

Quasi-likelihood for Spatial Point Processes

Yongtao Guan

Miami, USA

Abdollah Jalilian

Kermanshah, Iran

Rasmus Waagepetersen†

Aalborg, Denmark

Summary.

Fitting regression models for intensity functions of spatial point processes is of great interest in ecological and epidemiological studies of association between spatially referenced events and geographical or environmental covariates. When Cox or cluster process models are used to accommodate clustering not accounted for by the available covariates, likelihood based inference becomes computationally cumbersome due to the complicated nature of the likelihood function and the associated score function. It is therefore of interest to consider alternative more easily computable estimating functions. We derive the optimal estimating function in a class of first-order estimating functions. The optimal estimating function depends on the solution of a certain Fredholm integral equation which in practise is solved numerically. The derivation of the optimal estimating function has close similarities to the derivation of quasi-likelihood for standard data sets. The approximate solution is further equivalent to a quasi-likelihood score for binary spatial data. We therefore use the term quasi-likelihood for our optimal estimating function approach. We demonstrate in a simulation study and a data example that our quasi-likelihood method for spatial point processes is both statistically and computationally efficient.

Keywords: Estimating function, Fredholm integral equation, Godambe information, Intensity function, Regression model, Spatial point process.

1. INTRODUCTION

In many applications of spatial point processes it is of interest to fit a regression model for the intensity function. In case of a Poisson point process, maximum likelihood estimation of regression parameters is rather straightforward with a user-friendly implementation available in the R package `spatstat` (Baddeley and Turner, 2005). However, if e.g. Cox or cluster point process models are used to accommodate clustering not explained by a Poisson process, then maximum likelihood estimation is in general difficult from a computational point of view (see e.g. Møller and Waagepetersen, 2004). Alternatively, one may follow composite likelihood arguments (e.g. Møller and Waagepetersen, 2007) to obtain an estimating function that is equivalent to the score of the Poisson likelihood function. This provides a computationally tractable estimating function and theoretical properties of the

†*Address for correspondence:* Department of Mathematical Sciences, Aalborg University, Fredrik Bajersvej 7G, DK-9220 Aalborg, Denmark
E-mail: rw@math.aau.dk

resulting estimator are well understood, see e.g. Schoenberg (2005), Waagepetersen (2007) and Guan and Loh (2007).

A drawback of the Poisson score function approach is the loss of efficiency since possible dependence between points is ignored. In the context of intensity estimation, it appears that only Mrkvička and Molchanov (2005) and Guan and Shen (2010) have tried to incorporate second-order properties in the estimation so as to improve efficiency. Mrkvička and Molchanov (2005) show that their proposed estimator is optimal among a class of linear, unbiased intensity estimators, where the word ‘optimal’ refers to minimum variance. However, their approach is confined to a very restrictive type of intensity function known up to a one-dimensional scaling factor. In contrast, Guan and Shen (2010) propose a weighted estimating equation approach that is applicable to intensity functions in more general forms. A similar optimality result can on the other hand not be established for their approach.

In this paper we derive an optimal estimating function that not only takes into account possible spatial correlation but also is applicable for point processes with a general regression model for the intensity function. In the spirit of generalized linear models the intensity is given by a differentiable function of a linear predictor depending on spatial covariates. The optimal estimating function depends on the solution of a certain Fredholm integral equation and reduces to the likelihood score in case of a Poisson process. We show in Section 3.2 that the optimality result in Mrkvička and Molchanov (2005) is a special case of our more general result, and that the estimation method in Guan and Shen (2010) is only a crude approximation of our new approach. Apart from being computationally efficient, our estimating function only requires specification of the intensity function and the so-called pair correlation function, which is another advantage compared with maximum likelihood estimation.

For many types of correlated data other than spatial point patterns, estimating functions have been widely used for model fitting when maximum likelihood estimation is computationally challenging. Examples of such data include longitudinal data (Liang and Zeger, 1986), time series data (Zeger, 1988), clustered failure time data (Gray, 2003) and spatial binary or count data (Gotway and Stroup, 1997; Lin and Clayton, 2005). For most of these methods, the inverse of a covariance matrix is used to account for the correlation in data, and optimality can be established when the so-called quasi-score estimating functions are used (Wedderburn, 1974; Heyde, 1997). For a quasi-score estimating function the inverse covariance matrix contributes to an optimal linear transformation of the residual vector. For point processes, an analogue of residuals is given by the so-called residual measure and our optimal estimating function can be viewed as an optimal linear transformation of the residual measure. Moreover, the numerical implementation of our method is closely related to the quasi-likelihood for spatial data considered in Gotway and Stroup (1997) and Lin and Clayton (2005). Our work hence not only lays the theoretical foundation for optimal intensity estimation, but also fills in a critical gap between existing literature on spatial point processes and the well-established quasi-likelihood estimation method. We therefore adopt the term quasi-likelihood for our approach.

Following some background material on point processes and estimating functions, we derive our optimal quasi-likelihood score estimating function and discuss the practical implementation of it based on a numerical solution of the Fredholm integral equation. Asymptotic properties of the resulting parameter estimator is then considered and the superior performance of the quasi-likelihood method compared with existing ones is demonstrated through a simulation study. We finally illustrate the practical use of the quasi-likelihood in a data example of three tropical tree species.

2. BACKGROUND

In this section we provide background on the intensity and pair correlation function of a spatial point process and we state the basic assumptions on these needed for our quasi-likelihood method. Also reviews on composite likelihood estimation and estimating functions are provided. Throughout the presentation, we use \mathbb{E} , Var and Cov to denote expectation, variance and covariance, respectively.

2.1. Basic Assumptions on the Intensity and Pair Correlation Function

Let X be a point process on \mathbb{R}^2 and let $N(B)$ denote the number of points in $X \cap B$ for any bounded set $B \subset \mathbb{R}^2$. We assume that X has an intensity function $\lambda(\cdot)$ and a pair correlation function $g(\cdot, \cdot)$ whereby the first- and second-order moments of the counts $N(B)$ are given by

$$\mathbb{E}N(B) = \int_B \lambda(\mathbf{u})d\mathbf{u} \quad (1)$$

and

$$\text{Cov}[N(A), N(B)] = \int_{A \cap B} \lambda(\mathbf{u})d\mathbf{u} + \int_A \int_B \lambda(\mathbf{u})\lambda(\mathbf{v})[g(\mathbf{u}, \mathbf{v}) - 1]d\mathbf{u}d\mathbf{v} \quad (2)$$

for bounded sets $A, B \subseteq \mathbb{R}^2$ (Møller and Waagepetersen, 2004).

We assume that the intensity function is given in terms of a parametric model $\lambda(\mathbf{u}) = \lambda(\mathbf{u}; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ is a vector of regression parameters. The intensity function is further assumed to be positive and differentiable with respect to $\boldsymbol{\beta}$ with gradient $\boldsymbol{\lambda}'(\mathbf{u}; \boldsymbol{\beta}) = d\lambda(\mathbf{u}; \boldsymbol{\beta})/d\boldsymbol{\beta}$. A popular example is the log linear model $\log \lambda(\mathbf{u}; \boldsymbol{\beta}) = \mathbf{z}(\mathbf{u})\boldsymbol{\beta}^\top$, where $\mathbf{z}(\mathbf{u}) = (z_1(\mathbf{u}), \dots, z_p(\mathbf{u}))$ is a covariate vector for each $\mathbf{u} \in \mathbb{R}^2$. For convenience of exposition we assume that $g(\mathbf{u}, \mathbf{v})$ only depends on the difference $\mathbf{u} - \mathbf{v}$ since this is the common assumption in practise. This implies that X becomes second-order re-weighted stationary (Baddeley et al., 2000). In the following we thus let $g(\mathbf{r})$ denote the pair correlation function for two points \mathbf{u} and \mathbf{v} with $\mathbf{u} - \mathbf{v} = \mathbf{r}$. However, our proposed optimal estimating function is applicable also in the case of a non-translation invariant pair correlation function. For the derivations in Section 3 we further need that $g(\cdot)$ is continuous and that $g(\cdot) - 1$ is a non-negative definite function.

For given functions $\lambda(\cdot)$ and $g(\cdot)$, in addition to the assumptions specified above, it is of course required that these functions are indeed respectively an intensity function and a pair correlation function of some spatial point process. To be an intensity function, $\lambda(\cdot)$ just needs to be non-negative and integrable. We are not aware of simple necessary and sufficient conditions that ensure that a function $g(\cdot)$ is a pair correlation function. We hence restrict attention to functions $g(\cdot)$ which conform with pair correlation functions of existing point process models. From a practical point of view, we also need a computationally tractable expression for $g(\cdot)$. This precludes pair correlation functions of Markov and Gibbs point processes (e.g. Møller and Waagepetersen, 2007). On the other hand, a wide range of shot-noise Cox processes, log Gaussian Cox processes and Poisson cluster processes have pair correlation functions given in closed form (Møller and Waagepetersen, 2004) and satisfying the assumptions stated in the previous paragraph (see also Section 3.1 regarding the condition of non-negative definiteness).

2.2. Composite Likelihood

A first-order log composite likelihood function (Schoenberg, 2005; Waagepetersen, 2007) for estimation of β is given by

$$\sum_{\mathbf{u} \in X \cap W} \log \lambda(\mathbf{u}; \beta) - \int_W \lambda(\mathbf{u}; \beta) d\mathbf{u}, \quad (3)$$

where $W \subset \mathbb{R}^2$ is the observation window. This can be viewed as a limit of log composite likelihood functions for binary variables $Y_i = 1[N(B_i) > 0]$, $i = 1, \dots, m$, where the cells B_i form a disjoint partitioning of W and $1[\cdot]$ is an indicator function (e.g. Møller and Waagepetersen, 2007). The limit is obtained when the number of cells tends to infinity and the areas of the cells tend to zero. In case of a Poisson process, the composite likelihood function coincides with the likelihood function.

The composite likelihood is computationally simple and enjoys considerable popularity in particular in studies of tropical rain forest ecology where spatial point process models are fitted to spatial point pattern data sets of locations of thousands of rain forest trees (see e.g. Shen et al., 2009; Lin et al., 2011; Renner and Warton, 2013). However, it is not statistically efficient for non-Poisson data since possible correlations between counts of points are ignored.

