# An Algorithm for Nonparametric Bayesian Estimation of a Poisson Intensity[0]

Elja Arjas[1,3], Juha Heikkinen[2,4]

[1] Department of Mathematical Sciences, University of Oulu, Linnanmaa, FIN-90570 Oulu, Finland
[2] Department of Statistics, University of Jyväskylä, P.O. Box 35, FIN-40351 Jyväskylä, Finland

**Summary**

A new algorithm is introduced for the nonparametric Bayesian estimation of the intensity of a non-homogeneous Poisson process. The method is based on a model approximation, where the approximating intensities have the structure of a piecewise constant function. Both the number and the locations of the jump points are treated as random variables. Smoothing between nearby intensity values is applied in the spirit of Bayesian image analysis. Performance of the method is illustrated in two examples with simulated data.

**Keywords:** Hazard rate, Intensity, Markov chain Monte Carlo, Nonparametric inference

[0]This paper has been published in *Computational Statistics* (1997) **12**, 385–402. This is an identical copy of the published version except that the typo in equation (2.5) has been corrected and references updated Aug 19, 1998.

[3]Current address: Rolf Nevanlinna Institute, P.O. Box 4, FIN-00014 University of Helsinki, Finland

[4]Current address: Finnish Forest Research Institute, Unioninkatu 40 A, FIN-00170 Helsinki, Finland

# 1 Introduction

Arjas and Gasbarra (1994) introduced a new approach to the nonparametric Bayesian estimation of the intensity (or hazard rate) of a non-homogeneous Poisson process. 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 lead, however, to an infinite-dimensional parameter space. Examples of such random step functions have been plotted in Figure 2 (Section 4).

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 and 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 don't need to form a piecewise constant function since the jump times are variable, and indeed the posterior mean is typically a smooth continuous function (see Figure 1 in Secion 4). Further discussion on the topic can be found in the papers cited above and in Arjas (1996).

In Arjas and 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 main motivation of the work presented here was to modify the method of Arjas and Gasbarra (1994) so that it could also be generalized to the estimation of spatial intensities. In this modification random step functions are generated via 'center points' of regions of constant intensity rather than via the jump points; (one dimensional) Voronoi tessellations are applied. This also simplifies the structure so that corresponding to each generating point there is exactly one intensity level; in other words, step functions are specified by marked point patterns. Instead of the martingale model we use (one dimensional) Markov random fields: Conditional distributions of levels are specified given both the preceding and the following level, instead of building the prior distribution sequentially in time as in Arjas and Gasbarra (1994). This approach is also in better correspondence with the role of the prior as a smoother. Finally, we use the more general reversible jump algorithm of Green (1995) to replace the version of Gibbs sampler developed by Arjas

and Gasbarra (1994) for sampling from the variable dimensional posterior distribution.

Development and application of this approach to the estimation of spatial intensities is reported in Heikkinen and Arjas (1998). Green (1995) presented two examples where a similar approach was taken to the estimation of an intensity function on the real line and of a surface on the plane. 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, and perhaps better suited for prediction. Related concurrent work includes also Denison et al. (1998), where piecewise polynomials were used instead of our step functions.

# 2 Model

Suppose we observe a non-homogeneous Poisson process during observation time $\Delta_{\rm obs} \subset \mathbf{R}$, which may, in general, be a finite union of disjoint intervals. The likelihood of observing sequence $\mathbf{T} = (T_1, \ldots, T_N) \subset \Delta_{\rm obs}$, $T_1 < T_2 < \cdots < T_N$, of event times is

$$p(\mathbf{T}\,|\lambda) = \exp\left[-\int_{\Delta_{\rm obs}} \lambda(t)\,dt\right] \prod_{n=1}^{N} \lambda(T_n), \tag{2.1}$$

where $\lambda : \mathbf{R} \to [\mathbf{0}, \infty)$ is the *intensity function* of the process. Given data $\mathbf{T}$ we consider the inference concerning the restricion of $\lambda$ to a finite interval $\Delta_{\rm tot} = [T_{\rm min}, T_{\rm max})$ containing $\Delta_{\rm obs}$. We may choose $\Delta_{\rm tot}$ to be longer than $\Delta_{\rm obs}$ for the purpose of prediction; see Section 5.

We construct a prior distribution for $\lambda$ on the set of positive valued step functions

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k \mathbf{1}_{\Delta_k}(t), \tag{2.2}$$

where intervals $\Delta_1, \ldots, \Delta_K$ form a partition of $\Delta_{\rm tot}$, and $\lambda_1, \ldots, \lambda_K \in (0, \infty)$ are the corresponding intensity levels. Random partitions $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$ are generated by sequences $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$ of generating points $\xi_k \in \Delta_{\rm tot}$ so that the subinterval $\Delta_k$ consists of those points of $\Delta_{\rm tot}$ which are closer to $\xi_k$ than to any other point of $\boldsymbol{\xi}$; in other words $\boldsymbol{\Delta}$ is the (one dimensional) Voronoi tessellation of pattern $\boldsymbol{\xi}$. The explicit formulae for the intervals are

$$\Delta_k = \Delta_k(\boldsymbol{\xi}) = \begin{cases} \left[T_{\rm min}, \frac{1}{2}(\xi_1 + \xi_2)\right) & k = 1, \\ \left[\frac{1}{2}(\xi_{k-1} + \xi_k), \frac{1}{2}(\xi_k + \xi_{k+1})\right) & k = 2, \ldots, K-1, \\ \left[\frac{1}{2}(\xi_{K-1} + \xi_K), T_{\rm max}\right) & k = K. \end{cases} \tag{2.3}$$

An advantage of this parametrization, as opposed to defining the partition by the endpoints of the subintervals, is the one-to-one correspondence between generating points $\xi_k$ and intensity values $\lambda_k$. A step function is specified by a pattern of marked points $(\xi_k, \lambda_k)$. This makes it possible to extend the construction to more general spaces than the real line. A slight defect is that all partitions of the real line can not be represented as a Voronoi tessellation. This should not, however, have any essential effect on the practical performance of our method.

