 Vol 25: 435–450, 1998

Non-parametric Bayesian Estimation of a Spatial Poisson Intensity

JUHA HEIKKINEN

Finnish Forest Research Institute

ELJA ARJAS

Rolf Nevanlinna Institute

ABSTRACT. A method introduced by Arjas & Gasbarra (1994) and later modified by Arjas & Heikkinen (1997) for the non-parametric Bayesian estimation of an intensity on the real line is generalized to cover spatial processes. The method is based on a model approximation where the approximating intensities have the structure of a piecewise constant function. Random step functions on the plane are generated using Voronoi tessellations of random point patterns. Smoothing between nearby intensity values is applied by means of a Markov random field prior in the spirit of Bayesian image analysis. The performance of the method is illustrated in examples with both real and simulated data.

Key words: Markov chain Monte Carlo, Markov random fields, non-parametric Bayesian inference, spatial point processes, Voronoi tessellations

1. Introduction

Arjas & Gasbarra (1994) introduced a new approach to the non-parametric Bayesian estimation of the intensity (or hazard rate) of a non-homogeneous Poisson process on the real line. The basic idea was to use piecewise constant functions with a random number and random locations of jump times to approximate "real" (smooth) intensity functions. In this way an intensity defined on a finite interval was parametrized by a finite number of real numbers. Variability of this number led, however, to an infinite-dimensional parameter space.

The form of piecewise constant intensities was chosen as a convenient way of arriving at a simple model formulation and straightforward calculation for the posterior. Since Bayesian inference is not concerned with selecting a point estimate (here single intensity function) from the postulated model class, the precise functional form of its individual members is not as crucial as in the frequentist approach. More important is that the integrals of test functions of interest (e.g. predictive densities or probabilities) w.r.t. the posterior distribution obtained from the approximate model are close to those obtained from the "true" model (see Arjas & Andreev, 1996). Furthermore, a "Bayesian point estimate", the posterior mean, does not necessarily belong to the model class. In the present case pointwise posterior means do not need to form a piecewise constant function since the jump times are variable, and indeed the posterior mean is typically a smooth continuous function. Further discussion on the topic can be found in the papers cited above and in Arjas (1996).

In Arjas & Gasbarra (1994) a prior distribution on the space of random step functions, or jump processes, was specified in terms of the corresponding local characteristics. A martingale structure was assumed, which penalizes large differences between nearby function values. The aim was, besides smoothing the oscillations, to have the change points concentrated on the areas where the intensity is changing most rapidly.

The purpose of this work is to develop a similar method for the estimation of the intensity of a spatial Poisson process. There are three non-trivial problems in doing this. First, how to generate

random step functions in a flexible way that allows for local updating? We use Voronoi tessellations of random point patterns to provide partitions to domains of constant intensity. Hence, step functions are defined through marked point patterns, where the mark of a point gives the intensity value inside the associated Voronoi tile. The second question is, how to model similarity of nearby intensity values? The martingale model of Arjas & Gasbarra is inherently one-dimensional. We apply the scheme of conditional autoregression, by specifying the conditional distributions of individual intensity levels given the values of all others. Finally, there is the problem of sampling from the posterior distribution, which has a variable dimension because of the random number of steps. Arjas & Gasbarra modified the Gibbs sampler in a way that relied on a simple ordering of the generating points. In spatial point process literature simulation methods for variable dimensional patterns already have a relatively long history (see, e.g. Preston, 1977; Ripley, 1977). More recently, Geyer & Møller (1994) developed a Metropolis–Hastings algorithm for that purpose. Relying on the marked point process interpretation of our posterior, these methods would be applicable here as well. Green (1995) formulated a more general framework for variable dimensional problems, and also presented applications closely related to ours. We have followed Green's approach in designing the sampler.

Green (1995) presented two examples where the idea of "dynamic step functions" was applied to the estimation of an intensity function on the real line and of a surface on the plane. In the latter one two-dimensional step functions were also generated from Voronoi tessellations. The main concern in these examples was in finding change-points and boundaries in functions that are truly discontinuous, and accordingly independence was assumed between values of the step functions in different regions. In this sense our method is more general. In addition, concurrent work on non-parametric Bayesian curve estimation includes Denison *et al.* (1998), where sequences of piecewise polynomials are used instead of our step functions. Møller *et al.* (1998) is also closely related to the present paper, particularly its sec. 8 dealing with empirical Bayesian inference for an intensity surface of a point process. A one-dimensional version of the algorithm introduced here can be found in Arjas & Heikkinen (1997).

Much of earlier work (until 1994) on non-parametric Bayesian curve estimation is nicely reviewed in Hjort (1996), although that paper is specifically concerned with density estimation. Three alternative approaches are discussed there. The first group of ideas uses the Dirichlet process, or some of its relatives, as the main tool for specifying a prior. The second approach is to place a prior on the coefficients of series expansions, and the third to use locally parametric approximations to the true curve, and then place priors on these local parameters. In a sense, our approach is closest to this third one.

The plan of this paper is as follows. In section 2 we describe in detail the Bayesian model. Section 3 explains the algorithm for sampling from the posterior, and section 4 provides some examples based on both real and synthetic data. The paper concludes with a short discussion in section 5.

2. Model

Suppose we observe a non-homogeneous Poisson process within a bounded sampling window $S \subset \mathbf{R}^2$. The likelihood $p(\mathbf{x}|\lambda)$ (the density with respect to the unit intensity Poisson process) of point pattern $\mathbf{x} = \{x_1, \ldots, x_N\} \subset S$ is proportional to

$$\exp\left\{-\int_S \lambda(x)\nu(dx)\right\} \prod_{n=1}^{N} \lambda(x_n), \tag{2.1}$$

where λ: $\mathbf{R}^2 \to [0, \infty)$ is the *intensity function* of the process, and ν is the Lebesgue measure

on $\mathbf{R}^2$. We consider the inference concerning λ on a domain E containing S. We may choose E to be larger than S to reduce edge effects, and also for prediction outside S (see section 5).

We construct a prior distribution for λ by considering the set of positive valued step functions on E as a support. We use random partitions of E to determine domains of constant intensity; we also let the number of such domains be random. Partitions $\mathscr{E}(\boldsymbol{\xi}) = \{E_k(\boldsymbol{\xi})\}$ of E are generated via patterns $\boldsymbol{\xi}$ of generating points $\xi_k \in E$ so that the domain $E_k(\boldsymbol{\xi})$ consists of those points of E which are closer to ξ_k than to any other point of $\boldsymbol{\xi}$. That is $\mathscr{E}(\boldsymbol{\xi})$ is the *Voronoi tessellation* of $\boldsymbol{\xi}$. A step function

$$\lambda = \sum_k \lambda_k \mathbf{1}_{E_k} \tag{2.2}$$

is then defined by attaching to each generating point ξ_k a mark λ_k, the intensity level on the corresponding Voronoi tile E_k.

The main reason for using Voronoi tessellations is that they can be updated locally: when a new generating point is added to a tessellated pattern (or when one is deleted), only the tiles adjacent to the new (or deleted) one need to be modified (see Fig. 1). This property is essential for the simulation algorithm of section 3 to be computationally feasible.

