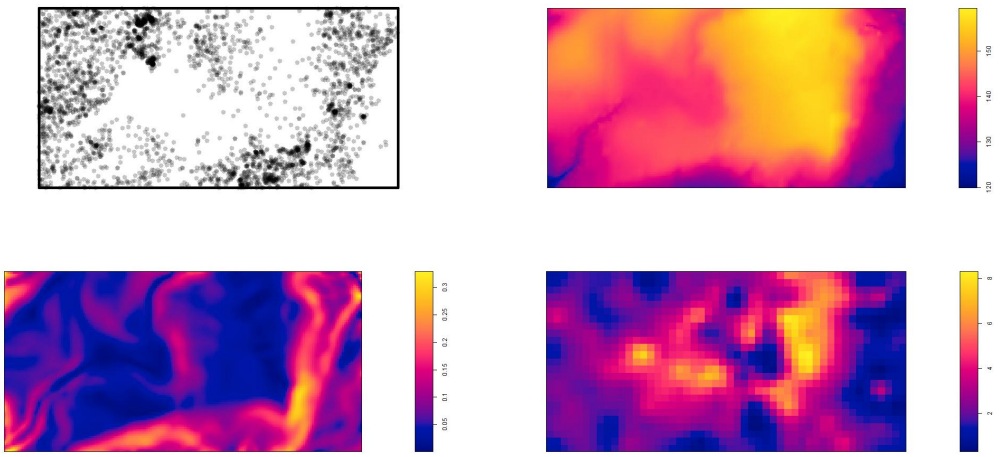

Figure 3.2: Maps of locations of BPL trees (top left), elevation (top right), slope (bottom left), and concentration of phosporus (bottom right).

We are particularly interested in analyzing the locations of 3,604 *Beilschmiedia pendula Lauraceae* (BPL) tree stems. We model the intensity of BPL trees as a log-linear function of two topological attributes and 13 soil properties as the covariates. Figure 3.2 contains maps of the locations of BPL trees, elevation, slope, and concentration of Phosporus. BPL trees seem to appear in greater abundance in the areas of high

elevation, steep slope, and low concentration of Phosporus. The covariates maps are depicted in Figure 3.4.

Table 3.10: Barro Colorado Island data analysis: Parameter estimates of the regression coefficients for Beilschmiedia pendula Lauraceae trees applying regularized (un)weighted Poisson and logistic regression likelihoods with adaptive lasso regularization.

	Unweighted method		Weighted method	
	Poisson estimates	Logistic estimates	Poisson estimates	Logistic estimates
Elev	0.39	0.40	0.41	0.45
Slope	0.26	0.32	0.51	0.60
Al	0	0	0	0
B	0.30	0.30	0	0
Ca	0.10	0.15	0	0
Cu	0.10	0.12	0	0
Fe	0.05	0	0	0
K	0	0	0	0
Mg	-0.17	-0.18	0	0
Mn	0.12	0.13	0.23	0.24
P	-0.60	-0.60	-0.50	-0.52
Zn	-0.43	-0.46	-0.35	-0.37
N	0	0	0	0
N.min	-0.12	-0.10	0	0
pH	-0.14	-0.14	0	0
Nb of cov.	12	11	5	5

We apply the regularized (un)weighted Poisson and the logistic likelihoods, combined with adaptive lasso regularization to select and estimate parameters. Since we do not deal with datasets which have very large number of points, we can set the default number of dummy points for Poisson likelihood as in the `spatstat` package, i.e. the number of dummy points can be chosen to be larger than the number of points, to perform quickly and efficiently. It is worth emphasizing that we center and scale the 15 covariates to observe which one has the largest effect on the intensity. The results are presented in Table 3.10: 12 covariates for the Poisson likelihood and 11 for the lo-

gistic method are selected out of the 15 covariates using the unweighted methods while only 5 covariates (both for the Poisson and logistic methods) are selected using the weighted versions. The unweighted methods tend to overfit the model by overselecting unimportant covariates.

The weighted methods tend to keep out the uninformative covariates. Both Poisson and logistic estimates own similar selection and estimation results. First, we find some differences on estimation between the unweighted and the weighted methods, especially for slope and Manganese (Mn), for which the weighted methods have approximately two times larger estimators. Second, we may loose some nonzero covariates when we apply the weighted methods, even though it is only for the covariates which have relatively small coefficient. Boron (B) has high correlation with many of the other covariates, particularly with those which are not selected. This is possibly why Boron, which is selected and may have nonnegligible coefficient in the unweighted methods, is not chosen in the model. This may explain why the weighted methods introduce extra biases. However, since the situation appears to be quite close to the scenario 3 from the simulation study, the weighted methods are more favorable in terms of both selection and prediction.

In this application, we do not face any computational problem. Nevertheless, if we have to model a species of trees with much more points, the default value for `nd` will lead to numerical problems. In such a case, the logistic likelihood would be a good alternative.

These results suggest that BPL trees favor to live in areas of higher elevation and slope. This result is different from the findings by Waagepetersen (2007) and Guan and Loh (2007) which concluded based on standard error estimation that BPL trees do not really prefer either high or low altitudes. However, we have the same conclusion with the analysis by Guan and Shen (2010) and Thurman et al. (2015) that BPL trees prefer to live on higher altitudes. Further, higher levels of Manganese (Mn) and lower levels of both Phosporus (P) and Zinc (Zn) concentrations in soil are associated with higher appearance of BPL trees.

3.9 Conclusion and discussion

We develop regularized versions of estimating equations based on Campbell theorem derived from the Poisson and the logistic likelihoods. Our procedure is able to estimate intensity function of spatial point processes, when the intensity is a function of many covariates and has a log-linear form. Furthermore, our procedure is also generally easy to implement in `R` since we need to combine `spatstat` package with `glmnet` and `ncvreg` packages. We study the asymptotic properties of both regularized weighted Poisson and logistic estimates in terms of consistency, sparsity, and normality distribution. We find that, among the regularization methods considered in this work, adaptive lasso, adaptive elastic net, SCAD, and MC+ are the methods that can satisfy our theorems.

We carry out some scenarios in the simulation study to observe selection and prediction properties of the estimates. We compare the penalized Poisson likelihood (PL) and the penalized weighted Poisson likelihood (WPL) with different penalty functions. From the results, when we deal with covariates having a complex covariance matrix and when the point pattern looks quite clustered, we recommend to apply the penalized WPL combined with adaptive lasso regularization. Otherwise, the regularized PL with adaptive lasso is more preferable. The further and more careful investigation to choose the tuning parameters may be needed to improve the selection properties. We note the bias increases quite significantly when the regularized WPL is applied. When the penalized WPL is considered, a two-step procedure may be needed to improve the prediction properties: (1) use the penalized WPL combined with adaptive lasso to chose the covariates, then (2) use the selected covariates to obtain the estimates. This post-selection inference procedure has not been investigated in this study.

We also compare the estimates obtained from the Poisson and the logistic likelihoods. When the number of dummy points can be chosen to be either similar to or larger than the number of points, we recommend the use of the Poisson likelihood method. Nevertheless, when the number of dummy points should be chosen to be smaller than the number of points, the logistic method is more favorable.

A further work would consist in studying the situation when the number of the

covariates is much larger than the sample size. In such a situation, the coordinate descent algorithm used in this study may cause some numerical troubles. The Dantzig selector procedure introduced by Candes and Tao (2007) might be a good alternative as the implementaion for linear models (and for generalized linear models) results in a linear programming. It would be interesting to bring this approach to spatial point process setting.

3.10 Maps of covariates

In this section, we present the maps of covariates used in the simulation studies (scenarios 2 and refsce3) and in the application.

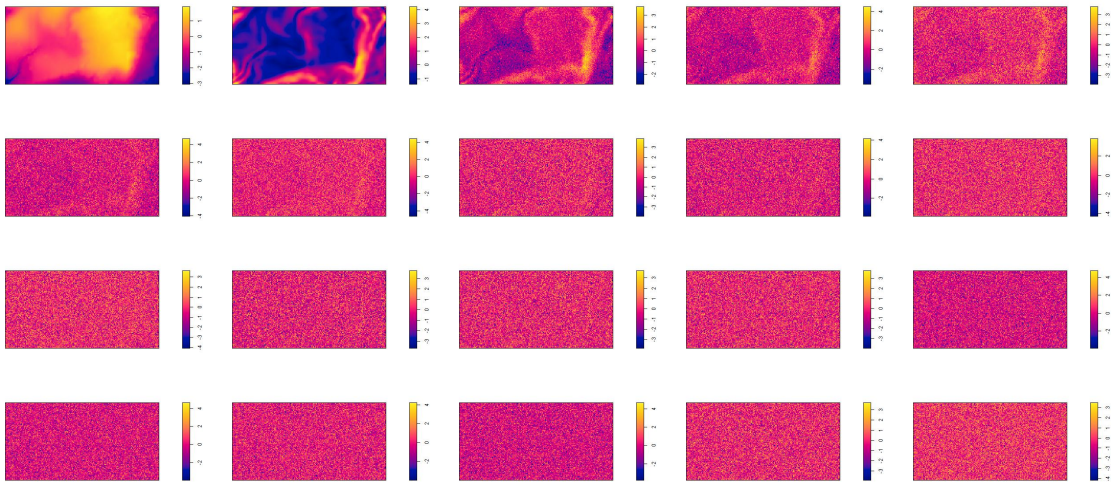

Figure 3.3: Maps of covariates designed in scenario 2. The first two top left images are the elevation and the slope. The other 18 covariates are generated as standard Gaussian white noise but transformed to get multicollinearity.

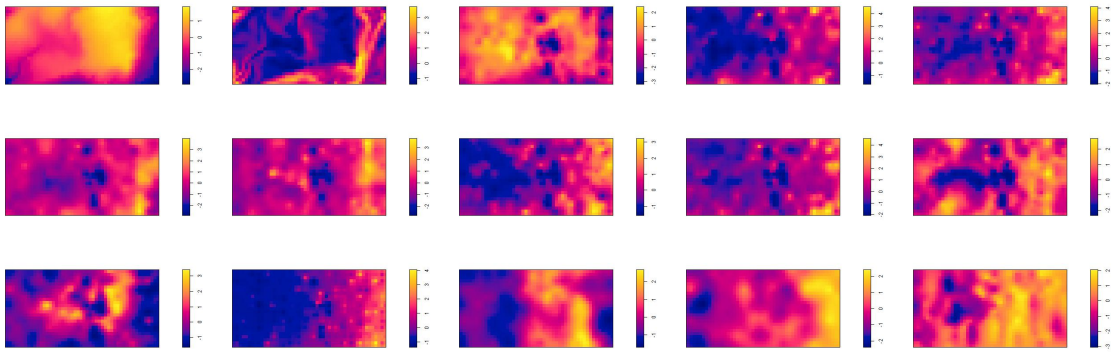

Figure 3.4: Maps of covariates used in scenario 3 and in application. From left to right: Elevation, slope, Aluminium, Boron, and Calcium (1st row), Copper, Iron, Potassium, Magnesium, and Manganese (2nd row), Phosporus, Zinc, Nitrogen, Nitrigen mineralisation, and pH (3rd row).

3.11 Proofs of the main results

3.11.1 Auxiliary Lemma

The following result is used in the proof of Theorems 3.6.1-3.6.2. 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 3.11.1. *Under the conditions (C.1)-(C.6), the following convergence holds in distribution as* $n \to \infty$

$$\{\mathbf{B}_n(w;\boldsymbol{\beta}_0) + \mathbf{C}_n(w;\boldsymbol{\beta}_0)\}^{-1/2}\ell_n^{(1)}(w;\boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \tag{3.21}$$

Moreover as $n \to \infty$,

$$|D_n|^{-\frac{1}{2}}\ell_n^{(1)}(w;\boldsymbol{\beta}_0) = O_{\mathrm{P}}(1). \tag{3.22}$$

Proof. Let us first note that using Campbell Theorems (5.2)-(5.3)

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

The proof of (3.21) 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

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

where

$$Y_{i,n} = \sum_{u \in \mathbf{X} \cap C_{i,n}} w(u)\mathbf{z}(u) - \int_{C_{i,n}} w(u)\mathbf{z}(u) \exp(\boldsymbol{\beta}_0^\top \mathbf{z}(u)) \mathrm{d}u.$$

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

$$\sup_{n \geq 1} \sup_{i \in \mathscr{I}_n} \mathbb{E}(\|Y_{i,n}\|^{2+\delta}) < \infty. \tag{3.23}$$

If we combine (3.23) 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, to obtain (3.21) which also implies that

$$\{\mathbf{B}_n(w; \boldsymbol{\beta}_0) + \mathbf{C}_n(w; \boldsymbol{\beta}_0)\}^{-1/2} \ell_n^{(1)}(w; \boldsymbol{\beta}_0) = O_{\mathrm{P}}(1)$$

as $n \to \infty$. The second result (3.22) is deduced from condition ($\mathcal{C}$.6) which in particular implies that $|D_n|^{1/2}\{\mathbf{B}_n(w; \boldsymbol{\beta}_0) + \mathbf{C}_n(w; \boldsymbol{\beta}_0)\}^{-1/2} = O(1)$.

□

3.11.2 Proof of Theorem 3.6.1

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 $d_n = |D_n|^{-1/2} + a_n$, and $\mathbf{k} = \{k_1, k_2, \ldots, k_p\}^\top \in \mathbb{R}^p$. We remind the reader that the estimate of $\boldsymbol{\beta}_0$ is defined as the maximum of the function Q (given by (3.7)) over Θ, an open convex bounded set of $\mathbb{R}^p$. 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 3.6.1, we follow the main argument by Fan and Li (2001) and aim at proving that for any given $\epsilon > 0$, there exists $K > 0$ such that for n sufficiently large

$$\mathrm{P}\left(\sup_{\|\mathbf{k}\| = K} \Delta_n(\mathbf{k}) > 0 \right) \leq \epsilon, \qquad \text{where } \Delta_n(\mathbf{k}) = Q(w; \boldsymbol{\beta}_0 + d_n\mathbf{k}) - Q(w; \boldsymbol{\beta}_0). \tag{3.24}$$

Equation (3.24) 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 $\hat{\boldsymbol{\beta}}$ 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

$$T_1 = \ell_n(w; \boldsymbol{\beta}_0 + d_n\mathbf{k}) - \ell_n(w; \boldsymbol{\beta}_0)$$

$$T_2 = |D_n| \sum_{j=1}^{p} \Big(p_{\lambda_{n,j}}(|\beta_{0j}|) - p_{\lambda_{n,j}}(|\beta_{0j} + d_n k_j|) \Big).$$

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

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

Since Θ is convex and bounded and since $w(\cdot)$ and $\mathbf{z}(\cdot)$ are uniformly bounded by conditions ($\mathcal{C}$.2)-($\mathcal{C}$.3), there exists a nonnegative constant κ such that

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

Let $\nu_{\min}(\mathbf{M})$ be the smallest eigenvalue of a squared matrix $\mathbf{M}$. By condition ($\mathcal{C}$.7),

$$\check{\nu} := \liminf_{n\to\infty} \nu_{\min}(|D_n|^{-1} \mathbf{A}_n(w;\boldsymbol{\beta}_0)) = \liminf_{n\to\infty} \frac{\mathbf{k}^\top \left(|D_n|^{-1} \mathbf{A}_n(w;\boldsymbol{\beta}_0) \right) \mathbf{k}}{\|\mathbf{k}\|^2} > 0.$$

Hence

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

Regarding the term T_2,

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

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

Since $d_n |D_n|^{1/2} = O(1)$, then by ($\mathcal{C}$.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

$$\begin{aligned} -T_2' = & d_n |D_n| \sum_{j=1}^{s} k_j p'_{\lambda_{n,j}}(|\beta_{0j}|) \operatorname{sign}(\beta_{0,j}) + \frac{1}{2} d_n^2 |D_n| \sum_{j=1}^{s} k_j^2 p''_{\lambda_{n,j}}(|\beta_{0j}|) \\ & + \frac{1}{6} d_n^3 |D_n| \sum_{j=1}^{s} k_j^3 p'''_{\lambda_{n,j}}(|\beta_{0j} + t_j d_n k_j|). \end{aligned}$$

Now by definition of a_n and c_n and from condition ($\mathcal{C}$.8), we deduce that there exists κ such that

$$T_2' \leq a_n d_n |D_n| \, |\mathbf{k}^\top \mathbf{1}| + \frac{1}{2} c_n d_n^2 |D_n| \|\mathbf{k}\|^2 + \kappa d_n^3 |D_n|$$

$$\leq \sqrt{s}a_n d_n |D_n| \|\mathbf{k}\| + \frac{1}{2} c_n d_n^2 |D_n| \|\mathbf{k}\|^2 + \kappa d_n^3 |D_n|$$

from Cauchy-Schwarz inequality. Since $c_n = o(1)$, $d_n = o(1)$ and $a_n d_n |D_n| = O(d_n^2 |D_n|)$, then for n sufficiently large

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

We now return to (3.24): for n sufficiently large

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

Since $d_n |D_n| = O(|D_n|^{1/2})$, by choosing K large enough, there exists κ such that for n sufficiently large

$$\mathrm{P}\left(\sup_{\|\mathbf{k}\|=K} \Delta_n(\mathbf{k}) > 0 \right) \leq \mathrm{P}\left(\|\ell_n^{(1)}(w; \boldsymbol{\beta}_0)\| > \kappa |D_n|^{1/2} \right) \leq \epsilon$$

for any given $\epsilon > 0$ from (3.22).

□

3.11.3 Proof of Theorem 3.6.2

To prove Theorem 3.6.2(i), we provide Lemma 3.11.2 as follows.

