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

Adrian Baddeley, Jean-François Coeurjolly, Ege Rubak
and Rasmus Waagepetersen

A logistic regression estimating function for spatial Gibbs point processes

No. 02, February 2013

A logistic regression estimating function for spatial Gibbs point processes

Adrian Baddeley[1], Jean-François Coeurjolly[2], Ege Rubak[3] and Rasmus Waagepetersen[4]

[1] CSIRO Mathematics, Informatics and Statistics, Perth, Western Australia, `Adrian.Baddeley@csiro.au`
[2] Laboratory Jean Kuntzmann, Grenoble University, France, `Jean-Francois.Coeurjolly@upmf-grenoble.fr`
[3,4] Department of Mathematical Sciences, Aalborg University, `rubak@math.aau.dk`, `rw@math.aau.dk`

Abstract

We propose a computationally efficient logistic regression estimating function for spatial Gibbs point processes. The sample points for the logistic regression consist of the observed point pattern together with a random pattern of dummy points. The estimating function is closely related to the pseudolikelihood score. However, unlike common implementations of maximum pseudolikelihood, our approach does not suffer from bias due to numerical quadrature. The developed method is implemented in `R` code and will be added to future versions of the package `spatstat`. We demonstrate its efficiency and practicability on a real dataset and in a simulation study. Finally, focusing on stationary models, we prove that the estimator derived from the estimating function is strongly consistent and satisfies a central limit theorem. Moreover, we provide a consistent estimate of the asymptotic covariance matrix which allows to construct asymptotic confidence intervals.

Keywords: confidence intervals, estimating functions, exponential family models, Georgii-Nguyen-Zessin formula, logistic regression, pseudolikelihood.

1 Introduction

Spatial Gibbs and Markov point processes form major classes of models for spatial dependence in point patterns. For such models, popular options for parameter estimation include maximum likelihood (e.g. Ogata and Tanemura, 1981; Møller and Waagepetersen, 2004), maximum pseudolikelihood (e.g. Besag, 1977; Jensen and Møller, 1991; Baddeley and Turner, 2000; Billiot et al., 2008) and Takacs-Fiksel (e.g. Takacs, 1983; Fiksel, 1984; Billiot, 1997; Coeurjolly et al., 2012) estimation. For all three methods, the associated estimating functions are unbiased.

However, in practice, approximate versions of these estimating functions are almost always used. In the likelihood function the normalizing constant is not available in a tractable form and it is typically approximated by stochastic methods like Markov chain Monte Carlo (MCMC) (Ripley, 1979; Huang and Ogata, 1999; Geyer, 1999; Møller and Waagepetersen, 2004). The score of the pseudolikelihood and the Takacs-Fiksel estimating function contain an integral over the (typically two or three-dimensional) spatial domain where the point process is observed, and this must usually be approximated using numerical quadrature.

Maximum pseudolikelihood and Takacs-Fiksel methods offer enormous savings in computation time, compared to MCMC maximum likelihood estimation. Another advantage is that they can often be implemented using standard software for Generalized Linear Models (GLMs), with all the attached benefits of numerical stability, efficient optimisation procedures, and flexible model specification by the user. However, lingering doubts remain about the bias inherent in these methods, due to the numerical approximations. Indeed, experiments suggest that the bias can be substantial.

One strategy for numerical approximation is to discretise the spatial domain onto a fine grid of pixels (Tukey, 1972) and to consider the random field of binary variables indicating presence or absence of points in the pixels. This is used extensively in Geographical Information Systems (GIS) to fit spatial Poisson process models (Agterberg, 1974; Bonham-Carter, 1995; Baddeley et al., 2010; Warton and Shepherd, 2010). The discrete approximation to the Poisson process likelihood is a binomial regression based on the binary presence/absence variables, making it possible to rely on standard software. Approximation error can be controlled using a fine discretisation, but this leads to numerical instability, arising because the overwhelming majority of pixels do not contain a data point. In practice, the approximation is further modified by using only a randomly-selected subset of the absence pixels. The pixel discretisation approach can be extended to Gibbs processes (Clyde and Strauss, 1991), although this has not been widely used in practice. For a given choice of grid the binary random field pseudolikelihood again takes the form of a logistic regression likelihood. The spatial point process pseudolikelihood function may be viewed as a limit of binary random field pseudolikelihood functions (Besag, 1975, 1977; Besag et al., 1982; Clyde and Strauss, 1991).

Another popular strategy for numerical approximation is the sparse quadrature approximation pioneered by Berman and Turner (1992) for maximum likelihood estimation of spatial Poisson processes and extended to maximum pseudolikelihood estimation of Gibbs processes by Baddeley and Turner (2000). The approximate pseudolikelihood is equivalent to a Poisson loglinear regression likelihood which can be implemented using standard GLM software. The sparse quadrature approximation involves a sum over the observed data points together with a set of "dummy" points. While it was originally envisaged that the dummy points might be generated at random (Berman and Turner, 1992; Baddeley and Turner, 2000), the standard software implementation in the `spatstat` package (Baddeley and Turner, 2005) generates a regular grid of dummy points if no dummy points are provided by the user.

When unbiased estimating functions are approximated using *deterministic* numerical approximations, the resulting estimating functions are not in general unbiased, and it may be difficult to quantify the error due to the approximations. It can therefore be advantageous to replace deterministic numerical quadrature with Monte Carlo approximations which can provide both unbiased results and the possibility of quantifying the Monte Carlo error. Rathbun et al. (2007) and Waagepetersen (2007) introduced Monte Carlo approximation based on random dummy points for the cases of maximum likelihood estimation of Poisson processes and composite likelihood for Neyman-Scott point processes, respectively. The estimating function in Waagepetersen (2007) obtained with so-called 'Dirichlet' weights (Baddeley and Turner, 2000) takes the form of a conditional logistic regression, equivalent to the case-control conditional likelihoods considered for epidemiological data in Diggle and Rowlingson (1994), and closely related to logistic regression in GIS where the absence pixels are subsampled (Bonham-Carter, 1995).

In this paper we introduce a logistic regression estimating function for the wide class of Gibbs point processes and we provide a detailed study of its properties both from a practical and a theoretical point of view. The logistic regression estimating function has several advantages. First, the estimating function is unbiased. Second, since the estimating function takes the form of a logistic regression score, parameter estimates can easily be obtained using existing software for GLMs. Third, due to a decomposition of variance it is possible to quantify the proportion of variance due to using random dummy points. By the third property, the user can establish how large a dummy point sample is needed in order to achieve a certain level of accuracy (e.g. that variance due to random dummy points does not exceed a certain percentage of the total variance). Apart from being attractive in its own right, our estimating function can be further motivated by its close relation to pseudolikelihood and a time-invariance estimating function obtained from Barker dynamics.

A brief background on spatial point processes in Section 2 is followed in Section 3 by an introduction and discussion of our logistic regression estimating function. Section 4 presents our asymptotic results while Section 5 contains simulation studies and a data example. Proofs and technical details are deferred to the appendices.

2 Background on spatial point processes and notation

Let Λ be a Borel subset of $\mathbb{R}^d$ and let $\mathbb{M}$ be an arbitrary space (e.g. a countable set or a Borel subset of $\mathbb{R}^k$ for some $k \geq 1$). A marked point is a pair $u = (\overline{u}, m)$ where $\overline{u} \in \Lambda$ and $m \in \mathbb{M}$ represent respectively the location and some other characteristic of an object observed in Λ. For example $\overline{u}$ might be the spatial location of a tree, and m its diameter at breast height. A marked point process Y on $\mathbb{S} = \Lambda \times \mathbb{M}$ is a locally finite random subset of $\Lambda \times \mathbb{M}$ meaning that $Y \cap (W \times \mathbb{M})$ is finite whenever W is a bounded subset of Λ. The set of all locally finite marked point configurations is denoted by Ω. We equip $\Lambda \times \mathbb{M}$ with the product measure $\mathcal{L}^d \otimes \mu$ where $\mathcal{L}^d$ is the Lebesgue measure on $\mathbb{R}^d$ and μ is a probability distribution on the mark space $\mathbb{M}$ that serves as the reference measure on $\mathbb{M}$. For simplicity we write $\mathrm{d}u = \mathcal{L}^d(\mathrm{d}\overline{u}) \otimes \mu(\mathrm{d}m)$ for a marked point $u = (\overline{u}, m)$.

The region Λ can be bounded or unbounded depending on the application. However, we consider $\Lambda = \mathbb{R}^d$ in Section 4 to derive asymptotic results. The notation W will be reserved for a bounded Borel set of $\mathbb{R}^d$. For a marked point configuration $y \in \Omega$, $y_W = y \cap (W \times \mathbb{M})$, i.e. the subset of marked points where the 'location part' falls in W. Finally, $|\cdot|$ will be used to denote, depending on the context, either the cardinality of a finite set or the volume of a bounded Borel set or the supremum norm of a vector. We let $n(y)$ denote the (possibly infinite) number of points in $y \in \Omega$. We only consider the case of a locally finite Y which has an intensity function α with respect to $\mathcal{L}^d \otimes \mu$. Then Campbell's Theorem holds (e.g. Møller and Waagepetersen (2004)):

$$\mathrm{E} \sum_{u \in Y} h(u) = \int h(u)\alpha(u)\,\mathrm{d}u \tag{2.1}$$

for any real Borel function h defined on $\mathbb{S}$ such that $h\alpha$ is absolutely integrable (with respect to $\mathcal{L}^d \otimes \mu$).

2.1 Marked Gibbs point processes

Let $\lambda_\theta(u, X)$ be the Papangelou conditional intensity of a spatial marked Gibbs point process X on $\mathbb{S}$ parametrized by $\theta \in \Theta \subset \mathbb{R}^p$ (for some $p \geq 1$). Intuitively, $\lambda_\theta(u, X)\mathrm{d}u$ is the conditional probability that a marked point u occurs in a small ball of volume $\mathrm{d}u$ around u given the rest of the point process X. Implicitly, we assume that the conditional intensity exists in the sense that the fundamental Georgii-Nguyen-Zessin equation (GNZ formula), which also characterizes the distribution of X, is satisfied:

$$\mathrm{E} \sum_{u \in X} f(u, X \setminus u) = \mathrm{E} \int f(u, X)\lambda_\theta(u, X)\mathrm{d}u \tag{2.2}$$

for non-negative functions $f : \mathbb{S} \times \Omega \to \mathbb{R}$ (see Georgii, 1976, for a general presentation). We therefore exclude in our study certain non-hereditary Gibbs processes (Dereudre and Lavancier, 2009) as well as deterministic point patterns. For further background material and measure theoretical details on point processes, we refer to Daley and Vere-Jones (2003) and Møller and Waagepetersen (2004).

We will consider the case of a log-linear conditional intensity,

$$\lambda_\theta(u, X) = H(u, X)e^{\theta^\top t(u,X)}, \tag{2.3}$$

where $t(u, X) = (t_1(u, X), \ldots, t_p(u, X))^\top$ for functions $t_i : \mathbb{S} \times \Omega \to \mathbb{R}$, $i = 1, \ldots, p$ and where $H : \mathbb{S} \times \Omega \to [0, \infty)$ is a fixed, known function which may serve as a model offset or may be used to model a hard-core effect.

Remark 2.1 (Multitype point processes)**.** In the important special case of a multitype point process where $\mathbb{M}$ is a finite discrete set of K elements, say, μ is typically the uniform distribution on $\mathbb{M}$. However, we could also have used counting measure as a reference mark measure in which case the intensity function α_c with respect to Lebesgue-counting product measure would become $\alpha_c(u) = \alpha(u)/K$.

3 An unbiased logistic regression estimating function

Let W denote the bounded observation window of X. In this section, we assume $\Lambda = W$ implying that X is a finite point process.

3.1 Estimating function

Our logistic regression estimating function involves a 'dummy point process' D on $\mathbb{S}$ independent of X, with positive intensity function ρ. The estimating function is

$$s_W(X, D; \theta) = \sum_{u \in X_W} \frac{\rho(u)\lambda_\theta^{(1)}(u, X \setminus u)}{\lambda_\theta(u, X \setminus u)[\lambda_\theta(u, X \setminus u) + \rho(u)]} - \sum_{u \in D_W} \frac{\lambda_\theta^{(1)}(u, X)}{\lambda_\theta(u, X) + \rho(u)} \tag{3.1}$$

where $\lambda_\theta^{(1)}$ denotes the p-dimensional gradient vector of λ_θ with respect to θ.

By the GNZ formula (2.2) for X and the Campbell formula (2.1) for D given X, respectively, we obtain

$$\mathrm{E} \sum_{u \in X_W} \frac{\rho(u)\lambda_\theta^{(1)}(u, X \setminus u)}{\lambda_\theta(u, X \setminus u)[\lambda_\theta(u, X \setminus u) + \rho(u)]} = \mathrm{E} \int_{W \times \mathbb{M}} \frac{\rho(u)\lambda_\theta^{(1)}(u, X)}{\lambda_\theta(u, X) + \rho(u)} \mathrm{d}u \tag{3.2}$$

and

$$\mathrm{E}\Big[\sum_{u \in D_W} \frac{\lambda_\theta^{(1)}(u, X)}{\lambda_\theta(u, X) + \rho(u)} \Big| X \Big] = \int_{W \times \mathbb{M}} \frac{\rho(u)\lambda_\theta^{(1)}(u, X)}{\lambda_\theta(u, X) + \rho(u)} \mathrm{d}u. \tag{3.3}$$

It follows that $s_W(X, D; \theta)$ is an unbiased estimating function (where the expectation is taken over both X and D).

The score (3.1) is the derivative of the function $\mathsf{LRL}_W(X; \theta)$ (where LRL stands for the logistic regression likelihood) given by

$$\mathsf{LRL}_W(X; \theta) = \sum_{u \in X_W} \log \frac{\lambda_\theta(u, X \setminus u)}{\lambda_\theta(u, X \setminus u) + \rho(u)} + \sum_{u \in D_W} \log \frac{\rho(u)}{\lambda_\theta(u, X) + \rho(u)} \tag{3.4}$$

which conditional on $X \cup D$ is formally equivalent to the log-likelihood function for Bernouilli trials $Y(u) = \mathbf{1}[u \in X]$ with $P(Y(u) = 1) = \lambda(u, X \setminus u; \theta)/[\lambda_\theta(u, X \setminus u) + \rho(u)]$. Note in this connection that $\lambda_\theta(u, X) = \lambda_\theta(u, X \setminus u)$ for $u \notin X$. If the Papangelou conditional intensity is log-linear (2.3), then our estimating function precisely takes the form of the score of a logistic regression with $-\log \rho(u)$ as an offset term. This means that estimation can be implemented straightforwardly using standard software for GLMs. Further, for log-linear models, (3.4) is a concave function of θ since

$$\frac{\mathrm{d}}{\mathrm{d}\theta^\top} s_W(X, D; \theta) = - \sum_{u \in (X \cup D)_W} t(u, X \setminus u) t(u, X \setminus u)^\top \frac{\rho(u)\lambda_\theta(u, X \setminus u)}{[\lambda_\theta(u, X \setminus u) + \rho(u)]^2} \tag{3.5}$$

is negative-semidefinite.

