Pseudo Bayesian inference for intensity-dependent point processes

Kasper K. Berthelsen[1]

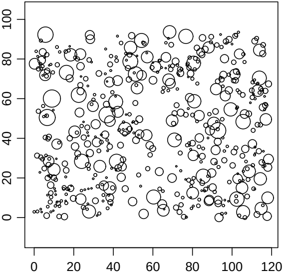
[1]Department of Mathematical Sciences
Aalborg University

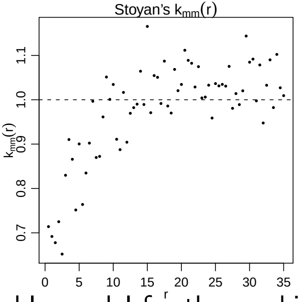
Joint work with Mari Myllymäki

Avignon, May 2012

Motivation

We are considering marked point patterns $\{(x_i, m_i)\}$, where $\{x_i\}$ denotes the locations of objects (trees) in "window" W, and $\{m_i\}$ denotes the corresponding marks (stem diameter at breast height (DBH)).

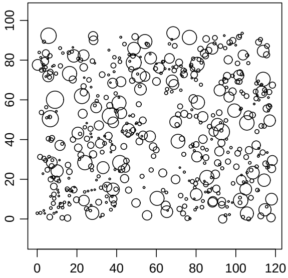


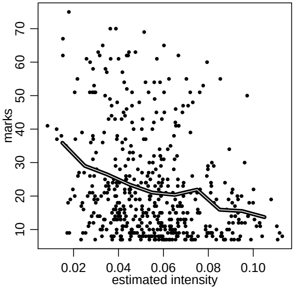


- We want to construct a reasonable model for the marking (and distribution) of points

Hainich data

Data: Location of 650 trees marked by dbh in a $118.5m \times 93.75m$ region. The trees belong to a mixed broad-leaved forest in Hainich in Western Thuringia (Germany), as so-called selection forest (Plenterwald).





- Left plot suggests inhomogeneous point distribution.
- Right plot suggests mark distribution depends on point intensity.

Intensity dependent marking

We consider a situation where there is a relation between the marks and the intensity of the point pattern.

Two examples where this is relevant

- Preferential sampling: One makes more measurements where the measured value (i.e. the mark) is high, e.g. pollution, see [Diggle et al., 2010]
- Density-dependence in plant ecology: In areas with relatively many trees the trees tend to be small, and vice versa. See [Myllymäki and Penttinen, 2009].

The model

We consider a model with a density

$$\pi((x_i, m_i\}|\beta) = \frac{1}{c(\beta, \theta_m, \theta_\varphi)} \prod_{i=1}^{n} \beta(x_i)\pi(m_i|\beta(x_i)) \\ \times \prod_{i<j} \varphi((x_i, m_i), (x_j, m_j); \theta_\varphi), \quad (1)$$

w.r.t. a Poisson process on $W \times \mathbb{R}_+$.

$\beta : W \to \mathbb{R}_+$ is the first order term.

Conditional on β and $\{x_i\}$ the marks are then distributed as

$$m_i|x_i, \beta \sim \pi(m_i|\beta(x_i), \theta_m),$$

i.e. the distribution of mark m_i depends on β evaluated at the location x_i and parameters θ_m.

Here where $\varphi : (W \times M) \times (W \times M) \to [0, 1]$ is the interaction function.

Specifying the interaction function φ

Specifically we choose

$$\varphi((x_i, m_i), (x_j, m_j)) = \begin{cases} \gamma & \text{if } \| x_i - x_j \| \leq R(m_i + m_j) \\ 1 & \text{otherwise,} \end{cases}$$

where $R \geq 0$ controls the interaction range and $\gamma \in [0, 1]$ controls the strength of the interaction.

Interpretation:

- Circular influence zones, where the diameter of the influence zone centred at x_i is proportional to m_i (DBH).
- The interaction parameter γ specifies the degree of "penalty" on each pair of overlapping influence zones.