The actual prior distribution among step functions parametrized as explained above, that is, the joint prior distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ is specified through

$$p(\boldsymbol{\xi}, \boldsymbol{\lambda}) = p(\boldsymbol{\xi}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}) \tag{2.4}$$

The prior distribution of $\boldsymbol{\xi}$ is taken to be the homogeneous Poisson process on $\Delta_{\mathrm{tot}}$ with a given intensity $\lambda_{\boldsymbol{\xi}} \in (0, \infty)$, and with zero probability assigned to the empty pattern. Hence, density $p(\boldsymbol{\xi})$ is proportional to $\lambda_{\boldsymbol{\xi}}^K$ if $K > 0$ and $T_{\min} < \xi_1 < \cdots < \xi_K < T_{\max}$, and equal to 0 otherwise.

The prior of $\boldsymbol{\lambda}$ (given $\boldsymbol{\xi}$) will reflect an assumption of smoothness in the sense that the differences $|\lambda_k - \lambda_{k-1}|$ between two consecutive levels are expected to be small. A multivariate Gaussian prior

$$p(\boldsymbol{\eta}|\boldsymbol{\xi}) \propto (2\pi)^{-K/2} |\mathbf{Q}|^{1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\mu})^T \mathbf{Q} (\boldsymbol{\eta} - \boldsymbol{\mu})\} \tag{2.5}$$

is assigned to the $K$-vector $\boldsymbol{\eta}$ 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.

The expectations $\mu_k$ may be chosen according to any prior knowledge of local intensities. They could, for example, be functions of covariate values attached to the corresponding intervals $\Delta_k$. Here we will assume, however, that they have been chosen to be all equal, $\mu_k = \mu$ for all $k$, to simplify the notation.

Following the Markov random field approach we specify the covariance structure via the local characteristics $p(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi})$, where $\boldsymbol{\eta}_{-k}$ denotes the sequence $\boldsymbol{\eta}$ with $\eta_k$ removed. With the assumption of multivariate normality (2.5) and the first-order Markov property

$$p(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi}) = p(\eta_k|\eta_j, |j-k| = 1, \boldsymbol{\xi}) \tag{2.6}$$

the conditional distributions $p(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi})$ become Gaussian with expectations

$$\mathrm{E}(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi}) = \mu + \sum_{j:|j-k|=1} \beta_{kj}(\eta_j - \mu), \tag{2.7}$$

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

$$\mathrm{Var}(\eta_k|\boldsymbol{\eta}_{-k}, \boldsymbol{\xi}) = \sigma_k^2 = Q_{kk}^{-1}. \tag{2.8}$$

In specifying the joint distribution (2.5) via the local characteristics (2.7) and (2.8), that is, in choosing the parameters $\beta_{kj}$ and $\sigma_k^2$, some consistency conditions must be imposed. The symmetry of $\mathbf{Q}$ requires that

$$\beta_{kj}\sigma_j^2 = \beta_{jk}\sigma_k^2. \tag{2.9}$$

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

$$\sum_{j:|j-k|=1} \beta_{kj} < 1, \qquad \text{for all } k. \tag{2.10}$$

The role of the prior as smoother becomes now apparent as

$$\mathrm{E}(\eta_k|\boldsymbol{\eta}_{-k},\boldsymbol{\xi}) = \left(1 - \sum_{j:|j-k|=1} \beta_{kj}\right)\mu + \sum_{j:|j-k|=1} \beta_{kj}\eta_j \tag{2.11}$$

is a weighted average of the prior expectation $\mu$ and the neighbouring levels $\eta_j, |j-k|=1$.

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

$$\beta_{kj} = \frac{l_{kj}}{l_k}\beta \qquad \text{and} \qquad \sigma_k^2 = \frac{\sigma^2}{l_k} \tag{2.12}$$

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

$$l_{kj} = l_{jk} \qquad \text{and} \qquad \sum_{j:|j-k|=1} l_{kj} \le l_k. \tag{2.13}$$

The easiest choice would be to have all $l_{kj}$ equal ($=1$) and $l_k$ equal to the number of neighbours of $k$ (1 for $k \in \{1,K\}$, 2 otherwise). We wish, however, to encourage adaptivity by allowing larger jumps for shorter intervals, and hence choose $l_k$ to be the length of $\Delta_k$. The requirements (2.13) are then met by choosing $l_{kj} = \frac{1}{2}|\xi_k - \xi_j|$. In the corresponding expression (2.5) of the joint distribution we now have $\mathbf{Q} = \frac{1}{\sigma^2}\boldsymbol{\Gamma}$, where

$$\Gamma_{kj} = \begin{cases} l_k & \text{if } j = k, \\ -\beta l_{kj} & \text{if } |j-k| = 1, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

Our 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, $\mu$ gives the expected overall level of log-intensity, $\beta$ determines the weighting between $\mu$ and the neighbouring levels,

and $\sigma^2$ between the prior and the data. Consider, for a moment, a slight modification of $p(\boldsymbol{\eta}|\boldsymbol{\xi})$, where $l_1 = l_{1,2}$ and $l_K = l_{K,K-1}$. As $\beta$ approaches 1, this distribution tends to the improper pairwise difference prior (Besag 1989, Besag et al. 1995)

$$p(\boldsymbol{\eta}|\boldsymbol{\xi}) \propto (2\pi\sigma^2)^{-K/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j\sim k} l_{kj} (\eta_k - \eta_j)^2 \right\}, \tag{2.15}$$

and $\mu$ disappears. Hence, in the case where prior knowledge of the intensity level is vague, we can give $\beta$ a value close to 1, and the choice of $\mu$ is not crucial. We are then left with two hyperparameters, $\lambda_{\boldsymbol{\xi}}$ and $\sigma^2$, which control the degree of smoothing. Our experience suggests that a moderate value of $\lambda_{\boldsymbol{\xi}}$ is sufficient; see the comments on Figure 4. Also, the posterior level of $K$ appears to get higher as we take smaller values of $\sigma^2$. Naturally, there is the option of treating (some of) these parameters as random variables by building one more level of hierarchy.