Note that (3.1) is applicable both for homogeneous and inhomogeneous Gibbs point processes. We also emphasise that (3.2) and (3.3) are true if $W \subset \Lambda$ (and even if $\Lambda = \mathbb{R}^d$) but in such a case the score (3.1) cannot be computed since it depends on $X_{\Lambda \setminus W}$, and an edge correction such as the border correction should then be applied. This is done in Section 4 where we focus on stationary Gibbs models and stationary dummy point processes in order to derive the asymptotic distribution of the logistic regression estimate $\widehat{\theta}$ obtained from (3.1) and an estimate of the asymptotic covariance matrix for $\widehat{\theta}$.

3.2 Relation to existing methods

Below we comment on the relation between our logistic regression estimating function and various existing alternatives.

3.2.1 Pseudolikelihood

If we rearrange (3.1) as

$$s_W(X, D; \theta) = \sum_{u \in X_W} \frac{\lambda_\theta^{(1)}(u, X \setminus u)}{\lambda_\theta(u, X \setminus u)} - \sum_{u \in (X \cup D)_W} \frac{\lambda_\theta^{(1)}(u, X \setminus u)}{\lambda_\theta(u, X \setminus u) + \rho(u)} \tag{3.6}$$

and apply the GNZ formula and (3.3) to the last term in (3.6), we obtain

$$\mathrm{E} \sum_{u \in (X \cup D)_W} \frac{\lambda_\theta^{(1)}(u, X \setminus u; \theta)}{\lambda_\theta(u, X \setminus u) + \rho(u)} = \mathrm{E} \int_{W \times \mathbb{M}} \lambda_\theta^{(1)}(u, X) \mathrm{d}u. \tag{3.7}$$

Thus, if the last term in (3.6) is replaced by its integral compensator $\int_{W \times \mathbb{M}} \lambda_\theta^{(1)}(u, X) \mathrm{d}u$, the score

$$\sum_{u \in X_W} \frac{\lambda_\theta^{(1)}(u, X \setminus u)}{\lambda_\theta(u, X \setminus u)} - \int_{W \times \mathbb{M}} \lambda_\theta^{(1)}(u, X) \mathrm{d}u \tag{3.8}$$

of the pseudolikelihood is obtained (Jensen and Møller, 1991). Hence our estimating function may be viewed as a Monte Carlo approximation of the pseudolikelihood score to which it converges (in mean square) when $\inf_{u \in W} \rho(u) \to \infty$. Similarly, the logistic regression log-likelihood (3.4) is equivalent to

$$\sum_{u \in X_W} \log \lambda_\theta(u, X \setminus u) + \sum_{u \in (X \cup D)_W} \log \frac{\rho(u)}{\lambda_\theta(u, X) + \rho(u)}$$

which converges (in mean square) to the log-pseudolikelihood as $\inf_{u \in W} \rho(u) \to \infty$.

3.2.2 Time-invariance estimating functions

As detailed in Appendix A the following unbiased estimating function can be derived as a time-invariance estimating function (Baddeley, 2000) associated with Barker dynamics (Barker, 1969):

$$b_W(X;\theta,B) = \sum_{u\in X_W} \frac{B(u)\lambda_\theta^{(1)}(u,X\setminus u)}{\lambda_\theta(u,X\setminus u)[\lambda_\theta(u,X\setminus u)+B(u)]} - \int_{W\times\mathbb{M}} \frac{B(u)\lambda_\theta^{(1)}(u,X)}{\lambda_\theta(u,X)+B(u)}\mathrm{d}u \tag{3.9}$$

where $B(\cdot)$ is a positive function. Asymptotic results for time-invariance estimators are available (Qi, 2008). If $\inf_{u\in W} B(u) \to \infty$ then (3.9) converges (a.s.) to the pseudolikelihood score (3.8).

If the integral is replaced by a Monte Carlo approximation, the approximate Barker estimating function

$$\sum_{u\in X_W} \frac{B(u)\lambda_\theta^{(1)}(u,X\setminus u)}{\lambda_\theta(u,X\setminus u)[\lambda_\theta(u,X\setminus u)+B(u)]} - \sum_{u\in D_W} \frac{B(u)\lambda_\theta^{(1)}(u,X)}{\rho(u)[\lambda_\theta(u,X)+B(u)]} \tag{3.10}$$

takes the form of a logistic regression score and coincides with (3.1) if $B = \rho$. Simulation results in Section 5.1 indicate in the stationary case that for a given ρ it is optimal to let $B = \rho$. This is further supported by an analogy to results by Rathbun (2012) which imply that $B = \rho$ is optimal in the case of an inhomogeneous Poisson point process. This also suggests that it is sub-optimal to approximate (3.8) by replacing the integral term by the simple Monte Carlo approximation $\sum_{u\in D_W} \lambda_\theta^{(1)}(u,X)/\rho(u)$ since this corresponds to (3.10) with $B(u) = \infty$.

3.2.3 Berman-Turner device

Baddeley and Turner (2000) extended the Berman and Turner (1992) device for spatial Poisson processes to the case of Gibbs processes. They approximated the integral on the right hand side of (3.7) using numerical quadrature with quadrature points given by the union of data points and (possibly random) dummy points Q. For each $u \in X \cup Q$ we denote the associated quadrature weight $w(u)$. Then (3.8) is approximated by

$$\begin{aligned} &\sum_{u\in X_W} \frac{\lambda_\theta^{(1)}(u,X\setminus u)}{\lambda_\theta(u,X\setminus u)} - \sum_{u\in X_W\cup Q_W} \lambda_\theta^{(1)}(u,X\setminus u)w(u) \\ &= \sum_{u\in(X\cup Q)_W} w(u)\Big[Y(u)\frac{\lambda_\theta^{(1)}(u,X\setminus u)}{\lambda_\theta(u,X\setminus u)} - \lambda_\theta^{(1)}(u,X\setminus u)\Big] \end{aligned} \tag{3.11}$$

where $Y(u) = 1[u \in X]/w(u)$. This is formally equivalent to the score function of weighted Poisson regression with log link. In general, for $u \in X$, $\lambda_\theta(u,X) \neq \lambda_\theta(u,X\setminus u)$. This 'discontinuity' (Baddeley and Turner, 2000) can lead to considerable bias if the number of dummy points is not high (see Section 5). The `ppm` function in the `R` package `spatstat` (Baddeley and Turner, 2005, 2006) implements the above approach to approximate maximum pseudolikelihood estimation.

3.2.4 Advantages of the logistic likelihood

Apart from the fact that it is unbiased and is easy to implement using GLM software, the logistic regression estimating function (3.1) has other advantages over methods that require numerical integration.

A powerful advantage is that the logistic regression estimating function typically requires fewer evaluations of the conditional intensity as demonstrated in the practical examples in Section 5. In higher dimensions (including higher spatial dimensions, space-time, and marked point patterns) numerical integration becomes increasingly costly due to the 'curse of dimensionality'. Even in two dimensions there are models such as the area-interaction process (Baddeley and Lieshout, 1995) for which the evaluation of the conditional intensity is so costly that there is a substantial advantage in using the logistic regression estimating function.

Even if we did have sufficient computing resources to perform numerical integration with a high density of quadrature points, this might still be unsatisfactory: it can cause computational and numerical problems such as overflow/underflow, numerical instability, and the failure of the Taylor approximation underpinning statistical procedures (Hauck and Donner, 1977).

In a field experiment in forestry or ecology, the data required for the logistic regression approach may be easier to collect. Following a remark by Comas and Mateu (2011) in case of Takacs-Fiksel estimation, to fit (for example) a stationary Strauss process with interaction radius R (see Section 5.1) using logistic regression, we only need to count the number of R-neighbours of each data or dummy point, which could be performed in the field. To fit the same model using maximum pseudolikelihood we would need to compute the same information for a fine grid of dummy points, which would typically not be feasible in the field. Instead we would have to record the exact spatial location of each data point, and perform the equivalent calculation in software.

3.3 Choice of dummy point distribution

For ease of implementation the dummy point process D should be both easy to simulate and also mathematically tractable. Obvious choices are marked versions of a homogeneous Poisson point process, a binomial point process or a stratified binomial point process on W. A stratified binomial point process is obtained by partitioning W into a regular grid and then generating independent binomial point processes within each of the grid cells. In particular, each of these binomial point processes might consist of just one random point. In the following we refer for convenience to a stratified single point binomial point process as a stratified point process. We typically assign independent random marks with a common density q with respect to the reference mark distribution μ. The intensity function of the marked dummy point process then becomes q times the constant intensity of the unmarked dummy point process. In particular, if $q = 1$ the mark distribution coincides with the reference mark distribution and the constant intensity ρ of the marked dummy point process then equals the intensity of the unmarked dummy point process. We discuss the choice of D in more detail in Section 4 and Section 5.

As a rough rule of thumb inspired by an analogous rule of thumb in `spatstat` we suggest to use $\rho = 4n(X_W)/|W|$. In our simulation studies this usually resulted in moderate additional variance due to using random dummy points. Moreover, this choice can be used as a starting point for a data driven approach to determine ρ (see next section and Section 5.4).

3.4 Variance decomposition

The score (3.1) can be rewritten as the sum of two random vectors $T_{1,W}(X)$ and $T_{2,W}(X, D)$ where $T_{1,W}(X)$ is given by (3.9) with $B(u) = \rho(u)$ and

$$\begin{aligned} T_{2,W}(X, D) &= s_W(X, D; \theta) - T_{1,W}(X) \\ &= \int_{W\times\mathbb{M}} \frac{\rho(u)\lambda_\theta^{(1)}(u, X)}{\lambda_\theta(u, X) + \rho(u)} \mathrm{d}u - \sum_{u\in D_W} \frac{\lambda_\theta^{(1)}(u, X)}{\lambda_\theta(u, X) + \rho(u)}. \end{aligned} \tag{3.12}$$

Under appropriate conditions (to be formalized in Section 4), the variance of $|W|^{1/2}\widehat{\theta}$ can be approximated by $S^{-1}GS^{-1}$ where $G = |W|^{-1} \operatorname{Var} s_W(X, D; \theta)$ and the sensitivity matrix S is the expected value

$$S = -|W|^{-1} \operatorname{E} \frac{\mathrm{d}}{\mathrm{d}\theta^\top} s_W(X, D; \theta)$$

of the negated and normalized version of the derivative given in (3.5). Since $T_{1,W}$ is a centered random vector depending only on X and since the expectation of $T_{2,W}$ given X is zero, $T_{1,W}$ and $T_{2,W}$ are uncorrelated. Hence the variance of $|W|^{1/2}\widehat{\theta}$ can be decomposed into the sum of

$$\Sigma_1 = S^{-1}G_1S^{-1} \quad \text{and} \quad \Sigma_2 = S^{-1}G_2S^{-1}$$

where $G_1 = |W|^{-1} \operatorname{Var} T_{1,W}(X; \theta)$ and $G_2 = |W|^{-1} \operatorname{Var} T_{2,W}(X, D; \theta)$. In the case where ρ is constant, asymptotic results in Section 4 and simulation studies in Section 5 suggest that Σ_2 is approximately proportional to $1/\rho$. Furthermore, in the simulation studies the estimated Σ_1 is close to the covariance matrix of the MPLE. We can thus quantify the increase in estimation variance due to the use of the random dummy points D relative to the estimation variance of the exact MPLE. This also allows us to determine how large a ρ should be used in order to achieve a certain accuracy relative to the variance of the MPLE (see Section 5.4).

4 Theoretical results for stationary models

In this section we focus on exponential family models of stationary marked Gibbs point processes that are defined on $\mathbb{S} = \mathbb{R}^d \times \mathbb{M}$ ($\Lambda = \mathbb{R}^d$) and we derive asymptotic properties for the logistic regression estimate.

By stationarity, the conditional intensity is translation-equivariant, i.e.

$$\lambda_\theta(u, x) = \lambda_\theta(0, \tau_{\overline{u}} x) \tag{4.1}$$

where $\tau_{\overline{u}} x = \{(\overline{v} - \overline{u}, m) \mid (\overline{v}, m) \in x\}$ is the translation of the locations of x by the vector $-\overline{u}$. We assume that λ_θ has finite interaction range $R \geq 0$, i.e.

$$\lambda_\theta(u, x) = \lambda_\theta(u, x_{\mathcal{B}(\overline{u}, R)}) \tag{4.2}$$

where $\mathcal{B}(\overline{u}, R)$ is the Euclidean ball centered at $\overline{u}$ with radius R. We further assume that X is observed in a sequence of bounded observation windows $W_n^+ \subset \mathbb{R}^d$, $n \geq 1$.

Under the assumption (4.2) of finite range, a logistic regression estimate $\widehat{\theta}_n$ of θ is obtained for each n by maximizing $\mathsf{LRL}_{W_n}(X;\theta)$, where $W_n = W_n^+ \ominus R$ is the erosion of W_n^+ by R:

$$W_n = \{\overline{v} \in W_n^+ \mid \mathcal{B}(\overline{v}, R) \subseteq W_n^+\}. \tag{4.3}$$

This corresponds to using minus sampling to correct for edge effects (Miles, 1974). We assume that $(W_n)_{n\geq 1}$ is a sequence of increasing cubes such that $W_n \to \mathbb{R}^d$ as $n \to \infty$. Appendix C lists some further technical assumptions. None of these are very restrictive and they are satisfied by a large class of models including the Strauss process, its multiscale and multitype generalizations, Geyer's triplet process, the area-interaction process and Geyer's saturation process.

In the following we consider three different choices of the stationary marked dummy point process D of constant intensity $\rho > 0$. In all cases the marks are assigned independently of the locations of the points according to the reference mark distribution μ. First, for the homogeneous marked Poisson process $\mathcal{P}(\mathbb{R}^d, \rho)$ the locations constitute a homogeneous Poisson process. Second, for the marked binomial point process we assume that $\rho|W_n|$ is integer. The marked binomial point process D_n on W_n then consists of $\rho|W_n|$ uniform and independent marked points with locations in W_n. In case of marked binomial dummy points, we abuse notation and let $D = \cup_{n=1}^{\infty}\{D_n\} \sim \mathcal{B}(\mathbb{R}^d, \rho)$ and $D_{W_n} = D_n$. Finally, the marked stratified point process on $\mathbb{S}$ requires a more detailed definition:

Definition 4.1. *Let $\mathbb{R}^d$ be decomposed as $\cup_{k\in\mathbb{Z}^d} C_k$ where the cells C_k are disjoint cubes centered at $k/\rho^{1/d}$ with volume $1/\rho$. For $k \in \mathbb{Z}^d$ let $U_k = (\overline{U}_k, M_k)$ where the random point $\overline{U}_k$ is uniform on C_k, $M_k \sim \mu$ and all U_k and M_k are independent. Then $D = \cup_{k\in\mathbb{Z}^d}\{U_k\}$ is referred to as a marked stratified binomial point process $\mathcal{SB}(\mathbb{R}^d, \rho)$ on $\mathbb{S}$.*

The following result is required in our proofs.

