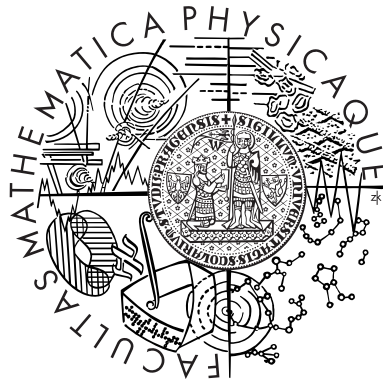


BAYESIAN AND MAXIMUM LIKELIHOOD  
NONPARAMETRIC ESTIMATION IN  
MONOTONE AALEN MODEL

JANA TIMKOVÁ



A DISSERTATION  
PRESENTED TO THE FACULTY OF MATHEMATICS AND PHYSICS  
OF CHARLES UNIVERSITY  
IN CANDIDACY FOR THE DEGREE  
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE  
BY THE DEPARTMENT OF  
PROBABILITY AND MATHEMATICAL STATISTICS  
ADVISOR: DOC. PETR VOLF, CSc.

PRAGUE 2014

# Abstract

This work is devoted to seeking methods for analysis of survival data with the Aalen model under special circumstances. We suppose, that all regression functions and all covariates of the observed individuals are nonnegative. Hence, every covariate is expected to impose an additional risk to the baseline hazard rate. As opposed to the classic Aalen model, the cumulative regression functions are always nondecreasing. We named this special case the *monotone Aalen model*. The main objective was to establish solid methods for estimation of the unknown functional parameters of the hazard rate. Three methods are presented in this work.

First, we considered two likelihood based approaches with assumption of discontinuous cumulative regression functions, namely the nonparametric maximum likelihood method and the Bayesian analysis using Beta processes as the priors for the unknown cumulative regression functions. Both methods led to well defined estimators. Study of their large-sample properties showed that under general conditions both nonparametric likelihood estimators and the Bayesian estimators were inconsistent. The consistency was found only with one-covariate model without a baseline hazard function.

In the third method the Bayesian framework was utilized again. As opposes to the previous, here the analysis was based on the assumption of continuity of the hazard rate. This allowed to estimate the regression functions directly instead of estimating their cumulative versions. The method relied on the correlated prior approach where the regression functions were supposed to be jump processes with a martingale structure. Due to the complexity of the posterior distribution, the evaluation of the estimators was obtained via sample paths generated by Gibbs sample. The performance, especially the consistency, was assessed by the simulation study.

The three methods were demonstrated on two real datasets: the famous Danish malignant melanoma data and a never before analysed dataset on delay times of patients with myocardial infarction.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ..... signature of the author

Název práce: Bayesovské odhady a odhady metodou maximální věrohodnosti v monotonním Aalenově modelu

Autor: Jana Timková

Katedra / Ústav: Katedra pravděpodobnosti a matematické statistiky

Vedoucí doktorské práce: Doc. Petr Volf, CSc.

Abstrakt: Tato dizertační práce se zabývá vývojem metod v analýze přežití v rámci Aalenova modelu za platnosti speciálních podmínek. Předpokládali jsme, že všechny regresní funkce a všechny pozorované proměnné jsou nezáporné. Tento typ modelu jsme nazvali monotonní Aalenův model. K odhadům regresních funkcí jsme použili metody založené na maximální věrohodnosti, konkrétně neparametrickou metodu maximální věrohodnosti, Bayesovskou analýzu s Beta procesy jako apriorními procesy pro kumulované regresní funkce a Bayesovskou analýzu s korelovanými apriorními procesy pro nekumulované regresní funkce.

Klíčová slova: monotonní Aalenův model, maximální věrohodnost, Beta proces, korelovaný skokový prior

Title: Bayesian and Maximum Likelihood Nonparametric Estimation in monotone Aalen model

Author: Jana Timková

Department / Institute: Department of Probability and Mathematical Statistics  
Supervisor of the doctoral thesis: Doc. Petr Volf, CSc.

Abstract: This work is devoted to seeking methods for analysis of survival data with the Aalen model under special circumstances. We supposed, that all regression functions and all covariates of the observed individuals were nonnegative and we named this class of models monotone Aalen models. To find estimators of the unknown regression functions we considered three maximum likelihood based approaches, namely the nonparametric maximum likelihood method, the Bayesian analysis using Beta processes as the priors for the unknown cumulative regression functions and the Bayesian analysis using a correlated prior approach, where the regression functions were supposed to be jump processes with a martingale structure.

Keywords: monotone Aalen model, maximum likelihood, Beta process, correlated stepwise prior

## Acknowledgements

I would like to thank the people at the Department of probability and mathematical statistics at Charles University in Prague who provided me with plentiful of valuable knowledge and mathematical background during my master and doctoral years. The special thank you goes to my supervisor Petr Volf who always pointed me to the right direction and never hesitated to give a helping hand.

I am very thankful to the Department of Statistics at University of Oslo in Norway who made it possible for me to spend one extremely exciting year on its premises during my doctoral studies. I particularly appreciate the opportunity to work under the supervision of Professor Nils Lid Hjort during my stay at University of Oslo, who is a great inspiration to me both academically and personally. A great part of findings in Chapter 2 and 3 are results of the joint work with Professor Hjort.

I would like to thank Dr. Víchová and Dr. Mořovská who kindly provided me with the real dataset on the time-delay of the patients with myocardial infarction.

The work done in this thesis was supported by the grants of GA ČR 201/05/H007 and by GA AV IAA101120604 and by the Yggdrasil grant for EU students studying in Norway under the project number 202841.

I would like to thank my parents who always stand beside me and without whom my life would not be possible. And last but not least I would like to thank Melda for his infinite support in every aspect of my academic, working and personal life.

To my family.

# Contents

|   |           |
|---|-----------|
| Abstract . . . . .                                    | ii        |
| Acknowledgements . . . . .                            | v         |
| List of Tables . . . . .                              | ix        |
| List of Figures . . . . .                             | x         |
| <b>1 Introduction</b>                                 | <b>1</b>  |
| 1.1 Survival data . . . . .                           | 3         |
| 1.2 Counting processes . . . . .                      | 5         |
| 1.3 Monotone Aalen model . . . . .                    | 10        |
| 1.4 Life-time models . . . . .                        | 13        |
| 1.5 Real dataset . . . . .                            | 15        |
| 1.6 Outline of the thesis . . . . .                   | 20        |
| 1.7 Current state on cases of inconsistency . . . . . | 22        |
| <b>2 Nonparametric MLE</b>                            | <b>25</b> |
| 2.1 Model formulation . . . . .                       | 26        |
| 2.2 Case $p = 1$ . . . . .                            | 28        |
| 2.3 Case $p > 1$ . . . . .                            | 37        |
| 2.4 Average consistency . . . . .                     | 45        |
| 2.5 An example . . . . .                              | 47        |
| <b>3 Beta process prior</b>                           | <b>51</b> |
| 3.1 Beta process prior . . . . .                      | 55        |
| 3.2 Posterior distribution . . . . .                  | 58        |
| 3.3 Bayesian estimators . . . . .                     | 64        |
| 3.4 Case $p = 1$ . . . . .                            | 70        |
| 3.5 Case $p > 1$ . . . . .                            | 72        |
| 3.6 Average consistency . . . . .                     | 74        |
| 3.7 Neutral to the right processes . . . . .          | 75        |
| 3.8 An example II . . . . .                           | 77        |
| 3.9 Discussion . . . . .                              | 81        |

---

|          |   |            |
|----------|---|------------|
| <b>4</b> | <b>Correlated prior</b>   | <b>85</b>  |
| 4.1      | Prior distribution . . . . .                                    | 86         |
| 4.2      | Posterior distribution . . . . .                                | 89         |
| 4.2.1    | Posterior distribution of regression functions levels . . . . . | 89         |
| 4.2.2    | Posterior distribution of jump times . . . . .                  | 93         |
| 4.3      | Simulations . . . . .   | 95         |
| 4.3.1    | Computational aspects of the estimation . . . . .               | 100        |
| 4.3.2    | Choice of the hyperparameters . . . . .                         | 103        |
| 4.4      | Discussion . . . . .  | 105        |
| <b>5</b> | <b>Real data analysis</b>                                       | <b>110</b> |
| 5.1      | Time-delay dataset . . . . .                                    | 110        |
| 5.2      | Danish malignant melanoma . . . . .                             | 114        |
| <b>6</b> | <b>Discussion</b>   | <b>120</b> |
| 6.1      | Future research directions . . . . .                            | 122        |
|          | <b>Bibliography</b>   | <b>124</b> |



# List of Tables

- 4.1 Results of simulation study: average values of measures of precision. 97
- 4.2 Results of simulation study: average values of simultaneous coverage of 95% and 99% pointwise credibility and confidence bands. . . 98
- 4.3 Comparison of exact and approximated posterior distribution. . . 101

# List of Figures

|     |   |     |
|-----|---|-----|
| 1.1 | Real dataset: logged estimated cumulative baseline hazards and Schoenfeld residuals for Cox model. . . . .              | 17  |
| 1.2 | Real dataset: The estimated cumulative regression functions from the classic Aalen model for the time-delay. . . . .    | 19  |
| 2.1 | Approximations of the NPML estimator. . . . .   | 30  |
| 2.2 | Two examples of the function in Lemma 1. . . . .  | 39  |
| 2.3 | NPML estimation in a simulated Aalen model. . . . .   | 49  |
| 3.1 | Bayesian estimation with Beta process prior in a simulated Aalen model. . . . .   | 78  |
| 3.2 | Bayesian estimation with Beta process prior in a simulated Aalen model - construction of the credibility bands. . . . . | 80  |
| 4.1 | Results of simulation study: Graphs of the pointwise averages of the estimators. . . . .                                | 99  |
| 4.2 | Estimators based on exact versus approximated posterior distribution. . . . .   | 102 |
| 4.3 | The MCMC trace of the regression functions. . . . .   | 103 |
| 4.4 | The MCMC trace of the trajectories at time point $t = 0.2$ . . . . .  | 104 |
| 4.5 | Estimated regression functions based on Bayesian analysis with correlated prior - PRIOR 1. . . . .                      | 106 |
| 4.6 | Estimated regression functions based on Bayesian analysis with correlated prior - PRIOR 2. . . . .                      | 107 |
| 4.7 | Estimated regression functions based on Bayesian analysis with correlated prior - PRIOR 3. . . . .                      | 108 |
| 5.1 | Real dataset: NPML estimation and Bayesian estimation with Beta process prior. . . . .                                  | 111 |
| 5.2 | Real dataset: Bayesian estimation with correlated prior for regression functions. . . . .                               | 112 |

5.3 Real dataset: Bayesian estimation with correlated prior for cumulative regression functions. . . . . 113

5.4 Melanoma dataset: NPML estimation and Bayesian estimation with Beta process prior. . . . . 118

5.5 Melanoma dataset: Bayesian estimation with correlated prior. . . 119

# Chapter 1

## Introduction

This thesis is devoted to exploring new possibilities in the estimation of unknown quantities in the survival analysis models. The main focus is on developing methods to estimate the unknown regression functions in a special case of the Aalen additive model. We consider two approaches to the estimation of the nonnegative regression functions. The first one is based on the assumption of a discontinuous cumulative hazard rate. We introduced two new nonparametric estimators of the cumulative regression functions derived from the maximum likelihood methodology and from the Bayesian framework with Beta process as a prior. This work is motivated by the Hjort paper on Beta processes in estimation of the cumulative hazard functions, [19], the path further extended by Kim and Lee to NII processes, [27], [31]. On the contrary, in the second approach we assume that the cumulative hazard rate is continuous and we model the regression functions using the priors as in Arjas and Gasbarra's work, [6].

