NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Approximate Bayesian Inference for nonhomogeneous Poisson processes with application to survival analysis

by

Rupali Akerkar, Sara Martino and Håvard Rue

PREPRINT
STATISTICS NO. 3/2012

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2012/S3-2012.pdf
Rupali akerkar has homepage: http://www.math.ntnu.no/∼akerkar
E-mail: akerkar@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Approximate Bayesian Inference for nonhomogeneous Poisson processes with application to survival analysis

Rupali Akerkar, Sara Martino and Håvard Rue
Department of Mathematical Sciences
NTNU, Norway

March 6, 2012

Abstract

Multiple event data occur in survival analysis when two or more events occur to a subject in the study. Examples include the occurrence of asthma attacks in respirology trials, the recurrence of tumours after surgical removal in cancer studies and recurrent heart attacks of coronary patients being treated for heart disease. We discuss a Bayesian semiparametric model for such multiple event data. We assume that multiple events occur according to a nonhomogeneous Poisson process. We decompose the intensity function into the product of various terms. First is a baseline intensity, second a term which includes the effect of various covariates, and third a frailty term to take care of heterogeneity among different individuals with respect to their tendency to develop events. We model the baseline intensity using a piecewise constant model. We demonstrate that we can rewrite the model into a latent Gaussian model which allows us to perform Bayesian inference using integrated nested Laplace approximations (INLA) ((Rue et al., 2009)). The big benefit is computational speed, as most models do not require more than a few seconds to run, but also accuracy in the results as the errors in the approximation are relative and not additive as with Monte Carlo based inference. We illustrate our approach using both simulated and real life data.

1 Introduction

In many research studies, the event of interest can only occur once for a given subject, for example death. It is sometime of interest to study events that may occur several times for a given subject. Such type of data is called multiple event data and arises in many fields, such as manufacturing and industrial reliability, biomedical studies, criminology and demography among others. Multiple events can be of two types, one when identical events are involved and the other when events considered are not identical. We concentrate on identical multiple events. A few examples are the occurrence of asthma attacks in respirology trials, the recurrence of tumour after surgical removal in cancer studies, recurrent heart attacks of coronary patients being treated for heart disease, and the discovery of a bug in an operating system.

Multiple events data have been studied by many authors in different contexts, for early accounts see (Gail et al. (1980); Andersen and Gill (1982); Lawless (1987); Oakes (1992)) and for recent review see (Manda and Meyer (2005); Cook and Lawless (2007)).

In a Bayesian approach, Sinha (1993) propose a semiparametric model for multiple event-time data. According to this model, the events arise in the ith, $i = 1, \cdots, N$ subject as a conditional Poisson process with

conditional intensity function

$$h(t|\boldsymbol{z}_i, w_i) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{z}_i)w_i \tag{1}$$

Here w_i is the subject specific random effect (frailty), $\boldsymbol{z}_i^T$is a vector of covariates, $\boldsymbol{\beta}$ a vector of unknown parameters, and $h_0(t)$ is the baseline intensity function. The frailty takes care of heterogeneity among different subjects. It is assumed that given the unobserved frailty, the intensity function for an individual does not depend on the number of previous events experienced by the individual. Sinha (1993) models the baseline cumulative intensity by a Gamma process with unknown parameters, while the frailty w_i is assumed to be Gamma with mean 1 and unknown variance. A Gibbs sampler is used to sample from the joint posterior distribution of the unknown parameters. This model is useful when the focus is on the regression parameter $\boldsymbol{\beta}$ and the frailties w_i.

In this report, we modify the methodology of Sinha (1993) to discuss the semiparametric model for multiple event data. We consider the conditional intensity function as given in (1) and use the fact that the conditional distribution of new events occurring for a particular subject in an interval is Poisson. We model the baseline intensity using a piecewise constant function. The main purpose of this report is to demonstrate that we can rewrite the nonhomogeneous Poisson processes into a latent Gaussian model which allows us to perform Bayesian inference using integrated nested Laplace approximations (INLA) (Rue et al., 2009). INLA provides fast and accurate deterministic alternative to Markov chain Monte Carlo (MCMC), which at moment is the standard tool for inference in such models. INLA compute approximations to posterior marginals for each component in the model, from which posterior expectation and standard deviations can easily be found. The software is open source and is available for Unix, Windows and Mac. It can be downloaded from website `www.r-inla.org`. On the same web site documentation details and applications are also provided.

The remainder of the report is organized as follows. Section 2 introduces latent Gaussian models with a short description of the INLA methodology. In section 3, we discuss the model and prior structure. In Section 4 we apply our proposed methodology to simulated data. In section 5, we illustrate our approach using two real life examples based on data sets from Gail et al. (1980) and Kvaløy and Skogvoll (2007). Section 6 contains some discussions.

2 Latent Gaussian models and INLA

Latent Gaussian models are a subset of Bayesian additive models with a structured additive predictor. For such models, the likelihood function for response variable y_i is related to the covariates through some structured additive predictor η_i

$$\eta_i = \beta_0 + \sum_{j=1}^{n_f} f^j(u_{ji}) + \sum_{k=1}^{n_\beta}\beta_k z_{ki} + \varepsilon_i \tag{2}$$

Here, the $\{f^{(j)}(\cdot)\}$s are unknown functions of the covariates $\boldsymbol{u}$, the $\{\beta_k\}$s are the linear effect of covariates $\boldsymbol{z}$ and ε_is are unstructured terms. A latent Gaussian model is obtained by assigning $\boldsymbol{x} = \{\{f^{(j)}(\cdot)\}, \{\beta_k\}, \{\eta_i\}\}$ a Gaussian prior. The density $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}_1)$ is assume to be Gaussian with (assumed) mean zero and precision matrix $\boldsymbol{Q}(\boldsymbol{\theta_1})$. The density of $\boldsymbol{x}$ is controlled by vector of hyperparameters $\boldsymbol{\theta}_1$, which are not necessarily Gaussian.

Let the distribution for response variable $\boldsymbol{y} = \{y_i : i = 1, \cdots, N\}$ be denoted by $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_2)$ and assume that y_is are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}_2$, some additional hyperparameters in the likelihood, then

the posterior distribution is given by

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_i \pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) \mid \boldsymbol{Q}(\boldsymbol{\theta}) \mid^{n/2} \exp\Big(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_i \log \pi(y_i|x_i, \boldsymbol{\theta})\Big)\end{aligned} \tag{3}$$

Here, $\boldsymbol{\theta} = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$ with $\dim(\boldsymbol{\theta}) = m$. This posterior density is not analytically tractable as the likelihood is not Gaussian. INLA (Rue et al. (2009)) builds approximations to the posterior marginals of $\pi(x_i|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ assuming two basic properties, First, the latent field $\boldsymbol{x}$ (often of large dimension) admits conditional independence properties, it is a Gaussian Markov random filed (GMRF) with a sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_1)$,(Rue and Held, 2005). Secondly, the number of hyperparameters $\boldsymbol{\theta}$ is not very large (say $m \leq 20$). Finally each point y_i should depend on the latent field $\boldsymbol{x}$ only through the predictor η_i, i.e. $\pi(y_i|\boldsymbol{x}, \boldsymbol{\theta}_1) = \pi(y_i|\eta_i, \boldsymbol{\theta}_1)$.