Proposition 4.2. *The second order product density $\rho^{(2)}$ of the marked stratified binomial point process exists and is given for any $u, v \in \mathbb{S}$ by*

$$\begin{aligned}\rho^{(2)}(u,v) &= \rho^2 \sum_{k\in\mathbb{Z}^d} \mathbf{1}(\overline{u} \in C_k, \overline{v} \in \mathbb{R}^d \setminus C_k) \\ &= \rho^2\, \mathbf{1}(\overline{u} \textit{ and } \overline{v} \textit{ not in the same cell})\end{aligned} \tag{4.4}$$

Let $\theta^\star$ denote the true parameter vector. The score $s_{W_n}(X, D; \theta^\star)$ evaluated at $\theta^\star$ is the sum of $T_{1,W_n}(X)$ and $T_{2,W_n}(X, D)$ given by (3.9) and (3.12) with $\theta = \theta^\star$ where now the intensity function ρ is just a constant, that is

$$T_{1,W_n}(X) = \sum_{u\in X_{W_n}} w_{\theta^\star}(u, X\setminus u) - \int_{W_n\times\mathbb{M}} w_{\theta^\star}(u,X)\lambda_{\theta^\star}(u,X)\mathrm{d}u \tag{4.5}$$

$$T_{2,W_n}(X,D) = \int_{W_n\times\mathbb{M}} w_{\theta^\star}(u,X)\lambda_{\theta^\star}(u,X)\mathrm{d}u - \sum_{u\in D_{W_n}} \frac{w_{\theta^\star}(u,X)\lambda_{\theta^\star}(u,X)}{\rho} \tag{4.6}$$

where for any $\theta \in \Theta$, $u \in \mathbb{S}$ and $x \in \Omega$

$$w_\theta(u,x) = \frac{\rho\, t(u,x)}{\lambda_\theta(u,x) + \rho}$$

with $t(u,x)$ as given in (2.3). Each component of the vector $T_{1,W_n}(X)$ is a special case of innovations for spatial Gibbs point processes introduced by Baddeley et al. (2005) with variances studied by Baddeley et al. (2008) and asymptotic results provided by Coeurjolly and Lavancier (2012) and Coeurjolly and Rubak (2012). Based on these tools, we show in Appendix D.2 that as $n \to \infty$

$$|W_n|^{-1/2} T_{1,W_n}(X) \xrightarrow{d} \mathcal{N}(0, G_1) \tag{4.7}$$

where $G_1 = \sum_{i=1}^{3} A_i(w_{\theta^\star}, w_{\theta^\star})$ and where for $i = 1,2,3$ the $p \times p$ matrix $A_i(g,h)$ is defined for two functions $g,h : \mathbb{S} \times \Omega \to \mathbb{R}^p$ by

$$\begin{aligned}
A_1(g,h) &= \mathrm{E}\left[g(0^M, X) h(0^M, X)^\top \lambda_{\theta^\star}(0^M, X) \right] \\
A_2(g,h) &= \mathrm{E} \int_{B(0,R)\times\mathbb{M}} g(0^M, X) h(v, X)^\top (\lambda_{\theta^\star}(0^M, X)\lambda_{\theta^\star}(v, X) - \lambda_{\theta^\star}(\{0^M, v\}, X)) \mathrm{d}v \\
A_3(g,h) &= \mathrm{E} \int_{B(0,R)\times\mathbb{M}} \Delta_v g(0^M, X) \Delta_{0^M} h(v, X)^\top \lambda_{\theta^\star}(\{0^M, v\}, X) \mathrm{d}v
\end{aligned}$$

where $0^M = (0, M)$ with $M \sim \mu$ and where for $\theta \in \Theta$, $u, v \in \mathbb{S}$,

$$\begin{aligned}
\lambda_\theta(\{u,v\}, X) &= \lambda_\theta(u, X \cup v)\lambda_\theta(v, X) = \lambda_\theta(v, X \cup u)\lambda_\theta(u, X) \\
\Delta_v g(u, X) &= g(u, X \cup v) - g(u, X).
\end{aligned}$$

Regarding the term T_{2,W_n} which involves the point process D, conditional on X, a Lindeberg central limit theorem is available. Using this we show in Appendix D.2 that given X,

$$|W_n|^{-1/2} T_{2,W_n}(X, D) \xrightarrow{d} \mathcal{N}(0, G_2) \tag{4.8}$$

where G_2 is given by

$$G_2 = \begin{cases} G_2^{\mathrm{p}} = \frac{1}{\rho} \mathrm{E}[w^\lambda_{\theta^\star}(0^M, X) w^\lambda_{\theta^\star}(0^M, X)^\top] & \text{if } D \sim \mathcal{P}(\mathbb{R}^d, \rho) \\ G_2^{\mathrm{b}} = \frac{1}{\rho} \mathrm{Var}[w^\lambda_{\theta^\star}(0^M, X)] = \frac{1}{\rho} \mathrm{Var}[w^\lambda_{\theta^\star}(U_0, X)] & \text{if } D \sim \mathcal{B}(\mathbb{R}^d, \rho) \\ G_2^{\mathrm{sb}} = \frac{1}{\rho} \mathrm{E}\,\mathrm{Var}[w^\lambda_{\theta^\star}(U_0, X) \mid X] & \text{if } D \sim \mathcal{SB}(\mathbb{R}^d, \rho) \end{cases} \tag{4.9}$$

where for $\theta \in \Theta$, $u \in \mathbb{S}$ and $x \in \Omega$ we write $w^\lambda_{\theta,j}(u,x)$ for $w_{\theta,j}(u,x)\lambda_\theta(u,x)$, and U_0 is as in Definition 4.1. We can easily check that $G_2^{\mathrm{sb}} \le G_2^{\mathrm{b}} \le G_2^{\mathrm{p}}$ where for two square matrices A and B, $A \le B$ means that $B - A$ is a positive-semidefinite matrix. Therefore, among the three choices of random dummy points, the marked stratified point process seems to be the optimal choice.

The following almost sure convergence is also proved to hold as $n \to \infty$

$$\begin{aligned}
-|W_n|^{-1} \frac{\mathrm{d}}{\mathrm{d}\theta^\top} s_{W_n}(X, D; \theta^\star) \to S &= \mathrm{E}\left[\frac{\rho t(0^M, X) t(0^M, X)^\top}{\lambda_{\theta^\star}(0^M, X) + \rho} \lambda_{\theta^\star}(0^M, X) \right] \\
&= \frac{1}{\rho} A_1\big(w_{\theta^\star}\sqrt{\lambda_{\theta^\star} + \rho}, w_{\theta^\star}\sqrt{\lambda_{\theta^\star} + \rho}\big)
\end{aligned} \tag{4.10}$$

where S is the sensitivity matrix.

We also define for two functions $g, h : \mathbb{S} \times \Omega \to \mathbb{R}^p$ the computationally fast empirical estimates (Coeurjolly and Rubak, 2012) of $A_i(g,h)$ for $i = 1, \ldots, 3$ by

$$\widehat{A}_1(X,D,g,h) = \frac{1}{|W_n|} \sum_{u \in (X \cup D)_{W_n}} g(u, X \setminus u) h(u, X \setminus u)^\top \frac{\lambda_{\widehat{\theta}}(u, X \setminus u)}{\lambda_{\widehat{\theta}}(u, X \setminus u) + \rho} \tag{4.11}$$

$$\begin{aligned}\widehat{A}_2(X,g,h) = \frac{1}{|W_n|} \sum_{\substack{u,v \in X_{W_n} \\ u \neq v, \|\overline{u} - \overline{v}\| \leq R}} & g(u, X \setminus \{u,v\}) h(v, X \setminus \{u,v\})^\top \\ & \times \left(\frac{\lambda_{\widehat{\theta}}(u, X \setminus \{u,v\}) \lambda_{\widehat{\theta}}(v, X \setminus \{u,v\})}{\lambda_{\widehat{\theta}}(\{u,v\}, X \setminus \{u,v\})} - 1 \right)\end{aligned} \tag{4.12}$$

$$\widehat{A}_3(X,g,h) = \frac{1}{|W_n|} \sum_{\substack{u,v \in X_{W_n} \\ u \neq v, \|\overline{u} - \overline{v}\| \leq R}} \Delta_v g(u, X \setminus \{u,v\}) \Delta_u h(v, X \setminus \{u,v\})^\top. \tag{4.13}$$

Combining the above results in a standard fashion we obtain the following main result where we denote by $\widehat{\theta} = \widehat{\theta}_n(X,D)$ the logistic regression score estimate based on the observation of X on W_n^+.

Theorem 4.3. *As $n \to \infty$, $\widehat{\theta}$ is a strongly consistent estimate of $\theta^\star$. Assume that G_1 and G_2 are positive-definite matrices, then $|W_n|^{1/2}(\widehat{\theta} - \theta^\star)$ tends to a Gaussian distribution with covariance matrix $\Sigma = S^{-1}(G_1 + G_2)S^{-1}$ which is consistently estimated by $\widehat{\Sigma} = \widehat{S}^{-1}(\widehat{G}_1 + \widehat{G}_2)\widehat{S}^{-1}$ specified below. In other words, as $n \to \infty$*

$$|W_n|^{1/2} \widehat{\Sigma}^{-1/2} (\widehat{\theta} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, I_p).$$

The matrices $\widehat{S}$ and $\widehat{G}_1$ are defined by

$$\widehat{S} = \frac{1}{\rho} \widehat{A}_1\big(X, D, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}} + \rho}, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}} + \rho}\big) \tag{4.14}$$

$$\widehat{G}_1 = \widehat{A}_1(X, D, w_{\widehat{\theta}}, w_{\widehat{\theta}}) + \widehat{A}_2(X, w_{\widehat{\theta}}, w_{\widehat{\theta}}) + \widehat{A}_3(X, w_{\widehat{\theta}}, w_{\widehat{\theta}}). \tag{4.15}$$

The matrix G_2 is consistently estimated as follows:

(i) if $D \sim \mathcal{P}(\mathbb{R}^d, \rho)$ the estimate $\widehat{G}_2^{\mathrm{p}}$ is defined by

$$\frac{1}{\rho} \widehat{A}_1(X, D, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}}}, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}}}). \tag{4.16}$$

(ii) if $D \sim \mathcal{B}(\mathbb{R}^d, \rho)$ the estimate $\widehat{G}_2^{\mathrm{b}}$ is defined by

$$\begin{aligned}\frac{1}{\rho} \Big\{ & \kappa_n \widehat{A}_1(X, D, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}}}, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}}}) \\ & - \widehat{A}_1(X, D, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}}}, \sqrt{\lambda_{\widehat{\theta}}}) \widehat{A}_1(X, D, w_{\widehat{\theta}} \sqrt{\lambda_{\widehat{\theta}}}, \sqrt{\lambda_{\widehat{\theta}}})^T \Big\}\end{aligned} \tag{4.17}$$

where $\kappa_n = |W_n|^{-1} \sum_{u \in (X \cup D)_{W_n}} (\lambda_{\widehat{\theta}}(u, X \setminus u) + \rho)^{-1}$.

(iii) if $D \sim \mathcal{SB}(\mathbb{R}^d, \rho)$ the estimate $\widehat{G}_2^{\mathrm{sb}}$ is defined by

$$\frac{1}{2\rho^2|W_n|} \sum_{\substack{\ell \in \mathbb{Z}^d: \\ C_\ell \cap W_n \neq \emptyset}} \big(w_{\widehat{\theta}}^{\lambda}(U_\ell, X) - w_{\widehat{\theta}}^{\lambda}(U'_\ell, X)\big)\big(w_{\widehat{\theta}}^{\lambda}(U_\ell, X) - w_{\widehat{\theta}}^{\lambda}(U'_\ell, X)\big)^\top \quad (4.18)$$

where $D' = \cup_{k \in \mathbb{Z}^d} \{U'_k\}$ is a marked stratified point process independent of D, observed in W_n.

Remark 4.4 (Ergodicity)**.** We point out that Theorem 4.3 does not require the assumption that $P_{\theta^\star}$ is ergodic and thus can be applied even if the Gibbs measure is not unique and exhibits a phase transition.

Remark 4.5 (on κ_n)**.** The variable κ_n in (4.17) converges to 1 as $n \to \infty$. It has been introduced to ensure that the estimate $\widehat{G}_2^{\mathrm{b}}$ is a positive-semidefinite matrix. This can be checked easily using the Cauchy-Schwarz inequality. By definition, $\widehat{G}_2^{\mathrm{p}}$ and $\widehat{G}_2^{\mathrm{sb}}$ also fulfill this property.

Remark 4.6 (On the definition of $\widehat{S}$)**.** If we had followed the strategy proposed in Coeurjolly and Rubak (2012), the estimate of $A_1(w_{\theta^\star}, w_{\theta^\star})$ would have been based only on X. We include the dummy point pattern D to get a more accurate estimate and we emphasize that no new numerical computations are required since $\widehat{S}$, using (4.11), depends only on the quantities $t_j(u, X \setminus u)$ for $j = 1, \dots, p$ and $u \in (X \cup D)_{W_n}$ which have already been stored when computing the estimate $\widehat{\theta}$. The estimates (4.12) and (4.13) involve second order characteristics which have not been computed before and are therefore defined using only the data point pattern X.

Remark 4.7 (On the definition of $\widehat{G}_2$)**.** Following the previous remark, we have proposed, in the Poisson case and the binomial case, estimates of G_2 based on $X \cup D$. As for $\widehat{S}$, the estimates $\widehat{G}_2^{\mathrm{p}}$ and $\widehat{G}_2^{\mathrm{b}}$ do not involve new numerical computations. The estimate of G_2^{sb} is more awkward to handle and requires an extra dummy point process D'. As pairs of points are involved in (4.18), we could not include the data points without adding second order characteristics computations for X and this has not been investigated.

As D' is required to define (4.18), we may also use this point process to improve the accuracy of the logistic regression estimate in the stratified case. An aggregated estimate is

$$\widehat{\theta}^{\mathrm{agg}} = \tfrac{1}{2}\big(\widehat{\theta}_n(X, D) + \widehat{\theta}_n(X, D')\big) \quad (4.19)$$

and we obtain the following result.