The prior distribution of the generating point pattern $\boldsymbol{\xi}$ is taken to be the homogeneous Poisson process on E with a given intensity $\lambda_{\boldsymbol{\xi}} \in (0, \infty)$, and with zero probability assigned to the empty pattern. For a pattern $\boldsymbol{\xi}$ with K points, the prior density $p(\boldsymbol{\xi})$ w.r.t. the unit intensity Poisson process is then proportional to $\lambda_{\boldsymbol{\xi}}^K$ if $K > 0$, and 0 otherwise.

The prior distributions of marks λ_k is developed conditionally on the unmarked pattern $\boldsymbol{\xi}$, and it will contain an assumption of smoothness in the sense that the differences $|\lambda_k - \lambda_j|$ between intensities on adjacent tiles are expected to be small. Given a pattern $\boldsymbol{\xi}$ with K points, let us index the points arbitrarily as $\xi_1, \ldots, \xi_K$, and treat $\boldsymbol{\xi}$ as a vector $(\xi_1, \ldots, \xi_K)$. A multivariate Gaussian prior

$$p(\boldsymbol{\eta}|\boldsymbol{\xi}) \propto (2\pi)^{-K/2} |\mathbf{Q}|^{1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{Q} (\boldsymbol{\eta} - \boldsymbol{\mu})\} \tag{2.3}$$

is assigned to the corresponding vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ of log-intensities $\eta_k = \log \lambda_k$. Here K naturally depends on $\boldsymbol{\xi}$, but also the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and the precision matrix $\mathbf{Q}$ may in general be functions of $\boldsymbol{\xi}$, although we will suppress these dependencies from the notation for clarity.

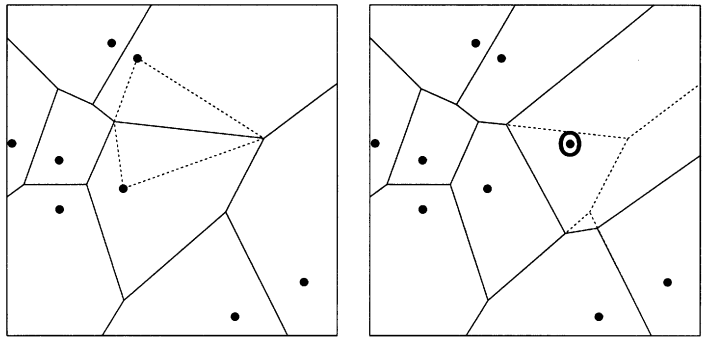

Fig. 1. Left: a Voronoi tessellation (solid lines); broken lines delineate two "sector triangles" referred to on page 438. The edge between them is perpendicular to the line passing through the two generating points and it is half way between those points. Right: when a new generating point (the circled one) is added, its Voronoi tile consists of parts "conquered" from the neighbouring tiles, other tiles are not affected.

The expectations μ_k may be chosen according to any prior knowledge of local intensities that may be available. They could, for example, be functions of covariate values attached to the corresponding tiles E_k. From now on we will assume, however, that they have been chosen to be all equal, $\mu_k = \mu$ for all k.

The covariance structure is specified by modeling $p(\boldsymbol{\eta}|\boldsymbol{\xi})$ as a Markov random field on the *Delaunay graph* of $\boldsymbol{\xi}$: Sites k and j are *neighbours*, $k \sim j$ (or, more precisely, $k \overset{\boldsymbol{\xi}}{\sim} j$; reference to $\boldsymbol{\xi}$ will be made only when there is a possibility of confusion) if tiles E_k and $E_j^{\boldsymbol{\xi}}$ of the Voronoi tessellation of $\boldsymbol{\xi}$ share a common edge, and the conditional distribution $p(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi})$, where $\boldsymbol{\eta}_{-k}$ denotes the sequence $\boldsymbol{\eta}$ with η_k removed, depends only on the neighbouring values η_j, $j \sim k$ (and on E_k). This implies that only those elements Q_{kj} of $\mathbf{Q}$, for which $j = k$ or $j \sim k$, may be non-zero, and that the conditional distributions $p(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi})$ are Gaussian with expectations

$$E(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi}) = \mu + \sum_{j:j\sim k} \beta_{kj}(\eta_j - \mu), \tag{2.4}$$

where $\beta_{kj} = -Q_{kj}/Q_{kk}$, and with variances

$$\mathrm{var}\,(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi}) = \sigma_k^2 = Q_{kk}^{-1}. \tag{2.5}$$

In specifying the joint distribution (2.3) via the local characteristics (2.4) and (2.5), that is, in choosing the parameters β_{kj} and σ_k^2, some consistency conditions must be imposed. The symmetry of $\mathbf{Q}$ requires that

$$\beta_{kj}\sigma_j^2 = \beta_{jk}\sigma_k^2. \tag{2.6}$$

The matrix $\mathbf{Q}$ must also be positive definite, for which a simple sufficient condition (Besag & Kooperberg, 1995) is that the β_{kj} are all non-negative and

$$\sum_{j:j\sim k} \beta_{kj} < 1, \quad \text{for all } k. \tag{2.7}$$

The role of the prior as smoother now becomes apparent as

$$E(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi}) = \left(1 - \sum_{j:j\sim k} \beta_{kj}\right)\mu + \sum_{j:j\sim k} \beta_{kj}\eta_j \tag{2.8}$$

is a weighted average of the prior expectation μ and the neighbouring levels η_j, $j \sim k$.

A simple and yet rather flexible scheme satisfying the above restrictions is given by

$$\beta_{kj} = \frac{l_{kj}}{l_k}\beta \quad \text{and} \quad \sigma_k^2 = \frac{\sigma^2}{l_k} \tag{2.9}$$

with hyperparameters $\beta \in [0, 1)$ and $\sigma^2 > 0$. Here l_{kj} and l_k are some simple functions of the generating points $\boldsymbol{\xi}$ satisfying

$$l_{kj} = l_{jk} \quad \text{and} \quad \sum_{j:j\sim k} l_{kj} \leqslant l_k. \tag{2.10}$$

The simplest choice would be to have all l_{kj} equal ($= 1$) and l_k equal to the number of neighbours of k. We wish, however, to encourage adaptivity by allowing rapid changes where the tiles are small and, on the other hand, have strong correlation between those neighbours which share a long edge. Accordingly, we choose l_k to be the area of tile E_k. In order for (2.10) to hold we then let l_{kj} be the area of the "sector triangle", which has the common edge of tiles E_k and E_j as one of its sides, and a vertex at the generating point ξ_k (see Fig. 1). Owing to the basic properties of Voronoi tessellations, the corresponding sector of E_j is a mirror image,

hence the symmetry $l_{jk} = l_{kj}$. Furthermore, the common edge is perpendicular to the line passing through the two generating points, and hence the area can be easily calculated as one-fourth of the product of the length of the common edge and the distance between the two generating points. Another natural choice would have been to let l_{kj} be the length of the common edge. In that case l_k should be the perimeter of tile E_k.