The approximations $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ and $\tilde{\pi}(x_i|\boldsymbol{\theta}_k, \boldsymbol{y})$, $i = 1, ..., n$ are based on a clever use of Laplace approximations. Rue et al. (2009) describe three different approximations for $\tilde{\pi}(x_i|\boldsymbol{\theta}_k, \boldsymbol{y})$, namely a Gaussian, a Simplified Laplace and a Laplace. The default option in the INLA library is the Simplified Laplace approximation and this is used in all the examples in this report. Posterior marginals for the latent variables $\tilde{\pi}(x_i|\boldsymbol{y})$ is obtained by numerical integration. For more detail we refer to Rue et al. (2009).

A small example of latent Gaussian model is as follows:

- Let $\boldsymbol{t} = \{t_i : i = 1, \cdots, N\}$ be Weibull distributed response variables, such that the hazard function is

  $$h(t) = \lambda_i s t_i^{s-1}, \quad t > 0, s > 0, \lambda_i > 0$$

  where s is the shape parameter and λ_i is the scale parameter.

- Let $\lambda_i = \exp(\beta_0 + \beta_1^{'} z_i)$, where z_i is an observed covariate and (β_0, β_1) are the unknown parameters of interest.

- Let $\eta_i = \beta_0 + \beta_1^{'} z_i$

Such model can be written as a latent Gaussian field if we assume Gaussian priors for β_0 and β_1. Then the vector $\boldsymbol{x} = (\beta_0, \beta_1, \eta_1, \cdots, \eta_N)$ has a joint Gaussian distribution and takes the role of Gaussian field. There is only one hyperparameter $\boldsymbol{\theta}_2 = s$, for which we assume a Gamma(a,b) prior distribution with known mean and variance.

In the Weibull model, described above, the likelihood for t_i depends on the latent field $\boldsymbol{x}$ only through predictor η_i and therefore INLA can be applied directly in such cases. The graph of latent Gaussian model for this example is shown in Figure 1.

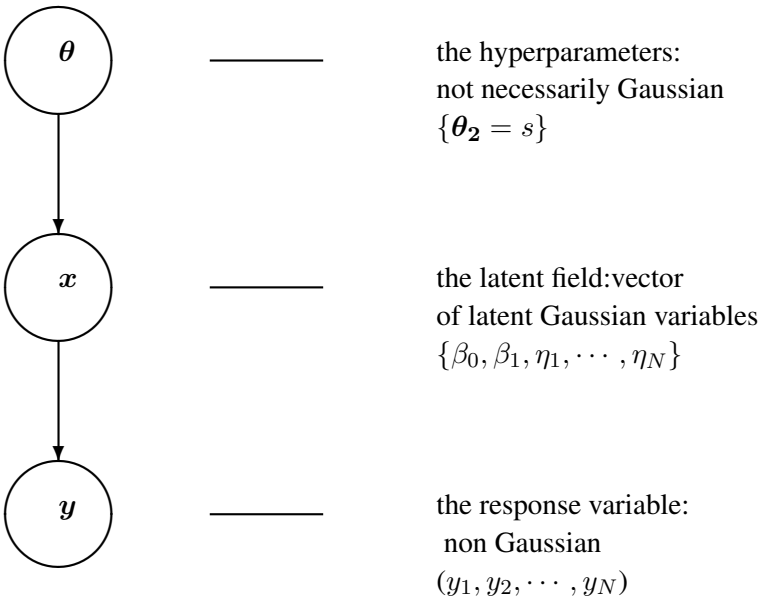


Figure 1: The structure of Latent Gaussian model for Weibull example

3 Nonhomogeneous Poisson process model

For survival data the most common and widely used approach is the Cox proportional hazards model (Cox (1972)), which describes the hazard for an individual with covariate vector $\boldsymbol{z}$ by the equation

$$h(t|\boldsymbol{z}) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{z}) \tag{4}$$

where $h_0(\cdot)$ is the baseline hazard function and $\boldsymbol{\beta}$ is the vector of parameters associated with covariates $\boldsymbol{z}$. The model in (4) is used for time to event data but can be extended for multiple events as suggested by Lawless (1987) and Sinha (1993).

To construct such a model, we partition the time axis into K non overlapping intervals, $0 = s_0 < s_1 < ...s_K = T$, define the k-th interval as $I_k = (s_{k-1}, s_k]$. We assume the baseline intensity to be constant in each interval:

$$h_0(t) = \lambda_k \quad \text{for} \quad t \in I_k = (s_{k-1}, s_k]$$

Suppose that there are N subjects under observation, let $E_i(T)$ denote the number of events occurring to subject i by time T. Let $E_{ik} = E_i(s_k) - E_i(s_{k-1})$ be the number of events occurring to subject i in interval I_k. Then the total number of events occurring to subject i during the study time is $E_i = \sum_k E_{ik}$. We assume $\{E_{ik}; i = 1, \cdots N, k = 1, \cdots, K\}$ to be independent Poisson distributed random variables. For $E_i(t)$, we assume a conditional non-homogeneous Poisson process given the covariate vector $\boldsymbol{z}_i$, and unobserved random

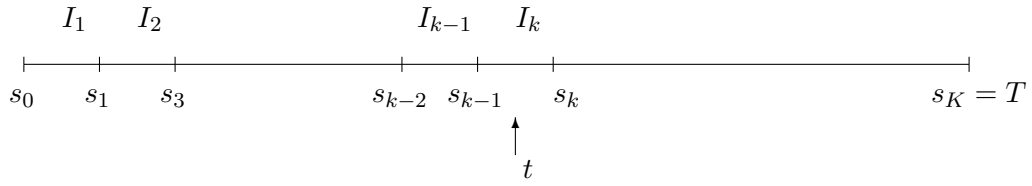


Figure 2: The time line is partitioned into K non overlapping intervals.

frailty w_i (Sinha (1993)). With all the specifications, the conditional proportional intensity function for subject i in the interval I_k is given by

$$\begin{aligned} h(t|\boldsymbol{z}_i, w_i) &= h_0(t) w_i \exp(\boldsymbol{\beta}^T \boldsymbol{z}_i), \qquad t \in I_k = (s_{k-1}, s_k] \\ &= \exp(\log(h_0(t)) + \log(w_i) + \boldsymbol{\beta}^T \boldsymbol{z}_i) \\ &= \exp(b_k + \alpha_i + \boldsymbol{\beta}^T \boldsymbol{z}_i) \end{aligned} \tag{5}$$

where, $\eta_{ik} = b_k + \alpha_i + \boldsymbol{\beta}^T \boldsymbol{z}_i$ with $b_k = \log(\lambda_k)$ and $\alpha_i = \log(w_i)$.