As it soon becomes clear, the estimators derived in this thesis turned out to be inconsistent. The reason of inconsistency of these estimators is not obvious but it seems to be one of the problematic cases when even the reliable methods like nonparametric maximum likelihood estimation and Bayesian approach crush if an infinite-dimensional parameter estimation is involved. Still, it is felt that the proposed work is beneficial to some degree. It is not known to the author that any kind of Bayesian analysis of the general Aalen model has been done, apart from the frailty models, see [40]. Second, it introduces a special case of Aalen model which has interesting interpretation and good potential in data analysis.

In survival analysis non- or semi-parametric approaches have become widely used. Functional parameters of models are often estimated as 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, [26], or Breslow estimator of cumulative baseline hazard

function in Cox proportional hazard model, [9]. In most famous models like Cox 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 general Aalen model was first suggested by Aalen in the eighties, [2] and [3]. He proposed nonparametric estimators of regression functions based on the least squares method. These were further extended to the weighted least squares estimators by McKeague and Huffer, [21]. Both the least squares and weighted least squares estimators are again jump processes with discontinuities located in failure times.

Bayesian approach to survival data analysis has become a popular alternative to the aforementioned estimators which 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 processes (NII), a wider class of processes incorporating Gamma and Beta processes (and also Dirichlet processes via the known relationship  $H(t) = \int_0^t dF(s)/(1 - F(s_-))$  between the cumulative hazard function  $H$  and the distribution function  $F$ ). A process, let us 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  $\lim_{t \rightarrow \infty} H(t) = \infty$ , and obviously it induces a proper cumulative hazard function. For more details see [12] or [31]. 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 [31], [28], [10] respectively. For a good overview of Bayesian analysis in survival models see e.g. [41].

In Section 1.1 and Section 1.2 we give a summary of the basic theory related to the survival data analysis, counting processes and martingales, which will be useful in next chapters. In Section 1.3 and Section 1.4 we introduce the monotone Aalen model and compare it to the classic Aalen model. Furthermore, we talk about other popular models within the survival analysis, in particular about the well known Cox model. Next in Section 1.5 we introduce a (never

before analysed) real dataset containing observed delay-times of patients with the myocardial infarction which will be later used to demonstrate the obtained estimators. The chapter is finished with an outline of the main body of this thesis in Section 1.6 and an overview of the current state of knowledge in the field in Section 1.7.

## 1.1 Survival data

Let  $T_i^0, i = 1, \dots, n$ , be survival times that come from observing  $n$  independent individuals or objects, with distribution functions  $F_i$ . As usual in applications, the survival times can be and often are right-censored with random variables  $C_i$ . We suppose that the censoring mechanism is independent from the failure times  $T_i^0$ . The actual observed times are  $T_i = \min(T_i^0, C_i)$ .

We will prefer to work with counting processes instead of the survival times. Let us consider a multivariate counting process  $N(t) = (N_1(t), N_2(t), \dots, N_n(t))^T$  observed on a time interval  $[0, \tau]$ , where  $\tau = \max_{1 \leq i \leq n} T_i < \infty$ . All processes start from zero,  $N_i(0) = 0$ , and  $N_i(t)$  increases by 1 when the  $i$ -th object happens to meet an event of interest. No two components of  $N(t)$  jump at the same time with probability 1. For now let us suppose that the distribution of the survival times is absolutely continuous and the densities  $f_i$  of the distribution functions  $F_i, i = 1, \dots, n$  exist.

We assume *the multiplicative intensity model*, meaning that the intensity takes the form  $Y_i(t)h_i(t)$ , where  $h_i(t)$  is a deterministic bounded nonnegative continuous hazard function and  $Y_i(t)$  is a predictable  $\{0, 1\}$ -valued process indicating whether the  $i$ -th individual is at risk of event whenever  $Y_i(t) = 1$ . The indicator process has its importance when the censoring is present or when the occurrence of an event implies the end of the observing of the object. In this work we will consider only the case when every subject can experience the event of interest only once.

*The hazard function* or *the hazard rate* related to the  $i$ -th object,  $h_i(t)$ , is the instantaneous rate of an event occurring at time  $t$  defined as

$$h_i(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(N_i(t + \Delta t) - N_i(t) = 1 | \mathcal{F}_t),$$

where  $\mathcal{F}_t := \sigma\{N_i(s), 0 \leq s \leq t\}$  is a  $\sigma$ -algebra of the history of the  $i$ -th individual up to time moment  $t$ . After multiplying  $h_i(t)$  with  $\Delta t$  we get an approximate probability of  $i$ -th subject failing in  $(t, t + \Delta t]$ .

Switching back to  $T_i$  in the definition of the hazard rate we can easily see that

$$h_i(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{P(t < T_i \leq t + \Delta t, T_i = T_i^0)}{P(T_i > t)} = \frac{f_i(t)}{S_i(t)} = - \frac{d}{dt} \log S_i(t)$$

and after integrating both sides and exponentiating them we get

$$S_i(t) = \exp \left\{ - \int_0^t h_i(s) ds \right\}.$$

With  $S_i(t) = 1 - F_i(t)$  we denoted the survival function, which is the probability of the  $i$ th individual surviving till time  $t$ . From  $h_i(t) = f_i(t)/S_i(t)$ , the expression for the probability density function of  $T_i$  emerges

$$f_i(t) = h_i(t) \exp \left\{ - \int_0^t h_i(s) ds \right\}. \quad (1.1)$$

Nevertheless, unlike in other fields of statistics it is not usual to describe the distribution of  $T_i$  by densities or the distribution functions. One is often more interested in the (cumulative) hazard rate or the survival function.

Most nonparametric methods within the survival data analysis are unable to estimate the hazard functions directly and focus on estimation of the cumulative hazard rate  $H_i(t) = \int_0^t h_i(s) ds$  instead. In homogeneous case, when  $h_i \equiv h, \forall i$ , the aggregated process  $\bar{N}(s) = \sum_{i=1}^n N_i(s)$  is sufficient for estimating  $H$ . When no parametric assumptions are imposed on the functional form of  $h$ , the cumulative hazard rate  $H$  is estimated using the traditional Nelson-Aalen estimator (see [1]),

$$\hat{H}(t) = \int_0^t \sum_{i=1}^n \frac{dN_i(s)}{\sum_{j=1}^n Y_j(s)} = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)},$$

where we denoted  $\bar{Y}(s) = \sum_{i=1}^n Y_i(s)$ .

There is a one-to-one relation between the distribution function  $F_i$  and the cumulative hazard function  $H_i$ , precisely

$$H_i(t) = \int_0^t \frac{dF_i(s)}{1 - F_i(s_-)} \quad \text{and} \quad F_i(t) = 1 - \prod_{s \in [0, t]} \{1 - dH_i(s)\}. \quad (1.2)$$

The estimator of the distribution function  $F_i$  as well as the survival function  $S_i$  is then easy to obtain once we have an estimator for  $H_i$ . The relationships in (1.2) remain intact and therefore applicable even when the distribution is not continuous and the distribution function has no density. The symbol  $\prod_{s \in [0, t]}$  in the right hand part of (1.2) is the product integration and it is a continuous version of the ordinary product. We say that  $Y(t) = \prod_{s \in [0, t]} \{1 + dX(s)\}$  is *the*

product integral of a càdlàg function  $X$  if

$$Y(t) = \lim_{\max |t_i - t_{i-1}| \rightarrow 0} \prod_{i=1}^m (1 + X(t_i) - X(t_{i-1})),$$

where  $0 = t_0 < t_1 < \dots < t_m = t$  is a partition of the time interval  $[0, t]$ . If  $X$  is a step function, we can write  $X = X^c + \sum \Delta X$ , where  $X^c$  is the continuous part and  $\Delta X = X - X_-$  are the steps. For step functions the product integral separates into two parts

$$\prod_{s \in [0, t]} \{1 + dX\} = \exp\{X^c\} \prod (1 + \Delta X).$$

Obviously, when  $F_i$  is continuous, the latter of (1.2) reduces to  $F_i(t) = 1 - \exp\{-H_i(t)\}$ .

Under the multiplicative intensity model the likelihood of the observed data  $N_1, \dots, N_n$  in time interval  $[0, \tau]$ , where  $\tau = \max_i \{T_i\}$ , is

$$L = \prod_{i=1}^n \prod_{t \in [0, \tau]} \left\{ dH_i(t)^{dN_i(t)} (1 - dH_i(t))^{Y_i(t) - dN_i(t)} \right\}.$$

In the continuous case, when  $h_i$ -s exist, using the features of the product integral simplifies the above expression down to

$$L = \prod_{i=1}^n h_i(T_i)^{dN_i(T_i)} \exp\{-H_i(T_i)\}.$$

This is in agreement with  $L = \prod_{i=1}^n f_i(T_i)^{dN_i(T_i)} S_i(T_i)^{1 - dN_i(T_i)}$ , where the first part represents the individuals who failed at their  $T_i$ -s, while the second part corresponds to the survival of all individuals up to their  $T_i$ -s regardless if they failed or were censored.

## 1.2 Counting processes and martingale theory

Since the introduction of the martingale approach and counting processes to the statistical analysis of time-to-event data in 1970s it has become remarkably useful in dealing with most problems arising in the survival data analysis. The central limit theorem for martingales as stated by Rebolledo in [38] made a clear path to proving the weak convergence of an estimator that could often be expressed as a stochastic integral of a predictable process with respect to a local martingale. In



next pages let us make a quick journey into the world of the martingale approach to survival data.

Let  $\mathcal{T} = [0, \tau)$  and  $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \in \mathcal{T}\}, P)$  be a complete stochastic basis, i.e. a probability space equipped with a filtration  $\{\mathcal{F}_t : t \in \mathcal{T}\}$ , an increasing right-continuous family of sub- $\sigma$ -algebras of the  $\sigma$ -algebra  $\mathcal{F}$ . We assume that the components of the multivariate counting process  $N$  defined in Section 1.1 are càdlàg processes existing in the probability space  $(\Omega, \mathcal{F})$  and adapted to the filtration  $\{\mathcal{F}_t : t \in \mathcal{T}\}$ . The filtration is supposed to fulfil the usual conditions (see e.g. [4], pp. 60) and it can either be defined up front or generated by the process  $N(t)$ . Let us denote  $\overline{\mathcal{T}} = \mathcal{T} \cup \{\tau\}$ . Most of the empirical processes used in this context live on the space of càdlàg functions  $D(\overline{\mathcal{T}})$  called the Skorohod space. Since most of the theoretical results in next chapters focus on the weak convergence of the empirical processes, we will endow the space  $D(\overline{\mathcal{T}})$  with Skorohod topology.

Let us first introduce several basic terms. A process is called *càdlàg* if its sample paths are right-continuous and have left-hand limits. A process  $X$  is called *adapted* to the filtration  $\{\mathcal{F}_t : t \in \mathcal{T}\}$  if  $X(t)$  is  $\mathcal{F}_t$  measurable for each  $t$ . If a process is measurable with respect to the  $\sigma$ -algebra generated by all left-continuous adapted processes, then it is called *predictable*. Finally, the process  $X$  is *bounded* if there exist a finite constant  $\Gamma$  such that  $\sup_{t \in \mathcal{T}} |X(t)| < \Gamma$  almost surely.