Corollary 4.8. As $n \to \infty$, the aggregated estimate $\widehat{\theta}^{\mathrm{agg}}$ is a strongly consistent estimate of $\theta^\star$ and $|W_n|^{1/2}(\widehat{\theta}^{\mathrm{agg}} - \theta^\star)$ tends to a Gaussian distribution with covariance matrix $\Sigma^{\mathrm{agg}} = S^{-1}(G_1 + G_2/2)S^{-1}$ which is consistently estimated by

$\widehat{\Sigma}^{\text{agg}} = (\widehat{S}^{\text{agg}})^{-1}(\widehat{G}_1^{\text{agg}} + \widehat{G}_2^{\text{sb}}/2)(\widehat{S}^{\text{agg}})^{-1}$ where $\widehat{S}^{\text{agg}}$ and $\widehat{G}_1^{\text{agg}}$ are the $p \times p$ matrices

$$\widehat{S}^{\text{agg}} = \frac{1}{2\rho}\Big(\widehat{A}_1(X, D, w_{\widehat{\theta}^{\text{agg}}}\sqrt{\lambda_{\widehat{\theta}^{\text{agg}}} + \rho}, w_{\widehat{\theta}^{\text{agg}}}\sqrt{\lambda_{\widehat{\theta}^{\text{agg}}} + \rho})$$
$$+ \widehat{A}_1(X, D', w_{\widehat{\theta}^{\text{agg}}}\sqrt{\lambda_{\widehat{\theta}^{\text{agg}}} + \rho}, w_{\widehat{\theta}^{\text{agg}}}\sqrt{\lambda_{\widehat{\theta}^{\text{agg}}} + \rho})\Big)$$
$$\widehat{G}_1^{\text{agg}} = \tfrac{1}{2}\big(\widehat{A}_1(X, D, w_{\widehat{\theta}^{\text{agg}}}, w_{\widehat{\theta}^{\text{agg}}}) + \widehat{A}_1(X, D', w_{\widehat{\theta}^{\text{agg}}}, w_{\widehat{\theta}^{\text{agg}}})\big) + \sum_{i=2}^{3} \widehat{A}_i(X, w_{\widehat{\theta}^{\text{agg}}}, w_{\widehat{\theta}^{\text{agg}}}).$$

5 Simulation studies and data example

In this section we use simulations to study the performance of our logistic regression estimating function using different dummy point distributions. We also study the variance decomposition discussed in Section 3.4 and compare with the `spatstat` `ppm` implementation of the Berman-Turner device, c.f. Section 3.2.3 and equation (3.11), using the default settings of `ppm`. The performance of (3.10) with varying B is considered briefly. We further consider an application to an inhomogeneous multitype point pattern of cell centres from the mucous membrane of a rat. Finally we consider coverage properties of asymptotic confidence intervals and how asymptotic covariance matrix estimates may be used to determine a suitable value of ρ.

In the simulation study $W = \Lambda$ is the unit square and we specify the intensity of points using a parameter n_d. The default behaviour of `ppm` is to use a deterministic grid of dummy points where a one-dimensional n_d specifies the dimension of the grid in each spatial direction while a two-dimensional n_d specifies different dimensions of the grid in each direction. That is, the total number of dummy points is ${n_d}^2$. We have implemented the logistic regression estimate as an option for `ppm` and then ${n_d}^2$ specifies the expected number of dummy points in case of Poisson or binomial dummy points while n_d specifies the grid dimensions in case of stratified dummy points. The rule of thumb mentioned in Section 3.3 then corresponds to choosing the `ppm` default value $n_d = 2\sqrt{n}$ where n is the observed number of points.

In general one may expect that random dummy points lead to less bias. To enable a more fair comparison between `ppm` and the logistic regression estimating function we have therefore chosen to use stratified dummy points instead of a deterministic grid for `ppm`. However, we stress that very similar results were obtained using a deterministic grid. In the following we for ease of presentation refer to the described version of `ppm` as 'default' `ppm`.

5.1 Simulation study

In the simulation study we generate simulations of a Strauss process specified by a conditional intensity of the form (2.3) with $t(u, X) = (1, n_R(u, X))$ and $\theta = (\theta_1, \theta_2)$ where $n_R(u, X)$ is the number of neighbouring points in X of distance from u less than or equal to R. The parameter values used for the simulations are $\theta_1 = \log 1000$, $\theta_2 = \log 0.5$ and $R = 0.01$. The interaction distance R is treated as a known parameter. We generate 1000 simulations of the specified Strauss process and estimate θ using Poisson, binomial or stratified dummy points as well as with default ppm and with n_d equal to 10, 20, 40, 80 or 160. In the particular case of a homogeneous Strauss process it is also possible (Baddeley and Turner, 2000) to obtain the exact maximum pseudolikelihood estimate (MPLE) which we also consider for comparison (see Baddeley and Turner, 2013, for details on the implementation and properties of the exact MPLE).

Figure 1 shows boxplots of the parameter estimates for the different estimation methods. It is obvious that the default ppm estimate is strongly biased even

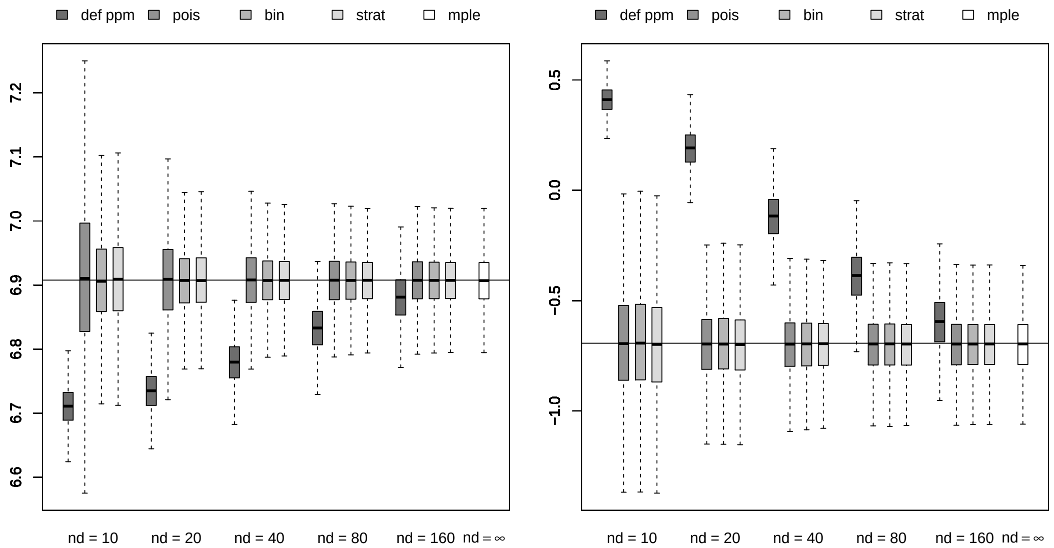

Figure 1: Boxplots of parameter estimates for increasing values of n_d for the different estimation methods. Horizontal lines show true parameter values. For comparison the exact MPLE is included as the rightmost box ($n_d = \infty$) in both subfigures.

with n_d=80 while the logistic regression estimate is essentially unbiased for all n_d. Moreover with n_d equal to 80 or 160 the variance of the logistic regression estimate seems very close to that of the exact MPLE. For small values of n_d the variance for default ppm is much smaller than for the logistic regression estimate. However, the strong bias means that the root mean square error (RMSE) is always largest for the ppm estimate, see Table 1. Table 1 also shows that for each n_d, the lowest estimation variance is obtained with stratified dummy points. With n_d equal to 80 or 160 and considering θ_2, the increase in RMSE relative to exact MPLE is just 0.9 % respectively 0.07 % when stratified dummy points are used. In the remainder we only consider stratified dummy points.

As mentioned in Section 3.4, the variance of the logistic regression estimator can be decomposed into two terms Σ_1 and Σ_2 where the last term is due to the use of the random dummy points. To investigate this we considered 500 simulations from the Strauss model and for each simulation we refitted the model 10 times using independent realizations of the dummy process. A one-way analysis of variance can then be used to partition the total estimation variance into Σ_1 and Σ_2. Results from the analysis of variance are given in Table 2. Here we use the generic notation $\sigma^2 = \sigma_1^2 + \sigma_2^2$ for the variance of a univariate parameter where σ_1^2 and σ_2^2 are extracted from the diagonals of Σ_1 and Σ_2. For n_d greater than or equal to 40 the relative increase in estimation standard deviation $(\sigma - \sigma_1)/\sigma_1$ due to using random dummy

Table 1: Percentage increase of the root mean squared errors of each estimator relative to the exact maximum pseudolikelihood estimate.

	θ_1				θ_2			
n_d	ppm	pois	bin	strat	ppm	pois	bin	strat
10	373.53	196.76	75.01	74.99	713.24	87.61	88.59	86.50
20	318.52	69.77	21.80	21.66	551.25	25.16	26.51	25.16
40	216.45	21.62	6.09	5.06	328.73	7.25	6.97	5.30
80	100.87	6.05	1.76	0.53	140.83	2.09	1.77	0.90
160	16.48	1.80	0.52	0.08	19.57	0.62	0.54	0.07

Table 2: Decomposition of variance for the logistic estimator using stratified dummy points with increasing values of n_d. The columns show the standard deviation of the estimator σ and the two contributions σ_1 and σ_2 as well as the percentage of variance increase due to random dummy points. For the exact MPLE the standard deviations are 0.042 and 0.136 for θ_1 and θ_2.

	θ_1				θ_2			
n_d	σ	σ_1	σ_2	$\frac{\sigma-\sigma_1}{\sigma_1}(\%)$	σ	σ_1	σ_2	$\frac{\sigma-\sigma_1}{\sigma_1}(\%)$
10	0.074	0.044	0.059	66.283	0.251	0.139	0.209	79.991
20	0.052	0.043	0.028	19.212	0.171	0.138	0.100	23.207
40	0.045	0.043	0.013	4.350	0.144	0.136	0.046	5.483
80	0.043	0.043	0.005	0.630	0.137	0.136	0.018	0.837
160	0.043	0.043	0.002	0.085	0.136	0.136	0.006	0.113

points is less than 5.5 %. For both parameters the standard error σ_1 quickly converges to a constant value (the standard error of the MPLE) as n_d increases. The reduction in variance as n_d increases thus mainly occurs for the σ_2^2 term. This justifies regarding σ_2^2 as the increase in variance additional to the MPLE variance due to using the random dummy points. Note also that σ_2 is approximately halved each time n_d is doubled.

We finally apply (3.10) with ρ corresponding to n_d=10 and $B(u)$ constant equal to $c\rho$ where c is 0.1, 0.5, 1, 10 or 100. Recall that $c = 1$ corresponds to our logistic regression estimating function. Based on the previous 1000 simulation of the Strauss model, Figure 2 shows the estimation variance for the varying values of c relative to the estimation variance for $c = 1$. The smallest estimation variance is obtained with $c = 1$.

5.2 Data example

We consider the mucous membrane data shown in Figure 1.3 in Møller and Waagepetersen (2004) and our analyses are inspired by Example 9.3 and Example 9.5 therein. The dataset used is a subset of the `mucosa` dataset available in `spatstat` and consists of the locations of two types of cells in an observation window $W = [0, 1] \times [0, 0.7]$.

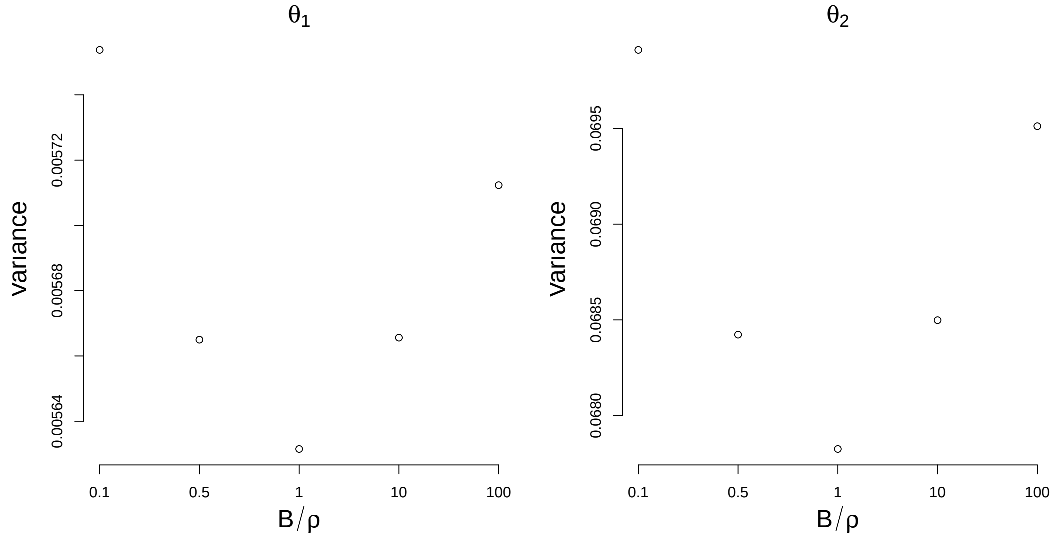

Figure 2: Variance of parameter estimates when using Barker dynamics with fixed ρ and different values of B/ρ.

There are 87 points of type 1 and 806 points of type 2. We fit an inhomogeneous multitype Strauss process with conditional intensity

$$\lambda_\theta(u, X) = \exp[q_m(y, \theta) + \theta_{11} n_R(u, X)]$$

where $u = (x, y, m)$, $m = 1$ or 2 denotes the cell type, $q_m(y, \theta)$, $m = 1, 2$, are fourth order polynomials with coefficients depending on the type of points and $\theta \in \mathbb{R}^{11}$ consists of the 10 polynomial regression coefficients and the interaction parameter $\theta_{11} \leq 0$. Notice that the polynomials only depond on y since the point pattern is considered homogeneous in the x-direction. As before $n_R(u, X)$ denotes the total number of neighbouring points and we use $R = 0.008$ as in Møller and Waagepetersen (2004). One question of interest is whether the two types of points share the same large scale polynomial trends, which is equivalent to that the two polynomials q_1 and q_2 only differ by a constant.

Recall that the logistic regression estimate requires a marked version of the stratified point process, which is generated by assigning a uniformly sampled random mark (1 or 2) to each point where each mark is independent of all other variables. In `spatstat` multitype point processes are specified with respect to counting measure on the mark space. To comply with this choice and following Remark 2.1 we therefore specify the dummy point intensity as 0.5 times $\rho = n_d^2/0.7$ in our implementation of the logistic regression estimating function, where we for this dataset use $nd = 60$ according to the rule of thumb.

To obtain confidence intervals for the fitted polynomials and the interaction parameter we use a parametric bootstrap based on 1000 simulations generated under the fitted model (still using n_d=60 when estimating parameters for each simulation). Furthermore, to enable empirical decomposition of estimation variance we use two replications of the dummy point process for each simulated dataset.

The estimated coefficients of the fourth order polynomials vary considerably but this is not so for the values of the resulting polynomials. We have therefore chosen to focus on estimated values of the polynomial for distinct y values in the range $[0, 0.7]$. Figure 3 shows the estimated polynomials without the constant term as well as bootstrap confidence intervals at the selected set of y values. This plot gives some indication that the two trends are significantly different.

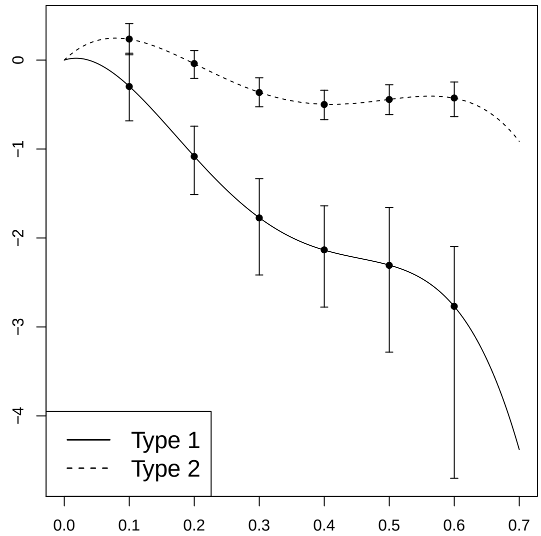

Figure 3: Fitted fourth order polynomials for the mucous membrane data without the constant term. Bootstrap confidence intervals are given at selected values.