2.3. Primer on Estimating Functions and Quasi-likelihood

Referring to the previous Section 2.2, the composite likelihood estimator of β is obtained by maximizing the log composite likelihood (3). This is equivalent to solving the following equation:

$$\mathbf{e}(\beta) = \mathbf{0}, \quad (4)$$

where

$$\mathbf{e}(\beta) = \sum_{\mathbf{u} \in X \cap W} \frac{\lambda'(\mathbf{u}; \beta)}{\lambda(\mathbf{u}; \beta)} - \int_W \lambda'(\mathbf{u}; \beta) d\mathbf{u}. \quad (5)$$

Equations in the form of (4) are typically referred to as estimating equations and functions like $\mathbf{e}(\beta)$ are called estimating functions (Heyde, 1997). Note that many other statistical estimation procedures, such as maximum likelihood estimation, moment based estimation and minimum contrast estimation, can all be written in terms of estimating functions.

We defer rigorous asymptotic details to Section 5 and here just provide an informal overview of properties of an estimator $\hat{\beta}$ based on an estimating function $\mathbf{e}(\beta)$. By a first-order Taylor series expansion at $\hat{\beta}$,

$$\mathbf{e}(\beta) \approx \mathbf{e}(\hat{\beta}) + [\hat{\beta} - \beta] \mathbf{S} = (\hat{\beta} - \beta) \mathbf{S},$$

where $\mathbf{S} = -\mathbb{E}d\mathbf{e}(\beta)/d\beta^T$ is the so-called sensitivity matrix (e.g. page 62 in Song, 2007) and the equality is due to $\mathbf{e}(\hat{\beta}) = \mathbf{0}$ as required by (4). It then follows immediately that $\hat{\beta} \approx \beta + \mathbf{e}(\beta) \mathbf{S}^{-1}$. Thus, with β equal to the true parameter value, $\hat{\beta}$ is approximately unbiased if $\mathbb{E}\mathbf{e}(\beta) = \mathbf{0}$, i.e. $\mathbf{e}(\beta)$ is an unbiased estimating function. Moreover, $\text{Var}\hat{\beta} \approx \mathbf{S}^{-1} \mathbf{\Sigma} \mathbf{S}^{-1}$ where $\mathbf{\Sigma} = \text{Vare}(\beta)$ and $\mathbf{S}^{-1} \mathbf{\Sigma} \mathbf{S}^{-1}$ is the asymptotic covariance matrix when the size of the data set goes to infinity in a suitable manner (Section 5). The inverse of $\mathbf{S}^{-1} \mathbf{\Sigma} \mathbf{S}^{-1}$, i.e. $\mathbf{S} \mathbf{\Sigma}^{-1} \mathbf{S}$, is called the Godambe information (e.g. Definition 3.7 in Song, 2007).

Suppose that two competing estimating functions $\mathbf{e}_1(\boldsymbol{\beta})$ and $\mathbf{e}_2(\boldsymbol{\beta})$ with respective Godambe informations \mathbf{I}_1 and \mathbf{I}_2 are used to obtain the estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. Then $\mathbf{e}_1(\boldsymbol{\beta})$ is said to be superior to $\mathbf{e}_2(\boldsymbol{\beta})$ if $\mathbf{I}_1 - \mathbf{I}_2$ is positive definite, since this essentially means that $\hat{\boldsymbol{\beta}}_1$ has a smaller asymptotic variance than $\hat{\boldsymbol{\beta}}_2$. If $\mathbf{I}_1 - \mathbf{I}_2$ is positive definite for all possible $\mathbf{e}_2(\boldsymbol{\beta})$, then we say that $\mathbf{e}_1(\boldsymbol{\beta})$ has the maximal Godambe information and is an optimal estimating function. The resulting estimator $\hat{\boldsymbol{\beta}}_1$ is then the asymptotically most efficient.

Consider an m -dimensional data vector \mathbf{Y} with covariance matrix \mathbf{V} and mean vector $\boldsymbol{\mu}$, a differentiable function of some p -dimensional parameter vector $\boldsymbol{\beta}$. Let $\mathbf{D} = d\boldsymbol{\mu}^\top/d\boldsymbol{\beta}$ be the $m \times p$ matrix of partial derivatives $d\mu_i/d\beta_j$. The quasi-likelihood score function is then

$$(\mathbf{Y} - \boldsymbol{\mu})\mathbf{V}^{-1}\mathbf{D}, \quad (6)$$

which is optimal among all estimating functions of the form $(\mathbf{Y} - \boldsymbol{\mu})\mathbf{A}$ for some $m \times p$ matrix \mathbf{A} (e.g. Heyde, 1997).

3. AN OPTIMAL FIRST-ORDER ESTIMATING EQUATION

The estimating function given in (5) can be rewritten as

$$\mathbf{e}_f(\boldsymbol{\beta}) = \sum_{\mathbf{u} \in X \cap W} \mathbf{f}(\mathbf{u}) - \int_W \mathbf{f}(\mathbf{u})\lambda(\mathbf{u}; \boldsymbol{\beta})d\mathbf{u}, \quad (7)$$

where $\mathbf{f}(\mathbf{u}) = \boldsymbol{\lambda}'(\mathbf{u}; \boldsymbol{\beta})/\lambda(\mathbf{u}; \boldsymbol{\beta})$. In general, $\mathbf{f}(\mathbf{u})$ can be any $1 \times p$ real vector valued function, where p is the dimension of $\boldsymbol{\beta}$. We call (7) a first-order estimating function. Our aim is to find a function ϕ so that \mathbf{e}_ϕ is optimal within the class of first-order estimating functions; in other words, the resulting estimator of $\boldsymbol{\beta}$ associated with \mathbf{e}_ϕ is asymptotically most efficient.

The estimating function (7) can be further re-expressed in terms of the residual measure (Baddeley et al., 2005; Waagepetersen, 2005) defined for bounded $B \subset \mathbb{R}^2$ as

$$R(B) = \sum_{\mathbf{u} \in X \cap W} 1[\mathbf{u} \in B] - \int_B \lambda(\mathbf{u}; \boldsymbol{\beta})d\mathbf{u}.$$

Thus $\mathbf{e}_f(\boldsymbol{\beta}) = \int_W \mathbf{f}(\mathbf{u})R(d\mathbf{u})$ so our estimating function can be viewed as a linear transformation of the residual measure. Hence just like the quasi-likelihood score for ordinary numerical data is the optimal linear transformation of the residual vector (Section 2.3), our optimal estimating function will be the optimal linear transformation of the residual measure.

Let $\boldsymbol{\Sigma}_f = \text{Var}\mathbf{e}_f(\boldsymbol{\beta})$, $\mathbf{J}_f = -d\mathbf{e}_f(\boldsymbol{\beta})/d\boldsymbol{\beta}^\top$ and $\mathbf{S}_f = \mathbb{E}\mathbf{J}_f$. Note that $\boldsymbol{\Sigma}_f$, \mathbf{J}_f and \mathbf{S}_f all depend on $\boldsymbol{\beta}$ but we suppress the dependence on $\boldsymbol{\beta}$ in this section for ease of presentation. Recalling the definition of optimality in Section 2.3, for \mathbf{e}_ϕ to be optimal we must have that

$$\mathbf{S}_\phi \boldsymbol{\Sigma}_\phi^{-1} \mathbf{S}_\phi - \mathbf{S}_f \boldsymbol{\Sigma}_f^{-1} \mathbf{S}_f \quad (8)$$

is non-negative definite for all $\mathbf{f} : W \rightarrow \mathbb{R}^p$. A sufficient condition for this is

$$\boldsymbol{\Sigma}_{\phi f} = \mathbf{S}_f \quad (9)$$

for all \mathbf{f} where $\boldsymbol{\Sigma}_{\phi f} = \text{Cov}[\mathbf{e}_\phi(\boldsymbol{\beta}), \mathbf{e}_f(\boldsymbol{\beta})]$. To understand the intuition behind (9), view \mathbf{e}_ϕ as the score function from maximum likelihood estimation (MLE) and \mathbf{e}_f as an arbitrary

unbiased estimating function not necessarily in the form of (7). It is then well known that (9) holds and in fact leads to the optimality of the MLE score function among all unbiased estimating functions. In our setting, (9) therefore suggests that \mathbf{e}_ϕ plays the role of the MLE score function and is expected to be optimal within the class of first-order estimating functions \mathbf{e}_f defined by (7). This type of condition is also provided in Theorem 2.1 in Heyde (1997) for both discrete and continuous vector-valued data. In Appendix A, we give a short self-contained proof of the sufficiency of (9) in our setting.

By the Campbell formulae (e.g. Møller and Waagepetersen, 2004, Chapter 4),

$$\begin{aligned}\Sigma_{\phi f} &= \int_W \mathbf{f}^\top(\mathbf{u})\phi(\mathbf{u})\lambda(\mathbf{u};\boldsymbol{\beta})d\mathbf{u} + \int_{W^2} \mathbf{f}^\top(\mathbf{u})\phi(\mathbf{v})\lambda(\mathbf{u};\boldsymbol{\beta})\lambda(\mathbf{v};\boldsymbol{\beta})[g(\mathbf{u}-\mathbf{v})-1]d\mathbf{u}d\mathbf{v}, \\ \mathbf{S}_f &= \int_W \mathbf{f}^\top(\mathbf{u})\lambda'(\mathbf{u};\boldsymbol{\beta})d\mathbf{u}.\end{aligned}$$