The conditional distributions of the number of events, E_{ik} (for $k = 1, \cdots, K$) given w_i and $\boldsymbol{z}_i$, are independent Poisson and can be expressed as

$$E_{ik} \sim \mathcal{P}(\exp(\eta_{ik})(s_k - s_{k-1})) \tag{6}$$

And we assume that if $E_{ik} \perp E_{ik'} | w_i, \boldsymbol{z}_i$ for $k \neq k'$. Under non informative censoring, the log-likelihood contribution of subject i is given by

$$l_i \propto \sum_{k=1}^{K} \Big\{ E_{ik} \log \Big(\exp(\eta_{ik})(s_k - s_{k-1}) \Big) - \exp(\eta_{ik})(s_k - s_{k-1}) \Big\} \tag{7}$$

This is the contribution from subject i over many short intervals. The data enters the likelihood only through the number of events happening in each small interval.

Our aim is to rewrite these non-homogeneous Poisson processes into a latent Gaussian model which allow us to perform Bayesian inference using integrated nested Laplace approximations (INLA).

We assume Gaussian priors with large variance for $\boldsymbol{\beta}$. For the log frailty term α_i we assume Gaussian prior with unknown precision τ_α. For the log-baseline intensity, b_k, we assume correlated prior process, namely an intrinsic first-order random walk (RW1) model (Rue and Held (2005), Ch.3) with precision τ_b. RW1 models are built by first assuming that the location k of the nodes are all positive integers, i.e. $k = 1, \cdots, K$ so that the distance between nodes is constant and equal to 1. Then, increments $b_{k+1} - b_k$ are assume independent and identically distributed.

$$b_{k+1} - b_k \sim N(0, \tau_b^{-1}), \qquad k = 1, \cdots K - 1 \tag{8}$$

We assume Gamma priors with known parameters for the hyperparameters, $\boldsymbol{\theta} = (\tau_\alpha, \tau_b)$. Conditioned on $\boldsymbol{\theta}$, the latent field

$$\boldsymbol{x} = \{\alpha_1, \cdots, \alpha_m, b_1, ..b_K, \boldsymbol{\beta}, \eta_{11}, \cdots\} \sim N(\mathbf{0}, \boldsymbol{Q}^{-1}) \tag{9}$$

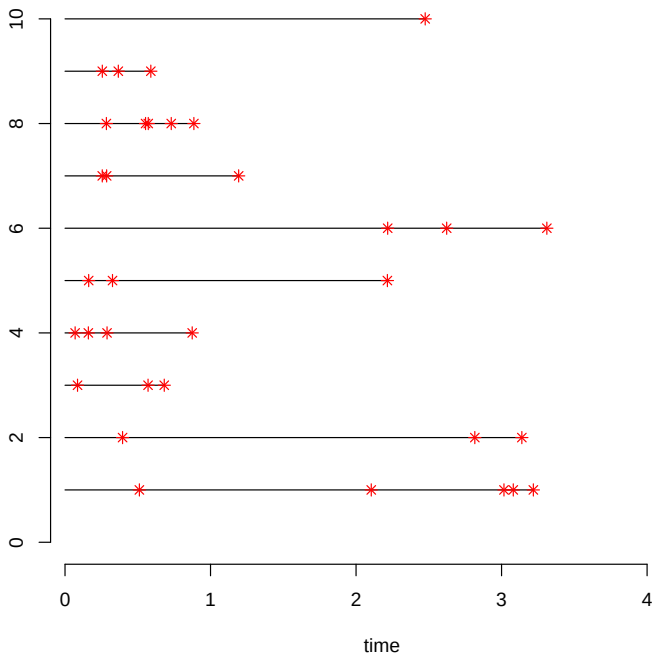


Figure 3: A sample of simulated data (* denotes detecting of an event).

has Gaussian distribution with precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$. By doing this we see that conditional on $\boldsymbol{\theta}$, the likelihood of number of events occurring to subject i in interval k depends on the latent Gaussian field $\boldsymbol{x}$ only through the predictor η_{ik}. Which is a requirement in INLA for computational purpose.

4 Simulation

A simulation study is conducted to assess the performance of our model. We simulate multiple events for 1000 cases of a non-homogeneous Poisson process using the thinning (random sampling) approach as discussed in Ross (2002). We use Example 3.24 from Rizzo (2008) with intensity function

$$h(t) = 3cos^2(t)$$

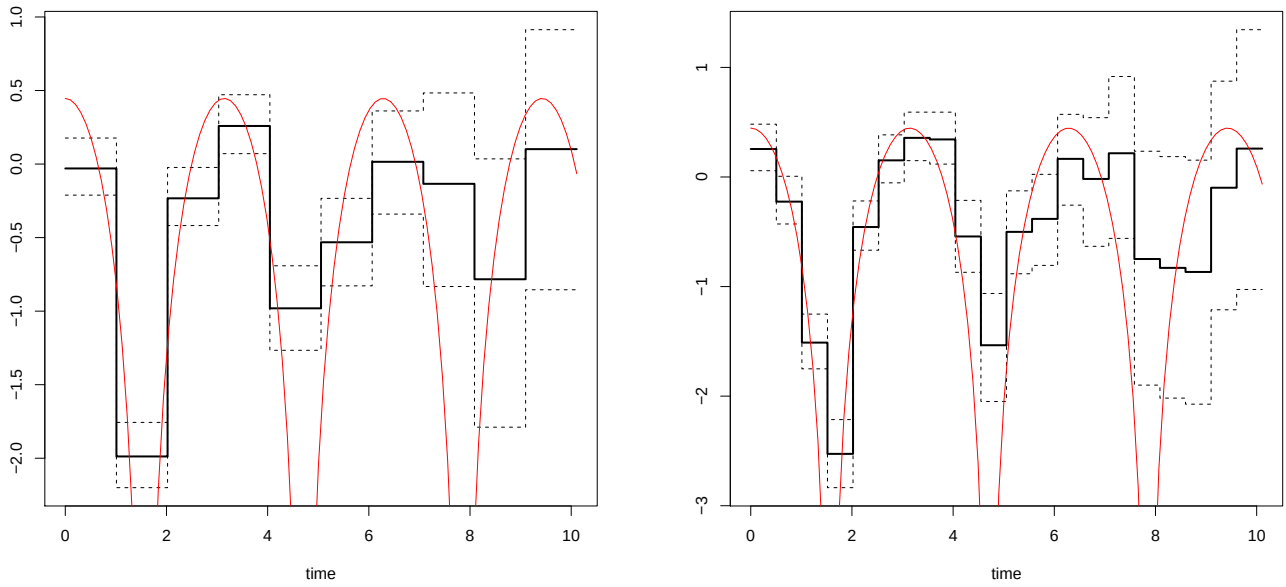
For this model the baseline intensity is $h_0(t) = cos^2(t)$, which is a function of time, and the covariate is fixed. A sample of simulated data set is given in Figure 3.

To evaluate the efficiency of the estimates of the baseline intensity from our model we divide the time axis into K intervals. We consider 4 different cases by assuming K = 10, 20, 40 and 80. Then we model the number of events $E_{ik}, i = 1, \cdots, 1000$ and $k = 1, \cdots, K$, according to our proposed non homogeneous Poisson process model in section 3. For log baseline intensity we assume RW1 prior.