The first column in Table 3 shows the estimated values of the polynomials for the selected y values and the estimate of the interaction parameter. The three next columns show the standard deviation σ of the parameter estimates, the standard deviation σ_1 due to T_1, the standard deviation due to T_2 and the relative increase $(\sigma - \sigma_1)/\sigma_1$ in standard deviation due to T_2. Comparing the fitted polynomials, the smallest standard deviations are obtained for the more abundant type 2 cells. For this reason also the largest relative increases (up to 16 %) in estimation standard error due to the random dummy points are obtained for the type 2 cells. We also applied our logistic regression estimate with n_d=120 to the simulated bootstrap datasets (see the three columns labeled $n_d = 120$ in Table 3) and this brings the maximal relative increase in standard deviation down to 4 %. Note again that doubling n_d leads to approximately halving σ_2.

To test the hypothesis of equal polynomials more formally we fitted the null model with a common (up to a constant) fourth order polynomial. We then generated 1000 simulations under the fitted null model and calculated for each simulation $-2 \log LR$ where LR is the ratio of the likelihoods for the logistic regressions corresponding to the null model and the original model. The 1000 values of this test statistic were between 0.1 and 16.3. The observed value of 28.6 is thus highly significant.

As a supplement to the simulation study in Section 5.1 we also applied default `ppm` with n_d=60 to the simulated bootstrap data sets generated previously under the fitted multitype Strauss model. Note that default ppm actually uses 2 times 3600 dummy points since each unmarked dummy point appears both with the mark 1 and

Table 3: The first five columns show estimated values of the two polynomials and the interaction parameter with decomposed standard deviations and relative increases for the logistic regression estimator with $n_d = 60$ based on 1000 simulations from the fitted model. The next three columns show the decomposition of the standard deviations and relative increases when $n_d = 120$ based on the same 1000 simulations. The final three columns show a decomposition of the standard deviations for $n_d = 60$ based on a Taylor approximation as well as the predicted relative increases for $n_d = 120$ (see Section 5.4 for details). All relative increases are given in percent.

	$n_d = 60$					$n_d = 120$			Taylor		
	est.	σ	σ_1	σ_2	$\frac{\sigma-\sigma_1}{\sigma_1}$	σ_1	σ_2	$\frac{\sigma-\sigma_1}{\sigma_1}$	σ_1	σ_2	$\frac{\sigma-\sigma_1}{\sigma_1}$
$q_1(0.1)$	6.00	0.20	0.19	0.05	3.61	0.19	0.02	0.81	0.21	0.05	0.72
$q_1(0.2)$	5.22	0.20	0.20	0.04	2.03	0.20	0.02	0.48	0.19	0.04	0.52
$q_1(0.3)$	4.53	0.27	0.26	0.04	1.33	0.26	0.02	0.30	0.26	0.04	0.30
$q_1(0.4)$	4.16	0.29	0.29	0.04	0.84	0.29	0.02	0.19	0.28	0.04	0.23
$q_1(0.5)$	3.99	0.41	0.40	0.04	0.56	0.40	0.02	0.15	0.37	0.04	0.16
$q_1(0.6)$	3.53	0.74	0.74	0.07	0.41	0.75	0.03	0.09	0.47	0.04	0.11
$q_2(0.1)$	7.80	0.09	0.08	0.05	15.62	0.08	0.02	3.80	0.08	0.05	4.18
$q_2(0.2)$	7.52	0.08	0.07	0.04	13.67	0.07	0.02	3.76	0.07	0.04	3.72
$q_2(0.3)$	7.20	0.08	0.08	0.04	10.92	0.07	0.02	2.59	0.07	0.04	2.95
$q_2(0.4)$	7.07	0.08	0.08	0.04	10.41	0.07	0.02	2.56	0.07	0.04	2.93
$q_2(0.5)$	7.12	0.09	0.08	0.04	10.56	0.08	0.02	2.82	0.07	0.04	2.89
$q_2(0.6)$	7.14	0.10	0.09	0.04	11.36	0.09	0.02	2.65	0.09	0.04	2.69
θ_{11}	−2.59	0.34	0.34	0.05	0.97	0.34	0.02	0.20	0.31	0.05	0.30

Table 4: Comparison of root mean squared error (RMSE) and bias relative to RMSE (in percent) for the four estimators.

	RMSE				**Relative bias (%)**			
	logistic		**ppm**		**logistic**		**ppm**	
	n_d=60	n_d=120	n_d=60	n_d=120	n_d=60	n_d=120	n_d=60	n_d=120
$q_1(0.1)$	0.20	0.19	0.20	0.18	−0.21	−1.00	−37.13	4.49
$q_1(0.2)$	0.21	0.20	0.21	0.20	−15.13	−14.80	−36.30	−10.48
$q_1(0.3)$	0.27	0.27	0.26	0.27	−16.26	−16.06	−11.70	−21.36
$q_1(0.4)$	0.29	0.29	0.29	0.29	−7.71	−7.73	−6.73	−10.62
$q_1(0.5)$	0.41	0.41	0.43	0.40	−11.56	−11.81	−24.34	−6.95
$q_1(0.6)$	0.80	0.81	0.77	0.83	−36.50	−36.40	−40.64	−34.91
$q_2(0.1)$	0.09	0.08	0.21	0.09	1.07	0.29	−94.33	−55.89
$q_2(0.2)$	0.08	0.07	0.18	0.08	−1.54	−2.89	−93.73	−49.12
$q_2(0.3)$	0.08	0.08	0.13	0.08	−1.99	−4.61	−84.55	−51.62
$q_2(0.4)$	0.08	0.08	0.11	0.08	−1.49	−3.58	−80.69	−45.91
$q_2(0.5)$	0.08	0.08	0.15	0.08	−1.98	−1.67	−87.02	−28.23
$q_2(0.6)$	0.10	0.09	0.15	0.09	−5.58	−3.96	−84.52	−33.64
θ_{11}	0.35	0.35	0.56	0.35	−19.72	−19.44	79.22	20.64

the mark 2. The `ppm` estimate with n_d=60 is strongly biased and to reduce the bias we also used standard `ppm` with n_d=120. Table 4 compares the different estimators in terms of root mean square error (RMSE) and bias (relative to the RMSE). The RMSEs for the logistic regression estimate with n_d equal to 60 and 120 and default `ppm` with n_d=120 are quite similar and in case of q_2 and θ_{11} much smaller than for `ppm` with n_d=60. However, for both versions of default `ppm` the proportion of RMSE due to bias is in general much larger than for the logistic method. For n_d equal to 120 this is especially the case for the estimated values of the q_2 polynomial where the RMSE is smallest.

The main drawback of using the default `ppm` is that it is not a priori clear how large n_d must be used to avoid severe bias and sometimes the required value of n_d may even be computationally prohibitive. In the present example we obtained reliable results for the logistic estimator with n_d=60 and the estimation for 1000 datasets took 2 minutes. For default `ppm` we need n_d=120 to avoid strong bias and in this case the 1000 estimations required over 20 minutes of computing time.

5.3 Coverage of approximate confidence intervals

In this section we study finite sample coverage properties of approximate confidence intervals based on the asymptotic normality demonstrated in Theorem 4.3. Simulations are generated from Strauss processes (see Section 5.1), multiscale (or piecewise) Strauss processes and Geyer's saturation processes with saturation threshold 1. The two latter classes of point processes are specified by conditional intensities of the form (2.3) with respectively $t(u,X) = (1, n_{R_1}(u,X), n_{R_2}(u,X) - n_{R_1}(u,X))$ and $t(u,X) = (1, \sum_{v \in X \cup u} \mathbf{1}(d(v, X \cup u) \leq R) - \sum_{v \in X} \mathbf{1}(d(v,X) \leq R))$ where $d(v,X)$

Table 5: Coverage rates for the logistic estimator using stratified dummy points with increasing values of n_d when W is a square with sidelength ℓ. The first column contains the average empirical intensities for the models. The results are based on 1000 realizations from each of the models.

		$\ell = 1$			$\ell = 2$		
	$n/\lvert W\rvert$	$n_d = 20$	$n_d = 40$	$n_d = 80$	$n_d = 20$	$n_d = 40$	$n_d = 80$
S1	87.5	93.9	95.5	94.7	94.7	95.7	95.8
S2	65.0	94.8	94.5	95.1	95.9	95.3	95.3
M1	52.9	93.8	93.6	94.3	94.7	94.4	95.1
M2	40.8	92.6	93.9	92.1	94.4	95.2	95.2
G1	56.8	94.3	94.8	95.8	95.0	94.8	94.7
G2	44.8	96.2	95.1	95.1	94.6	94.8	95.8

denotes the distance from v to the nearest point in X without v. More specifically we consider two Strauss processes with $R = 0.05$ and $\theta_1 = \log(100)$, where models S1 and S2 respectively have $\theta_2 = \log(0.8)$ and $\theta_2 = \log(0.2)$, two multiscale Strauss processes with $R_1 = 0.05$, $R_2 = 0.1$, and $\theta_1 = \log(100)$, where models M1 and M2 respectively have $(\theta_2, \theta_3) = (\log(0.2), \log(0.8))$, and $(\theta_2, \theta_3) = (\log(0.8), \log(0.2))$, and two Geyer's saturation processes with saturation parameter $s = 1$, $R = 0.05$ and $\theta_1 = \log(50)$, where models G1 and G2 respectively have $\theta_2 = \log(1.2)$ and $\theta_2 = \log(0.8)$. For all models we use relatively small values of θ_1 to illustrate that the asymptotic results can be applied even for point pattern data with small numbers of points.

For all the models the observation window is $W^+ = [-R, \ell + R]^2$, $\ell = 1, 2$, where R is the interaction range of each model. Note that due to edge effects, the simulations of X_{W^+} are in fact not realizations of stationary processes. To obtain approximate realizations of stationary processes we use the `spatstat` default settings and simulate a finite process on W^+ expanded by a border of size $2R$ and consider the restriction to W^+. For each simulation we obtain parameter estimates using the logistic regression estimating function with stratified dummy points with $n_d = 20$, $n_d = 40$ and $n_d = 80$. Subsequently we record whether or not the parameter vector falls within the approximate 95 % ellipsoidal confidence region

$$\{\theta \mid \||W|^{1/2}\widehat{\Sigma}^{-1/2}(\widehat{\theta} - \theta)\|^2 \leq \chi^2_{0.95}(p)\}.$$

The results given in Table 5 show that the coverage rates are in general close to the nominal 95 % for all the models. Model M2 is one exception where the coverage rates are consistently too low when $\ell = 1$ suggesting that there are too few points to rely on asymptotic results. This agrees with the fact that M2 has the lowest empirical intensity. The estimated Monte Carlo errors are of the order 0.5 %–1 % so the remaining deviations from the nominal 95 % are not not worrying. As can be expected, the closeness to the nominal level does not appear to depend on n_d.

5.4 Determination of ρ

As mentioned in Section 3.4, the variance σ^2 of a parameter estimate is the sum of a term σ_1^2 which is roughly constant as a function of ρ and a term σ_2^2 which is roughly proportional to ρ, $\sigma_2^2 = \bar{\sigma}_2^2/\rho$, say. For a given choice of ρ (e.g. using the rule of thumb) our asymptotic results provide estimates $\hat{\sigma}_1^2$ and $\hat{\bar{\sigma}}_2^2$ of these quantities. Suppose now that we wish to find a ρ_p so that the dummy point additional variance $\bar{\sigma}_2^2/\rho_p$ is less than a specified fraction p of σ_1^2. Then we may approximately determine ρ_p as $\rho_p = \hat{\bar{\sigma}}_2^2/(\hat{\sigma}^2 p)$. Alternatively, this relation can also be used to approximately determine $p_\rho = \hat{\bar{\sigma}}_2^2/(\hat{\sigma}^2 \rho)$ for a given value of ρ. In practice we may rewrite these relations in terms of standard deviations such that p gives the relative increase of the standard deviation (and thereby of the confidence interval length) due to random dummy points.

To exemplify this procedure we return to the mucous membrane data of Section 5.2 . Our Theorem 4.3 only covers stationary models but here we nevertheless still apply the Taylor series approximation of the variance of $\hat{\theta}$ given by $S^{-1}[G_1 + G_2]S^{-1}/|W|$, see Section 3.4. We next replace S, G_1 and G_2 by their unbiased estimates (4.14), (4.15) and (4.18) (using $n_d = 60$) and obtain the estimated standard deviations in the second and third columns from the right in Table 3. These estimates agree very well with the bootstrap estimates for $n_d = 60$ (columns 3 and 4 from the right) except for $q_1(0.6)$. The disagreement for $q_1(0.6)$ may be explained by the fact that there is a low number of type 1 points for large y values. The rightmost column shows the approximate value of p (in percent) for $n_d = 120$ which was determined based on the variance matrices estimated with $n_d = 60$. The agreement with the bootstrap results for $n_d = 120$ (column 4 from the left) is very good.

Acknowledgements

We thank Professor Antonietta Mira for drawing our attention to the connection with Barker dynamics. This research was done when J.-F. Coeurjolly was a Visiting Professor at Department of Mathematical Sciences, Aalborg University, February-July 2012. He would like to thank the members of the department for their kind hospitality. The research of J.-F. Coeurjolly was supported by Joseph Fourier University of Grenoble (project "SpaComp"). Ege Rubak 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. Jean-François Coeurjolly, Rasmus Waagepetersen and Ege Rubak were further supported by a grant from the Institut Français du Danemark.

References

F. P. Agterberg. Automatic contouring of geological maps to detect target areas for mineral exploration. *Journal of the International Association for Mathematical*

Geology, 6:373–395, 1974.

A. Baddeley. Time-invariance estimating equations. *Bernoulli*, 6(5):783–808, 2000.

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

A. Baddeley and R. Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005. URL: `www.jstatsoft.org`, ISSN: 1548-7660.

A. Baddeley and R. Turner. Modelling spatial point patterns in R. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan, editors, *Case Studies in Spatial Point Pattern Modelling*, number 185 in Lecture Notes in Statistics, pages 23–74. Springer-Verlag, New York, 2006. ISBN: 0-387-28311-0.

A. Baddeley and R. Turner. Bias correction for parameter estimates of spatial point process models. *Journal of Statistical Computation & Simulation*, 2013. Accepted for publication.

A. Baddeley, J. Møller, and A. G. Pakes. Properties of residuals for spatial point processes. *Annals of the Institute of Statistical Mathematics*, 60(3):627–649, 2008.