Lemma 3.11.2. *Assume the conditions (C.1)-(C.6) and condition (C.8) hold. If* $a_n = O(|D_n|^{-1/2})$ *and* $b_n |D_n|^{1/2} \to \infty$ *as* $n \to \infty$, *then with probability tending to* 1, *for any* $\boldsymbol{\beta}_1$ *satisfying* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = O_{\mathrm{P}}(|D_n|^{-1/2})$, *and for any constant* $K_1 > 0$,

$$Q\Big(w; ({\boldsymbol{\beta}_1}^\top, \mathbf{0}^\top)^\top\Big) = \max_{\|\boldsymbol{\beta}_2\| \leq K_1 |D_n|^{-1/2}} Q\Big(w; ({\boldsymbol{\beta}_1}^\top, {\boldsymbol{\beta}_2}^\top)^\top\Big).$$

Proof. It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = O_{\mathrm{P}}(|D_n|^{-1/2})$, for some small $\varepsilon_n = K_1 |D_n|^{-1/2}$, and for $j = s+1, \ldots, p$,

$$\frac{\partial Q(w; \boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for } 0 < \beta_j < \varepsilon_n, \text{ and} \tag{3.25}$$

$$\frac{\partial Q(w; \boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for } -\varepsilon_n < \beta_j < 0. \tag{3.26}$$

First note that by (3.22), we obtain $\|\ell_n^{(1)}(w; \boldsymbol{\beta}_0)\| = O_{\mathrm{P}}(|D_n|^{1/2})$. Second, by con-

ditions ($\mathcal{C}$.2)-($\mathcal{C}$.3), there exists $t \in (0,1)$ such that

$$\begin{aligned}\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + t\sum_{l=1}^{p} \frac{\partial^2 \ell_n(w;\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}-\boldsymbol{\beta}_0))}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{0l}) \\ &= O_{\mathrm{P}}(|D_n|^{1/2}) + O_{\mathrm{P}}(|D_n||D_n|^{-1/2}) = O_{\mathrm{P}}(|D_n|^{1/2}).\end{aligned}$$

Third, let $0 < \beta_j < \varepsilon_n$ and b_n the sequence given by (3.17). By condition ($\mathcal{C}$.8), b_n is well-defined and since by assumption $b_n|D_n|^{1/2} \to \infty$, in particular, $b_n > 0$ for n sufficiently large. Therefore, for n sufficiently large,

$$\begin{aligned}\mathrm{P}\left(\frac{\partial Q(w;\boldsymbol{\beta})}{\partial \beta_j} < 0\right) &= \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} - |D_n| p'_{\lambda_{n,j}}(|\beta_j|)\operatorname{sign}(\beta_j) < 0\right) \\ &= \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} < |D_n| p'_{\lambda_{n,j}}(|\beta_j|)\right) \\ &\geq \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} < |D_n| b_n\right) \\ &= \mathrm{P}\left(\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} < |D_n|^{1/2}|D_n|^{1/2} b_n\right).\end{aligned}$$

$\mathrm{P}\left(\partial Q(w;\boldsymbol{\beta})/\partial \beta_j < 0\right) \to 1$ as $n \to \infty$ since $\partial \ell_n(w;\boldsymbol{\beta})/\partial \beta_j = O_{\mathrm{P}}(|D_n|^{1/2})$ and $b_n|D_n|^{1/2} \to \infty$. This proves (3.25). We proceed similarly to prove (3.26). □

Proof. We now focus on the proof of Theorem 3.6.2. Since Theorem 3.6.2(i) is proved by Lemma 3.11.2, we only need to prove Theorem 3.6.2(ii), which is the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$. As shown in Theorem 3.6.1, there is a root-$|D_n|$ consistent local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(w;\boldsymbol{\beta})$, and it can be shown that there exists an estimator $\hat{\boldsymbol{\beta}}_1$ in Theorem 3.6.1 that is a root-$(|D_n|)$ consistent local maximizer of $Q\Big(w;({\boldsymbol{\beta}_1}^\top, \mathbf{0}^\top)^\top\Big)$, which is regarded as a function of $\boldsymbol{\beta}_1$, and that satisfies

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

There exists $t \in (0,1)$ and $\breve{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + t(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})$ such that

$$\begin{aligned}0 =& \frac{\partial \ell_n(w;\hat{\boldsymbol{\beta}})}{\partial \beta_j} - |D_n| p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\operatorname{sign}(\hat{\beta}_j) \\ =& \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \frac{\partial^2 \ell_n(w;\breve{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) - |D_n| p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\operatorname{sign}(\hat{\beta}_j) \\ =& \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \frac{\partial^2 \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) + \sum_{l=1}^{s} \Psi_{n,jl}(\hat{\beta}_l - \beta_{0l}) \\ & - |D_n| p'_{\lambda_{n,j}}(|\beta_{0j}|)\operatorname{sign}(\beta_{0j}) - |D_n|\phi_{n,j}, \end{aligned} \tag{3.27}$$

where

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

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

$$\begin{aligned}
\phi_{n,j} &= p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\,\mathrm{sign}(\hat{\beta}_j) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\,\mathrm{sign}(\beta_{0j}) \\
&= \Big(p'_{\lambda_{n,j}}(|\hat{\beta}_j|) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\Big)\,\mathrm{sign}(\beta_{0j}) + p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\Big(\mathrm{sign}(\hat{\beta}_j) - \mathrm{sign}(\beta_{0j})\Big) \\
&\leq \kappa\Big||\hat{\beta}_j| - |\beta_{0j}|\Big| + 2a_n \\
&\leq \kappa|\hat{\beta}_j - \beta_{0j}| + 2a_n. \qquad (3.28)
\end{aligned}$$

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 $\breve{\beta}_j = \hat{\beta}_j + t(\beta_{0j} - \hat{\beta}_j)$ such that

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

where the latter equation ensues from Theorem 3.6.1 and condition ($\mathcal{C}$.8). Again, from Theorem 3.6.1, $\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \xrightarrow{L^1} 1$ which implies that $\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \xrightarrow{\mathrm{P}} 1$, so $T_1 = p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\Big(1 + o_{\mathrm{P}}(1)\Big) + O_{\mathrm{P}}(|D_n|^{-1})$.

Regarding the term T_2, we have by (3.28)

$$T_2 \leq \{\kappa|\hat{\beta}_j - \beta_{0j}| + 2a_n\}\,\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}).$$

By Theorem 3.6.1, $|\hat{\beta}_j - \beta_{0j}| = O_{\mathrm{P}}(|D_n|^{-1/2})$ and $\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}) = o_{\mathrm{P}}(1)$, since $a_n = O(|D_n|^{-1/2})$, we obtain $T_2 = o_{\mathrm{P}}(|D_n|^{-1/2})$, and we deduce that

$$\phi_{n,j} = p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\Big(1 + o_{\mathrm{P}}(1)\Big) + o_{\mathrm{P}}(|D_n|^{-1/2}). \qquad (3.29)$$

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

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

We rewrite both sides of (3.27) as

$$\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0)+\ell^{(2)}_{n,1}(w;\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})+\boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})-|D_n|\mathbf{p}'_n-|D_n|\boldsymbol{\phi}_n=0. \quad (3.30)$$

By definition of $\boldsymbol{\Pi}_n$ given by (3.20) and from (3.29), we obtain $\boldsymbol{\phi}_n=\boldsymbol{\Pi}_n(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})\big(1+o_{\mathrm{P}}(1)\big)+o_{\mathrm{P}}(|D_n|^{-1/2})$. Using this, we deduce, by premultiplying both sides of (3.30) by $\mathbf{M}_n$, that

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

The condition ($\mathcal{C}$.6) implies that there exists an $s\times s$ positive definite matrix $\mathbf{I}''_0$ such that for all sufficiently large n, we have $|D_n|^{-1}(\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0)+\mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0))\geq\mathbf{I}''_0$, hence $\|\mathbf{M}_n\|=O(|D_n|^{-1/2})$.

Now, $\|\boldsymbol{\Psi}_n\|=O_{\mathrm{P}}(|D_n|^{1/2})$ by conditions ($\mathcal{C}$.2)-($\mathcal{C}$.3) and by Theorem 3.6.1, and $\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}\|=O_{\mathrm{P}}(|D_n|^{-1/2})$ by Theorem 3.6.1 and by Theorem 3.6.2(i). Finally, since by assumption $a_n=o(|D_n|^{-1/2})$, we deduce that

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

Therefore, we have that

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

From (3.21), Theorem 3.6.2(i) and by Slutsky's Theorem, we deduce that

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

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

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

where $\boldsymbol{\Sigma}_n(w,\boldsymbol{\beta}_0)$ is given by (3.19). □

CHAPTER 4

Regularized Poisson and logistic regression methods for spatial point processes intensity estimation with a diverging number of covariates

4.1 Introduction

Intensity estimation for inhomogeneous spatial point processes have become one of the main interests in many applications and it is often assumed that the intensity can be modeled as a parametric function of certain covariates (see e.g. Møller and Waagepetersen, 2007; Renner and Warton, 2013; Yue and Loh, 2015). For parametric estimation, while maximum likelihood estimation (e.g. Berman and Turner, 1992; Rathbun and Cressie, 1994) has been widely implemented for Poisson point process models, estimating equation-based methods (e.g. Waagepetersen, 2007, 2008; Guan and Shen, 2010; Baddeley et al., 2014) are more simple to implement for more general spatial point process 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 may become undesirable. First, these methods cannot perform variable selection which leads to hard interpretation of the model. Second, as the number of covariates is large, employing these methods will yield large variance for parameter estimates.

In this study, we consider feature selection procedures for spatial point processes

intensity estimation. We model the intensity as a log-linear form of some covariates:

$$\rho(u;\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u)). \tag{4.1}$$

More precisely, $\mathbf{z}(u) = \{z_1(u), \ldots, z_p(u)\}^\top$ are the p spatial covariates measured at location u and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^\top$ is a real p-dimensional parameter. Such variable selection procedures have been previously investigated, for example, by Renner and Warton (2013); Thurman and Zhu (2014) for Poisson point process model and by Thurman et al. (2015), and in Chapter 2, for more general spatial point processes model. The main idea is to regularize the estimating functions based on Campbell theorem by a chosen penalty function. Theoretical study has been conducted by Thurman et al. (2015) who established the asymptotic properties of the estimates in terms of consistency, sparsity, and normality distribution using regularized estimating functions based on Campbell theorem derived from Poisson likelihood with adaptive lasso penalty, while we studied in Chapter 2 the same asymptotic properties by penalizing the estimating functions derived from both Poisson and logistic regression likelihoods considering more general penalty functions.

It is worth emphasizing that the previously mentioned procedures only considered a finite number of covariates p. In recent decades, with the advancement of technology and huge investment in data collection, more complex spatial data with a plenty of covariates have been rapidly available, so the setting when the number of parameter diverges should be considered. For example, 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, censuses have been carried out where 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 more than 300 species (see Condit, 1998; Hubbell et al., 1999, 2005). At the same region, many environmental covariates such as topographical attributes and soil properties have been also collected. Modeling the intensity of a specific tree species as a function of environmental covariates and their possible interactions can increase the number of covariates considerably. This chapter intends to extend the

results we obtained in Chapter 2 to the case where the number of parameters diverges.

Asymptotic properties which considers a diverging number of parameters for M-estimators have long been studied (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 regression estimator under finite number of parameters setting are encouraging, there are many naive and simple model selection procedures which possess those properties. As importance of the validity of these asymptotic properties for a diverging number of parameters setting, we consider to study this type of asymptotic properties in spatial point processes setting.

We investigate in this chapter the asymptotic properties of the estimates obtained from regularized Poisson and logistic regression methods studied in Chapter 2 but considering the setting where the number of covariates diverges as the domain of observation increases. We show that under some conditions, if the number of covariates does not grow too fast with respect to the observation domain, our estimates satisfy consistency, sparsity, and normality distribution. It is worth noticing that we do not make any assumption on the distribution of spatial point process, making our results available for large classes of spatial point process. Furthermore, our procedure does not require further effort for computational implementation since we combine the `spatstat` (Baddeley et al., 2015) `R` package with the two `R` packages `glmnet` (Friedman et al., 2010) and `ncvreg` (Breheny and Huang, 2011).

The rest of this chapter is organized as follows. In Section 4.2, we introduce brief background on spatial point processes as well as regularization methods for spatial point processes intensity estimation. Section 4.3 presents our asymptotic results. We investigate in Section 4.4 the finite sample performance of the proposed methods in a simulation study and in an application to tropical forestry datasets. Conclusion and discussion are presented in Section 4.5. Proofs of the main results are postponed to Sections 4.6.1-4.6.3.

4.2 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 in Chapter 2 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. 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 $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$

$$\mathbb{E} \sum_{u \in \mathbf{X}} k(u) = \int k(u)\rho(u)\mathrm{d}u \tag{4.2}$$

$$\mathbb{E} \sum_{u,v \in \mathbf{X}}^{\neq} k(u, v) = \int \int k(u, v)\rho^{(2)}(u, v)\mathrm{d}u\mathrm{d}v. \tag{4.3}$$

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 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).

Let $\ell(w; \boldsymbol{\beta})$ be the weighted Poisson likelihood (e.g. Guan and Shen, 2010) or the weighted logistic regression likelihood (e.g. Baddeley et al., 2014) given respectively by

$$\ell_{\mathrm{PL}}(w; \boldsymbol{\beta}) = \sum_{u \in \mathbf{X} \cap D} w(u) \log \rho(u; \boldsymbol{\beta}) - \int_D w(u)\rho(u; \boldsymbol{\beta})\mathrm{d}u, \tag{4.4}$$

$$\begin{aligned} \ell_{\mathrm{LRL}}(w; \boldsymbol{\beta}) = & \sum_{u \in \mathbf{X} \cap D} w(u) \log \left(\frac{\rho(u; \boldsymbol{\beta})}{\delta(u) + \rho(u; \boldsymbol{\beta})} \right) \\ & - \int_D w(u)\delta(u) \log \left(\frac{\rho(u; \boldsymbol{\beta}) + \delta(u)}{\delta(u)} \right) \mathrm{d}u, \end{aligned} \tag{4.5}$$

where $w(\cdot)$ is a weight function depending on the first and the second-order characterictics of $\mathbf{X}$ and $\delta(\cdot)$ is a nonnegative real-valued function. We recommend the interested

readers to look at the paper by Guan and Shen (2010) for further details on the weight function $w(\cdot)$ and the paper by Baddeley et al. (2014) for the role of function $\delta(\cdot)$. The Poisson estimator (resp. the logistic regression estimator) can be obtained by maximizing (4.4) (resp. (4.5)). Note that these methods cannot perform variable selection. To do so, regularization methods (see e.g. Chapter 2) are introduced by maximizing a penalized version of (4.4)-(4.5)

$$Q(w;\boldsymbol{\beta}) = \ell(w;\boldsymbol{\beta}) - |D|\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{4.6}$$

where $\ell(w;\boldsymbol{\beta})$ is either the Poisson likelihood (4.4) or the logistic regression likelihood (4.5). We refer the second term of (4.6) to as a penalization term. In this term, we have mainly two parts: (1) a penalty function p_λ parameterized by $\lambda \geq 0$ which can change for each component $j, j = 1, \cdots, p$, and (2) the volume of the observation domain $|D|$ which plays the role as the sample size in spatial point process framework. For any nonnegative λ, we say that $p_\lambda(\cdot)$ is a penalty function if p_λ is a nonnegative function with $p_\lambda(0) = 0$. Some examples, described in Table 4.1, include l_2 penalty (Hoerl and Kennard, 1988), l_1 penalty (Tibshirani, 1996; Zou, 2006), elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009), SCAD (Fan and Li, 2001), and MC+ (Zhang, 2010). See, for example, Friedman et al. (2008) for further backgrounds about penalty function, regularization methods, and related materials for more general objectives.

Table 4.1: Examples of penalty function.

Penalty	$p_\lambda(\theta)$
l_2 penalty	$\frac{1}{2}\lambda\theta^2$
l_1 penalty	$\lambda\lvert\theta\rvert$
Elastic net	$\lambda\{\gamma\lvert\theta\rvert + \frac{1}{2}(1-\gamma)\theta^2\}$
SCAD	$\lambda\theta\mathbb{I}(\theta \leq \lambda) + \frac{\gamma\lambda\theta - \frac{1}{2}(\theta^2+\lambda^2)}{\gamma-1}\mathbb{I}(\lambda \leq \theta \leq \gamma\lambda) + \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}\mathbb{I}(\theta \geq \gamma\lambda)$
MC+	$\left(\lambda\theta - \frac{\theta^2}{2\gamma}\right)\mathbb{I}(\theta \leq \gamma\lambda) + \frac{1}{2}\gamma\lambda^2\mathbb{I}(\theta \geq \gamma\lambda)$

4.3 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 of intensity function given by (4.1) for which the dimension of parameter $\boldsymbol{\beta}$, denoted now by p_n, diverges to ∞ as $n \to \infty$. We provide notation and conditions, and discuss the differences with the setting where p is fixed in Section 4.3.1. Our main results are presented in Section 4.3.2. 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 4.3.2).

4.3.1 Notation and conditions

Throughout this section and Sections 4.6.1-4.6.3, let

$$\begin{aligned} \ell_n(w;\boldsymbol{\beta}) =& \ell_{n,\mathrm{PL}}(w;\boldsymbol{\beta}) \\ =& \sum_{u \in \mathbf{X} \cap D_n} w(u) \log \rho(u;\boldsymbol{\beta}) - \int_{D_n} w(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u, \end{aligned} \tag{4.7}$$

$$Q_n(w;\boldsymbol{\beta}) = \ell_n(w;\boldsymbol{\beta}) - |D_n| \sum_{j=1}^{p_n} p_{\lambda_{n,j}}(|\beta_j|), \tag{4.8}$$

be respectively the weighted Poisson likelihood and the penalized likelihood.

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 of true coefficients, where $\boldsymbol{\beta}_{01}$ is the s-dimensional vector of nonzero coefficients and $\boldsymbol{\beta}_{02}$ is the $(p_n - s)$-dimensional vector of zero coefficients. We assume that the number of nonzero 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 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

$$\alpha_{k,l}(q) = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathscr{F}(\Lambda_1), B \in \mathscr{F}(\Lambda_2), \\ \Lambda_1 \in \mathscr{B}(\mathbb{R}^d), \Lambda_2 \in \mathscr{B}(\mathbb{R}^d), |\Lambda_1| \leq k, |\Lambda_2| \leq l, d(\Lambda_1, \Lambda_2) \geq q\}, \tag{4.9}$$

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 a pair correlation function indicating the attraction (or repulsion) among points 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 by $\mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)$ (resp. $\mathbf{B}_{n,11}(w; \boldsymbol{\beta}_0)$, $\mathbf{C}_{n,11}(w; \boldsymbol{\beta}_0)$) the $s \times s$ top-left corner of $\mathbf{A}_n(w; \boldsymbol{\beta}_0)$ (resp. $\mathbf{B}_n(w; \boldsymbol{\beta}_0)$, $\mathbf{C}_n(w; \boldsymbol{\beta}_0)$). 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 on p_n.

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

$$a_n = \max_{j=1,\ldots,s} |p'_{\lambda_{n,j}}(|\beta_{0j}|)|, \tag{4.10}$$

$$b_n = \inf_{j=s+1,\ldots,p_n} \inf_{\substack{|\theta| \leq \epsilon_n \\ \theta \neq 0}} p'_{\lambda_{n,j}}(\theta), \text{ for } \epsilon_n = K_1|D_n|^{-1/2}, \tag{4.11}$$

$$c_n = \max_{j=1,\ldots,s} |p''_{\lambda_{n,j}}(|\beta_{0j}|)|, \tag{4.12}$$

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, where o denotes the origin of $\mathbb{R}^d$:

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

($\mathcal{C}$.2) We assume that the intensity function has the log-linear specification given by (4.1) 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 $\sup_{u \in \mathbb{R}^d} \rho(u; \boldsymbol{\beta}) < \infty$ for all $\boldsymbol{\beta} \in \Xi(\boldsymbol{\beta}_0)$.

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

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

($\mathcal{C}$.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 (4.9), 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) We assume that $\nu_{\min}\Big(|D_n|^{-1}\{\mathbf{B}_{n,11}(w; \boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w; \boldsymbol{\beta}_0)\}\Big) > 0$ for all sufficiently large n, where $\nu_{\min}(\mathbf{M}_n)$ is the smallest eigenvalue of a squared matrix $\mathbf{M}_n$.

($\mathcal{C}$.7) We assume that, for all sufficiently large 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 nonnegative 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 $\sqrt{|D_n|/p_n}\tilde{r}_{n,j} \to \infty$ 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) We assume that $p_n^3/|D_n| \to 0$ as $n \to \infty$.

