KYBERNETIKA — VOLUME *50* (2014), NUMBER 6, PAGES 849–868

BAYESIAN NONPARAMETRIC ESTIMATION OF HAZARD RATE IN MONOTONE AALEN MODEL

Jana Timková

This text describes a method of estimating the hazard rate of survival data following monotone Aalen regression model. The proposed approach is based on techniques which were introduced by Arjas and Gasbarra [4]. The unknown functional parameters are assumed to be a priori piecewise constant on intervals of varying count and size. The estimates are obtained with the aid of the Gibbs sampler and its variants. The performance of the method is explored by simulations. The results indicate that the method is applicable on small sample size datasets.

Keywords: monotone Aalen model, Bayesian estimation, Gibbs sampler, small sample size

Classification: 62N02, 62G05

1. INTRODUCTION

In survival analysis non- or semi-parametric approaches have become widely used. A functional parameter of a model is often estimated as a piecewise constant function with jumps at every failure time, rather than a pre-specified parametric function. Typical examples are the Kaplan–Meier estimator of survival function in homogeneous case, [12], or Breslow estimator of cumulative baseline hazard function in Cox's proportional hazard model, [5]. In most famous models like Cox's proportional model or Aalen's additive model these estimators have proved to be consistent, their asymptotic features are known and no need to impose a functional form in advance makes them perfect candidates for usage in data analysis. The dimension of the functional space from which the estimators are drawn is fixed to the number of observed failures. Fixed discontinuities located at the failure times could be viewed as a sole drawback of the estimators. The nonparametric estimators of regression functions based on least squares or weighted least squares in Aalen additive model, [1, 2] and [11], are of exactly this type.

Bayesian approach to survival data analysis has become a popular alternative to the aforementioned estimators which allows for more flexible estimation. Furthermore, it enables one to solve concrete problems as integrals with respect to the posterior distribution. Nowadays, the computational feasibility is less of an issue and the inference from complicated models can be obtained using MCMC algorithm. Popular priors for functional parameters of survival models are the nondecreasing independent increment

DOI: 10.14736/kyb-2014-6-0849

processes (NII), a wider class of processes incorporating Gamma and Beta processes (and also Dirichlet processes via the known relationship $H(t) = \int_0^t \mathrm{d}F(s)/1 - F(s_-)$ between the survival function H and the distribution function F). A process, let's say H, is a NII process if H is a nondecreasing right-continuous function having $H(0) = 0$, jumps $\Delta H(t) \leq 1$ and either $\Delta H(t) = 1$ for some t or $\lim_{t\to\infty} H(t) = \infty$, and obviously it induces a proper cumulative hazard function. For more details see [8] or [14]. Lately it was shown by several authors that the estimators of functional parameters based on these priors are consistent and asymptotically equivalent to the standard nonparametric estimators in the homogeneous case, the Cox model and the competing risk model, see [6, 13, 14]. For a good overview of Bayesian analysis in survival models see e. g. [15].

Arjas and Gasbarra [4] suggested Bayesian inference of homogeneous lifetime data using a simple piecewise constant process with dependent increments for prior for hazard function, i. e. a random function which is piecewise constant on some intervals. Number of the intervals and variation of the function from one interval to another is controlled by four hyperparameters. This setting includes desirable possibility of changing the dimension of the model in favor of best fit according to the data while moderated by the prior information. The inference was conducted using Gibbs sampler resulting in a set of piecewise constant trajectories of a process ruled by the posterior distribution of the hazard function. Using these trajectories allowed one to approximate the posterior expectation of the hazard function as well as the cumulative hazard function/survival function or any other integrable function on space of the parameter trajectories.

In this paper we study Aalen additive model of Aalen [1] and [2], on a dataset of form $(T_i, \delta_i, (x_{i,0}, \ldots, x_{i,p})^\top)_{i=1}^n$, where $T_i = \min(T_i^0, C_i)$ are observed right-censored survival times, $\delta_i = I\{T_i = T_i^0\}$ and $(x_{i,0}, \ldots, x_{i,p})^\top$ are $(p+1)$-dimensional covariate vectors. The number of the covariates p is usually quite small, for example up to $p = 3$. T_i^0 is a real lifetime of ith individual with distribution function $F_i = F(\cdot|(x_{i,0}, \ldots, x_{i,p})^\top)$ and C_i is a censoring variable independent on T_i^0. Aalen additive model assumes that the hazard rate for ith object is

$$h_i(t) = \sum_{j=0}^{p} x_{i,j}\alpha_j(t),\ i = 1, \ldots, n, \tag{1}$$

where $\alpha_0, \ldots, \alpha_p$ are unknown regression functions. Typically $x_{i,0} \equiv 1, \forall i$, and α_0 represents baseline risk of failure common for all individuals if there is no other risk factor present. Aalen studied the model assuming that α_js take real values and only the overall hazard function h_i needs to be nonnegative. He estimated the cumulative versions of regression function $A_j(t) = \int_0^t \alpha_j(s)\,\mathrm{d}s, j = 0, \ldots, p$ by a least squares estimator. Let us introduce processes $N_i(t) = I\{T_i \leq t, \delta_i = 1\}$, $Y_i(t) = I\{T_i \geq t\}$. We denote $\alpha(t) = (\alpha_0(t), \ldots, \alpha_p(t))^\top$, $A(t) = (A_0(t), \ldots, A_p(t))^\top$, $z_i = (x_{i,0}, \ldots, x_{i,p})^\top$, $N(s) = (N_1(s), \ldots, N_n(s))^\top$ and $Z(s) = (z_1 Y_1(s), \ldots, z_n Y_n(s))^\top$. Then the Aalen least squares estimator equals to

$$\hat{A}(t) = \int_0^t (Z(s)^\top Z(s))^{-1} Z(s)^\top \mathrm{d}N(s). \tag{2}$$

Huffer and McKeague [11] introduced a two-stage estimator which is in core a weighted least squares estimator with a matrix of weights $V^\star = diag\{Z\alpha^\star\}^{-1}$, where $\alpha^\star$ is obtained in the first stage as a smoothed OLS estimator via kernel estimation. If certain

mild conditions on the kernel function and bandwidth are fulfilled, $\alpha^\star$ is an uniformly consistent estimator of the vector of regression functions $(\alpha_0, \ldots, \alpha_p)^\top$. In the second stage the regression processes are estimated by

$$A^\star(t) = \int_0^t (Z(s)^\top V^\star(s) Z(s))^{-1} Z(s)^\top V^\star(s)\, \mathrm{d}N(s).$$

Both Aalen's and Huffer and McKeague's estimators are under certain regularity conditions consistent and their asymptotic distributions are $(p+1)$-dimensional zero-mean Gaussian martingales. Furthermore, as shown in [3], section VIII.4.4., the Huffer and McKeague's WLS estimator is asymptotically efficient in the sense that asymptotic distribution of any other estimator satisfying certain regularity conditions cannot be more concentrated around the true value A, and therefore the WLS estimator is optimal.

In next we work with a submodel of Aalen model. First, let us suppose that we have the intercept included in the model, $x_{i,0} \equiv 1$ for all i. Second, we suppose that all the covariates $x_{i,j}$ are nonnegative. When working with an actual dataset, this can be achieved by shifting the covariates to the positive values (and keeping this adjustment in mind when interpreting the results). Finally, we assume that the regression functions $\alpha_j, j = 0, \ldots, p$, are nonnegative and we will call this model a *monotone Aalen model*. The most obvious impact of this restriction is that the cumulative regression functions are always positive valued and nondecreasing (hence they are monotone and inspiring the name *monotone* Aalen model). Furthermore, it rules out the problematic issue with non-monotonicity of the estimated survival functions when the standard Aalen model approach is used (see bottom of pg. 910 in [2]).