Mark distribution

Regarding the mark distribution, we assume

$$m_i - m_0|\theta, \beta(x_i) \sim \Gamma\left[c, \frac{1}{c}\left(a + \frac{b}{\sqrt{\beta(x_i)}}\right)\right],$$

where $m_0 \geq 0$ is the minimum mark size, and $\Gamma(k, \theta)$ denotes the gamma distribution with shape parameter k and scale parameter θ. Hence

$$\mathbb{E}[m_i - m_0|\theta, \beta(x_i)] = a + \frac{b}{\sqrt{\beta(x_i)}} \quad \text{and} \quad \frac{\mathbb{V}\mathrm{ar}[m_i - m_0|\theta, \beta(x_i)]}{(\mathbb{E}[m_i - m_0])^2} = \frac{1}{c}.$$

The special case, where $m_0 = 0$ and $a = 0$ we obtain a situation which is similar to location dependent scaling considered by [Hahn et al., 2003].

Bayesian inference

We perform Bayesian posterior inference for

- β the first order term
- a, b, c parameters of the mark distribution
- R, γ the interaction parameters

Priors

- For a, b, c, R and γ we assume uniform priors on a bounded interval.
- For β we assume a non-parametric approach

Prior distribution for β

As a prior on β we use a shot noise style prior

$$\beta(x) = \sum_{c \in \mathcal{C}} \lambda \mathcal{K}(x - c),$$

where $\lambda > 0$, $\mathcal{C}$ is a Poisson process on $\mathbb{R}^2$ and $\mathcal{K}$ is a kernel, i.e. a probability density on $\mathbb{R}^2$. This is the prior used by [Berthelsen and Møller, 2008] (in the 1-dimensional case).

One alternative is a log Gaussian random field. This is the prior considered by H&S (2008) and M&P (2009)

Approximative prior

For the remainder we focus of the shot-noise prior:

$$\beta(x) = \sum_{c \in \mathcal{C}} \lambda \mathcal{K}(x - c).$$

For simulation purposes we replace the Poisson process $\mathcal{C}$ on $\mathbb{R}^2$ by a Poisson process $\mathcal{C}_+$ on an extended window

$$W_+ = \{x \in \mathbb{R}^2 : \delta(x, W) \leq \Delta\}, \quad \Delta \geq 0,$$

where

$$\delta(A, B) = \inf_{x \in A, y \in B} \| x - y \|, \quad A, B \subseteq \mathbb{R}^2.$$

Further, we assume C_+ has intensity $\beta_{\mathcal{C}}$, and that $\mathcal{K}$ is the density of a bivariate normal distribution with covariance matrix $\sigma^2 I$.

How to choose Δ

The prior mean of β is $[\beta(x)] = \lambda\beta_{\mathcal{C}}$. When restricting $\mathcal{C}$ to W_+ the prior mean is (obviously) reduced. But by how much?

Let D denoted the (missed) contribution for kernels centred outside W_+:

$$D = \int_W \sum_{c\in\mathcal{C}\setminus W_+} \lambda\mathcal{K}(x, c)\mathrm{d}c.$$

How to choose Δ

The prior mean of β is $[\beta(x)] = \lambda\beta_{\mathcal{C}}$. When restricting $\mathcal{C}$ to W_+ the prior mean is (obviously) reduced. But by how much?

Let D denoted the (missed) contribution for kernels centred outside W_+:

$$D = \int_W \sum_{c \in \mathcal{C} \setminus W_+} \lambda \mathcal{K}(x, c) \mathrm{d}c.$$

Then the expected value of D is

$$\mathbb{E}[D] = \lambda\beta_{\mathcal{C}} \int_{\mathbb{R}^2 \setminus W_+} \int_W \mathcal{K}(x, c) \mathrm{d}x \mathrm{d}c$$

How to choose Δ

The prior mean of β is $[\beta(x)] = \lambda\beta_{\mathcal{C}}$. When restricting $\mathcal{C}$ to W_+ the prior mean is (obviously) reduced. But by how much?

Let D denoted the (missed) contribution for kernels centred outside W_+:

$$D = \int_W \sum_{c \in \mathcal{C} \setminus W_+} \lambda \mathcal{K}(x, c) \mathrm{d}c.$$

Then the expected value of D is

$$\begin{aligned}
\mathbb{E}[D] &= \lambda\beta_{\mathcal{C}} \int_{\mathbb{R}^2 \setminus W_+} \int_W \mathcal{K}(x, c) \mathrm{d}x \mathrm{d}c \\
&\leq \lambda\beta_{\mathcal{C}} \int_{\mathbb{R}^2 \setminus W_+} \int_W k(x, c) \mathrm{d}x \mathrm{d}c,
\end{aligned}$$

where

$$k(x, c) \geq \mathcal{K}(x, c) \text{ for all } (x, c) \in W \times (\mathbb{R}^2 \setminus W_+)$$

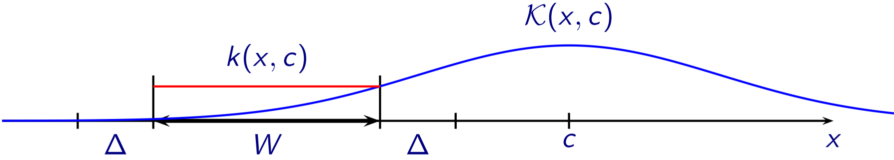
is chosen to make integration easier.