Conditions ($\mathcal{C}$.1)-($\mathcal{C}$.8) are quite similar to the ones we require in Chapter 2 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 as n is large enough. However, to obtain an estimator which is consistent and has two other properties: sparsity and normality distribution, we need that the number of covariates does not grow too fast with respect to the volume of the observation domain stated by condition ($\mathcal{C}$.9). This condition is similar to that of Fan and Peng (2004) when $|D_n|$ is simply replaced by n (the sample size in their context).

4.3.2 Main results

We state our main results here. Proofs are relegated to Sections 4.6.1-4.6.3.

We first show in Theorem 4.3.1 that the penalized weighted Poisson likelihood estimator converges in probability and exhibits its rate of convergence.

Theorem 4.3.1. *Assume the conditions ($\mathcal{C}$.1)-($\mathcal{C}$.5) and ($\mathcal{C}$.7)-($\mathcal{C}$.9) hold. Let a_n and c_n be given respectively by (4.10) and (4.12). 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, if $a_n = O(|D_n|^{-1/2})$ and $c_n = o(1)$, our estimator is root-$(|D_n|/p_n)$ consistent. Note that the convergence rate of our estimator is the $\sqrt{p_n}$ times the convergence rate of the estimator obtained assuming finite number of parameters we studied in Chapter 2. In addition, when we compare our results to that under regularized likelihood estimation developed 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, in which both $|D_n|$ in current setting and n in their setting play the same role as the sample size.

Furthermore, we demonstrate in Theorem 4.3.2 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 4.3.2. *Assume the conditions (C.1)-(C.9) are satisfied. If* $a_n|D_n|^{1/2} \to 0$, $b_n\sqrt{|D_n|/p_n^2} \to \infty$ *and* $\sqrt{p_n}c_n \to 0$ *as* $n \to \infty$, *the root-*$(|D_n|/p_n)$ *consistent local maximizers* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$ *in Theorem 4.3.1 satisfy:*

(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

$$\begin{aligned}\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0) =& |D_n|\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\boldsymbol{\Pi}_n\}^{-1}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\} \\ & \{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0) + |D_n|\boldsymbol{\Pi}_n\}^{-1}, \qquad (4.13)\\ \boldsymbol{\Pi}_n =& \mathrm{diag}\{p''_{\lambda_{n,1}}(|\beta_{01}|), \ldots, p''_{\lambda_{n,s}}(|\beta_{0s}|)\}. \qquad (4.14)\end{aligned}$$

As a consequence, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ *is the asymptotic covariance matrix of* $\hat{\boldsymbol{\beta}}_1$. *Note that* $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}$ *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* $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\left(\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\right)^\top = \boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$.

Remark 4.3.1. *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.2) and (C.3),* $|D_n|\,\|\boldsymbol{\Pi}_n\|$ *is asymptotically negligible with respect to* $\|\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\|$.

Remark 4.3.2. *Theorems 4.3.1 and 4.3.2 remain true for the regularized logistic regression likelihood estimates 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 (C.3) by adding* $\sup_{u\in\mathbb{R}^d}\delta(u) < \infty$.

We show in Theorem 4.3.2 that the sparsity and the normality distribution are still valid when the number of parameters diverges. By Remark 4.3.1, when n is large enough, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ in (4.13) becomes approximately

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

This means that we have the same efficiency as 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 4.3.2, we require that $a_n|D_n|^{1/2} \to 0$, $b_n\sqrt{|D_n|/p_n^2} \to \infty$ and $\sqrt{p_n}c_n \to 0$ 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 to represent the sample size. However, we require a slightly stronger condition 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 $0 < K_2 < \infty$, $\nu_{\max}\Big(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0)\Big) < K_2$, where $\nu_{\max}(\mathbf{M}_n)$ is the largest eigenvalue of a squared matrix $\mathbf{M}_n$. 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 4.2. For the ridge regularization method, $b_n = 0$, preventing from applying Theorem 4.3.2 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 $a_n|D_n|^{1/2} \to 0$ 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 nonconvex penalties SCAD and MC+, for which λ_n can be adjusted. For the regularization methods we consider in this study, the condition $\sqrt{p_n}c_n \to 0$ is implied by the condition $a_n|D_n|^{1/2} \to 0$ as $n \to \infty$ and condition ($\mathcal{C}$.9).

4.4 Numerical results

This section is devoted to present numerical results. More precisely, we conduct simulation experiments in Section 4.4.1 to assess the finite sample peformance of our estimates and apply our method to an application in ecology in Section 4.4.2. We apply the regularized Poisson likelihood (PL) and the regularized weighted Poisson likelihood (WPL) to select covariates and estimate their coefficients simultaneously. Similar approach can be used easily for the regularized (un)weighted logistic regression (see Chapter 2, Section 2.5).

To numerically evaluate the parameters estimates, we apply Berman-Turner method (Berman and Turner, 1992) combined with coordinate descent algorithm (Friedman

Table 4.2: Details of the sequences a_n, b_n and c_n for a given regularization method.

Method	a_n	b_n	c_n
Ridge	$\lambda_n \max_{j=1,\ldots s}\{\lvert\beta_{0j}\rvert\}$	0	λ_n
Lasso	λ_n	λ_n	0
Enet	$\lambda_n\Big[(1-\gamma)\max_{j=1,\ldots s}\{\lvert\beta_{0j}\rvert\}+\gamma\Big]$	$\gamma\lambda_n$	$(1-\gamma)\lambda_n$
AL	$\max_{j=1,\ldots s}\{\lambda_{n,j}\}$	$\min_{j=s+1,\ldots p}\{\lambda_{n,j}\}$	0
Aenet	$\max_{j=1,\ldots s}\{\lambda_{n,j}\Big((1-\gamma)\lvert\beta_{0j}\rvert+\gamma\Big)\}$	$\gamma\min_{j=s+1,\ldots p}\{\lambda_{n,j}\}$	$(1-\gamma)\max_{j=1,\ldots,s}\{\lambda_{n,j}\}$
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 $\sqrt{|D_n|/p_n^2}\lambda_n \to \infty$ for n sufficient large

et al., 2007) to perform variable selection and parameter estimation. Berman-Turner device allows to show that maximizing (4.4) is equivalent to fitting a weighted Poisson generalized linear model, so the standard software for generalized linear model can be used. This in fact has been exploited in the `spatstat` `R` package (Baddeley et al., 2015). Coordinate descent algorithm, which has been implemented in the `glmnet` (Friedman et al., 2010) for some convex penalties and in the `ncvreg` (Breheny and Huang, 2011) for some nonconvex penalties, is used to compute the regularization paths solutions. More details for computational strategies have been discussed in detail in Chapter 2.

Our methods rely on the tuning parameter λ. Some previous studies suggested to use BIC-type method to select the tuning parameter in order to obtain selection consistent estimator (see e.g. Zou et al., 2007; Wang et al., 2007b, 2009). In this study, we select λ which minimizes WQBIC(λ), a weighted 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 nonzero

regression coefficients and $|D|$ is the volume of observation domain. To implement the adaptive methods (i.e. adaptive lasso and adaptive elastic net), we 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 Chapter 2, 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).

4.4.1 Simulation study

In this section, we aim to observe our estimates behaviour in different situations when a large number of covariates for fitting spatial point process intensity estimation is involed. We intend to extend the setting considered in Chapter 2. We start with relatively complex situation where strong multicollinearity is present (Scenarios 1a and 2a) and we then consider 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 set to be different in each setting.

We set the spatial domain to be $D = [0, 1000] \times [0, 500]$ and set the mean number of points over D to be 1600. The true intensity function is set to be $\rho(u; \boldsymbol{\beta}_0) = \exp(\mathbf{z}(u)^\top \boldsymbol{\beta}_0)$, where $\mathbf{z}(u) = \{1, z_1(u), \ldots, z_{50}(u)\}^\top$ and $\boldsymbol{\beta}_0 = \{\beta_0, \beta_{01}, \cdots, \beta_{050}\}$. Here, we do not estimate β_0 since it is chosen such that each realization has 1600 points in average. 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` (R Core Team, 2016), 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 par-

ticular, we define $\mathbf{z}(u) = \mathbf{V}^\top \mathbf{x}(u)$, where $\mathbf{x}(u) = \{x_1(u), \ldots, x_{50}(u)\}^\top$. More precisely, $\mathbf{V}$ is such that $\mathbf{\Omega} = \mathbf{V}^\top \mathbf{V}$, 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.

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 extra 45 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 $\mathbf{z}(u) = \mathbf{V}^\top \mathbf{x}(u)$, where $\mathbf{V}$ is such that $\mathbf{\Omega} = \mathbf{V}^\top \mathbf{V}$, 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.

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 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 present in Table 4.3 the selection properties of our estimates. We consider the true positive rate (TPR), the false positive rate (FPR), and the positive predictive value (PPV) to evaluate the selection properties of the estimates. 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, and MC+) and which cannot satisfy (lasso, elastic net) our theorems is that the methods which cannot satisfy our theorems tend to overselect covariates, leading to suffer 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, adaptive lasso is slightly better. The difference in these results compared with the ones obtained in Chapter 2 is that adaptive elastic net appears to be another alternative (appart from adaptive lasso) to perform variable selection for

Table 4.3: 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 4.4: 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
Aenet	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

the setting involving large number of covariates. By combining l_1 and l_2 penalties, the adaptive elastic net becomes more appropriate in the situation where the number of covariates is large with the potential existance of strong multicollinearity. This may explain why adaptive elastic net perform better in our setting than that we considered in Chapter 2.

We do not consider in this study the extra covariates generated as standard Gaussian white noise independently as we considered in Chapter 2 as in Scenario 1. We are still able to show in the current chapter 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. Furthermore, we cannot see clearly which one in which particular cases we could recommend between the regularized PL and the regularized WPL or vice versa. This is quite different from what we suggested in Chapter 2 that we would recommend the use of regularized WPL with adaptive lasso for a very structured and clustered case while we prefer with penalized PL with adaptive lasso for the other cases. In the experiments we conduct in this chapter, 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 4.3, we would recommend in general to combine the regularized WPL with adaptive lasso to perform variable selection.

Table 4.4 gives the prediction properties of the estimates in terms of biases, standard deviations (SD), and square root of mean squared errors (RMSE), some criterions 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). This is different from the results we obtained in Chapter 2 recommending to use adaptive lasso as the best method. We find here that adaptive elastic net appears to be more competitive than adaptive lasso especially in the complex situation: large number of covariates, strong multicollinearity, 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 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 seems more favorable. From Table 4.4, when the focus is on prediction, we would recommend to combine the regularized WPL with adaptive elastic net when we deal with a clustered spatial point process which have covariates with complex spatial structure while we would recommend the regularized PL combined with adaptive elastic net for a clustered process with no complex spatial structure. However, if we are faced with a less clustered process, the regularized PL combined with adaptive lasso is slightly more preferable for the case with no complex spatial structure in the covariates while the regularized WPL combined with adaptive lasso is slightly more recommended for the case with complex structure in the covariates. Note that the adaptive lasso is slightly better than the adaptive elastic net for a few cases. Thus, as general advice, we would recommend to use adaptive elastic net (instead of adaptive lasso) if the focus is for prediction.

Note that the combination between the regularized WPL and adaptive lasso is more preferable if the focus is on variable selection while adaptive elastic net is more favorable if the focus is for prediction. To have a more general recommendation, we would recommend to apply adaptive elastic net when we are faced with complex situation: large number of covariates, strong multicollinearity, and complex spatial structure.

By combining l_1 and l_2 penalties, adaptive elastic net seems to balance the selection and the 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.

4.4.2 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 4.1. In particular, we are interested to study the spatial distribution of 3,604 locations of *Beilschmiedia pendula Lauraceae* (BPL) trees by estimating its intensity. There have been 93 available covariates which can be considered including 2 topological attributes, 13 soil properties, and 78 interactions between two soil nutrients. We model the intensity of BPL tree as a log-linear function of these 93 covariates. 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 lasso, adaptive lasso, and SCAD. Note that we center and scale all the covariates to observe which covariates owing relatively large effect on the intensity.

Table 4.5: Number of selected and non-selected covariates among 93 covariates by regularized Poisson likelihood with lasso, adaptive lasso, and SCAD regularization.

Method	Regularized PL		Regularized WPL	
	#Selected	#No	#Selected	#No
LASSO	77	16	45	48
AL	50	43	10	83
SCAD	58	35	3	90

We present in Table 4.5 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 less covariates are selected by regularized WPL.

Table 4.6: 10 common covariates selected

Covariates	Regularized PL			Regularized WPL		
	LASSO	AL	SCAD	LASSO	AL	SCAD
Elev	0.32	0.40	0.33	0.40	0.32	0
Slope	0.39	0.40	0.36	0.42	0.44	0
Cu	0.56	0.31	0.61	0.39	0.33	0
Mn	0.14	0.14	0.09	0.15	0.22	0
P	-0.48	-0.43	-0.54	-0.33	-0.57	-1.07
Zn	-0.75	-0.66	-0.83	-0.58	-0.40	0
Al:P	-0.30	-0.29	-0.31	-0.28	-0.16	0
Mg:P	0.62	0.29	0.45	0.48	0.42	0
Zn:N	0.21	0.35	0.30	0.01	0	0.62
N.Min:pH	0.44	0.44	0.49	0.25	0.27	0

The regularized PL seems to overfit model by overselecting less informative covariates.

Regarding lasso method, 77 covariates are selected by regularized PL method while 45 covariates are selected by regularized WPL. Lasso tends to keep less important covariates in the model even if we consider weighted method. This may explain why lasso cannot satisfy our Theorem 4.3.2. Different from lasso, adaptive lasso and SCAD, which satisfy our Theorem 4.3.2, seem to perform better by keeping less informative covariates out from the model. However, regularized WPL combined with SCAD seems to underfit model by removing some potentially important covariates. Regularized WPL with adaptive lasso seems to outperform the other methods.

Table 4.6 gives the information regarding 10 covariates commonly selected among six combination methods. Although the magnitudes of the estimates can be slightly different, the signs all agree with each other. Some covariates suspected to have relatively high influence to the intensity of BPL include: elevation, slope, concentration of Copper, Phosphorus and Zinc in the soil as well as the interaction between Magnesium and Phosphorus. SCAD may loose five (out of six) potentially important covariates by

removing them from the model.

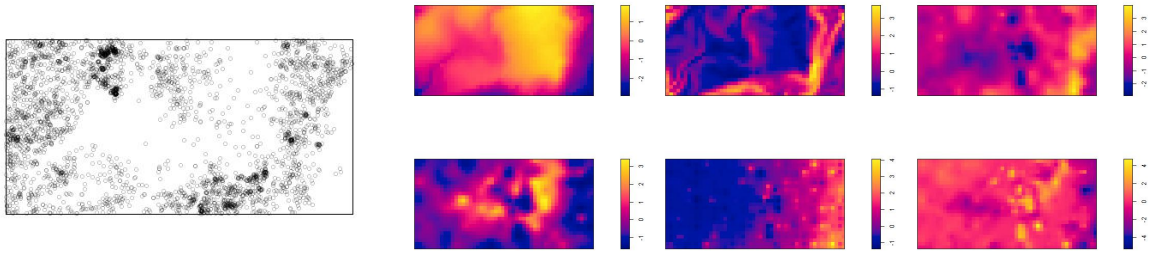

Figure 4.1: Maps of 3,604 locations of BPL trees and the six covariates suspected to have high influence on the intensity of BPL trees, row 1: elevation, slope, and Copper, row 2: Phosphorus Zinc and the interaction between Magnesium and Phosphorus.

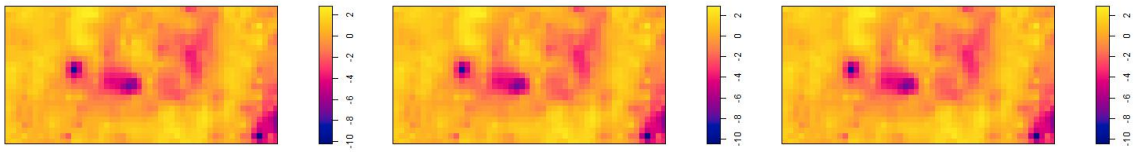

Figure 4.2: Estimates of BPL tree intensity (log scale) for each method: row 1: regularized PL, row 2: regularized WPL, column 1 (resp. 2 and 3): lasso (resp. adaptive lasso and SCAD).

These results suggest that BPL trees favor to live in the areas of higher elevation and slope with high concentration of Copper 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 Magnesium and Phosphorus gives positive association with the appearance of BPL trees. The maps of 3,604 locations of BPL trees as well as six most influencing covariates are depicted in Figure 4.1. We also present the estimate of the intensity (log scale) for each of the six methods in Figure 4.2.

4.5 Conclusion and discussion

We consider feature selection procedures for spatial point processes intensity estimation by regularizing the estimating functions 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 regularity conditions, we prove that the estimates obtained from such setting satisfy consistency, sparsity, and normality distribution. Our results are available for large classes of spatial point processes and for many penalty functions.

We conduct some 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 the focus is for prediction, the regularized WPL combined with adaptive elastic net is more preferable for a clustered process considering large number of covariates with the present of strong multicollinearity and complex spatial structure. Otherwise, we would recommend to combine the regularized PL with adaptive elastic net. For more general advice, we would recommend to use the adaptive elastic net than the adaptive lasso since adaptive elastic net is able to balance the selection and the prediction properties by combining the l_1 and the l_2 penalties.

As proposed in Chapter 2, combination between `spatstat R` package with two `R` packages `glmnet` and `ncvreg` can work quite fast even when we consider a significantly larger number of covariates. 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., 2007b, 2009). We have numerical evidence from simulation studies that this criterion can satisfy the selection consistency when regularization methods satisfying our theorems are considered. Such a criterion is also used under spatial point process setting by Thurman et al. (2015) for practical implementation. However, theoretical justification may be needed under spatial point process setting to support our theoretical results. We leave this direction for further 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

lasso seems to outperform the other methods. Among 93 covariates, we find six spatial covariates which may have high influence to the appearance of BPL trees, including two topological attributes: elevation and slope and four soil nutrients: Copper, Phosphorus, Zinc and the interaction between Magnesium and Phosphorus.

A further work would consider to include other 296 species of trees, which were surveyed in the same observation region as BPL was observed on, 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 the linear models (and generalized linear models) results in a linear programming. Thus, more competitive algorithms is available. It would be interesting to bring this approach to spatial point process setting and investigate both its theoretical properties and computational implementation.