Another advantage is that the monotone Aalen model is more natural in interpretation of the estimated regression functions. If the model is formulated in a way that an individual with covariates $x_{i,j} = 0, j \geq 1$, represents an average healthy individual then their hazard rate is contained in the regression function α_0. The non-zero covariates account for presence of additional risk factors, such as smoking, stressful lifestyle or excess weight, contributing to the normal level. This formulation of the model can be interpreted as a competing risks model with $p+1$ cause-specific hazard functions. The overall hazard function of the competing risks model is the same as in (1) and the observed outcome is the failure due to one of the $p+1$ independent causes. Then T_i^0 would be viewed as the minimum of the $p+1$ independent life-time variables with hazard rates $\alpha_0, x_{i,1}\alpha_1, \ldots, x_{i,p}\alpha_p$. Unlike in the competing risk model we only have information on the failure (if present: $\delta_i = 1$, else $\delta_i = 0$) and we do not know which of the present risks caused the outcome. Furthermore, with the monotone Aalen model the failure can also be a result of collective additive effect of the risk factors. Hence, the statistical methods which apply well in the competing risks models cannot be used in the monotone Aalen model.

In practice, it often happens that only little data contribute to the estimation at the end of the observation window. When using the general Aalen model, it might happen that a cumulative regression function corresponding to a covariate which is expected to have a harmful effect, exhibits a distinctive decline or even runs into negative values. If the knowledge on the particular risk factor known before the study strongly antagonizes this kind of behaviour, then this undesired outcome is most likely attributed to the general instability of the estimates at the end of the observation window. The monotone

Aalen model is of a good use in case if we would like to utilize also the ending of the time window and we need a nonnegative estimator. Further advantage of the restriction imposed by the monotone Aalen model is that it can produce narrower confidence bands around the estimators as it rules out the negative values. It is though important to consider whether the assumption of nonnegativity for α_js is truly justified for the particular dataset in hand. The decision about the usage of the monotone Aalen model should be based on the beforehand knowledge of the effects of covariates on the outcome (using results of previous studies, a mechanism of the experiment, etc.). Furthermore, this decision should be done before looking in data as otherwise we might artificially increase the precision of estimators by imposing the unsubstantiated restriction on monotonicity.

The estimation in the monotone Aalen model can be done using the classic Aalen methodology. With small datasets, however, there is a risk of running into negative values, what is in conflict with the model interpretation. Obviously, for large n the consistency of these estimators is a certain guarantee of obtaining proper nonnegative estimators. Hjort and Timková analysed the monotone Aalen model using two likelihood based approaches with assumption of discontinuous cumulative regression functions (results not published yet). The likelihood of the data used in deriving the estimators was

$$\prod_{i=1}^{n}\prod_{s>0}\left\{\prod_{j=0}^{p}\{1-\mathrm{d}A_j(s)\}^{x_{i,j}(Y_i(s)-\mathrm{d}N_i(s))}\left(1-\prod_{j=0}^{p}\{1-\mathrm{d}A_j(s)\}^{x_{i,j}}\right)^{\mathrm{d}N_i(s)}\right\}.$$

First of the applied methods was the nonparametric maximum likelihood method which led to estimators for the cumulative regression functions A_j of following form

$$\int_0^t \sum_{i=1}^{n} \frac{V_{ij}(s)}{R_j(s)}\,\mathrm{d}N_i(s) + o_p(n^{-1/2}).$$

Here, $R_j(s) = \sum_{i=1}^{n} Y_i(s)x_{i,j}$ and $V_{ij}(s), i = 1,\dots,n, j = 0,\dots,p$. V_{ij} served as (random) weight functions of certain form satisfying $\sum_{j=0}^{p} V_{ij}(s) = 1$. When we sent $n \to \infty$, the estimators became closer and closer to a function of type $\int_0^t b_j(s)\,\mathrm{d}s$ where $b_j \neq \alpha_j$. The second method was Bayesian where we assumed that a priori the cumulative regression functions A_j were distributed as Beta processes, see [10]. We arrived at estimators for A_js that were of the same form as NPML estimators, only with different weight functions. Also these estimators proved to be inconsistent and they tended to different functions than the NPML estimators. The reason of inconsistency of these estimators is not clear to authors but it seems to be one of the problematic cases when even the reliable methods like NPML estimation and Bayesian approach crush if an infinite-dimensional parameter estimation is involved.

The motivation to seek a different way to estimation of regression parameters is obvious. The main objective of this paper is to conduct yet another type of Bayesian modelling. We assume that the regression functions are continuous, i. e. α_js exist, and the piecewise constant process suggested by Arjas and Gasbarra [4] is applied as priors for the regression functions. The advantage of this Bayesian approach as opposed to

the analysis with Beta processes as priors is that the method estimates the regression functions directly.

On the following pages such type of modelling is demonstrated. The method approximates the baseline hazard rate and the regression functions using piecewise constant functions with a random number and locations of jump times. In the next section the process used as prior to regression function is explained. In Section 3 the posterior distribution under Aalen model is derived and followed by explanation of the MCMC algorithm used for estimation. Section 4 is devoted to simulation study conducted to explore the performance of the method.

2. PRIOR DISTRIBUTION

Based on [4], we model the unknown regression functions $\alpha_0(t), \ldots, \alpha_p(t)$ in observed time window $[0, \tau]$, where $\tau = \max\{T_i\}$, as a correlated piecewise constant function. The values of regression function α_j are assumed to be constant within $m^{(j)} + 1$ intervals which emerge from dividing the time window $[0, \tau]$ by $m^{(j)}$ jump times $W_1^{(j)}, \ldots, W_{m^{(j)}}^{(j)}$. The value of regression function α_j within the interval $[W_{k-1}^{(j)}, W_k^{(j)})$ is denoted as $\lambda_k^{(j)}$. The number of jump times $m^{(j)}$ varies among the iterations of the Gibbs sampler through adding and deleting jumps. The regression function α_j can be expressed as a simple jump process

$$\alpha_j(t) = \sum_{k=1}^{m^{(j)}+1} I_{\{W_{k-1}^{(j)} \le t < W_k^{(j)}\}} \, \lambda_k^{(j)}, \tag{3}$$

where $W_0^{(j)} = 0$ and $W_{m^{(j)}+1}^{(j)} = \tau$. The elements of the prior distribution of each regression function $\alpha_j, j = 0, \ldots, p$ are specified as follows:

- $m^{(j)}$ jump times $W_1^{(j)}, \ldots, W_{m^{(j)}}^{(j)}$ are a realization of an inhomogeneous Poisson process with rate $\mu(t) = d \exp\{-ct\}, t \ge 0, \, c \ge 0, d > 0$
- $m^{(j)} + 1$ parameters $\lambda_1^{(j)}, \ldots, \lambda_{m^{(j)}+1}^{(j)}$ are gamma distributed random variables

$$\lambda_1^{(j)} \sim \Gamma(a_0, b_0)$$

$$\lambda_k^{(j)} \sim \Gamma(a, a/\lambda_{k-1}^{(j)}), \quad k = 2, \ldots, m^{(j)} + 1.$$

The a_0, b_0, a, c and d are the pre-specified *hyperparameters*. The convention for the Gamma distribution parametrization here is that if $X \sim \Gamma(a, b)$ then the density is $\gamma(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$ and for mean and variance we have $\mathrm{E}\, X = \frac{a}{b}$, $\mathrm{Var}\, X = \frac{a}{b^2}$. The parameters of the prior distribution for λ_ks are chosen as suggested by Arjas and Gasbarra. Obviously, the prior and hence the posterior distribution of the level $\lambda_k^{(j)}$ is dependent on the value in the previous interval $\lambda_{k-1}^{(j)}$, thus we incorporate a martingale structure into the model. It is easily seen from the properties of gamma distribution

that the conditional mean of $\lambda_k^{(j)}$ is set by the value in the previous interval

$$\mathrm{E}\,(\lambda_k^{(j)}|\,\lambda_{k-1}^{(j)}) = \frac{a}{a/\lambda_{k-1}^{(j)}} = \lambda_{k-1}^{(j)}, \quad k = 2, \ldots, m^{(j)} + 1,$$

while the conditional variation from the mean is adjusted by hyperparameter a

$$\mathrm{Var}\,(\lambda_k^{(j)}|\,\lambda_{k-1}^{(j)}) = \frac{a}{\left(a/\lambda_{k-1}^{(j)}\right)^2} = \frac{\left(\lambda_{k-1}^{(j)}\right)^2}{a}, \quad k = 2, \ldots, m^{(j)} + 1.$$

In case the hyperparameter a is small, the regression function α_j may change greatly from one interval to another, while bigger a keeps the regression function more compact and avoids huge jumps in it.