We say that a càdlàg process  $M = \{M(t) : t \in \mathcal{T}\}$  is a *martingale* with respect to the filtration  $\{\mathcal{F}_t : t \in \mathcal{T}\}$  if  $M$  is adapted to  $\{\mathcal{F}_t : t \in \mathcal{T}\}$ ,  $E|M(t)| < \infty$  for all  $t \in \mathcal{T}$  and

$$E\{M(t) | \mathcal{F}_s\} = M(s) \quad \text{a.s.} \quad \text{for all } s, t \in \mathcal{T}, s \leq t.$$

If the equality in the last expression is replaced by  $E\{M(t) | \mathcal{F}_s\} \geq M(s)$  then  $M$  is called a *submartingale*. Further, the process  $M$  is *square integrable* if  $\sup_{t \in \mathcal{T}} E\{M(t)\}^2 < \infty$ .

Looking at the multivariate counting process we defined in previous section, we can see that every component of  $N(t)$  is a nonnegative right-continuous local submartingale. The attribute *local* means that there exists a *localization*, i.e. a sequence of random variables  $\{\tau_m\}_{m=1}^\infty$  in  $\overline{\mathcal{T}}$  such that  $\{\tau_m \leq t\} \in \mathcal{F}_t$ ,  $P(\tau_m \geq t) \rightarrow 1$  as  $m \rightarrow \infty$  for all  $t \in \mathcal{T}$  and the stopped process  $I(\tau_m > 0)N_i(t \wedge \tau_m)$  is a submartingale:

$$E(I(\tau_m > 0)N_i(t \wedge \tau_m) | \mathcal{F}_s) \geq I(\tau_m > 0)N_i(s \wedge \tau_m), \quad \forall s \leq t, \quad \forall i.$$

Hence according to the Doob-Meyer decomposition (see [4], pp. 66-67), there exists a càdlàg nondecreasing predictable process  $H_i^Y$  to every  $N_i$ , such that  $H_i^Y$  has finite variation and  $H_i^Y(0) = 0$ . Moreover, we have that the difference between  $N_i$  and  $H_i^Y$ ,

$$M_i(t) = N_i(t) - H_i^Y(t), \quad t \in \mathcal{T}, \quad (1.3)$$

is a zero-mean right-continuous local martingale.  $M_i(t)$  is in fact the difference between the number of the events occurred up to time  $t$  and the expected number of events. The process  $H_i^Y$  is called a *compensator* of the process  $N_i$  and once it exists, it is unique. It can be shown that  $H_i^Y(t) = H_i(t \wedge \max\{s \in \mathcal{T} : Y_i(s) = 1\})$ , i.e.  $H_i^Y$  is the cumulative intensity process of  $N_i$  and in the absolutely continuous case  $H_i^Y(t) = \int_0^t Y_i(s)h_i(s)ds$ . The important consequence of this application of Doob-Meyer decomposition is the fact, that for every  $N_i(t)$  defined as above the uniquely determined cumulative intensity process exists, which means the cumulative hazard rate  $H_i$  exists.

In the next, certain properties of the processes like boundedness or integrability will hold only *locally* which means that a property is satisfied by the stopped process with appropriately chosen localization. As  $N_i$  and  $H_i^Y$  are both locally bounded, combining their localizing times it can be shown that  $M_i$  from (1.3) is square integrable. By Jensen's inequality we have that  $M_i^2$  is also a local submartingale. Applying the Doob-Meyer decomposition on  $M_i^2$  we have the existence of a compensator to  $M_i^2$  that we denote  $\langle M_i, M_i \rangle$  or often just  $\langle M_i \rangle$ . This unique nondecreasing right-continuous predictable process is called a *predictable variation process* of  $M_i$ . For any two  $M_i$  and  $M_j$ ,  $i \neq j$ ,  $M_i M_j$  is again a local submartingale with compensator  $\langle M_i, M_j \rangle$  called a *predictable covariation process*.

If the compensators  $H_i^Y$  of the counting processes  $N_i$ ,  $i = 1, \dots, n$ , are continuous, then for the predictive variation and covariation processes of  $M_i = N_i - H_i^Y$  we have

$$\langle M_i \rangle = H_i^Y \quad \forall i \quad \text{and} \quad \langle M_i, M_j \rangle = 0 \quad i \neq j. \quad (1.4)$$

Furthermore, let us consider a process  $\sum_{i=1}^n \int U_i dM_i$  where  $U_i$ ,  $i = 1, \dots, n$  are predictable locally bounded processes and  $M_i$ ,  $i = 1, \dots, n$  are from (1.3). Processes of this type arises often in practise in statistical testing and estimation within survival analysis, viz. the log-rank tests. The process  $\sum_{i=1}^n \int U_i dM_i$  has nice features and it is a local square integrable zero-mean martingale with compensator equal to

$$\left\langle \sum_{i=1}^n \int_0^t U_i(s) dM_i(s), \sum_{i=1}^n \int_0^t U_i(s) dM_i(s) \right\rangle$$

$$= \sum_{i=1}^n \sum_{j=1}^n \int_0^t U_i(s)U_j(s)d \langle M_i, M_j \rangle (s), \quad \forall t \in \mathcal{T}.$$

If  $H_i^Y$  are continuous then by applying (1.4) and substituting  $Y_i(s)h_i(s)ds$  instead of  $dH_i^Y(s)$  we see that the compensator equals to

$$\sum_{i=1}^n \sum_{j=1}^n \int_0^t U_i(s)U_j(s)d \langle M_i, M_j \rangle (s) = \sum_{i=1}^n \int_0^t U_i^2(s)Y_i(s)h_i(s)ds, \quad \forall t \in \mathcal{T}.$$

If  $E \int_0^t U_i^2 d \langle M_i \rangle (s) < \infty$ , for all  $i$ , then  $\sum_{i=1}^n \int U_i dM_i$  is a zero-mean martingale over  $[0, t]$  and

$$E \left\{ \sum_{i=1}^n \int_0^t U_i(s)dM_i(s) \right\}^2 = E \sum_{i=1}^n \int_0^t U_i^2(s)Y_i(s)h_i(s)ds.$$

Since  $E \sum_{i=1}^n \int U_i dM_i = 0$ , we can see that  $E \langle \sum_{i=1}^n \int U_i dM_i \rangle$  is the variance of the process  $\sum_{i=1}^n \int U_i dM_i$  at  $t$ .

Finally, we recall the martingale central limit theorem which will be useful later on. The version we state here was given by Rebolledo in [38], hence it is called the *Rebolledo theorem*. Let us reformulate the theorem for the processes which are the products of the stochastic integration with respect to martingales. Denote

$$V^{(n)}(t) = \sum_{i=1}^n \int_0^t U_i^{(n)}(s)dM_i^{(n)}(s), \quad (1.5)$$

where  $M_i^{(n)} = N_i^{(n)} - H_i^{Y,(n)}$ . The superscript  $(n)$  is to emphasize the dependence of the processes on the sample size  $n$ . The following is the central limit theorem for the process  $V^{(n)}$  as stated in [14], Th. 5.3.5.

**Theorem 1 (Martingale central limit theorem for  $V^{(n)}$ )** *Assume that the compensators  $H_i^{Y,(n)}$  of the counting processes  $N_i^{(n)}$  are continuous for  $\forall i$ ,  $U_i^{(n)}$  are locally bounded predictable processes for  $\forall i$ . For any  $\epsilon > 0$  let us denote the process*

$$V_\epsilon^{(n)}(t) = \sum_{i=1}^n \int_0^t U_i^{(n)}(s)I_{\{|U_i^{(n)}(s)| \geq \epsilon\}} dM_i^{(n)}(s).$$

*Let  $V^{(\infty)}$  be a zero-mean Gaussian process with independent increments,  $V^{(\infty)}(0) = 0$  and  $E\{V^{(\infty)}(t)\}^2 = C(t)$ , where  $C$  is a continuous function. Suppose that for all  $t \in \overline{\mathcal{T}}$ , as  $n \rightarrow \infty$ ,*

$$\langle V^{(n)}, V^{(n)} \rangle (t) \xrightarrow{P} C(t)$$

and

$$\langle V_\epsilon^{(n)}, V_\epsilon^{(n)} \rangle(t) \xrightarrow{P} 0, \quad \text{for any } \epsilon > 0.$$

Then

$$V^{(n)} \xrightarrow{\mathcal{D}} V^{(\infty)} \quad \text{in } D(\overline{\mathcal{T}}) \quad \text{as } m \rightarrow \infty.$$

Often we need a convergence result of a functional of a convergent process instead of the process itself. As long as the functional in question is continuous, the convergence to the functional of the limiting process can be achieved easily by applying the continuous mapping theorem. Here we state the theorem as it is in [14], Th. B.1.1.

**Theorem 2 (Continuous mapping theorem)** *Suppose  $g$  is a continuous mapping from one metric space  $(\Gamma, \mathcal{S}_0)$  to another  $(\Gamma', \mathcal{S}'_0)$ . If for random elements  $X^{(n)}$  and  $X$  defined on  $(\Gamma, \mathcal{S}_0)$  we have  $X^{(n)} \xrightarrow{\mathcal{D}} X$  in  $(\Gamma, \mathcal{S}_0)$  then  $g(X^{(n)}) \xrightarrow{\mathcal{D}} g(X)$  in  $(\Gamma', \mathcal{S}'_0)$ .*

A practical application of the continuous mapping theorem is for example the convergence of any set of the finite-dimensional distributions of  $V^{(n)}$  to the corresponding set of finite-dimensional distributions of  $V^{(\infty)}$ . Often used functional is also a supremum of the process over some time interval, i.e. the respective convergence result would be as following

$$\sup_{0 \leq s \leq t} V^{(n)}(t) \xrightarrow{\mathcal{D}} \sup_{0 \leq s \leq t} V^{(\infty)}(t).$$

This section covers only a basic knowledge on the martingale approach to survival data analysis to provide a foundation for findings in the next chapters. More details and very good overview on the theory of counting processes and their connection to martingales can be found either in [14] or [4], Ch. II.

The last theorem stated here is an inequality which is not necessarily connected with stochastic processes but it will be crucial for one of the proofs in Chapter 2.

**Theorem 3 (McDiarmid inequality, [35])** *Let  $X_1, \dots, X_m$  be independent random variables defined in a set  $\mathcal{X}$ . Further, let  $g : \mathcal{X}^m \mapsto \mathbb{R}$  be a function of  $X_1, \dots, X_m$  that satisfies*

$$\left| g(x_1, \dots, x_i, \dots, x_m) - g(x_1, \dots, x'_i, \dots, x_m) \right| \leq c_i \quad \forall i, \forall x_1, \dots, x_m, x'_i \in \mathcal{X}.$$

Then for all  $\epsilon > 0$ ,

$$P\left(\left|g(x_1, \dots, x_m) - \mathbb{E}g(x_1, \dots, x_m)\right| \geq \epsilon\right) \leq 2 \exp\left\{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right\}.$$

Before we move on to another section, let us remind what big  $O_p$  and small  $o_p$  stand for in probability notation.