4.6 Proofs of the main results

4.6.1 Auxiliary Lemma

The following Lemma is used in the proof of Theorem 4.3.1 and Lemma 4.6.2 (which includes Lemma 4.6.3 and Theorem 4.3.2). 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 4.6.1. *Under conditions (C.1)-(C.5), the following result holds as* $n \to \infty$

$$\ell_n^{(1)}(w;\boldsymbol{\beta}_0) = O_{\mathrm{P}}\left(\sqrt{p_n|D_n|}\right). \tag{4.15}$$

Proof. Using Campbell Theorems (4.2)-(4.3), 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 (C.4)-(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 $\mathrm{Var}[Y_n]$, $Y_n = O_{\mathrm{P}}(\sqrt{\mathrm{Var}[Y_n]})$. □

4.6.2 Proof of Theorem 4.3.1

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 (4.8)) 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 4.3.1, we aim at proving that for any given $\epsilon > 0$, there exists sufficiently large $K > 0$ such that for n sufficiently large

$$\mathrm{P}\left(\sup_{\|\mathbf{k}\|=K} \Delta_n(\mathbf{k}) > 0\right) \leq \epsilon, \qquad \text{where } \Delta_n(\mathbf{k}) = Q_n(w;\boldsymbol{\beta}_0 + d_n\mathbf{k}) - Q_n(w;\boldsymbol{\beta}_0). \tag{4.16}$$

Equation (4.16) 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 nonnegative 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, by condition ($\mathcal{C}$.7),

$$\check{\nu} := \liminf_{n\to\infty} \nu_{\min}(|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0)) = \liminf_{n\to\infty} \frac{\mathbf{k}^\top (|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0))\,\mathbf{k}}{\|\mathbf{k}\|^2} > 0.$$

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 nonnegative 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

$$\begin{aligned}
T_{2,1}' &= d_n|D_n|\sum_{j=1}^{s} k_j p'_{\lambda_{n,j}}(|\beta_{0j}|)\,\mathrm{sign}(\beta_{0,j}) \leq \sqrt{s}a_n d_n|D_n|\,\|\mathbf{k}\| \leq d_n^2|D_n|\,\|\mathbf{k}\|,\\
T_{2,2}' &= \frac{1}{2}d_n^2|D_n|\sum_{j=1}^{s} k_j^2 p''_{\lambda_{n,j}}(|\beta_{0j}|) \leq c_n d_n^2|D_n|\|\mathbf{k}\|^2,\\
T_{2,3}' &= \frac{1}{6}d_n^3|D_n|\sum_{j=1}^{s} k_j^3 p'''_{\lambda_{n,j}}(|\beta_{0j}+t_j d_n k_j|) \leq \kappa d_n^3|D_n|.
\end{aligned}$$

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

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

and then

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

We now return to (4.16): for n sufficiently large

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

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 (4.15) in Lemma 4.6.1.

□

4.6.3 Proof of Theorem 4.3.2

Before proving Theorem 4.3.2, we present Lemmas 4.6.2-4.6.3. Lemma 4.6.2 is used to prove Theorem 4.3.2(i) while Lemma 4.6.3 is used to derive Theorem 4.3.2(ii).

Lemma 4.6.2. *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 $K_1>0$,*

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

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 $j=s+1,\ldots,p_n$

$$\frac{\partial Q_n(w;\boldsymbol{\beta})}{\partial\beta_j}<0\quad\text{for } 0<\beta_j<\varepsilon_n,\ \text{and}\tag{4.17}$$

$$\frac{\partial Q_n(w;\boldsymbol{\beta})}{\partial\beta_j}>0\quad\text{for } -\varepsilon_n<\beta_j<0.\tag{4.18}$$

From (4.7),

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

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 4.6.1, 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 $j = s+1, \ldots, p_n$

$$\frac{\partial \ell_n(w;\boldsymbol{\beta})}{\partial \beta_j} = O_{\mathrm{P}}(\sqrt{|D_n| p_n^2}). \tag{4.19}$$

Now, we want to prove (4.17). Let $0 < \beta_j < \varepsilon_n$ and b_n be the sequence given by (4.11). By condition ($\mathcal{C}$.8), b_n is well-defined and since by 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 (4.17) is therefore deduced from (4.19) and from assumption that $b_n\sqrt{|D_n|/p_n^2} \to \infty$ as $n \to \infty$. We proceed similarly to prove (4.18). □

Lemma 4.6.3. *Under the conditions ($\mathcal{C}$.1)-($\mathcal{C}$.8) and the conditions required in Lemma 4.6.2, the following convergence holds in distribution as $n \to \infty$*

$$\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_{01}) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_{01})\}^{-1/2} \ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_{01}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_s), \tag{4.20}$$

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 4.6.2 and by using Campbell Theorems (4.2)-(4.3),

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

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

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

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 $(\mathcal{C}.4)$,

$$\sup_{n \geq 1} \sup_{i \in \mathscr{I}_n} \mathbb{E}(\|Y_{i,n}\|^{2+\delta}) < \infty. \tag{4.21}$$

If we combine (4.21) 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 4.3.2. Since Theorem 4.3.2(i) is proved by Lemma 4.6.2, we only need to prove Theorem 4.3.2(ii), which is the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$. As shown in Theorem 4.3.1, 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 4.3.1 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 $j = 1, \cdots, s$

$$\begin{aligned}
0 =& \frac{\partial \ell_n(w;\hat{\boldsymbol{\beta}})}{\partial \beta_j} - |D_n| p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\,\mathrm{sign}(\hat{\beta}_j) \\
=& \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \frac{\partial^2 \ell_n(w;\tilde{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) - |D_n| p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\,\mathrm{sign}(\hat{\beta}_j) \\
=& \frac{\partial \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \frac{\partial^2 \ell_n(w;\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l - \beta_{0l}) + \sum_{l=1}^{s} \Psi_{n,jl}(\hat{\beta}_l - \beta_{0l}) \\
& - |D_n| p'_{\lambda_{n,j}}(|\beta_{0j}|)\,\mathrm{sign}(\beta_{0j}) - |D_n|\phi_{n,j},
\end{aligned} \tag{4.22}$$

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

$$\begin{aligned}
\phi_{n,j} &= p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\,\mathrm{sign}(\hat{\beta}_j) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\,\mathrm{sign}(\beta_{0j}) \\
&= \Big(p'_{\lambda_{n,j}}(|\hat{\beta}_j|) - p'_{\lambda_{n,j}}(|\beta_{0j}|)\Big)\,\mathrm{sign}(\beta_{0j}) + p'_{\lambda_{n,j}}(|\hat{\beta}_j|)\Big(\mathrm{sign}(\hat{\beta}_j) - \mathrm{sign}(\beta_{0j})\Big) \\
&\leq \kappa\Big||\hat{\beta}_j| - |\beta_{0j}|\Big| + 2a_n \\
&\leq \kappa|\hat{\beta}_j - \beta_{0j}| + 2a_n. \qquad (4.23)
\end{aligned}$$

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

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

where the latter equation ensues from Theorem 4.3.1 and condition ($\mathcal{C}$.8). Again, from Theorem 4.3.1, $\mathbb{I}(|\hat{\beta}_j - \beta_{0j}| \leq \tilde{r}_{n,j}) \xrightarrow{L^1} 1$ 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 (4.23)

$$T_2 \leq \{\kappa|\hat{\beta}_j - \beta_{0j}| + 2a_n\}\ \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}).$$

Now, we want to prove that $T_2 = o_{\mathrm{P}}(|D_n|^{-1/2})$. Let $S_n = |\hat{\beta}_j - \beta_{0j}|\ \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j})$ and $T_n = \mathbb{I}(S_n > |D_n|^{-1/2})$. If $\mathbb{E}T_n \xrightarrow{\mathrm{P}} 0$ then $S_n = o_{\mathrm{P}}(|D_n|^{-1/2})$ which implies that, by combining with the condition on a_n, $T_2 = o_{\mathrm{P}}(|D_n|^{-1/2})$. 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 $T_n = \mathbb{I}(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j})$. Hence,

$$\mathbb{E}T_n = \mathrm{P}\left(|\hat{\beta}_j - \beta_{0j}| > \tilde{r}_{n,j}\right) = \mathrm{P}\left(|\hat{\beta}_j - \beta_{0j}| > \frac{\sqrt{p_n}}{\sqrt{|D_n|}}\frac{\tilde{r}_{n,j}\sqrt{|D_n|}}{\sqrt{p_n}}\right),$$

which implies the result since $|\hat{\beta}_j - \beta_{0j}| = O_{\mathrm{P}}(\sqrt{p_n/|D_n|})$ from Theorem 4.3.1 and since $\sqrt{|D_n|/p_n}\tilde{r}_{n,j}$ given by condition ($\mathcal{C}$.8). We now deduce that

$$\phi_{n,j} = p''_{\lambda_{n,j}}(|\beta_{0j}|)(\hat{\beta}_j - \beta_{0j})\Big(1 + o_{\mathrm{P}}(1)\Big) + O_{\mathrm{P}}(p_n/|D_n|) + o_{\mathrm{P}}(|D_n|^{-1/2}). \tag{4.24}$$

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

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

We rewrite both sides of (4.22) as

$$\ell^{(1)}_{n,1}(w;\boldsymbol{\beta}_0) + \ell^{(2)}_{n,1}(w;\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) + \boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) - |D_n|\mathbf{p}'_n - |D_n|\boldsymbol{\phi}_n = 0. \tag{4.25}$$

By definition of $\boldsymbol{\Pi}_n$ given by (4.14) and from (4.24), we obtain $\boldsymbol{\phi}_n = \boldsymbol{\Pi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\Big(1 + o_{\mathrm{P}}(1)\Big) + O_{\mathrm{P}}(p_n/|D_n|) + o_{\mathrm{P}}(|D_n|^{-1/2})$. Using this, we deduce, by premultiplying both sides of (4.25) by $\mathbf{M}_n$, that

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

Now, $\|\mathbf{M}_n\| = O(1/\sqrt{|D_n|})$ by condition ($\mathcal{C}$.6), $\|\boldsymbol{\Psi}_n\| = O_{\mathrm{P}}(\sqrt{p_n|D_n|})$ by conditions ($\mathcal{C}$.2)-($\mathcal{C}$.3) and by Theorem 4.3.1, and $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\| = O_{\mathrm{P}}(\sqrt{p_n/|D_n|})$ by Theorem 4.3.1 and by Theorem 4.3.2(i). Finally, since by assumptions $a_n|D_n|^{1/2} \to 0$ and $c_n\sqrt{p_n} \to 0$ as $n \to \infty$, we deduce that

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

$$\|\mathbf{M}_n \boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})\| = O_{\mathrm{P}}\left(\sqrt{\frac{p_n^2}{|D_n|}}\right) = o_{\mathrm{P}}(1).$$

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

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

By (5.41) in Lemma 4.6.3 and by Slutsky's Theorem, we deduce that

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

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

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

where $\boldsymbol{\Sigma}_n(w, \boldsymbol{\beta}_0)$ is given by (4.13). □

CHAPTER 5

The adaptive linearized Dantzig selector for spatial point processes intensity estimation

5.1 Introduction

Several recent applications involve the observation of spatial point pattern data together with spatial covariates (see e.g. Waagepetersen, 2007; Møller and Waagepetersen, 2007; Thurman et al., 2015; Renner et al., 2015). 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 modelling of the spatial distribution of trees species in a forest related to some environmental factors (e.g. Waagepetersen, 2007; Renner and Warton, 2013). We focus in this chapter on the log-linear model for the intensity function defined by

$$\rho(u;\boldsymbol{\beta}) = \exp(\mathbf{z}(u)^\top \boldsymbol{\beta}), u \in D \subset \mathbb{R}^d, \tag{5.1}$$

where $\mathbf{z}(u) = \{z_1(u), \ldots, z_p(u)\}^\top$ are the p spatial covariates measured at location u and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^\top$ is a real p-dimensional parameter. Parametric approaches to estimate $\boldsymbol{\beta}$ include likelihood-based methods (e.g. Berman and Turner, 1992; Rathbun and Cressie, 1994; Møller and Waagepetersen, 2004), estimating equation-based methods (e.g. Waagepetersen, 2007, 2008; Guan and Shen, 2010; Baddeley et al., 2014; Guan et al., 2015), and variational approaches (e.g. Baddeley and Dereudre, 2013; Coeurjolly and Møller, 2014). All these methods are not appropriate when the number of covariates is large. To tackle this problem, regularized versions of likelihood-based

and estimating equation-based methods have been recently proposed. 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 fuction (see Renner and Warton, 2013; Thurman and Zhu, 2014) such as l_1 penalty. For more general point process models, instead of using the likelihood of the processes which often requires computational intensive MCMC methods (Møller and Waagepetersen, 2004), penalized versions of estimating functions based on Campbell theorem derived both from Poisson and logistic regression likelihoods have been developped (see Chapters 2 and 3). Some examples of penalty functions are ℓ_2 penalty (e.g. Hoerl and Kennard, 1988), ℓ_1 penalty (e.g. Tibshirani, 1996; Zou, 2006), elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009), SCAD (Fan and Li, 2001), and MC+ (Zhang, 2010).

Apart from regularization techniques, the Dantzig selector (Candes and Tao, 2007) has appeared to enrich the class of variable selection methods for linear regression models. Unlike most other variable selection methods such as lasso, SCAD, and MC+, which minimize the sum of squared errors subject to a penalty function, the Dantzig selector minimizes the l_1 norm of the coefficients subject to a constraint on the error terms. Similarly to the lasso method or regularization method involving the SCAD or MC+ penalty functions, the Dantzig selector is able to perform variable selection and parameter estimation simultaneously. Nevertheless, unlike the other methods, the standard linear programming can be used to compute the solution to the Dantzig selector optimization problem, providing a computationally efficient algorithm. Candes and Tao (2007) also provided sharp non-asymptotic bounds on the l_2 norm of estimated coefficients error and showed that the error is within a factor of $\log p$ of the error that would be achieved if the locations of the non-zero coefficients were known. As $\log p$ grows very slowly, the Dantzig selector only pays a small price for adaptively choosing the significant variables and is then very suitable for a very large dataset. Some extended studies have been conducted. For example, James and Radchenko (2009) studied the computational implementation of the Dantzig selector for generalized linear models. Antoniadis et al. (2010) extended the theoretical results and the implementa-

tion of the Dantzig selector for the class of Cox's proportional hazards model. Dicker (2010) provided a large sample study of the Dantzig selector and proposed its adaptive version to establish interesting asymptotic properties of the estimates. Finally, Li et al. (2014) developed the Dantzig selector for censored linear regression models and evaluated its asymptotic properties. To our knowledge, there have not been conducted a study which develops the Dantzig selector-type approaches for spatial point processes intensity estimation.

This chapter considers the Dantzig selector-type methods based on estimating equations to obtain intensity estimates for spatial point processes. In particular, we propose a modified version of the Dantzig selector based on linear approximation in the constraint vector which we call by the adaptive linearized Dantzig selector. Although our proposed methods may work for a very general estimating function, we focus on estimating functions derived from Poisson and logistic regression likelihoods, which, as detailed in Chapter 2, have strong links with Poisson and logistic regressions. We study the asymptotic properties of our estimates and prove, under some conditions, that they satisfy sparsity and asymptotic normality. In addition, we find a closed form expression of our estimates. We also prove that the complex optimization problem can be reduced to a linear programming problem, which potentially produces a computationally efficient algorithm. In this study, we do not investigate finite sample properties like oracle inequalities obtained by Candes and Tao (2007); Bühlmann and Van De Geer (2011), for example, for simpler models and in the independent case. Even if it seems feasible to derive finite sample size properties for Poisson point process using for instance concentration inequalities obtained by Reynaud-Bouret (2003), it is not straightforwardly applicable for more general spatial point processes due to the lack of such concentration inequalities for general spatial point processes. By focusing on asymptotic properties, we are able to make our results available for very large classes of spatial point processes which exhibit strong dependence (i.e. very clustered or repulsive point processes).

The rest of this chapter is structured as follows. Section 5.2 presents our general setting. Section 5.3 details methodology, establishes its asymptotic properties, discusses the results we obtain and the conditions we impose, and presents computational

strategy. Section 5.4 reports numerical results from a simulation study. Finally, Section 5.5 provides conclusion and discussion. Proofs of the main results as well as the derivation of the dual to the Dantzig selector optimization problem are deferred to Sections 5.6.1-5.6.5.

5.2 Preliminaries

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. 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 $k : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$

$$\mathbb{E} \sum_{u \in \mathbf{X}} k(u) = \int k(u)\rho(u)\mathrm{d}u \tag{5.2}$$

$$\mathbb{E} \sum_{u,v \in \mathbf{X}}^{\neq} k(u, v) = \int \int k(u, v)\rho^{(2)}(u, v)\mathrm{d}u\mathrm{d}v. \tag{5.3}$$

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 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 material on spatial point processes, see for example Møller and Waagepetersen (2004); Illian et al. (2008).

Given a $p \times p$ matrix $\mathbf{M}$ and a $p \times 1$ vector $\mathbf{V}$, we denote the transposition and the complement of matrix $\mathbf{M}$ (resp. vector $\mathbf{V}$) by $\mathbf{M}^\top$ and $\mathbf{M}^c$ (resp. by $\mathbf{V}^\top$ and $\mathbf{V}^c$). Let $\|\mathbf{V}\|_r = (\sum_{j=1}^p |V_j|^r)^{1/r}$ for $0 < r < \infty$, $\|\mathbf{V}\|_0 = \#\{j, V_j \neq 0\}$ and $\|\mathbf{V}\|_\infty = \max_{1 \leq j \leq p} |V_j|$. Let $\|\mathbf{M}\| = \|\mathbf{M}\|_2 = (\sum_{i=1}^p \sum_{j=1}^p M_{ij}^2)^{1/2}$.

Finally, 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

$$\begin{aligned}\alpha_{k,l}(q) = \sup\{|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)| : A \in \mathscr{F}(\Lambda_1), B \in \mathscr{F}(\Lambda_2),\\ \Lambda_1 \in \mathscr{B}(\mathbb{R}^d), \Lambda_2 \in \mathscr{B}(\mathbb{R}^d), |\Lambda_1| \leq k, |\Lambda_2| \leq l, d(\Lambda_1, \Lambda_2) \geq q\},\end{aligned} \tag{5.4}$$

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 $\mathbb{R}^d$.

5.3 The adaptive linearized Dantzig selector for spatial point processes

In this section, we present the Dantzig selector-type methods, based on linear approximation in the constraint vector, to estimate the intensity of spatial point processes. More precisely, we describe the methodology in Section 5.3.1 and present the asymptotic properties in Section 5.3.2. Computational aspects are discussed in Section 5.3.3.

5.3.1 Methodology

Consider a log-linear form of the intensity $\rho(\cdot;\boldsymbol{\beta})$ given by (5.1). One of the traditional ways to obtain the estimates is by maximizing the likelihood function $\ell(\boldsymbol{\beta})$ or by solving $\mathbf{U}(\boldsymbol{\beta})=0$, where $\mathbf{U}(\boldsymbol{\beta})$ is an unbiased estimating function. In this study, we consider estimating equations derived from Poisson (Waagepetersen, 2007; Guan and Shen, 2010) and logistic regression (Baddeley et al., 2014) likelihoods defined respectively by

$$\mathbf{U}_{\mathrm{PL}}(w;\boldsymbol{\beta}) = \sum_{u\in\mathbf{X}\cap D} w(u)\mathbf{z}(u) - \int_D w(u)\mathbf{z}(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u, \tag{5.5}$$

$$\mathbf{U}_{\mathrm{LRL}}(w;\boldsymbol{\beta}) = \sum_{u\in\mathbf{X}\cap D} \frac{w(u)\mathbf{z}(u)\delta(u)}{\delta(u)+\rho(u;\boldsymbol{\beta})} - \int_D \frac{w(u)\mathbf{z}(u)\rho(u;\boldsymbol{\beta})\delta(u)}{\delta(u)+\rho(u;\boldsymbol{\beta})}\mathrm{d}u. \tag{5.6}$$

Their corresponding likelihoods are

$$\ell_{\mathrm{PL}}(w;\boldsymbol{\beta}) = \sum_{u\in\mathbf{X}\cap D} w(u)\log\rho(u;\boldsymbol{\beta}) - \int_D w(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u, \tag{5.7}$$

$$\begin{aligned}\ell_{\mathrm{LRL}}(w;\boldsymbol{\beta}) = &\sum_{u\in\mathbf{X}\cap D} w(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})}{\delta(u)+\rho(u;\boldsymbol{\beta})}\right)\\ &- \int_D w(u)\delta(u)\log\left(\frac{\rho(u;\boldsymbol{\beta})+\delta(u)}{\delta(u)}\right)\mathrm{d}u,\end{aligned} \tag{5.8}$$

where $w(\cdot)$ is a weight function depending on the first and the second-order characterictics of $\mathbf{X}$ and $\delta(\cdot)$ is a nonnegative real-valued function. We refer the reader 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 $\delta(\cdot)$. Note that, as explained in Chapter 2 (Section 2.3), (5.5) and (5.6) are unbiased estimating equations.