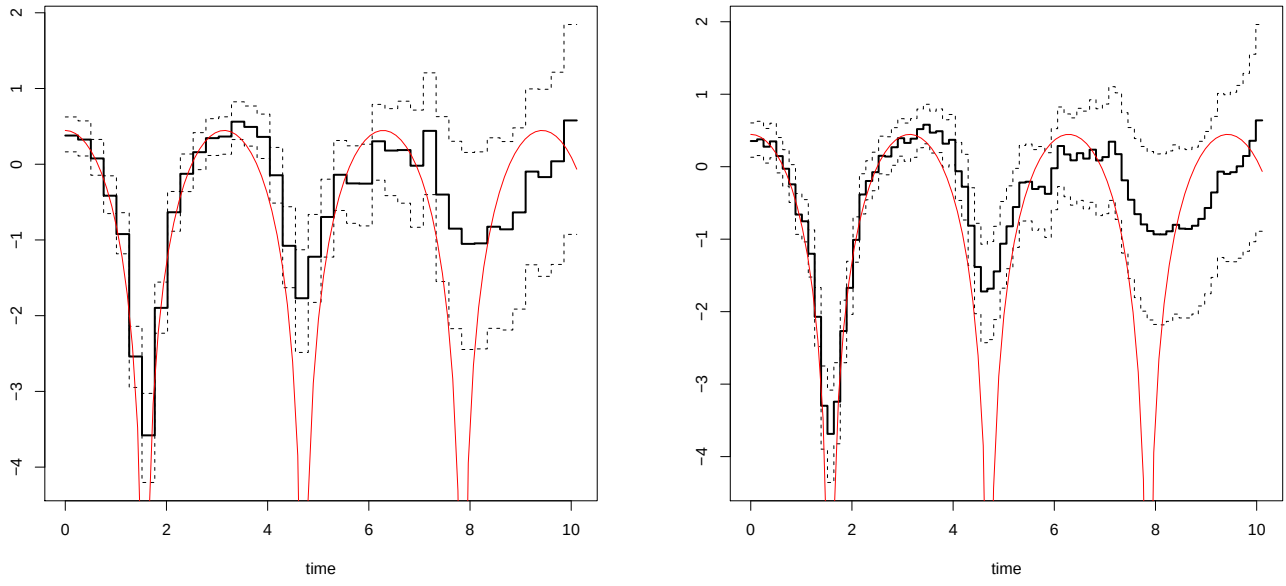
The true curve for log-baseline intensity and posterior estimates of the mean of the log-baseline along with 95% credible intervals are shown in Figure 4. The pattern of $cos^2(t)$ is reasonably captured by piecewise

constant baseline intensity. By increasing the number of intervals we get better approximations for the baseline intensity.

We observe that the estimates of log baseline intensity are quite reasonable for the intervals where there is sufficient data. There is difference in the true values and the estimates on the right end of the plots in Figure 4(a), Figure 4(b), Figure 4(c) and Figure 4(d). Also the 95% credible intervals are very broad as there are not so many values.



(a) (b)



(c) (d)

Figure 4: Simulation study: the true log baseline intensity (red curve) with posterior estimates of means (solid line black) and 95% credible intervals (dashed lines), when the time axis is divided into K = 10 (Panel (a)), K = 20 (Panel(b)), K = 40 (Panel(c)) and K = 80 (Panel(d)) intervals.

5 Applications

In this section we re analyse the mammary tumour data from Gail et al. (1980) and cardiac arrests data from Kvaløy and Skogvoll (2007) using the proposed model. For mammary tumour data, we also compare results obtained by INLA and MCMC. The examples are run on a dual-core 2.5GHz laptop and the execution times refer to such machine.

5.1 Example: Mammary tumour data

We consider times to development of mammary cancer in 48 rats given by Gail et al. (1980). These animals were injected with a carcinogen and then were randomly assigned to receive either the treatment or control. The occurrence of tumours are noted twice a week from day 62 till day 182. All animals were right censored after 182nd day.

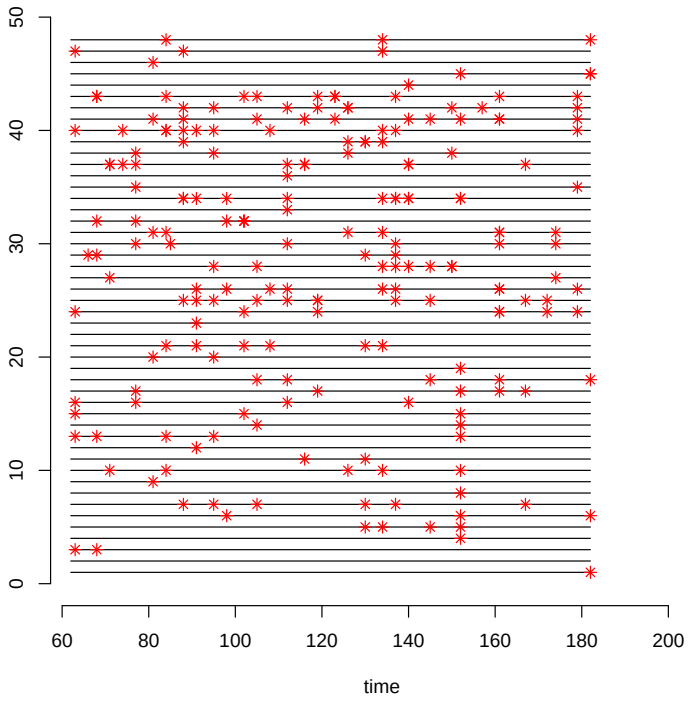
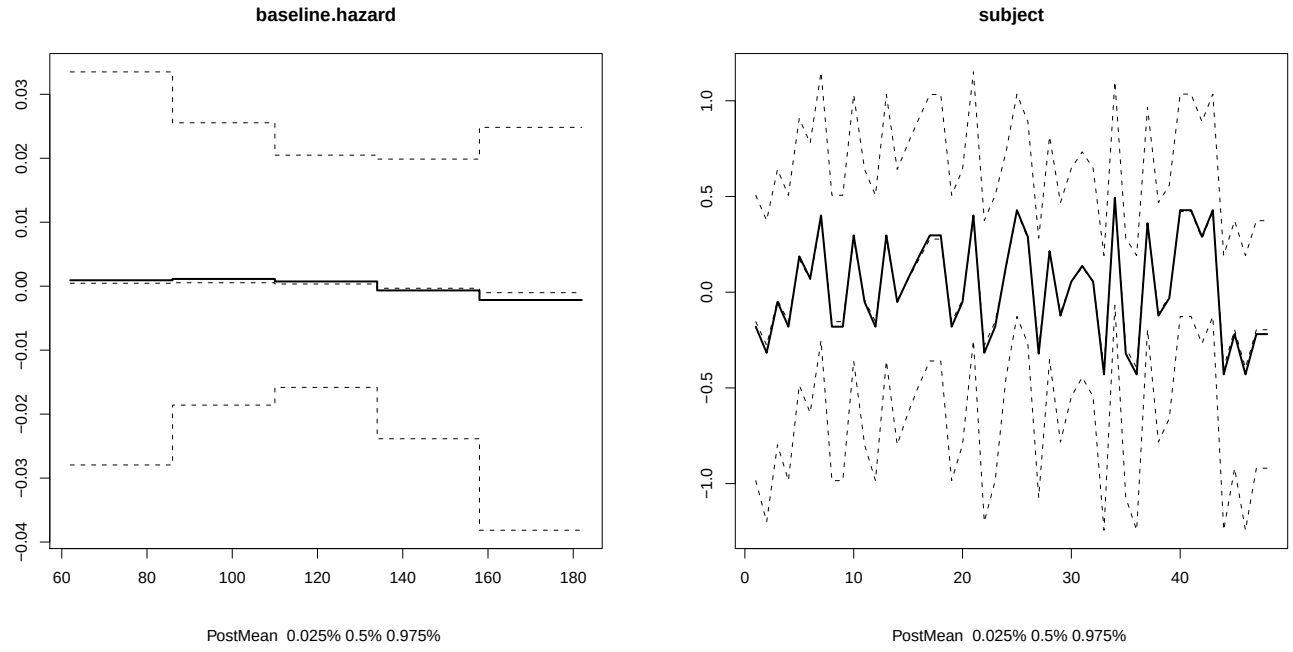


Figure 5: Display of events for tumour occurrences in 48 Rats.