Hence, (9) is equivalent to

$$\int_W \mathbf{f}^\top(\mathbf{u})\left\{\lambda'(\mathbf{u};\boldsymbol{\beta}) - \phi(\mathbf{u})\lambda(\mathbf{u};\boldsymbol{\beta}) - \lambda(\mathbf{u};\boldsymbol{\beta}) \int_W \phi(\mathbf{v})\lambda(\mathbf{v};\boldsymbol{\beta})[g(\mathbf{u}-\mathbf{v})-1]d\mathbf{v}\right\}d\mathbf{u} = \mathbf{0}$$

for all $\mathbf{f} : W \rightarrow \mathbb{R}^p$, which is true if

$$\lambda'(\mathbf{u};\boldsymbol{\beta}) - \phi(\mathbf{u})\lambda(\mathbf{u};\boldsymbol{\beta}) - \lambda(\mathbf{u};\boldsymbol{\beta}) \int_W \phi(\mathbf{v})\lambda(\mathbf{v};\boldsymbol{\beta})[g(\mathbf{u}-\mathbf{v})-1]d\mathbf{v} = 0 \quad (10)$$

for all $\mathbf{u} \in W$. Since $\lambda > 0$, (10) implies that ϕ is a solution to the Fredholm integral equation (e.g. Hackbusch, 1995, Chapter 3)

$$\phi = \frac{\lambda'}{\lambda} - \mathbf{T}\phi, \quad (11)$$

where \mathbf{T} is the operator given by

$$(\mathbf{T}\mathbf{f})(\mathbf{u}) = \int_W t(\mathbf{u}, \mathbf{v})\mathbf{f}(\mathbf{v})d\mathbf{v} \quad \text{with } t(\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{v};\boldsymbol{\beta})[g(\mathbf{u}-\mathbf{v})-1]. \quad (12)$$

By continuity of g , \mathbf{T} is compact in the space of continuous functions on W (Hackbusch, 1995, Theorem 3.2.5). Moreover, -1 is not an eigenvalue of \mathbf{T} since $g(\cdot) - 1$ is positive definite (Section 3.1). It then follows by Theorem 3.2.1 in Hackbusch (1995) that (11) has a unique solution

$$\phi = (\mathbf{I} + \mathbf{T})^{-1} \frac{\lambda'}{\lambda},$$

where \mathbf{I} is the identity operator (or, depending on context, the identity matrix) and $(\mathbf{I} + \mathbf{T})^{-1}$ is the bounded linear inverse of $\mathbf{I} + \mathbf{T}$. We define

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{e}_\phi(\boldsymbol{\beta}) = \sum_{\mathbf{u} \in X \cap W} \phi(\mathbf{u}) - \int_W \phi(\mathbf{u})\lambda(\mathbf{u};\boldsymbol{\beta})d\mathbf{u}, \quad (13)$$

$$\boldsymbol{\Sigma} = \mathbb{V}\text{are}(\boldsymbol{\beta}), \quad \mathbf{J} = -d\mathbf{e}(\boldsymbol{\beta})/d\boldsymbol{\beta}^\top, \quad \mathbf{S} = \mathbb{E}\mathbf{J},$$

where by the above derivations,

$$\mathbf{S} = \boldsymbol{\Sigma} = \int_W \phi^\top(\mathbf{u})\lambda'(\mathbf{u};\boldsymbol{\beta})d\mathbf{u}. \quad (14)$$

In the Poisson process case where $g(\cdot) = 1$, (13) reduces to the Poisson likelihood score (5).

We develop a more explicit expression for ϕ by using Neumann series expansion in Appendix B. The Neumann series expansion is also useful for checking the conditions for our asymptotic results; see Section 1 in the supplementary material. However, it is not essential for our approach so we omit the detailed discussion here.

3.1. Condition for non-negative eigenvalues of \mathbf{T}

In general it is difficult to assess the eigenvalues of \mathbf{T} given by (12). However, since $g - 1$ is non-negative definite, \mathbf{T}^s is a positive operator (i.e., $\int_W \mathbf{f}^\top(\mathbf{u})(\mathbf{T}^s \mathbf{f})(\mathbf{u}) d\mathbf{u} \geq 0$) where \mathbf{T}^s is given by the symmetric kernel $t^s(\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}; \boldsymbol{\beta})^{1/2} \lambda(\mathbf{v}; \boldsymbol{\beta})^{1/2} [g(\mathbf{u} - \mathbf{v}) - 1]$. Then all eigenvalues of \mathbf{T}^s are non-negative (Lax, 2002, Corollary 1, p. 320). In particular, -1 is not an eigenvalue. The same holds for \mathbf{T} since it is easy to see that the eigenvalues of \mathbf{T} coincide with those of \mathbf{T}^s .

For the wide class of second-order re-weighted stationary Cox point processes, $g(\mathbf{r}) = 1 + \text{Cov}[\Lambda(\mathbf{u}), \Lambda(\mathbf{u} + \mathbf{r})] / [\lambda(\mathbf{u})\lambda(\mathbf{u} + \mathbf{r})]$ where Λ denotes the random intensity function of the Cox process. Hence $g(\cdot) - 1$ is non-negative definite for this class of processes.

3.2. Relation to Existing Methods

Guan and Shen (2010) consider a subset of first-order estimating functions of the form

$$\sum_{\mathbf{u} \in X \cap W} w(\mathbf{u}) \frac{\boldsymbol{\lambda}'(\mathbf{u}; \boldsymbol{\beta})}{\lambda(\mathbf{u}; \boldsymbol{\beta})} - \int_W w(\mathbf{u}) \boldsymbol{\lambda}'(\mathbf{u}; \boldsymbol{\beta}) d\mathbf{u} \quad (15)$$

obtained by introducing a weight function $w(\cdot)$ in the composite likelihood estimating function (5). They then seek to minimize the parameter estimation variance with respect to $w(\cdot)$ and obtain an approximate solution of this minimization problem. In contrast, our estimating function (13) is optimal among all first-order estimating functions including those of the form (15).

An approximate version of our estimating function coincides with the one obtained by Guan and Shen (2010). This follows by approximating the operator \mathbf{T} by

$$(\mathbf{T}\mathbf{f})(\mathbf{u}) = \int_W \mathbf{f}(\mathbf{v}) \lambda(\mathbf{v}; \boldsymbol{\beta}) [g(\mathbf{u} - \mathbf{v}) - 1] d\mathbf{v} \approx \lambda(\mathbf{u}; \boldsymbol{\beta}) \mathbf{f}(\mathbf{u}) \int_W [g(\mathbf{u} - \mathbf{v}) - 1] d\mathbf{v}. \quad (16)$$

This is justified if $\mathbf{f}(\mathbf{v}) \lambda(\mathbf{v}; \boldsymbol{\beta})$ is close to $\mathbf{f}(\mathbf{u}) \lambda(\mathbf{u}; \boldsymbol{\beta})$ for the \mathbf{v} where $g(\mathbf{u} - \mathbf{v}) - 1$ differs substantially from zero. Then the Fredholm integral equation (11) can be approximated by $\phi = \frac{\boldsymbol{\lambda}'}{\lambda} - \lambda A \phi$, where

$$A(\mathbf{u}) = \int_W [g(\mathbf{u} - \mathbf{v}) - 1] d\mathbf{v}. \quad (17)$$

We hence obtain an approximate solution $\phi = w \boldsymbol{\lambda}' / \lambda$ with $w(\mathbf{u}) = [1 + \lambda(\mathbf{u}; \boldsymbol{\beta}) A(\mathbf{u})]^{-1}$. Using this approximation in (13) the resulting estimating function is precisely of the form (15) suggested by Guan and Shen (2010).

Mrkvička and Molchanov (2005) derived optimal intensity estimators in the situation of $\lambda(\mathbf{u}; \rho) = \rho \gamma(\mathbf{u})$ for some known function $\gamma(\mathbf{u})$ and unknown parameter $\rho > 0$. Since ρ is the only unknown parameter, a direct application of (11) yields

$$\rho \phi(\mathbf{u}) + \rho^2 \int_W \phi(\mathbf{v}) \gamma(\mathbf{v}) [g(\mathbf{u} - \mathbf{v}) - 1] d\mathbf{v} = 1,$$

which is essentially Corollary 3.1 of Mrkvička and Molchanov (2005). It is uncommon for an intensity function to be known up to a one-dimensional scaling factor. In contrast, our proposed modeling framework for the intensity function closely mimics that used in classical regression analysis and is more general. As a result, our method of derivation is completely different from that in Mrkvička and Molchanov (2005).

4. IMPLEMENTATION

In this section we discuss practical issues concerning the implementation of our proposed optimal estimating function. In particular we show in Section 4.2 that a particular numerical approximation of our optimal estimating function is equivalent to a quasi-likelihood for binary spatial data for which an iterative generalized least squares solution can be implemented. An R implementation `ql.ppm()` will appear in future releases of `spatstat` (Baddeley and Turner, 2005).

4.1. Numerical Approximation

To estimate ϕ , consider the numerical approximation

$$(\mathbf{T}\phi)(\mathbf{u}) = \int_W t(\mathbf{u}, \mathbf{v})\phi(\mathbf{v})d\mathbf{v} \approx \sum_{i=1}^m t(\mathbf{u}, \mathbf{u}_i)\phi(\mathbf{u}_i)w_i, \quad (18)$$

where $\mathbf{u}_i, i = 1, \dots, m$, are quadrature points with associated weights w_i . Inserting this approximation in (11) with $\mathbf{u} = \mathbf{u}_l$ we obtain estimates $\hat{\phi}(\mathbf{u}_l)$ of $\phi(\mathbf{u}_l)$, $l = 1, \dots, m$, by solving the system of linear equations,

$$\phi(\mathbf{u}_l) + \sum_{i=1}^m t(\mathbf{u}_l, \mathbf{u}_i)\phi(\mathbf{u}_i)w_i = \frac{\lambda'(\mathbf{u}_l; \boldsymbol{\beta})}{\lambda(\mathbf{u}_l; \boldsymbol{\beta})}, \quad l = 1, \dots, m.$$

Then $(\mathbf{T}\phi)(\mathbf{u}) \approx \sum_{i=1}^m t(\mathbf{u}, \mathbf{u}_i)\hat{\phi}(\mathbf{u}_i)w_i$ and plugging this further approximation into (11), the Nyström approximate solution of (11) directly becomes

$$\hat{\phi}(\mathbf{u}) = \frac{\lambda'(\mathbf{u}; \boldsymbol{\beta})}{\lambda(\mathbf{u}; \boldsymbol{\beta})} - \sum_{i=1}^m t(\mathbf{u}, \mathbf{u}_i)\hat{\phi}(\mathbf{u}_i)w_i. \quad (19)$$

In (13) we replace ϕ by $\hat{\phi}$ and we approximate the integral term applying again the quadrature rule used to obtain $\hat{\phi}$. This leads to

$$\hat{\mathbf{e}}(\boldsymbol{\beta}) = \sum_{\mathbf{u} \in X \cap W} \hat{\phi}(\mathbf{u}) - \sum_{i=1}^m \hat{\phi}(\mathbf{u}_i)\lambda(\mathbf{u}_i; \boldsymbol{\beta})w_i. \quad (20)$$

To estimate $\boldsymbol{\beta}$, we solve $\hat{\mathbf{e}}(\boldsymbol{\beta}) = 0$ iteratively using Fisher scoring. Suppose that the current estimate is $\boldsymbol{\beta}^{(l)}$. Then $\boldsymbol{\beta}^{(l+1)}$ is obtained by the Fisher scoring update

$$\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} + \hat{\mathbf{e}}(\boldsymbol{\beta}^{(l)})\hat{\mathbf{S}}^{-1}, \quad (21)$$

where

$$\hat{\mathbf{S}} = \sum_{i=1}^m \hat{\phi}(\mathbf{u}_i)\mathbf{T}\lambda'(\mathbf{u}_i; \boldsymbol{\beta}^{(l)})w_i \quad (22)$$

is the numerical approximation of the sensitivity matrix $\mathbf{S} = \int_W \phi^\top(\mathbf{u})\lambda'(\mathbf{u}; \boldsymbol{\beta}^{(l)})d\mathbf{u}$.

