Bayesian Binary Segmentation Procedure for a Poisson Process with Multiple Changepoints

TAE YOUNG YANG and LYNN KUO *

Abstract

We observe n events occurring in $(0, T]$ taken from a Poisson process. The intensity function of the process is assumed to be a step function with multiple changepoints. We propose a Bayesian binary segmentation procedure for locating the changepoints and the associated heights of the intensity function. We conduct a sequence of nested hypothesis tests using the Bayes factor or the BIC approximation to the Bayes factor. At each comparison in the binary segmentation steps, we only need to compare a single-changepoint model to a no-changepoint model. Therefore, this method circumvents the computational complexity we would normally face in problems with an unknown (large) number of dimensions. A simulation study and an analysis on a real data set are given to illustrate our methods.

Key Words: Bayes factor; Hypothesis testing; Schwarz information criterion.

1. INTRODUCTION

We observe n event times taken from a Poisson process. We assume the intensity function (failure rate) of the Poisson process is a piecewise constant function with an unknown rate λ_k between unknown changepoints $(c_{k-1}, c_k]$. Moreover, the number of changepoints K is also unknown. That is, we assume the failure rate of the Poisson process to be

$$\lambda(t) = \sum_{k=1}^{K+1} I_{\{c_{k-1} < t \leq c_k\}} \lambda_k; \tag{1}$$

*Tae Young Yang is Associate Professor, Department of Mathematics, Myongji University, Kyunggi, Korea. Lynn Kuo is Professor, Department of Statistics, U-120, University of Connecticut, Storrs, CT 06269-3120. Yang's research was supported by a research fund of Myongji University.

where $I_{\{E\}}$ is the indicator function of the event E, and $0 = c_0 < c_1 < \cdots < c_K < c_{K+1} = T$ is an increasing sequence of unknown changepoints, and $\lambda_1, \ldots, \lambda_{K+1}$ are unknown positive constants. The purpose of this paper is to develop a *Bayesian binary segmentation procedure* for locating the multiple changepoints $c_1, c_2, \ldots$ and for estimating the corresponding failure rates $\lambda_1, \lambda_2, \ldots$ in (1).

The observed n event times taken from this Poisson process are denoted by $D_{(0,T]} = (x_1, \ldots, x_n)$ with $0 < x_1 \leq \cdots \leq x_n < T$. A simple nonparametric estimator of $\lambda(t)$ when all the observed event times are distinct is

$$\hat{\lambda}(t) = \sum_{k=1}^{n+1} I_{\{x_{k-1} < t \leq x_k\}} \frac{1}{x_k - x_{k-1}}, \tag{2}$$

where $x_0 = 0$ and $x_{n+1} = T$. Given the data, the most changepoints we can detect is n. Therefore, we assume that K in (1) is at most n. There are essentially at most $2n+1$ parameters (n changepoints and $n+1$ constant failure rates) to be estimated for the most saturated model. The interesting question to ask is whether there exist nested sub-models of (1) with fewer changepoints that can fit the data reasonably well. The Bayes factor or the Bayesian Information Criterion (BIC) approximation to the Bayes factor are commonly used for model selection. The computation of them gets more and more complex as K increases, and is essentially infeasible for a large model with many changepoints. To circumvent the computational difficulties, we explore a binary segmentation method. We convert the model selection problem into several nested hypothesis testing problems; in each step we only need to compare a constant failure rate model with no changepoints to a model with exactly one changepoint. The computation is made easy. Our simulation runs suggest that our binary segmentation method yields satisfactory results.

Vostrikova (1981) proposes a binary segmentation procedure, and proves its consistency for locating the number of changepoints in a multi-dimensional random process. Chen and Gupta (1997) propose a binary procedure with the BIC to locate multiple variance changepoints in a sequence of independent Gaussian random variables with known common mean. Our procedure is similar to that of Vostrikova. The procedure can be described briefly as follows: at the first level, we compare the models respectively with no changepoints and a single changepoint in $(0, T]$ by using the Bayes factor. We assume the single changepoint τ is a continuous random variable with range in $(0, T]$. If the test is in favor of the no-changepoint model, we estimate the failure rate based on $D_{(0,T]}$, and stop the procedure. If not, then we locate the changepoint denoted by $\hat{\tau}$ by using numerical integrations or a sampling based approach. Then we divide the data into two parts: one denoted by $D_{(0,\hat{\tau}]}$ contains all the event times between 0 and $\hat{\tau}$ inclusive; and the other denoted by $D_{(\hat{\tau},T]}$ contains all the event times between $\hat{\tau}$ and T inclusive. Then we run two

Bayes factor tests similar to what we have done before; one is based on $D_{(0,\hat{\tau}]}$ and the other is based on $D_{(\hat{\tau},T]}$. If at anytime a test suggests that there are no changepoints in that subsegment, we immediately estimate the constant failure rate in the subsegment based on its data. If a test suggests a changepoint, we locate the changepoint and continue splitting the data according to the changepoint and continue testing for each of the new parts. We continue testing until no more changepoints are found in all of the further subdivisions. Using this procedure we need only compare the no-changepoint model to the single-changepoint model. Moreover, when we determine there is no single changepoint in a subsegment, we do not need to continue testing for the data in that subsegment. This cuts the sample size down quite significantly for locating changepoints in the remaining region. The procedure is efficient and quite easy to implement.

Study of changepoint problems is an active research area. The book edited by Carlstein, Müller, and Siegmund (1994) contains a wide range of articles on changepoint problems mostly from frequentist perspectives with applications in quality control, survival analysis, time series, and image analysis. If we restrict our attention to Bayesian statistical analysis for a single changepoint in the Poisson process, we can find Raftery and Akman (1986); Carlin, Gelfand, and Smith (1992); and Raftery (1994). There is a dearth of work in the area of multiple changepoints problems, perhaps due to their complexity. Recently, Green (1995) has proposed a reversible jump Markov chain Monte Carlo (MCMC) method for estimating the multiple changepoints and multiple failure rates for a Poisson process. Chib (1998) reformulates the changepoint models in terms of hidden Markov models and proposes an MCMC algorithm for computation. Both the reversible jump MCMC algorithm by Green and the MCMC algorithm by Chib are advanced techniques that are challenging to understand and implement. Instead, we explore the binary segmentation technique that can be implemented easily.