A. Baddeley, M. Berman, N. I. Fisher, A. Hardegen, R. K. Milne, D. Schuhmacher, R. Shah, and R. Turner. Spatial logistic regression and change-of-support for Poisson point processes. *Electronic Journal of Statistics*, 4:1151–1201, 2010. `doi: 10.1214/10-EJS581`.

A. J. Baddeley and M. N. M. Van Lieshout. Area-interaction point processes. *Ann. Inst. Statist. Math.*, 47(4):601–619, 1995.

A. J. Baddeley, R. Turner, J. Møller, and M. Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B*, 67:617–666, 2005.

A. A. Barker. Monte Carlo calculations of the radial distribution function of the proton-electron plasma. *Australian Journal of Physics*, 18:119–133, 1969.

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

E. Bertin, J.-M. Billiot, and R. Drouilhet. R-local Delaunay inhibition model. *Journal of Statistical Physics*, 132(4):649–667, 2008.

J. Besag. Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, 47:77–92, 1977.

J. Besag, R. K. Milne, and S. Zachary. Point process limits of lattice processes. *Journal of Applied Probability*, 19:210–216, 1982.

J. E. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

J.-M. Billiot. Asymptotic properties of Takacs-Fiksel estimation method for Gibbs point processes. *Statistics*, 30:68–89, 1997.

J.-M. Billiot, J.-F. Coeurjolly, and R. Drouilhet. Maximum pseudolikelihood estimator for exponential family models of marked Gibbs point processes. *Electronic Journal of Statistics*, 2:234–264, 2008.

G. Bonham-Carter. *Geographic Information Systems for geoscientists: modelling with GIS*. Number 13 in Computer Methods in the Geosciences. Pergamon Press/ Elsevier, Kidlington, Oxford, UK, 1995.

M. Clyde and D. Strauss. Logistic regression for spatial pair-potential models. In A. Possolo, editor, *Spatial Statistics and Imaging*, volume 20 of *Lecture Notes - Monograph series*, chapter II, pages 14–30. Institute of Mathematical Statistics, 1991. ISBN 0-940600-27-7.

J.-F. Coeurjolly and F. Lavancier. Residuals for stationary marked Gibbs point processes. *To appear in Journal of the Royal Statistical Society, Series B*, 2012. `http://arxiv.org/abs/1002.0857`.

J.-F. Coeurjolly and E. Rubak. Fast covariance estimation for innovations computed from a spatial Gibbs point process. Research Report R-2012-3, Centre for Stochastic Geometry and advanced Bioimaging, Aarhus University, 2012. Submitted.

J.-F. Coeurjolly, D. Dereudre, R. Drouilhet, and F. Lavancier. Takacs-Fiksel method for stationary marked Gibbs point processes. *Scandinavian Journal of Statistics*, 49(3):416–443, 2012.

C. Comas and J. Mateu. Statistical inference for Gibbs point processes based on field observations. *Stochastic Environmental Research and Risk Assessment*, 25 (2):287–300, 2011.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods.* Springer-Verlag, New York, second edition, 2003.

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

P. J. Diggle and B. Rowlingson. A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, series A (Statistics in Society)*, 157(3):433–440, 1994.

T. Fiksel. Estimation of parameterized pair potentials of marked and nonmarked Gibbsian point processes. *Elektronische Informationsverarbeitung und Kypernetik*, 20:270–278, 1984.

H. O. Georgii. Canonical and grand canonical Gibbs states for continuum systems. *Communications in Mathematical Physics*, 48(1):31–51, 1976.

C. J. Geyer. Likelihood inference for spatial point processes. In O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. van Lieshout, editors, *Stochastic Geometry: Likelihood and Computation*, number 80 in Monographs on Statistics and Applied Probability, chapter 3, pages 79–140. Chapman and Hall/CRC, Boca Raton, Florida, 1999.

X. Guyon. *Random fields on a network. Modeling, Statistics and Applications*. Springer Verlag, New York, 1995.

W.W. Hauck, Jr. and A. Donner. Wald's test as applied to hypotheses in logit analysis. *J. Amer. Statist. Assoc.*, 72(360, part 1):851–853, 1977.

F. Huang and Y. Ogata. Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics*, 8(3):510–530, 1999.

J. L. Jensen and H. R. Künsch. On asymptotic normality of pseudo likelihood estimates for pairwise interaction processes. *Annals of the Institute of Statistical Mathematics*, 46:475–486, 1994.

J. L. Jensen and J. Møller. Pseudolikelihood for exponential family models of spatial point processes. *Ann. Appl. Probab.*, 1:445–461, 1991.

J. L. Jensen and J. Møller. Pseudolikelihood for exponential family models of spatial point processes. *Annals of Applied Probability*, 3:445–461, 1991.

R. E. Miles. On the elimination of edge effects in planar sampling. In E. F. Harding and D. G. Kendall, editors, *Stochastic Geometry: A Tribute to the Memory of Rollo Davidson*, pages 228–47. John Wiley and Sons, Ltd, Chichester, 1974.

J. Møller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes.* Chapman and Hall/CRC, Boca Raton, 2004.

X. X. Nguyen and H. Zessin. Ergodic theorems for spatial processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 48:133–158, 1979.

Y. Ogata and M. Tanemura. Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Annals of the Institute of Statistical Mathematics*, 33:315–338, 1981.

C. J. Preston. *Random fields.* Springer Verlag, 1976.

C. J. Preston. Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, 46:371–391, 1977.

X. Qi. A functional central limit theorem for spatial birth and death processes. *Advances in Applied Probability*, 40:759–797, 2008.

S. L. Rathbun. Optimal estimation of Poisson intensity with partially observed covariates. *Biometrika*, 2012. To appear.

S. L. Rathbun, S. Shiffman, and C. J. Gwaltney. Modelling the effects of partially observed covariates on Poisson process intensity. *Biometrika*, 94:153–165, 2007.

B. D. Ripley. Simulating spatial patterns: dependent samples from a multivariate density. *Applied Statistics*, 28:109–112, 1979.

R. Takacs. Estimator for the pair-potential of a Gibbsian point process. Institutsbericht 238, Institut für Mathematik, Johannes Kepler Universität Linz, Austria, 1983.

J. W. Tukey. Discussion of paper by F. P. Agterberg and S. C. Robinson. *Bulletin of the International Statistical Institute*, 44(1):596, 1972. Proceedings, 38th Congress, International Statistical Institute.

R. Waagepetersen. Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, 95:351–363, 2007.

D. I. Warton and L. C. Shepherd. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics*, 4:1383–1402, 2010.

Appendices

A Estimating equation from Barker dynamics

This appendix is based partly on discussions with Prof. Antonietta Mira and Dr. Pavel Grabarnik.

As in Section 3 we consider the case of a bounded $\Lambda = W$ and construct a spatial birth-and-death process (Preston, 1977) $(Y_t)_{t>0}$ whose equilibrium distribution coincides with the distribution of X. Such a process is a continuous-time Markov process, whose states are point patterns x, and whose only transitions are instantaneous deaths $x \mapsto x \setminus \{x_i\}$ and instantaneous births $x \mapsto x \cup \{u\}$.

We start with a proposal mechanism. The proposals are instantaneous deaths with rate $q(x \mapsto x \setminus \{x_i\}) = 1$ for each existing point $x_i \in x$, and instantaneous births with rate $q(x \mapsto x \cup \{u\}) = B(u)$ for $u \in W$, with respect to Lebesgue measure on W, where $B(u) > 0$ and $\int_W B(u) \mathrm{d}u < \infty$.

Barker dynamics (Barker, 1969) are defined to have acceptance probabilities

$$A(x \mapsto x \cup \{u\}) = \frac{\lambda_\theta(u,x)/B(u)}{1 + \lambda_\theta(u,x)/B(u)} \quad \text{and} \quad A(x \mapsto x \setminus \{x_i\}) = \frac{1}{1 + \lambda_\theta(x_i,x)/B(u)}$$

so that the transition rates are

$$r(x \mapsto x \cup \{u\}) = B(u) \frac{\lambda_\theta(u,x)/B(u)}{1 + \lambda_\theta(u,x)/B(u)} \quad \text{and} \quad r(x \mapsto x \setminus \{x_i\}) = \frac{1}{1 + \lambda_\theta(x_i,x)/B(u)}.$$

It can easily be checked that the birth-death process $(Y_t)_{t>0}$ with these transition rates is in detailed balance with equilibrium distribution given by the distribution of X.

The so-called infinitesimal generator of $(Y_t)_{t>0}$ is given by

$$(\mathcal{A}_\theta f)(x) = \lim_{t \to 0} \frac{1}{t} \left[\mathrm{E}[f(Y_t) | Y_0 = x] - f(x) \right]$$

for real functions f on Ω. The infinitesimal generator yields a class of time-invariance estimating functions (Baddeley, 2000) which are unbiased

$$\mathrm{E}[(\mathcal{A}_\theta f)(X)] = 0$$

for any choice of f.

Assume that $t(u,x) = w(x \cup \{u\}) - w(x)$ for some statistic $w(x)$. Then in case of the spatial birth-death process with Barker dynamics we obtain the estimating function

$$(\mathcal{A}_\theta w)(x) = -\sum_i \frac{t(x_i, x \setminus x_i)}{1 + \lambda_\theta(x_i, x \setminus x_i)/B(u)} + \int_W \frac{t(u,x)\lambda_\theta(u,x)}{1 + \lambda_\theta(u,x)/B(u)} \, \mathrm{d}u \qquad \text{(A.1)}$$

which negated corresponds to (3.9) in case of a log-linear conditional intensity.

B Proof of Proposition 4.2

For $A, B \subseteq \mathbb{R}^d \times \mathbb{M}$ we have

$$
\begin{aligned}
\mathrm{E} \sum_{u,v \in D}^{\neq} \mathbf{1}(u \in A, v \in B) &= \sum_{k,k' \in \mathbb{Z}^d, k \neq k'} \mathrm{E}\, \mathbf{1}(U_k \in A, U_{k'} \in B) \\
&= \sum_{k,k' \in \mathbb{Z}^d, k \neq k'} \int_{C_k \times \mathbb{M}} \int_{C_{k'} \times \mathbb{M}} \mathbf{1}(u \in A, v \in B) \rho^2 \mathrm{d}u\, \mathrm{d}v \\
&= \sum_{k \in \mathbb{Z}^d} \int_A \int_B \mathbf{1}(\overline{u} \in C_k, \overline{v} \in \mathbb{R}^d \setminus C_k) \rho^2 \mathrm{d}u\, \mathrm{d}v \\
&= \rho^2 \int_A \int_B \sum_{k \in \mathbb{Z}^d} \mathbf{1}(\overline{u} \in C_k, \overline{v} \in \mathbb{R}^d \setminus C_k) \mathrm{d}u\, \mathrm{d}v
\end{aligned}
$$

which leads to the stated result.

C Conditions for asymptotic results

In addition to stationarity and finite range, we need the further conditions below in order to verify Theorem 4.3. These conditions have already been considered in Billiot et al. (2008) and are not very restrictive.

Let P_θ denote the distribution of a (well-defined) stationary hereditary marked Gibbs point process with Papangelou conditional intensity λ_θ, and let $X \sim P_{\theta^\star}$ where $\theta^\star$ denotes the true parameter. We will assume

(i) $\Theta \subset \mathbb{R}^p$ is an open convex set.

(ii) For any $\theta \neq \theta^\star$, the following (identifiability) condition holds

$$P_{\theta^\star}\left((\theta - \theta^\star)^\top t(0^M, X)\, \mathbf{1}(H(0^M, X) > 0) \neq 0 \right) > 0. \tag{C.1}$$

(iii) For all $u \in \mathbb{S}$, $x \in \Omega$ and $i = 1, \dots, p$ there exists a constant $\kappa \geq 0$ such that $H(u, x) \leq \kappa$ and at least one of the following two assumptions is satisfied:

$$\theta_i \leq 0 \quad \text{and} \quad -\kappa \leq t_i(u, x) \leq \kappa\, n(x_{\mathcal{B}(\overline{u}, R)}) \tag{C.2}$$

or

$$-\kappa \leq t_i(u, x) \leq \kappa \tag{C.3}$$

where R is the range of interaction defined in (4.2).

D Proof of Theorem 4.3

We present the proof when D corresponds to a marked stratified point process with intensity ρ, which seems to us the most interesting and challenging case. We claim that Theorem 4.3 remains true when $D \sim \mathcal{P}(\mathbb{R}^d, \rho)$ or $D \sim \mathcal{B}(\mathbb{R}^d, \rho)$. Before detailing the proofs, we let

$$\mathcal{K}_n^{\inf} = \left\{k \in \mathbb{Z}^d : C_k \subseteq W_n\right\} \text{ and } \mathcal{K}_n^{\sup} = \left\{k \in \mathbb{Z}^d : C_k \cap W_n \neq \emptyset\right\}$$

where the cells C_k are described in Definition 4.1. We denote for $\bullet = \inf, \sup$ by $W_n^\bullet$ the bounded domain $\cup_{k \in \mathcal{K}_n^\bullet} C_k$. We point out that $W_n^{\inf} \subseteq W_n \subseteq W_n^{\sup}$, that as $n \to \infty$ $|W_n^\bullet| \sim |W_n|$ and that $n(D_{W_n^\bullet}) = |\mathcal{K}_n^\bullet|$. In the following proofs c stands for a generic positive constant (which may depend on ρ) that may be different from line to line and U_k is a marked point distributed as described in Definition 4.1. Finally, we recall that the notation 0^M stands for a marked point at location 0 with mark distribution μ and that $\mathrm{d}u$ is short for $\mathcal{L}^d(\mathrm{d}\overline{u}) \otimes \mu(\mathrm{d}m)$.

D.1 Consistency of the logistic regression estimate

Lemma D.1. *For any $\theta, \theta' \in \Theta$, any $j = 1, \ldots, p$ and any $q \geq 1$*

$$\mathrm{E}[\,|t_j(0^M, X)|^q \lambda_\theta(0^M, X)] < \infty \quad \textit{and} \quad \mathrm{E}[\,|w_{\theta',j}(0^M, X)|^q \lambda_\theta(0^M, X)] < \infty.$$

Proof. Using (C.2) and (C.3), we can show that the Gibbs point process is locally stable, i.e. there exists a constant c such that for any $\theta \in \Theta$, $u \in \mathbb{S}$ and any $x \in \Omega$, $\lambda_\theta(u, x) \leq c$. Furthermore, the largest expectation to control is of the form $\mathrm{E}\, n(X_{\mathcal{B}(0,R)})^q \mathrm{e}^{cn(X_{\mathcal{B}(0,R)})}$. Now, we can invoke Proposition 9 in Bertin et al. (2008) which proves that such expectations are finite under the local stability property. $\square$

Since any stationary Gibbs measure can be represented as a mixture of ergodic measures (see Preston, 1976), it is sufficient to prove the consistency for ergodic measures. We therefore assume that $P_{\theta^\star}$ is ergodic. From (3.5), $\theta \mapsto -\mathsf{LRL}_{W_n}(x, \theta)$ is a convex function for every $x \in \Omega$. Therefore from Guyon (1995, Theorem 3.4.4), we only have to prove that almost surely

$$\lim_{n \to \infty} K_n(\theta, \theta^\star) := \lim_{n \to \infty} |W_n|^{-1} \big(\mathsf{LRL}_{W_n}(X; \theta^\star) - \mathsf{LRL}_{W_n}(X; \theta)\big) = K(\theta, \theta^\star)$$

where $\theta \mapsto K(\theta, \theta^\star)$ is a non-negative function which vanishes only at $\theta = \theta^\star$.