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

$$p(\mathbf{T}|\boldsymbol{\xi}, \boldsymbol{\eta}) = \exp\left\{ \sum_{k=1}^{K} \ell_k(\boldsymbol{\eta}) \right\}, \tag{2.16}$$

where

$$\ell_k(\boldsymbol{\eta}) = N(\Delta_k)\eta_k - |\Delta_k \cap \Delta_{\text{obs}}| e^{\eta_k}, \tag{2.17}$$

$N(A)$ is the number of events of $\mathbf{T}$ during time $A$, and $|A|$ is the length of $A$. The inference concerning the intensity is now based on sampling from the resulting posterior distribution

$$\begin{aligned} & p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{T}) \\ & \quad \propto p(\boldsymbol{\xi}) p(\boldsymbol{\eta}|\boldsymbol{\xi}) p(\mathbf{T}|\boldsymbol{\xi}, \boldsymbol{\eta}) \\ & \quad \propto \lambda_{\boldsymbol{\xi}}^K (2\pi\sigma^2)^{-\frac{1}{2}K} |\boldsymbol{\Gamma}|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{\eta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma} (\boldsymbol{\eta} - \boldsymbol{\mu}) + \sum_k \ell_k(\boldsymbol{\eta}) \right\}. \end{aligned} \tag{2.18}$$

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

# 3 Simulation of the posterior

Our MCMC algorithm follows the ideas of Green (1995, sections 4 and 5); motivation behind some of the choices made below is also discussed there more thoroughly. The move types considered here are

1. change of level in a subinterval,

2. birth of a new generating point, and

3. death of an existing generating point,

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

$$b_K = \begin{cases} c & \text{if } K \le \lambda_{\boldsymbol{\xi}} |\Delta_{\rm tot}| - 1, \\ c\frac{\lambda_{\boldsymbol{\xi}} |\Delta_{\rm tot}|}{K+1} & \text{if } K > \lambda_{\boldsymbol{\xi}} |\Delta_{\rm tot}| - 1, \end{cases} \tag{3.1}$$

$$d_K = \begin{cases} 0 & \text{if } K = 1, \\ c\frac{K}{\lambda_{\boldsymbol{\xi}} |\Delta_{\rm tot}|} & \text{if } 1 < K \le \lambda_{\boldsymbol{\xi}} |\Delta_{\rm tot}|, \\ c & \text{if } K > \lambda_{\boldsymbol{\xi}} |\Delta_{\rm tot}|, \end{cases} \tag{3.2}$$

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: We have $b_K + d_K \in [c, 2c]$ for all K.

In a type 1 move an index $k$ is sampled from the uniform distribution on $\{1, \ldots, K\}$, and a proposal $\eta'_k$ for a new log-level is drawn from the uniform distribution on $[\eta_k - \delta, \eta_k + \delta)$, where $\eta_k$ is the current value, and $\delta$ is a given sampler parameter. Since the proposal kernel is symmetric, the acceptance probability is simply $\min\{1, p(\boldsymbol{\xi}, \boldsymbol{\eta}'|\mathbf{T})/p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{T})\}$, and the posterior ratio turns out to be

$$\frac{p(\boldsymbol{\xi}, \boldsymbol{\eta}'|\mathbf{T})}{p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{T})} = \exp\Bigg[ -\frac{\eta'_k - \eta_k}{\sigma^2} \Bigg\{ \Gamma_{kk} \left( \frac{\eta'_k + \eta_k}{2} - \mu \right) + \sum_{j:|j-k|=1} \Gamma_{kj} (\eta_j - \mu) \Bigg\} + \ell_k(\boldsymbol{\eta}') - \ell_k(\boldsymbol{\eta}) \Bigg]. \tag{3.4}$$

Moves of type 2 and 3 are designed to form pairs of reversible jumps. Considering first a birth move, suppose that there are currently $K$ generating points forming a sequence $\boldsymbol{\xi}$. A proposal $\xi'$ for the location of a new generating point is drawn from the uniform distribution on $\Delta_{\rm tot}$. Let $\boldsymbol{\xi}'$ be the sequence of order statistics of $\boldsymbol{\xi} \cup \xi'$, and suppose that $\xi'_k = \xi'$ (i.e., $\xi' \in (\xi_{k-1}, \xi_k)$).

To simplify notation in comparing $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, let us re-index the elements of the current sequence $\boldsymbol{\xi}$ as $\xi_1, \ldots, \xi_{k-1}, \xi_{k+1}, \ldots, \xi_{K'}$, where $K' = K + 1$; then $\xi'_j = \xi_j$ for $j \ne k$; similar re-indexing will naturally be applied to the partition $\boldsymbol{\Delta}$, the log-level sequence $\boldsymbol{\eta}$ and the matrix $\boldsymbol{\Gamma}$. Also, we will not make explicit the modifications near the ends of $\Delta_{\rm tot}$ every time they are needed. The general rule is: Whenever an index falls outside the range $1, \ldots, K$ (or $1, \ldots, K'$, as appropriate) in the formulae (3.5) through (3.17) below, then simply ignore the corresponding term.

Let then $s_-$ and $s_+$ be the lengths of the intervals which $\Delta'_k$ 'conquers' from its two neighbours, that is,

$$s_- = |\Delta_{k-1}| - |\Delta'_{k-1}| \qquad \text{and} \qquad s_+ = |\Delta_{k+1}| - |\Delta'_{k+1}|, \tag{3.5}$$

yielding $|\Delta'_k| = s_- + s_+$ for the new interval. The log-level proposed for the new interval is then $\eta'_k = \widetilde{\eta}_k + \varepsilon$, where $\widetilde{\eta}_k$ is the weighted average

$$\widetilde{\eta}_k = \frac{s_-}{|\Delta'_k|}\eta_{k-1} + \frac{s_+}{|\Delta'_k|}\eta_{k+1}, \tag{3.6}$$

and perturbation $\varepsilon$ is drawn from the density

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

where $C$ is yet another parameter of the sampler (in addition to $c$ and $\delta$); 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

$$\eta'_{k-1} = \frac{|\Delta_{k-1}|}{|\Delta'_{k-1}|}\eta_{k-1} - \frac{s_-}{|\Delta'_{k-1}|}\eta'_k \text{ and } \eta'_{k+1} = \frac{|\Delta_{k+1}|}{|\Delta'_{k+1}|}\eta_{k+1} - \frac{s_+}{|\Delta'_{k+1}|}\eta'_k \tag{3.8}$$