It is worth emphasizing that maximizing (5.7)-(5.8) cannot perform variable selection. To do so, regularization methods (see Chapter 2) are usually introduced by maximizing a penalized version of (5.7)-(5.8)

$$Q(w;\boldsymbol{\beta}) = \ell(w;\boldsymbol{\beta}) - |D|\sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{5.9}$$

where $\ell(w;\boldsymbol{\beta})$ is either the Poisson likelihood (5.7) or the logistic regression likelihood (5.8), p_λ is a penalty function parameterized by $\lambda \geq 0$ which can change for each component $j, j = 1, \cdots, p$, and $|D|$ the volume of the observation domain which plays the role as the sample size in the spatial point process framework. We call the parameter λ by tuning parameter. Note that if p_λ is a l_1 penalty, it corresponds to lasso regularization method (Tibshirani, 1996; Zou, 2006), and (5.9) becomes equivalent to

$$\min \sum_{j=1}^{p} \tilde{\lambda}_j|\beta_j| \text{ subject to } |\ell(w;\boldsymbol{\beta})| \leq \kappa, \tag{5.10}$$

where $\tilde{\lambda}_j = |D|\lambda_j, j = 1, 2, \ldots, p$, are the modified versions of λ_j given by (5.9) and κ is a positive constant.

The Dantzig selector has recently been proposed by Candes and Tao (2007) as an alternative to lasso-type techniques. In this chapter, we consider the adaptive version of the Dantzig selector for spatial point processes intensity estimation. The estimator is defined as the solution to

$$\min \sum_{j=1}^{p} \tilde{\lambda}_j|\beta_j| \text{ subject to } |\mathbf{U}_j(w;\boldsymbol{\beta})| \leq \tilde{\lambda}_j \text{ for } j = 1, \ldots, p, \tag{5.11}$$

where and $\mathbf{U}_j(w;\boldsymbol{\beta})$ is the jth element of either (5.5) or (5.6). When the specific need is unnecessary, we denote the estimating function by $\mathbf{U}(w;\boldsymbol{\beta})$ to simplify notation. Of course we could have simply used the notation λ_j in (5.11) instead of $\tilde{\lambda}_j$. We decided to make a difference between λ_j and $\tilde{\lambda}_j = \lambda_j|D|$ so that we can further compare the assumptions we impose on $\tilde{\lambda}_j$ for the ALDS and the assumptions we made on λ_j when we focused on the adaptive lasso method.

Now, it looks more clearly that (5.11) is quite similar to (5.10). With the same objective function, the adaptive Dantzig selector uses the score vector as a constraint, while the adaptive lasso uses the likelihood function as a constraint. Similarities between lasso and Dantzig selector in linear regression contexts have been studied by Meinshausen et al. (2007); Bickel et al. (2009); James et al. (2009); Asif and Romberg (2010).

It is to be noticed that (5.11) can be rewritten in a matrix form as

$$\min \|\boldsymbol{\Lambda}\boldsymbol{\beta}\|_1 \text{ subject to } \left\|\boldsymbol{\Lambda}^{-1}\mathbf{U}(w;\boldsymbol{\beta})\right\|_\infty \leq 1, \tag{5.12}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_p\}$. One possible challenge of (5.12) is that it may be a non-convex optimization problem due to the possible nonconvexity of the constraint vector. To ease some theoretical difficulties, similarly to Dicker (2010, Chapter 3), we propose a convex version of (5.12) based on a linear approximation to the constraint vector. We call this approach by the adaptive linearized Dantzig selector (ALDS) and we define the criterion by

$$\min \|\boldsymbol{\Lambda}\boldsymbol{\beta}\|_1 \text{ subject to } \left\|\boldsymbol{\Lambda}^{-1}\Big\{\mathbf{U}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\Big\}\right\|_\infty \leq 1, \tag{5.13}$$

where $\tilde{\boldsymbol{\beta}}$ is an initial estimator and $-\mathbf{A}(w;\tilde{\boldsymbol{\beta}})$ is the derivative of $\mathbf{U}(w;\tilde{\boldsymbol{\beta}})$ evaluated at $\tilde{\boldsymbol{\beta}}$. We provide more details on its role in the following section. For the rest of this chapter, we denote by $\hat{\boldsymbol{\beta}}$ the solution of (5.13).

5.3.2 Asymptotic results

In this section, we present asymptotic results for the adaptive linearized Dantzig selector estimator when $|D_n| \to \infty$ as $n \to \infty$ with a fixed number of parameters p. More precisely, we consider increasing domain asymptotic, i.e. $\mathbf{X}$ is 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 $n \to \infty$. The modified tuning parameters $\tilde{\lambda}_j = \tilde{\lambda}_{n,j}$ for $j = 1, \ldots, p$ are now indexed by n, so $\boldsymbol{\Lambda}_n = \text{diag}\{\tilde{\lambda}_{n,1}, \cdots, \tilde{\lambda}_{n,p}\}$. We only present in this section the results for the methods considering estimating equations derived from the Poisson like-

lihood. For sake of conciseness, we do not present asymptotic results for the methods using estimating equations derived from the logistic regression likelihood. The results are very similar. The main difference is lying in the conditions ($\mathcal{C}.6$)-($\mathcal{C}.8$) for which the matrices $\mathbf{A}_n, \mathbf{B}_n$, and $\mathbf{C}_n$ have a different expression (see Remark 5.3.3).

5.3.2.1 Notation and assumptions

Throughout this section and Sections 5.6.1-5.6.5, let

$$\mathbf{U}_n(w;\boldsymbol{\beta}) = \mathbf{U}_{n,\mathrm{PL}}(w;\boldsymbol{\beta}) = \sum_{u\in\mathbf{X}\cap D_n} w(u)\mathbf{z}(u) - \int_{D_n} w(u)\mathbf{z}(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u \tag{5.14}$$

be the score vector of the weighted Poisson likelihood.

We denote by $\boldsymbol{\beta}_0 = \{\beta_{01},\dots,\beta_{0s},\beta_{0(s+1)},\dots,\beta_{0p}\}^\top = \{\boldsymbol{\beta}_{01}^\top,\boldsymbol{\beta}_{02}^\top\}^\top = (\boldsymbol{\beta}_{01}^\top,\mathbf{0}^\top)^\top$ the p-dimensional vector of true coefficients, where $\boldsymbol{\beta}_{01}$ is the s-dimensional vector of nonzero coefficients and $\boldsymbol{\beta}_{02}$ is the $(p-s)$-dimensional vector of zero coefficients. Let $\boldsymbol{\Lambda}_{n,1} = \mathrm{diag}\{\tilde{\lambda}_{n,1},\cdots,\tilde{\lambda}_{n,s}\}$ and $\boldsymbol{\Lambda}_{n,2} = \mathrm{diag}\{\tilde{\lambda}_{n,s+1},\cdots,\tilde{\lambda}_{n,p}\}$ denote the decomposition of $\boldsymbol{\Lambda}_n$.

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

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

where $g(u,v)$ is the pair correlation function indicating the attraction (or repulsion) among points 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 by $\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)$ (resp. $\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0)$, $\mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)$) the $s\times s$ top-left corner

of $\mathbf{A}_n(w;\boldsymbol{\beta}_0)$ (resp. $\mathbf{B}_n(w;\boldsymbol{\beta}_0), \mathbf{C}_n(w;\boldsymbol{\beta}_0)$).

For any $\boldsymbol{\beta} \in \Theta$, we decompose $\mathbf{A}_n(w;\boldsymbol{\beta})$ by

$$\mathbf{A}_n(w;\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{A}_{n,1}(w;\boldsymbol{\beta}) \\ \\ \mathbf{A}_{n,2}(w;\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{n,11}(w;\boldsymbol{\beta}) & \mathbf{A}_{n,12}(w;\boldsymbol{\beta}) \\ \\ \mathbf{A}_{n,21}(w;\boldsymbol{\beta}) & \mathbf{A}_{n,22}(w;\boldsymbol{\beta}) \end{bmatrix}, \tag{5.15}$$

where $\mathbf{A}_{n,1}(w;\boldsymbol{\beta})$ (resp. $\mathbf{A}_2(w;\boldsymbol{\beta})$) is the first $s \times p$ (resp. the following $(p-s)\times p$) components of $\mathbf{A}_n(w;\boldsymbol{\beta})$ and $\mathbf{A}_{n,11}(w;\boldsymbol{\beta})$ (resp. $\mathbf{A}_{n,12}(w;\boldsymbol{\beta})$, $\mathbf{A}_{n,21}(w;\boldsymbol{\beta})$, and $\mathbf{A}_{n,22}(w;\boldsymbol{\beta})$) is the $s \times s$ top-left corner (resp. the $s \times (p-s)$ top-right corner, the $(p-s) \times s$ bottom-left corner, and the $(p-s) \times (p-s)$ bottom-right corner) of $\mathbf{A}_n(w;\boldsymbol{\beta})$.

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) We assume that the intensity function has the log-linear specification given by (5.1) where $\boldsymbol{\beta} \in \Theta$ and Θ is an open convex bounded set of $\mathbb{R}^p$.

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

$$\sup_{u \in \mathbb{R}^d} ||\mathbf{z}(u)|| < \infty \quad \text{and} \quad \sup_{u \in \mathbb{R}^d} |w(u)| < \infty.$$

($\mathcal{C}$.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 (5.4), 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) There exists a $p \times p$ positive definite matrix $\mathbf{I}_0$ such that for all sufficiently large n, $|D_n|^{-1}\{\mathbf{B}_n(w;\boldsymbol{\beta}_0) + \mathbf{C}_n(w;\boldsymbol{\beta}_0)\} \geq \mathbf{I}_0$.

($\mathcal{C}$.7) There exists a $p \times p$ positive definite matrix $\mathbf{I}_0'$ such that for all sufficiently large n, we have $|D_n|^{-1}\mathbf{A}_n(w;\boldsymbol{\beta}_0) \geq \mathbf{I}_0'$.

($\mathcal{C}$.8) The initial estimate $\tilde{\boldsymbol{\beta}}$ satisfies $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_{\mathrm{P}}(|D_n|^{-1/2})$. Furthermore, we assume that $\left\|\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1} - \mathbf{I}_s\right\| = o_{\mathrm{P}}(1)$, where $\mathbf{I}_s$ is an identity matrix with dimension s.

($\mathcal{C}$.9) Consider the sequences $\tilde{a}_n = \max_{j=1,\ldots,s} \tilde{\lambda}_{n,j}$ and $\tilde{b}_n = \min_{j=s+1,\ldots,p} \tilde{\lambda}_{n,j}$. We assume that $\tilde{a}_n|D_n|^{-1/2} \to 0$ and $\tilde{b}_n|D_n|^{-1/2} \to \infty$ as $n \to \infty$.

Conditions ($\mathcal{C}$.1)-($\mathcal{C}$.7) are similar to the ones required in Chapter 2. Although condition ($\mathcal{C}$.8) looks quite strong, the initial estimates required by this condition can be satisfied by considering, for example, an estimate obtained from estimating equation-based methods developed by Guan and Shen (2010); Baddeley et al. (2014), i.e. estimating equations described in Chapter 2 (Section 2.3) when no regularization is considered. Condition ($\mathcal{C}$.9) is also imposed in Chapter 2 since the sequences $\tilde{a}_n$ and $\tilde{b}_n$ used in the current chapter are similar to the ones defined in Chapter 2 when the adaptive lasso penalty is considered. For a moment, denote $a_n = \tilde{a}_n|D_n|$ and $b_n = \tilde{b}_n|D_n|$, the condition ($\mathcal{C}$.9) becomes $a_n\sqrt{|D_n|} \to 0$ and $b_n\sqrt{|D_n|} \to \infty$, which is exactly the assumption we imposed on the regularization methods using the adaptive lasso technique (see Theorem 3.6.2 in Chapter 2).

5.3.2.2 Main results

We state our main results here. Proofs are relegated to Sections 5.6.1-5.6.5.

The optimization problem (5.13) allows us to characterize the solution $\hat{\boldsymbol{\beta}}$ in terms of primal and dual feasibility and complementary slackness conditions stated in Lemma 5.3.1.

Lemma 5.3.1. *If there exists $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ such that*

$$\left\|\boldsymbol{\Lambda}_n^{-1}\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\}\right\|_\infty \leq 1, \tag{5.16}$$

$$\|\boldsymbol{\Lambda}_n^{-1}\mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}}\|_\infty \leq 1, \tag{5.17}$$

$$\hat{\boldsymbol{\alpha}}^\top\boldsymbol{\Lambda}_n^{-1}\mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}} = \|\boldsymbol{\Lambda}_n\hat{\boldsymbol{\beta}}\|_1, \tag{5.18}$$

$$\hat{\boldsymbol{\alpha}}^\top\boldsymbol{\Lambda}_n^{-1}\Big\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\Big\} = \|\hat{\boldsymbol{\alpha}}\|_1, \tag{5.19}$$

then $\hat{\boldsymbol{\beta}}$ *is the solution to (5.13).*

In particular, conditions (5.16) and (5.17) are primal and dual feasibility conditions, while (5.18) and (5.19) are complementary slackness conditions. Inspired by Asif (2008), we follow the main arguments by Boyd and Vandenberghe (2004, section 5.2) to derive the Dantzig selectors' dual optimization problem to (5.13), given by (5.31), and present them in detail in Section 5.6.2. If the vector $\hat{\boldsymbol{\alpha}}$ in Lemma 5.3.1 is defined as the solution to (5.31), Lemma 5.3.1 implies that conditions (5.16)-(5.19) are necessary and sufficient for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ to be the unique primal-dual solution pair to (5.13) and (5.31).

Considering Lemma 5.3.1, we provide in Theorem 5.3.1 the primal-dual pair solution $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$. Furthermore, under some conditions, we find a closed form expression for $\hat{\boldsymbol{\beta}}$ which is valid with probability tending to 1.

Theorem 5.3.1. *Suppose the conditions (C.1)-(C.3) and (C.7)-(C.9) hold. Then, with probablity tending to 1, the solutions to (5.13) and (5.31) denoted respectively by* $\hat{\boldsymbol{\beta}} = \{\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top\}^\top$ *and* $\hat{\boldsymbol{\alpha}} = \{\hat{\boldsymbol{\alpha}}_1^\top, \hat{\boldsymbol{\alpha}}_2^\top\}^\top$ *are given by*

$$\hat{\boldsymbol{\beta}}_1 = \{\mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})\}^{-1}\Big\{\mathbf{U}_{n,1}(w; \tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,1}(w; \tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\alpha}}_1)\Big\}, \tag{5.20}$$

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{0}, \tag{5.21}$$

$$\hat{\boldsymbol{\alpha}}_1 = \boldsymbol{\Lambda}_{n,1}\{\mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})\}^{-1}\boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\beta}}_1), \tag{5.22}$$

$$\hat{\boldsymbol{\alpha}}_2 = \mathbf{0}. \tag{5.23}$$

Remark 5.3.1. *Condition (C.9) is the key in Theorem 5.3.1. Suppose we consider a constant* $\tilde{\lambda}_n$ *as used by the original Dantzig selector (Candes and Tao, 2007). This implies that* $\tilde{a}_n = \tilde{b}_n = \tilde{\lambda}_n$*, so the two conditions* $\tilde{a}_n |D_n|^{-1/2} \to 0$ *and* $\tilde{b}_n |D_n|^{-1/2} \to \infty$ *as* $n \to \infty$ *cannot be satisfied simultaneously. This justifies the introduction of the adaptive version of the Dantzig selector.*

In Theorem 5.3.1, by the main argument stated by condition (C.9), our estimator possesses sparsity, meaning that our estimator will correctly set $\boldsymbol{\beta}_2$ to zero with probability tending to 1 as $n \to \infty$. Furthermore, we demonstrate in Theorem 5.3.2 that the first s elements of the estimator proposed by Theorem 5.3.1, that is $\hat{\boldsymbol{\beta}}_1$, is asymptotically normal.

Theorem 5.3.2. *Assume the conditions (C.1)-(C.9) hold. Then, $\hat{\boldsymbol{\beta}}_1$ defined by Theorem 5.3.1 satisfies:*

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

where

$$\begin{aligned}\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0) =& |D_n|\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\} \\ & \{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1}. \end{aligned} \tag{5.24}$$

As a consequence, $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$. Note that $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}$ 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 $\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\left(\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{1/2}\right)^\top = \boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)$.

Remark 5.3.2. *We observe that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is the same one as the one derived from the Poisson likelihood regularized by an adaptive lasso penalty, see Theorem 3.6.2 in Chapter 2.*

Remark 5.3.3. *Theorems 5.3.1 and 5.3.2 remain true for the adaptive linearized Dantzig selector using estimating functions derived from logistic regression likelihood if we extend the condition (C.3) by adding $\sup_{u\in\mathbb{R}^d}\delta(u) < \infty$, and replace in the expression of the matrices $\mathbf{A}_n, \mathbf{B}_n$, and $\mathbf{C}_n$, $w(u)$ by $w(u)\delta(u)/(\rho(u;\boldsymbol{\beta}_0) + \delta(u)), u \in D_n$.*

Condition (C.8) has also to be enriched; we require that $\tilde{\boldsymbol{\beta}}$ is such that $\Delta_n(\tilde{\boldsymbol{\beta}}) = O_{\mathrm{P}}(1)$, where $\Delta_n(\tilde{\boldsymbol{\beta}})$ is defined by (5.39) in Section 5.6.4 with $\mathbf{U}_{n,1}(\boldsymbol{\beta}_0)$ is given by (5.8) and $\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)$ is the corresponding $s \times s$ top-left corner of $\mathbf{A}_n(w;\boldsymbol{\beta}_0)$.

We show in Theorems 5.3.1 and 5.3.2 that the sparsity and the asymptotic normality are valid for the adaptive linearized Dantzig selector esimates. Since (5.24) is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_1$, this implies that we have the same efficiency as 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.

5.3.3 Computations

This section discusses the numerical aspects used to compute the adaptive linearized Dantzig selector estimates. We aim to solve (5.13) using an iterative algorithm which is quite similar to the one proposed by James and Radchenko (2009) for computing the Dantzig selector estimator for generalized linear models. Another computational strategy seems also possible using primal-dual pursuit algorithm as proposed by Asif

(2008) for linear models. An extension to spatial point process setting may be feasible. This has not been investigated in this chapter.