Provided the quadrature scheme is convergent, it follows by Lemma 4.7.4, Lemma 4.7.6 and Theorem 4.7.7 in Hackbusch (1995) that $\|\phi - \hat{\phi}\|_\infty$ converges to zero as $m \rightarrow \infty$ where $\|\cdot\|_\infty$ denotes supremum norm of a function. This justifies the use of the Nyström method to obtain an approximate solution of the Fredholm integral equation.

4.2. Implementation as quasi-likelihood

Suppose that we are using simple Riemann quadrature in (18). Then the w_i 's correspond to areas of sets B_i that partition W and for each i , $\mathbf{u}_i \in B_i$. Let Y_i denote the number of points from X falling in B_i and define $\mu_i = \lambda(\mathbf{u}_i; \boldsymbol{\beta})w_i$. If the B_i 's are sufficiently small so that the Y_i 's are binary then (20) is approximately equal to

$$\sum_{i=1}^m \hat{\phi}(\mathbf{u}_i)(Y_i - \mu_i). \quad (23)$$

Further, by (1) and (2), $\mathbb{E}Y_i \approx \mu_i$ and

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= 1(i=j) \int_{B_i} \lambda(\mathbf{u}; \boldsymbol{\beta})d\mathbf{u} + \int_{B_i \times B_j} \lambda(\mathbf{u}; \boldsymbol{\beta})\lambda(\mathbf{v}; \boldsymbol{\beta})[g(\mathbf{u} - \mathbf{v}) - 1]d\mathbf{u}d\mathbf{v} \\ &\approx V_{ij} = \mu_i 1(i=j) + \mu_i \mu_j [g(\mathbf{u}_i, \mathbf{u}_j) - 1]. \end{aligned}$$

Define $\mathbf{Y} = (Y_i)_i$, $\boldsymbol{\mu} = (\mu_i)_i$ and $\mathbf{V} = [V_{ij}]_{ij}$. Then $\mathbb{E}\mathbf{Y} \approx \boldsymbol{\mu}$ and $\text{Cov}\mathbf{Y} \approx \mathbf{V}$. Moreover, from (19), $[\hat{\phi}(\mathbf{u}_i)]_i = \mathbf{V}^{-1}\mathbf{D}$ where $\mathbf{D} = d\boldsymbol{\mu}^\top/d\boldsymbol{\beta}$ is the $m \times p$ matrix of partial derivatives $d\mu_i/d\beta_j$. Hence, (22) becomes

$$\mathbf{D}^\top \mathbf{V}^{-1} \mathbf{D} \quad (24)$$

and (23) becomes

$$(\mathbf{Y} - \boldsymbol{\mu})\mathbf{V}^{-1}\mathbf{D}, \quad (25)$$

which is formally a quasi-likelihood score for spatial data \mathbf{Y} with mean $\boldsymbol{\mu}$ and covariance matrix \mathbf{V} (Gotway and Stroup, 1997).

4.3. Computational details

With the quasi-likelihood formulation discussed in the previous section, $\hat{\mathbf{S}}$ and $\hat{\mathbf{e}}$ in (21) are given by respectively (24) and (25). The Fisher scoring updates (21) thus take the form of generalized least squares updates

$$(\boldsymbol{\beta}^{(l+1)} - \boldsymbol{\beta}^{(l)})\mathbf{D}(\boldsymbol{\beta}^{(l)})^\top \mathbf{V}(\boldsymbol{\beta}^{(l)})^{-1} \mathbf{D}(\boldsymbol{\beta}^{(l)}) = [\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(l)})]\mathbf{V}(\boldsymbol{\beta}^{(l)})^{-1} \mathbf{D}(\boldsymbol{\beta}^{(l)}), \quad (26)$$

where we here use the notation $\mathbf{D}(\boldsymbol{\beta})$, $\mathbf{V}(\boldsymbol{\beta})$ and $\boldsymbol{\mu}(\boldsymbol{\beta})$ to emphasize the dependence of \mathbf{D} , \mathbf{V} , and $\boldsymbol{\mu}$ on $\boldsymbol{\beta}$.

Let $\mathbf{V} = \mathbf{V}_\mu^{1/2}(\mathbf{I} + \mathbf{G})\mathbf{V}_\mu^{1/2}$ where $\mathbf{V}_\mu = \text{Diag}(\mu_i)$ and $G_{ij} = \sqrt{\mu_i \mu_j} [g(\mathbf{u}_i, \mathbf{u}_j) - 1]$ so that $\mathbf{G} = [G_{ij}]_{ij}$ is the matrix analogue of the symmetric operator \mathbf{T}^s from Section 3.1. A computational difficulty in (26) arise from the inversion of the high-dimensional matrix $\mathbf{I} + \mathbf{G}$. However, we can approximate \mathbf{G} by a sparse matrix $\mathbf{G}_{\text{taper}}$ obtained using tapering (e.g. Furrer et al., 2006). Thus in (26) \mathbf{V} is replaced by $\mathbf{V}_{\text{taper}} = \mathbf{V}_\mu^{1/2}(\mathbf{I} + \mathbf{G}_{\text{taper}})\mathbf{V}_\mu^{1/2}$. A sparse matrix Cholesky decomposition $\mathbf{I} + \mathbf{G}_{\text{taper}} = \mathbf{L}\mathbf{L}^\top$ is obtained using the `R Matrix` package

by Doug Bates and Martin Maechler. The matrix product $\mathbf{V}_{\text{taper}}^{-1}\mathbf{D}$ can then be efficiently computed by solving the equation $\mathbf{x}\mathbf{V}_{\mu}^{1/2}\mathbf{L}\mathbf{L}^{\top} = \mathbf{V}_{\mu}^{-1/2}\mathbf{D}$ with respect to \mathbf{x} by using forward and backward substitution for the sparse Cholesky factors \mathbf{L} and \mathbf{L}^{\top} , respectively.

In practise, it is often assumed that $g(\mathbf{r}) = g_0(\|\mathbf{r}\|)$ for some function g_0 . If g_0 is a decreasing function of $\|\mathbf{r}\|$ then we may define the entries in $\mathbf{G}_{\text{taper}}$ as $G_{ij}1[\|\mathbf{u}_i - \mathbf{u}_j\| \leq d_{\text{taper}}]$, where d_{taper} solves $[g_0(d) - 1]/[g_0(0) - 1] = \epsilon$ for some small ϵ . That is, we replace entries G_{ij} by zero if $g_0(\|\mathbf{u}_i - \mathbf{u}_j\|) - 1$ is below some small percentage of the maximal value $g_0(0) - 1$. In general g inside $\mathbf{G}_{\text{taper}}$ is unknown and must be replaced by an estimate. We replace β and g in $\mathbf{G}_{\text{taper}}$ by preliminary estimates (see Section 4.4) which are fixed during the generalized least squares iterations. This yields further computational simplification since the Cholesky factorization of $\mathbf{I} + \mathbf{G}_{\text{taper}}$ then only needs to be computed once.

By the asymptotic result (29) in Section 5 the asymptotic covariance matrix of $\hat{\beta}$ is given by the inverse sensitivity where the sensitivity is estimated by (24). When \mathbf{V} is replaced by $\mathbf{V}_{\text{taper}}$ in (26) we need the following adjusted estimate of the covariance matrix of $\hat{\beta}$:

$$\mathbf{S}_{\text{taper}}^{-1}\mathbf{D}^{\top}\mathbf{V}_{\text{taper}}^{-1}\mathbf{V}\mathbf{V}_{\text{taper}}^{-1}\mathbf{D}\mathbf{S}_{\text{taper}}^{-1}, \quad (27)$$

where $\mathbf{S}_{\text{taper}} = \mathbf{D}^{\top}\mathbf{V}_{\text{taper}}^{-1}\mathbf{D}$. Parameter standard errors are given by the square roots of the diagonal elements in (27). Note that it is not required to invert the non-sparse covariance matrix \mathbf{V} in order to compute (27). Nevertheless, the computation of (27) can still be computationally intensive as shown in the practical examples in Section 6.2.

The discretization of W should be chosen as fine as possible and the tapering parameter ϵ as small as possible given the available computational resources. Typically the discretization of W is generated by the cells B_i of a regular grid. In the practical data example in Section 6.2 we investigate the sensitivity of the parameter estimates and parameter standard errors to the choice of grid size and ϵ . In the data example very similar results are obtained with $\epsilon = 0.002$, 0.01 and $\epsilon = 0.05$ suggesting that $\epsilon = 0.05$ is in fact sufficiently small. Tapering entails a loss of statistical efficiency relative to quasi-likelihood estimation without tapering. However, valid parameter estimates and standard errors from the adjusted covariance matrix (27) are obtained regardless of the chosen value of the tapering tuning parameter ϵ .