Choosing k

Following "B&M 2008", the function $k(x, c)$ is chosen so that it is constant on W:

$$k(x,c) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\delta(c,W)^2}{2\sigma^2}\right)$$

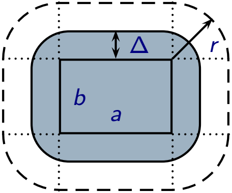
Illustration of the 1-dimensional case:



Note: The 1-dimensional case is consider by B&M (2008) where the introduction of bounding function k is not needed.

Bound on $\mathbb{E}[D]$

$$\mathbb{E}[D] \leq \lambda\beta_{\mathcal{C}} \int_{\mathbb{R}^2 \setminus W_+} \int_W k(x,c)\mathrm{d}x\mathrm{d}c$$
$$= \lambda\beta_{\mathcal{C}}|W| \int_{\Delta}^{\infty} \left[2(a+b)/(2\pi\sigma^2) + r/\sigma^2\right] e^{-r^2/(2\sigma^2)} dr$$



The proportion contribution missed:

$$\frac{\mathbb{E}[D]}{\int_W \mathbb{E}[\beta(x)]\mathrm{d}x} = \int_{\Delta}^{\infty} \left[2(a+b)/(2\pi\sigma^2) + r/\sigma^2\right] e^{-r^2/(2\sigma^2)}\mathrm{d}r$$

Finally, Δ is determined using numerical methods.

Posterior simulations

We want to explore the posterior distribution, $\pi(\theta, \beta|x) \propto \pi(x|\theta, \beta)\pi(\theta, \beta)$, using MCMC.

For convenience we write the likelihood as

$$\pi((\mathbf{x}, \mathbf{m})|\theta, \beta) = c^{-1}(\theta, \beta) f(x|\theta, \beta),$$

where c^{-1} is the unknown normalising constant of $f(y|\theta)$.

Posterior simulations

We want to explore the posterior distribution, $\pi(\theta, \beta|x) \propto \pi(x|\theta, \beta)\pi(\theta, \beta)$, using MCMC.

For convenience we write the likelihood as

$$\pi((\mathbf{x}, \mathbf{m})|\theta, \beta) = c^{-1}(\theta, \beta) f(x|\theta, \beta),$$

where c^{-1} is the unknown normalising constant of $f(y|\theta)$.

Using (conventional) Metropolis-Hastings updates involves evaluating the Hastings ratio:

$$H(\theta, \theta') = \frac{c^{-1}(\theta', \beta') f(x; \theta', \beta') \pi(\theta', \beta') q(\theta', \beta'; \theta, \beta)}{c^{-1}(\theta, \beta) f(x; \theta, \beta) \pi(\theta, \beta) q(\theta, \beta; \theta', \beta')}$$

Notice this involves evaluating a ratio of unknown normalising constants.

Difficulty of avoiding the normalising constant

- There are several ways of circumventing the problem of ratios of unknown normalising constants, e.g. the approaches by [Møller et al., 2006] or [Murray et al., 2006].
- For both of these approaches, each MCMC step involves simulating (perfectly) a realisation of the mark point process conditional on the proposed values of β, γ, R, a and b.
- These perfect realisations be achieved by perfect sampling (dominating coupling from the past [Kendall and Møller, 2000]).
- For many relevant problems however, perfect sampling is infeasible. Instead we we consider a *Pseudo Bayesian* approach.

The Pseudo likelihood

In a pseudo Bayesian approach the likelihood is (simply) replaced by the pseudo likelihood:

$$PL(\theta|(\mathbf{x},\mathbf{m})) = \prod_{i=1}^{n} \lambda_\theta((x_i, m_i); (\mathbf{x},\mathbf{m}))\times$$
$$\exp\left(-\int_W \int_M \lambda_\theta((y, l); (\mathbf{x},\mathbf{m})) \mathrm{d}l \mathrm{d}y\right),$$

where λ_θ is the Papangelou conditional intensity:

$$\lambda_\theta((y, l), (\mathbf{x},\mathbf{m}))$$
$$= \beta(y)\pi(l|\beta(y), \theta) \times \prod_{i=1}^{n} \varphi((y, l), (x_i, m_i)).$$

Usually the integral in the PL-function is approximated using some discretisation scheme.