In this section, we present the computational implementation for the method using the estimating functions derived from the Poisson likelihood. A similar approach can be derived for the version using the logistic regression likelihood.

We first review in Section 5.3.3.1 the method proposed by Berman and Turner (1992) who proved that fitting a Poisson point process is nearly equivalent to fitting a weighted Poisson generalized linear models. Second, we present in Section 5.3.3.2 the computational algorithm to compute the adaptive linearized Dantzig selector estimates based on the results obtained in Section 5.3.3.1. Third, the final procedure depends on a tuning parameter, a problem discussed in Section 5.3.3.3.

5.3.3.1 Weighted Poisson regression

Berman and Turner (1992) developed a numerical quadrature method to approximate maximum likelihood estimation for an inhomogeneous Poisson point process. Suppose we approximate the integral term in (5.5) by Riemann sum approximation

$$\int_D w(u)\mathbf{z}(u)\rho(u;\boldsymbol{\beta})\mathrm{d}u \approx \sum_{i=1}^{M} v_i w_i \mathbf{z}(u_i)\rho(u_i;\boldsymbol{\beta})$$

where $u_i, i = 1, \ldots, M$ are points in D consisting of the m data points and $M - m$ dummy points. To simplify notation, note that v_i (resp. w_i) means $v(u_i)$ (resp. $w(u_i)$). The quadrature weights $v_i > 0$ are such that $\sum_i v_i = |D|$. To implement this method, the domain is firstly partitioned into M rectangular pixels of equal area, denoted by a. Then one dummy point is placed in the center of the pixel. Let Δ_i be an indicator whether the point is an event of point process ($\Delta_i = 1$) or a dummy point ($\Delta_i = 0$). Without loss of generality, let $u_1, \ldots, u_m$ be the observed events and $u_{m+1}, \ldots, u_M$ be the dummy points. Thus, (5.5) can be approximated and rewritten as

$$U(\boldsymbol{\beta}) \approx \sum_{i}^{M} v_i w_i \mathbf{z}(u_i)\{y_i - \rho(u_i;\boldsymbol{\beta})\}, \text{ where } y_i = v_i^{-1}\Delta_i. \tag{5.25}$$

Equation (5.25) is formally equivalent to the weighted score function of independent

Poisson variables y_i with weights $v_i w_i$. Thus, standard statistical software for generalized linear models can be used to obtain the estimates. This fact is implemented in the `spatstat` `R` package by `ppm` function (Baddeley et al., 2015).

5.3.3.2 The adaptive linearized Dantzig selector algorithm

Before presenting the main algorithm, let us first consider the following matrices and vectors. We define the $M \times p$ matrix $\mathbf{z}$ and the $M \times M$ matrix $\mathbf{v}$ by

$$\mathbf{z} = \begin{bmatrix} z_1(u_1) & z_2(u_1) & \dots & z_p(u_1) \\ \vdots & \vdots & \ddots & \vdots \\ z_1(u_M) & z_2(u_M) & \dots & z_p(u_M) \end{bmatrix}$$

$$\mathbf{v} = \text{diag}\{v_1 w_1, \dots, v_M w_M\},$$

where $v_1, \dots, v_M$ and $w_1, \dots, w_M$ are, respectively, the quadrature weights and the weight function obtained from (5.25). We remind the readers that M is the total number of observed points and dummy points which plays the same role as the number of observations in classical regression analysis. We also denote the ith row and the jth column of the matrix $\mathbf{z}$ by the vectors $\mathbf{z}_{i\cdot}$ and $\mathbf{z}_{\cdot j}$, respectively, given by

$$\mathbf{z}_{i\cdot} = \{\mathbf{z}_1(u_i), \dots, \mathbf{z}_p(u_i)\}^\top, \text{ for } i = 1, \dots, M,$$
$$\mathbf{z}_{\cdot j} = \{\mathbf{z}_j(u_1), \dots, \mathbf{z}_j(u_M)\}^\top, \text{ for } j = 1, \dots, p.$$

By convention, $z_{ij} = \mathbf{z}_j(u_i)$.

Now, by (5.25), the adaptive linearized Dantzig selector criterion defined in (5.13) can be rewritten as

$$\min \sum_{j=1}^{p} \tilde{\lambda}_j |\beta_j| \text{ subject to } \left| \mathbf{z}_{\cdot j}^\top \mathbf{v} \Big(\mathbf{y} - \exp(\mathbf{z}\tilde{\boldsymbol{\beta}}) \Big) + \mathbf{z}_{\cdot j}^\top \mathbf{v} \exp(\mathbf{z}\tilde{\boldsymbol{\beta}}) \mathbf{z}_{\cdot j}^\top (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right| \leq \tilde{\lambda}_j \quad (5.26)$$

where $\mathbf{y} = \{y_1, \cdots, y_M\}^\top$ is given by (5.25).

Note that, given an initial estimate $\tilde{\boldsymbol{\beta}}$, (5.26) is a linear optimization problem which can be solved direcly by linear programming. Form a theoretical point of view, we

have already mentioned that the initial estimates obtained by maximizing the Poisson likelihood or the logistic regression likelihood satisfy condition ($\mathcal{C}.8$) which is sufficient to derive the asymptotic results. In practice, these proposed initial estimates can be computed easily by `ppm` function in the `spatstat R` package. However, these choices cannot always perform well when the number of parameters to estimate is large due to unstability and convergence issues. Another alternative is to start with $\tilde{\boldsymbol{\beta}}$ obtained as the ridge regression estimate, i.e. maximizing $\ell(w;\boldsymbol{\beta}) - \sum_{j=1}^{p} \tilde{\lambda}\beta_j^2$, to obtain more stable initial estimate.

In this study, instead of using a single initial estimate, we propose an iterative algorithm to compute the estimate for which it is always updated until convergence criterion is held. This will guarantee the stability of the results, hence, improve the prediction. More precisely, our iterative algorithm is divided into three steps, explained as follows. Our approach is quite similar to the ones proposed by James and Radchenko (2009) and can be viewed as its extension as we consider the weights $v_i w_i, i = 1, \ldots, M$ in Step 2 and use the adaptive version of the Dantzig selector in Step 3.

Step 0. Set the initial estimates $\tilde{\boldsymbol{\beta}}^{(0)}$.

Step 1. Note that (k) denotes the corresponding estimate from the kth iteration. At the $(k+1)$th iteration, compute $\Omega_i = \exp(\mathbf{z}_{i\cdot}^{\top}\tilde{\boldsymbol{\beta}}^{(k)})$ and $Y_i = \sum_{j=1}^{p} z_{ij}\tilde{\beta}_j^{(k)} + \Big(y_i - \exp(\mathbf{z}_{i\cdot}^{\top}\tilde{\boldsymbol{\beta}}^{(k)})\Big)/\Omega_i$, where y_i is given by (5.25).

Step 2. Define $\tilde{Y}_i = Y_i\sqrt{v_i w_i \Omega_i}$ and $\tilde{z_{ij}} = z_{ij}\sqrt{v_i w_i \Omega_i}$.

Step 3. Use the adaptive Dantzig selector for linear models

$$\min \sum_{j=1}^{p} \tilde{\lambda}_j|\beta_j| \text{ subject to } |\tilde{\mathbf{z}}_{\cdot j}^{\top}(\tilde{\mathbf{Y}} - \tilde{\mathbf{z}}\boldsymbol{\beta})| \leq \tilde{\lambda}_j \text{ for } j = 1, \ldots, p,$$

to compute $\tilde{\boldsymbol{\beta}}^{(k+1)}$.

Step 4. Repeat Steps 1-3 until convergence.

More precisely, Step 3 is equivalent to

$$\min \sum_{j=1}^{p} \tilde{\lambda}_j |\beta_j| \text{ subject to } |\mathbf{z}_{.j}^{\top} \mathbf{v}\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{z}\boldsymbol{\beta})| \leq \tilde{\lambda}_j \text{ for } j = 1, \ldots, p, \tag{5.27}$$

where $\boldsymbol{\Omega}$ is the $M \times M$ matrix $\boldsymbol{\Omega} = \text{diag}\{\Omega_1, \ldots, \Omega_M\}$. Note that (5.27) can be solved by linear programming. Furthermore, for any given $\tilde{\boldsymbol{\beta}}$, (5.27) is the same as the ones defined by (5.26), so the solution to (5.27) is in the feasible region for (5.26).

5.3.3.3 Tuning parameter selection

Our procedure relies on a tuning parameter $\tilde{\lambda}_j = j = 1, \cdots, p$. Similarly to Chapters 2 and 3, we define $\tilde{\lambda}_j = \tilde{\lambda}|\hat{\beta}_j(ridge)|^{-1}$, where $\tilde{\lambda}$ is the modified tuning parameter and $\hat{\boldsymbol{\beta}}(ridge)$ is the estimate obtained from ridge regression. Therefore, we need to choose an appropriate $\tilde{\lambda}$, which is a scalar, in order to achieve good performance. Some previous studies have been conducted by Zou et al. (2007); Wang et al. (2007b, 2009) who suggested to use BIC-type methods to select the tuning parameter. We follow these works and suggest to choose $\tilde{\lambda}$ which minimizes WQBIC($\tilde{\lambda}$), a weighted version of the BIC criterion, defined by

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

where $\ell(w; \hat{\boldsymbol{\beta}}(\tilde{\lambda}))$ is the Poisson likelihood function, $s(\tilde{\lambda}) = \sum_{j=1}^{p} \mathbb{I}\{\hat{\beta}_j(\tilde{\lambda}) \neq 0\}$ is the number of selected covariates with nonzero regression coefficients and $|D|$ is the volume of observation domain.

5.4 Simulation study

In this section, we aim to highlight the finite sample properties of our estimates in a simulation experiment and compare them to the estimates obtained by maximizing the Poisson likelihood penalized by adaptive lasso penalty previously developped in Chapters 2 and 3. We consider a setting similar to the ones used in Chapters 2 and 3. We start with relatively complex situation where strong multicollinearity is present (Scenarios 1a and 2a) and we then consider 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 set to be different in each setting.

We set the spatial domain to be $D = [0, 1000] \times [0, 500]$ and set the mean number of points over D to be 1600. The true intensity function is set to be $\rho(u; \boldsymbol{\beta}_0) = \exp(\mathbf{z}(u)^\top \boldsymbol{\beta}_0)$, where $\mathbf{z}(u) = \{1, z_1(u), \ldots, z_{50}(u)\}^\top$ and $\boldsymbol{\beta}_0 = \{\beta_0, \beta_{01}, \cdots, \beta_{050}\}$. Here, we do not estimate β_0 since it is chosen such that each realization has 1600 points in average. 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` (R Core Team, 2016), 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 $\mathbf{z}(u) = \mathbf{V}^\top \mathbf{x}(u)$, 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.

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 extra 45 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 $\mathbf{z}(u) = \mathbf{V}^\top \mathbf{x}(u)$, 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.

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 fewer number of parents (see e.g., Møller and Waagepetersen, 2004, for further details regarding the Thomas point process).

We present in Table 5.1 the selection properties of our estimates. We consider the true positive rate (TPR), the false positive rate (FPR), and the positive predictive value (PPV) to evaluate the selection properties of the estimates. 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

Table 5.1: 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. The Poisson likelihood (PL) and the weighted Poisson likelihood (WPL) are combined with two feature selection procedures: the adaptive lasso (AL) and the adaptive linearized Dantzig selector (ALDS).

Method	PL			WPL			PL			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											
AL	100[1]	1	92	97	0[1]	96	95	3	70	70	0[1]	98
ALDS	100[1]	0[1]	95	50	0[1]	100[1]	92	2	80	50	0[1]	93
	Scenario 1b											
AL	100[1]	26	19	95	4	83	99	85	5	29	6	39
ALDS	100[1]	16	31	51	0[1]	100[1]	98	61	6	19	1	31
	Scenario 2a											
AL	96	48	19	85	36	27	96	64	15	45	17	31
ALDS	84	19	36	0[1]	0[1]	0[1]	82	26	33	0[1]	0[1]	1
	Scenario 2b											
AL	56	27	19	54	26	19	59	33	18	48	27	13
ALDS	46	14	28	0[1]	0[1]	0[1]	46	14	30	0[1]	0[1]	0[1]

[1] Approximate value

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.

We do not consider in this chapter the extra covariates generated as standard Gaussian white noise independently as considered in Chapter 2 (Scenario 1). We are still able to show that even when the strong multicollinearity exists such as in Scenario 1a, our proposed methods work well especially using unweighted methods. 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.

Table 5.2: 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. The Poisson likelihood (PL) and the weighted Poisson likelihood (WPL) are combined with two feature selection procedures: the adaptive lasso (AL) and the adaptive linearized Dantzig selector (ALDS).

Method	PL			WPL			PL			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											
AL	0.05	0.18	0.19	0.15	0.25	0.29	0.20	0.59	0.63	0.57	0.56	0.80
ALDS	0.05	0.19	0.20	0.82	0.19	0.84	0.22	0.60	0.64	0.87	0.52	1.02
	Scenario 1b											
AL	1.26	6.13	6.26	1.63	1.81	2.43	0.18	0.71	0.73	0.28	0.40	0.49
ALDS	0.68	3.92	3.98	1.80	0.65	1.92	0.16	0.54	0.56	0.87	0.20	0.90
	Scenario 2a											
AL	1.56	1.76	2.36	2.09	2.10	2.96	1.44	4.89	5.10	4.66	3.85	6.05
ALDS	2.35	1.59	2.84	7.41	0.12	7.41	2.64	3.61	4.47	7.40	0.18	7.41
	Scenario 2b											
AL	3.65	1.62	3.99	3.71	1.85	4.14	6.96	2.21	7.30	7.10	1.88	7.35
ALDS	3.94	1.19	4.12	7.41	0.07	7.41	7.03	1.60	7.21	7.42	0.00	7.42

In general, the best selection properties are obtained from larger κ (5×10^{-4}) which indicates weaker spatial dependence. To compare between the adaptive lasso (AL) and the adaptive linearized Dantzig selector (ALDS), we find that ALDS seems to be sparser than AL. AL and ALDS (combined with PL) perform quite similar for Scenario 1a, but ALDS (combine with PL) performs better for a more complex situation. However, the combination between the WPL and ALDS tends to underfit model by selecting very few covariates. From Table 5.1, we would recommend in general to combine the PL with ALDS to perform variable selection.

Table 5.2 gives the prediction properties of the estimates in terms of biases, standard deviations (SD), and square root of mean squared errors (RMSE), some criterions 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. To compare the two variable selection procedures, combined with the PL, ALDS has a smaller variance than AL but has larger biases. This may come from the fact that ALDS is sparser than AL. The biases are in general not too large yielding a better performance in terms of RMSE for the ALDS. When the WPL is considered, ALDS still has smaller variance but much higher biases, so the AL is more preferable. This is because ALDS (combined with WPL) selects much fewer covariates. From table 5.2, when our focus is on prediction, as a general advice, we would recommend combining the PL with ALDS.

5.5 Conclusion and discussion

We propose the adaptive linearized Dantzig selector (ALDS), a modified version of the Dantzig selector based on linear approximation on the constraint vector, for spatial point processes intensity estimation. Under some conditions, we prove that the estimates obtained from such procedures satisfy sparsity and asymptotic normality. We find that the asymptotic properties we derive under ALDS are similar to the ones we develop under adaptive lasso (AL) procedures as studied in Chapter 2. For computational point of view, as we make links between spatial point processes intensity estimation and generalized linear models (GLMs), we only need to deal with feature selection procedures for GLMs which are easy to implement and computationally fast.

We make a simulation study with some different scenarios to assess the finite sample performances of the estimates obtained from ALDS compared to the ones from AL. In general, the best two methods are between the weighted Poisson likelihood (WPL) with AL and the Poisson likelihood (PL) with ALDS. Since the combination between

the PL and ALDS performs slightly better and looks more stable, we would recommend using the PL combined with ALDS.

Future work could consist of considering combination between l_1 and l_2 penalties in a fashion similar to the adaptive elastic net but combined with the Dantzig selector method. In real applications, the setting which considers the involvement of a lot of covariates with complex spatial structure and the presence of strong multicollinearity is easily found. In such situations, the adaptive elastic net may outperform adaptive lasso as it is the case for linear regression (see e.g. Zou and Zhang, 2009). We have preliminary results, as considered in the simulation experiment in Chapter 3, that the adaptive elastic net is more favorable than the adaptive lasso for such situations. It would be interesting to bring this approach in the Dantzig selector-type methods.

5.6 Supplementary materials

5.6.1 Auxiliary Lemma

The following Lemma is used in the proof of Theorems 5.3.1 and 5.3.2. 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 $\|\mathbf{M}_n\| = O(x_n)$. We follow the convention that $\|\cdot\|$ means $\|\cdot\|_2$.