In original Arjas and Gasbarra paper [4] a homogeneous Poisson process was utilized as the prior process for jump times splitting the observation window into disjoint intervals. Here, as it will be clear from derivation in Section 3, the computational evaluation of the posterior distribution of λ_ks gets highly demanding, even intractable, within the intervals with larger amount of uncensored events (for instance > 15). To avoid the occurrence of extensive amount of observations in one interval it is wise to split the observation window more frequently in the beginning where the observations usually prevail. Hence, the inhomogeneous Poisson process with decreasing hazard rate $\mu(t) = d\exp\{-ct\}, t \geq 0, c \geq 0, d > 0$ is a natural choice for the prior distribution for jump times positions while careful setting of hyperparameters c and d allows one to control the number of jumps and their positions across the observation window. The likelihood of a realization $(W_1^{(j)}, \ldots, W_{m^{(j)}}^{(j)})$ of the Poisson process with rate $\mu(t) = d\exp\{-ct\}, t \geq 0$, such that $W_i^{(j)} < \tau, \forall i$, is equal to

$$\exp\left\{-\int_0^\tau \mu(t)\,\mathrm{d}t\right\} \prod_{i=1}^{m^{(j)}} \mu(W_i^{(j)}) = \exp\left\{-\frac{d}{c}(1 - e^{-c\tau})\right\} \prod_{i=1}^{m^{(j)}} d\exp\{-cW_i^{(j)}\}.$$

The number of the jumps contained within the time interval $[0, \tau]$ is a random variable with Poisson distribution with parameter $\frac{d}{c}(1 - e^{-c\tau})$. The number of intervals of jump functions is influenced by the choice of both hyperparameters c and d. Parameter c defines the shape of the rate function with larger values implying higher concentration of jump times close to the beginning of the observation window. Setting $c = 0$ gives a homogeneous Poisson process with rate equal to d, i. e. the jump times are spread across the observation window independently on time. Surely, the decreasing rate $\mu(t)$ is merely a recommendation based on the authors findings. There are many other possibilities of how to choose the rate $\mu(t)$, e. g. one might be particularly interested in behaviour of the regression functions in a certain part of $[0, \tau]$, hence he would choose a function with greater values within the relevant region.

Finally, the conditional prior distribution for the jth regression function α_j given the values of a_0, b_0, a, c and d is proportional to

$$\exp\left\{-\frac{d}{c}(1 - e^{-c\tau})\right\} \prod_{i=1}^{m^{(j)}} d\exp\{-cW_i^{(j)}\}\ \gamma(\lambda_1^{(j)}, a_0, b_0) \prod_{k=2}^{m^{(j)}+1} \gamma(\lambda_k^{(j)}, a, a/\lambda_{k-1}^{(j)}).$$

To obtain the posterior distribution the prior information is combined with the likelihood of the observed data which under the hazard function h_i as specified in (1) is proportional to the following formula

$$\begin{aligned} L((T_i, z_i, \delta_i), i = 1, \ldots, n) &\propto \prod_{i=1}^{n} h_i(T_i)^{\delta_i} \exp\Big\{ - \int_0^{T_i} h_i(t)\, \mathrm{d}t \Big\} \\ &= \prod_{i=1}^{n} \Big[\sum_{j=0}^{p} \alpha_j(T_i) x_{i,j} \Big]^{\delta_i} \exp\Big\{ - \int_0^{T_i} \sum_{j=0}^{p} \alpha_j(t) x_{i,j}\, \mathrm{d}t \Big\}. \end{aligned}$$

3. THE POSTERIOR DISTRIBUTION AND THE GIBBS SAMPLER

Let us denote the set of parameters determining the jump function described in previous section

$$H^{(j)} = (\lambda_1^{(j)}, W_1^{(j)}, \ldots, \lambda_{m^{(j)}}^{(j)}, W_{m^{(j)}}^{(j)}, \lambda_{m^{(j)}+1}^{(j)}), \quad j = 0, \ldots, p.$$

In every iteration of MCMC we gain a new trajectory characterized by $H^{(j)}$ for every of the regression function. When creating a new history $H^{(j)}$ for α_j we proceed sequentially by updating the pairs $(\lambda_k^{(j)}, W_k^{(j)}), k = 1, \ldots, m^{(j)}$ conditionally on the rest of parameters in $H^{(j)}$ and conditionally on current states of $\alpha_l, l \neq j$. The last interval is treated differently as we allow a change of the number of the intervals induced by either adding new jump times or discarding the last jump time if favourable for better fit. If we denote by η the number of added intervals, then altogether we have $2(m^{(j)} + \eta) + 1$ steps within every iteration for α_j. According to the MCMC methodology we provide as many iterations as necessary to reach certain stability in obtained trajectories, then we throw away several of the starting iterations (burn-in part) and use the rest to calculate a mean/median curve which represents desired estimator of the unknown regression function. This is done in pointwise fashion on a sufficiently thin net on interval $[0, \tau]$. Similarly we can obtain pointwise 95 % credibility bands for the estimator taking 0.025 and 0.975 quantile of the values in every point of the net from all the MCMC trajectories but the burn-in part. Furthermore, by using the simulated histories it is possible to approximate the posterior expectation of any integrable function of $H^{(0)}, \ldots, H^{(p)}$ with respect to the posterior distribution, as is the predictive hazard function or survival function of an individual with certain risk factors.

The sampling itself is done by Gibbs sampler with the simulation from a distribution with the density proportional to $\exp(vW_k^{(j)})$ on a bounded interval for jump times and the rejection sampling method for sampling the $\lambda_k^{(j)}$s within the intervals. The steps of the sampling are explained in detail in next pages with overall summary of the algorithm at the end of the section.

3.1. Posterior distribution of regression functions' Levels within intervals

The values of the regression functions are tied together in the likelihood of the data, however, we can derive the posterior distribution separately for a regression function, say α_j, as long as we work conditionally on all the other regression functions $\alpha_l, l \neq j$.

In particular, we will look step-by-step into every level $\lambda_k^{(j)}$ of the regression function α_j within the intervals created by the corresponding realization of the jump times. We will evaluate the posterior distribution conditionally on the jump times and the other levels of α_j and all the characteristics of all other α_ls. Hence, the part of the likelihood of the data containing the information within the examined interval is sufficient for specifying the posterior distribution of single level $\lambda_k^{(j)}$.

The posterior probability of the level $\lambda_k^{(j)}$ of regression function α_j in interval $I_k^{(j)} = [W_{k-1}^{(j)}, W_k^{(j)})$ is proportional to

$$\begin{aligned}
&p(\lambda_k^{(j)} |\ \lambda_1^{(j)}, \dots, \lambda_{k-1}^{(j)}, \lambda_{k+1}^{(j)}, \dots, \lambda_{m^{(j)}+1}^{(j)}, W_1^{(j)}, \dots, W_{m^{(j)}}^{(j)}, \{h_i\}_{i=1}^n, a_0, b_0, a, c, d, \text{data}) \\
&\quad = p(\lambda_k^{(j)} |\ \lambda_{k-1}^{(j)}, \lambda_{k+1}^{(j)}, W_{k-1}^{(j)}, W_k^{(j)}, \{(T_i, \delta_i, z_i, h_{i(-k)}) :\ T_i \geq W_{k-1}^{(j)}\}, a_0, b_0, a) \\
&\quad \propto p(\lambda_k^{(j)}, \lambda_{k+1}^{(j)}, \{(T_i, \delta_i, z_i, h_{i(-k)}) :\ T_i \geq W_{k-1}^{(j)}\} |\ \lambda_{k-1}^{(j)}, W_{k-1}^{(j)}, W_k^{(j)}, a_0, b_0, a) \\
&\quad = \gamma(\lambda_k^{(j)}; a, a/\lambda_{k-1}^{(j)}) \gamma(\lambda_{k+1}^{(j)}; a, a/\lambda_k^{(j)}) \qquad (4) \\
&\qquad\quad \times \prod_{i:T_i \in I_k^{(j)}} \Big(\lambda_k^{(j)} x_{i,j} + h_{i(-j)}(T_i)\Big)^{\delta_i} \\
&\qquad\quad \times \exp\Big\{ - \sum_{i:T_i \in I_k^{(j)}} \lambda_k^{(j)} x_{i,j} (T_i - W_{k-1}^{(j)}) - \sum_{i:T_i \geq W_k^{(j)}} \lambda_k^{(j)} x_{i,j} (W_k^{(j)} - W_{k-1}^{(j)}) \Big\}
\end{aligned}$$