**Definition 1** Let  $X_n$  be a set of random variables and  $a_n$  be a corresponding set of constants.

- (a) The notation  $X_n = O_p(a_n)$  means that random variables  $X_n/a_n$  are stochastically bounded, i.e. for  $\forall \epsilon > 0$  there exists a finite constant  $M$  such that

$$P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \epsilon \quad \forall n.$$

- (b) The notation  $X_n = o_p(a_n)$  means that random variables  $X_n/a_n$  converge to 0 in probability with  $n \rightarrow \infty$ , i.e. for  $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_n}{a_n}\right| > \epsilon\right) = 0.$$

### 1.3 Monotone Aalen model

In this work we study the Aalen additive model of Aalen, [2] and [3], on a dataset of form  $(T_i, \delta_i, (x_{i,1}, \dots, x_{i,p})^\top)_{i=1}^n$ , where  $T_i = \min(T_i^0, C_i)$  are observed right-censored survival times,  $\delta_i = I\{T_i = T_i^0\}$  is the indicator of noncensored observation and  $(x_{i,1}, \dots, x_{i,p})^\top$ , are  $p$ -dimensional real-valued 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  $i$ th individual with distribution function  $F_i = F(\cdot | (x_{i,1}, \dots, x_{i,p})^\top)$  and  $C_i$  is a censoring variable independent on  $T_i^0$ . The Aalen model assumes that the hazard rate for  $i$ th object is

$$h_i(t) = \sum_{j=1}^p x_{i,j} \alpha_j(t), \quad i = 1, \dots, n, \quad (1.6)$$

where  $\alpha_1, \dots, \alpha_p$  are unknown regression functions. Often  $x_{i,1} \equiv 1, \forall i$ , and  $\alpha_1$  represents a baseline risk of failure common for all individuals if there is no other risk factor present. Typically, the Aalen model with a baseline regression function is formulated as  $h_i(t) = \alpha_0(t) + \sum_{j=1}^p x_{i,j} \alpha_j(t)$ , but for our needs it will be more convenient to stick to the formulation in (1.6) and set  $x_{i,1} \equiv 1$  if the baseline hazard is involved. Even though we allow the covariates  $x_{i,j}$  to be real, hence even equal to zero, it is clearly not admissible that an individual has all  $x_{i,j} = 0, j = 1, \dots, p$ .

Aalen studied the model assuming that  $\alpha_j$ -s take real values and only the overall hazard function  $h_i$  needs to be nonnegative. He estimated the cumulative versions of regression function  $A_j(t) = \int_0^t \alpha_j(s) ds, j = 1, \dots, p$  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  $\alpha(t) = (\alpha_1(t), \dots, \alpha_p(t))^\top$ ,  $A(t) = (A_1(t), \dots, A_p(t))^\top$ ,  $z_i = (x_{i,1}, \dots, x_{i,p})^\top$ ,  $N(s) = (N_1(s), \dots, N_n(s))^\top$  and  $Z(s) = (z_1 Y_1(s), \dots, z_n Y_n(s))^\top$ . Then the Aalen least squares estimator is equal to

$$A^a(t) = \int_0^t (Z(s)^\top Z(s))^{-1} Z(s)^\top dN(s).$$

From the theory of martingales and the Doob-Meier decomposition we have existence of a zero-mean martingale  $M_i(t) = N_i(t) - \int_0^t Y_i(s) z_i^\top dA(s)$ . The estimator is motivated by the fact that

$$E dN(t) = E Z dA(t)$$

and intuitively

$$A^a(t) = \int_0^t Z(s)^- dN(s).$$

The particular choice of pseudoinverse  $Z^- = (Z^\top Z)^{-1} Z^\top$  leads to Aalen estimator. The unweighted least squares method, however, does not take into account the fact that the variances of  $M_i$ -s might be unequal. Huffer and McKeague [21] introduced a two-stage estimator which is essentially a weighted least squares estimator with a matrix of weights  $V^* = \text{diag}\{Z\alpha^*\}^{-1}$ , where  $\alpha^*$  is obtained in the first stage as a smoothed OLS estimator via kernel estimation. Under some conditions on the kernel function and the bandwidth,  $\alpha^*$  is a uniformly consistent estimator of the vector of regression functions  $(\alpha_1, \dots, \alpha_p)^\top$ . Essentially,  $V^*$  is a matrix with estimators of  $\text{Var}(dM_i(s))$  on diagonal. In the second stage the regression processes are estimated by

$$A^*(t) = \int_0^t (Z(s)^\top V^*(s) Z(s))^{-1} Z(s)^\top V^*(s) dN(s). \quad (1.7)$$

Both Aalen's and Huffer and McKeague's estimators are under certain regularity conditions consistent and their asymptotic distributions are  $p$ -dimensional zero-mean Gaussian martingales. Furthermore, as shown in [4], 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 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). Second, we assume that the regression functions  $\alpha_j, j = 1, \dots, p$ , 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 non-decreasing (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 p. 910 in [3]).

Another advantage is that the monotone Aalen model is more natural in interpretation of the estimated regression functions. Let us assume, that we have the intercept included in the model,  $x_{i,1} \equiv 1$  for all  $i$ . Hence we can formulate the model in a way that an individual with covariates  $x_{i,j} = 0, j > 1$ , represents an average healthy individual and their hazard rate is contained in the regression function  $\alpha_1$ . 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  $p$  cause-specific hazard functions. The overall hazard function of the competing risks model is the same as in (1.6) and the observed outcome is the failure due to one of the  $p$  independent causes. Then  $T_i^0$  would be viewed as the minimum of the  $p$  independent life-time variables with hazard rates  $\alpha_1, x_{i,2}\alpha_2, \dots, x_{i,p}\alpha_p$ . Under the assumption of independence of the competing risks we have

$$1 - F(t|z_i) = \{1 - G_1(t)\} \prod_{j=2}^p \{1 - G_j(t)\}^{x_{i,j}},$$

where  $G_1, \dots, G_p$  are distribution functions of random variables with cumulative hazard rates equal to  $A_1, \dots, A_p$ . 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 as we do not observe the type of failure, i.e. which of the risks caused the outcome.

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 of 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 it is most likely caused by the general instability of the estimates at the end of the observation window. The monotone Aalen model is of good use if we would like to utilize also the ending of the time window and we need a nonnegative estimator as well. 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  $\alpha_j$ -s 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 made before looking into 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.

## 1.4 Other life-time models

Among the most popular semiparametric models belongs the well-known *Cox regression model* of Cox, [9], which assumes the hazard rate  $h_i$  has following form,

$$h_i(t) = h(t; z_i) = \exp\{\beta^\top z_i\} h_0(t). \quad (1.8)$$

Here  $\beta$  is a column vector of  $p$  unknown regression coefficients and  $h_0$  is an unknown and unspecified baseline hazard rate common for all individuals (the hazard rate function for individual with  $z = (0, \dots, 0)^\top$ ).

The traditional approach to the regression parameter estimation is via the partial maximum likelihood theory. It uses the fact that likelihood can be written as a product of two components and only one of these components contains information about  $\beta$ . The estimator  $\hat{\beta}$  of  $\beta$  is defined as a solution of  $U(\beta, \tau) = 0$ ,



where  $U(\beta, t), t \in [0, \tau]$ , is the score process equal to

$$U(\beta, t) = \sum_{i=1}^n \int_0^t \left( z_i - \frac{\sum_{j=1}^n Y_j(s) z_j \exp\{\beta^\top z_j\}}{\sum_{j=1}^n Y_j(s) \exp\{\beta^\top z_j\}} \right) dN_i(s).$$

Once we have the estimator for  $\beta$ , the cumulative baseline hazard function  $h_0(t)$  can be estimated using the Breslow estimator of similar nature as Nelson-Aalen estimator in homogeneous case,

$$\hat{H}_0(t) = \int_0^t \left[ \sum_{i=1}^n Y_i(s) \exp\{\hat{\beta}^\top z_i\} \right]^{-1} d\bar{N}(s). \quad (1.9)$$

Andersen and Gill, [5], extended the proportional effect of the covariates on the intensity process of a counting process and established consistency of  $\hat{\beta}$  and weak convergence of  $\hat{H}_0$  using the martingale approach. Tsiatis, [45], proved strong consistency under the time-constant covariates.

Another class of models are *accelerated failure time models*, that assume that for  $i$ th object following is true

$$\log T_i = -Z_i^\top \beta + \epsilon_i, \quad i = 1, \dots, n, \quad (1.10)$$

where  $\beta$  is an unknown  $p$ -dimensional regression parameter and  $\epsilon_i, i = 1, \dots, n$  are the error terms that are ruled by a common, in general unknown, distribution. The model was introduced by Kalbfleisch and Prentice in 1980, [25].

Other famous models are the already mentioned competing risks models, Cox-Aalen models which combine both additive and multiplicative effect of covariates on the hazard rate, frailty models and many more.

Even though many datasets are analysed using both the Cox model and the Aalen model there are several differences between them. The main is contained in fact that the effect of covariates on the hazard rate is multiplicative in the Cox model and additive in the Aalen model. The important aspect of Cox model is the log-linearity and the proportionality of the hazard meaning that the hazard functions of two individuals should not cross and one should be a multiple of the other. This follows from the fact that for time-constant covariates the ratio of the hazard rates of two individuals is

$$\frac{h_i(t)}{h_j(t)} = \frac{h(t; z_i)}{h(t; z_j)} = \exp\{\beta^\top (z_i - z_j)\} \equiv \text{const}, \quad \forall t \in [0, \tau].$$

Naturally, the case of time-dependent covariates damages this property. When the condition on the constant proportional hazard is not fulfilled, either the strat-

ification of the dataset should be considered or a variant of the Cox model with functions  $\beta_j(t)$  instead of the coefficients  $\beta_j$  can be applied. Another option is the additive Aalen model where the regression functions are left completely unspecified what allows for arbitrary shape of the hazard rate. If the estimates of the cumulative regression functions in the Aalen model look like a straight line, it can be assumed that the effect of the covariates is constant over time, hence as long as the log-linearity is fulfilled the Cox model can be a sufficient choice.

There are several reasons why the Cox model is the most popular choice for the time-to-event analysis. It produces estimators which are easy to interpret and the model fitting and diagnostics are well-known and available in most commonly used computer packages such as SAS, SPLUS and R. Still, the assumption of the constant proportional hazards needs to be fulfilled to make this estimators effective. On the other hand, the Aalen model is more flexible and gives an appealing understanding how the effects of the covariates develop over time.

## 1.5 Real dataset: Time-delay of the patients with STEMI

Myocardial infarction, or commonly known as heart attack, is an event when blood stops flowing properly into a part of the heart and causes damage to the heart muscle due to the oxygen deprivation. According to the information stated at the webpage [www.cdc.gov](http://www.cdc.gov), heart disease is the leading cause of death worldwide with about 25% of deaths attributable to any heart related disease. In particular, about 15% of patients who experience acute myocardial infarction die of it.

A great part of the patients who are hospitalised with an incidence of the heart attack are classified with ST-elevated myocardial infarction (STEMI) and treated with reperfusion therapy which restores circulation of blood to the heart. To fully enjoy the benefits from the reperfusion therapy it is crucial that the therapy is initiated at early stage of the incident. Hence, it is important that the patient with suspicion of undergoing heart attack presents early. The term *time-delay* stands for the total duration from the very onset of the myocardial infarction to the surgery itself. The onset of the myocardial infarction is defined as the onset of the symptoms typical for the heart failure like the chest pain, discomfort and shortness of breath. According to the present-day knowledge, increase in the time-delay of a patient is associated mainly with female gender, higher age, low intensity of the symptoms, daytime when the incident happens, delay at the transfer to the hospital, etc.

The data analysed here were collected during three years from 2009 to 2011 at the Royal Vinohrady Teaching Hospital in Prague in the Czech Republic by Dr. Víchová and Dr. Mořovská who kindly provided the author with the dataset. The dataset contains entries about 649 patients with the highest value of the observed time-delay equal to 5895 minutes. We decided to consider only the patients who arrived before the first 24 hours (1440 minutes) as the data after this time-point are too sparse. The final dataset concerns 622 patients, from which 425 are males and 197 are females. There are no censored observations. A great amount of medical, life-style and system delay factors associated with every patient were collected.