Integral over W

- A discretisation of the (location) space W is done by dividing W into disjoint "cells" W_j and associating each cell with a **dummy point**

$$y_j \in W, j = 1, \ldots, J.$$

- The integral over W is then approximated by assuming that the integrand is constant on each cell, with a value obtained at the corresponding dummy point.
- Notice that the usual "practical pseudo-likelihood approach" of including the data point in the grid of dummy points introduces bias (unless you do a clever correction).

Integral over M

The integral over the markspace $M = \mathbb{R}_+$ is

$$\int_M \pi(l|\beta(x), \theta_m) \prod_{i=1}^{n} \varphi((y_j, l), (x_i, m_i)) \mathrm{d}l$$

The product of interaction functions can be written as

$$\prod_{i=1}^{n} \varphi((y_j, l), (x_i, m_i)) = \gamma^{S_R((y_j, l), (\mathbf{x}, \mathbf{m}))}. (**)$$

where

$$S_R((y_j, l), (\mathbf{x}, \mathbf{m})) = \sum_{i=1}^{n} \mathbf{1}\left(\frac{\| y_j - x_i \|}{m_i + l} < R \right)$$

In other word, (**) is a decreasing step function of l, where each step is a factor γ lower than the previous step.

Steps

In summary

$$\prod_{i=1}^{n} \varphi((y_j, l), (x_i, m_i)) = \gamma^{S_R((y_j,l),(\mathbf{x},\mathbf{m}))}$$

is a step function.

For the dummy point y_j steps happen at $d_{j,1}, d_{j,2}, \ldots, d_{j,n}$, where

$$d_{j,i} = \max\left\{0, \frac{\| x_i - y_j \|}{R} - m_i\right\}$$

In the following we assume that $d_{j,1} \leq d_{j,2} \leq \ldots \leq d_{j,n}$, and $d_{j,0} = 0$ and $d_{j,n+1} = \infty$.

Rewriting integral

Assume $F(m|\theta_m, \beta(x))$ is the distribution function corresponding to the mark density $\pi(m|\theta_m, \beta(x))$.

The integral

$$\int_M \pi(l|\beta(x), \theta_m) \prod_{i=1}^{n} \varphi((y_j, l), (x_i, m_i)) \mathrm{d}l$$

can now be written as

$$\sum_{i=1}^{n+1} \left(F(d_{j,i}; \beta(y_j), \theta_m) - F(d_{j,i-1}; \beta(y_j), \theta_m) \right) \gamma^{i-1}.$$

The $d_{j,i}$s are to be pre-calculated for each dummy point.

Approximate pseudo likelihood

The pseudo likelihood can now be approximated by

$$PL(\theta) \approx \prod_{i=1}^{n} \lambda_\theta((x_i, m_i), (\mathbf{x}, \mathbf{m})) \times$$

$$\exp\Big(-\sum_{j=1}^{J} |W_j| \beta(y_j)$$

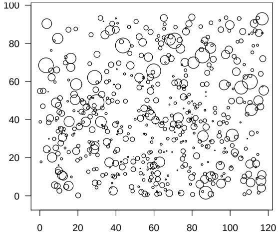
$$\Big[\sum_{i=1}^{n+1} (F(d_{j,i}; \beta(y_j), \theta_m) - F(d_{j,i-1}; \beta(y_j), \theta_m)) \gamma^{i-1} \Big] \Big)$$

With this in place we turn to an example.

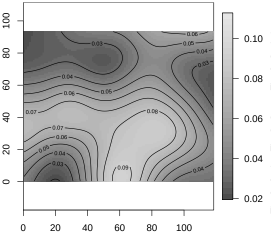
Simulation example

In this example $W = [0, 118.5] \times [0, 93.7]$, $a = 0.2$, $b = 2$, $c = 2.5$, $\gamma = 0.1$ $R = 0.02$, $\lambda = 20$, $\beta_{\mathcal{C}} = 0.003$.