4.4. Preliminary Estimation of Intensity and Pair Correlation

To obtain a preliminary estimate of g we assume that $g(\mathbf{r}) = g(\mathbf{r}; \psi)$ where $g(\cdot; \psi)$ is a translation invariant parametric pair correlation function model. We replace ψ and β inside \mathbf{G} by preliminary estimates $\hat{\beta}$ and $\hat{\psi}$ which are fixed during the iterations (26). The estimates $\hat{\beta}$ and $\hat{\psi}$ can be obtained using the two-step approach in Waagepetersen and Guan (2009) where $\hat{\beta}$ is obtained from the composite likelihood function and $\hat{\psi}$ is a minimum contrast estimate based on the K -function. That is,

$$\hat{\psi} = \underset{\psi}{\operatorname{argmin}} \int_0^{r_{\max}} [K(t; \psi) - \hat{K}(t)]^{1/4} dt$$

where

$$K(t; \psi) = \int_{\|\mathbf{r}\| \leq t} g(\mathbf{r}; \psi) d\mathbf{r} \quad (28)$$

is the K -function and \hat{K} is a non-parametric estimate of the K -function (Baddeley et al., 2000). If translation invariance can not be assumed, ψ may instead be estimated using a second-order composite likelihood as in Jalilian et al. (2013).

5. ASYMPTOTIC THEORY

Let $W_n \subset \mathbb{R}^2$ be an increasing sequence of observation windows in \mathbb{R}^2 . Following Section 4.4 we assume that the true pair correlation function is given by a parametric model $g(\mathbf{r}) = g(\mathbf{r}; \boldsymbol{\psi})$ for some unknown parameter vector $\boldsymbol{\psi} \in \mathbb{R}^q$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}) \in \mathbb{R}^{p+q}$. We denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$. In what follows, \mathbb{E} and Var denote expectation and variance under the distribution corresponding to $\boldsymbol{\theta}^*$.

Introducing the dependence on n and $\boldsymbol{\theta}$ in the notation from Section 3, we have

$$\phi_{n,\boldsymbol{\theta}}(\mathbf{u}, \boldsymbol{\beta}) = \left[(\mathbf{I} + \mathbf{T}_{n,\boldsymbol{\theta}})^{-1} \frac{\lambda'(\cdot; \boldsymbol{\beta})}{\lambda(\cdot; \boldsymbol{\beta})} \right](\mathbf{u}), \quad (\mathbf{T}_{n,\boldsymbol{\theta}} \mathbf{f})(\mathbf{u}) = \int_{W_n} t_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}) \mathbf{f}(\mathbf{v}) d\mathbf{v}$$

and

$$t_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{v}; \boldsymbol{\beta}) [g(\mathbf{u} - \mathbf{v}; \boldsymbol{\psi}) - 1].$$

Following Section 4.4 we replace $\boldsymbol{\theta}$ in the kernel $t_{\boldsymbol{\theta}}$ by a preliminary estimate $\tilde{\boldsymbol{\theta}}_n = (\tilde{\boldsymbol{\beta}}_n, \tilde{\boldsymbol{\psi}}_n)$. The estimating function (13) then becomes $\mathbf{e}_{n,\tilde{\boldsymbol{\theta}}_n}(\boldsymbol{\beta})$ where

$$\mathbf{e}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta}) = \sum_{\mathbf{u} \in X \cap W_n} \phi_{n,\boldsymbol{\theta}}(\mathbf{u}, \boldsymbol{\beta}) - \int_{W_n} \phi_{n,\boldsymbol{\theta}}(\mathbf{u}, \boldsymbol{\beta}) \lambda(\mathbf{u}; \boldsymbol{\beta}) d\mathbf{u}.$$

Let $\hat{\boldsymbol{\beta}}_n$ denote the estimator obtained by solving $\mathbf{e}_{n,\tilde{\boldsymbol{\theta}}_n}(\boldsymbol{\beta}) = 0$. Further, define

$$\bar{\boldsymbol{\Sigma}}_n = |W_n|^{-1} \text{Vare}_{n,\boldsymbol{\theta}^*}(\boldsymbol{\beta}^*), \quad \mathbf{J}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta}) = -\frac{d}{d\boldsymbol{\beta}} \mathbf{e}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta}) \quad \text{and} \quad \bar{\mathbf{S}}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta}) = |W_n|^{-1} \mathbb{E} \mathbf{J}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta}).$$

Note that $\bar{\boldsymbol{\Sigma}}_n$ and $\bar{\mathbf{S}}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta})$ are ‘averaged’ versions of $\boldsymbol{\Sigma}_n = \text{Vare}_{n,\boldsymbol{\theta}^*}(\boldsymbol{\beta}^*)$ and $\mathbf{S}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta}) = \mathbb{E} \mathbf{J}_{n,\boldsymbol{\theta}}(\boldsymbol{\beta})$.

In Section 2 in the supplementary material we verify the existence of a $|W_n|^{1/2}$ consistent sequence of solutions $\hat{\boldsymbol{\beta}}_n$, i.e., $|W_n|^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$ is bounded in probability. We further show in Section 3 in the supplementary material that $|W_n|^{-1/2} \mathbf{e}_{n,\tilde{\boldsymbol{\theta}}_n}(\boldsymbol{\beta}^*) \bar{\boldsymbol{\Sigma}}_n^{-1/2}$ is asymptotically standard normal. The conditions needed for these results are listed in Section 1 in the supplementary material. By a Taylor series expansion,

$$|W_n|^{-1/2} \mathbf{e}_{n,\tilde{\boldsymbol{\theta}}_n}(\boldsymbol{\beta}^*) \bar{\boldsymbol{\Sigma}}_n^{-1/2} = |W_n|^{1/2} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \frac{\mathbf{J}_{n,\tilde{\boldsymbol{\theta}}_n}(\mathbf{b}_n)}{|W_n|} \bar{\boldsymbol{\Sigma}}_n^{-1/2}$$

for some $\mathbf{b}_n \in \mathbb{R}^p$ satisfying $\|\mathbf{b}_n - \boldsymbol{\beta}^*\| \leq \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\|$. Invoking further the convergence $\mathbf{J}_{n,\tilde{\boldsymbol{\theta}}_n}(\mathbf{b}_n)/|W_n| \rightarrow \bar{\mathbf{S}}_{n,\boldsymbol{\theta}^*}(\boldsymbol{\beta}^*)$ (results R2 and R3 in Section 2 in the supplementary material) and that $\bar{\boldsymbol{\Sigma}}_n = \bar{\mathbf{S}}_{n,\boldsymbol{\theta}^*}(\boldsymbol{\beta}^*)$ by (14), we obtain

$$|W_n|^{1/2} (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \bar{\mathbf{S}}_{n,\boldsymbol{\theta}^*}(\boldsymbol{\beta}^*)^{1/2} \rightarrow N_p(0, \mathbf{I}). \quad (29)$$

Hence, for a fixed n , $\hat{\boldsymbol{\beta}}_n$ is approximately normal with mean $\boldsymbol{\beta}^*$ and covariance matrix estimated by $|W_n|^{-1} \bar{\mathbf{S}}_{n,(\tilde{\boldsymbol{\psi}}_n, \hat{\boldsymbol{\beta}}_n)}^{-1}(\hat{\boldsymbol{\beta}}_n)$.

6. SIMULATION STUDY AND DATA EXAMPLE

To examine the performance of our optimal quasi-likelihood estimator relative to composite likelihood and weighted composite likelihood, we carry out a simulation study under the Guan and Shen (2010) setting. We refrain from a comparison with maximum likelihood estimation due to the lack of a computationally feasible implementation of this method. In addition to the simulation study we demonstrate the practical usefulness of our method and discuss computational issues in a tropical rain forest data example.

6.1. Simulation Study

Following Guan and Shen (2010), realizations of Cox processes are generated on a square window W . Each simulation involves first the generation of a zero-mean Gaussian random covariate field $\mathbf{Z} = \{Z(\mathbf{u})\}_{\mathbf{u} \in W}$ with exponential covariance function $c(\mathbf{u}) = \exp(-\|\mathbf{u}\|/\gamma)$, $\gamma > 0$, and then the generation of an inhomogeneous Thomas point process given \mathbf{Z} with intensity function $\lambda(\mathbf{u}; \boldsymbol{\beta}) = \exp[\beta_0 + \beta_1 Z(\mathbf{u})]$ and pair correlation function

$$g(\mathbf{r}) = 1 + \exp[-\|\mathbf{r}\|^2/(4\omega^2)]/(4\pi\omega^2\kappa), \quad (30)$$

where $\kappa > 0$ is the intensity of the parent process and $\omega > 0$ is the dispersal parameter (Waagepetersen, 2007). For each simulation, preliminary estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\psi}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{\psi} = (\kappa, \omega)$ are obtained using the two-step method in Waagepetersen and Guan (2009) (i.e. $\tilde{\boldsymbol{\beta}}$ is the composite likelihood (CL) estimate of $\boldsymbol{\beta}$). These preliminary estimates are then used to further obtain weighted composite likelihood (WCL) and quasi-likelihood (QL) estimates, see Section 3.2 and Section 4.3-4.4 for details.

The root mean square error (RMSE) of the CL, WCL and QL estimates is computed from 1000 simulations for different settings of parameters and observation windows $W = [0, 1]^2$ or $W = [0, 2]^2$. We consider covariate spatial correlation scales $\gamma^* = 0.05, 0.1, 0.2$. The combinations of parent point intensities $\kappa^* = 100, 200$ and dispersal parameters $\omega^* = 0.02, 0.04$ further create a wide selection of clustering behaviours given \mathbf{Z} as reflected by the corresponding pair correlation functions. We moreover consider different inhomogeneity levels $\beta_1^* = 0.5, 1$ and adjust the intercept β_0^* so that the expected number of points is always 400 in the case of $W = [0, 1]^2$ and 1600 in the case of $W = [0, 2]^2$. The integral terms in the CL, WCL and QL estimating functions are approximated using a 50×50 grid for $W = [0, 1]^2$ and a 100×100 grid for $W = [0, 2]^2$. Tapering for QL is carried out as described in Section 4.4 using d_{taper} obtained with $\epsilon = 0.01$ for each estimated pair correlation function $g(\cdot; \tilde{\boldsymbol{\psi}})$. For WCL we use $A(\mathbf{u}) \approx K(d_{\text{taper}}; \tilde{\boldsymbol{\psi}}) - \pi d_{\text{taper}}^2$ where $A(\mathbf{u})$ is given by (17) and K by (28).