where we denoted by $h_{i(-j)}(t) = h_i(t) - \alpha_j(t) x_{i,j}$ the complement of the hazard function for ith subject to term $\alpha_j(t) x_{i,j} = \lambda_k^{(j)} x_{i,j}$ (conditionally on terms in h_i from latest iteration of the MCMC simulation). Breaking down the product in the expression we get a sum of functions $\sum_{r=0}^{R} \beta_r f_r(\lambda_k^{(j)})$, where $R = \sum_{i:T_i \in I_k^{(j)}} \delta_i$,

$$\begin{aligned}
f_r(\lambda_k^{(j)}) = [\lambda_k^{(j)}]^r \gamma(\lambda_k^{(j)}; a, a/\lambda_{k-1}^{(j)}) \gamma(\lambda_{k+1}^{(j)}; a, a/\lambda_k^{(j)}) \\
\times \quad \exp\Big\{ - \sum_{i:T_i \in I_k^{(j)}} \lambda_k^{(j)} x_{i,j} (T_i - W_{k-1}^{(j)}) \\
- \sum_{i:T_i \geq W_k^{(j)}} \lambda_k^{(j)} x_{i,j} (W_k^{(j)} - W_{k-1}^{(j)}) \Big\}, \quad r = 0, \dots, R.
\end{aligned}$$

Let us have $T_1^\star, \dots, T_R^\star$ as an auxiliary notation for the set of the failure times in the interval $I_k^{(j)}$ corresponding to the subjects with jth covariate equal to $x_{1,j}^\star, \dots, x_{1,j}^\star$. Then the constants in the sum of functions are

$$\beta_r = \sum_{l_1=1}^{R} \sum_{l_2=l_1+1}^{R} \cdots \sum_{l_{R-r}=l_{R-r-1}+1}^{R} \left[\prod_{\substack{i=1 \\ i \notin \{l_1, \dots, l_{R-r}\}}}^{R} x_{i,j}^\star \right] h_{l_1(-j)}(T_{l_1}^\star) \cdots h_{l_{R-r}(-j)}(T_{l_{R-r}}^\star). \qquad (5)$$

This case of distribution can be viewed as a mixture of distributions proportional to f_r weighted by factors β_r. In particular, note that every term $\beta_r f_r$ represents a case, when r

individuals of total R individuals who failed in interval $[W_{k-1}^{(j)}, W_k^{(j)})$, died because of the risk imposed by factor $\alpha_j(T_i)x_{i,j}$ while the rest $R-r$ individuals died of any other factor $h_{i(-j)}(T_i) = h_i(T_i) - \alpha_j(T_i)x_{i,j}$. This corresponds to aforementioned interpretation of the monotone Aalen model when all the covariates in the model represent an additional risk of death to the baseline risk α_0 while every of the covariates increases the probability of failure, however, only one causes the death. Generating a sample from this kind of distribution can be done using classical approaches to mixtures of distributions. First we calculate the weights $w_r = \beta_r / \sum_{s=0}^{R} \beta_s$ and then we generate a sample from $U[0,1]$. If the sampled value falls in the interval $[\sum_{s=0}^{r-1} w_s, \sum_{s=0}^{r} w_s)$ then we sample from the distribution proportional to f_r.

The simulation from the distribution proportional to function f_r is done similarly as in Arjas and Gasbarra's work in [4]. Assuming $\xi > 0$ we could rewrite the function f_r in following form

$$f_r(\lambda_k^{(j)}) = d_{r,\xi}(\lambda_k^{(j)}) g_{r,\xi}(\lambda_k^{(j)})$$

where

$$\begin{aligned} d_{r,\xi}(\lambda) &= \lambda^{\xi+r-1} \exp\Big\{ -\lambda \Big[\frac{a}{\lambda_{k-1}^{(j)}} + \sum_{i:T_i \in I_k^{(j)}} x_{i,j}(T_i - W_{k-1}^{(j)}) \\ &\qquad + \sum_{i:T_i \geq W_k^{(j)}} x_{i,j}(W_k^{(j)} - W_{k-1}^{(j)}) \Big] \Big\} \\ g_{r,\xi}(\lambda) &= \frac{1}{\lambda^{\xi}} \exp\Big\{ -\frac{1}{\lambda} a \lambda_{k+1}^{(j)} \Big\}. \end{aligned}$$

The first function $d_{r,\xi}(\cdot)$ is the probability density function of gamma distribution $\Gamma(\xi + r, \frac{a}{\lambda_{k-1}^{(j)}} + \sum_{i:T_i \in I_k^{(j)}} x_{i,j}(T_i - W_{k-1}^{(j)}) + \sum_{i:T_i \geq W_k^{(j)}} x_{i,j}(W_k^{(j)} - W_{k-1}^{(j)}))$. The function $g_{r,\xi}(\cdot)$ is the density of the distribution known as the inverse-gamma distribution with parameters $\xi + 1$ and $a\lambda_{k+1}^{(j)}$.

As the following holds

$$d_{r,\xi}(\lambda) g_{r,\xi}(\lambda) \leq d_{r,\xi}(\lambda) \max_{\lambda} g_{r,\xi}(\lambda) = d_{r,\xi}(\lambda) g_{r,\xi}(a\lambda_{k+1}^{(j)}/\xi),$$

the rejection sampling method may be directly applied. All we need is to simply sample from gamma distribution with density $d_{r,\xi}$ as long as necessary to reach the acceptance. To increase the probability of acceptance we set the value of ξ to let the modes of both $d_{r,\xi}$ and $g_{r,\xi}$ equal. This is guaranteed when ξ satisfies following equation:

$$\frac{\xi + r - 1}{\frac{a}{\lambda_{k-1}^{(j)}} + \sum_{i:T_i \in I_k^{(j)}} x_{i,j}(T_i - W_{k-1}^{(j)}) + \sum_{i:T_i \geq W_k^{(j)}} x_{i,j}(W_k^{(j)} - W_{k-1}^{(j)})} = \frac{a\lambda_{k+1}^{(j)}}{\xi}.$$

The special case is the simulation of $\lambda_{m^{(j)}+1}^{(j)}$ in the very last interval $I_{m^{(j)}+1}^{(j)} = [W_{m^{(j)}}^{(j)}, \tau)$. The value of $\lambda_{m^{(j)}+1}$ no more influences any subsequent level of the hazard

function and therefore the posterior distribution for $\lambda_{m^{(j)}+1}$ simplifies to a mixture of gamma distributions, symbolically written as

$$\sum_{r=0}^{R} \beta_r \; \gamma\Big(a+r,\; a/\lambda^{(j)}_{m^{(j)}} + \sum_{i:T_i \in I^{(j)}_{m^{(j)}+1}} x_{i,j}(T_i - W^{(j)}_{m^{(j)}})\Big) \tag{6}$$

where β_r is as in (5) and again R being total of observed deaths in $I^{(j)}_{m^{(j)}+1}$.

It is typical with lifetime distribution that the incidents are clustered in the beginning of the observation window. However, if lots of observations fall into the examined interval, the evaluation of the weighting coefficients $\beta_r, r = 0, \ldots, R$ becomes a serious computational problem, as we need to consider every r-combination of total R observations within the interval. This is exactly $\binom{R}{r}$ possibilities of what caused the deaths occurred within the examined time interval: either the actual $\alpha_j(\cdot)x_{i,j}$ or the complementary $h_{i(-j)}(\cdot)$. However, the number of all r-combinations, $r = 0, \ldots, R$, equals to 2^R and while for $R = 10$ we have 1024 options to explore, for $R = 15$ we get up to circa $3 \cdot 10^5$ combinations. A feasible approximation to calculate the β_rs is in need. One of the options is for every r such that it produces larger number of combinations than a pre-fixed number then instead of using all the combinations in the evaluation of β_r we would randomly choose only L combinations, where $L \ll \binom{R}{r}$. To get the proportionally equal number it is necessary to multiply the obtained number by ratio $\binom{R}{r}/L$. Choice of the value for L is a question of balance of precision of the evaluation on the one hand and computational feasibility on the other hand.