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

$$\sum |\Delta'_j|\eta'_j = \sum |\Delta_j|\eta_j. \tag{3.9}$$

The death proposal reverses the above procedure: A random generating point $\xi_k$ is omitted, the current partition $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$ is updated to

$$\boldsymbol{\Delta}' = (\Delta_1, \ldots, \Delta_{k-2}, \Delta'_{k-1}, \Delta'_{k+1}, \Delta_{k+2}, \ldots, \Delta_K)$$

with

$$s_- = |\Delta'_{k-1}| - |\Delta_{k-1}| \quad \text{and} \quad s_+ = |\Delta'_{k+1}| - |\Delta_{k+1}|, \tag{3.10}$$

and new log-intensity levels

$$\eta'_{k-1} = \frac{|\Delta_{k-1}|}{|\Delta'_{k-1}|}\eta_{k-1} + \frac{s_-}{|\Delta'_{k-1}|}\eta_k \text{ and } \eta'_{k+1} = \frac{|\Delta_{k+1}|}{|\Delta'_{k+1}|}\eta_{k+1} + \frac{s_+}{|\Delta'_{k+1}|}\eta_k \tag{3.11}$$

are proposed.

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

$$\min\{1, (\text{posterior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian})\}. \tag{3.12}$$

Suppose that a birth move is proposed from $(\boldsymbol{\xi}, \boldsymbol{\eta})$ to $(\boldsymbol{\xi}', \boldsymbol{\eta}')$, the new generating point being $\xi_k'$, and $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ being re-indexed as explained above; $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}'$ are the corresponding dependence matrices as introduced in Section 2. Then the posterior ratio is

$$\begin{aligned} &R_{\text{post}}\{(\boldsymbol{\xi}, \boldsymbol{\eta}), (\boldsymbol{\xi}', \boldsymbol{\eta}')\} = \\ &\quad = \frac{p(\boldsymbol{\xi}', \boldsymbol{\eta}'|\mathbf{T})}{p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{T})} = \lambda_{\boldsymbol{\xi}} (2\pi\sigma^2)^{-\frac{1}{2}} \left(\frac{|\boldsymbol{\Gamma}'|}{|\boldsymbol{\Gamma}|}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} D_{\boldsymbol{\Gamma}} + D_{\mathbf{T}}\right], \end{aligned} \tag{3.13}$$

where

$$D_{\boldsymbol{\Gamma}} = (\boldsymbol{\eta}' - \boldsymbol{\mu})^T \boldsymbol{\Gamma}' (\boldsymbol{\eta}' - \boldsymbol{\mu}) - (\boldsymbol{\eta} - \boldsymbol{\mu})^T \boldsymbol{\Gamma} (\boldsymbol{\eta} - \boldsymbol{\mu}) \tag{3.14}$$

and

$$D_{\mathbf{T}} = \ell_k(\boldsymbol{\eta}') + \ell_{k-1}(\boldsymbol{\eta}') + \ell_{k+1}(\boldsymbol{\eta}') - \ell_{k-1}(\boldsymbol{\eta}) - \ell_{k+1}(\boldsymbol{\eta}). \tag{3.15}$$

Most terms in the two quadratic forms apperaring in equation (3.14) are equal, and therefore actually cancel across in the differencing.

The proposal ratio corresponding to the proposal mechanism introduced above is

$$R_{\text{prop}}\{(\boldsymbol{\xi}, \boldsymbol{\eta}), (\boldsymbol{\xi}', \boldsymbol{\eta}')\} = \frac{d_{K+1}/(K+1)}{b_K f(\eta_k' - \widetilde{\eta}_k)/|\Delta_{\text{tot}}|} = [f(\eta_k' - \widetilde{\eta}_k)\lambda_{\boldsymbol{\xi}}]^{-1}, \tag{3.16}$$

with the Jacobian

$$J\{(\boldsymbol{\xi}, \boldsymbol{\eta}), (\boldsymbol{\xi}', \boldsymbol{\eta}')\} = \left|\frac{\partial \boldsymbol{\eta}'}{\partial(\boldsymbol{\eta}, \varepsilon)}\right| = \frac{|\Delta_{k-1}|}{|\Delta_{k-1}'|} \frac{|\Delta_{k+1}|}{|\Delta_{k+1}'|}. \tag{3.17}$$

For the death proposal $(\boldsymbol{\xi}', \boldsymbol{\eta}')$ generated from $(\boldsymbol{\xi}, \boldsymbol{\eta})$ by removing $\xi_k$ the terms in the expression (3.12) of acceptance probability are

$$\begin{aligned} \text{posterior ratio} &= [R_{\text{post}}\{(\boldsymbol{\xi}', \boldsymbol{\eta}'), (\boldsymbol{\xi}, \boldsymbol{\eta})\}]^{-1}, \\ \text{proposal ratio} &= [R_{\text{prop}}\{(\boldsymbol{\xi}', \boldsymbol{\eta}'), (\boldsymbol{\xi}, \boldsymbol{\eta})\}]^{-1}, \text{ and} \\ \text{Jacobian} &= [J\{(\boldsymbol{\xi}', \boldsymbol{\eta}'), (\boldsymbol{\xi}, \boldsymbol{\eta})\}]^{-1}. \end{aligned}$$

## 3.1 Algorithm

Let us now summarize the contents of this Section in the form of a simulation algorithm. Further suggestions for practical implementation of step 1 are given afterwards.

1. Choose an initial configuration $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\eta}^{(0)}$, the sampler parameters $c$, $\delta$, $C$, and the number of iterations $I$.