The intercept β_0 is typically not of prime interest when it comes to statistical inference in regression models so in the following we focus on the results for β_1 . Table 1 shows for each setting of $(\gamma^*, \kappa^*, \omega^*, \beta_1^*)$ and size of window W the root mean square errors (RMSEs) of the QL estimates for β_1 as well as the increases in RMSE for WCL and CL relative to QL. The table also shows empirical standard errors (SD) for the QL estimates as well as the square root of the average of the estimated variances obtained for the QL estimates using (27) (column ASD in Table 1). The RMSE and the empirical standard errors coincide which confirms the unbiasedness of the QL estimates. There is further close agreement between SD and ASD as expected from the formula $\sqrt{\text{Var}\hat{\beta}_1} = \sqrt{\mathbb{E}\text{Var}[\hat{\beta}_1|\mathbf{Z}] + \text{Var}\mathbb{E}[\hat{\beta}_1|\mathbf{Z}]}$ since the expression (27) provides an estimate of $\text{Var}[\hat{\beta}_1|\mathbf{Z}]$ and $\text{Var}\mathbb{E}[\hat{\beta}_1|\mathbf{Z}]$ is zero by unbiasedness

Table 1. RMSE for QL estimates of β_1 and the increase (in %) in RMSE for WCL and CL relative to QL. The empirical standard error (SD) and asymptotic standard error (ASD) for the QL estimates are shown too.

ψ^*	γ^*	β_1^*	$W = [0, 1]^2 \quad \mathbb{E}[N(W)] = 400$					$W = [0, 2]^2 \quad \mathbb{E}[N(W)] = 1600$				
			QL	WCL	CL	SD	ASD	QL	WCL	CL	SD	ASD
(100, .02)	.05	.5	.08	21	24	.08	.08	.04	21	26	.04	.04
		1	.09	22	44	.09	.08	.04	25	59	.04	.04
	.1	.5	.09	13	21	.09	.09	.04	15	25	.04	.04
		1	.10	17	39	.10	.10	.05	17	61	.05	.05
	.2	.5	.11	11	15	.11	.10	.05	11	20	.05	.05
		1	.12	13	31	.12	.12	.06	12	49	.06	.06
(100, .04)	.05	.5	.06	23	20	.06	.06	.03	19	21	.03	.03
		1	.07	29	26	.07	.07	.03	31	40	.03	.03
	.1	.5	.08	23	23	.08	.07	.04	22	28	.04	.04
		1	.08	29	34	.08	.08	.04	24	42	.04	.04
	.2	.5	.09	14	20	.09	.09	.05	18	25	.05	.04
		1	.11	16	25	.11	.10	.05	14	38	.05	.05
(200, .02)	.05	.5	.07	7	9	.07	.07	.03	10	13	.03	.03
		1	.08	12	23	.08	.07	.04	21	34	.04	.04
	.1	.5	.08	7	10	.08	.08	.04	7	11	.04	.04
		1	.09	6	19	.09	.08	.04	13	43	.04	.04
	.2	.5	.09	4	7	.09	.09	.04	6	11	.04	.04
		1	.11	8	19	.11	.10	.05	4	30	.05	.05
(200, .04)	.05	.5	.06	9	8	.06	.07	.03	11	10	.03	.03
		1	.07	20	10	.07	.06	.03	25	21	.03	.03
	.1	.5	.07	9	10	.07	.07	.03	11	13	.03	.03
		1	.08	13	9	.08	.07	.04	16	25	.04	.03
	.2	.5	.08	7	8	.08	.08	.04	9	9	.04	.04
		1	.10	14	11	.10	.08	.04	11	23	.04	.04

of $\hat{\beta}_1$ given \mathbf{Z} . The standard errors increase with increasing γ^* but there is not a strong dependence of the standard errors on $(\kappa^*, \omega^*, \beta_1^*)$. For the larger observation window the standard errors are half as big as for the smaller. This is in accordance with the asymptotic result (29) which implies that the standard deviation is inversely proportional to the square root of the observation window size.

As expected from the theoretical results, the QL estimator has superior performance compared with both CL and WCL in all cases. The improvement over CL is especially substantial in the more clustered (corresponding to small κ^* and ω^*) and more inhomogeneous (corresponding to $\beta_1^* = 1$) cases where CL has up to 44% larger RMSE than QL in case of $W = [0, 1]$ and up to 61% larger RMSE in case of $W = [0, 2]^2$. As we alluded in Section 3.2, the performance of WCL may rely on the validity of the approximation (16). In case of a longer dependence range, the approximation is expected to be less accurate and this explains the large drop in efficiency of WCL relative to CL when ω^* increases from 0.02 to 0.04. In particular, WCL even performs worse than CL when $\psi^* = (200, 0.04)$. In contrast, QL still fares better than CL with increases in RMSE of 8-11% ($W = [0, 1]^2$) and 9-25% ($W = [0, 2]^2$) for CL relative to QL. There is not a clear pattern regarding the dependence on γ^* of the relative increases.

6.2. Data Example

A fundamental problem in biological research is to understand the very high biodiversity in tropical rain forests. One explanation is the niche assembly hypothesis, which states that different species coexist by adapting to different environmental niches. Data available for studying this hypothesis consist of point patterns of locations of trees as well as observations of environmental covariates. Figure 1 shows the spatial locations of three tree species, *Acalypha diversifolia* (528 trees), *Lonchocarpus heptaphyllus* (836 trees) and *Capparis frondosa* (3299 trees), in a $1000m \times 500m$ observation window on Barro Colorado Island (Condit et al., 1996; Condit, 1998; Hubbell and Foster, 1983). Also one example of an environmental variable (potassium content in the soil) is shown.

In order to study the niche assembly hypothesis we use our quasi-likelihood method to fit log-linear regression models for the intensity functions depending on environmental variables. In addition to soil potassium content (K , divided by 1000), we consider nine other covariates for the intensity functions: pH, elevation (\mathbf{dem} , divided by 100), slope gradient (\mathbf{grad}), multi-resolution index of valley bottom flatness (\mathbf{mrvbf}), incoming mean solar radiation (\mathbf{solar}), topographic wetness index (\mathbf{twi}) as well as soil contents of copper (\mathbf{Cu}), mineralized nitrogen (\mathbf{Nmin} , divided by 100) and phosphorus (\mathbf{P}). The quasi-likelihood estimation was implemented as in the simulation study using a 100×50 grid for the numerical quadrature and tapering tuning parameter $\epsilon = 0.01$.

6.2.1. Modeling of pair correlation function

For each species we initially fit using the two-step procedure in Waagepetersen and Guan (2009) the following pair correlation functions of normal variance mixture type (Jalilian et al., 2013):

$$g(\mathbf{r}; \boldsymbol{\psi}) = 1 + c(\mathbf{r}; \boldsymbol{\psi}), \quad \mathbf{r} \in \mathbb{R}^2,$$

where the covariance function $c(\mathbf{r}; \boldsymbol{\psi})$ is either Gaussian

$$c(\mathbf{r}; (\sigma^2, \alpha)) = \sigma^2 \exp[-(\|\mathbf{r}\|/\alpha)^2],$$

Matérn

$$c(\mathbf{r}; (\sigma^2, \alpha, \nu)) = \sigma^2 \frac{(\|\mathbf{r}\|/\alpha)^\nu K_\nu(\|\mathbf{r}\|/\alpha)}{2^{\nu-1} \Gamma(\nu)},$$

(K_ν is the modified Bessel function of the second kind) or Cauchy

$$c(\mathbf{r}; (\sigma^2, \alpha)) = \sigma^2 [1 + (\|\mathbf{r}\|/\alpha)^2]^{-3/2}.$$

These covariance functions represent very different tail behaviour ranging from light (Gaussian), exponential (Matérn), to heavy tails (Cauchy). The pair correlation function obtained with the Gaussian covariance function is just a re-parametrization of the Thomas process pair correlation function (30). For the Matérn covariance we consider three different values of the shape parameter $\nu = 0.25, 0.5$ and 1 . With $\nu = 0.5$ the exponential model $c[\mathbf{r}; (\sigma^2, \alpha, 0.5)] = \sigma^2 \exp(-\|\mathbf{r}\|/\alpha)$ is obtained while $\nu = 0.25$ and 1 yields respectively a log convex and a log concave covariance function.

The solid curves in Figure 2 show $g(\cdot; \hat{\boldsymbol{\psi}})$ for the best fitting (in terms of the minimum contrast criterion for the corresponding K -function) pair correlation functions: Cauchy for *Acalypha* ($\hat{\boldsymbol{\psi}} = (15.4, 4.6)$), Matérn ($\hat{\boldsymbol{\psi}} = (2.3, 15.4, 0.5)$) for *Lonchocarpus* and Matérn ($\hat{\boldsymbol{\psi}} = (1.3, 22.9, 0.25)$) for *Capparis*. The tapering distances corresponding to $\epsilon = 0.01$ are

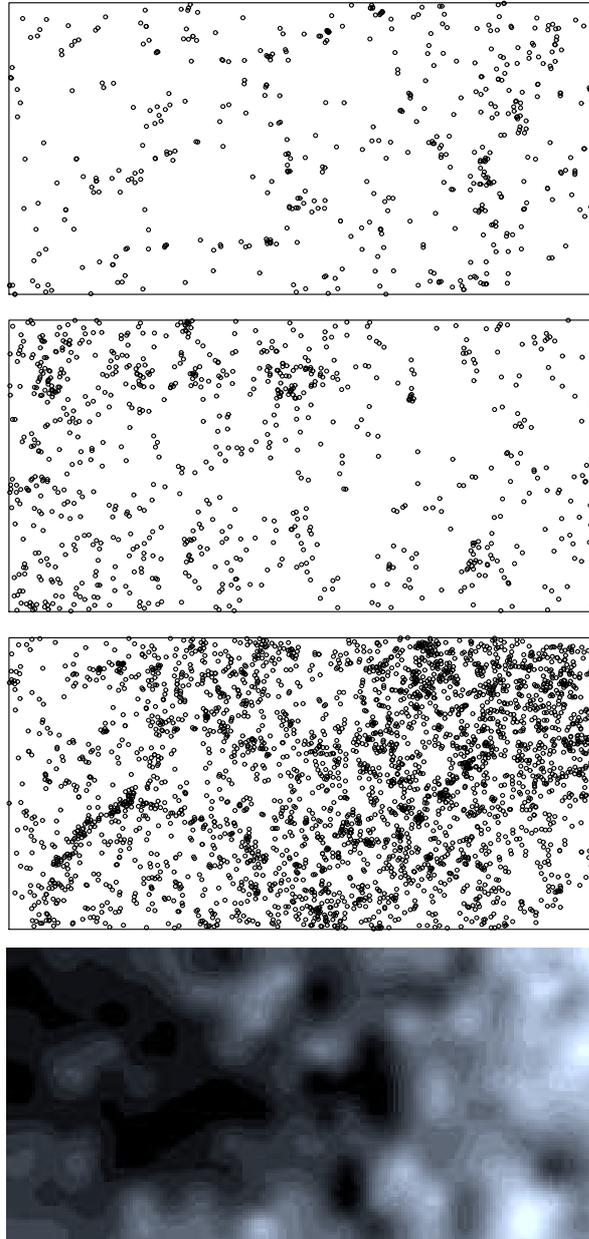


Fig. 1. Locations of *Acalypha*, *Lonchocarpus*, and *Capparis* trees and image of interpolated potassium content in the surface soil (from top to bottom).

respectively 20.1, 71.3 and 84.9 for the three species. Hence *Capparis* is the computationally most challenging case.