Other works on problems related to changepoints in the hazard rate studied from the Bayesian perspective include Achcar and Bolfarine (1989); Arjas and Gasbarra (1994); Ghosh, Joshi, and Mukhopadhyay (1998); and Ebrahimi *et al.* (1997). All of them treat single changepoint problems, except the one by Arjas and Gasbarra where multiple changepoint problems are studied.

The outline of this paper is as follows: Section 2 studies the hypothesis testing problem of no changepoints versus a single changepoint. Section 3 discusses the binary segmentation method. The simulation results and a numerical example based on the coal-mining disasters data set of Maguire, Pearson, and Wynn (1952) and Jarrett (1979) are given in Sections 4 and 5.

2. MODEL DETERMINATION BETWEEN NO CHANGEPOINTS AND SINGLE CHANGEPOINT

The likelihood function for the data $D_{(0,T]}$ with the model of $\lambda(\cdot)$ given in (1) can be written as

$$L(c_1, \ldots, c_K, \lambda_1, \ldots, \lambda_{K+1} | D_{(0,T]}) = \left(\prod_{i=1}^{n} \lambda(x_i) \right) \exp\left(-\int_0^T \lambda(s) ds \right),$$

cf. Cox and Lewis (1966), Basawa and Prakasa Rao (1980), and Crowder *et al.* (1991).

Let us consider the model M_1, the single-changepoint model. The intensity function for the model can be written as

$$\lambda(t) = \begin{cases} \lambda_1 & \text{if } 0 < t \leq \tau; \\ \lambda_2 & \text{if } t > \tau. \end{cases}$$

Therefore, the likelihood of the data can be written as

$$\mathrm{pr}(D|\lambda_1, \lambda_2, \tau) = \lambda_1^{N(\tau)} \exp(-\lambda_1 \tau) \lambda_2^{n-N(\tau)} \exp[-\lambda_2 (T - \tau)], \tag{3}$$

where $N(\tau)$ denotes the number of events in $(0, \tau]$, and the original data sequence $D_{(0,T]}$ is abbreviated by D.

On the other hand, the likelihood function for the no-changepoint model M_0 with the constant failure rate λ_0 is

$$\mathrm{pr}(D|\lambda_0) = \lambda_0^n \exp(-\lambda_0 T).$$

We will develop two tests for comparing a single-changepoint model versus a no-changepoint model. The first test is based on calculating the Bayes factor B_{10} for the single-changepoint model M_1 against the model with no changepoints M_0. The second test is based on the BIC approximation to the Bayes factor. In the following, we assume our data sequence D consists of distinct observations. That is, $x_1 < x_2 < \cdots < x_n$.

2.1 Bayes Factor

The model selection between the model M_1 and the model M_0 is determined by using the posterior odds ratio; that is, we select the model M_1 if $\mathrm{pr}(M_1|D)/\mathrm{pr}(M_0|D) > 1$. Note that

$$\begin{aligned} \mathrm{pr}(M_1|D)/\mathrm{pr}(M_0|D) &= \{\mathrm{pr}(M_1)/\mathrm{pr}(M_0)\} \times \{\mathrm{pr}(D|M_1)/\mathrm{pr}(D|M_0)\} \\ &= \text{prior odds ratio} \times B_{10}. \end{aligned}$$

Suppose that a priori, we assume the model M_0 is equally likely to the model M_1. Then we reject model M_0 if $B_{10} > 1$, where B_{10} (the Bayes factor of the model M_1 against the model M_0) is defined to be

$$B_{10} = \mathrm{pr}(D|M_1)/\mathrm{pr}(D|M_0). \tag{4}$$

Given the model M_1, we assume the prior distribution on λ_1, λ_2, and τ are independent with $\lambda_i \sim G(a_i, b_i)$ (with mean a_i/b_i) and $\tau \sim U(0,T)$. We choose the gamma distribution G for its flexibility to model our prior belief. Moreover, the computational convenience of using the gamma prior makes it more desirable. The assumption of a uniform distribution on τ can be relaxed, because we will rely on numerical methods for evaluating a one-dimensional integral (or a sampling based approach to update τ). So

$$\begin{aligned} \mathrm{pr}(D|M_1) &= \int_0^T \int_0^\infty \int_0^\infty \mathrm{pr}(D|\lambda_1,\lambda_2,\tau)\pi(\lambda_1,\lambda_2,\tau)d\lambda_1 d\lambda_2 d\tau \\ &= \sum_{j=1}^{n+1} \int_{[x_{j-1},x_j)} \frac{\Gamma(a_1+j-1)\Gamma(a_2+n-j+1)b_1^{a_1}b_2^{a_2}}{\Gamma(a_1)\Gamma(a_2)(b_1+\tau)^{a_1+j-1}(b_2+T-\tau)^{a_2+n-j+1}T}d\tau, \end{aligned} \tag{5}$$

where $x_0 = 0$ and $x_{n+1} = T$. If the prior on τ is not $U(0,T)$, we can easily change the above expression from $1/Td\tau$ to $\pi(\tau)d\tau$ where $\pi(\tau)$ is the prior on τ.

Given the model M_0, we assume $\lambda_0 \sim G(a_0, b_0)$. Then we compute

$$\begin{aligned} \mathrm{pr}(D|M_0) &= \int_0^\infty \lambda_0^n \exp(-\lambda_0 T)\pi(\lambda_0)d\lambda_0 \\ &= \frac{\Gamma(a_0+n)b_0^{a_0}}{\Gamma(a_0)(b_0+T)^{a_0+n}}. \end{aligned} \tag{6}$$

If $B_{10} < 1$, we accept model M_0, estimate λ_0 to be

$$\hat{\lambda}_0 = (a_0+n)/(b_0+T), \tag{7}$$

and stop the procedure. Otherwise, we accept model M_1 and estimate τ by using

$$\hat{\tau} = E(\tau|D) = A/B, \tag{8}$$

where

$$A = \sum_{j=1}^{n+1} \int_{[x_{j-1},x_j)} \frac{\Gamma(a_1+j-1)\Gamma(a_2+n-j+1)\tau}{(b_1+\tau)^{a_1+j-1}(b_2+T-\tau)^{a_2+n-j+1}}d\tau$$

and

$$B = \sum_{j=1}^{n+1} \int_{[x_{j-1},x_j)} \frac{\Gamma(a_1+j-1)\Gamma(a_2+n-j+1)}{(b_1+\tau)^{a_1+j-1}(b_2+T-\tau)^{a_2+n-j+1}}d\tau.$$