3.2. Posterior distribution of jump times

The posterior distribution for the particular jump time $W^{(j)}_k$ in the level of the regression function α_j is again determined only by these parts of the likelihood and prior information which are affected by $W^{(j)}_k$ itself. The posterior probability of jump time $W^{(j)}_k$ can be written as

$$\begin{aligned}
&p(W^{(j)}_k |\; W^{(j)}_1, \ldots, W^{(j)}_{k-1}, W^{(j)}_{k+1}, \ldots, W^{(j)}_{m^{(j)}}, \lambda^{(j)}_1, \ldots, \lambda^{(j)}_{m^{(j)}+1}, \{h_i\}_{i=1}^n, a_0, b_0, a, c, d, \text{data}) \\
&\quad = \; p(W^{(j)}_k |\; W^{(j)}_{k-1}, W^{(j)}_{k+1}, \lambda^{(j)}_k, \lambda^{(j)}_{k+1}, \{(T_i, \delta_i, z_i, h_i) :\; T_i \geq W^{(j)}_{k-1}\}, c, d) \\
&\quad \propto \; p(W^{(j)}_k,\; W^{(j)}_{k-1}, W^{(j)}_{k+1}, \lambda^{(j)}_k, \lambda^{(j)}_{k+1}, \{(T_i, \delta_i, z_i, h_i) :\; T_i \geq W^{(j)}_{k-1}\}, c, d) \\
&\quad \propto \; d \exp\{-cW^{(j)}_k\} \prod_{i:T_i \in I^{(j)}_k} h_i(T_i)^{\delta_i} \prod_{l:T_l \in I^{(j)}_{k+1}} h_l(T_l)^{\delta_l} \qquad (7) \\
&\qquad \times \; \exp\Big\{ - \sum_{i:T_i \geq W^{(j)}_{k+1}} \Big[\lambda^{(j)}_k x_{i,j}(W^{(j)}_k - W^{(j)}_{k-1}) + \lambda^{(j)}_{k+1} x_{i,j}(W^{(j)}_{k+1} - W^{(j)}_k)\Big] \\
&\qquad\qquad - \sum_{i:T_i \in I^{(j)}_{k+1}} \Big[\lambda^{(j)}_k x_{i,j}(W^{(j)}_k - W^{(j)}_{k-1}) + \lambda^{(j)}_{k+1} x_{i,j}(T_i - W^{(j)}_k)\Big] \\
&\qquad\qquad - \sum_{i:T_i \in I^{(j)}_k} \lambda^{(j)}_k x_{i,j}(T_i - W^{(j)}_{k-1})\Big\}.
\end{aligned}$$

The expression is in core similar to the result of Arjas and Gasbarra [4]. In the examined interval the posterior distribution is between the observation times proportional to $u\exp(vW_k^{(j)})$. The sample for a new jump position can be generated from this piecewise continuous distribution for example by using inverse sampling. A special case is when we update the last jump time $W_{m^{(j)}}^{(j)}$ where the simulation is on $[W_{m^{(j)}-1}^{(j)}, \infty)$ and the probability of a jump falling out of $[W_{m^{(j)}-1}^{(j)}, \tau)$ is proportional to

$$\prod_{i:T_i\in[W_{m^{(j)}-1}^{(j)},\tau]} h_i(T_i)^{\delta_i} \exp\Big\{ - \sum_{i:T_i=\tau} \lambda_{m^{(j)}}^{(j)} x_{i,j}(\tau - W_{m^{(j)}-1}^{(j)}) - \sum_{i:T_i\in[W_{m^{(j)}-1}^{(j)},\tau)} \lambda_{m^{(j)}}^{(j)} x_{i,j}(T_i - W_{m^{(j)}-1}^{(j)})\Big\}.$$

If an updated jump is generated outside the window $[W_{m^{(j)}-1}^{(j)}, \tau)$, this jump is simply discarded and the iteration is ended. However, if this updated jump $W_{m^{(j)}}^{(j)} < \tau$ then we try to sample another new jump $W_{m^{(j)}+1}^{(j)}$ on the interval $[W_{m^{(j)}}^{(j)}, \tau)$ and if this jump falls into the observation window we keep it and instead of $[W_{m^{(j)}}^{(j)}, \tau)$ we introduce two intervals $[W_{m^{(j)}}^{(j)}, W_{m^{(j)}+1}^{(j)})$ and $[W_{m^{(j)}+1}^{(j)}, \tau)$ into the sets of the intervals. We set $m^{(j)} \leftarrow m^{(j)} + 1$ and sample value $\lambda_{m^{(j)}+1}^{(j)}$ for the newly created interval at the end of the observation window. Summed up, in one iteration we either add one or more new jumps into the estimator or we erase one jump. For detailed explanation of the algorithm see pp. 512–513 in Arjas and Gasbarra [4].

Another option is to use the Metropolis–Hasting algorithm. Let us denote the conditional posterior distribution of $W_k^{(j)}$ from (7) with $p^{post}(W_k^{(j)})$. As the proposal density we may consider the density of the uniform distribution on interval $[W_{k-1}^{(j)}, W_{k+1}^{(j)})$. Then the proposal acceptance density of new jump time located in W^{new} equals to

$$\alpha^{post}(W_k^{(j)}, W^{new}) = \min\left\{1, \frac{p^{post}(W^{new})}{p^{post}(W_k^{(j)})}\right\}.$$

Apart from sampling new positions of jump times from posterior distribution we would like to allow adding a new jump into the last interval or deleting the very last jump $W_{m^{(j)}}^{(j)}$. The problem of adding/discarding of a jump can be formulated as birth and death, i. e. a special case of reversible jump problem (for details see e. g. [9]). The set of jump times represents the finite point process within the interval $[0, \tau]$ with the density (proportional to the posterior density of jump time) with respect to the unit intensity Poisson process. Hence we may adopt the birth-death Metropolis–Hastings algorithm to provide desired steps of adding or deleting particular jump times.

Now let U be the total number of the iterations of the Gibbs sampler and let us denote

$$H^{(j)}(u) = \left(\lambda_1^{(j)}(u), W_1^{(j)}(u), \ldots, \lambda_{m^{(j)}(u)}(u), W_{m^{(j)}(u)}^{(j)}(u), \lambda_{m^{(j)}(u)+1}^{(j)}(u)\right),\ u = 0, \ldots, U,$$

the uth member of the Markov chain $\{H^{(j)}(u)\}_{u=0}^{U}$ generated in uth iteration of the Gibbs sampler. The chain $\{H^{(j)}(u)\}_{u=0}^{U}$ corresponds to jth regression function $\alpha_j, j = 0, \ldots, p$. The steps of the algorithm can be summarized as follows:

Sampling algorithm:

- generate a starting trajectory $H^{(j)}(0)$ for $j = 1, \ldots, p$ from the prior distribution; $m^{(j)}(0)$ let be the random number of jumps which comes from the inhomogeneous Poisson process simulation of jump times
- for uth iteration, where $u \in \{1, \ldots, U\}$, do
  - for $j = 1, \ldots, p$ do
    1. set $m^{(j)}(u) \leftarrow m^{(j)}(u-1)$,
    2. $k \leftarrow 1$,
    3. sample $\lambda_k^{(j)}(u)$ from posterior distribution in (4) (sampling from the mixture distribution),
    4. sample $W_k^{(j)}(u)$ from posterior distribution in (7), $k \leftarrow k+1$
    5. repeat steps 3. and 4. until $k = m^{(j)}(u)$,
    6. sample $\lambda_{m^{(j)}(u)}^{(j)}(u)$ from posterior distribution in (4),
    7. sample $W_{m^{(j)}(u)}^{(j)}(u)$ from posterior distribution in (7), if $W_{m^{(j)}(u)}^{(j)}(u) \geq \tau$ then discard it, else set $m^{(j)}(u) \leftarrow m^{(j)}(u)+1$ and repeat steps 6 and 7,
    8. sample $\lambda_{m^{(j)}(u)+1}^{(j)}(u)$ from posterior distribution in (6).