The outcome of the interest here is the time-delay reported by each patient or their relatives. The observations are naturally subject to certain inaccuracy induced by absence of the symptoms or imprecise memory of the time when the incidence started. The bias from the true value, however, is most likely not systematic. We will search for the relationship of the time-delay of a patient in connection to four factors:

GENDER: Male/Female,

FIRST CONTACT DELAY: Yes/No; a delay induced by the first contact medical ward, e.g. by misinterpretation of the symptoms,

DAYTIME: Day/Night; the time of the day when the symptoms appear,

WORKING STATUS: Employed/Unemployed or retired.

It is well known, that women are less prone to the myocardial infarction incidence, hence there seem to be a negligence towards the heart attack symptoms from both female patients and physicians. According to the findings in several medical studies the chance that a female patient shows up later than three hours from the onset is about 1.5 times greater than for a male patient. The prolonging effect of the system-delay caused by the first contact physicians on the total time-delay is obvious. The factor DAYTIME was included as it is common that patients tend to wait till morning to visit their practising physician, if the symptoms occur during the night time. The last factor is again crucial as there is greater chance of getting immediate help or encouragement to call the ambulance from co-workers. Furthermore, it carries along the information about the age of the patient and higher age is again associated with greater time-delay.

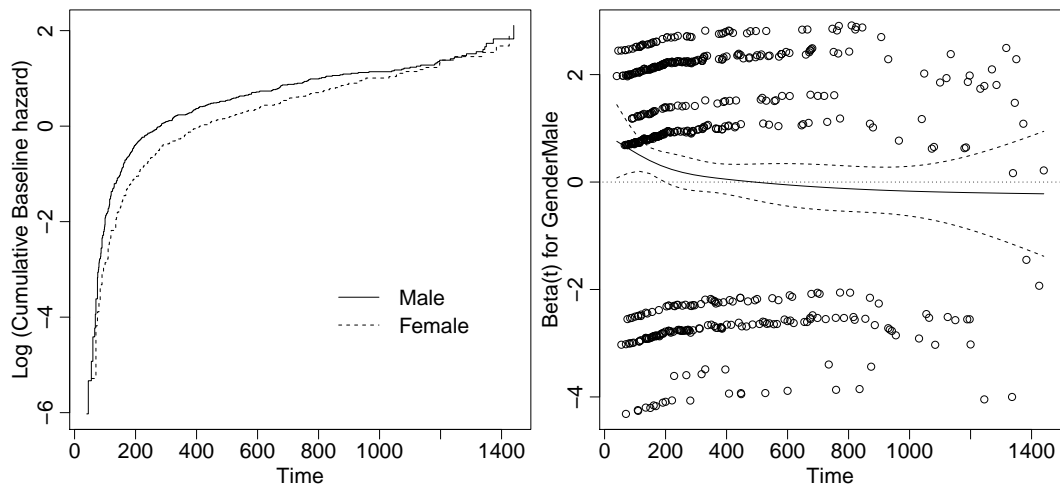


Figure 1.1: *Left:* The logged estimated cumulative baseline hazards for the time-delay separately for male and female gender, the estimation was done using stratified Cox model. *Right:* Standardized Schoenfeld residuals plotted against event time from non-stratified Cox model.

Hence, the female gender, presence of the first contact delay, onset at the night time and unemployment are all expected to have positive effect on prolonging the time-delay. Stating the same in the opposite way, the hazard rate of the delay time increases (i.e. the time-delay is shorter) for working male patients with the onset of the myocardial infarction during the day and without any hold-up caused by the first contact medical ward. To make use of the monotone Aalen model interpretation later on, we set the baseline hazard rate to correspond to an average non-working female patient with the onset of symptoms at night and who gets delayed at the first contact treatment. According to the hypothesis given by the results of most cardiological studies, this is the worst scenario and implies the greatest time-delay, so in fact the smallest hazard rate. Any deviation from this baseline setting, e.g. male gender or the onset of symptoms during daytime, is expected to have an increasing effect on the hazard rate leading to a smaller time-delay. So as opposed to the monotone Aalen model interpretation proposed in Section 1.3, here an individual with the greatest health risk is contained in the baseline hazard and every additional factor represents a step towards less risky situation with regards to the impact on health of the individual. Still, this setting is fully valid as we expect the regression functions to be positive.

We first analysed the dataset using the Cox model. The estimated parameters were all significantly greater than zero what supports our hypothesis about all the factors having increasing effect on the hazard rate. However, the proportional hazard assumption seems to be invalid for the two genders, see Figure 1.1.

The graph on the left hand side shows the logged cumulative baseline hazards estimated in the Cox model stratified according to the gender. We tested the proportional hazard assumption using the Grambsch and Therneau test based on rescaled Schoenfeld residuals, see [17]. Let us consider a process

$$M(\beta, t) = \frac{\sum_{i=1}^n Y_i(t) \exp\{\beta^\top z_i\} z_i}{\sum_{i=1}^n Y_i(t) \exp\{\beta^\top z_i\}}.$$

$M(\beta, t)$  is in fact a weighted mean of the covariate vector at time  $t$  conditionally on the parameter vector  $\beta$ . Let us suppose that the number of the failure events is  $l \leq n$  and let us denote these events by  $t_1, \dots, t_l$  and the covariate vectors corresponding to the individuals with these failure times by  $z_{(1)}, \dots, z_{(l)}$ . Then the Schoenfeld residuals are defined as

$$r_i(\beta) = z_{(i)} - M(\beta, t_i), \quad i = 1, \dots, l.$$

Let us consider that the data truly comes from a nonproportional Cox model with a hazard rate equal to  $\lambda_0(t) \exp\{\beta^*(t)^\top z_i\}$ , with  $\beta^* = (\beta_1^*, \dots, \beta_p^*)^\top$  and  $\beta_j^*(t) = \beta_j + \theta_j g_j(t)$  for some predictable  $g_j$ . It shows (details in Section 2 of [17]) that under this model  $E r_i(\beta) \simeq V(\beta, t_i) G(t_i) \theta$ , where  $G(t_i) = \text{diag}\{g_1(t_i), \dots, g_p(t_i)\}$ ,  $\theta = (\theta_1, \dots, \theta_p)^\top$  and  $V(\beta, t_i)$  is a conditional variance of the covariate vector under the original model  $\lambda_0(t) \exp\{\beta^\top z_i\}$  (see Section 1 in [17]). Clearly, when there is no departure from the original model  $\lambda_0(t) \exp\{\beta^\top z_i\}$  then the Schoenfeld residuals  $r_i(\beta)$  will randomly vary around the x-axis. If there is however a functional trend visible when a smoothed line is added to the plot of  $r_i(\beta_j)$  against  $t_i$ , for some  $j$ , it shows the functional form for  $\beta_j^*(t)$  which should be considered in the model instead of the constant  $\beta_j$ .

Apart from the visual check, Grambsch and Therneau, [17], derived also an asymptotic test of a hypothesis  $H_0 : \theta = 0$ . Following their work it turns out that  $n^{-1/2} \sum_{i=1}^l G(t_i) r_i(\hat{\beta})$  is asymptotically distributed as  $\chi^2$  distribution with  $p$  degrees of freedom. The estimated value  $\hat{\beta}$  in the test statistics is the estimate of the covariate vector derived from the Cox model under  $H_0$  (i.e. under the proportional hazard assumption).

The Schoenfeld residuals are plotted on the right hand side graph of Figure 1.1 together with the smoothed estimation of the trend (the smoothed line is a natural spline fit indicating a departure from proportionality once it is not flat and straight). The smoothed line based on the Schoenfeld residuals suggests that the covariate  $\beta_j$  should linearly decrease towards zero in time. The p-value of the  $\chi^2$  test of the non-zero slope was 0.02.

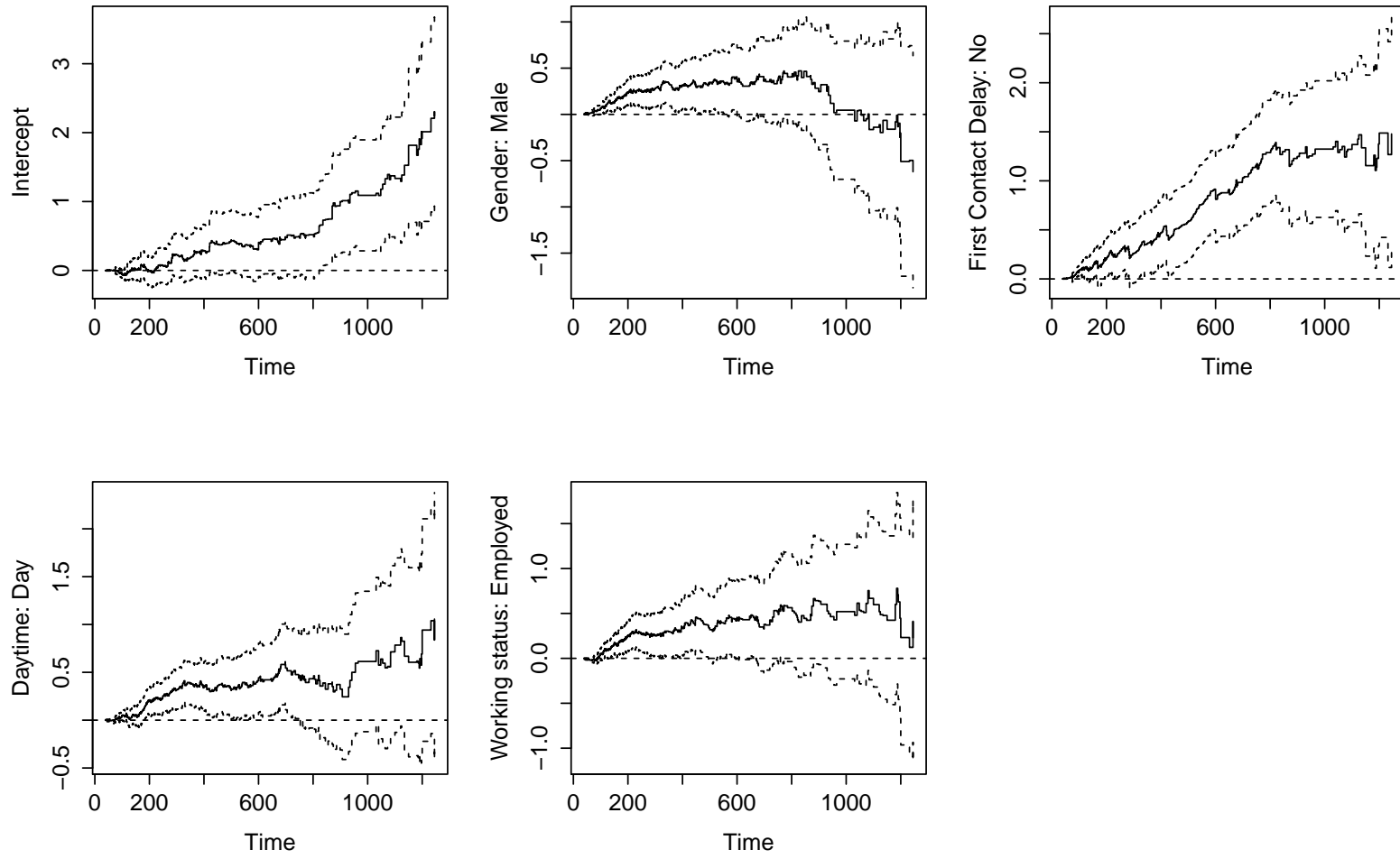


Figure 1.2: The estimated cumulative regression functions from the classic Aalen model for the time-delay.