Note that the posterior distribution of τ given D in this situation is a mixture distribution $\sum_{j=1}^{n+1} w_j f_j(\tau)$, where

$$f_j(\tau) = \frac{1}{c_j(b_1+\tau)^{a_1+j-1}(b_2+T-\tau)^{a_2+n-j+1}} I_{\{\tau\in[x_{j-1},x_j)\}},$$

with

$$c_j = \int_{[x_{j-1},x_j)} \frac{1}{(b_1+\tau)^{a_1+j-1}(b_2+T-\tau)^{a_2+n-j+1}} d\tau$$

and

$$w_j = \frac{\Gamma(a_1+j-1)\Gamma(a_2+n-j+1)c_j}{\sum_{k=1}^{n+1}\Gamma(a_1+k-1)\Gamma(a_2+n-k+1)c_k}.$$

Therefore, an alternative method for computing $\hat{\tau}$ would be: Generate an index variable J with values in $1,\ldots,n+1$ where the probability of $J=j$ is w_j. Then generate a variable τ from the density f_J. Replicate this procedure for a large number of times and take the average of the τ values.

Instead of using the posterior mean of τ to estimate the change point, we could use the posterior mode. In fact, the posterior mode should be a more significant estimator because the posterior distribution of τ is discontinuous with jumps at the data points.

Remark: The above prior formulation excludes the case $\lambda_1 = \lambda_2$ in the M_1 model (that is, $\text{pr}(\lambda_1=\lambda_2)=0$). An alternative prior formulation that contains the model M_0 as nested within the model M_1 could be considered. Let the joint prior on (λ_1,λ_2) be, independent of the prior on τ,

$$\pi(\lambda_1,\lambda_2) = \pi_0 I_{\{\lambda_1=\lambda_2\}} f(\lambda_1|a_0,b_0) + (1-\pi_0)f(\lambda_1|a_1,b_1)f(\lambda_2|a_2,b_2)$$

with f denoting the gamma density and π_0 denoting the prior probability of $\lambda_1=\lambda_2$, that is the no-changepoint model. The posterior distribution of the $\pi_0,\lambda_1,\lambda_2$, and τ can be derived. Furthermore, the posterior odds ratio $[1-E(\pi_0|D)]/E(\pi_0|D)$ can be used for model selection.

2.2 BIC Approximation

Schwarz (1978) proposes the BIC approximation to the Bayes factor. He suggests selecting the model with the largest $\log L(\hat{\boldsymbol{\theta}}|D) - \frac{1}{2}p\log n$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE) of the unknown parameter $\boldsymbol{\theta}$ in the likelihood function L and p is the dimension of the $\boldsymbol{\theta}$.

For the model M_1, observe the likelihood function in (3) is maximized by setting $\hat{\lambda}_1 = N(\tau)/\tau$ and $\hat{\lambda}_2 = [n-N(\tau)]/(T-\tau)$ for each τ. Then we need to maximize the following profile likelihood of τ (we have dropped the constant $\exp(-n)$ from the profile likelihood expression):

$$L(\tau) = \left(\frac{N(\tau)}{\tau}\right)^{N(\tau)} \left(\frac{n - N(\tau)}{T - \tau}\right)^{n-N(\tau)}. \tag{9}$$

This likelihood function takes the value 0 for $\tau < x_1$ and $\tau \geq x_n$. Therefore, we need only examine the function for $x_1 \leq \tau < x_n$. Observe the likelihood function is a piecewise continuous function of τ with discontinuities located at the data points $x_1, \ldots, x_n$. When $x_j \leq \tau < x_{j+1}$, for each $j = 1, \ldots, n-1$, the likelihood function restricted to this region can be written as

$$L(\tau) I_{\{x_j \leq \tau < x_{j+1}\}} = \left(\frac{j}{\tau}\right)^j \left(\frac{n-j}{T-\tau}\right)^{n-j} I_{\{x_j \leq \tau < x_{j+1}\}}.$$

It is straightforward to show this function is a convex function of τ. Therefore, this function is maximized at one of its boundaries x_j or x_{j+1}^-. We use the notation x^- to denote the left-hand limit of x. To find the maximum of $L(\tau)$ for $x_1 \leq \tau < x_n$, we only need to compare the profile likelihood evaluated at $x_1, x_2^-, x_2, x_3^-, \ldots, x_n^-$ and pick $\hat{\tau}$ to be the argument that maximizes these values. We only need to compare the following values at $j = 1, \ldots, n$ excluding $L(x_1^-)$ and $L(x_n)$:

$$L(x_j^-) = \left(\frac{j-1}{x_j^-}\right)^{j-1} \left(\frac{n-j+1}{T-x_j^-}\right)^{n-j+1}$$

and

$$L(x_j) = \left(\frac{j}{x_j}\right)^j \left(\frac{n-j}{T-x_j}\right)^{n-j}.$$

Now let us consider the null hypothesis with no changepoints. The likelihood function is given by

$$L(\lambda_0) = \lambda_0^n \exp(-\lambda_0 T).$$

It is maximized by $\hat{\lambda}_0 = n/T$ with a maximum value of $(n/T)^n \exp(-n)$.

So

$$\log B_{10} \approx \log \mathrm{pr}(D|\hat{\boldsymbol{\theta}}_1, M_1) - \log \mathrm{pr}(D|\hat{\boldsymbol{\theta}}_0, M_0) - \frac{1}{2}(p_1 - p_0)\log n,$$

where $\hat{\boldsymbol{\theta}}_1 = \{\hat{\lambda}_1, \hat{\lambda}_2, \hat{\tau}\}$, $\hat{\boldsymbol{\theta}}_0 = \hat{\lambda}_0$, $p_1 = 3$, and $p_0 = 1$. The right-hand side, denoted by $\log B_{10}^{BIC}$, can be simplified to

$$\log B_{10}^{BIC} = \begin{cases} i\log(\frac{i}{x_i}) + (n-i)\log(\frac{n-i}{T-x_i}) - n\log(\frac{n}{T}) - \log n, & \text{if } x_i \text{ is the maximum;} \\ (i-1)\log(\frac{i-1}{x_i}) + (n-i+1)\log(\frac{n-i+1}{T-x_i}) - n\log(\frac{n}{T}) - \log n, & \text{if } x_i^- \text{ is the maximum.} \end{cases} \tag{10}$$

If the above expression is positive, then we would reject the no-changepoint hypothesis and estimate the changepoint at the argument that achieves the maximum.