The precision matrix of our multivariate Gaussian prior $p(\boldsymbol{\eta}|\boldsymbol{\xi})$ can now be written as $\mathbf{Q} = (1/\sigma^2)\boldsymbol{\Gamma}$, where

$$\Gamma_{kj} = \begin{cases} l_k & \text{if } j = k, \\ -\beta l_{kj} & \text{if } j \sim k, \text{ and:} \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

The entire prior distribution $p(\boldsymbol{\xi}, \boldsymbol{\eta}) = p(\boldsymbol{\xi})p(\boldsymbol{\eta}|\boldsymbol{\xi})$ has four hyperparameters: $\lambda_{\boldsymbol{\xi}}$ controls the resolution, μ gives the expected overall level of log-intensity, β determines the weighting between μ and the neighbouring levels, and σ^2 between the prior and the data. Consider, for a moment, a scheme where $l_k = \sum_{j:j\sim k} l_{kj}$. As β approaches 1, this prior distribution tends to the improper pairwise difference prior (Besag, 1989; Besag *et al.*, 1995)

$$p(\boldsymbol{\eta}|\boldsymbol{\xi}) \propto (2\pi\sigma^2)^{-K/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{(k,j):k\sim j} l_{kj}(\eta_k - \eta_j)^2 \right\}, \tag{2.12}$$

and μ disappears. Hence, in the case where prior knowledge of the intensity level is vague, we can give β a value close to 1, and the choice of μ is not crucial. We are then left with two hyperparameters, $\lambda_{\boldsymbol{\xi}}$ and σ^2, which control the degree of smoothing. In principle, $\lambda_{\boldsymbol{\xi}}$ determines how fine details of the intensity surface are shown. Very large values may result in wiggly intensities that are following the data too closely, although this can to some extent be counterbalanced by decreasing σ^2. Also the computing time increases with $\lambda_{\boldsymbol{\xi}}$. Naturally, there is the option of treating (some of) these parameters as random variables, by building one more level of hierarchy into the model. As pointed out by one of the referees, this might also improve the mixing of the MCMC-sampler; varying σ^2, for example, can be seen as a form of simulated tempering (see e.g. Geyer & Thompson, 1995). On the other hand, it may be useful to try out different degrees of smoothing and use σ^2 as a control variable.

For the piecewise constant log-intensity function $\sum \eta_k \mathbf{1}_{E_k}$ determined by $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, the Poisson likelihood (2.1) can be written as

$$p(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\eta}) = \exp\left[\sum_{k=1}^{K} \{ N(E_k)\eta_k - \nu(E_k \cap S) \exp(\eta_k) \} \right], \tag{2.13}$$

where $N(A)$ is the number of points of $\mathbf{x}$ located within domain A. Our inference concerning the intensity is now based on sampling from the resulting posterior distribution

$$\begin{aligned} p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{x}) &\propto p(\boldsymbol{\xi})p(\boldsymbol{\eta}|\boldsymbol{\xi})p(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\eta}) \\ &\propto \lambda_{\boldsymbol{\xi}}^K (2\pi\sigma^2)^{-1/2K} |\boldsymbol{\Gamma}|^{1/2} \exp\left[\sum_k \left\{ -\frac{1}{2\sigma^2} \left(\Gamma_{kk}(\eta_k - \mu)^2 \right.\right.\right. \\ &\quad \left.\left.\left. + \sum_{j:j\sim k} \Gamma_{kj}(\eta_k - \mu)(\eta_j - \mu) \right) + N(E_k)\eta_k - \nu(E_k \cap S) \exp(\eta_k) \right\} \right]. \end{aligned} \tag{2.14}$$

by means of a Markow chain Monte Carlo (MCMC) algorithm. The details of the algorithm are given in section 3.

3. Simulation of the posterior

Our MCMC algorithm is adopted from Green (1995, sec. 5); motivation behind some of the choices made below is also discussed there more thoroughly. The move types considered here are:

1. change of intensity level on one tile;
2. birth of a new tile (generating point); and
3. death of an existing tile (generating point);

with proposal probabilities h_K, b_K, and d_K, respectively, depending on the current number of tiles. We take

$$b_K = \begin{cases} c & \text{if } K \leqslant \lambda_\xi \nu(E) - 1, \\ c\dfrac{\lambda_\xi \nu(E)}{K+1} & \text{if } K > \lambda_\xi \nu(E) - 1, \end{cases} \tag{3.1}$$

$$d_K = \begin{cases} 0 & \text{if } K = 1, \\ c\dfrac{K}{\lambda_\xi \nu(E)} & \text{if } 1 < K \leqslant \lambda_\xi \nu(E), \\ c & \text{if } K > \lambda_\xi \nu(E), \end{cases} \tag{3.2}$$

and

$$h_K = 1 - b_K - d_K. \tag{3.3}$$

The constant $c \in (0, \frac{1}{2})$ controls the rate at which changes are proposed to the number of generating points: For $K > 1$, we have $b_K + d_K \in [c, 2c]$. Introducing these three move types in detail below, we always denote by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ the current (arbitrarily ordered) sequences of generating points and their marks, and by $\boldsymbol{\xi}' = (\xi'_1, \ldots, \xi'_{K'})$ and $\boldsymbol{\eta}' = (\eta'_1, \ldots, \eta'_{K'})$ the proposed ones.

In a type 1 move an index k is sampled from the uniform distribution on $\{1, \ldots, K\}$, and a proposal η'_k for a new log-level is drawn from the uniform distribution $[\eta_k - \delta, \eta_k + \delta)$, where η_k is the current value, and δ is a given sampler parameter. Since the proposal kernel is symmetric, the acceptance probability is simply $\min\{1, p(\boldsymbol{\xi}, \boldsymbol{\eta}'|\mathbf{x})/p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{x})\}$, and the posterior ratio turns out to be

$$\frac{p(\boldsymbol{\xi}, \boldsymbol{\eta}'|\mathbf{x})}{p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{x})} = \exp\left[-\frac{\eta'_k - \eta_k}{\sigma^2}\left\{ \Gamma_{kk}\left(\frac{\eta'_k + \eta_k}{2} - \mu\right) + \sum_{j: j \sim k} \Gamma_{kj}(\eta_j - \mu)\right\}\right.$$
$$\left. + (\eta'_k - \eta_k)N(E_k) - (\exp(\eta'_k) - \exp(\eta_k))\nu(E_k \cap S)\right]. \tag{3.4}$$

Moves of type 2 and 3 are designed to form pairs of reversible jumps. Considering first a birth move, a proposal ξ' for the location of a new generating point is drawn from the uniform distribution on E. Let

$$\xi'_k = \begin{cases} \xi_k & \text{if } k \leqslant K \\ \xi' & \text{if } k = K'(= K+1); \end{cases}$$

this ordering will naturally be applied to the proposed partition $\mathscr{E}'$, to the mark sequence $\boldsymbol{\eta}'$, and to the corresponding $\boldsymbol{\Gamma}$-matrix $\boldsymbol{\Gamma}'$ introduced in section 2.