The data set consists of times to tumour in days for each rat. The only covariate used is group, group=1 denotes treatment and group=2 denotes control. Figure 5 displays the occurrence of tumours for all 48 rats under study, in the figure $*$ denotes a detected tumour. We analyse the data by assuming piecewise constant baseline intensity function. Analysis starts by partitioning the time axis into 5 intervals of equal length. Thus for this data $i = 1, \cdots, 48$ and $k = 1, \cdots, 5$. The conditional proportional intensity for animal i in time interval k is

$$h(t|z_i, w_i) = \exp\{\beta_0 + \text{group}_i\beta_1 + b_k + \alpha_i\}, \quad t \in I_k \tag{10}$$



(a) (b)

Figure 6: Posterior means by INLA (a) log baseline intensity and (b) log frailty.

where $s_k - s_{k-1}$ is constant as the time intervals are same. Moreover, we assume are as follows, $\beta_0 \sim \mathcal{N}(0, 0.001^{-1})$, $\beta_1 \sim \mathcal{N}(0, 0.001^{-1})$, $\alpha_i \sim \mathcal{N}(0, \tau_\alpha^{-1})$, $\log(\boldsymbol{\lambda}) = \boldsymbol{b} \sim RW1(\tau_b)$, further we assign Gamma priors, $\Gamma(a, b)$ with mean (a/b) and variance (a/b^2) for the hyperparameters, $\tau_\alpha \sim \Gamma(1, 0.001)$ and $\tau_b \sim \Gamma(1, 0.001)$.

To implement the model in INLA, we define the formula and the `inla()` function as follows:

```
cutpoints =  seq( 62,182, len=6)
formula = inla.surv(time,event, subject=subject) ~ group
        +  f(subject, model="iid",param=c(1,0.001))
model = inla(formula,family="coxph",control.hazard=list(cutpoints
      =cutpoints), control.inla = list(
        int.strategy="grid", diff.logdens=15, dz=0.2), data=data)
   h  = inla.hyperpar(model,dz = 0.2,  diff.logdens = 15 )
```

`inla.hyperpar` is used to improve the estimate of the marginal posterior densities of the hyperparameters. For more details about `inla.hyperpar()` we refer to Rue et al. (2009). `inla.surv()` function is discussed in Akerkar et al. (2010).

Estimates of the log-baseline intensity and log frailty are shown in Figure 6. The log-baseline intensity is constant. The frailty effect provides strong evidence of heterogeneity among the rats. Thus some of the rats are more prone to tumours as compared to others in the same group. The estimated mean and standard deviation of group (treatment effect) are calculated as 0.8 and 0.2. Which indicates the significance of the treatment effect. Our results agree with the results given in Sinha (1993).

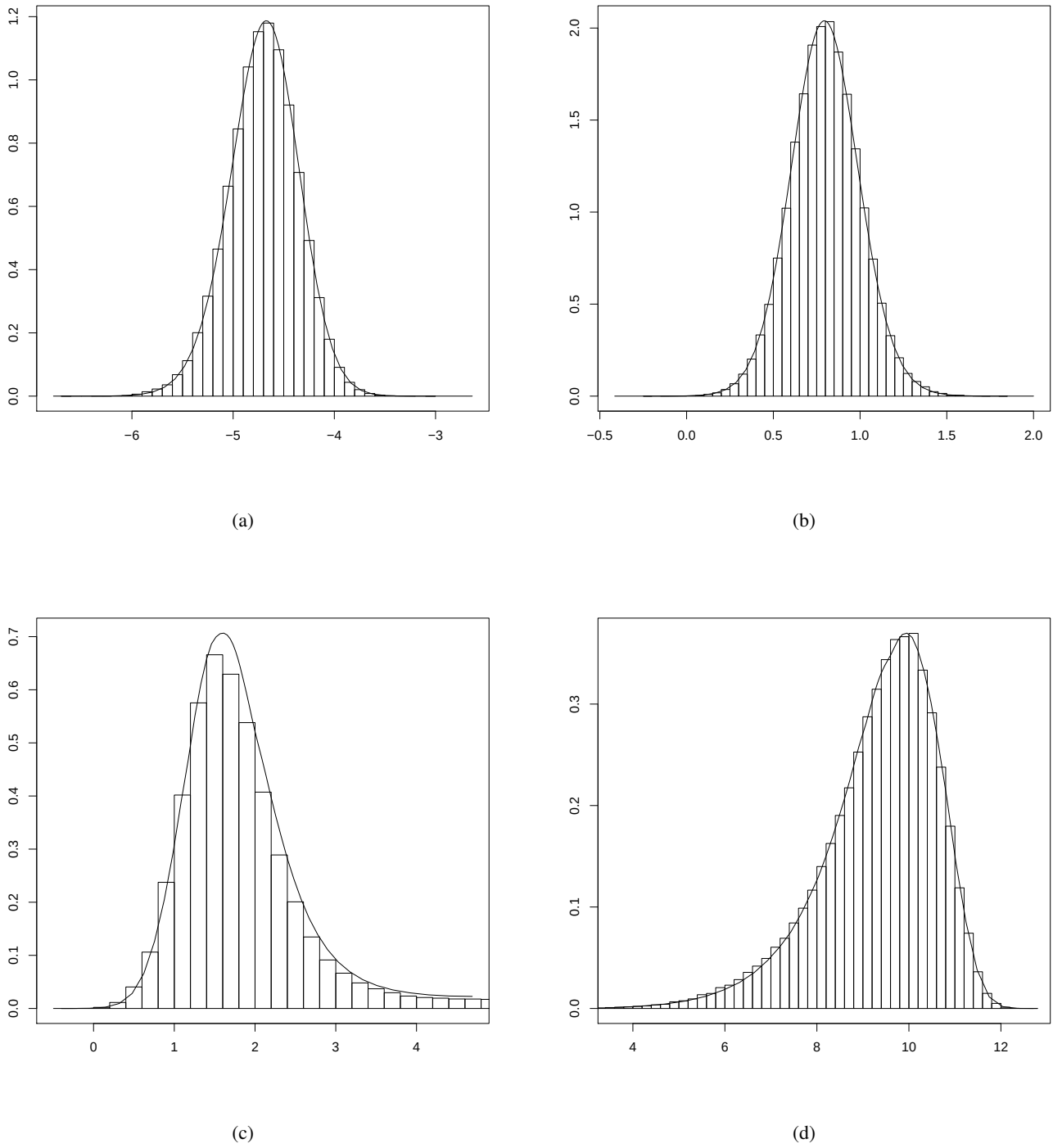


Figure 7: Posterior marginals distributions approximated by INLA(solid line) and MCMC based density estimates (histogram) (a) β_0, (b) β_1, (c) τ_b and (d) τ_α.