2. Iterate steps 3 and 4 for $i = 1, \ldots, I$.

3. Let $\boldsymbol{\xi} = \boldsymbol{\xi}^{(i-1)}$, $\boldsymbol{\eta} = \boldsymbol{\eta}^{(i-1)}$, and let $K$ be the number of points in $\boldsymbol{\xi}$. Draw a uniform random number $u \in [0, 1]$, and proceed accordingly:

   - If $u \leq h_K$ then
     (a) Choose a random index $k \in \{1, \ldots, K\}$.
     (b) Let $\boldsymbol{\xi}' = \boldsymbol{\xi}$. Draw a proposal $\eta_k'$ from the uniform distribution on $[\eta_k - \delta, \eta_k + \delta)$, and let $\eta_j' = \eta_j$ for $j \neq k$.
     (c) Let $R = p(\boldsymbol{\xi}', \boldsymbol{\eta}'|\mathbf{T})/p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathbf{T})$ as given by equation (3.4).
   - Else if $u \leq (h_K + b_K)$ then
     (a) Draw a new generating point $\xi'$ from the uniform distribution on $\Delta_{\text{tot}}$.
     (b) Create $\boldsymbol{\xi}'$, and re-index $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ as explained on page 7 (following equation (3.4) ($\xi'$ becomes $\xi_k'$).
     (c) Create $\boldsymbol{\eta}'$ as explained between equations (3.5) and (3.9) ($\eta_j' = \eta_j$ for $j \notin \{k-1, k, k+1\}$).
     (d) Let

$$R = R_{\text{post}}\{(\boldsymbol{\xi}, \boldsymbol{\eta}), (\boldsymbol{\xi}', \boldsymbol{\eta}')\} R_{\text{prop}}\{(\boldsymbol{\xi}, \boldsymbol{\eta}), (\boldsymbol{\xi}', \boldsymbol{\eta}')\} \\ J\{(\boldsymbol{\xi}, \boldsymbol{\eta}), (\boldsymbol{\xi}', \boldsymbol{\eta}')\},$$

     the terms being given by equations (3.13), (3.16), and (3.17), respectively.
   - Else
     (a) Choose a random index $k \in \{1, \ldots, K\}$.
     (b) Let $\boldsymbol{\xi}' = (\xi_j)_{j \neq k}$.
     (c) Create $\boldsymbol{\eta}'$ as explained between equations (3.9) and (3.11) ($\eta_j' = \eta_j$, for $j \notin \{k-1, k+1\}$).
     (d) Let

$$R = [R_{\text{post}}\{(\boldsymbol{\xi}', \boldsymbol{\eta}'), (\boldsymbol{\xi}, \boldsymbol{\eta})\} R_{\text{prop}}\{(\boldsymbol{\xi}', \boldsymbol{\eta}'), (\boldsymbol{\xi}, \boldsymbol{\eta})\} \\ J\{(\boldsymbol{\xi}', \boldsymbol{\eta}'), (\boldsymbol{\xi}, \boldsymbol{\eta})\}]^{-1},$$

     the terms being given by equations (3.13), (3.16), and (3.17), respectively.

4. - If $R \geq 1$, then let $\boldsymbol{\xi}^{(i)} = \boldsymbol{\xi}'$ and $\boldsymbol{\eta}^{(i)} = \boldsymbol{\eta}'$.
   - Else draw a uniform random number $u \in [0, 1]$. If $R \geq u$, let $\boldsymbol{\xi}^{(i)} = \boldsymbol{\xi}'$ and $\boldsymbol{\eta}^{(i)} = \boldsymbol{\eta}'$, otherwise let $\boldsymbol{\xi}^{(i)} = \boldsymbol{\xi}$ and $\boldsymbol{\eta}^{(i)} = \boldsymbol{\eta}$.

The choice of the initial configuration is not crucial, although it may save some computation time to choose one that should be a 'likely' realization of the posterior. In our simulations we have simply taken $K^{(0)} = 1$ and drawn $\xi_1^{(0)}$ from the uniform distribution on $\Delta_{\mathrm{tot}}$. If $\lambda_{\boldsymbol{\xi}}\Delta_{\mathrm{tot}}$ were large, it would perhaps be more efficient to simulate $\boldsymbol{\xi}^{(0)}$ from the homogeneous Poisson($\lambda_{\boldsymbol{\xi}}$) process on $\Delta_{\mathrm{tot}}$. For the $\eta_k^{(0)}$ logical choices are $\log(N(\Delta_k^{(0)})/|\Delta_k^{(0)}|)$.

Selection of a sufficient number of iterations for an MCMC sampler is always a difficult issue, and it is usually done by monitoring the evolution of important summary statistics in a few pilot runs. Some convergence diagnostics can be found in relevant chapters of Gilks et al. (1996) (see also our Figure 3 in Section 4, and its explanation).

Careful choice of the sampler parameters $c$, $\delta$ and $C$ is essential for fast convergence. If $\delta$ is small and $C$ large, only small changes are proposed to the intensity levels, and it takes a long time to explore the parameter space thoroughly enough. On the other hand, if $\delta$ is too large ($C$ too small), then the advantage of using MCMC is lost: Proposals become almost arbitrary, and consequently the acceptance probabilities small. The proportions of accepted moves among proposed ones are simple and useful diagnostics for adjusting the sampler parameters. It is often suggested that proportions around 0.5 (or maybe even a bit smaller) should indicate a reasonably well mixing sampler.

# 4 Examples

Figure 1 gives condensed summaries from two test runs, Simulation 1 and Simulation 2, on two data sets, Data 1 and Data 2. The data were simulated from Poisson processes with intensity functions $500f$ for Data 1 and $100f$ for Data 2 (the dotted lines in Figure 1), where $f$ is a density function on $[0, 1]$ constructed as follows. First, a mixture $f' = w_1 f_1 + w_2 f_2$ was taken from two log-normal densities, $f_1$ and $f_2$, with origins at 1/4 and 2/3, modes at $m_1 = 1/3$ and $m_2 = 5/6$, variances $2^2$ and $(1/12)^2$, and mixture weights $w_1$ and $w_2$ such that $w_1 f_1(m_1) = w_2 f_2(m_2)$. Then $f$ was set to have a constant value $f_0 = f'(m')/3$, where $m'$ is the mode of $f'$, on the interval $[0, t_0)$, where $t_0$ is the first time $f'$ reaches the level $f_0$. On the interval $[t_0, 1]$, $f$ was set equal to $f'$, and finally, $f$ was scaled to a density.