Let $\mathscr{K}'$ denote the neighbourhood $\{k: k \sim_{\xi'} K'\}$ of the proposed new site, and let ν'_k, $k \in \mathscr{K}'$, be the areas the new tile $E'_{K'}$ conquers from its neighbours, that is,

$$\nu'_k = \nu(E_k) - \nu(E'_k) \tag{3.5}$$

yielding $\nu(E'_{K'}) = \sum_{k\in\mathscr{K}'} \nu'_k$. The log-level proposed for the new tile is then $\eta'_{K'} = \tilde{\eta}_{K'} + \varepsilon$, where $\tilde{\eta}_{K'}$ is the weighted average

$$\tilde{\eta}_{K'} = \sum_{k\in\mathscr{K}'} \frac{\nu'_k}{\nu(E'_{K'})}\eta_k, \tag{3.6}$$

and perturbation $\varepsilon \in \mathbf{R}$ is drawn from the density

$$f(\varepsilon) = C\exp(C_\varepsilon)/\{1+\exp(C_\varepsilon)\}^2, \tag{3.7}$$

where C is yet another parameter of the sampler (in addition to c and δ); these can be tuned to improve mixing. Reasons for the form of density (3.7) are symmetry and easy sampling by the inversion method: the inverse function of the cumulative probability is simply

$$F^{-1}(u) = C^{-1}\log\left(\frac{u}{1-u}\right).$$

The proposal also includes modifications

$$\eta'_k = \frac{\nu(E_k)}{\nu(E'_k)}\eta_k - \frac{\nu'_k}{\nu(E'_k)}\eta'_{K'} \tag{3.8}$$

to the neighbouring log-intensity values, whereby the integral of η remains unchanged in this type of move, that is,

$$\sum \nu(E'_k)\eta'_k = \sum \nu(E_k)\eta_k.$$

The death proposal reverses the above procedure: $\boldsymbol{\xi}'$ is $\boldsymbol{\xi}$ minus one random point. Suppose, for simpler notation, that death of ξ_K is proposed, and let $\xi'_k = \xi_k$, $k \leqslant K' = K-1$. The current partition $\mathscr{E} = (E_1, \ldots, E_K)$ is updated to $\mathscr{E}' = (E'_1, \ldots, E'_{K'})$ with

$$\nu'_k = \nu(E'_k) - \nu(E_k), \qquad k \underset{\boldsymbol{\xi}}{\sim} K, \tag{3.9}$$

and new log-intensity levels

$$\eta'_k = \frac{\nu(E_k)}{\nu(E'_k)}\eta_k + \frac{\nu'_k}{\nu(E'_k)}\eta_K \tag{3.10}$$

are proposed for $k \underset{\boldsymbol{\xi}}{\sim} K$.

Applying the terminology of Green (1995), the acceptance probabilities are of the form

$$\min\{1, \text{(posterior ratio)} \times \text{(proposal ratio)} \times \text{(Jacobian)}\}. \tag{3.11}$$

Suppose that a birth move is proposed from $(\boldsymbol{\xi}, \boldsymbol{\eta})$ to $(\boldsymbol{\xi}', \boldsymbol{\eta}')$. Then the posterior ratio is

$$\frac{p(\boldsymbol{\xi}', \boldsymbol{\eta}'|\mathbf{x})}{p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{x})} = \lambda_{\boldsymbol{\xi}}(2\pi\sigma^2)^{-1/2}\left(\frac{|\boldsymbol{\Gamma}'|}{|\boldsymbol{\Gamma}|}\right)^{1/2}\exp\left(-\frac{1}{2\sigma^2}D_{\boldsymbol{\Gamma}} + D_{\mathbf{x}}\right), \tag{3.12}$$

where

$$D_{\boldsymbol{\Gamma}} = \Gamma'_{K'K'}(\eta'_{K'} - \mu)^2 + \sum_{k\in\mathscr{K}'}\Bigg[\Gamma'_{kk}(\eta'_k-\mu)^2 - \Gamma_{kk}(\eta_k-\mu)^2 + 2\Gamma'_{K'k}(\eta'_{K'}-\mu)(\eta'_k-\mu)$$

$$+ \sum_{j\in\mathscr{K}: j\underset{\boldsymbol{\xi}}{\sim}k}\{\Gamma'_{kj}(\eta'_k-\mu)(\eta'_j-\mu) - \Gamma_{kj}(\eta_k-\mu)(\eta_j-\mu)\} + 2\sum_{j\notin\mathscr{K}: j\underset{\boldsymbol{\xi}}{\sim}k}\Gamma_{kj}(\eta'_k-\eta_k)(\eta_j-\mu)\Bigg], \tag{3.13}$$

since addition of one generating point alters the Delaunay graph only among the neighbours of the new tile (some connections can break), and

$$D_{\mathbf{x}} = N(E'_{K'})\eta'_{K'} + \sum_{k\in\mathscr{K}'}\{N(E'_k)\eta'_k - N(E_k)\eta_k\}$$

$$- \nu(E'_{K'}\cap S)\exp(\eta'_k) - \sum_{k\in\mathscr{K}'}\{\nu(E'_k\cap S)\exp(\eta'_k) - \nu(E_k\cap S)\exp(\eta_k)\}. \qquad (3.14)$$

Hence, evaluating the posterior ratio requires only local calculations except for the ratio of determinants. We use a simple local approximation to this ratio obtained by ignoring everything beyond the second order neighbourhood

$$\mathscr{K}'' = \bigcup_{k\in\mathscr{K}'}\left\{E_k \cup \bigcup_{j\sim k} E_j\right\} \subset E$$

of the new tile $E'_{K'}$, that is, by replacing the determinants $|\boldsymbol{\Gamma}|$ and $|\boldsymbol{\Gamma}'|$ by determinants of submatrices consisting of rows and columns corresponding to the indices in $\mathscr{K}''$ (and K' for $|\boldsymbol{\Gamma}'|$).

The proposal ratio corresponding to the proposal mechanism introduced above is

$$\frac{d_{K+1}/(K+1)}{b_K f(\eta'_{K'} - \tilde{\eta}_{K'})/\nu(E)} = [f(\eta'_{K'} - \tilde{\eta}_{K'})\lambda_{\boldsymbol{\xi}}]^{-1}, \qquad (3.15)$$

with the Jacobian

$$\left|\frac{\partial\boldsymbol{\eta}'}{\partial(\boldsymbol{\eta}, \varepsilon)}\right| = \prod_{k\in\mathscr{K}'}\frac{\nu(E_k)}{\nu(E'_k)}. \qquad (3.16)$$

For the corresponding death proposal (from $(\boldsymbol{\xi}', \boldsymbol{\eta}')$ to $(\boldsymbol{\xi}, \boldsymbol{\eta})$ in the current notation), the terms in the expression (3.11) of acceptance probability are simply the inverses of (3.12), (3.15) and (3.16).