Lemma 5.6.1. *Under conditions (C.1)-(C.5), the following result holds as* $n \to \infty$

$$\mathbf{U}_n(w; \boldsymbol{\beta}_0) = O_{\mathrm{P}}\left(\sqrt{|D_n|}\right). \tag{5.28}$$

Proof. Using Campbell Theorems (5.2)-(5.3), the score vector $\mathbf{U}_n(w; \boldsymbol{\beta}_0)$ has variance

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

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 ($\mathcal{C}$.2)-($\mathcal{C}$.3) that

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

The result is proved since for any centered real-valued stochastic process Y_n with finite

variance $\text{Var}[Y_n]$, $Y_n = O_{\text{P}}(\sqrt{\text{Var}[Y_n]})$. □

5.6.2 Formulation of the Dantzig selector's dual optimization problem to (5.13)

We follow the main arguments by Boyd and Vandenberghe (2004, section 5.2) to derive the Dantzig selectors' dual optimization problem to (5.13). First, the Lagrangian form associated with the problem (5.13) is

$$L(\boldsymbol{\beta}; \boldsymbol{\alpha}) = \|\boldsymbol{\Lambda}_n \boldsymbol{\beta}\|_1 + \boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \Big\{ \mathbf{U}_n(w; \tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Big\} - \|\boldsymbol{\alpha}\|_1, \tag{5.29}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the dual vector (which can be viewed as a lagrange multiplier) and $\tilde{\boldsymbol{\beta}}$ is any given initial estimator. To simplify the presentation, although this is not always required, we refer to $\tilde{\boldsymbol{\beta}}$ as the initial estimator which satisfies condition ($\mathcal{C}$.8). Second, the dual function of (5.29) is defined by

$$\begin{aligned} \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}; \boldsymbol{\alpha}) =& \boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \Big\{ \mathbf{U}_n(w; \tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) \tilde{\boldsymbol{\beta}} \Big\} - \|\boldsymbol{\alpha}\|_1 \\ & + \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \Big(\text{sign}(\boldsymbol{\beta}) - \boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) \boldsymbol{\Lambda}_n^{-1} \Big) \boldsymbol{\Lambda}_n \boldsymbol{\beta} \Big\} \\ =& \begin{cases} \boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \Big\{ \mathbf{U}_n(w; \tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) \tilde{\boldsymbol{\beta}} \Big\} - \|\boldsymbol{\alpha}\|_1, & \text{if } \|\boldsymbol{\Upsilon}_n\|_\infty \leq 1, \\ -\infty & \text{otherwise,} \end{cases} \end{aligned} \tag{5.30}$$

where $\boldsymbol{\Upsilon}_n = \boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) \boldsymbol{\Lambda}_n^{-1}$. Note that $\text{sign}(\boldsymbol{\beta})^\top \boldsymbol{\Lambda} \boldsymbol{\beta} = \|\boldsymbol{\Lambda} \boldsymbol{\beta}\|_1$ since $\boldsymbol{\Lambda}_n$ is a diagonal matrix with positive entries. The solution set $\boldsymbol{\beta}$ is feasible if $\|\boldsymbol{\Upsilon}_n\|_\infty \leq 1$. Third, the Lagrange dual of (5.29) maximizes the dual function given by $\sup_{\boldsymbol{\alpha} \in \mathbb{R}^p} \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}; \boldsymbol{\alpha})$. This can be reformulated as a dual problem to (5.13), by including the dual feasibility condition as a constraint, given by

$$\begin{aligned} & \max \Big(\boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \Big\{ \mathbf{U}_n(w; \tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) \tilde{\boldsymbol{\beta}} \Big\} - \|\boldsymbol{\alpha}\|_1 \Big) \\ & \text{subject to } \|\boldsymbol{\alpha}^\top \boldsymbol{\Lambda}_n^{-1} \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) \boldsymbol{\Lambda}_n^{-1}\|_\infty \leq 1. \end{aligned} \tag{5.31}$$

5.6.3 Proof of Lemma 5.3.1

Proof. Consider (5.29) and its dual given by (5.30). By arguments similar to the ones derived in Section 5.6.2, we can show that the dual function of the Lagrange form associated with the problem (5.31) is equivalent to

$$\begin{aligned}\sup_{\boldsymbol{\alpha}\in\mathbb{R}^p} L(\boldsymbol{\beta};\boldsymbol{\alpha}) &= \|\boldsymbol{\Lambda}_n\boldsymbol{\beta}\|_1 + \sup_{\boldsymbol{\alpha}\in\mathbb{R}^p} \boldsymbol{\alpha}^\top\Big(\boldsymbol{\Lambda}_n^{-1}\Big\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\Big\} - \operatorname{sign}(\boldsymbol{\alpha})\Big) \\ &= \begin{cases} \|\boldsymbol{\Lambda}_n\boldsymbol{\beta}\|_1, & \text{if } \left\|\boldsymbol{\Lambda}_n^{-1}\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\}\right\|_\infty \le 1, \\ \infty & \text{otherwise,} \end{cases} \end{aligned} \tag{5.32}$$

if we set $-\boldsymbol{\Lambda}_n\boldsymbol{\beta}$ as the dual vector in the Lagrange form associated with the problem (5.31). The solution set $\boldsymbol{\alpha}$ is feasible if $\left\|\boldsymbol{\Lambda}_n^{-1}\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\}\right\|_\infty \le 1$. This shows that the dual function of (5.31) is equivalent to the primal problem (5.13).

Now, let us take any $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$. By conditions (5.17), (5.19), and (5.18) respectively

$$\begin{aligned}\inf_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta};\hat{\boldsymbol{\alpha}}) &= \hat{\boldsymbol{\alpha}}^\top\boldsymbol{\Lambda}_n^{-1}\Big\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}}\Big\} - \|\hat{\boldsymbol{\alpha}}\|_1 \\ &= \hat{\boldsymbol{\alpha}}^\top\boldsymbol{\Lambda}_n^{-1}\mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}} \\ &= \|\boldsymbol{\Lambda}_n\hat{\boldsymbol{\beta}}\|_1.\end{aligned}$$

Thus, it is valid that

$$\|\boldsymbol{\Lambda}_n\hat{\boldsymbol{\beta}}\|_1 = \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta};\hat{\boldsymbol{\alpha}}) \le \sup_{\boldsymbol{\alpha}\in\mathbb{R}^p}\inf_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta};\boldsymbol{\alpha}) \le \sup_{\boldsymbol{\alpha}\in\mathbb{R}^p} L(\bar{\boldsymbol{\beta}};\boldsymbol{\alpha}), \tag{5.33}$$

for any given $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$. Regarding (5.32) and (5.33), we show that $\|\boldsymbol{\Lambda}_n\hat{\boldsymbol{\beta}}\| \le \|\boldsymbol{\Lambda}_n\boldsymbol{\beta}\|$ whenever the solution set $\boldsymbol{\alpha}$ is feasible. Thus, by condition (5.16), it is proved that $\hat{\boldsymbol{\beta}}$ solves (5.13). In addition, since $\sup_{\boldsymbol{\alpha}\in\mathbb{R}^p} L(\hat{\boldsymbol{\beta}};\boldsymbol{\alpha}) = \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta};\hat{\boldsymbol{\alpha}}) = L(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\alpha}}) = \|\boldsymbol{\Lambda}_n\hat{\boldsymbol{\beta}}\|_1$ by conditions (5.16)-(5.19), strong duality holds, meaning that $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ is a unique primal dual solution to the problems (5.13) and (5.31). □

5.6.4 Proof of Theorem 5.3.1

Proof. To prove Theorem 5.3.1, it is sufficient to show that with probability tending to 1, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ given by (5.20)-(5.23) satisfy conditions (5.16)-(5.19).

We derive the following results which will be used in the proofs of Theorems 5.3.1 and 5.3.2. Consider the matrix $\mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})$ and their corresponding partitions given by (5.15), and the matrices $\boldsymbol{\Lambda}_{n,1}$ and $\boldsymbol{\Lambda}_{n,2}$. Note that by condition ($\mathcal{C}$.9),

$$\boldsymbol{\Lambda}_{n,1} = o(|D_n|^{1/2}), \tag{5.34}$$

$$\boldsymbol{\Lambda}_{n,2}^{-1} = o(|D_n|^{-1/2}). \tag{5.35}$$

and by conditions ($\mathcal{C}$.1)-($\mathcal{C}$.3) and ($\mathcal{C}$.7)-($\mathcal{C}$.8), $\sup_{\beta\in\Theta}\sup_{u\in\mathbb{R}^d}\rho(u,\beta) < \infty$ so in partic-

ular for any $t \in [0,1]$ and $\check{\beta} = \tilde{\beta} + t(\beta_0 - \tilde{\beta})$

$$A_n(w; \check{\boldsymbol{\beta}}) = O(|D_n|), \tag{5.36}$$

which also implies that

$$\mathbf{A}_n(w; \tilde{\boldsymbol{\beta}}) = O(|D_n|). \tag{5.37}$$

Note that (5.37) also implies that all the partitions of $\mathbf{A}_n(w; \tilde{\boldsymbol{\beta}})$ are bounded by $|D_n|$. Now, note that condition ($\mathcal{C}$.7) implies that $\mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)^{-1} = O(|D_n|^{-1})$. Thus, by conditions ($\mathcal{C}$.7)-($\mathcal{C}$.8),

$$\begin{aligned} \mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})^{-1} &= \mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)^{-1} + \mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)^{-1}\{\mathbf{A}_{n,11}(w; \boldsymbol{\beta}_0)\mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})^{-1} - \mathbf{I}_s\} \\ &= o_{\mathrm{P}}(|D_n|^{-1}). \end{aligned} \tag{5.38}$$

Finally, we define the $p \times 1$ vector $\Delta_n(w; \tilde{\boldsymbol{\beta}})$ by

$$\Delta_n(w; \tilde{\boldsymbol{\beta}}) = \mathbf{U}_n(w; \tilde{\boldsymbol{\beta}}) - \mathbf{U}_n(w; \boldsymbol{\beta}_0) + \mathbf{A}_n(w; \boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{5.39}$$

with $\Delta_{n,1}(w; \tilde{\boldsymbol{\beta}})$ and $\Delta_{n,2}(w; \tilde{\boldsymbol{\beta}})$ are respectively the first s elements and the last $p - s$ elements of $\Delta_n(w; \tilde{\boldsymbol{\beta}})$. Note that

$$\Delta_n(w; \tilde{\boldsymbol{\beta}}) = -\int_{D_n} w(u)\mathbf{z}(u)\{\rho(u; \tilde{\boldsymbol{\beta}}) - \rho(u; \boldsymbol{\beta}_0)\}\mathrm{d}u + \int_{D_n} w(u)\mathbf{z}(u)\mathbf{z}(u)^\top \rho(u; \boldsymbol{\beta}_0)\mathrm{d}u(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

Again, from condition ($\mathcal{C}$.2), there exists $t \in (0,1)$ such that

$$\begin{aligned} \rho(u; \tilde{\boldsymbol{\beta}}) - \rho(u; \boldsymbol{\beta}_0) &= (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{z}(u)\rho(u; \boldsymbol{\beta}_0) \\ &\quad + \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{z}(u)\mathbf{z}(u)^\top (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\rho(u; \boldsymbol{\beta}_0 + t(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \end{aligned}$$

which yields that

$$\begin{aligned} \Delta_n(w; \tilde{\boldsymbol{\beta}}) &= -\frac{1}{2}\int_{D_n} w(u)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{z}(u)\mathbf{z}(u)^\top (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\rho(u; \boldsymbol{\beta}_0 + t(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\mathrm{d}u \\ &= O_{\mathrm{P}}(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 |D_n|) = O_{\mathrm{P}}(1), \end{aligned} \tag{5.40}$$

from conditions ($\mathcal{C}$.1)-($\mathcal{C}$.3) and ($\mathcal{C}$.8). (5.40) also implies that $\Delta_{n,1}(w; \tilde{\boldsymbol{\beta}}) = O_{\mathrm{P}}(1)$ and $\Delta_{n,2}(w; \tilde{\boldsymbol{\beta}}) = O_{\mathrm{P}}(1)$.

Now we focus to prove Theorem 5.3.1. By (5.20)-(5.23),

$$\begin{aligned} \hat{\boldsymbol{\alpha}}^\top \boldsymbol{\Lambda}_n^{-1} \mathbf{A}_n(w; \tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1} \mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}_1 \\ &= \mathrm{sign}(\hat{\boldsymbol{\beta}}_1)^\top \boldsymbol{\Lambda}_{n,1}\Big\{\mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})\Big\}^{-1} \boldsymbol{\Lambda}_{n,1}\boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,11}(w; \tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}_1 \\ &= \mathrm{sign}(\hat{\boldsymbol{\beta}}_1)^\top \boldsymbol{\Lambda}_{n,1}\hat{\boldsymbol{\beta}}_1 \\ &= \|\boldsymbol{\Lambda}_{n,1}\hat{\boldsymbol{\beta}}_1\|_1 = \|\boldsymbol{\Lambda}_n\hat{\boldsymbol{\beta}}\|_1, \end{aligned}$$

so, (5.18) is satisfied. Now, we want to show that (5.19) holds. We have

$$\hat{\boldsymbol{\alpha}}^\top \boldsymbol{\Lambda}_n^{-1}\left\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\right\} = \mathbf{I} + \mathbf{II},$$

where

$$\begin{aligned}
\mathbf{I} &= \hat{\boldsymbol{\alpha}}^\top \boldsymbol{\Lambda}_n^{-1}\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}),\\
\mathbf{II} &= \hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}_1\\
&= \hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}}\\
&\quad - \hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1}\{\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\alpha}}_1)\}\\
&= -\hat{\boldsymbol{\alpha}}_1^\top \boldsymbol{\Lambda}_{n,1}^{-1}\{\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) - \boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\alpha}}_1)\},
\end{aligned}$$

from (5.20)-(5.23). By summing **I** and **II**, we deduce that (5.19) holds.

To prove (5.17) holds, let us decompose the vector $\boldsymbol{\Lambda}_n^{-1}\mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}}$ by

$$\boldsymbol{\Lambda}_n^{-1}\mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}} = \begin{bmatrix}\mathbf{I}'\\ \mathbf{II}'\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}}\\ \boldsymbol{\Lambda}_{n,2}^{-1}\mathbf{A}_{n,2}(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}}\end{bmatrix}.$$

Now, consider $\mathbf{I}'$. By (5.22)-(5.23), we can show that

$$\begin{aligned}
\|\mathbf{I}'\|_\infty &= \|\boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}}_1\|_\infty\\
&= \|\operatorname{sign}(\hat{\boldsymbol{\beta}}_1)\|_\infty = 1.
\end{aligned}$$

Regarding $\mathbf{II}'$, by (5.22)-(5.23), we have

$$\begin{aligned}
\mathbf{II}' &= \boldsymbol{\Lambda}_{n,2}^{-1}\mathbf{A}_{n,21}(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\hat{\boldsymbol{\alpha}}_1\\
&= \boldsymbol{\Lambda}_{n,2}^{-1}\mathbf{A}_{n,21}(w;\tilde{\boldsymbol{\beta}})\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Lambda}_{n,1}\{\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})\}^{-1}\boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\beta}}_1)\\
&= \boldsymbol{\Lambda}_{n,2}^{-1}\mathbf{A}_{n,21}(w;\tilde{\boldsymbol{\beta}})\{\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})\}^{-1}\boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\beta}}_1)\\
&= o(|D_n|^{-1/2})O(|D_n|)o_{\mathrm{P}}(|D_n|^{-1})o(|D_n|^{1/2})O_{\mathrm{P}}(1) = o_{\mathrm{P}}(1),
\end{aligned}$$

where the last line is obtained from (5.34)-(5.38). Hence, (5.17) is satisfied with probability tending to 1. We finally focus on (5.16). Note that

$$\begin{aligned}
\boldsymbol{\Lambda}_n^{-1}\{\mathbf{U}_n(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_n(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\} &= \begin{bmatrix}\tilde{\mathbf{I}}\\ \tilde{\mathbf{II}}\end{bmatrix}\\
&= \begin{bmatrix}\boldsymbol{\Lambda}_{n,1}^{-1}\{\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\}\\ \boldsymbol{\Lambda}_{n,2}^{-1}\{\mathbf{U}_{n,2}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,2}(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\}\end{bmatrix}.
\end{aligned}$$

Regarding $\tilde{\mathbf{I}}$, from (5.20)-(5.21),

$$\begin{aligned}\tilde{\mathbf{I}} =& \|\boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}_1\|_\infty \\ =& \|\boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \boldsymbol{\Lambda}_{n,1}^{-1}\mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} \\ & - \boldsymbol{\Lambda}_{n,1}^{-1}\{\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\Lambda}_{n,1}\,\mathrm{sign}(\hat{\boldsymbol{\alpha}}_1)\}\|_\infty \\ =& \|\,\mathrm{sign}(\hat{\boldsymbol{\alpha}}_1)\|_\infty = 1.\end{aligned}$$

Now, consider $\tilde{\mathbf{II}}$. By Taylor expansion, we can show that by considering $\Delta_{n,2}(w;\tilde{\boldsymbol{\beta}})$ defined by (5.40) and by conditions ($\mathcal{C}$.1)-($\mathcal{C}$.3) and ($\mathcal{C}$.8), we have

$$\begin{aligned}\mathbf{U}_{n,2}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,2}(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) &= \mathbf{U}_{n,2}(w;\boldsymbol{\beta}_0) + \Delta_{n,2}(w;\tilde{\boldsymbol{\beta}}) \\ &\quad + \{\mathbf{A}_{n,2}(w;\tilde{\boldsymbol{\beta}}) - \mathbf{A}_{n,2}(w;\boldsymbol{\beta}_0)\}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \\ &= \mathbf{U}_{n,2}(w;\boldsymbol{\beta}_0) + O_{\mathrm{P}}(1).\end{aligned}$$

Hence, by noting that $\mathbf{U}_{n,2}(w;\boldsymbol{\beta}_0) = O_{\mathrm{P}}(|D_n|^{1/2})$ from Lemma 5.6.1 and by (5.35)

$$\begin{aligned}\tilde{\mathbf{II}} &= \boldsymbol{\Lambda}_{n,2}^{-1}\{\mathbf{U}_{n,2}(w;\boldsymbol{\beta}_0) + O_{\mathrm{P}}(1)\} \\ &= o(|D_n|^{-1/2})\{O_{\mathrm{P}}(|D_n|^{1/2}) + O_{\mathrm{P}}(1)\} \\ &= o_{\mathrm{P}}(1).\end{aligned}$$

□

5.6.5 Proof of Theorem 5.3.2

Proof. Before proving Theorem 5.3.2, we provide Lemma 5.6.2 stated below. This Lemma has been discovered in Chapter 3 and the proofs are omitted here.

Lemma 5.6.2. *Under the conditions ($\mathcal{C}$.1)-($\mathcal{C}$.9), the following convergence holds in distribution as $n \to \infty$*

$$\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_s), \tag{5.41}$$

where $\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0)$ corresponds to the first s components of $\mathbf{U}_n(w;\boldsymbol{\beta}_0)$ 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 $\mathbf{C}_n(w;\boldsymbol{\beta}_0)$).

Let us consider the vector $\mathbf{T}_n$ given by

$$\begin{aligned}\mathbf{T}_n &= |D_n|^{1/2}\boldsymbol{\Sigma}_n(w;\boldsymbol{\beta}_0)^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}), \\ &= \{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}),\end{aligned}$$

where $\Sigma_n(w;\boldsymbol{\beta}_0)$ in the first line is given by (5.24).

Now, by (5.20)-(5.21), by conditions ($\mathcal{C}$.1)-($\mathcal{C}$.3), ($\mathcal{C}$.6)-($\mathcal{C}$.8) and by $\Delta_{n,1}(w;\tilde{\boldsymbol{\beta}})$ given

by (5.40), we can show using a Taylor expansion that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} &= \mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1}\Big\{\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}} - \boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\alpha}}_1)\Big\} - \boldsymbol{\beta}_{01} \\ &= \mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1}\Big\{\mathbf{U}_{n,1}(w;\tilde{\boldsymbol{\beta}}) + \mathbf{A}_{n,1}(w;\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \boldsymbol{\Lambda}_{n,1}\operatorname{sign}(\hat{\boldsymbol{\alpha}}_1)\Big\} \\ &= \mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1}\Big\{\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) + \Delta_{n,1}(w;\tilde{\boldsymbol{\beta}}) + O_{\mathrm{P}}(1) + o_{\mathrm{P}}(|D_n|^{1/2})\Big\} \\ &= \mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1}\Big\{\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) + \Delta_{n,1}(w;\tilde{\boldsymbol{\beta}}) + o_{\mathrm{P}}(|D_n|^{1/2})\Big\}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbf{T}_n =& \{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1} \\ & \Big\{\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) + \Delta_{n,1}(w;\tilde{\boldsymbol{\beta}}) + o_{\mathrm{P}}(|D_n|^{1/2})\Big\}.\end{aligned}$$

Let us now decompose $\mathbf{T}_n = \mathbf{T}_{1,n} + \mathbf{T}_{2,n}$, where, for an identity matrix with dimension s denoted by $\mathbf{I}_s$,

$$\begin{aligned}\mathbf{T}_{1,n} =& \{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\Big\{\mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1} - \mathbf{I}_s\Big\} \\ & \Big\{\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) + \Delta_n(w;\tilde{\boldsymbol{\beta}}) + o_{\mathrm{P}}(|D_n|^{1/2})\Big\}, \\ \mathbf{T}_{2,n} =& \{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\Big\{\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) + \Delta_n(w;\tilde{\boldsymbol{\beta}}) + o_{\mathrm{P}}(|D_n|^{1/2})\Big\}.\end{aligned}$$

By conditions ($\mathcal{C}$.6) and ($\mathcal{C}$.8), we have

$$\begin{aligned}\{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2} &= O(|D_n|^{-1/2}) \\ \mathbf{A}_{n,11}(w;\boldsymbol{\beta}_0)\mathbf{A}_{n,11}(w;\tilde{\boldsymbol{\beta}})^{-1} - \mathbf{I}_s &= o_{\mathrm{P}}(1),\end{aligned}$$

and let us remind that $\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) = O_{\mathrm{P}}(|D_n|^{1/2})$ from Lemma 5.6.1 and $\Delta_n(w;\tilde{\boldsymbol{\beta}}) + o_{\mathrm{P}}(|D_n|^{1/2}) = o_{\mathrm{P}}(|D_n|^{1/2})$ from (5.40). Hence,

$$\begin{aligned}\mathbf{T}_{1,n} &= O(|D_n|^{-1/2})o_{\mathrm{P}}(1)\Big(O_{\mathrm{P}}(|D_n|^{1/2}) + o_{\mathrm{P}}(|D_n|^{1/2})\Big) = o_{\mathrm{P}}(1), \\ \mathbf{T}_{2,n} &= \{\mathbf{B}_{n,11}(w;\boldsymbol{\beta}_0) + \mathbf{C}_{n,11}(w;\boldsymbol{\beta}_0)\}^{-1/2}\mathbf{U}_{n,1}(w;\boldsymbol{\beta}_0) + o_{\mathrm{P}}(1).\end{aligned}$$

Thus, the result is proved from Lemma 5.6.2 and Slutsky's theorem.