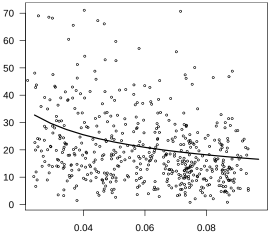
Data



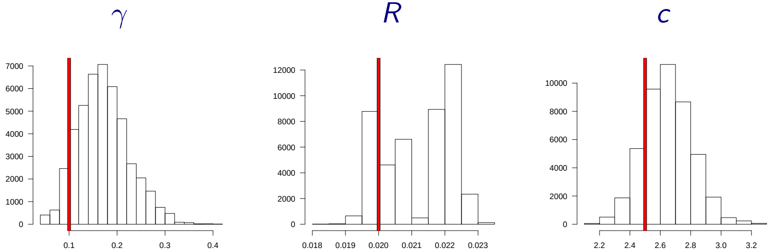
β



m_i vs $\beta(x_i)$



Posterior distribution: Interaction



In details: Posterior distribution of R

Recall pseudo likelihood:

$$PL(\theta) \approx \prod_{i=1}^{n} \lambda_\theta((x_i, m_i), (\mathbf{x}, \mathbf{m})) \times$$
$$\exp\Big(- \sum_{j=1}^{J} |W_j| \beta(y_j) \Big[\sum_{i=1}^{n+1} (F(d_{j,i}; \ldots) - F(d_{j,i-1}; \ldots)) \gamma^{i-1} \Big] \Big)$$

As a function of R, $\prod_{i=1}^{n} \lambda_\theta((x_i, m_i), (\mathbf{x}, \mathbf{m}))$ is a decreasing stepfunction.

In details: Posterior distribution of R

Recall pseudo likelihood:

$$PL(\theta) \approx \prod_{i=1}^{n} \lambda_\theta((x_i, m_i), (\mathbf{x}, \mathbf{m})) \times$$

$$\exp\Big(-\sum_{j=1}^{J} |W_j| \beta(y_j) \Big[\sum_{i=1}^{n+1} (F(d_{j,i}; \ldots) - F(d_{j,i-1}; \ldots)) \gamma^{i-1} \Big] \Big)$$

As a function of R, $\prod_{i=1}^{n} \lambda_\theta((x_i, m_i), (\mathbf{x}, \mathbf{m}))$ is a decreasing stepfunction. On the other hand, as a function of R

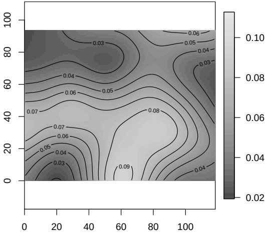
$$\sum_{i=1}^{n} (F(d_{j,i}; \beta(x), \theta_m) - F(d_{j,i-1}; \beta(x), \theta_m)) \gamma^{i-1}$$

is a continuous, decreasing function.

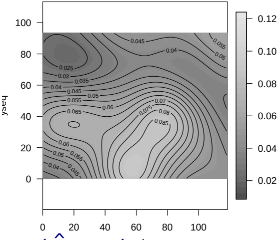
As a result PL becomes "saw toothed".

Posterior distribution: β

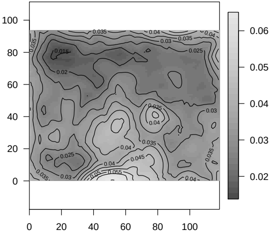
True β:



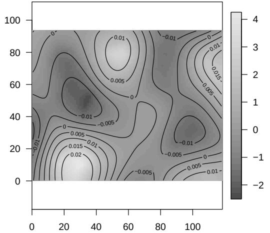
Posterior mean β:



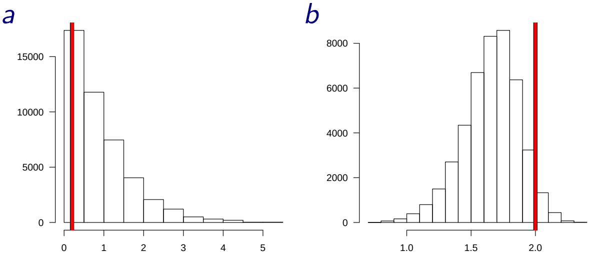
Std. error



$(\hat{\beta} - \beta)$/std. error

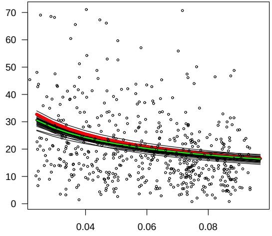


Posterior distribution: β dependence



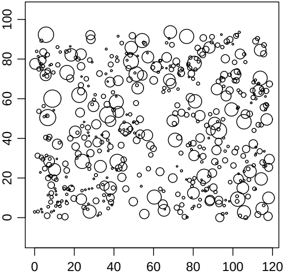
m_i vs $\hat{\beta}(x)$

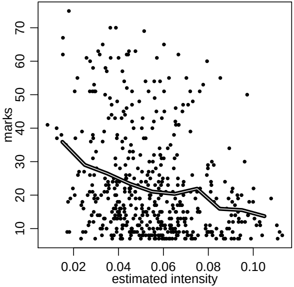
$\mathbb{E}[m_i]$ vs $a + \frac{b}{\sqrt{\hat{\beta}}}$