4. Examples

In this section we present three situations in which we have tested our method. In the first two examples the data were simulated from Poisson processes with given intensities on the unit square $S = [0, 1] \times [0, 1]$. The most detailed report is given from the first test (section 4.1), in which a rather complicated continuous intensity surface was used. The second intensity function (section 4.2), on the other hand, was a simple piecewise constant function with only two distinct values. In our third example (section 4.3) we used a real data set previously analysed by Ogata & Katsura (1988). The simulated data sets are available from the World Wide Web address `http://www.stat.jyu.fi/~jmhe`

We did not concentrate on optimizing the samplers. Rather, we ran them for so long that there appeared to be no reason to doubt the convergence. In each run the sample size was $M = 1000$, with a burn-in period of 100,000 basic update steps before starting the sampling, after which the realizations after every 500th step were saved to form the MCMC-sample $\lambda^{(1)}, \ldots, \lambda^{(M)}$ of piecewise constant intensity functions from the posterior distribution.

The basic summary in each case is the surface of MCMC-estimates

$$\hat{\lambda}(x) = \frac{1}{M}\sum_{m=1}^{M}\lambda^{(m)}(x) \qquad (4.1)$$

of the pointwise posterior expectations $E[\lambda(x)|\mathbf{x}]$ evaluated on a square grid of 50×50 points x spanning E. This is what we call *posterior mean estimate* below. We want to emphasize, however, that the real result of the Bayesian modelling and statistical analysis is the whole

posterior, here approximated by an MCMC-sample, and not some particular smooth function of x. Such a posterior is impossible to display in a graphical form, which is why we have mainly used the functions (4.1) in our illustrations. This should be kept in mind when comparing below the Bayesian estimates to the ones obtained by kernel smoothing.

4.1. Example 1

On the top row of Fig. 2 are the perspective and contour views of the intensity function λ from which the test data in this example were simulated. To study the effect of the amount

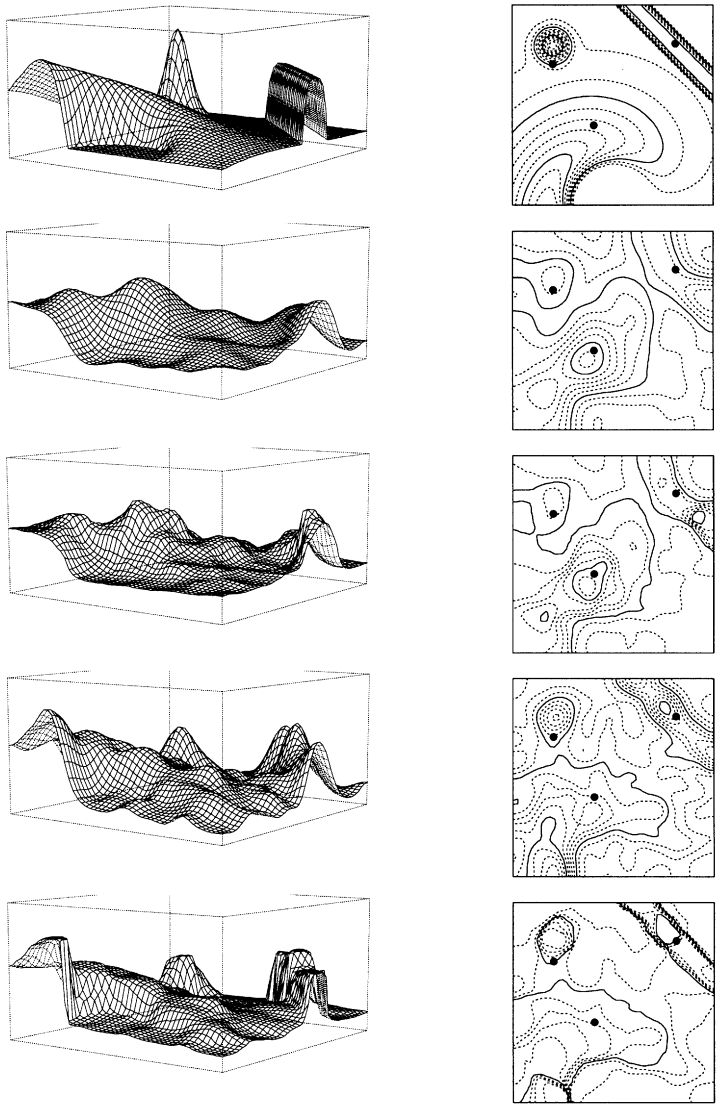

Fig. 2. Perspective plots and contour lines of the true intensity function in example 1 (top), and of the kernel estimates and our posterior mean estimates from data 1 (rows 2 and 3) and from data 2 (rows 4 and 5). The dots in the contour plots indicate locations of the diagnostic points referred to in Figs 5 and 6.

of data we used two different scalings, $\int \lambda = 500$ and $\int \lambda = 3000$, in producing the two point patterns (Fig. 3) to be analysed, data 1 (536 points) and data 2 (3035 points) respectively. As a benchmark, simple kernel density estimates multiplied by the number of data points were calculated for both data sets, with bandwidths 0.07 and 0.04, respectively, chosen by a few trials and errors. Reflective boundaries were applied to correct for the finite support. The kernel estimates are shown in the second and fourth rows of Fig. 2.

The hyperparameter values for our Bayesian procedure were chosen as $\lambda_{\xi} = 50$, $\mu = 7.5 \approx \log(1800)$, $\beta = 0.99$, and $\sigma^2 = 0.003$ after some experiments; the same values were used for both data sets. As discussed in section 2, the choice of μ has little effect when β is close to 1. This is illustrated by the fact that the same μ works reasonably well in both examples although the intensity levels in the latter are 6 times as large as in the former.

The posterior mean estimates from simulation 1 (with data 1) and simulation 2 (data 2) are shown on rows 3 and 5 of Fig. 2. We can see that piecewise constancy of the individual realizations (almost) disappears in the posterior mean estimates, which form rather smooth surfaces. For data 1 the two estimates seem almost identical, except for the somewhat smoother appearance of the kernel density estimate. Adaptivity of our method becomes apparent in estimates from data 2: on "flat" regions in the lower half of the square and further up the "valley" our method yields a smooth surface, but it also allows for sudden changes in the intensity level, for example on the steep slopes of the "ridge" in the north-east.

Table 1 makes some quantitative comparisons of our estimate from simulation 2 to kernel density estimates with various bandwidths. To measure the success in restoring the original

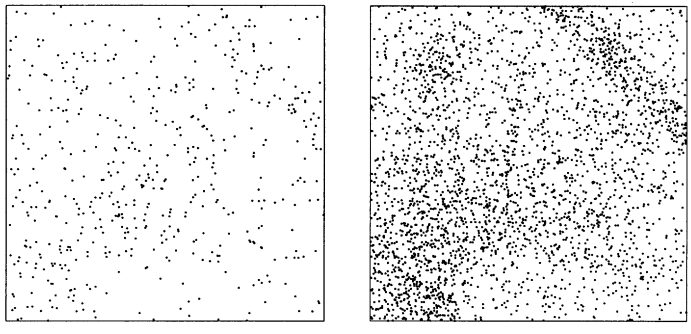

Fig. 3. The test data of example 1; pattern data 1 with 536 points (left) and data 2 with 3035 points (right).