On the left hand side graph of Figure 1.1 the logarithms of the estimated hazard rates for the two genders under the proportional hazard assumption is plotted. Apparently, the difference of the plotted lines is fairly big up to 600 minutes and then decreases monotonically in time, signaling the slight departure from the proportionality. This is in agreement with  $g_j$  being a decreasing linear function as indicated by the graph on the right hand side.

As the next step we fitted the classic Aalen model using the OLS technique. The estimated regression functions are plotted in Figure 1.2. Here all the cumulative regression functions' estimates, but the one corresponding to male gender, are positive. The estimated cumulative regression function for male grows the most intensively within the first 200 minutes and then exhibit only slight increase over the time. At the end of the time window, where only few observations contribute to the estimation, the decline of the estimated cumulative hazard function is distinctive. This can be given by the general instability of the estimates at the end of the observation window. It is also possible that this particular dataset exhibit a surprising twist in the patients' behaviour in comparison to other datasets analysed within this field. How likely is this to happen is to e.g. the cardiologists who run the study to assess. A monotone Aalen model can be of use if the decline at the end is attributed to the uncertainty at the end of the time window and if there is for various reasons a need for a nonnegative estimator.

## 1.6 Outline of the thesis

The whole thesis is devoted to the estimation in monotone Aalen model. The core of the developed methods is contained in three chapters. The main focus is aimed at the well-known approaches to the estimation within the survival data analysis, as is the nonparametric maximum likelihood and Bayesian analysis using the Beta process in Chapters 2 and 3. This work is a topic of a paper which will be submitted in imminent future. Chapter 4 is again devoted to the Bayesian modelling, although using different approach. The results of this chapter are summarized in the paper accepted for publication, [44]. In Chapter 5 we compare the results of the methods proposed in Chapters 2, 3 and 4 to the classic Aalen estimators using the famous Danish malignant melanoma dataset and we revisit the real data example introduced in Section 1.5.

In Chapter 2 and Chapter 3 we deal with estimation of integrated regression functions  $A_j$  in monotone Aalen additive model for right-censored data and we apply the nonparametric maximum likelihood method (NPML) and Bayesian

approach with Beta process prior. It is not known to authors that this kind of analysis has been done before and therefore in the first place it is vital to supply the existing reservoir of available methods. Secondly, as it turns out, for  $p > 1$  these estimators are inconsistent – an interesting and certainly disturbing result on its own, showing that two solid approaches like NPML and Bayesian analysis can mislead us. Both methods were based on the assumption of the discontinuity of the regression functions, i.e. the likelihood of the data used in deriving the estimators was

$$\prod_{i=1}^n \prod_{s>0} \left\{ \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j}(Y_i(s) - dN_i(s))} \left( 1 - \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j}} \right)^{dN_i(s)} \right\}. \quad (1.11)$$

The maximum likelihood estimation often leads to optimal estimators, as it is for example Nelson-Aalen estimator, further extended to Kaplan-Meier estimator and the baseline hazard estimator in Cox model, all of them possessing nice asymptotic features. Further, let us recall that in a finite-dimensional case and for regular systems the maximum likelihood estimator (MLE) is asymptotically efficient with variance reaching the Cramér-Rao's lower bound, when the sample size grows. A good comment on optimality of Nelson-Aalen estimator as the nonparametric maximum likelihood estimator (NPMLE) can be found in [4], section VIII.4.5. The reasoning goes along these lines: Nelson-Aalen as a solution of maximum likelihood technique is asymptotically equivalent to a linear combination of score equations, hence this estimator is asymptotically linear and under some additional assumptions on regularity this linearity induces asymptotic efficiency. Interestingly, in our case of the monotone Aalen model it is the method of weighted least squares which produces optimal estimator, not NPML.

We derive the NPML estimator for the monotone Aalen model in Chapter 2. We give the consistency and the asymptotic distribution results of this estimator for a case when the hazard rate equals  $h_i(t) = \alpha_1(t)x_{i,1}$ . Then we move on to a general case with  $p > 1$  and we explain how a rather complicated estimator is obtained. We show that this estimator converges to a function that does not equal the true unknown regression function. However, it is then demonstrated that there is present a feature we would like to call an average consistency even in the general  $p > 1$  case.

In Chapter 3 we devote to Bayesian analysis using the Beta processes. First we introduce the NII processes used as priors in survival data analysis and give details on the Beta processes. We derive the posterior distribution of the regression parameters under the Beta process prior and propose the Bayesian estimator as the expectation of this distribution. Then the consistency results alongside with



a Bernstein-von Mises type of statement are given for  $p = 1$  case. Similarly as in NPML case, the Bayesian estimators of the regression functions prove to be inconsistent for  $p > 1$ . Interestingly, they converge to a different function than the NPML estimator although of a similar type. Further, we generalize this result by showing that a Bayesian estimator based on any NII process as a prior is inconsistent. We finish the section by explaining an algorithm which can be used to produce the estimators.

The motivation to seek a different way to estimation of regression parameters is obvious. Therefore the next step is to conduct different Bayesian modelling assuming that unlike in (1.11) the regression functions are continuous, i.e.  $\alpha_j$ -s exist, and a piecewise constant process suggested by Arjas and Gasbarra [6] is applied as prior. Such approach is utilized in Chapter 4. The method approximates the baseline hazard rate and the regression functions using piecewise constant functions with a random number and locations of jump times. The process which is used as prior to regression function is explained and the posterior distribution is derived and followed by explanation of the MCMC algorithm used for estimation. The resulting estimators are of rather complicated structure not allowing one to directly assess if the estimators are consistent. Therefore we conduct a simulation study to explore the performance of the estimators.

In Chapter 5 we revisit the STEMI dataset introduced in Section 1.5 and demonstrate the proposed estimators in comparison to the classical least squares estimators. We also show the performance of the estimators on the Danish melanoma data.

The thesis is concluded with a discussion of the achieved results in Chapter 6.

## 1.7 Current state on cases of inconsistent estimators

As one of the results of this thesis is yet an example of the inconsistent estimators, we conclude this chapter with a summary on the current state of the knowledge concerning the inconsistency of the estimators developed within the time-to event framework. Obviously, the consistency of a estimator is essential to statisticians as otherwise instead of being sure that we are getting closer and closer to the truth with growing sample size, when the consistency is lacking we are more and more sure about being wrong and moreover not knowing how much.

Generally, the nonparametric maximum likelihood as a generalization of the classic maximum likelihood method to infinitely-dimensional problems not always carries along nice asymptotical features. Particularly, in survival analysis the results on consistency of a nonparametric maximum likelihood estimator (further as NPMLE) vary with different settings. Furthermore, some rather weak conditions in addition to consistency can provide a gate to asymptotic normality – e.g. Gill et al. in [16] using the von Mises method.

For *right-censored* data the results have been so far comforting. The NPMLE in the multiplicative intensity model without covariates is essentially equal to Nelson-Aalen estimator (see Johansen [23] and Jacobsen’s work [22]) and the same stands for the baseline hazard function in Cox model. The consistency of these estimators is fulfilled automatically from the asymptotic results of the Nelson-Aalen estimator and the Breslow estimator. Zeng and Lin [50] proposed an approximate NPML method for accelerated failure time models leading to consistent estimators when some regularity conditions are fulfilled.

The estimation gets complicated when we deal with other censoring schemes or a truncation is present. For the *interval-censored* data a nice overview was done by Huang and Wellner [20], saying that for i.i.d. data the NPMLE is consistent in both ”case 1” (current status data) and ”case k” interval censoring. However, the ML estimation for the joint distribution of ”case k” interval-censored survival times and continuous marks as i.i.d. pairs gives an inconsistent estimator (see [34]). Pan and Chappell in [36] accommodated ”case 1” interval-censored data with left truncation and pointed out that the NPMLE in such model is inconsistent while the estimator based on conditioning on truncating times proved to be consistent. Interestingly, purely left-truncated data (as well as only interval-censored data) induces consistent NPMLE, [48]. Huang and Wellner in their paper [20] dealt also with regression model under both interval-censoring schema in Cox and accelerated failure time models. They got consistency of NPMLE for both models under reasonable conditions in both ”case 1” and ”case k” interval censoring.

For *doubly-censored* data the consistency was shown by Turnbull [46] under the assumption that the times are observed only on a finite set (i.e. the check-ups are of a finite number).

Other cases of inconsistency emerged when a likelihood is maximized with respect to a constraint, e.g. the NPMLE of a distribution with increasing failure rate average, [8], or the NPMLE of a distribution which is uniformly stochastically smaller than a beforehand obtained standard, [39].

From the Bayesian point of view, the consistency of a Bayesian estimator is often impossible to check as the posterior distributions are often of very complicated structures. Still there have been attempts to derive the asymptotic features of some Bayesian estimators. The inconsistency problem within the Bayesian framework was discussed in various papers, among others especially by Diaconis and Freedman [11] and Ghosal [15]. Generally, in parametric models the convergence to the true value of parameter is ensured as long as the prior distribution is properly chosen. However, introducing a functional parameter into a model brings along difficulties not unfamiliar to statisticians (similarly for maximum likelihood method etc.). Wu and Ghosal [49] proved a fairly general result on consistency for some semiparametric cases, i.e. the Cox model, accelerated failure time models, under a Levy process prior. Less optimistic result was obtained by Zhou [51] who showed that for doubly or interval censored i.i.d. data a Bayesian estimator under Dirichlet process prior did not have the same limit as the NPML estimator. Although he did not directly discuss the consistency itself, the result is troubling as often Bayesian and ML techniques lead to an identical result asymptotically.

## Chapter 2

# Nonparametric maximum likelihood estimation

In principle, the maximum likelihood estimation in parametric case estimates the parameters of assumed distribution or model structure by finding such values that the corresponding distribution has most likely produced the studied data. This approach has been proven to be very efficient in most situations when an unknown finite dimensional quantity needs to be estimated.

When we move from a parametric to a nonparametric setting, then instead of estimating the unknown parameters of the distribution we seek the distribution itself from a class of suitable functions. The sought entity can be in form of a distribution function, a density or a (cumulative) hazard function. As we know, estimation of this kind can preserve nice asymptotic properties. The classic example would be an empirical distribution function, which by the Glivenko-Cantelli theorem converges to the sought distribution function uniformly on  $\mathbb{R}$ . Furthermore, a process equal to a difference of an empirical and a true distribution function multiplied by square root of the sample size converges in distribution to a zero-mean Gaussian process. Within the field of the survival data analysis, we would be interested in finding an estimator of a hazard function or a survival function. Already mentioned Kaplan-Meier estimator  $S^{K-M}$  of a survival function  $S$  in homogeneous case is a product limit estimator made up by inserting the Nelson-Aalen estimator of a cumulative hazard rate into the expression on the right hand side of the (1.2). It is again a nonparametric estimator based on no distributional assumptions and is uniformly consistent on interval  $[0, \tau]$  in which the true survival function and the censoring survival function is non-zero. And indeed, the process  $\sqrt{n}(S^{K-M}(t) - S(t))$  converges weakly, on interval  $[0, \tau]$ , to  $-S \cdot U$  where  $U$  is a zero-mean Gaussian martingale. For details on both consistency and weak convergence see e.g. [4], Section IV.3.2.

Even though we often suppose that our data comes from a continuous distribution, the classic nonparametric estimators like the empirical distribution function or the Kaplan-Meier estimator are discontinuous functions (step functions in particular). So when looking for a nonparametric estimator of the cumulative regression functions in Aalen model, we do not seek a suitable estimator in the class of continuous functions only. On the contrary, we enlarge the class of admissible functions by including discontinuous distributions into the original model and equip the extended space with suitable topology. The maximum likelihood estimator then is found by maximizing the probability of the observations in this extended space. The extension can be done in many ways and often the resulting NPMLs equal asymptotically, see e.g. [22]. Here, let us seek the estimator in the class of the càdlàg functions  $D[0, \tau]$  endowed with Skorohod topology.