The problem of ergodicity of every component $H^{(j)}(u)$ of the resulting Markov chain is similar to the original Arjas and Gasbarra's method as long as the other components $H^{(k)}(u), k \neq j$ are held fixed. If the birth-death Metropolis–Hastings algorithm is used for simulation of new jumps, the proposal density and the acceptance probability of adding/discarding a new jump needs to be specified in the manner which allows for the detailed balance condition to be fulfilled. The ergodicity is then ensured similarly as with standard Hastings algorithms. More details on the ergodicity and proper specification of the acceptance probability when switching between the subspaces can be found in [9].

4. SIMULATIONS

The posterior distribution of the method proposed in this paper is of rather complicated structure not allowing us to gain straightforward asymptotic features. It estimates the functional parameters or any integrable function of these parameters by approximating the posterior expectation, in fact by averaging a set of jump functions, each with a finite number of jumps. These jumps are not fixed through the iterations, hence the method provides us with an estimator resembling a continuous function. The choice of hyperparameters and no functional restriction allows for very flexible estimation. These features are the assets of the method, however, to assess the performance of the obtained estimators we have to rely on the aid of simulation techniques. The method was tested

on 300 datasets sampled from a model $h_i(t) = \alpha_0(t) + \alpha_1(t)x_{i,1} + \alpha_2(t)x_{i,2}$ of the hazard rate on interval $[0, 1]$ with regression functions equal to

$$\begin{aligned} \alpha_1(t) &= \sin(\pi t) + 1.5, \\ \alpha_2(t) &= \exp(-3t) + 1, \end{aligned}$$

and the baseline hazard rate $\alpha_0(t)$ was chosen to be a piecewise constant function with jumps in $(0.2, 0.35, 0.6, 0.7, 0.9)$ and values $(0.8, 2.2, 3, 0.9, 1.5, 2)$. The time-constant covariates were sampled randomly for every dataset from gamma distributions with parameters $\Gamma(2, 2)$ and $\Gamma(1, 2)$ for $x_{i,1}$ and $x_{i,2}$, respectively. We have chosen various shapes of the regression functions to compare how well different functions can be approximated by the proposed method. We estimated the regression functions under two different priors

$$\begin{aligned} &\text{PRIOR 1: } a_0 = 0.1,\ b_0 = 0.1,\ a = 0.5,\ c = 1,\ d = 25, \\ &\text{PRIOR 2: } a_0 = 0.1,\ b_0 = 0.1,\ a = 0.2,\ c = 0.5,\ d = 35. \end{aligned}$$

The parameters of prior 1 was chosen to produce jump functions with smaller variations from one level to another and less intervals while smaller a in prior 2 allowed for greater variability. The number of the jump times on $[0, 1]$ is a priori Poisson distributed with mean approximately equal to 16 and 28 for prior 1 and prior 2, respectively. When choosing the parameters for the prior distribution one should consider how much flexibility he or she requires. Due to the computational issues it is advisable to choose the parameters c and d so that the average number of jumps will not be a priori smaller than say the number of observations divided by 15. Also the parameters a_0 and b_0 defining the first level should allow for a great variability. The number of observations was $n = 25$, 50 and 80 and we generated 100 datasets for every n. The observations were independently right-censored with non-censoring rate equal to ≈ 0.8. If a generated failure time fell out of the interval $[0, 1]$, it was right-censored at time 1. For every dataset we calculated the estimators based on both PRIOR 1 and PRIOR 2. The expectations of the posterior distribution for regression functions were approximated by pointwise averages of members of gained Markov chains $H^{(j)}, j = 1, 2, 3$ after discarding the first 100 from total $U = 500$ iterations of the Gibbs sampler. The sampling of the levels $\lambda_j^{(k)}$ within intervals was done by rejection sampling as described in Section 3.1 and the jump times were generated from the piecewise continuous density proportional to $u \exp(vW_k^{(j)})$ using inverse sampling. Alongside Aalen's least squares estimators were calculated with 95 % pointwise confidence bands.

The choice of the burn-in and the total number of members of the Markov chain was based on the study of the MCMC traces in various time points in interval $[0, 1]$ (not displayed here). Most of the traces showed certain stability after 50 to 100 iterations. Furthermore, as explained at the end of Section 3.1, if there are more failures present within an interval, the computational demands become huge. We evaluated the parameters of the mixture distribution in exact fashion whenever the number of the failures within the interval was smaller than 11. In case that the total of the failures exceeded this number, we used the approximation suggested at the end of Section 3.1

	n		A_0 *Bayes*	A_0 *Aalen*	A_1 *Bayes*	A_1 *Aalen*	A_2 *Bayes*	A_2 *Aalen*
PRIOR 1	25	BIAS	0.013	-0.001	0.001	0.005	0.001	0.002
		MSE	0.001	0.017	0	0.016	0	0.028
		MAE	0.013	0.038	0.004	0.035	0.005	0.046
		Sup	0.131	0.44	0.051	0.431	0.054	0.598
		Surface	0.066	0.165	0.069	0.153	0.094	0.206
	50	BIAS	0.011	-0.004	-0.001	0.004	-0.001	-0.001
		MSE	0.001	0.005	0	0.005	0	0.011
		MAE	0.011	0.021	0.004	0.021	0.004	0.032
		Sup	0.114	0.273	0.054	0.291	0.05	0.413
		Surface	0.062	0.107	0.063	0.101	0.091	0.144
	80	BIAS	0.009	-0.003	-0.002	0.001	-0.002	-0.001
		MSE	0.001	0.004	0	0.003	0	0.006
		MAE	0.01	0.019	0.004	0.016	0.005	0.024
		Sup	0.105	0.229	0.056	0.214	0.051	0.312
		Surface	0.056	0.084	0.059	0.076	0.087	0.109
PRIOR 2	25	BIAS	0.013	-0.001	0.001	0.005	0.001	0.002
		MSE	0.002	0.017	0.003	0.016	0.004	0.028
		MAE	0.013	0.038	0.004	0.035	0.005	0.046
		Sup	0.131	0.44	0.051	0.431	0.055	0.598
		Surface	0.066	0.165	0.069	0.153	0.094	0.206
	50	BIAS	0.011	-0.004	-0.001	0.004	-0.001	-0.001
		MSE	0.003	0.005	0.001	0.005	0.002	0.011
		MAE	0.011	0.021	0.005	0.021	0.004	0.032
		Sup	0.116	0.273	0.056	0.291	0.058	0.413
		Surface	0.062	0.107	0.064	0.101	0.091	0.144
	80	BIAS	0.009	-0.003	-0.002	0.001	-0.003	-0.001
		MSE	0.002	0.004	0.001	0.003	0.002	0.006
		MAE	0.015	0.019	0.006	0.016	0.005	0.024
		Sup	0.105	0.229	0.058	0.214	0.053	0.312
		Surface	0.059	0.084	0.062	0.076	0.091	0.109

Tab. 1. Results of simulation study: average values of measures of precision calculated from 100 instances for every prior and every number of observations per dataset $n = 25, 50$ and 80. Statistics for Aalen estimators were calculated alongside.

with $L = 20$. In addition we conducted a thorough study of the efficiency of the approximation in a simple setting with no jumps and only one interval. We generated 1000 of datasets with 10 uncensored observations, calculated the exact posterior distributions (which are mixtures of exactly gamma distributions) and calculated the approximated posterior using $L = 20$ and $L = 50$. The posterior distributions and estimators obtained in the exact and approximated manner were in great agreement. The approximation using $L = 50$ did not show much improvement over the one with $L = 20$. The mean computation time needed for the calculation of the estimators for a simulated dataset with $n = 10$ was on average 43.3 seconds for the exact derivation of the posterior and 34.5 seconds for the approximated posteriors with $L = 20$. The difference in the time cost needed for evaluation of the exact and approximated posterior distribution will be more apparent once the number of observations within the intervals is greater than 15.