Table 1. *Statistics comparing our posterior mean estimate and various kernel estimates from data 2 of example 1. First three rows measure the distance between the estimate and the true intensity, the bottom row the fit to the data (see text for details)*

	Our post. mean	Kernel estimates, bandwidths					
		0.02	0.03	0.035	0.04	0.05	0.06
Mean absolute error	433	652	510	485	**476**	487	522
Root mean squared error	643	834	680	**666**	672	718	781
Mean relative squared error	11.7	16.1	13.2	**12.9**	12.9	13.5	14.5
χ^2-fit to data	2228	1876	2157	2228	2279	2350	2404

intensity surface λ we calculated average absolute differences $\sum_i |\hat{\lambda}(x_i) - \lambda(x_i)|/N$, root mean squared differences $[\sum_i \{\hat{\lambda}(x_i) - \lambda(x_i)\}^2/N]^{1/2}$, and average relative squared differences $[\sum_i \{\hat{\lambda}(x_i) - \lambda(x_i)\}^2/\lambda(x_i)]/N$ between the estimated surfaces $\hat{\lambda}$ and the true one with x_i ranging over the square grid of 50×50 points. All three statistics indicate that our estimate is closer to the true intensity surface than any one of the kernel estimates considered. Goodness of fit of the estimates to the data was measured by a χ^2-statistic $\sum_i (o_i - e_i)^2/e_i$ comparing the observed point counts $o_i = N(A_i)$ in square bins A_i of size $1/50 \times 1/50$ centred at x_i to the expected ones $e_i = \hat{\lambda}(x_i)/2500$. Our estimate seems to fit the data equally well as that kernel density estimate which best reproduces the original intensity (bandwidth 0.035); naturally the fit of kernel estimates improves as the bandwidth gets smaller.

Figure 4 describes the realized tessellations in simulation 2. Three generating point patterns $\boldsymbol{\xi}^{(m)}$, $m = 250$, 500, 750 and the corresponding partitions $\mathscr{E}(\boldsymbol{\xi}^{(m)}) = \{E_k^{(m)}\}$ are shown along with an image of average tile size over all sampled realizations as a function of location. More precisely, let $E_{k(x)}^{(m)}$ be that tile of the mth realization which contains point x. Then the average tile size $\overline{\nu(E_{k(x)})}$ at point x is defined to be the average of the sizes of tiles $E_{k(x)}^{(m)}$, $m = 1, \ldots, M$, and the shade in the image is lighter for larger $\overline{\nu(E_{k(x)})}$. The flexibility of our method is well illustrated: tiles are typically small in places where the intensity seems to change rapidly. Boundaries of the ridge are clearly discernible both in the individual tessellations and in the image of average tiles.

As an example of the convergence diagnostics we present Fig. 5. It contains the plots of $\lambda^{(m)}(x)$ against m (the left-hand column of Fig. 5) from simulation 1 at three reference points

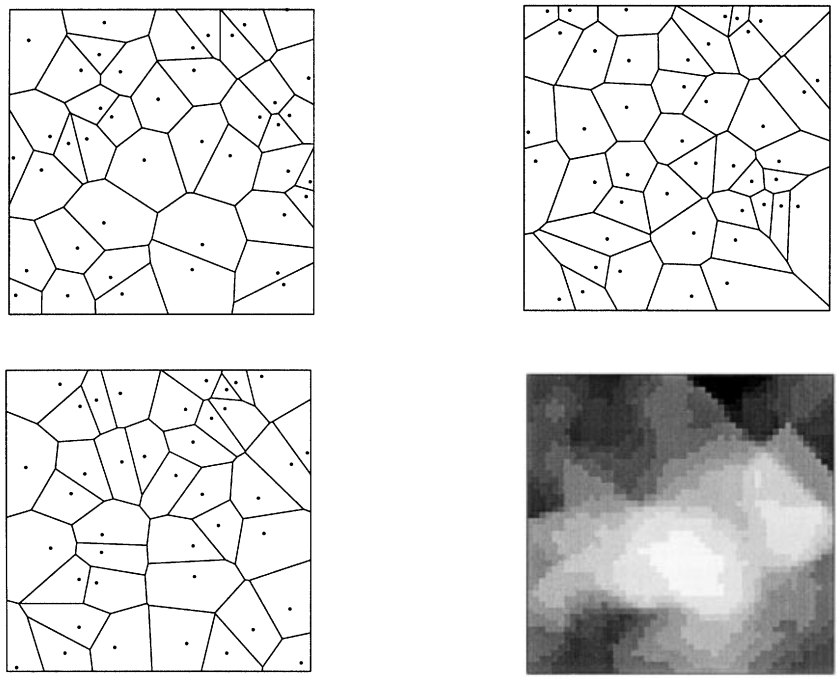

Fig. 4. Generating point patterns and corresponding Voronoi tessellations of realizations 250, 500 and 750 from the sample of simulation 2. The grey level image shows average sizes of tiles containing the reference point; the darker the smaller.

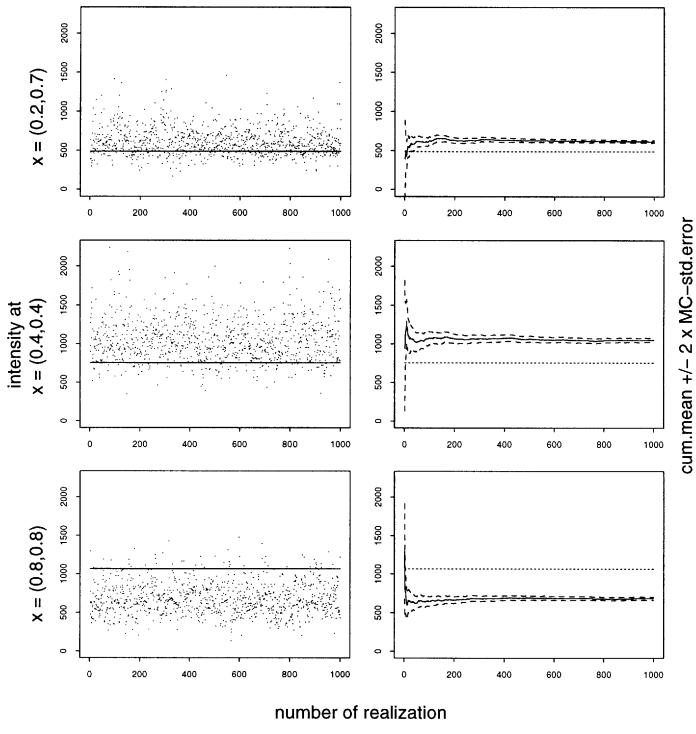

Fig. 5. Intensity values of the realizations of simulation 1 at points $x = (0.2, 0.7), (0.4, 0.4), (0.8, 0.8)$ (left hand column). Their cumulative means (solid lines) along with error bands (dashed lines) of width twice the estimated Monte Carlo standard deviation (right hand column). The straight horizontal lines show the true intensity levels.