The simulated data, 505 and 96 points, are shown as jittered dot plots. The dashed lines in Figure 1 are adaptive kernel density estimates multiplied by the observed number of points. The method of Abramson (1982) was used, in which the bandwidth $h_n$ of the kernel around the data point $T_n$ is chosen as $h/\sqrt{\hat{f}(T_n)}$, where $\hat{f}$ is a 'pilot estimate' of the density. We used the ordinary kernel estimation to obtain $\hat{f}$, and the parameters $h$ and $h_0$, the bandwidth for the pilot, were chosen by trial and error as $h_0 = .03$, $h = .1$ for Data 1 and $h_0 = .08$, $h = .15$ for Data 2. Reflective boundaries were applied to correct for the finite support: The density was estimated from the

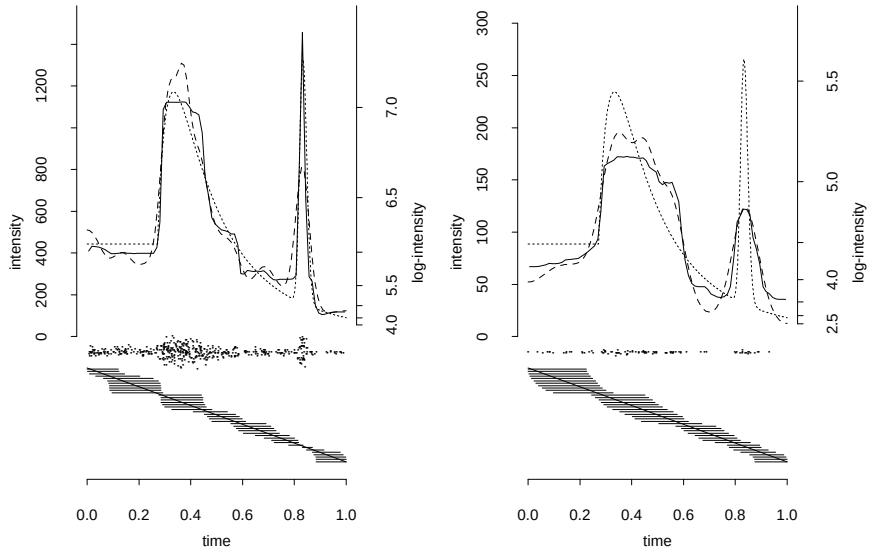


Figure 1: True intensity functions (dotted lines), data sets simulated from them (dot plots), kernel estimates (dashed lines), posterior mean estimates by our method (solid lines), and line segment diagrams illustrating the lengths of the steps in the piecewise constant intensity functions of the samples (see text for details); Data 1 and Simulation 1 on the left, Data 2 and Simulation 2 on the right

augmented data $(-\mathbf{T}, \mathbf{T}, 2-\mathbf{T})$, and the result was multiplied by 3.

The solid lines in Figure 1 show our estimates

$$\widehat{\lambda(t)} = \frac{1}{M}\sum_{m=1}^{M} \lambda^{(m)}(t) \tag{4.1}$$

of the pointwise posterior expectations $\mathrm{E}[\lambda(t)|\mathbf{T}]$, based on an MCMC-sample $\lambda^{(1)}, \ldots, \lambda^{(M)}$ from the posterior. The line segments at the bottom of Figure 1 illustrate average lengths of the subintervals $\Delta_k$. More precisely, let $\Delta^{(m)}_{k(t)}$ denote the subinterval of the $m$th realization which contains point $t$, and consider the line segment intersecting the diagonal line at point $t$ in the horizontal direction. The length of that segment to the left (right) of the diagonal is the average (over the sample) of the distances from $t$ to the left (right) ends of the intervals $\Delta^{(m)}_{k(t)}$. We can see, for example, that almost every realization of Simulation 1 has a jump point at $t = t_0 \approx 1/3$.

Figure 1 illustrates the flexibility of our method: Subintervals are short in places where the intensity seems to change rapidly. We can also see that piecewise constancy of the individual realizations (almost) disappears in the

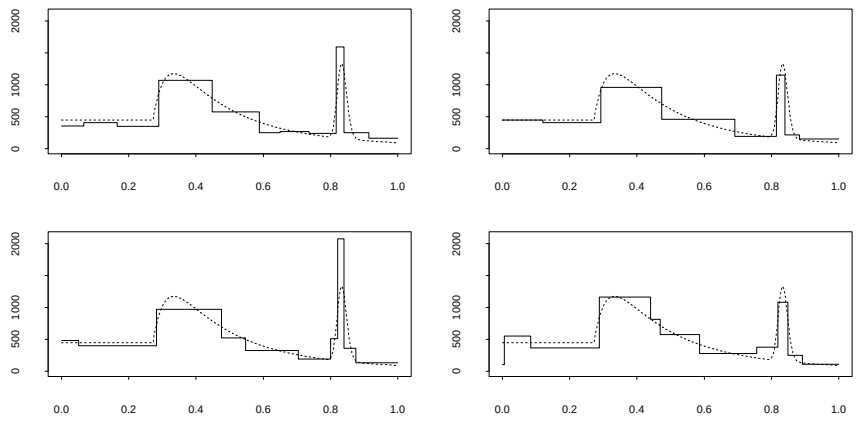


Figure 2: Realizations 250, 500, 750 and 1000 from the sample of Simulation 1; the true intensity function is shown as dotted lines

posterior mean estimates, which are rather smooth curves.