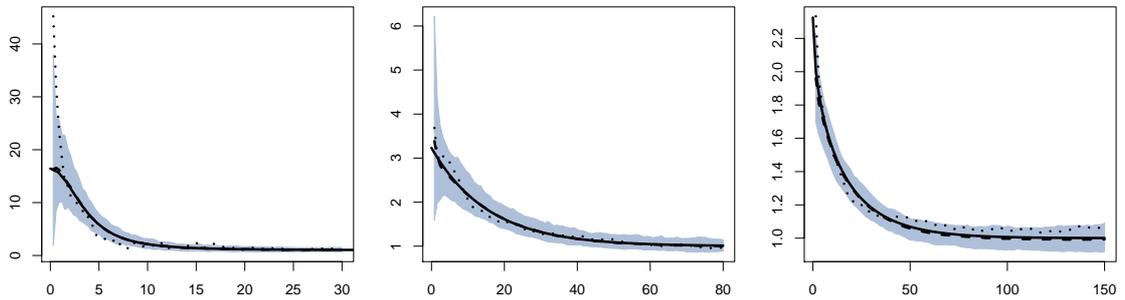


Fig. 2. Solid curves: best fitting pair correlation functions $g(\cdot; \hat{\psi}) - 1$ for *Acalypha* (left), *Lonchocarpus* (middle), and *Capparis* (right). Dotted curve: non-parametric estimate of the pair correlation function. Shaded area: 95% point-wise envelopes based on 199 simulations of fitted model. Dashed curve: mean of non-parametric estimates of pair correlation functions obtained from the simulations.

Table 2. Quasi-likelihood parameter estimates and standard errors for the selected intensity function models. The table also shows the composite and weighted composite likelihood estimates and the increases in percent of their standard errors relative to quasi-likelihood.

		Acalypha		Lonchocarpus			Capparis			
		K		Nmin	P		dem	grad	K	
Estm.	QL	-6.9	4.4	-6.5	-2.8	-0.15	-5.1	2.3	-2.0	4.1
	WCL	-6.9	4.3	-6.5	-2.8	-0.16	-5.1	2.8	-0.9	4.2
	CL	-6.9	4.0	-6.5	-2.7	-0.16	-5.1	2.9	-1.1	4.3
se.	QL	0.085	1.2	0.088	0.69	0.055	0.066	0.80	0.95	1.0
Incr. se.	WCL	0.0	0.1	0.9	1.4	2.1	1.3	4.6	11.5	5.3
	CL	0.0	2.3	1.2	7.0	4.5	1.5	8.5	17.1	7.4

6.2.2. Results of quasi-likelihood estimation and comparison with previous methods

Backward model selection based on the quasi-likelihood estimates is carried out for each species using significance level 5%. The selected covariates are potassium (K) for *Acalypha*, mineralized nitrogen (Nmin) and phosphorous (P) for *Lonchocarpus* and elevation (dem), gradient (grad) and potassium for *Capparis*. The associated parameter estimates and standard errors are shown in Table 2. The qualitative biological findings from the selected models and the associated parameter estimates are that *Acalypha* and *Capparis* both have a preference for niches with high levels of potassium. Regarding topography, *Capparis* prefers flat areas in high altitudes. Finally, *Lonchocarpus* appears as a frugal species which is adapted to low levels of mineralized nitrogen and phosphorous. Table 2 for comparison also shows the composite and weighted composite likelihood estimates which are in general quite similar to the quasi-likelihood estimates. One exception is the gradient (grad) regression parameter estimate for *Capparis* where the absolute value of the quasi-likelihood estimate is about twice as large as for the other estimates.

The advantage of using quasi-likelihood instead of composite likelihood or weighted composite likelihood is that we obtain more precise regression parameter estimates as reflected in smaller estimation standard errors and narrower confidence intervals. Table 2 also shows parameter estimation standard errors for quasi-likelihood and the increase in percent in the standard errors when composite or weighted composite likelihood is used instead. As demonstrated in the simulation study, the advantage of using quasi-likelihood depends much

Table 3. Computing times (T) in seconds (without computation of standard errors) and QL parameter estimates for different combinations of grid size and tapering.

Grid	ϵ	Acalypha			Lonchocarpus				Capparis				
		T	estm.			T	estm.			T	estm.		
100×50	0.05	0.3	-6.9	4.4	1.1	-6.5	-2.8	-0.16	1.0	-5.1	2.4	-2.0	4.3
	0.01	0.4	-6.9	4.4	2.6	-6.5	-2.8	-0.15	2.2	-5.1	2.3	-2.0	4.1
	.002	0.6	-6.9	4.4	4.4	-6.5	-2.8	-0.15	4.3	-5.1	2.3	-2.0	4.1
150×75	0.05	0.5	-6.9	4.3	8.5	-6.5	-2.8	-0.16	10.2	-5.1	2.4	-1.9	4.2
	0.01	1.8	-6.9	4.3	23.7	-6.5	-2.8	-0.15	25.0	-5.1	2.3	-1.9	4.1
	.002	5.3	-6.9	4.3	41.6	-6.5	-2.8	-0.15	108.2	-5.1	2.3	-1.9	4.0

on the strength of correlation as measured by the pair correlation function. For Acalypha where the pair correlation function drops off quickly, the standard errors for the composite likelihood approaches are at most 2.3% larger than for quasi-likelihood. On the other hand, for Capparis the increase in standard error is up to 17.1%. The differences in the parameter estimates and standard errors imply that gradient is a significant covariate on the 5% level according to the quasi-likelihood results but not according to the other two approaches.

We carry out model assessment using the empirical empty space (F) function and a non-parametric estimate of the pair correlation function adjusted for the inhomogeneous intensity function where 95% point-wise envelopes are computed using simulations under the fitted models (for details see e.g. Sections 4.3.4-4.3.6 in Møller and Waagepetersen, 2004). For Lonchocarpus and Capparis the non-parametric estimates of the pair correlation function are contained within the envelopes while a few minor excursions occur for Acalypha (Figure 2). Plots based on the F function (not shown) did not reveal any deficiencies of the fitted models.

6.2.3. Dependence on grid size and tapering parameter

The computing time for the QL estimation depends both on the grid used for the numerical quadrature and the tapering tuning parameter ϵ . We also tried out a 150×75 grid and $\epsilon = 0.05$ and 0.02 for the QL fitting of the final models. Parameter estimates and parameter estimation computing time (system plus CPU time on a 2.90 GHz lap top) for all combinations of grid sizes, ϵ and species are shown in Table 3. The computing time for the parameter estimation depends much on both grid sizes, ϵ and species (i.e. range of spatial dependence). Computing time including computation of standard errors is shown in Table 4, together with the computed standard errors for the parameter estimates in Table 3. The computing time with computation of standard errors is less sensitive to ϵ and species since in this case the main computational burden arises from the non-sparse matrix in (27). For the 100×50 grid and $\epsilon = 0.01$, the maximal computing time of 13.5 seconds (including computation of standard errors) occurs for Capparis. In contrast to large variations in the computing time, the parameter estimates and estimated standard errors for each species are very stable across the combinations of grid sizes and tapering parameter ϵ .

7. DISCUSSION

In contrast to maximum likelihood estimation our quasi-likelihood estimation method only requires the specification of the intensity function and a pair correlation function. Moreover, the estimation of the regression parameters can be expected to be quite robust toward

Table 4. Computing times (T) in seconds (including computation of standard errors) and estimated standard errors of QL parameter estimates for different combinations of grid size and tapering

Grid	ϵ	Acalypha			Lonchocarpus				Capparis				
		T	sd.		T	sd.			T	sd.			
100×50	0.05	12.1	0.085	1.2	22.4	0.088	0.69	0.055	13.0	0.067	0.86	1.1	1.0
	0.01	12.0	0.085	1.2	24.0	0.088	0.69	0.055	13.5	0.067	0.86	1.1	1.0
	.002	12.1	0.085	1.2	25.9	0.088	0.69	0.055	13.6	0.067	0.86	1.1	1.0
150×75	0.05	59.4	0.079	1.1	187.2	0.087	0.69	0.055	158.6	0.066	0.86	1.1	1.0
	0.01	58.9	0.079	1.1	204.6	0.087	0.69	0.055	169.5	0.066	0.86	1.1	1.0
	.002	63.6	0.079	1.1	226.5	0.087	0.69	0.055	170.7	0.066	0.86	1.1	1.0

misspecification of the pair correlation function since the resulting estimating equation is unbiased for any choice of pair correlation function. In the data example we considered pair correlation functions obtained from covariance functions of normal variance mixture type. Alternatively one might consider pair correlation functions of the log Gaussian Cox process type (Møller et al., 1998), i.e., $g(\mathbf{r}) = \exp[c(\mathbf{r})]$, where $c(\cdot)$ is an arbitrary covariance function.