$x = (0.2, 0.7), (0.4, 0.4), (0.8, 0.8)$ (see Fig. 2 for the locations), and the plots of corresponding cumulative means

$$\hat{\lambda}(x)_m = \frac{1}{m}\sum_{j=1}^{m} \lambda^{(j)}(x) \tag{4.2}$$

(solid lines) along with Monte Carlo error bands

$$\hat{\lambda}(x)_m \pm 2\sqrt{\sigma^2_{\mathrm{MC}}/m}, \tag{4.3}$$

where σ^2_{MC} is an initial monotone sequence estimate (Geyer, 1992) of the asymptotic Monte Carlo variance of $\sqrt{m}\hat{\lambda}(x)_m$ (the right hand column of Fig. 5). Our diagnostics do not indicate any problems with the convergence.

Various features of the posterior distributions can be studied from the MCMC-samples. As an example, Fig. 6 shows the estimates of the full marginal posterior densities of $\lambda(x)$ from the two simulations at the same reference points x as in Fig. 5. These are simple kernel density estimates applied to the samples $\{\lambda^{(1)}(x), \ldots, \lambda^{(M)}(x)\}$ from the marginal posteriors with bandwidth 150 applied for the samples from simulation 1 and 900 for simulation 2. We can see how the posterior distributions of simulation 1 are relatively more spread out as there are less data.

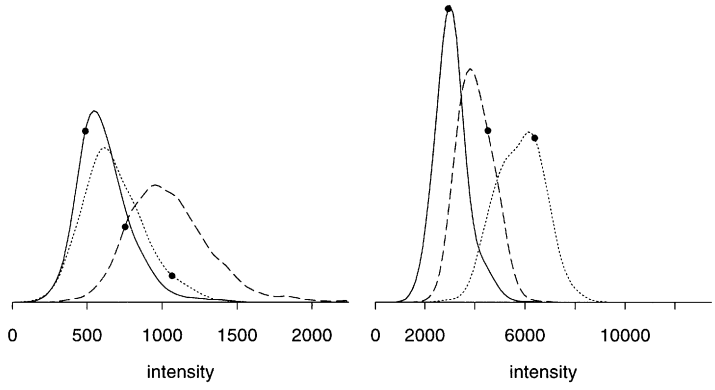

Fig. 6. Posterior density estimates for the intensity values at points $x = (0.2, 0.7)$ (solid lines), (0.4, 0.4) (dashed lines), and (0.8, 0.8) (dotted lines) from simulation 1 (left) and simulation 2 (right). The dots on the estimated density curves are horizontally located at the corresponding true intensity values.

4.2. Example 2

Here we divided the unit square to two halfs, with intensity 33 on the left hand half and 167 on the right. Figure 7 shows the simulated point pattern, and the contour lines at levels 50, 100 and 150 of the kernel density estimate with reflexive boundaries and bandwidth 0.1, and of our posterior mean estimate with prior parameters $\lambda_{\xi} = 5$, $\mu = 4.6 \approx \log(100)$, $\beta = 0.9$, and $\sigma^2 = 0.1$. Kernel estimation with constant bandwidth can not do much better than this, since the contour lines are too far apart due to too much smoothing, but on the other hand the bend in the contour line at level 150 suggests that more smoothing would be required there.

Figure 8 shows marginal posterior densities (kernel estimates with bandwidth 20 from the MCMC-sample) at points $x = (0.1, 0.1)$, (0.5, 0.5), and (0.9, 0.9). Note the bimodality of the posterior at (0.5, 0.5) lying on the edge between the two domains of constant intensity level.

4.3. Example 3

Finally, we studied a point pattern of 204 Japanese black pines in an area of 10×10 metres (Numata, 1964). Assuming a non-homogeneous Poisson process model, Ogata & Katsura (1988) presented an estimate of the intensity function and pointwise standard errors of the logarithm of the estimate as their fig. 2. We tried to produce comparable plots by

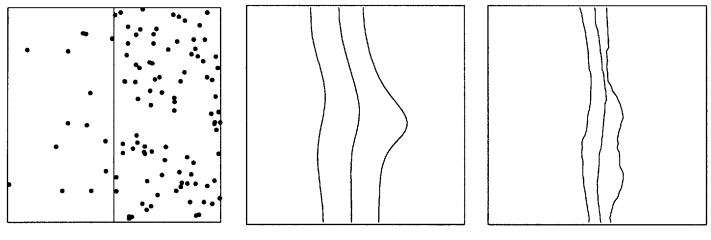

Fig. 7. Simulated data (left), kernel density estimate (middle), and our posterior mean estimate (right) from example 2. Contour lines are at levels 50, 100 and 150.

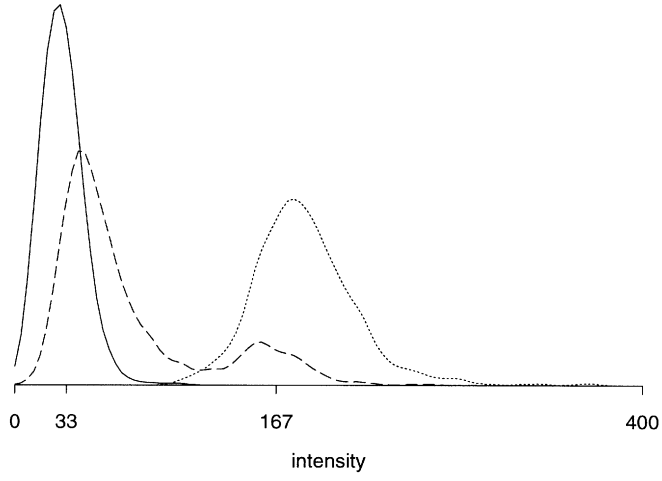

Fig. 8. Posterior density estimates of the intensity values at points $x = (0.1, 0.1)$ (solid line), (0.5, 0.5) (dashed line), and (0.9, 0.9) (broken line) in example 2.

tuning our prior parameters so that the range of estimated intensity values is about the same as in Ogata & Katsura (1988). The posterior mean estimate with $\lambda_\xi = 20$, $\mu = 5 \approx \log(150)$, $\beta = 0.99$, and $\sigma^2 = 0.02$ is shown as the top row of Fig. 9, and the bottom row shows the surface of standard deviations in the samples $\{\lambda^{(k)}(x)\}$ of intensity values