The values $\lambda_{\boldsymbol{\xi}} = 5$, $\mu = 4 \approx \log(100)$, $\beta = 0.9$ and $\sigma^2 = 0.05$ of the hyperparameters were chosen after some experiments; same values were used in both simulations. As discussed in Section 2, the choice of $\mu$ has little effect when $\beta$ is close to 1. This is illustrated by the fact that the same $\mu$ works reasonably well in both examples although the intesity levels in the latter are 5 times as large as in the former. The results appear to be rather insensitive to the choice of $\lambda_{\boldsymbol{\xi}}$ as well, and a value as low as 5 seems to offer enough flexibility in our example. Very large values of $\lambda_{\boldsymbol{\xi}}$ may result in wiggly intensity functions that are following the data too closely, although this can to some extent be counterbalanced by decreasing $\sigma^2$. Also the computing time increases with $\lambda_{\boldsymbol{\xi}}$. The parameter $\sigma^2$ has a similar role as the bandwidth in kernel smoothing, and can be tuned to control the degree of smoothing. The sampler parameters were adjusted by monitoring acceptance ratios for the different kinds of proposals; the values used in the final run were $c = 0.45$, $\delta = 1$, and $C = 5$. The sample size was $M = 1,000$, with a burn-in period of 50,000 basic update steps, and with the realizations after every 500th step saved to form the sample. This makes $I = 550,000$ basic update steps altogether, which took about 50 seconds on a Sun Ultra workstation.

Figure 2 shows four realizations from Simulation 1. As an example of the convergence diagnostics we present Figure 3 . It contains the plots of $\lambda^{(m)}(t)$ against $m$ from Simulation 2 at three reference points $t = .1, .79, 5/6$ located at the constant intensity area, just before the thin peak, and at the top of the thin peak, respectively (the left hand column of Figure 3), and the curves

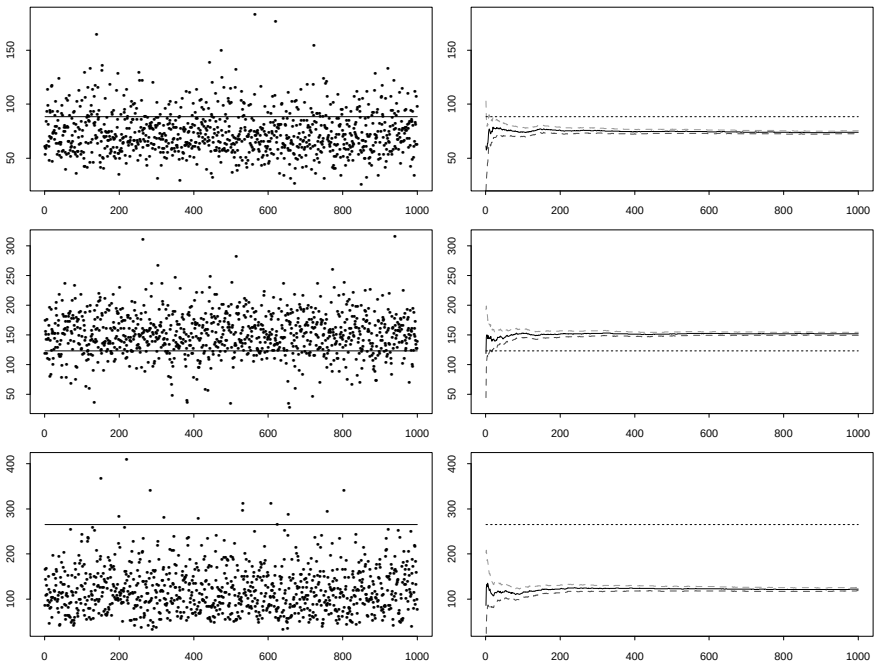


Figure 3: Intensity values of the realizations of Simulation 2 at points $t = .1, .79, 5/6$ (left hand column), and their cumulative means (solid lines) along with error bands (dashed lines) of width twice the estimated Monte Carlo standard deviation (right hand column); the horizontal lines show the true intensity values

of corresponding cumulative means

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

(solid lines) along with Monte Carlo error bands

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

where $\sigma^2_{\mathrm{MC}}$ is an initial monotone sequence estimate (Geyer 1992) of the asymptotic Monte Carlo variance of $\sqrt{m}\widehat{\lambda(t)}_m$ (the right hand column of Figure 3). Our diagnostics do not indicate any problems with the convergence. Apparently, we could have used much less iterations; the time series of Figure 3, for example, have almost no autocorrelation. It should perhaps be emphasized here that the Monte Carlo error bands reflect the variability caused by sampling based calculation. They do *not* tell us about the spread of the posterior distribution, which is illustrated in Figures 4–6.

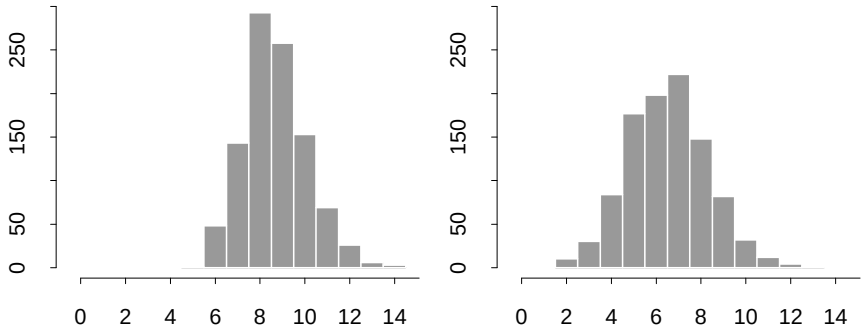


Figure 4: Distributions of $K$, the number of steps, in the samples of Simulation 1 (left) and Simulation 2 (right)