		A_0			A_1			A_2		
		Bayes		*Aalen*	*Bayes*		*Aalen*	*Bayes*		*Aalen*
		95%	99%	95%	95%	99%	95%	95%	99%	95%
PRIOR 1	n = 25	0.97	0.97	0.74	0.99	0.99	0.63	0.99	0.99	0.52
	n = 50	0.96	0.98	0.82	0.97	0.97	0.43	0.96	0.97	0.49
	n = 80	0.93	0.93	0.66	0.92	0.93	0.4	0.98	1	0.37
PRIOR 2	n = 25	0.97	0.97	0.74	0.99	0.99	0.63	0.99	0.99	0.52
	n = 50	0.97	0.98	0.82	0.97	0.97	0.43	0.96	0.98	0.49
	n = 80	0.94	0.94	0.66	0.92	0.92	0.4	0.99	1	0.37

Tab. 2. Results of simulation study: average values of coverage of 95 % and 99 % pointwise credibility bands for Bayesian estimation and 95 % pointwise confidence bands for Aalen. The values are calculated from 100 instances for every prior and every number of observations per dataset $n = 25$, 50 and 80.

We considered several measures of precision of both Bayesian and Aalen estimators, in detail the *functional BIAS*

$$BIAS(\hat{A}_j) = \int_0^{\tau^\star} \left(\hat{A}_j(t) - A_j(t)\right) \mathrm{d}t,$$

and analogically calculated *functional MSE*, *functional mean absolute error (MAE)*, *supremum* of the absolute differences between real and estimated regression functions and *surface*. The last characteristic is the surface of the area contained between 95 % pointwise credibility/confidence bands. The integrals were approximated by summation on a thin net of 100 time points within the interval $[0, \tau^\star]$. We chose the right end $\tau^\star$ so that the interval $[0, \tau^\star]$ represented the part of the whole interval $[0, 1]$ where in all instances the Aalen estimators were calculated. The minimal value of $\tau^\star$ for all datasets was equal to 0.17. The estimators proposed in this paper are able to estimate the unknown parameters on the whole observation window $[0, \tau]$ but similarly as the classic Aalen estimators they suffer from great instability at the end where few observations appear. Therefore we decided to evaluate the statistics only on the interval with enough observations in hand, where both Aalen and Bayesian estimators are stable. The average values of the statistics are displayed in Table 1. Further we examined the coverage of the pointwise credibility/confidence bands for Bayesian estimation and Aalen estimators on $[0, \tau^\star]$. The coverage was calculated as the proportion of the datasets where the true cumulative regression function was contained within the pointwise credibility/confidence bands (again, evaluated on a thin grid on $[0, \tau^\star]$). See Table 2 for the results.

From the results in Table 1 it is obvious that the Bayesian estimators in comparison to standard least squares Aalen estimators can suffer from greater functional BIAS, see in particular the estimator of A_0. Overall, the average of the functional BIAS of the Bayesian estimators does not suggest any discrepancy from the consistency, as also for A_0 it has a decreasing tendency for both priors. Interestingly, the proposed Bayesian estimators have consistently smaller functional MSE, functional MAE and supremum of the differences, suggesting that while on average these estimators might be for some regression functions less precise, the variation from the true value of the

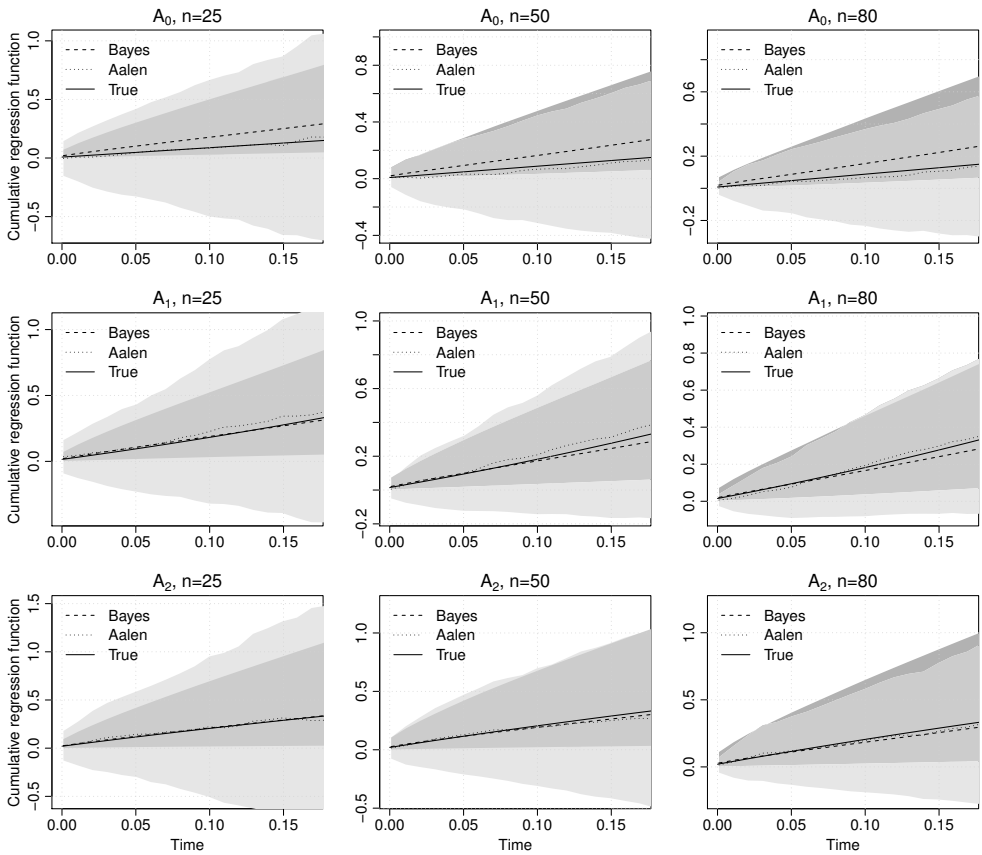

Fig. 1. Graphs of the pointwise averages of the estimators obtained from 100 repetitions for prior 1 and numbers of observations $n = 25$, 50 and 80. The true regression function is plotted in solid line, average of the Bayesian estimators in dashed line and average of the Aalen's estimators in dotted line. The average pointwise credibility/confidence bands are included: Bayesian credibility bands in dark gray and Aalen's confidence bands in light gray.

sought parameter is fairly small. Further, from the coverage results in Table 2 we see that the incidences when the real regression function is contained in the 95 % pointwise credibility bands on the shortened interval $[0, \tau^\star]$ varies from 92 % to 99 % of all cases, while the coverage of pointwise 95 % confidence bands based on the Aalen estimators was a lot worse with 40 % to 82 %. Obviously, these bands are pointwise, hence, they are not expected to fulfil the required 95 % coverage. Another appealing result is that the surface of the estimated credibility bands is for smaller datasets ($n = 25$) about half of the area contained within the Aalen pointwise confidence bands. With growing number of observations the surface of the Aalen pointwise confidence bands is rapidly

decreasing, however, not in the single case it reached the average surface of the Bayesian credibility bands. It is to be expected, though, that for datasets with several hundreds of observations the Aalen confidence bands would by narrower than the Bayesian credibility bands. This follows in the first place from the consistency of the Aalen estimators, secondly it is suggested by the rate of decline of the averaged surfaces of Aalen confidence bands displayed in the simulations in comparison to the Bayesian credibility bands.

The fact that the characteristics describing the variability of the estimators are smaller for Bayesian estimators than for the Aalen's least squares estimators is not a great surprise, though. The Bayesian estimators work with more information from the start as they are restricted to the positive values, while Aalen estimators span the whole real line at every time point.

The outcome from both priors is rather similar, with slightly larger values of characteristics on variation from the true values, what was expected from the setting of the parameters of the priors.