Let $W_0 = [0,1]^d$ and let $\theta \in \Theta$. From (4.1) and Lemma D.1, the general ergodic theorem for spatial point processes obtained by Nguyen and Zessin (1979) and the GNZ formula can be combined to prove the following almost sure convergence as $n \to \infty$

$$\begin{aligned} \frac{1}{|W_n|} \sum_{u \in X_{W_n}} \log \frac{\lambda_\theta(u, X \setminus u)}{\lambda_\theta(u, X \setminus u) + \rho} &\to |W_0|^{-1}\, \mathrm{E} \int_{W_0 \times \mathbb{M}} \lambda_{\theta^\star}(u, X) \log \frac{\lambda_\theta(u, X)}{\lambda_\theta(u, X) + \rho} \mathrm{d}u \\ &= \mathrm{E}\left[\lambda_{\theta^\star}(0^M, X) \log \frac{\lambda_\theta(0^M, X)}{\lambda_\theta(0^M, X) + \rho}\right] \end{aligned} \tag{D.1}$$

where to take into account the possible hard-core component (which could appear when $H(u, x) = 0$) we take the convention $0 \log 0 = 0$.

For any bounded domain W we let $V_W(X, D)$ denote the right term of (3.4), i.e. $V_W(X, D) = -\sum_{u \in D_W} \log \frac{\lambda_\theta(u, X) + \rho}{\rho}$. Since D is an ergodic process, (X, D) is ergodic. Now since $V_W(X,D)$ is additive and translation invariant (i.e. for any $y \in \mathbb{R}^d$ $V_{\tau_y W}(\tau_y X, \tau_y D) = V_W(X,D)$), then using Lemma D.1 we can still apply the ergodic theorem from Nguyen and Zessin (1979) to the process $|W_n|^{-1} V_{W_n}(X,D)$. Combined with Campbell's theorem and the stationarity of X, we obtain that almost surely as $n \to \infty$

$$\begin{aligned}\frac{1}{|W_n|} V_{W_n}(X,D) &\to -|W_0|^{-1} \operatorname{E} \sum_{u \in D_{W_0}} \log \frac{\lambda_\theta(u,X)+\rho}{\rho} \\ &= -|W_0|^{-1} \operatorname{E} \int_{W_0 \times \mathbb{M}} \rho \log \frac{\lambda_\theta(u,X)+\rho}{\rho} \mathrm{d}u \\ &= -\operatorname{E}\left[\rho \log \frac{\lambda_\theta(0^M,X)+\rho}{\rho}\right].\end{aligned}$$

Combining this with (D.1), we derive that $\theta \mapsto K(\theta, \theta^\star)$ is a well-defined function given by

$$\begin{aligned}K(\theta,\theta^\star) &= \operatorname{E}\left[\lambda_{\theta^\star}(0^M,X)\log\frac{\lambda_{\theta^\star}(0^M,X)}{\lambda_\theta(0^M,X)} - \big(\lambda_{\theta^\star}(0^M,X)+\rho\big)\log\frac{\lambda_{\theta^\star}(0^M,X)+\rho}{\lambda_\theta(0^M,X)+\rho}\right] \\ &= \operatorname{E}\left[\lambda_\theta(0^M,X)\left(A\log A - (A+B)\log\frac{A+B}{1+B}\right)\right]\end{aligned}$$

where $A := \mathbf{1}(H(0^M,X)>0)e^{(\theta^\star-\theta)^\top t(0^M,X)}$ and $B = \rho e^{-\theta^\top t(0^M,X)}\mathbf{1}(H(0^M,X)>0)$. For every $b>0$, the function $a \mapsto a\log a - (a+b)\log(a+b/1+b)$ is always non-negative for $a>0$ and equals 0 if and only if $a=1$. Therefore, if $P(A \neq 1) > 0$ which is implied by (C.1) then $K(\cdot,\theta^\star)$ vanishes only at $\theta = \theta^\star$.

D.2 Asymptotic normality of the logistic regression estimate

The asymptotic normality of $\widehat{\theta}$ will be achieved by applying Guyon (1995, Theorem 3.4.5) which is a general result on asymptotic normality for minimum contrast estimators. By arguments similar to those in Appendix D.1 and using Lemma D.1, we can show that the renormalized negative score $-|W_n|^{-1}s_{W_n}(X,D;\theta)$ ((3.5) with $\rho(\cdot)=\rho$) converges almost surely towards a matrix depending on θ (and ρ) and that (4.10) holds when $\theta = \theta^\star$. Therefore, according to Guyon (1995, Theorem 3.4.5), Theorem 4.3 will be proved if we can prove that as $n \to \infty$

(i) $|W_n|^{-1/2} s_{W_n}(X,D;\theta) \xrightarrow{d} \mathcal{N}(0, G_1 + G_2^{\mathrm{sb}})$.

(ii) $\widehat{G}_1, \widehat{G}_2^{\mathrm{sb}}$ and $\widehat{S}$ are consistent estimates of G_1, G_2^{sb} and S respectively.

The point (ii) will be examined in Appendix D.3. To prove (i) we assume that $P_{\theta^\star}$ is an ergodic measure, we consider the decomposition of the score as the sum of the vectors $T_{1,W_n}(X)$ and $T_{2,W_n}(X,D)$ defined by (4.5) and (4.6) and proceed in three streps.

Step 1. Central limit theorem for $T_{1,W_n}(X)$: proof of (4.7).

We note that the jth component of $T_{1,W_n}(X)$ corresponds to the $(w_{\theta^\star,j})$-innovations of a spatial point process, a notion defined by Baddeley et al. (2005) which we now recall: the h-innovations of a spatial point process (for a function $h : \mathbb{S} \times \Omega \to \mathbb{R}$) computed in a bounded domain W is the centered random variable defined by

$$I_W(X; h) := \sum_{u \in X_W} h(u, X \setminus u) - \int_{W \times \mathbb{M}} h(u, X) \lambda_{\theta^\star}(u, X) \mathrm{d}u. \tag{D.2}$$

Asymptotic properties for innovations have been considered in Coeurjolly et al. (2012) and Coeurjolly and Lavancier (2012). In particular, Coeurjolly et al. (2012, Lemma 3) gives conditions ensuring the asymptotic normality. For every $j = 1, \ldots, p$ and any bounded domain W it is required that (a) $I_W(X; w_{\theta^\star,j})$ depends only on $X_{W \oplus R}$ and (b) $\mathrm{E}\,|I_W(X; w_{\theta^\star,j})|^3 < \infty$. The condition (a) is implied by (4.2) and (b) follows by (C.2), (C.3) and Lemma D.1. Finally, the form of the matrix G_1 is obtained by applying Coeurjolly and Rubak (2012, Proposition 3.1).

Step 2. Central limit theorem for $T_{2,W_n}(X, D)$: proof of (4.8).

Using the Cramér-Wold device, this step is achieved if we prove that given X and for any $y \in \mathbb{R}^p \setminus \{0\}$, we have $|W_n|^{-1/2}\, y^\top T_{2,W_n}(X, D) \xrightarrow{d} \mathcal{N}(0, y^\top G_2^{\mathrm{sb}} y)$ as $n \to \infty$. For $\theta \in \Theta$, $u \in \mathbb{S}$ and $x \in \Omega$, we recall the notation $w_\theta^\lambda(u, x) = w_\theta(u, x)\lambda_\theta(u, x)$ and we define $Z(U_k, x; y) = \int_{C_k \times \mathbb{M}} y^\top w_{\theta^\star}^\lambda(u, x) \mathrm{d}u - \frac{1}{\rho} y^\top w_{\theta^\star}^\lambda(U_k, x)$. We start with the following decomposition

$$y^\top T_{2,W_n}(X, D) = y^\top T_{2,W_n^{\mathrm{inf}}}(X, D) + y^\top T_{2,W_n \setminus W_n^{\mathrm{inf}}}(X, D) \tag{D.3}$$

where we recall that W_n^{inf} is such that $n(D_{W_n^{\mathrm{inf}}}) = |\mathcal{K}_n^{\mathrm{inf}}|$. Then, $y^\top T_{2,W_n^{\mathrm{inf}}}(X, D) = \sum_{k \in \mathcal{K}_n^{\mathrm{inf}}} Z(U_k, X; y)$. We let $s_n^2 = \sum_{k \in \mathcal{K}_n^{\mathrm{inf}}} \mathrm{Var}[Z(U_k, X; y)|X]$ and state the following lemma.

Lemma D.2. *For any bounded domain $W \subset \mathbb{R}^d$*

$$\begin{aligned}\mathrm{Var}[T_{2,W}(X, D)|X] =& \frac{1}{\rho^2}\Big\{ \int_{W \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) w_{\theta^\star}^\lambda(u, X)^\top \rho du \\ & - \sum_{k \in \mathbb{Z}^d} \Big(\int_{(W \cap C_k) \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) \rho du \Big) \Big(\int_{(W \cap C_k) \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) \rho du \Big)^\top \Big\}.\end{aligned}$$

Proof. Using standard results on the variance of functionals for point processes and Proposition 4.2, we get

$$\begin{aligned}\mathrm{Var}[T_{2,W}(X, D)|X] =& \frac{1}{\rho^2}\Big\{ \int_{W \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) w_{\theta^\star}^\lambda(u, X)^\top \rho \mathrm{d}u \\ & + \int_{W \times \mathbb{M}} \int_{W \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) w_{\theta^\star}^\lambda(v, X)^\top (\rho^{(2)}(u, v) - \rho^2) \mathrm{d}u \, \mathrm{d}v \Big\}.\end{aligned}$$

Rearranging the second integral leads to the result. ☐

Using Lemma D.2 with $W = W_n^{\mathrm{inf}}$ we have that $s_n^2 = y^\top G_n(X) y$ where

$$\begin{aligned}G_n(X) =& \frac{1}{\rho^2}\Big\{ \int_{W_n^{\mathrm{inf}} \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) w_{\theta^\star}^\lambda(u, X)^\top \rho \mathrm{d}u \\ & - \sum_{k \in \mathcal{K}_n^{\mathrm{inf}}} \Big(\int_{C_k \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) \rho \mathrm{d}u \Big) \Big(\int_{C_k \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) \rho \mathrm{d}u \Big)^\top \Big\}.\end{aligned}$$

As $n \to \infty$, using the ergodic theorem from Nguyen and Zessin (1979) we have that

$$
\begin{aligned}
&|W_n^{\inf}|^{-1}\frac{1}{\rho^2}\int_{W_n^{\inf}\times\mathbb{M}} w_{\theta^\star}^\lambda(u,X)w_{\theta^\star}^\lambda(u,X)^\top \rho \mathrm{d}u \\
&\xrightarrow{a.s.} \frac{1}{\rho}\operatorname{E}\int_{C_0\times\mathbb{M}} w_{\theta^\star}^\lambda(u,X)w_{\theta^\star}^\lambda(u,X)^\top \rho \mathrm{d}u \\
&= \frac{1}{\rho}\operatorname{E}\operatorname{E}\left[w_{\theta^\star}^\lambda(U_0,X)w_{\theta^\star}^\lambda(U_0,X)^\top \mid X\right]
\end{aligned}
\tag{D.4}
$$

and using the mean ergodic theorem Guyon (1995, Theorem 3.1.1) we have

$$
\begin{aligned}
&|\mathcal{K}_n^{\inf}|^{-1}\sum_{k\in\mathcal{K}_n^{\inf}}\Big(\int_{C_k\times\mathbb{M}} w_{\theta^\star}^\lambda(u,X)\rho\mathrm{d}u\Big)\Big(\int_{C_k\times\mathbb{M}} w_{\theta^\star}^\lambda(u,X)\rho\mathrm{d}u\Big)^\top \\
&\xrightarrow{a.s.} \operatorname{E}\Big[\Big(\int_{C_0\times\mathbb{M}} w_{\theta^\star}^\lambda(u,X)\rho\mathrm{d}u\Big)\Big(\int_{C_0\times\mathbb{M}} w_{\theta^\star}^\lambda(u,X)\rho\mathrm{d}u\Big)^\top\Big] \\
&= \operatorname{E}\Big[\operatorname{E}\left[w_{\theta^\star}^\lambda(U_0,X)|X\right]\operatorname{E}\left[w_{\theta^\star}^\lambda(U_0,X)|X\right]^\top\Big]
\end{aligned}
\tag{D.5}
$$

Since $|\mathcal{K}_n^{\inf}| = \rho|W_n^{\inf}|$ and since X is stationary, we assert that combining (D.4) and (D.5) leads to

$$
|W_n^{\inf}|^{-1}G_n(X) \xrightarrow{a.s.} \frac{1}{\rho}\operatorname{E}\operatorname{Var}[w_{\theta^\star}^\lambda(U_0,X)|X] = G_2^{\mathrm{sb}}. \tag{D.6}
$$

Since for any $y\in\mathbb{R}^p$, $y^\top G_2^{\mathrm{sb}}y \ge 0$ we get in particular that almost surely for n large enough

$$
|W_n^{\inf}|^{-1}s_n^2 - y^\top G_2^{\mathrm{sb}}y \ge -\frac{1}{2}y^\top G_2^{\mathrm{sb}}y
$$

which leads to

$$
|W_n^{\inf}|^{-1}s_n^2 \ge \frac{1}{2}y^\top G_2^{\mathrm{sb}}y \ge \|y\|^2\frac{1}{2}\inf_{y\in\mathbb{R}^p\setminus\{0\}} y^\top G_2^{\mathrm{sb}}y/\|y\|^2 = \|y\|^2\frac{1}{2}\nu_{\min} > 0 \tag{D.7}
$$

where $\nu_{\min}$ is the smallest eigenvalue of G_2^{sb} (assumed to be positive-definite). Furthermore, for any $k\in\mathcal{K}_n^{\inf}$ we have from (C.2) and (C.3)

$$
Z(U_k,X;y)^4 \le c\|y\|_\infty^4\Big(1+n(X_{\mathcal{B}(U_k,R)}) + \int_{C_k\times\mathbb{M}} n(X_{\mathcal{B}(\overline{u},R)})\mathrm{d}u\Big)^4.
$$

By the Cauchy-Schwarz inequality $(|z_1|+|z_2|+|z_3|)^4 \le 27(z_1^4+z_2^4+z_3^4)$ for any $z=(z_1,z_2,z_3)^\top\in\mathbb{R}^3$. Hence,

$$
Z(U_k,X;y)^4 \le 27c\|y\|_\infty^4\Big(1+n(X_{\mathcal{B}(U_k,R)})^4 + \Big(\int_{C_k\times\mathbb{M}} n(X_{\mathcal{B}(\overline{u},R)})\mathrm{d}u\Big)^4\Big)
$$

which allows us to write

$$
\operatorname{E}[Z(U_k,X;y)^4|X] \le c\Big(1+\int_{C_k\times\mathbb{M}} n(X_{\mathcal{B}(\overline{u},R)})^4\rho\mathrm{d}u + \Big(\int_{C_k\times\mathbb{M}} n(X_{\mathcal{B}(\overline{u},R)})\mathrm{d}u\Big)^4\Big).
$$