3. THE BINARY SEGMENTATION PROCEDURE

3.1 The Procedure

The overall procedure using the Bayes factor criterion for inference for the intensity function is given as follows:

(1) Level 1 analysis:

We follow the procedure in subsection 2.1 using the whole data and region $(0, T]$ to test the model of no changepoints (M_0) versus single-changepoint (M_1). If we decide there are no changepoints, then we estimate the constant failure rate to be $\hat{\lambda}_0 = (a_0 + n)/(b_0 + T)$ as in (7) and stop. If not, we continue to the level 2 analysis.

(2) Level 2 analysis:

We estimate the changepoint $\hat{\tau}$ using (8) and divide the data into two parts, D_{21} and D_{22}, where D_{21} consists of all event times observed in $(0, \hat{\tau}]$ with sample size n_{21}; and D_{22} consists of all event times observed in $(\hat{\tau}, T]$ with sample size n_{22}. Then we apply two tests similar to the one in subsection 2.1; one is based on D_{21} and the other is based on D_{22}. For the first test, we need to replace D with D_{21}, change T to $\hat{\tau}$ and n to n_{21}. For the second test, we need to relabel the data D_{22} to $x_1, \ldots, x_{n_{22}}$. Then we apply the test with the region $(0, T]$ changed to $(\hat{\tau}, T]$, $1/T d\tau$ changed to $1/(T - \hat{\tau}) d\tau$, and n changed to n_{22}. If both of these tests select the null models, then we stop the procedure and estimate the constant failure rate $\hat{\lambda}_{21} = (a_0 + n_{21})/(b_0 + \hat{\tau})$ for the segment $(0, \hat{\tau}]$ and $\hat{\lambda}_{22} = (a_0 + n_{22})/(b_0 + T - \hat{\tau})$ for the segment $(\hat{\tau}, T]$. That is, we estimate the intensity function to be $\hat{\lambda}(t) = \hat{\lambda}_{21} I_{\{0<t\leq\hat{\tau}\}} + \hat{\lambda}_{22} I_{\{\hat{\tau}<t\leq T\}}$. If any of the tests suggests that there is a changepoint, then we would proceed to the next level.

(3) Continue testing:

We would estimate the changepoint as in (8) with appropriate changes. Then we would continue testing until no more splitting is allowed. Any time a null model of no changepoints is determined, we would estimate the constant failure rate at that region to be $\hat{\lambda} = (a_0 + \text{current sample size})/[b_0 + \text{length(current segment)}]$ and cease further testing in the subregion.

For the BIC, we follow the above procedure with the criteria given in subsection (2.2). Every time that model M_0 is selected, we estimate the corresponding failure rate to be $\hat{\lambda}$=(the current sample size)/(length of the current segment).

3.2 Heuristic Argument

Let us suppose the intensity function has changepoints at c_1, c_2, and c_3 with the constant failure rates $\lambda_1, \lambda_2, \lambda_3$, and λ_4 where any of the adjacent λ's are not equal. Let us call this the saturated (S) model. Then our level 1 test is essentially testing $H_0 : \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4$ versus $H_a : \lambda_1 = \lambda_2 = \lambda_3 \neq \lambda_4$ or $\lambda_1 = \lambda_2 \neq \lambda_3 = \lambda_4$ or $\lambda_1 \neq \lambda_2 = \lambda_3 = \lambda_4$. Our level 1 test would most likely reject the H_0 hypothesis, because the BF_{S0} would tend to be larger than the BF_{S1} and $\mathrm{BF}_{10} = \mathrm{BF}_{S0}/\mathrm{BF}_{S1}$. We use BF_{S0} to denote the Bayes factor for testing the saturated model S against the model in H_0, and BF_{S1} to denote the Bayes factor for testing the saturated model S against the model in H_a. When the single-changepoint model is detected, we would most likely be able to locate one of the changepoints well with a large sample. If c_1 is detected, then our binary segmentation test should suggest no more splits to the left, and either c_2 or c_3 is detected to the right. Then a third level test will determine the remaining changepoint. If c_2 is detected, then both level 2 tests should be able to pick up the changepoints c_1 and c_3. Similar results will follow when c_3 is detected first, where a level 3 test will be needed.

4. SIMULATION

We generate event times from a Poisson process with the rate function

$$\lambda(s) = 3I_{\{0<s\leq 20\}} + 0.5I_{\{20<s\leq 90\}} + 5I_{\{90<s\leq 100\}}. \tag{11}$$

There are two changepoints $c_1 = 20$ and $c_2 = 90$, and three corresponding rates, $\lambda_1 = 3$, $\lambda_2 = 0.5$, and $\lambda_3 = 5$. We randomly sample 145 event times in $(0, 100]$ from the Poisson process with rate function in (11). This is done by using the corresponding mixture distribution. We first generate an index variable I with $\mathrm{pr}(I = 1) = 60/145$, $\mathrm{pr}(I = 2) = 35/145$, and $\mathrm{pr}(I = 3) = 50/145$. Then we replicate this process 145 times and count the numbers of observations with values of 1, 2, or 3. The numbers are denoted by n_1, n_2, and n_3. Then we generate n_1 event times from a Poisson process with rate 3 using the IMSL RNNPP routine. Similarly, we generate n_2 observations from a Poisson process with rate 0.5 and translate the event times up by 20 units, and n_3 event times with rate 5 and moved up by 90 units. The simulated data are plotted in Figure 1 where the event label i is plotted against the event time x_i for $i = 1, \ldots, 145$.