If a log Gaussian Cox process is deemed appropriate, a computationally feasible alternative to our approach is to use the method of integrated nested Laplace approximation (INLA, Rue et al., 2009; Illian et al., 2012) to implement Bayesian inference. However, in order to apply INLA it is required that the Gaussian field can be approximated well by a Gaussian Markov random field and this can limit the choice of covariance function. For example, the accurate Gaussian Markov random field approximations in Lindgren et al. (2011) of Gaussian fields with Matérn covariance functions are restricted to integer ν in the planar case. In contrast, our approach is not subject to such limitations and is not restricted to log Gaussian Cox processes.

We finally note that for the Nyström approximate solution of the Fredholm equation we used the simplest possible quadrature scheme given by a Riemann sum for a fine grid. This entails a minimum of assumptions regarding the integrand but at the expense of a typically high-dimensional covariance matrix \mathbf{V} . There may be scope for further development considering more sophisticated numerical quadrature schemes.

Acknowledgements

We thank the editors and a referee for their helpful and constructive comments. Abdollah Jalilian and Rasmus Waagepetersen’s research was supported by the Danish Natural Science Research Council, grant 09-072331 ‘Point process modeling and statistical inference’, Danish Council for Independent Research — Natural Sciences, Grant 12-124675, ‘Mathematical and Statistical Analysis of Spatial Data’, and by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Yongtao Guan’s research was supported by NSF grant DMS-0845368, by NIH grant 1R01DA029081-01A1 and by the VELUX Visiting Professor Programme.

The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell: DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Insti-

tute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Celera Foundation, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

The BCI soils data set were collected and analyzed by J. Dalling, R. John, K. Harms, R. Stallard and J. Yavitt with support from NSF DEB021104, 021115, 0212284, 0212818 and OISE 0314581, STRI and CTFS. Paolo Segre and Juan Di Trani provided assistance in the field. The covariates `dem`, `grad`, `mrvcf`, `solar` and `twi` were computed in SAGA GIS by Tomislav Hengl (<http://spatial-analyst.net/>).

References

- Baddeley, A. and R. Turner (2005). `spatstat`: An R package for analyzing spatial point patterns. *Journal of Statistical Software* 12(6), 1–42.
- Baddeley, A. J., J. Møller, and R. Waagepetersen (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* 54, 329–350.
- Baddeley, A. J., R. Turner, J. Møller, and M. Hazelton (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B* 67, 617–666.
- Condit, R. (1998). *Tropical Forest Census Plots*. Berlin, Germany and Georgetown, Texas: Springer-Verlag and R. G. Landes Company.
- Condit, R., S. P. Hubbell, and R. B. Foster (1996). Changes in tree species abundance in a neotropical forest: impact of climate change. *Journal of Tropical Ecology* 12, 231–256.
- Furrer, R., M. G. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* 15, 502–523.
- Gotway, C. A. and W. W. Stroup (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics* 2, 157–178.
- Gray, Robert, J. (2003). Weighted estimating equations for linear regression analysis of clustered failure time data. *Lifetime Data Analysis* 9(2), 123–138.
- Guan, Y. and J. M. Loh (2007). A thinned block bootstrap procedure for modeling inhomogeneous spatial point patterns. *Journal of the American Statistical Association* 102, 1377–1386.
- Guan, Y. and Y. Shen (2010). A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika* 97, 867–880.
- Hackbusch, W. (1995). *Integral equations - theory and numerical treatment*. Birkhäuser.
- Heyde, C. C. (1997). *Quasi-likelihood and its application - a general approach to optimal parameter estimation*. Springer Series in Statistics. Springer.

- Hubbell, S. P. and R. B. Foster (1983). Diversity of canopy trees in a neotropical forest and implications for conservation. In S. L. Sutton, T. C. Whitmore, and A. C. Chadwick (Eds.), *Tropical Rain Forest: Ecology and Management*, pp. 25–41. Oxford: Blackwell Scientific Publications.
- Illian, J. B., S. H. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Annals of Applied Statistics* 6, 1499–1530.
- Jalilian, A., Y. Guan, and R. Waagepetersen (2013). Decomposition of variance for spatial Cox processes. *Scandinavian Journal of Statistics* 40, 119–137.
- Lax, P. D. (2002). *Functional analysis*. Wiley.
- Liang, K. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
- Lin, P.-S. and M. K. Clayton (2005). Analysis of binary spatial data by quasi-likelihood estimating equations. *Annals of Statistics* 33, 542–555.
- Lin, Y.-C., L.-W. Chang, K.-C. Yang, H.-H. Wang, and I.-F. Sun (2011). Point patterns of tree distribution determined by habitat heterogeneity and dispersal limitation. *Oecologia* 165, 175–184.
- Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society B* 73, 423–498.
- Møller, J., A. R. Syversveen, and R. P. Waagepetersen (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics* 25, 451–482.
- Møller, J. and R. P. Waagepetersen (2004). *Statistical inference and simulation for spatial point processes*. Boca Raton: Chapman and Hall/CRC.
- Møller, J. and R. P. Waagepetersen (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics* 34, 643–684.
- Mrkvička, T. and I. Molchanov (2005). Optimisation of linear unbiased intensity estimators for point processes. *Annals of the Institute of Statistical Mathematics* 57, 71–81.
- Renner, I. W. and D. I. Warton (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* 69, 274–281.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B* 71, 319–392.
- Schoenberg, F. P. (2005). Consistent parametric estimation of the intensity of a spatial-temporal point process. *Journal of Statistical Planning and Inference* 128, 79–93.
- Shen, G., M. Yu, X.-S. Hu, X. Mi, H. Ren, I.-F. Sun, and K. Ma (2009). Species-area relationships explained by the joint effects of dispersal limitation and habitat heterogeneity. *Ecology* 90, 3033–3041.

- Song, P. X.-K. (2007). *Correlated data analysis: modeling, analytics, and applications*. Springer Series in Statistics. New York, NY: Springer.
- Waagepetersen, R. (2005). Discussion of ‘Residual analysis for spatial point processes’. *Journal of the Royal Statistical Society, Series B* 67, 662.
- Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* 63, 252–258.
- Waagepetersen, R. and Y. Guan (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society, Series B* 71, 685–702.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss - Newton method. *Biometrika* 61(3), 439–447.
- Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* 75, 621–629.

APPENDIX A. Condition for optimality

To show that (9) implies non-negative definiteness of (8), let $\hat{\mathbf{e}}_\phi(\boldsymbol{\beta}) = \mathbf{e}_\mathbf{f}(\boldsymbol{\beta})\boldsymbol{\Sigma}_\mathbf{f}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}\phi}$ be the optimal linear predictor of $\mathbf{e}_\phi(\boldsymbol{\beta})$ given $\mathbf{e}_\mathbf{f}(\boldsymbol{\beta})$. Then

$$\text{Var}[\hat{\mathbf{e}}_\phi(\boldsymbol{\beta}) - \mathbf{e}_\phi(\boldsymbol{\beta})] = \boldsymbol{\Sigma}_\phi - \boldsymbol{\Sigma}_{\phi\mathbf{f}}\boldsymbol{\Sigma}_\mathbf{f}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}\phi}$$

is non-negative definite whereby

$$\mathbf{S}_\phi\boldsymbol{\Sigma}_\phi^{-1}\mathbf{S}_\phi - \mathbf{S}_\phi\boldsymbol{\Sigma}_\phi^{-1}\boldsymbol{\Sigma}_{\phi\mathbf{f}}\boldsymbol{\Sigma}_\mathbf{f}^{-1}\boldsymbol{\Sigma}_{\mathbf{f}\phi}\boldsymbol{\Sigma}_\phi^{-1}\mathbf{S}_\phi$$

is non-negative definite too. Hence, (8) is non-negative definite provided

$$\mathbf{S}_\phi\boldsymbol{\Sigma}_\phi^{-1}\boldsymbol{\Sigma}_{\phi\mathbf{f}} = \mathbf{S}_\mathbf{f}$$

which follows from (9) (in particular, (9) implies $\boldsymbol{\Sigma}_\phi = \boldsymbol{\Sigma}_{\phi\phi} = \mathbf{S}_\phi$).

APPENDIX B. SOLUTION USING NEUMANN SERIES EXPANSION

Suppose that $\|\mathbf{T}\|_{\text{op}} = \sup\{\|\mathbf{T}\mathbf{f}\|_\infty/\|\mathbf{f}\|_\infty : \|\mathbf{f}\|_\infty \neq 0\} < 1$. Then we can obtain the solution ϕ of (11) using a Neumann series expansion which may provide additional insight on the properties of ϕ . More specifically,

$$\phi = \sum_{k=0}^{\infty} (-\mathbf{T})^k \frac{\lambda'}{\lambda}. \quad (31)$$

If the infinite sum in (31) is truncated to the first term ($k = 0$) then (13) becomes the Poisson score. Note that

$$\|\mathbf{T}\|_\infty \leq \sup_{\mathbf{u} \in W} \int_W |t(\mathbf{u}, \mathbf{v})| d\mathbf{v}.$$

Hence, a sufficient condition for the validity of the Neumann series expansion is

$$\sup_{\mathbf{u} \in W} \lambda(\mathbf{u}; \boldsymbol{\beta}) \int_{\mathbb{R}^2} |g(\mathbf{r}) - 1| d\mathbf{r} < 1. \quad (32)$$

Condition (32) roughly requires that $g(\mathbf{r}) - 1$ does not decrease too slowly to zero and/or that λ is moderate. For example, suppose that g is the Thomas process pair correlation function (30). Then,

$$\int_{\mathbb{R}^2} |g(\mathbf{r}) - 1| d\mathbf{r} = \frac{1}{4\pi\kappa\omega^2} \int_{\mathbb{R}^2} \exp\left(-\frac{\|\mathbf{r}\|^2}{4\omega^2}\right) d\mathbf{r} = 1/\kappa$$

and (32) is equivalent to $\sup_{\mathbf{u} \in W} \lambda(\mathbf{u}; \boldsymbol{\beta}) < \kappa$. In this case, Condition (32) can be quite restrictive. However, the Neumann series expansion is not essential for our approach and we use it only for checking the conditions for asymptotic results; see Section 1 in the supplementary material.