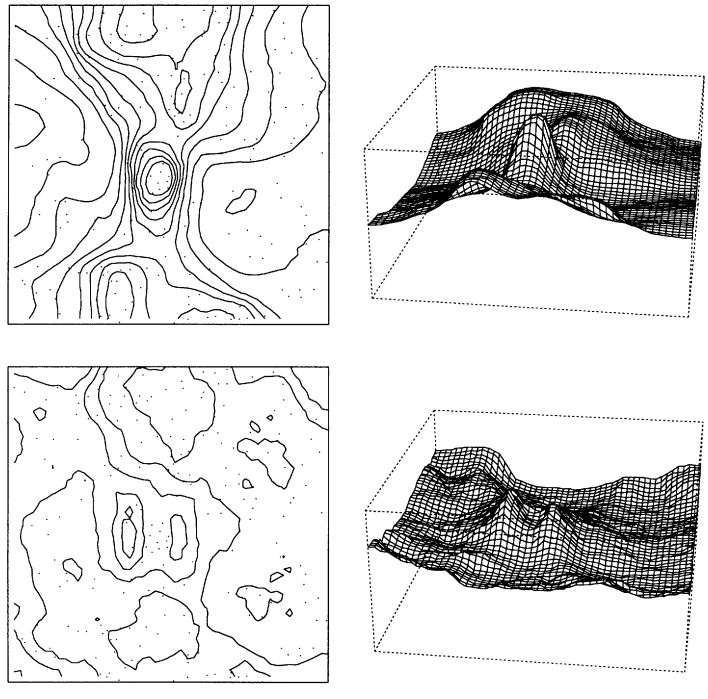

Fig. 9. Contour lines and perspective plots of the posterior mean estimate (top row, contour levels 1.2, 1.4, . . ., 3.2) and pointwise posterior standard deviations (bottom row, contour levels 0.25, 0.3, . . ., 0.45) in example 3. The data points are plotted on the contour plots.

from the posterior. Although the ranges of intensity estimates are almost equal, the surfaces of Ogata & Katsura (1988) are much smoother. Our standard errors are of the same order, but somewhat larger on average, than those of Ogata & Katsura.

5. Discussion

We have introduced a Bayesian method for non-parametric estimation of a spatial intensity. The underlying model is built from simple elements, and the prior distribution is easy to understand and to quantify. There is also built-in adaptivity, which was shown to work in practice in our examples.

Prediction outside the sample window S can be directly implemented to our method. Suppose, for example, that we are asked to predict the number of points within domain A, which is not included in S. Then we simply choose E so that it contains both S and A, and approximate the probability

$$\Pr(N(A) = N|\mathbf{x}) = \int \Pr(N(A) = N|\lambda) \Pr(d\lambda|\mathbf{x}) = \int \{\exp(-\Lambda)\Lambda^N/N!\} \Pr(d\lambda|\mathbf{x}), \quad (5.1)$$

where $\Lambda = \int_A \lambda(x)\,dx$, by the average of $\exp(-\Lambda^{(m)})(\Lambda^{(m)})^N/N!$ over the Monte Carlo sample from the posterior. Here the correlation between intensity levels on neighbouring domains, implied by a positive value of β, is especially important.

It is not essential that the surface to be estimated is an intensity function. In particular, we can use the obvious and well-known connection between a Poisson process and independent random sampling to estimate bivariate densities: considering a bounded sampling window $S \subset \mathbf{R}^2$, and conditionally on having observed N points from a Poisson process with intensity function λ to be inside this window, we can view them as a simple random sample from a distribution having density

$$f_S(x) = \left\{ \int_S \lambda(y)\nu(dy) \right\}^{-1} \lambda(x), \quad x \in S.$$

Having produced an MCMC sample $\{\lambda^{(m)}: 1 \leqslant m \leqslant M\}$ of intensity functions from the posterior, all we have to do is to normalize these estimates, dividing each $\lambda^{(m)}$ by the corresponding integral $\int_S \lambda^{(m)}(y)\nu(dy)$, thereby arriving at a sample $\{f^{(m)}\}$ of probability densities.

More generally, our method can be applied to other surface estimation problems such as regression analysis or image restoration. It is tempting to view it (along with the one-dimensional version described in Arjas & Heikkinen, 1997) as just one module in a Bayesian "inference-machine" involving larger models with various parameter curves and surfaces. Such work is currently being pursued by the authors in the context of Poisson point processes with concomitant variables.

Our final remark concerns the link our estimation problem has to more classical inference from a Cox process. The fact that, in Bayesian inference, the intensity function is considered as a random process with respect to the prior means that the data points can be viewed as coming from a doubly stochastic Poisson process. The difference between the two problem formulations is that our estimation is concerned with finding out about the "true" intensity function, whereas classical inference from a Cox process would typically be more focused on the estimation of the underlying structural parameters governing the random intensity. A natural way to deal with such a problem in the present Bayesian context would be to add one more layer of parameters to the hierarchical model, interpreting the present hyperparameters as random variables drawn from an underlying prior, and then estimating them jointly with the intensity function.

Acknowledgements

We are most grateful to Yosihiko Ogata and Masaharu Tanemura for providing us the data for example 3. Several useful discussions with Jesper Møller are also gratefully acknowledged, as well as the helpful comments of an associate editor and two referees. This work was supported by a research grant from the Academy of Finland. Facilities were provided by the Department of Statistics, University of Jyväskylä, where JH worked during this study. Rolf Turner's ratfor-routines from StatLib-archive (`http://lib.stat.cmu.edu/general/delaunay`) were modified to perform Voronoi tessellations. Also from StatLib we found Guy Nason's S-function `kde` to do the two-dimensional kernel density estimation.

References

Arjas, E. (1996). Discussion of paper by Hartigan. In *Bayesian statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith) 221–222. Oxford University Press, Oxford.

Arjas, E. & Andreev, A. (1996). A note on histogram approximation in Bayesian density estimation. In *Bayesian statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith) 487–490. Oxford University Press, Oxford.

Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4**, 505–524.

Arjas, E. & Heikkinen, J. (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Comput. Statist.* **12**, 385–402.

Besag. J. (1989). Towards Bayesian image analysis. *J. Appl. Statist.* **16**, 395–407.

Besag, J. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **84**, 733–746.

Besag, J., Green, P. J., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10**, 3–66.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. Ser. B* **60**, 333–350.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statist. Sci.* **7**, 473–511.

Geyer, C. J. & Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* **21**, 359–373.

Geyer, C. J. & Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90**, 909–920.

Green, P. J. (1995). Reversible jump MCMC and Bayesian model determination. *Biometrika* **82**, 711–732.

Hjort, N. L. (1996). Bayesian approaches to non- and semi-parametric density estimation. In *Bayesian statistics 5* (eds M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith) 223–253. Oxford University Press, Oxford.

Møller, J., Syversveen, A. R. & Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scand. J. Statist.* **25**, 451–482.

Numata, M. (1964). Forest vegetation, particularly pine stands in the vicinity of Choshi-flora and vegetation at Choshi, Chiba Prefecture, VI. *Bull. Choshi Mar. Lab.* **6**, 27–37.

Ogata, Y. & Katsura, K. (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns. *Ann. Inst. Statist. Math.* **40**, 29–39.

Preston, C. J. (1977). Spatial birth-and-death processes. *Bull. Int. Statist. Inst.* **46**, 371–391.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 172–212.

Received July 1996, in final form June 1997

J. Heikkinen, Finnish Forest Research Institute, Unioninkatu 40 A, FIN-00170 Helsinki, Finland.