For illustration we included graphs of the pointwise averages of the estimators obtained from 100 repetitions for every size of dataset and for PRIOR 1, see Figure 1. The true regression functions are plotted in solid line, the averages of the proposed Bayesian estimators are in dashed line and the averages of the Aalen estimators are in dotted line. When looking at the graphs, the BIAS of the Bayesian estimator of A_0 from the true value is apparent. We added average pointwise credibility bands for Bayesian estimators (dark gray area) and confidence bands for Aalen estimators (light gray area) into the graphs. It can be seen that for small datasets (n=25) the classic Aalen's estimation and the proposed Bayesian solution on average produce similar estimators. The average credibility bands of the Bayesian estimator are a lot slimmer than the average Aalen estimators' confidence bands, i. e. the graphs support the results on smaller variation of proposed estimators from the true value. When looking only at the part with positive values, the Aalen confidence bands and Bayesian credibility bands take almost similar surface. With growing number of observations Aalen estimators apparently exhibit better fit. The graphs based on the results obtained from PRIOR 2 show the same trend and are not displayed here.

As pointed out by one of the referees, the high values (between 92 % and 99 %) of the simultaneous coverage of the pointwise 95 % credibility bands of the Bayesian estimators are rather curious in comparison to the coverage of the Aalen estimators. Indeed, if the pointwise coverage of the pointwise 95 % credibility bands were evaluated instead, the coverage would be even higher (close to 1 in most cases). There is no exact explanation for this phenomenon, perhaps just the smoothness of the Bayesian estimators in comparison to the variability of the Aalen least squares estimators could enhance the coverage. Furthermore, we could have had a look at the behaviour of the estimators after the $\tau^\star = 0.17$ to assess the closeness of the fit to the real regression functions in later times. The reason why it was not done is that the focus was on the part of the time interval where both Aalen least squares and Bayesian method provide a good estimation based on enough data. A differently designed simulation model with more data available in later times would be useful in this kind of study.

The estimation was conducted in program R version 3.0.2 on 64-bit Ubuntu 13.04 and on a computer with Intel Core i5-3470 CPU 3.20 GHz × 4 and 3.8 GiB RAM. The

average time of the computations was the same for both priors and it was about 3, 9 and 20 minutes for the total number of observations $n = 25$, 50, 80, respectively. It was observed, that the number of observations and choice of hyperparameters have the biggest impact on the computational time. The most crucial is the approximation of the mixture weights β_r.

5. DISCUSSION

The paper is devoted to deriving alternative estimators to least squares estimation in Aalen model with focus on monotonicity of sought cumulative regression functions A_js. There has been little work done in this direction as most interest in survival analysis is directed to the popular Cox model. The previous joint work of Hjort and the author revealed that the Bayesian inference using NII processes, as it was done in the case of homogeneous hazard function and in the Cox model by Kim, De Blasi, Hjort and others, [6, 13, 14] and [7], has proven to be inconsistent. Similar unsatisfactory result was obtained by using the maximum likelihood nonparametric estimation under assumption of discontinuity of the regression functions.

The estimation proposed in this paper was taken down a slightly different path as we worked with likelihood based on continuous regression functions. We proposed a sensible prior distribution with a martingale structure based on Arjas and Gasbarra's work [4]. We derived the posterior distribution for the parameters of the model and proposed a sampling machine for generation of the estimators via Gibbs sampling. The performance of the method was tested in the simulation study. The focus was in particular on the consistency of the estimators, which was lacking in previous work by the author. The method provides one with the estimators of the regression functions α_j directly, however, as the intention was to compare the features of the proposed Bayesian estimators with Aalen least squares estimation, the cumulative regression functions were assessed in the simulation study. It should be also noted, that the estimators of the cumulative versions are more stable that the noncumulative ones, hence, they might be preferred. On average the estimators of α_j are very good estimators, however, they are more sensitive to the choice of the prior parameters.

The results of the simulations suggest certain tendency of the Bayesian estimators towards the real values, but with a lot slower pace than the standard Aalen least squares estimators. The apparent advantage of the Bayesian estimators lies in the values of functional MSE and MAE and in the coverage performance of the pointwise credibility bands. The obtained numbers suggest that the proposed Bayesian estimators can be of better use with small sized datasets where the least squares estimation can be unstable and suffer from great variation. Furthermore, Aalen estimator can run into the negative values while we would like to abide by the monotonicity condition. Another advantage is that we can obtain the estimators of the α_js directly instead of the cumulative versions and these estimators are close to continuous functions. Apart from the possible bias, the main disadvantage of this method is certainly concentrated in the computational demands as well as the need for careful choice of the hyperparameters. Hence, for datasets with greater number of observations, it is recommended to reach for the classic Aalen or Huffer and McKeague estimation where the consistency is assured and computational demands are less overwhelming.

We will conclude this work by suggesting a few possibilities for a future work in this direction. It is clear that a greater simulation study is needed to obtain a closure about the consistency of the proposed estimators. Furthermore, there are several straightforward extensions of the proposed method into more complicated scenarios. First, it could be applied on data with recurrent events with only minor changes in the posterior distribution. Secondly, if we considered a prior distribution for $\lambda_k^{(j)}$ which would be contained on the whole real line, it would lead us to Bayesian analysis of the classic Aalen model. The other possibility is to create a hierarchical model by imposing a prior distribution on the parameters a_0, b_0, a, c and d instead of the fixed values. Both recurrent events and hierarchical model are employed in previous work of the author, see [16]. An option in the frequentist framework would be developing a monotone alternative to Aalen's and Huffer–McKeague estimators by using restricted least squares estimation with constraint that the increment $\Delta A_j \geq 0$. And yet another idea is to assume that α_j are piecewise constant functions with fixed numbers of equidistant jumps and estimate the values of these functions via maximum likelihood method.

(Received October 14, 2013)

REFERENCES

[1] O. O. Aalen: A model for nonparametric regression analysis of counting processes. Springer Lect. Notes in Statist. *2* (1980), 1–25.

[2] O. O. Aalen: A linear regression model for the analysis of life times. Statist. Med. *8* (1989), 907–925.

[3] P. K. Andersen, A. Borgan, R. D. Gill, and N. Kieding: Statistical Models Based on Counting Processes. Springer, New York 1993.

[4] E. Arjas and D. Gasbarra: Nonparametric bayesian inference from right censored survival data, using Gibbs sampler. Statist. Sinica *4* (1994), 505–524.

[5] D. R. Cox: Regression models and life-tables. J. Roy. Statist. Soc. *34* (1972), 2, 187–220.

[6] P. De Blasi and N. L. Hjort: Bayesian survival analysis in proportional hazard models with logistic relative risk. Scand. J. Statist. *34* (2007), 229–257.

[7] P. De Blasi and N. L. Hjort: The Bernstein–von Mises theorem in semiparametric competing risks models. J. Statist. Planning Inf. *34* (2009), 1678–1700.

[8] K. Doksum: Tailfree and neutral random probabilities and their posterior distributions. Ann. Statist. *2* (2006), 183–201.

[9] P. J. Green: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika *82* (1995), 711–732.

[10] N. L. Hjort: Nonparametric Bayes estimators based on beta processes in models for Life history data. Ann. Stat. *3* (1990), 1259 – 1294.

[11] F. W. Huffer and I. W. McKeague: Weighted least squares estimation for Aalen's additive risk model. J. Amer. Statist. Assoc. *86* (1991), 114–129.

[12] E. L. Kaplan and P. Meier: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. *53* (1958), 457–481.

[13] Y. Kim: The Bernstein-von Mises theorem for the proportional hazard model. Ann. Statist. *34* (2006), 1678–1700.

[14] Y. Kim and J. Lee: A Bernstein-von Mises theorem in the nonparametric right-censoring model. Ann. Statist. *32* (2004), 1492–1512.

[15] D. Sinha and K. D. Dipak: Semiparametric Bayesian analysis of survival data. J. Amer. Statist. Assoc. *92* (1997), 1195–1212.

[16] J. Timková: Bayesian nonparametric estimation of hazard rate in survival analysis using Gibbs sampler. In: Proc. WDS 2008, Part I: Mathematics and Computer Sciences, pp. 80–87.

Jana Timková, ÚTIA, Institute of Information Theory and Automation — Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: j.timkova@gmail.com