We fit the data with Bayes factor B_{10} for the single-changepoint model against the no-changepoint model. Calculation of the Bayes factor is based on numerical evaluation of a one-dimensional integral. At each stage, we consider a relatively diffuse Gamma prior $\Gamma(0.5, 0.0000001)$ for λ_i, the heights of the intensity function, so that the analysis is more focused on the likelihood. For the complete data in $(0, 100]$, the B_{10} is 3.2×10^7. The single-changepoint model is selected and the estimated changepoint is at 88.9. We then divide the data into two parts: data in the first segment $(0, 88.9]$ and data in the second segment $(88.9, 100]$. For the data in $(0, 88.9]$, the Bayes factor B_{10} is 2.6×10^9 and the estimated changepoint is at 20.0. We divide the data further. For the data in $(0, 20]$, the Bayes factor B_{10} is 0.11 and the estimated constant intensity function is at $\hat{\lambda}_1 = 3.02$. For the data in $(20, 88.9]$, the Bayes factor B_{10} is 0.58 and the estimated constant intensity function is at $\hat{\lambda}_2 = 0.5$. For the data in the segment of $(88.9, 100]$, the Bayes factor is 0.09 and the estimated constant intensity function is at $\hat{\lambda}_3 = 4.66$.

Next, we fit the data using BIC. The procedure of splitting the data is quite similar to that using the Bayes factor criterion. We obtain the estimated intensity function with heights of $\hat{\lambda}_1 = 3.05$ in $(0, 19.6]$, $\hat{\lambda}_2 = 0.51$ in $(19.6, 89.2]$, and $\hat{\lambda}_3 = 4.79$ in $(89.2, 100]$.

The true intensity function and the estimated intensity functions using the Bayes factor and the BIC approximation are plotted in Figure 2. They are marked in order by the line with plus signs, the dotted line, and the line with asterisks. The figure shows that both methods yield similar results and both capture the true intensity function.

5. NUMERICAL EXAMPLE

A data set of time intervals between British coal-mining disasters was initially gathered by Maguire *et al.* (1952). Then it was corrected and extended for 112 years from March 15, 1851 to March 22, 1962 by Jarrett (1979). There were 191 accidents in this period of 40,550 days and the accident time was recorded in days. There was a 0 value in this data set corresponding to a second accident in the same day. We group these two accidents into one and use the data of 190 accidents. The data are plotted in Figure 3 where the number of the accident i is plotted against the time of the accident x_i for $i = 1, \ldots, 190$. We can also apply our methods to the original data of 191 accidents by moving the time of the second accident in the same day up by a half day. Similar results are obtained and are omitted here.

The Jarrett data have been used by other authors that include Raftery and Akman (1986), Carlin *et al.* (1992), and Green (1995). Raftery and Akman (1986) extend the data set to the period between January 1, 1851 and December 31, 1962. They assume a single changepoint and use the uniform prior over the extended range for the changepoint. They estimate the posterior mode of the location of the single changepoint to be at 14,313 days between the 124^{th} accident and the 125^{th} accident. Carlin *et al.* (1992) also assume a single changepoint. They group the data by discretizing them into counts per year, and apply the Poisson process to the modified data. The changepoint is taken as a discrete variable in units of years. They estimate the posterior modal year for a single changepoint to be the year of 1891 (around the 127^{th} accident) by a hierarchical Bayesian approach. Green (1995) develops a reversible jump MCMC method for changepoints and intensity functions. According to Green's Figure 2, the posterior probabilities of one change, two changes, three changes, and four changes are about 0.16, 0.35, 0.26, and 0.15, respectively. Therefore, the model with two changepoints has the highest posterior probability. In the two changepoints scenario in Green's Figure 3, the posterior mode for the first (second) changepoint is about 14,000 (35,000) days. In the one changepoint scenario, the posterior mode for the single changepoint is about 14,000 days.

We fit the data with the Bayes factor criterion. For each stage, we consider a relatively diffuse Gamma prior $\Gamma(0.5, 0.0000001)$ for the heights of the intensity function. This mimics the diffuse (improper) prior $\Gamma(0.5, \epsilon)$ with $\epsilon \to 0$ used in Raftery and Akman (1986). For the complete data in $(0, 40, 550]$, B_{10} is 2.1×10^9, and the estimated changepoint is at 14,435. We divide the original data into two segments $(0, 14, 435]$ and $(14, 435, 40, 550]$. For the data in $(0, 14, 435]$, the Bayes factor B_{10} is 0.04. We estimate the constant intensity function to be at $\hat{\lambda}_1 = 0.0086$. For the data in $(14, 435, 40, 550]$, the Bayes factor B_{10} is 0.16 and so we estimate the corresponding constant intensity function to be at $\hat{\lambda}_2 = 0.0025$.

We fit the data with the BIC approximation for the single-changepoint model against the no-changepoint model. For the complete data in $(0, 40, 550]$, $\log B_{10}^{BIC}$ is 30.99, and the estimated changepoint is at $x_{124} = 14, 241$. We further divide the complete data into two segments. For the data in $(0, 14, 241]$, $\log B_{10}^{BIC}$ is -3.41 and the estimated constant intensity function is at 0.0087. For the data in $(14, 241, 40, 550]$, $\log B_{10}^{BIC}$ is 2.44, and the estimated changepoint is at $x_{186} = 35, 242$. So we continue dividing the data in the second segment into two parts: $(14, 241, 35, 242]$ and $(35, 242, 40, 550]$. For the data in $(14, 241, 35, 242]$, $\log B_{10}^{BIC}$ is -0.62, and the estimated height of the constant intensity function is 0.0029. For the data in $(35, 242, 40, 550]$, $\log B_{10}^{BIC}$ is -0.86 and the estimated height of the constant intensity function is 0.00085.

We summarize the two procedures. According to the Bayes factor criterion, there is only one changepoint at 14,435 days with the intensity function of heights of $\hat{\lambda}_1 = 0.0086$ and $\hat{\lambda}_2 = 0.0025$. The single changepoint at 14,435 (between the 124^{th} and 125^{th} accidents) is similar to that in

Raftery and Akman (1986) and Carlin *et al.* (1992). According to the BIC approximation criterion, there are two changepoints located at 14,241 (the 124^{th} accident) and 35,242 (the 186^{th} accident) with the corresponding heights of the intensity functions of 0.0087, 0.0029, and 0.00085, respectively, from left to right. Our result from the BIC approximation criterion is quite similar to that of the Green's two changepoints scenario. In Figure 4, we plot the two estimated intensity functions, the dotted line denoting the one using the Bayes factor and the line with the asterisks denoting the one using the BIC.

References

Achcar, J. A., and Bolfarine, H. (1989), "Constant Hazard Against a Change-Point Alternative: A Bayesian Approach with Censored Data," *Communications in Statistics, Part A-Theory and Methods,* 18, 3801-3819.