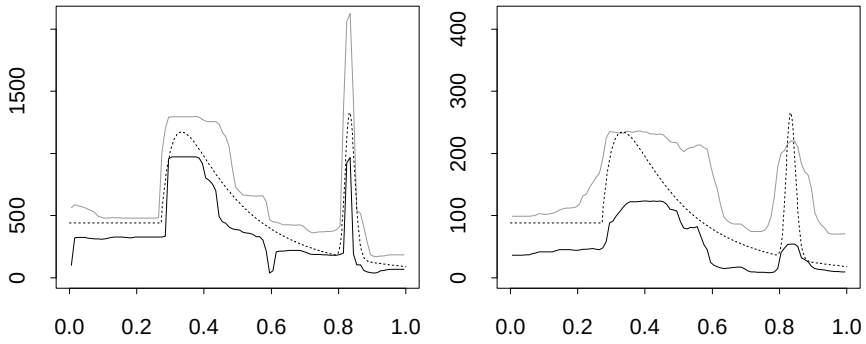


Figure 5: 5% and 95% quantiles of the pointwise distributions of intensity levels in the samples of Simulation 1 (left) and Simulation 2 (right); the true intensity functions are shown as dotted lines

Various features of the posterior distributions can be studied from the MCMC-samples. The distributions of $K$, the number of steps, are shown in Figure 4 . Note that these are shifted upwards from the prior distribution, Poisson(5), due to the complexity of the intensity function. Figure 5 shows simple interval estimates for the intensity function, pointwise 5% and 95% quantiles of the intensity levels in the samples. In Simulation 1 the true intensity lies entirely within the envelope, while in Simulation 2 the thin peak is somewhat smoothed down. In other regions, we can see how the envelope is relatively wider in Simulation 2, as there is less data. This is also illustrated in Figure 6 , where the approximations of the full marginal posterior distributions of $\lambda(t)$ from the two simulations are compared at the same reference points $t$ as in Figure 3.

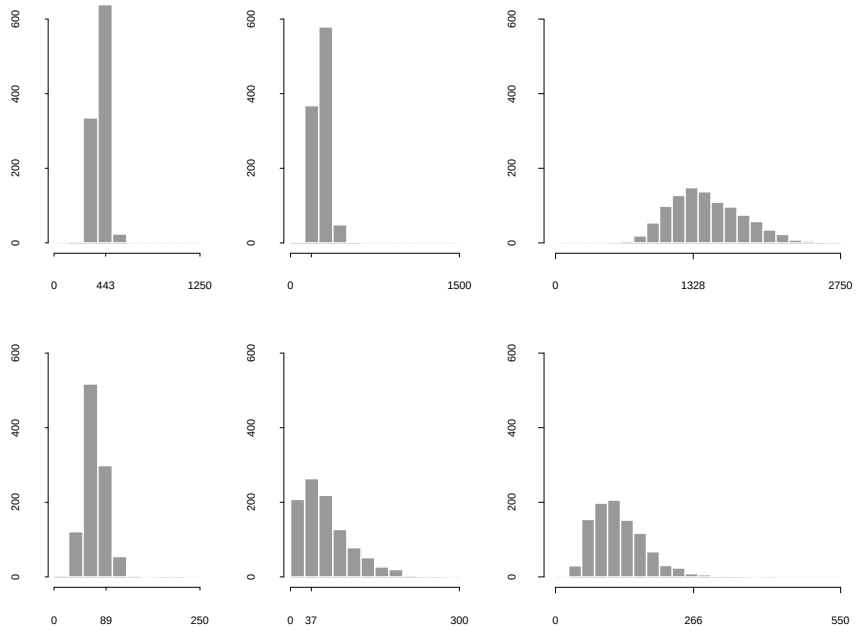


Figure 6: Distributions of the intensity values at points $t = .1, .79, 5/6$ in the samples of Simulation 1 (top) and Simulation 2 (bottom); the interior tick marks at the horizontal axes are located at the true intesity values

# 5 Discussion

Prediction beyond the observation time $\Delta_{\text{obs}}$ can also be directly implemented to our method. Suppose, for example, that we are asked to predict the number of events occurring during time $A$, which is not included in $\Delta_{\text{obs}}$. Then we simply choose $\Delta_{\text{tot}}$ so that it contains both $\Delta_{\text{obs}}$ and $A$, and approximate the probability

$$\begin{aligned} \Pr(N(A) = N|\mathbf{T}) &= \int \Pr(N(A) = N|\lambda)\ \Pr(d\lambda|\mathbf{T}) \\ &= \int \exp(-\Lambda)\Lambda^N/N!\ \Pr(d\lambda|\mathbf{T}), \end{aligned} \tag{5.1}$$

where $\Lambda = \int_A \lambda(t)\,dt$, 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 consecutive intervals, implied by a positive value of $\beta$, is especially important.

There is an obvious connection between the estimation of an intensity from Poisson data and the estimation of a density function on the real line

from a simple random sample. As is well known, given that there are $N$ data points on an interval $[0, T)$ from a Poisson process with intensity $\lambda$, their joint distribution is the same as that of the order statistics of a sample of size $N$ from the density on $[0, T)$ which is proportional to $\lambda$. Therefore, an intensity generated by our MCMC algorithm can always be immediately converted into a corresponding density function supported by $[0, T)$, thus providing a Bayesian method of estimating a density with a compact support.

Finally, it must be emphasized that the way in which the prior for $\lambda$ was specified here need not be the best, or even adequate, in some problems. For example, we could postulate that the means $\mu_k$ are not all equal, but that they form an increasing or a decreasing sequence, or first decreasing and then increasing ('bath-tub shape') as is commonly done in reliability problems.

## Acknowledgements

This work was supported by a research grant from the Academy of Finland. We are grateful to an associate editor and a referee for helpful comments.

## References

Abramson, I. S. (1982), 'On bandwidth variation in kernel estimates - a square root law', *The Annals of Statistics* **10**, 1217–1223.

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

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

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

Besag, J. (1989), 'Towards Bayesian image analysis', *Journal of Applied Statistics* **16**, 395–407.

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

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

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998), 'Automatic Bayesian curve fitting', *Journal of the Royal Statistical Society, Series B* **60**, 333–350.

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

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds (1996), *Practical Markov Chain Monte Carlo*, Chapman and Hall, London.

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

Heikkinen, J. and Arjas, E. (1998), 'Non-parametric Bayesian estimation of a spatial Poisson intensity', *Scandinavian Journal of Statistics* **25**. To appear.