□

CHAPTER 6

Summary and future extensions

6.1 Summary

This thesis focuses to develop feature selection procedures for estimating the intensity of spatial point processes which depends on spatial covariates. In Chapter 3, we adopt the lasso-type procedures based on convex and non-convex regularization techniques to perform variable selection and improve the prediction. In particular, we regularize by a penalty function the estimating equations based on Campbell theorem derived from the Poisson and logistic regression likelihoods. We show that when the observation domain goes to $\mathbb{R}^d$, the estimates obtained from such procedures are consistent, sparse and asymptotically normal if an appropriate regularization method is chosen such that the tuning parameter follows an appropriate rate. These results indicate that not only do our methods select automatically and consistently the important covariates, but also produce estimates which are as efficient as the estimates where the true set of covariates is known in advance. In Chapter 4, we liberate the assumption required in Chapter 3, which deals with a finite number of covariates, such that the number of parameters to estimate diverges as the observation region increases. We prove that the attractive properties gained in Chapter 3 are still valid for a diverging number of parameters setting, with a few additional assumptions to impose mainly requiring that the number of parameters does not go too fast with respect to the volume of the observation domain.

From a computational point of view regarding the methods developed in Chap-

ters 3 and 4, we make links between spatial point processes intensity estimation and generalized linear models (GLMs) previously proposed by Berman and Turner (1992); Waagepetersen (2008); Baddeley et al. (2014), so we only have to deal with feature selection procedures for GLMs. Hence, we clearly have many advantages: the existing `R` packages for variable selection for GLMs are available, have been carefully studied, easy to implement and computationally fast. This makes our approaches are highly applicable. In particular, we combine the `spatstat` `R` package (Baddeley et al., 2015) and the two `R` packages: `glmnet` (Friedman et al., 2010) and `ncvreg` (Breheny and Huang, 2011).

We conduct simulation experiments to assess the finite sample properties of our estimates. From the results, we have the following conclusions. First, among the regularization methods considered in Chapters 3 and 4, the methods which satisfy the sparsity and the asymptotic normality (i.e., adaptive lasso, adaptive elastic net, SCAD and MC+) perform better than the ones which cannot satisfy (i.e., ridge, lasso and elastic net). We consider that the total number of covariates used in Chapter 3 is 20 (and 15) while we decide to use 50 covariates for the simulation study set in Chapter 4. Adaptive lasso seems to perform best in the situations where the number of covariates is not too large such as the ones considered in the Chapter 3. However, when the number of covariates increases considerably, adaptive elastic net appears to be slightly more competitive than adaptive lasso, mainly because of the presence of the strong multicollinearity and the complex spatial structure of the covariates. Second, the Poisson likelihood (PL) and the weighted Poisson likelihood (WPL) perform differently depending on the situations where they are applied. In general, the PL produces less biased estimates but less efficient estimates than that of the WPL. The WPL gains smaller RMSE, in general, for clustered point processes. Third, the Poisson estimates and the logistic regression likelihood estimates (combined with adaptive lasso) perform quite similarly, especially when the number of dummy points used to approximate the integral term in the likelihood can be chosen to be either similar to or larger than the number of data points.

We apply our methods to model the intensity of *Beilschmiedia pendula Lauraceae* as

a log-linear function of 15 covariates in Chapter 3 consisting of 2 topological attributes and 13 soil nutrients, which is extended in Chapter 4 by adding the interactions between two soil properties such that we have 93 covariates in total. Among 93 covariates, we have 10 covariates which are selected commonly by each method, for which the 5 (out of 15) covariates selected in the application considered in Chapter 3 are also included in the 10 common selected covariates chosen in the application considered in Chapter 4. Furthermore, based on the magnitude of the coefficients, we find six spatial covariates which may have a high influence to the appearance of BPL trees, including two topological attributes: elevation and slope and four soil nutrients: Copper, Phosphorus, Zinc and the interaction between Magnesium and Phosphorus.

In Chapter 5, we study another feature selection procedure based on the development of the Dantzig selector-type methods. In particular, we propose the adaptive linearized Dantzig selector (ALDS) which is based on the linear approximation on the constraint vector. Compared to the ones without linearization, this approach leads to two advantages: in theoretical and computational aspects. In theory, the asymptotic results can be derived more easily than that without linearisation in the constraint vector. In practice, the optimization problem can be solved by linear programming. We find that the asymptotic properties of the estimates obtained from ALDS are similar to the ones we develop in Chapter 3 when adaptive lasso is considered. This may not be surprising since the similarities of the theoretical properties between the estimates of the Dantzig selector and the lasso have already been discovered in other contexts (see e.g., Meinshausen et al., 2007; Bickel et al., 2009; Dicker, 2010). In the simulation study, ALDS performs slightly better and looks more stable than adaptive lasso. In this thesis, we do not compare rigorously the computational cost between ALDS and adaptive lasso for spatial point processes intensity estimation. However, to have an intuition, James et al. (2009) developed an algorithm to compute the Dantzig selector solution for linear models which is quite similar to the one we propose in this study. The algorithm proposed by James et al. (2009) has same computational cost as LARS algorithm (Efron et al., 2004), a known algorithm to compute the entire path lasso solutions by cost $O(Mp^2)$, which also has the same computational cost as the least

squares fit, where M is the number of observations in the linear models' context. As we employ coordinate descent algorithm (Friedman et al., 2007) in Chapters 3 and 4, which is proved to be computationally faster than LARS algorithm (by cost $O(Mp)$), our ALDS may be slightly more computationally expensive than that developed in Chapters 3 and 4, in a similar conclusion to the comparison between LARS and coordinate descent algorithm (see e.g., Friedman et al., 2010). Even though we have preliminary numerical results for ALDS which are encouraging, one may prefer to implement a faster method such as adaptive lasso albeit it performs slightly under the performance of ALDS. This makes an open research area to develop computationally faster methods while keeping their good performances for the Dantzig-selector type methods.

6.2 Future extensions

This work has the potential to be extended in a number of directions. We draw some examples in the following sections.

6.2.1 Spatio-temporal point processes

We have not covered developments for spatio-temporal point processes in this thesis. In the application considered in this study, we find less interesting reasons to extend our study to the spatio-temporal point processes setting. First, we find that there is no much variation for the maps of *Beilschmiedia pendula Lauraceae* appearances among the different year observations. Second, the available covariates for topological attributes and soil properties are only available in space. Moreover, after personal contact with Prof. Jim Dalling (one of the PIs responsible for collecting and analyzing soil nutrients data, see `http://ctfs.si.edu/webatlas/datasets/bci/soilmaps/BCIsoil.html`), he finds that their results are very similar to the ones conducted more recently by his colleagues.

Note that in other applications such as spatio-temporal incidences of forest fires (e.g., Møller and Díaz-Avalos, 2010; Serra et al., 2014), it may be more useful and

interesting to consider extensions of our study in spatio-temporal point processes setting since such applications also consider a relatively large number of covariates. The intensity function naturally becomes

$$\rho(u,t;\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{z}(u,t)), \tag{6.1}$$

where $\mathbf{z}(u,t) = \{z_1(u,t),\ldots,z_k(u,t)\}^\top$ are the k covariates varying in space and time (see e.g., Diggle, 2013; González et al., 2016). In the previously mentioned applications, it is not easy to obtain the information of covariates in both space and time. Furthermore, it is also sometimes less realistic to consider spatio-temporal covariates. For example, in the application used by Møller and Díaz-Avalos (2010), temperature is more relevant to be treated as temporal covariates and elevation is treated as spatial covariates. In such a way, one possibility is to assume first-order spatio-temporal separability (e.g., Gabriel and Diggle, 2009; Møller and Ghorbani, 2012) :

$$\rho(u,t;\boldsymbol{\beta}) = \rho_1(u;\boldsymbol{\beta}_1)\rho_2(t;\boldsymbol{\beta}_2) = \exp(\boldsymbol{\beta}_1^\top \mathbf{z}(u) + \boldsymbol{\beta}_2^\top \mathbf{z}(t)), \tag{6.2}$$

where $\mathbf{z}(u) = \{z_1(u),\ldots,z_p(u)\}^\top$ and $\mathbf{z}(t) = \{z_1(t),\ldots,z_q(t)\}^\top$ are, respectively, the p spatial covariates observed at location u and the q temporal covariates collected at time t. $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the associated real p and q dimensional parameters which exhibit the effects of both spatial and temporal covariates to the intensity.

By assuming (6.2), the spatio-temporal processes may be regarded using spatial and temporal margins (e.g Møller and Díaz-Avalos, 2010). Thus, parameter estimates can be conducted by separating into Poisson likelihoods for the spatial margin and the temporal margin and then maximizing the corresponding Poisson likelihood. To perform the variable selection, the idea is, first, to regularize by a penalty function the Poisson likelihood of the temporal point process and, second, to apply a regularization technique for the Poisson likelihood of the spatial point process. This seems feasible in term of practical implementation. However, further study is needed regarding the theoretical justification.

6.2.2 Multivariate point processes

In the application we consider in this thesis, we focus to model the intensity of a single species of trees. Since the locations of each of hundreds of species of trees have been collected, a natural extension would consider assessing the hypotheses regarding biodiversity in the forest by involving as many as possible the species of trees in the analysis. In a such analysis, taking the form of the intensity has to be taken carefully in order to be able to embrace the effect of both spatial covariates and the interactions among trees (which can be distinguished into interaction intra-specific: interaction between individuals of the same species and interaction inter-specific: interaction between trees of different species). This leads to not only study the first-order characteristics such as the intensity, but also the second-order characteristics such as cross pair correlation function.

Related recent study has been conducted by Waagepetersen et al. (2016) who modeled the nine species of trees contained in BCI study (e.g., Condit, 1998; Hubbell et al., 2005) by multivariate log-Gaussian Cox processes. Waagepetersen et al. (2016) decomposed the random intensity into three parts which exhibits spatial inhomogeneity due to the dependence on the spatial covariates, correlation across species, and clustering due to species-specific factors such as seed dispersal. However, considerable challenges remain due to the very high number of parameters to estimate, especially when a large number of tree species is considered. Thus, in the similar motivation as of this thesis, similar feature selection procedures seem interesting to study more carefully in order to obtain more interpretable models and practically applicable statistical methods for point patterns with hundreds of types of points.

6.2.3 Tuning parameter selection

In this study, we develop methods whose estimates satisfy some interesting properties: consistency, sparsity, and asymptotic normality. Nevertheless, it is worth noticing that our methods rely on the tuning parameter λ. For practical implementation, we suggest

selecting λ which minimizes WQBIC(λ) defined by

$$\text{WQBIC}(\lambda) = -2\ell(w; \hat{\boldsymbol{\beta}}(\lambda)) + \|\hat{\boldsymbol{\beta}}\|_0 \log |D|, \tag{6.3}$$

where $\ell(w; \hat{\boldsymbol{\beta}}(\lambda))$ is the Poisson or logistic regression likelihoods, $\|\hat{\boldsymbol{\beta}}\|_0$ is the number of selected nonzero coefficients and $|D|$ is the volume of observation domain. It has been proved in the contexts of linear and generalized linear models that BIC-type methods satisfy selection consistency (e.g., Zou et al., 2007; Wang et al., 2007b, 2009; Zhang et al., 2010), meaning that it selects the correct model with probability tending to 1 in large samples when a set of candidate models contains the true model. It may be true that the criterion defined by (6.3) or its slightly modified form can also satisfy the selection consistency in the spatial point processes framework. However, this needs further investigation.

6.2.4 The generalized Dantzig selector

Regarding Chapter 5 in this manuscript, note that the adaptive Dantzig selector (5.11) can be written more generally as

$$\min |D| \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \text{ subject to } |\mathbf{U}_j(w; \boldsymbol{\beta})| \leq |D| p'_{\lambda_j}(|\beta_j|) \text{ for } j = 1, \ldots, p, \tag{6.4}$$

where $p_\lambda(|\theta|)$ is a penalty function which can be a convex or a non-convex function (see Chapter 3 as examples) and $p'_\lambda(|\theta|)$ is its derivative with respect to θ. The theoretical results may be able to derive in a similar way as developped in Chapter 5. However, the computational implementation may be much more challenging due to the possible non-convex optimization problem, if for example, SCAD or MC+ penalties are considered.

The more interesting extension is to consider the combination between l_1 and l_2 penalties in a fashion similar to the adaptive elastic net but adapted to the Dantzig selector setting. More precisely, in (6.4), $p_{\lambda_j}(|\beta_j|) = \lambda_j\{\gamma|\beta_j| + \frac{1}{2}(1-\gamma)\beta_j^2\}$ and $p'_{\lambda_j}(|\beta_j|) = \lambda_j\{\gamma(\text{sign}(\beta_j) - \beta_j)_+ + |\beta_j|\}$ for $j = 1, \cdots, p$, where $0 < \gamma < 1$. In real applications, the setting which considers the involvement of a lot of covariates with complex spatial structure and the presence of strong multicollinearity is easily found.

In such situations, the adaptive elastic net may outperform adaptive lasso as it is the case for linear regression (see e.g., Zou and Zhang, 2009). We have preliminary results, as considered in the simulation experiments in Chapter 4, that the adaptive elastic net is slightly more preferable than the adaptive lasso. As the adaptive linearized Dantzig selector has slightly better results than the ones from the regularization methods with the adaptive lasso, it is interesting to investigate the theoretical and computational aspects of the proposed ideas in the spatial point processes setting, which may also be regarded as an extension of the Dantzig selector-type approaches in a more general context.

Note that we consider a fixed number of parameters setting in Chapter 5. In a similar motivation, as we consider from Chapter 3 to Chapter 4, it seems also possible to extend the Chapter 5 in the situation where the number of parameters diverges.

Bibliography

Anestis Antoniadis, Piotr Fryzlewicz, and Frédérique Letué. The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):531–552, 2010.

M Salman Asif. Primal dual pursuit: A homotopy based algorithm for the Dantzig selector. Master's thesis, Georgia Institute of Technology, 2008.

M Salman Asif and Justin Romberg. On the lasso and dantzig selector equivalence. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1–6. IEEE, 2010.

Adrian Baddeley and David Dereudre. Variational estimators for the parameters of Gibbs point process models. *Bernoulli*, 19(3):905–930, 2013.

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

Adrian Baddeley and Rolf Turner. Spatstat: An R package for analyzing spatial point pattens. *Journal of Statistical Software*, 12(6):1–42, 2005.

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.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Erwin Bolthausen. On the central limit theorem for stationary mixing random fields. *The Annals of Probability*, 10(4):1047–1050, 1982.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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.

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.

Lee Dicker. *Regularized regression methods for variable selection and estimation*. PhD thesis, Harvard University, 2010.

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)*, pages 349–362, 1990.

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

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697, 2009.

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 Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

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

Janet Franklin. *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, 2010.

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. 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.

Edith Gabriel and Peter J Diggle. Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63(1):43–51, 2009.

Jonatan A González, Francisco J Rodríguez-Cortés, Ottmar Cronie, and Jorge Mateu. Spatio-temporal point process statistics: a review. *Spatial Statistics*, 18:505–544, 2016.

Yongtao Guan and Ji Meng Loh. A thinned block bootstrap variance estimation procedure for inhomogeneous spatial point patterns. *Journal of the American Statistical Association*, 102(480):1377–1386, 2007.

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

Yongtao Guan, Abdollah Jalilian, and Rasmus Plenge Waagepetersen. Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):677–697, 2015.

Xavier Guyon. *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media, 1995.

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*, pages 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.

Gareth M James and Peter Radchenko. A generalized Dantzig selector with shrinkage tuning. *Biometrika*, 96(2):323–337, 2009.

Gareth M James, Peter Radchenko, and Jinchi Lv. DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):127–142, 2009.

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

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):853–877, 2015.

Yi Li, Lee Dicker, and Sihai Dave Zhao. The Dantzig selector for censored linear regression models. *Statistica Sinica*, 24(1):251, 2014.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

Nicolai Meinshausen, Guilherme Rocha, and Bin Yu. Discussion: A tale of three cousins: Lasso, l2boosting and Dantzig. *The Annals of Statistics*, 35(6):2373–2384, 2007.

Jesper Møller and Carlos Díaz-Avalos. Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires. *Scandinavian Journal of Statistics*, 37(1):2–25, 2010.

Jesper Møller and Mohammad Ghorbani. Aspects of second-order analysis of structured inhomogeneous spatio-temporal point processes. *Statistica Neerlandica*, 66(4):472–491, 2012.

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

Jesper Møller and Rasmus Plenge Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

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*, pages 1298–1309, 1984.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

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

Ian W Renner. *Advances in presence-only methods in ecology*. PhD thesis, University of New South Wales, 2013.

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.

Patricia Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126(1):103–153, 2003.

Frederic Paik Schoenberg. Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference*, 128 (1):79–93, 2005.

Laura Serra, Marc Saez, Jorge Mateu, Diego Varga, Pablo Juan, Carlos Díaz-Ávalos, and Håvard Rue. Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of catalonia, 1994–2008. *Environmental and ecological statistics*, 21(3):531–563, 2014.

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.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

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.

Rasmus Plenge 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.

Rasmus Plenge Waagepetersen, Yongtao Guan, Abdollah Jalilian, and Jorge Mateu. Analysis of multispecies point patterns by using multivariate log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1): 77–96, 2016.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007a.

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

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.

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Yu Ryan Yue and Ji Meng Loh. Variable selection for inhomogeneous spatial point process models. *Canadian Journal of Statistics*, 43(2):288–305, 2015.

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

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 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 Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.

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, Robert Tibshirani, et al. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.