Arjas, E., and Gasbarra, D. (1994), "Nonparametric Bayesian Inference from Right Censored Survival Data, Using the Gibbs Sampler," *Statistica Sinica,* 4, 505-524.

Basawa, I.V., and Prakasa Rao, B.L.S. (1980), *Statistical Inference for Stochastic Processes.* Academic Press, London.

Carlin, B. P., Gelfand A. E., and Smith, A. F. M. (1992), "Hierarchical Bayesian Analysis of Change Point Problems," *Journal of the Royal Statistical Society C,* 41, 389-405.

Carlstein, E., Müller, H. G., and Siegmund, D. (1994), *Change-point Problems.* Institute of Mathematical Statistics, Hayward.

Chen, J., and Gupta, A. K. (1997), "Testing and Locating Variance Changepoints with Application to Stock Prices," *Journal of the American Statistical Association,* 92, 739-747.

Chib, S. (1998), "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics,* 86, 221-241.

Cox, D.R., and Lewis, P.A. (1966), *Statistical Analysis of Series of Events.* Methuen, London.

Crowder, M.J., Kimber, A.C., Smith, R.L., and Sweeting T.J. (1991), *Statistical Analysis of Reliability.* Chapman and Hall, London.

Ebrahimi, N., Gelfand, A.E., Ghosh, M., and Ghosh, S. (1997) "Bayesian Analysis of Change-Point Hazard Rate Models," Technical Report TR9708, University of Connecticut.

Ghosh, J. K., Joshi, S. N., and Mukhopadhyay C. (1998), "Inference About Sharp Changes in Hazard Rate," in *Frontiers in Reliability*, eds Basu, A. P., Basu, S. K., and Mukhopadhyay, S. World Scientific, Singapore, 133-142.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika,* 82, 711-732.

Jarrett, R. G. (1979), "A Note on the Intervals Between Coal-Mining Disasters," *Biometrika,* 66, 191-193.

Maguire, E. S., Pearson, B.A., and Wynn, A. H. A. (1952), "The Time Intervals Between Industrial Accidents, " *Biometrika,* 39, 168-180.

Raftery, A. E. (1994), "Change Point and Change Curve Modeling in Stochastic Processes and Spatial Statistics," *Journal of Applied Statistical Science,* 1, 403-424.

Raftery, A. E., and Akman, V. E., (1986), "Bayesian Analysis of a Poisson Process with a Change-Point," *Biometrika,* 73, 85-89.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Vostrikova, L. J., (1981), "Detecting 'Disorder' in Multidimensional Random Processes," *Soviet Mathematics Doklady,* 24, 55-59.

Captions for Figures

Figure 1. Graph for the simulated data.

Figure 2. Comparison among the true intensity function (plus line), the estimated intensity functions from the Bayes factor criterion (dotted line), and the BIC (asterisked line) for the simulated data.

Figure 3. Graph for the data of 190 coal mining disasters.

Figure 4. Comparison between the estimated intensity functions from the Bayes factor criterion (dotted line) and the BIC (asterisked line).

Figure 1:

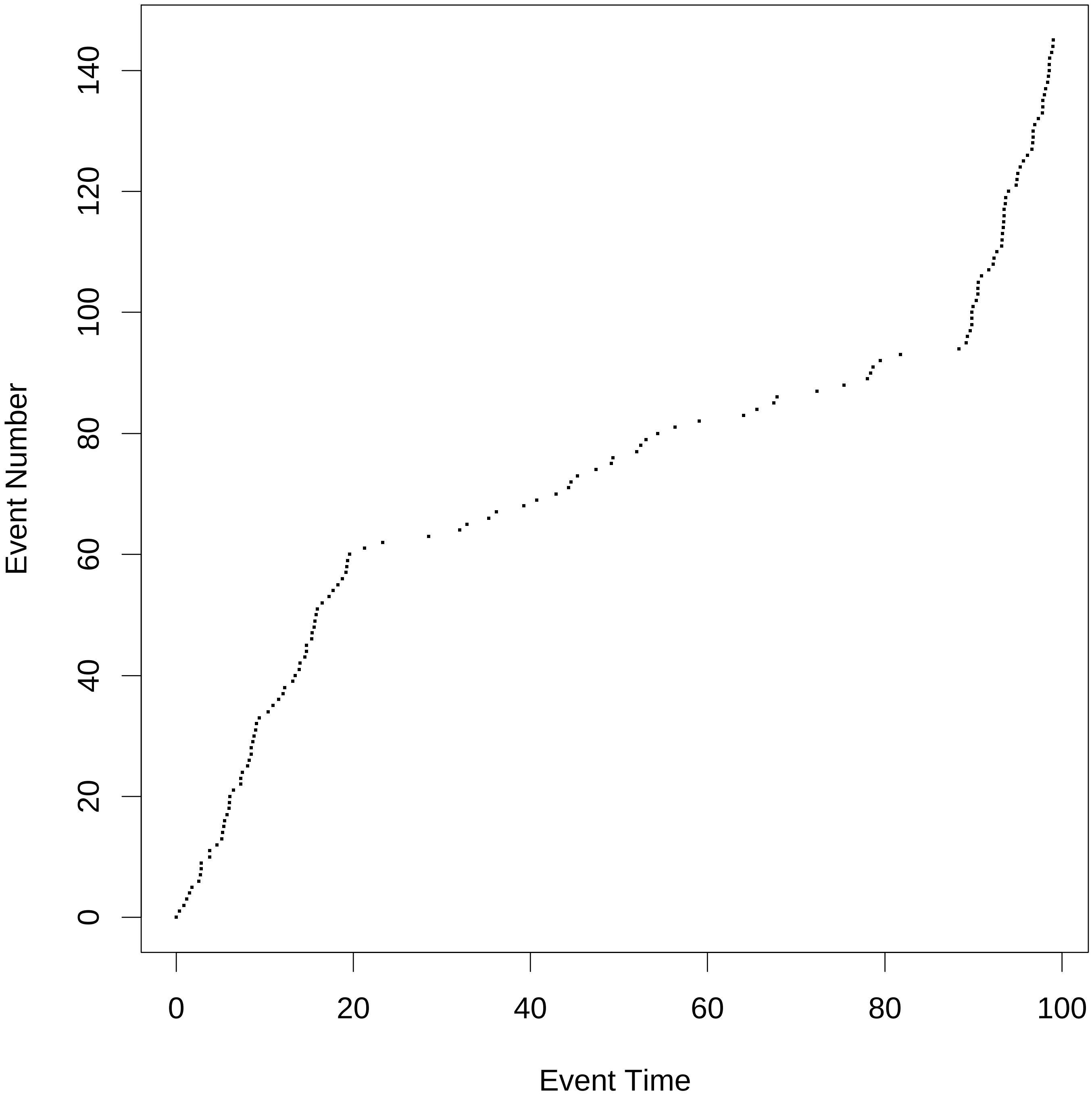

Event Number
0
20
40
60
80
100
120
140
0
20
40
60
80
100
Event Time

Figure 2:

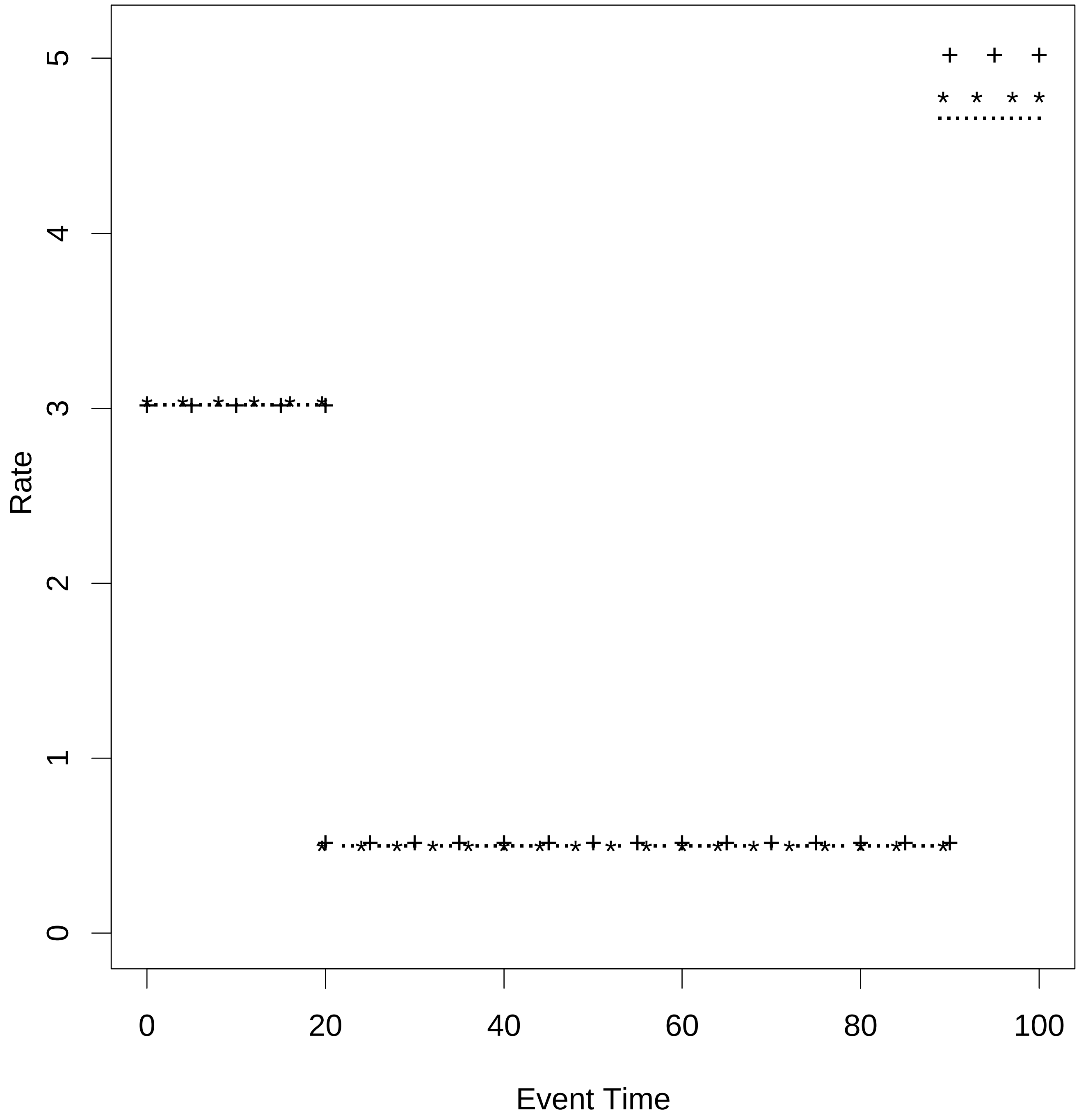


Figure 3:

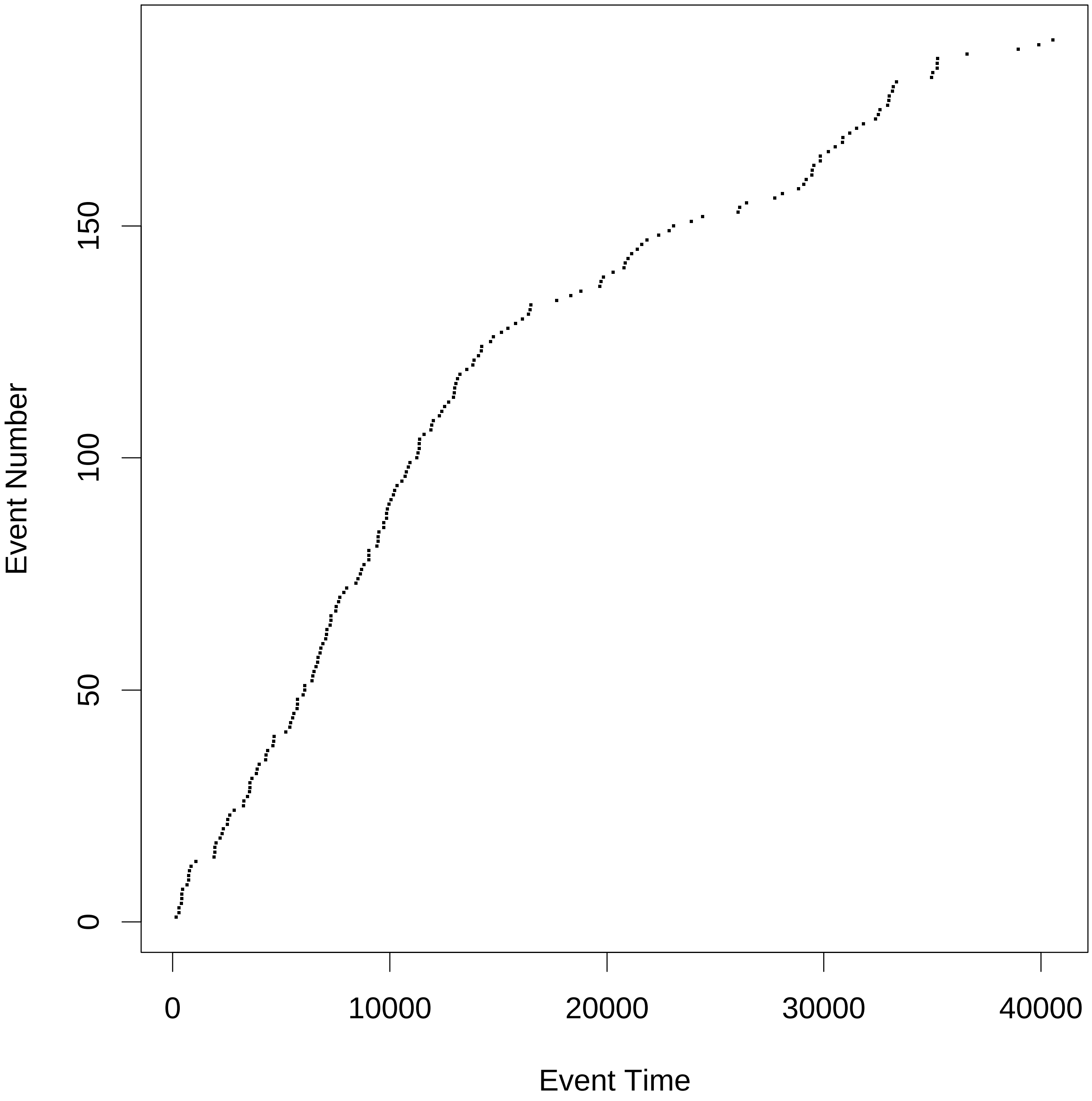
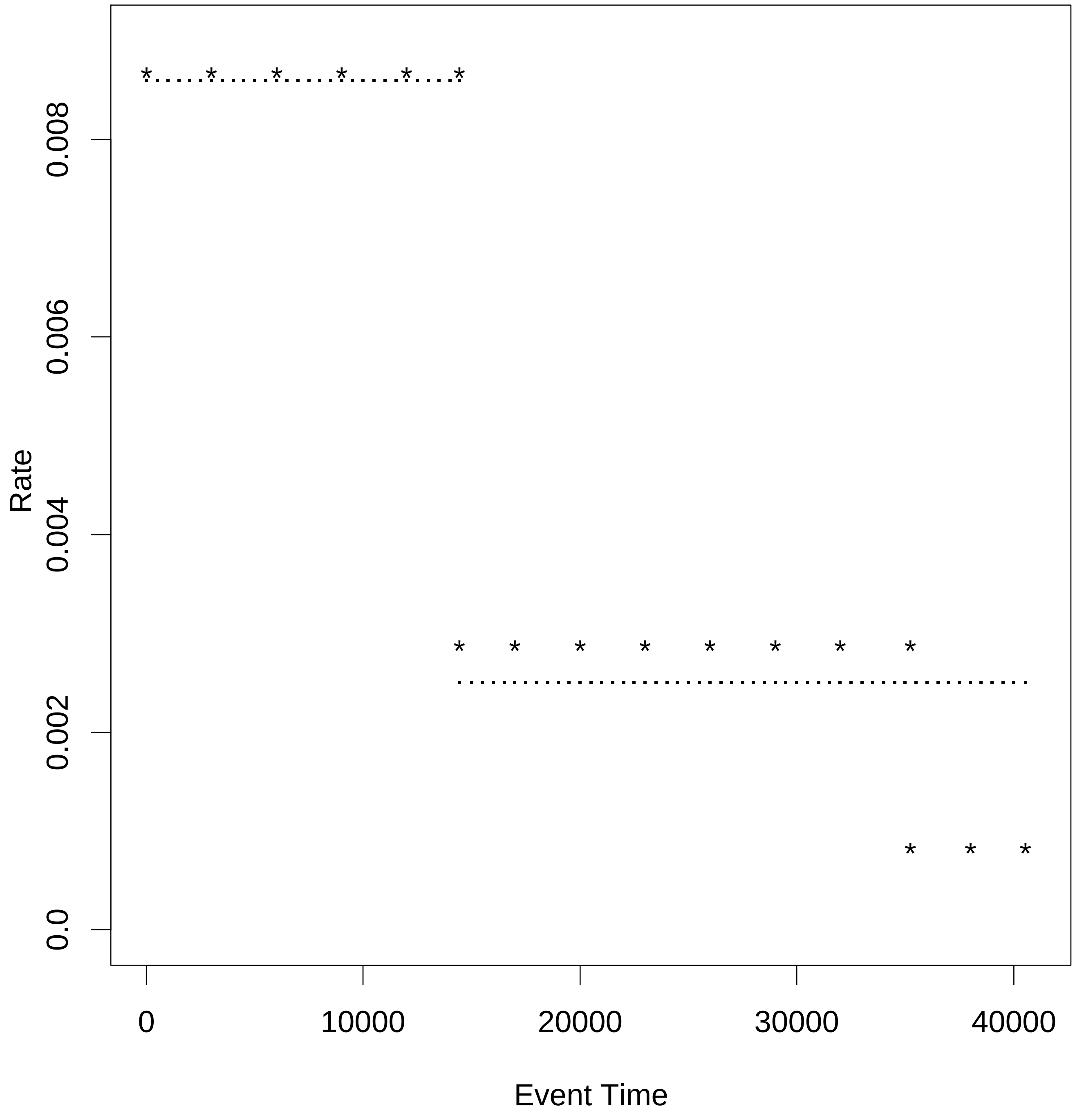

Event Number
0
50
100
150
0
10000
20000
30000
40000
Event Time

Figure 4:


0.0
0.002
0.004
0.006
0.008
Rate
0
10000
20000
30000
40000
Event Time