Next section introduces the reformulation of the Aalen model in the way that possible discontinuities in the distribution function are accommodated. More importantly we state several assumptions that will hold throughout whole Chapter 2 and Chapter 3. In Section 2.2 we introduce a NPML estimator for a simplified case of  $p = 1$  when the hazard rate equals  $h_i(t) = \alpha_1(t)x_{i,1}$ . We show that the proposed NPML estimator is uniformly consistent and weakly convergent process. We move onto the general case of  $p > 1$  in Section 2.3, where we derive the form of the NPML estimator and we prove that it is an inconsistent estimator. The NPML estimator, however, exhibits a so called average consistency feature and we deal with this matter in Section 2.4. The chapter is finished with a simulated example in Section 2.5.

## 2.1 Formulation of the model

As we work in an extended model we need to switch to *time-discrete framework* which is suitable for characterization of both discrete and continuous distributions, see e.g. [19]. Using the relationship between  $H_i$  and  $F_i$  in (1.2), we get that the survival function for  $i$ th observation equals

$$S_i(t) = 1 - F_i(t) = \prod_{[0,t]} \{1 - dH_i(s)\}.$$

Accommodating the formula (1.6) for the cumulative hazard function  $H_i$  under the monotone Aalen model to the discontinuous setting gives us

$$1 - dH_i(t) = \prod_{j=1}^p \{1 - dA_j(t)\}^{x_{i,j}}. \quad (2.1)$$

Hence, the likelihood of a sample of the time-to-event observations, transformed to counting processes  $(N_i, Y_i, z_i)_{i=1}^n$ , is in the time-discrete framework equal to

$$\prod_{i=1}^n \prod_{s>0} \{1 - dH_i(s)\}^{Y_i(s) - dN_i(s)} dH_i(s)^{dN_i(s)}$$

and under the monotone Aalen model it becomes

$$\prod_{i=1}^n \prod_{s>0} \left\{ \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j}(Y_i(s) - dN_i(s))} \left( 1 - \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j}} \right)^{dN_i(s)} \right\}. \quad (2.2)$$

If  $k$ th individual dies in  $[t, t + dt]$ , the contribution to the likelihood is

$$\prod_{j=1}^p \{1 - dA_j(t)\}^{R_j(t) - x_{k,j}} \left( 1 - \prod_{j=1}^p \{1 - dA_j(t)\}^{x_{k,j}} \right).$$

Here we denoted  $R_j(t) = \sum_{i=1}^n x_{i,j} Y_i(t)$ . We will work with this formulation of the likelihood of the data in the whole Chapter 2 and also in Chapter 3.

**Assumption (\*):** Before moving to the derivation of the estimators let us make assumptions which we suppose to hold throughout the whole Chapter 2 and Chapter 3. Firstly, we expect that the covariate vectors  $z_1, \dots, z_n$  are independent and ruled by an unknown distribution. In general we say that any  $z_i, i = 1, \dots, n$  is distributed similarly as a random vector  $z$ . In particular we say that every component  $x_{i,j}$  of  $z_i$  for any  $i$  is distributed as a random vector  $x_j$ . The distribution of  $x_j, j = 1, \dots, p$ , cannot be degenerated in zero.

We assume that the true underlying distribution function  $F_i$  for  $i = 1, \dots, n$  is absolutely continuous, i.e. there exist densities  $f_i$ -s and hazard functions  $h_i$ -s. Hence, there exist regression functions  $\alpha_j, j = 1, \dots, p$ . Furthermore, as  $\tau$  is the maximum observed time we suppose that there is at least one  $i$  such, that  $S_i(\tau) = 1 - F_i(\tau) > 0$ .

Next, let us suppose that the censoring mechanism is independent and the censoring times  $C_i$  are distributed according to a distribution with distribution function  $G(t)$ . Finally, we want that  $\overline{G}(\tau) = 1 - G(\tau_-) > 0$  for all  $t < \tau$ .

The assumption on the absolute continuity of the survival times distribution  $F_i$  might seem unnecessarily strict, considering that we will work in time-discrete framework and with discontinuous estimators in both Chapter 2 and Chapter 3. The results obtained in following chapters could be restated for distributions with discontinuities, however, the character of the counting processes theory and hence

also the theory developed in this work would become more complex. Due to the unfavourable results we eventually decided to stay within the range of simpler case with the absolutely continuous distributions.

## 2.2 Case $p = 1$

Let us start with the simplest case when only one covariate is considered and we observe a dataset of following form  $(N_i, Y_i, x_{i,1})_{i=1}^n$  on a fixed time interval  $[0, \tau], \tau < \infty$ . As we will see later, the outcome is rather different for a single covariate case and a multiple covariate case. The formula for the cumulative hazard rate abiding the Aalen model simplifies into  $H_i(s) = x_{i,1}A_1(s)$  and there is only one functional parameter  $A_1$  which needs to be estimated. Let us mention that in this case all covariates  $x_{i,1}, i = 1, \dots, n$ , need to be non-zero and positive. Assuming the discontinuous cumulative hazard rate, we have

$$1 - dH_i(s) = (1 - dA_1(s))^{x_{i,1}}, \quad i = 1, \dots, n,$$

and we seek a nonparametric maximum likelihood estimator for the unknown parameter  $A_1(t) = \int_0^t \alpha_1(s)ds$  by maximizing the likelihood in (2.2). We can do that separately for each  $s \in [0, \tau]$ . First let us consider the times between the observed failure times, i.e. when  $\sum_i dN_i(s) = 0$ . Then the likelihood is of the following form

$$\{1 - dA_1(s)\}^{R_1(s)},$$

where  $R_1(s) = \sum_{i=1}^n x_{i,1}Y_i(s)$ . Clearly, this expression is maximised for  $dA_1(s) = 0$ . In next we look into the times when one of the processes jumped, e.g. the  $i$ -th individual experienced the event and  $dN_i(s) = 1$ . This leads us to seeking maxima of the following expression

$$\{1 - dA_1(s)\}^{R_1(s) - x_{i,1}} [1 - \{1 - dA_1(s)\}^{x_{i,1}}].$$

It can be shown that within the admissible domain  $[0, 1]$  this function gains its maximal value when  $dA_1(s)$  equals

$$1 - \left\{1 - \frac{x_{i,1}}{R_1(s)}\right\}^{1/x_{i,1}} =: \hat{a}_{i,1}(s), \quad (2.3)$$

where we denoted the estimator of the jump at  $s$  by  $\hat{a}_{i,1}(s)$ . Note, that the estimator  $\hat{a}_{i,1}$  is valid for assumed  $x_{i,1} > 0$  and is contained in  $(0, 1)$  as long as

$x_{i,1} < R_1(s)$ . If the event observed in the very right-end of the time window  $\tau$  is a failure, then  $x_{i,1} = R_1(s)$  and the estimator in this time point equals  $\hat{a}_{i,1}(\tau) = 1$ .

Finally, we have a nonparametric maximum likelihood estimator for  $A_1$  of following form

$$\hat{A}_1(t) = \sum_{i=1}^n \int_0^t \hat{a}_{i,1}(s) dN_i(s). \quad (2.4)$$

Representing the estimator  $\hat{a}_{i,1}$  as a function of  $x_{i,1}$  by Taylor series at  $x_{i,1} = 1$  and using the fact that  $\log x \approx x - 1$  for  $x \approx 1$  gives

$$\hat{a}_{i,1}(s) = 1 - \left\{ 1 - \frac{x_{i,1}}{R_1(s)} \right\}^{1/x_{i,1}} = \frac{1}{R_1(s)} + \frac{x_{i,1} - 1}{2R_1^2(s)} + O_P\left(\frac{1}{R_1^3(s)}\right). \quad (2.5)$$

Approximating the estimator by two first members of the Taylor series is valid for  $x_{i,1}$  bounded and  $R_1 \gg 0$  and becomes more and more accurate with growing sample size. The behaviour of the approximation is exhibited in Figure 2.1, where the estimator  $\hat{a}_{i,1}$  from (2.3) is drawn in the solid line, the dotted line represents the second-order approximation from (2.5) and the dashed line shows the first-order approximation by  $1/R_1(s)$ . This is shown for  $R_1 = 10, 50, 100, 200$ . We have in mind here, that larger value of  $R_1$  implies greater sample size under the assumption that  $x_{i,1}$  are i.i.d. for all  $i$ . Clearly, greater  $R_1$  in comparison to  $x_{i,1}$  indicates higher accuracy of the approximation.

Using the approximation from (2.5) and assuming that the covariate distribution has finite  $E|x_1|^2$  we get that the NPML estimator equals

$$\hat{A}_1(t) = \int_0^t \frac{d\bar{N}(s)}{R_1(s)} + \epsilon_n, \quad \text{where } \epsilon_n = o_p(1/n^{1/2}).$$

Let us explain the order of the convergence in probability to zero of the remainder term  $\epsilon_n$ . Under a closer inspection we reveal that

$$\left| \sum_{i=1}^n \int_0^t \frac{1}{2} \frac{x_{i,1} - 1}{R_1^2(s)} dN_i(s) \right| \leq \frac{1}{2} \frac{n \max_{1 \leq k \leq n} x_{k,1}}{R_1^2(t)} \leq \text{const.} \frac{\max_{1 \leq k \leq n} x_{k,1}}{R_1(t)}$$

as  $R_1(s) \geq R_1(t)$  for  $s \leq t$ . If  $x_1$  is bounded then  $R_1(t) = O_p(n)$ , and this implies  $\epsilon_n = O_p(1/n)$ . Furthermore, we have  $\epsilon_n = o_p(1/n^{1/2})$  even when the covariates are not bounded but have finite second moments, i.e. when  $E|x_1|^2 < \infty$ .