To assess the quality of the INLA approximations, we compare them with MCMC estimates, obtained by a " one-block MCMC sampler" described in (Rue and Held (2005), Ch. 4). In Figure 6 the INLA posterior marginals for $\beta_0, \beta_1, \tau_b, \tau_\alpha$ are compared to histograms based on long MCMC runs. The processing time for INLA() was 57 seconds and MCMC results took about 2 hours for 10^6 updates. The results are quite comparable.

To check whether it is important to include the frailty effect in the model, we fit a model without considering frailty effect . The estimated mean and standard deviation of group effect is 0.80 and 0.15. The results obtained clearly support the treatment effect, but the standard error is underestimated. Our results are similar with those obtained by Lawless (1987).

5.2 Example: Cardiac arrest data

In this example, we re analyse the effect of weather conditions on cardiac arrests in a specified population. We consider the data set concerning the occurrence of cardiac arrests treated by the emergency medical service in Trondheim, a city in central Norway during the time period from the start of January 1990 until the end of November 1998 (Kvaløy and Skogvoll (2007)). The data is available for 3256 days and 809 cardiac arrests were reported during this period. The details about the number of events per day is given in Table 1. In the current example the number of events occurred are quite small and there are approx. 78 % days without any event. Kvaløy and Skogvoll (2007) consider several covariates but we include only those covariates, which they concluded are important.

cardiac arrest	0	1	2	3
number of days	2536	636	79	5

Table 1: The number of events per day.

We consider covariates air temperature, relative humidity, wind speed, precipitation, snowfall indicator and day number (day number of a year). Temperature, relative humidity and wind speed are recorded several times a day. For the analysis, we used the daily averages of these variables. Furthermore instead of using the direct snow depth covariate, we use indicator function being 1 with snowfall and 0 otherwise, which we denote by snow.

Following the suggestion by Kvaløy and Skogvoll (2007), we consider the time interval of 24 hours as it seems reasonable that the occurrence of cardiac arrest are affected by this cycle. For this example, the subject is a day and we analyze the effect of covariates (weather conditions) on the number of cardiac arrests in a day. Let E_{ik} represent the number of cardiac arrests on day i in the kth interval. We partition a time interval of 24 hours of a day in 12 equal intervals. The predictor function for ith day in kth interval is as follows

$$\begin{aligned}\eta_{ik} = \beta_0 + \beta_1 \text{windspeed}_i + \beta_2 \text{snow}_i + f^{(\text{temp})}(\text{temp}_i) + f^{(\text{precipitation})}(\text{precipitation}_i) \\ + f^{(\text{humidity})}(\text{humidity}_i) + f^{(\text{day})}(\text{day}_i) + b_k\end{aligned} \tag{11}$$

where $i = 1, \cdots, 3256$ and $k = 1, \cdots, 12$. As discussed in section 3, $b_k = \log(\lambda_k)$ is the log of baseline intensity in kth interval.

We assume linear effects for windspeed and snow, and smooth effects for temperature, precipitation, relative humidity, day number and log-baseline intensity. We assume RW1 prior for log baseline intensity,

and RW2 priors for temperature, precipitation, relative humidity, day number (Rue and Held (2005), Ch. 3). Moreover, since we have chronological data for nearly nine years, we consider day number as cyclic covariate. All hyperparameters are assign gamma priors with known precision.

The posterior estimates of windspeed and snow are summarized in Table 2.

covariate	mean	s.d.	0.025quant	0.975quant
windspeed	-0.005	0.015	-0.03	0.02
snow	0.162	0.105	-0.04	0.37

Table 2: Posterior estimates of windspeed and snow.

Estimates of temperature, precipitation, relative humidity and day number are given in Figure 8. The intensity of cardiac arrests is maximum when temperature is little lower than $0°$C and decreases with increase in temperature (Figure 8(a)). The effect of precipitation is clearly linear (Figure 8(b)). The effect of relative humidity seems constant (Figure 8(c)). It is quite clear from Figure 8(a), 8(b) and 8(c) that the uncertainty in the estimates is largest at the boundaries where there are less observations.

The covariate day number is used in the model to incorporate the seasonal changes. The effect of day number varies over whole year but is maximum during winter and least during spring (Figure 8(d)).

Figure 9(a) shows the histogram of actual cardiac arrests times. The number of cardiac arrests are least around 5 in the morning. It increases with activity level until 16 hours, then it decreases and remains stable. The estimates of the log baseline intensity along with 95% credible intervals are shown in Figure 9(b).

It is clear from Figure 8(a) and Figure 8(d) that the occurrence of cardiac arrest is affected by both temperature and day number. The occurrence of cardiac arrest is more in winter or when temperature is negative. Since temperature and day number are closely related, we want to study their effectiveness when modelled separately.

We model the number of cardiac arrests in two different models, in one we consider only day number and in the other only temperature. We assume smooth effect for both temperature and day number. The posteriors estimates along with 95 % credible intervals for temperature and day number are given in Figure 10.

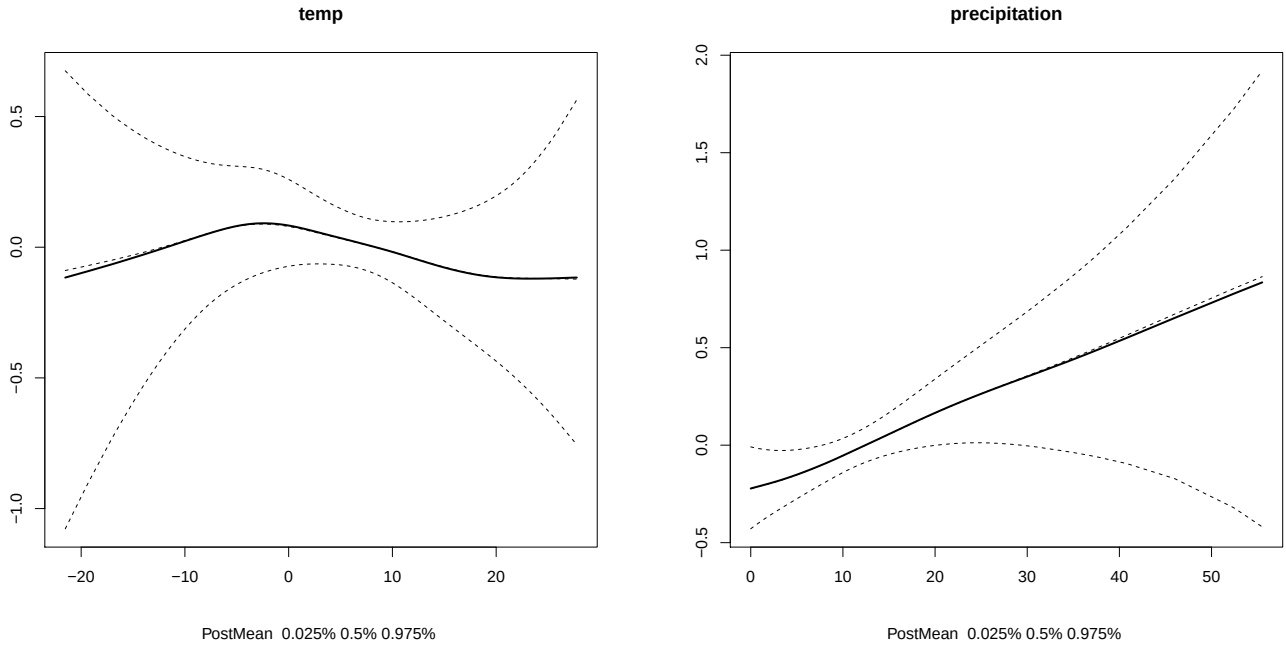
The significance of temperature and day number is evident from Figure 10(a) and Figure 10(b). The intensity of cardiac arrests is constant but higher, when temperature is less than $0°$C and decreases sharply with increase in the temperature. The effect of day number is varying all year. Though, the occurrence of cardiac arrest are more during winter, when the weather conditions are extreme (bad) and is lowest in the spring.