Now, we can invoke the mean ergodic theorem from Guyon (1995, Theorem 3.1.1) to assert that there exists $\kappa = \kappa(\rho, y)$ such that

$$|\mathcal{K}_n^{\inf}|^{-1} \sum_{k \in \mathcal{K}_n^{\inf}} \mathrm{E}[Z(U_k, X; y)^4 | X] \leq \kappa. \tag{D.8}$$

From (D.7) and (D.8), we derive that almost surely for n large enough

$$\frac{1}{s_n^4} \sum_{k \in \mathcal{K}_n^{\inf}} \mathrm{E}[Z(U_k, X; y)^4 | X] \leq \frac{4\kappa}{\nu_{\min}^2 \|y\|^4} \frac{|\mathcal{K}_n^{\inf}|}{|W_n^{\inf}|^2} = \mathcal{O}(|W_n^{\inf}|^{-1}).$$

Therefore, we can apply the Lindeberg-Feller theorem with Lyapunov's condition to get that given X,

$$\frac{y^\top T_{2,W_n^{\inf}}(X,D)}{(y^\top G_n(X) y)^{1/2}} \xrightarrow{d} \mathcal{N}(0, I_p)$$

which from (D.6) leads to $|W_n^{\inf}|^{-1/2} y^\top T_{2,W_n^{\inf}}(X,D) \xrightarrow{d} \mathcal{N}(0, y^\top G_2^{\mathrm{sb}} y)$. Now, since given X $T_{2,W}$ is a centered vector for any W, we leave the reader to show that using essentially Lemma D.2 applied with $W = W_n \setminus W_n^{\inf}$ there exists some constant c such that

$$\mathrm{Var}\left[y^\top T_{2,W_n \setminus W_n^{\inf}}(X,D)\right] = \mathrm{E}\,\mathrm{Var}[y^\top T_{2,W_n \setminus W_n^{\inf}}(X,D) \big| X] \leq c |W_n \setminus W_n^{\inf}|.$$

Since $|W_n \setminus W_n^{\inf}| \leq |W_n| - (|W_n|^{1/d} - 1/\rho^{1/d})^d = o(|W_n|)$, we get by Chebyshev's inequality that $|W_n|^{-1/2} y^\top T_{2,W_n \setminus W_n^{\inf}}(X,D)$ tends to zero in probability as $n \to \infty$. So from (D.3) and Slutsky's lemma

$$|W_n|^{-1/2} y^\top T_{2,W_n}(X,D) \xrightarrow{d} \mathcal{N}(0, y^\top G_2^{\mathrm{sb}} y)$$

which proves (4.8).

Step 3. Central limit theorem for the score function.

We denote by $\tilde{T}_{1,n}$ (resp. $\tilde{T}_{2,n}$) the normalized vector $|W_n|^{-1/2} T_{1,W_n}(X)$ (resp. $|W_n|^{-1/2} T_{2,W_n}(X,D)$). Let $z \in \mathbb{R}^p$ and let $\Phi_i(z)$ for $i = 1, 2$ be the characteristic function of a Gaussian vector with covariance matrix G_i. We now prove that the characteristic function of $\tilde{T}_{1,n} + \tilde{T}_{2,n}$, say $\Phi^{(n)}(z)$, tends to the one of a Gaussian vector with covariance matrix $G_1 + G_2^{\mathrm{sb}}$, that is towards $\Phi(z) := \Phi_1(z)\Phi_2(z)$ as $n \to \infty$. We have

$$\begin{aligned} \Phi^{(n)}(z) - \Phi(z) &= \mathrm{E}\Big[e^{iz^\top \tilde{T}_{1,n}} \big(e^{iz^\top \tilde{T}_{2,n}} - \Phi_2(z)\big)\Big] + \Phi_2(z)\big(\mathrm{E}[e^{iz^\top \tilde{T}_{1,n}}] - \Phi_1(z)\big) \\ &= \mathrm{E}\Big[e^{iz^\top \tilde{T}_{1,n}}\, \mathrm{E}\big[e^{iz^\top \tilde{T}_{2,n}} - \Phi_2(z) | X\big]\Big] + \Phi_2(z)\big(\mathrm{E}[e^{iz^\top \tilde{T}_{1,n}}] - \Phi_1(z)\big). \end{aligned}$$

Using in particular the fact that $|e^{iz^\top \tilde{T}_{1,n}}| < 1$ and $|\Phi_2(z)| \leq 1$ we derive

$$\big|\Phi^{(n)}(z) - \Phi(z)\big| \leq \mathrm{E}\left|\mathrm{E}\big[e^{iz^\top \tilde{T}_{2,n}} - \Phi_2(z) | X\big]\right| + \big|\mathrm{E}[e^{iz^\top \tilde{T}_{1,n}}] - \Phi_1(z)\big|.$$

The proof is finished since from (4.7) and (4.8), both terms of the right-hand side of the last inequality converge to 0 as $n \to \infty$ for any $z \in \mathbb{R}^p$.

D.3 Consistency of the asymptotic covariance matrix estimates

The consistency of $\widehat{G}_1$ follows from the consistency of $\widehat{A}_1(X, D, w_{\widehat{\theta}}, w_{\widehat{\theta}})$ (resp. $\widehat{A}_\ell(X, w_{\widehat{\theta}}, w_{\widehat{\theta}})$ for $\ell = 2, 3$). To show that $\widehat{A}_1(X, D, w_{\widehat{\theta}}, w_{\widehat{\theta}})$ (resp. $\widehat{A}_\ell(X, w_{\widehat{\theta}}, w_{\widehat{\theta}})$ for $\ell = 2, 3$) converges towards $A_1(w_{\theta^\star}, w_{\theta^\star})$ (resp. $A_\ell(w_{\theta^\star}, w_{\theta^\star})$ for $\ell = 2, 3$), we need to verify that for some neighbourhood $\mathcal{V}(\theta^\star)$ of $\theta^\star$, the variables $I_\ell(w_{\theta^\star,j}, w_{\theta^\star,k})$ for $\ell = 1, 2, 3$ and $j, k = 1, \dots, p$ defined by (3.3)-(3.5) in Coeurjolly and Rubak (2012) have finite expectations. This follows from (C.2) and (C.3) and Lemma D.1. The consistency of $\widehat{S}$ follows analogously.

We now focus on $\widehat{G}_2^{\mathrm{sb}}$. We define for any $\theta \in \Theta$ and any bounded domain W

$$G_{2,W}^{\mathrm{sb}}(X, D; \theta) = \frac{1}{2\rho^2 |W|} \sum_{u \in D_W} \sum_{v \in D'} \mathbf{1}(\overline{u} \text{ and } \overline{v} \text{ in the same cell}) g(u, v, X; \theta)$$

where for $u, v \in \mathbb{S}$, for any $x \in \Omega$ and any $\theta \in \Theta$, we define g as follows

$$g(u, v, x; \theta) = (w_\theta^\lambda(u, x) - w_\theta^\lambda(v, x))(w_\theta^\lambda(u, x) - w_\theta^\lambda(v, x))^\top.$$

We note that $\widehat{G}_2^{\mathrm{sb}}$ defined by (4.18) is nothing else than $G_{2,W_n}^{\mathrm{sb}}(X, D; \widehat{\theta})$.

By ergodicity, Lemma D.1 (which ensures finite expectation) and Campbell's Theorem,

$$\begin{aligned} G_{2,W_n}^{\mathrm{sb}}(X, D; \theta) &\to \frac{1}{2\rho^2} \frac{1}{|C_0|} \mathrm{E}\Big[\sum_{u \in D_{C_0}} \sum_{v \in D'_{C_0}} g(u, v, X; \theta) \Big] \\ &= \frac{\rho}{2} \mathrm{E} \int_{(C_0 \times \mathbb{M})^2} g(u, v, X; \theta) \mathrm{d}u \, \mathrm{d}v \end{aligned}$$

almost surely as $n \to \infty$.

Since $\widehat{\theta}$ is a strongly consistent estimate of $\theta^\star$ and since $\theta \mapsto w^\lambda(u, x; \theta)$ is a continuous function for every $u \in \mathbb{S}$ and $x \in \Omega$, we derive

$$G_2^{\mathrm{sb}}(\widehat{\theta}) \to \frac{\rho}{2} \mathrm{E} \int_{(C_0 \times \mathbb{M})^2} g(u, v, X; \theta^\star) \mathrm{d}u \, \mathrm{d}v$$

almost surely as $n \to \infty$. Now we can check that

$$\begin{aligned} \frac{\rho}{2} \mathrm{E} \int_{(C_0 \times \mathbb{M})^2} g(u, v, X; \theta^\star) \mathrm{d}u \, \mathrm{d}v &= \frac{1}{\rho} \mathrm{E} \int_{C_0 \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) w_{\theta^\star}^\lambda(u, X)^\top \rho \mathrm{d}u \\ &- \frac{1}{\rho} \mathrm{E}\Big[\Big(\int_{C_0 \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) \rho \mathrm{d}u \Big) \Big(\int_{C_0 \times \mathbb{M}} w_{\theta^\star}^\lambda(u, X) \rho \mathrm{d}u \Big)^\top \Big]. \end{aligned}$$

In other words,

$$\begin{aligned} \widehat{G}_2^{\mathrm{sb}} &\to \frac{1}{\rho} \mathrm{E} \left(\mathrm{E} \left[w_{\theta^\star}^\lambda(U_0, X) w_{\theta^\star}^\lambda(U_0, X)^\top | X \right] \right) \\ &\quad - \frac{1}{\rho} \mathrm{E} \Big(\mathrm{E} \left[w_{\theta^\star}^\lambda(U_0, X) | X \right] \mathrm{E} \left[w_{\theta^\star}^\lambda(U_0, X) | X \right]^\top \Big) \\ &= \frac{1}{\rho} \mathrm{E} \, \mathrm{Var}[w_{\theta^\star}^\lambda(U_0, X) | X]. \end{aligned}$$

As a conclusion, we have proved that under the assumption that $P_{\theta^\star}$ is an ergodic measure, there exists an empirical matrix $\widehat{G}$ such that $\widehat{G}^{-1/2}(\widehat{\theta}-\theta^\star) \xrightarrow{d} \mathcal{N}(0, I_p)$. We now use an argument developed in Jensen and Künsch (1994). If $P_{\theta^\star}$ is not an ergodic measure, it can be decomposed as a mixture of ergodic measures, see Preston (1976). Since any mixture of standard Gaussian distributions is a standard Gaussian distribution the result is true for non ergodic Gibbs measures.

D.4 Proof of Corollary 4.8

We begin by the following result establishing a Bahadur type representation for the logistic regression estimate.

Lemma D.3. *If $P_{\theta^\star}$ is ergodic*

$$S(\widehat{\theta}-\theta^\star) \;=\; |W_n|^{-1}s_{W_n}(X,D;\theta^\star) + o_P(|W_n|^{-1/2}).$$

Proof. Under the setting of Section 4 and since $\widehat{\theta}$ minimizes the score function, we can write the following Taylor expansion

$$\frac{s_{W_n}(X,D;\theta^\star)}{|W_n|} = \frac{s_{W_n}(X,D;\hat{\theta})}{|W_n|} + \underbrace{\Big(\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\theta^\top}\frac{s_{W_n}(X,D;\theta^\star+t(\widehat{\theta}-\theta^\star))}{|W_n|}\mathrm{d}t\Big)}_{:=F_{W_n}}(\widehat{\theta}-\theta^\star)$$

Therefore, since $s_{W_n}(X,D;\widehat{\theta})=0$

$$|W_n|^{1/2}(S(\widehat{\theta}-\theta^\star) - |W_n|^{-1}s_{W_n}(X,D;\theta^\star)) = (S-F_{W_n})|W_n|^{1/2}(\widehat{\theta}-\theta^\star).$$

By the strong consistency of $\widehat{\theta}$, we have almost surely as $n\to\infty$

$$F_{W_n} - |W_n|^{-1}\frac{\mathrm{d}}{\mathrm{d}\theta^\top}s_{W_n}(X,D;\theta^\star) \to 0 \text{ and } |W_n|^{-1}\frac{\mathrm{d}}{\mathrm{d}\theta^\top}s_{W_n}(X,D;\theta^\star) \to S.$$

Since $|W_n|^{1/2}(\widehat{\theta}-\theta^\star)$ tends to a Gaussian vector, the result follows from Slutsky's Theorem. □

Assume $P_{\theta^\star}$ ergodic. Let $D, D' \sim \mathcal{SB}(\mathbb{R}^d,\rho)$ be independent. Then using Lemma D.3 we derive

$$\begin{aligned} S(\widehat{\theta}^{\text{agg}}-\theta^\star) &= \tfrac{1}{2}\left(s_{W_n}(X,D)+s_{W_n}(X,D')\right) + o_P(|W_n|^{-1/2}) \\ &= T_{1,W_n}(X) + \tfrac{1}{2}\left(T_{2,W_n}(X,D)+T_{2,W_n}(X,D')\right) + o_P(|W_n|^{-1/2}). \quad \text{(D.9)} \end{aligned}$$

Let $\widetilde{T}_{2,W_n}(X,D,D') = \left(T_{2,W_n}(X,D)+T_{2,W_n}(X,D')\right)/2$. Given X, $T_{2,W_n}(X,D)$ and $T_{2,W_n}(X,D')$ are independent and (normalized by $|W_n|^{-1/2}$) both tend to a Gaussian vector with covariance matrix G_2^{sb}. Therefore as $n\to\infty$, we have that given X

$$|W_n|^{-1/2}\widetilde{T}_{2,W_n}(X,D,D') \xrightarrow{d} \mathcal{N}(0, G_2^{\text{sb}}/2).$$

Then, following Step 3 of D.2 we can prove that $|W_n|^{-1/2}(T_{1,W_n}(X)+\widetilde{T}_{2,W_n}(X,D,D'))$ tends to a Gaussian vector with covariance matrix $G_1+G_2^{\text{sb}}/2$. Thus, from (D.9)

$$|W_n|^{1/2}(\widehat{\theta}^{\text{agg}}-\theta^\star) \xrightarrow{d} \mathcal{N}\left(0, S^{-1}(G_1+G_2^{\text{sb}}/2)S^{-1}\right).$$

Following the proof of D.3, $\widehat{S}^{\text{agg}}$ is a consistent estimate of S. This leads to the statement $|W_n|^{1/2}(\widehat{G}^{\text{agg}})^{-1/2}(\widehat{\theta}-\theta^\star) \xrightarrow{d} \mathcal{N}(0,I_p)$. And we apply the same argument as previously to assert that the same result holds for non ergodic Gibbs measures.