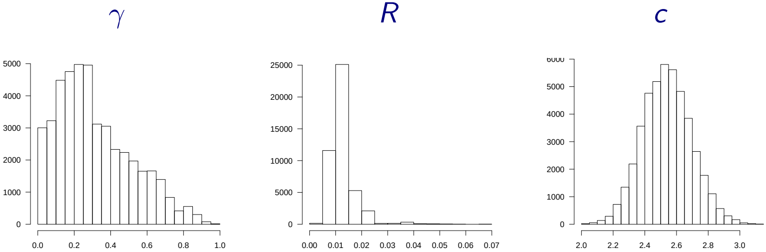
Hainich data

Data: Location of 650 trees marked by dbh in a $118.5m \times 93.75m$ region. The trees belong to a mixed broad-leaved forest in Hainich in Western Thuringia (Germany), as so-called selection forest (Plenterwald).

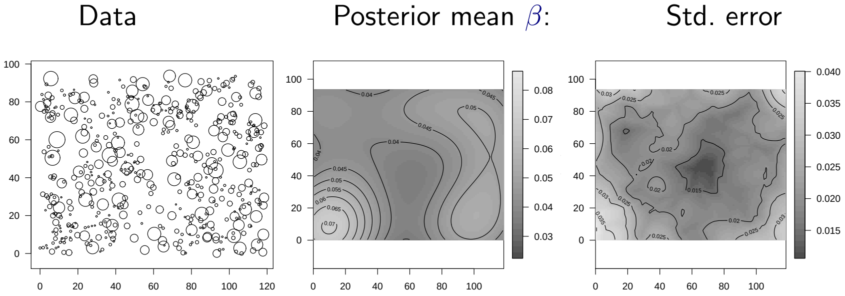




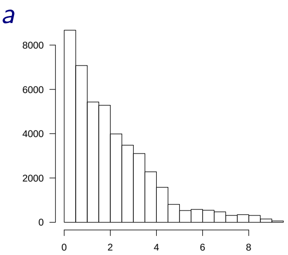
Posterior distribution: Interaction

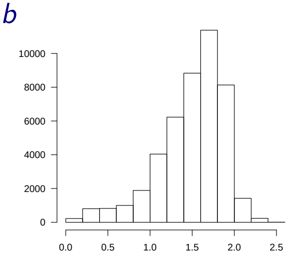


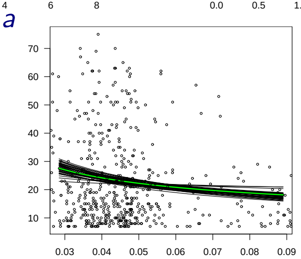
Posterior distribution: β



Posterior distribution: β dependence







Berthelsen, K. K. and Møller, J. (2008).
Non-parametric bayesian inference for inhomogeneous Markov point processes.
Australian & New Zealand Journal of Statistics, 50:257–272.

Diggle, P. J., Menezes, R., and Su, T.-L. (2010).
Geostatistical inference under preferential sampling.
Journal of the Royal Statistical Society, Ser. B, 59:191–232.

Hahn, U., Jensen, E. B. V., van Lieshout, M.-C., and Nielsen, L. S. (2003).
Inhomogeneous spatial point processes by location dependent scaling.
Adv. Appl. Prob., 35:319–336.

Ho, L. P. and Stoyan, D. (2008).
Modelling marked point patterns by intensity-marked cox processes.
Statistics and Probability Letters, 78:1194–1199.

Kendall, W. S. and Møller, J. (2000).
Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes.
Adv. Appl. Prob., 32:844–865.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006).
An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants.
Biometrika, 93:451–458.

Murray, I., Ghahramani, Z., and MacKay., D. J. C. (2006).
MCMC for doubly-intractable distributions.
In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia. AUAI Press.