We investigate a number of additional models to learn about significant covariates. We assume linear effect for covariates snow and windspeed, and assume smooth effects (RW2) for temperature, precipitation, relative humidity and day number. Here also we consider cyclic effect for day number.

To compare different models, we use the deviance information criterion (DIC) of Spiegelhalter et al. (2002). Details about some of the models along with the DIC and the effective number of parameters are given in Table 3.

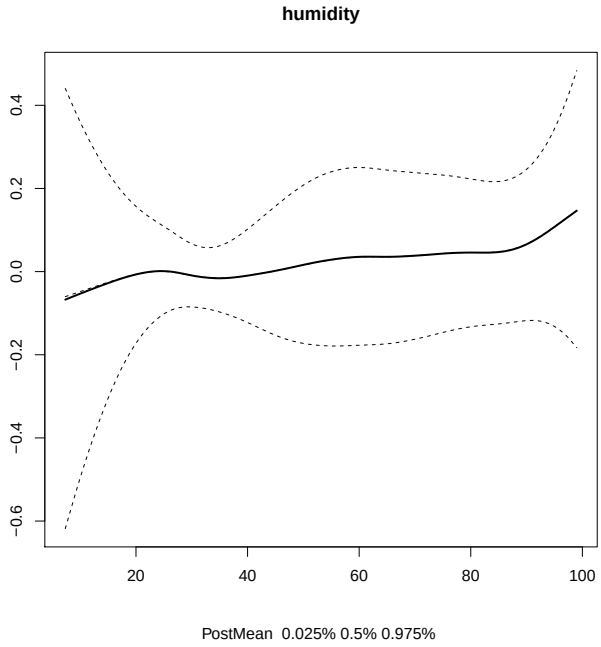
From the results mentioned in Table 3, The best model we obtain in terms of the DIC (7720.15) is by including temperature, precipitation and snow in the model. While comparing different models, we notice that the DIC of models with day number are more than the DIC of models with temperature, given the other covariates are same. Thus we believe that temperature is more significant than day number.

We conclude that to study the effect of weather covariates on intensity of cardiac arrests, it is sufficient to

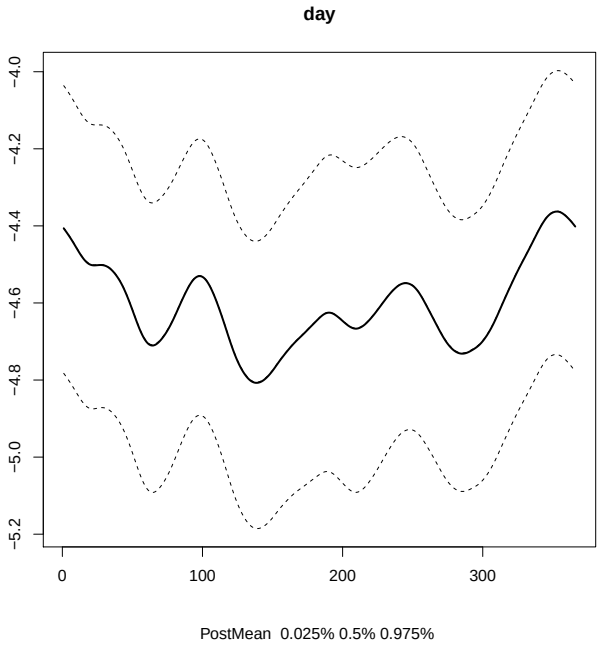


(a)

(b)

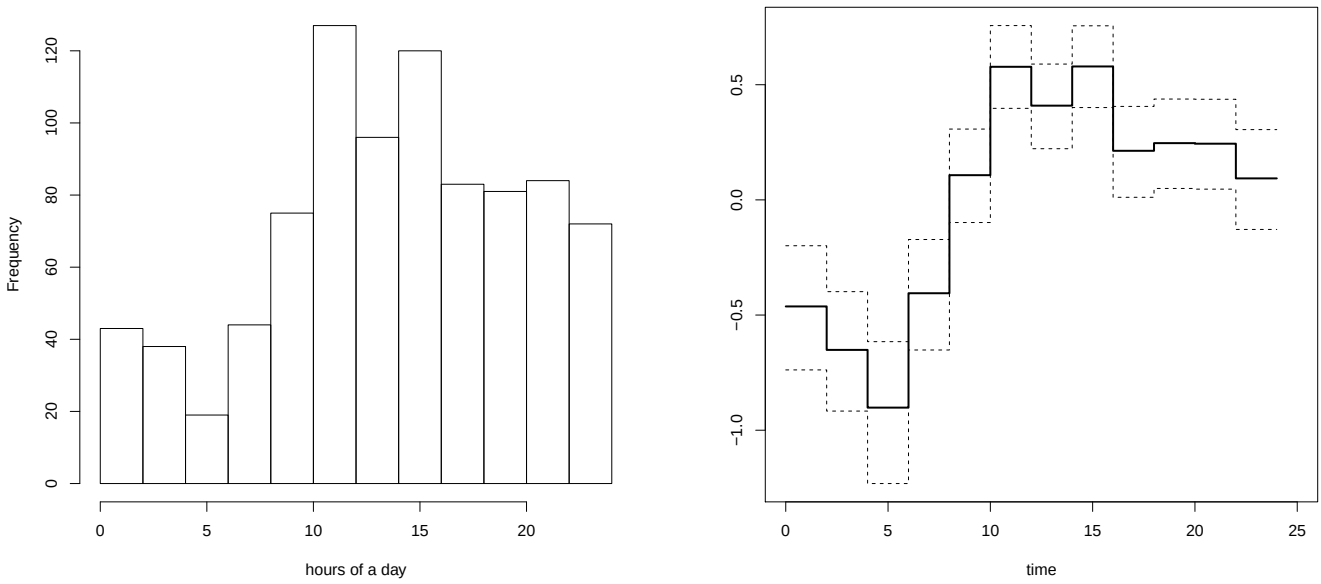


(c)



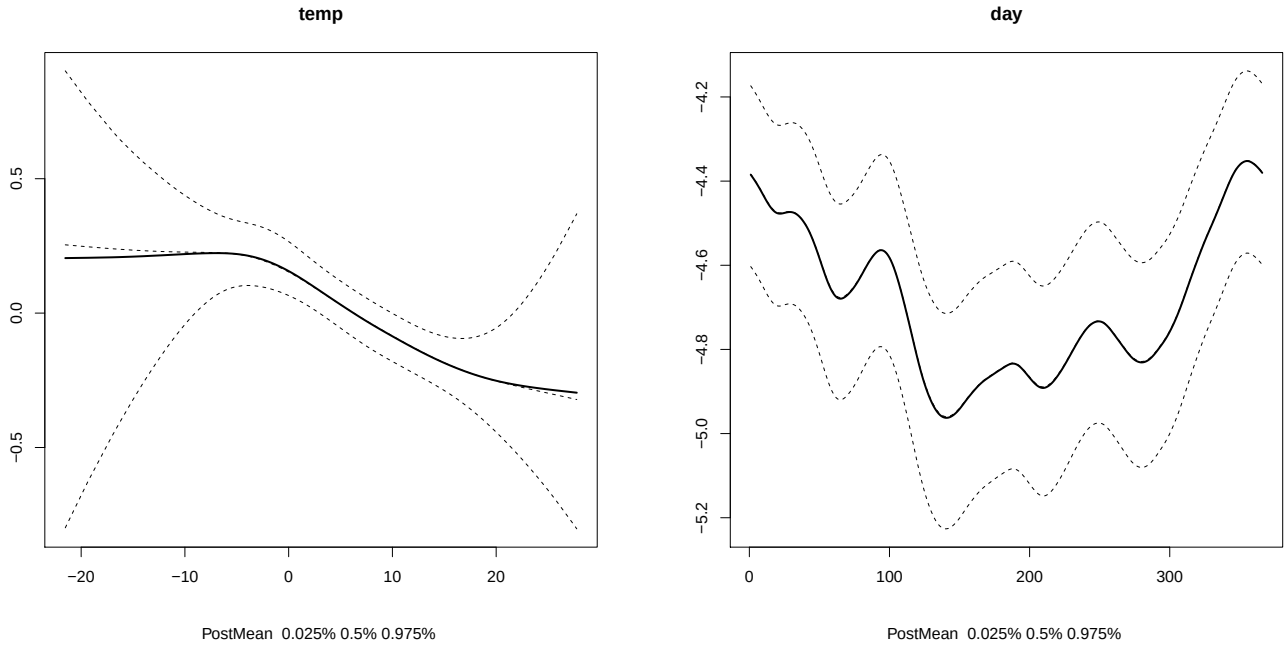
(d)

Figure 8: Posterior means by INLA for (a) temperature, (b) precipitation, (c) relative humidity, (d) day number, (e) log frailty



(a) (b)

Figure 9: (a) Histogram of actual cardiac arrest times. (b) Posterior means and 95% credible intervals (dashed lines) of log baseline intensity .



(a) (b)

Figure 10: Posterior means and 95% credible intervals (dashed lines) of (a) temperature (b) day number, when modelled separately.

Model	Covariates	effective number of parameters	DIC
1	-	9.67	7740.6
2	temperature	12.33	7726.62
3	precipitation	12	7737
4	day number	19.8	7729.3
5	temp + snow	13.17	7725.2
6	snow + day number	20.8	7728
7	temp + precip	14.5	7721.3
8	day number + precip	21.9	7726.27
9	temp + humidity	15.9	7730.15
10	day number + humidity	39.5	7748.9
11	temp + precip + snow	15.32	7720.15
12	day number + precip + snow	22.9	7724.25
13	temp + precip + humidity	18	7725.48
14	day number + precip + humidity	57.7	7759.25
15	temp + precip + snow + humidity + windspeed + day number	29.5	7732.24

Table 3: Cardiac arrests data: the effective number of parameters, DIC and the time used in seconds for different model specifications.

include covariates such as temperature, precipitation and snow in the model. Our results support the general belief about more cardiac arrests in extreme weather conditions.

In our final model, we include three weather variables, temperature, precipitation and snow. Figure 8(b) suggests that precipitation has linear effect, and Figure 10(a) suggest that effect of temperature is closer to linearity. So we assume linear effect for temperature and precipitation along with snow. The results of our final model are summarised in the Table 5. The DIC for the model is 7718.9 and is the minimum of all the DICs. Although the DIC of the model, when we assumed smooth prior for temperature and precipitation is not very different. We conclude that a linear effect for precipitation, temperature and snow are sufficient to describe the occurrence of cardiac arrest.

covariate	mean	s.d.	0.025quant	0.975quant
temperature	-0.015	0.01	-0.027	-0.003
precipitation	0.016	0.01	0.005	0.03
snow	0.17	0.07	-0.03	0.36

Table 4: Posterior estimates of snow.

6 Discussion

In this report, we discussed a Bayesian semiparametric model for multiple event time data based on the so-called proportional intensity model. Conditional on fixed covariate and the frailty random effect, multiple

events occur to a subject according to a non-homogeneous Poisson process .

We avoid parametric assumptions about the baseline and model it using piecewise constant function. We treated random effects (frailty terms) like regression coefficients. We demonstrated that we can rewrite the nonhomogeneous Poisson process model as latent Gaussian model, which allows us to do the approximate Bayesian inference using integrated nested Laplace approximations.

Acknowledgements

We would like to thank Bo Lindqvist for his suggestions, Eirik Skogvoll and Jan Terje Kvaløy for providing data set on cardiac arrests in Trondheim.

References

Akerkar, R., Martino, S., and Rue, H. (2010). Implementing approximate Bayesian inference for survival analysis using integrated nested Laplace approximations. Technical report 1, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression models for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120.

Cook, R. and Lawless, J. (2007). *The statistical analysis of recurrent events*. Springer Verlag.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Gail, M. H., Santner, T. J., and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times of tumor. *Biometrics*, 36:255–266.

Kvaløy, J. T. and Skogvoll, E. (2007). Modelling seasonal and weather dependency of cardiac arrests using the covariate method. *Statistics in Medicine*, 26:3315 – 3329.

Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82:808–815.

Manda, S. and Meyer, R. (2005). Bayesian inference for recurrent events data using time-dependent frailty. *Statistics in medicine*, 24(8):1263–1274.

Oakes, D. (1992). Frailty models for multiple event times. *in Survival Analysis: State of the Art, eds. J.P. Klein and P. K. Goel*, pages 371–379.

Rizzo, M. (2008). *Statistical computing with R*. Chapman & Hall.

Ross, S. M. (2002). *Simulation*. Academic Press, California.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

Sinha, D. (1993). Semiparametric bayesian analysis of multiple event time data. *Journal of the American Statistical Association*, 88:979–983.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.