Let us look at the proposed model from a slightly different angle. Transforming the covariate into  $x_{i,1} = \exp\{\beta w_{i,1}\}$  for some arbitrary  $\beta \in \mathbb{R} \setminus 0$ , we get the familiar expression of the hazard rate for one-covariate Cox model, see (1.8), with an unknown baseline hazard  $\alpha_1$  and a known one-dimensional parameter  $\beta$ . In the reformulated model we then have that the hazard rate for  $i$ -th individual



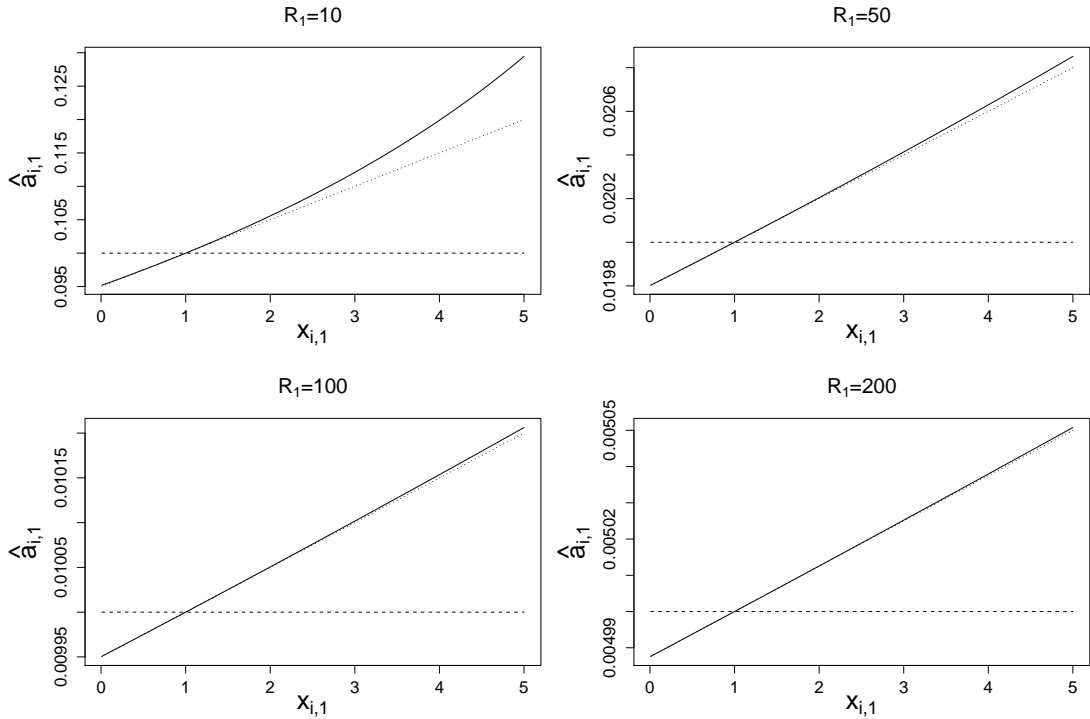


Figure 2.1: In the solid line the estimator  $\hat{a}_{i,1}$  from (2.3) is drawn, the dotted and the dashed line represents the second-order approximation from (2.5) and the first-order approximation by  $1/R_1(s)$ , respectively.

equals  $h_i(s) = \alpha_1(s)e^{\beta w_i}$ . We only need to estimate the cumulative baseline hazard rate  $A_1 = \int \alpha_1$  and using the Breslow estimator (1.9) introduced in Section 1.4, we have that the estimator of the cumulative baseline hazard rate equals to  $\int_0^t d\bar{N}(s) / \sum_{i=1}^n \exp\{\beta w_{i,1}\} Y_i(s) = \int_0^t d\bar{N}(s) / R_1(s)$ .

It is seen that for large  $n$  the proposed NPML estimator in Aalen model and the Breslow estimator are equivalent. We already know that the Breslow estimator of the cumulative baseline hazard function is weakly convergent, what gives us a certain idea about the existence of positive asymptotic results for the proposed NPML estimator in (2.4).

Let us start with the uniform consistency of the NPML estimator, which is stated in the next theorem.

**Theorem 4 (Consistency)** *Let us suppose that the conditions in Assumption (\*) in Section 2.1 are fulfilled. If  $\bar{Y}(\tau) \xrightarrow{P} \infty$  with  $n \rightarrow \infty$ ,  $\alpha_1$  is continuous and  $E|x_1|^2$  is finite, then the estimator  $\hat{A}_1$  is an uniformly, on  $[0, \tau]$ , consistent estimator of the true  $A_1$ .*

**PROOF** The proof of consistency utilizes the knowledge on the martingale theory for counting processes which is summarized in detail in Section 1.2. We need to

show that

$$\sup_{0 \leq t \leq \tau} \left| \hat{A}_1(t) - A_1(t) \right| \xrightarrow{P} 0.$$

We have that  $Y_i(s)x_{i,1}\alpha_1(s)ds$  is a compensator for  $dN_i(s)$  and  $M_i(s) = N_i(s) - \int_0^s Y_i(u)x_{i,1}\alpha_1(u)du$  is a martingale. Hence we can write

$$\left| \hat{A}_1(t) - A_1(t) \right| \leq \left| \left( \int_0^t \sum_{i=1}^n \hat{a}_{i,1}(s)Y_i(s)x_{i,1} - 1 \right) \alpha_1(s)ds \right| + \left| \int_0^t \sum_{i=1}^n \hat{a}_{i,1}(s)dM_i(s) \right|. \quad (2.6)$$

For the proof it is sufficient to show that both terms on the right side uniformly converge to 0 in probability. Using the representation for  $\hat{a}_{i,1}$  from (2.5) we have

$$\begin{aligned} & \sup_{0 \leq t \leq \tau} \left| \int_0^t \left( \sum_{i=1}^n \hat{a}_{i,1}(s)Y_i(s)x_{i,1} - 1 \right) \alpha_1(s)ds \right| \\ &= \sup_{0 \leq t \leq \tau} \left| \int_0^t \sum_{i=1}^n \left( \frac{1}{R_1(s)} + \frac{1}{2} \frac{x_{i,1} - 1}{R_1^2(s)} + O_p \left( \frac{1}{R_1^3(s)} \right) \right) Y_i(s)x_{i,1}\alpha_1(s)ds \right. \\ & \quad \left. - \int_0^t \alpha_1(s) ds \right| \\ &= \sup_{0 \leq t \leq \tau} \left| \int_0^t \sum_{i=1}^n \left( \frac{1}{2} \frac{x_{i,1} - 1}{R_1^2(s)} + O_p \left( \frac{1}{R_1^3(s)} \right) \right) Y_i(s)x_{i,1}\alpha_1(s)ds \right| \\ &\leq \sup_{0 \leq t \leq \tau} \left| \int_0^t \sum_{i=1}^n \frac{1}{2} \frac{\max_{1 \leq k \leq n} x_{k,1}}{R_1^2(s)} Y_i(s)x_{i,1}\alpha_1(s)ds \right| \\ & \quad + \sup_{0 \leq t \leq \tau} \left| \int_0^t O_p \left( \frac{1}{R_1^2(s)} \right) \alpha_1(s)ds \right| \\ &\leq \frac{1}{2} \frac{\max_{1 \leq k \leq n} x_{k,1}}{R_1(\tau)} A_1(\tau) + \int_0^\tau O_p \left( \frac{1}{R_1^2(s)} \right) \alpha_1(s)ds \end{aligned}$$

where under the assumptions  $E|x_1|^2 < \infty$ ,  $\alpha_1$  continuous and  $\bar{Y}(\tau) \xrightarrow{P} \infty$  the first term is  $o_p(1/n^{1/2})$  and the second converges to 0 in probability with  $n \rightarrow \infty$ . Hence, the first part of the right side of (2.6) converges to 0 in probability.

To show that

$$\sup_{0 \leq t \leq \tau} \left| \int_0^t \sum_{i=1}^n \hat{a}_{i,1}(s)dM_i(s) \right| \xrightarrow{P} 0$$

we use a corollary of Lenglart inequality ([14], Corollary 3.4.1.) which says that for any  $\epsilon, \eta > 0$ , the following relation is true

$$P \left\{ \sup_{0 \leq t \leq \tau} \left| \int_0^t \sum_{i=1}^n \hat{a}_{i,1}(s)dM_i(s) \right| > \eta \right\} \leq \frac{\epsilon}{\eta^2} + P \left\{ \int_0^\tau \sum_{i=1}^n \hat{a}_{i,1}^2(s)d\langle M_i \rangle(s) > \epsilon \right\}.$$

Here we use the fact that  $\hat{a}_{i,1}$  is adapted and predictable (from left-continuity) and  $\int \sum_i \hat{a}_{i,1} dM_i$  is therefore a martingale. For the variance process we have

$$\int_0^\tau \sum_{i=1}^n \hat{a}_{i,1}^2(s) d\langle M_i \rangle(s) = \int_0^\tau \sum_{i=1}^n \hat{a}_{i,1}^2(s) Y_i(s) x_{i,1} \alpha_1(s) ds.$$

Using the result in (2.5) we have

$$\hat{a}_{i,1}^2(s) = \frac{1}{R_1^2(s)} + \frac{x_{i,1} - 1}{R_1^3(s)} + O_p\left(\frac{1}{R_1^4(s)}\right),$$

hence the variance process is for large  $n$  close to

$$\int_0^\tau \sum_{i=1}^n \frac{1}{R_1^2(s)} Y_i(s) x_{i,1} \alpha_1(s) ds = \int_0^\tau \frac{1}{R_1(s)} \alpha_1(s) ds.$$

The latter converges to 0 in probability as long as  $\bar{Y}(\tau) \xrightarrow{P} \infty$  and this concludes the proof.

Q.E.D.

In following theorem we state the weak convergence result of  $\sqrt{n}(\hat{A}_1 - A_1)$  to a zero-mean Gaussian process.

**Theorem 5 (Asymptotic distribution)** *Let us suppose that the conditions in Assumption (\*) in Section 2.1 are fulfilled. If  $\bar{Y}(\tau) \xrightarrow{P} \infty$  with  $n \rightarrow \infty$ ,  $\alpha_1$  is continuous,  $E|x_1|^2$  is finite and if there exists a function  $r_1(s)$  on  $[0, \tau]$  such that*

$$\sup_{0 \leq s \leq \tau} \left| \frac{R_1(s)}{n} - r_1(s) \right| \xrightarrow{P} 0,$$

*then the process  $\sqrt{n}(\hat{A}_1 - A_1)$  converges weakly to a zero-mean Gaussian process with covariate process equal to*

$$C(t) = \int_0^t \frac{\alpha_1(s)}{r_1(s)} ds,$$

*consistently estimated by*

$$\hat{C}(t) = \int_0^t \sum_{i=1}^n \frac{n}{R_1^2(s)} dN_i(t).$$

**PROOF** The proof again relies on the martingale theory explained in Section 1.2. Utilizing the Doob-Meyer decomposition we can expand the expression of interest

into two parts,

$$\begin{aligned} \sqrt{n}(\hat{A}_1(t) - A_1(t)) &= \sqrt{n} \int_0^t \sum_{i=1}^n \hat{a}_{i,1}(s) dM_i(s) \\ &\quad + \sqrt{n} \int_0^t \left( \sum_{i=1}^n \hat{a}_{i,1}(s) Y_i(s) x_i - 1 \right) \alpha_1(s) ds. \end{aligned} \quad (2.7)$$

We need to show that the first term converges to a zero-mean Gaussian martingale with the variance process equal to  $C(t)$  while the latter converges uniformly in probability to zero, what is done using similar arguments as in the proof of consistency of  $A_1$ .

To handle the first part of the task we use the Rebolledo martingale central limit theorem (see Theorem 1). Here, the examined process is

$$V^{(n)}(t) = \sqrt{n} \int_0^t \sum_{i=1}^n \hat{a}_{i,1}(s) dM_i(s).$$

Since the processes  $\hat{a}_{i,1} = 1 - (1 - x_{i,1}/R_1(s))^{1/x_{i,1}}$  are adapted, predictable (all elements are known at  $s_-$ ) and bounded by 1, the process  $V^{(n)}$  is a martingale. When we apply the knowledge from Section 1.2 and the familiar representation for  $\hat{a}_{i,1}$  from (2.5), we get

$$\begin{aligned} \langle V^{(n)}, V^{(n)} \rangle(t) &= \int_0^t n \sum_{i=1}^n \hat{a}_{i,1}^2(s) Y_i(s) x_{i,1} \alpha_1(s) ds \\ &= \int_0^t n \sum_{i=1}^n \left( \frac{1}{R_1^2(s)} + \frac{x_{i,1} - 1}{R_1^3(s)} + O_p\left(\frac{1}{R_1^4}\right) \right) Y_i(s) x_{i,1} \alpha_1(s) ds \\ &= \int_0^t \frac{1}{\frac{R_1(s)}{n}} \alpha_1(s) ds + \int_0^t n \left( \frac{x_{i,1} - 1}{R_1^2(s)} + O_p\left(\frac{1}{R_1^3}\right) \right) \alpha_1(s) ds. \end{aligned}$$

From the assumption that  $\sup_{0 \leq s \leq \tau} |R_1(s)/n - r_1(s)| \xrightarrow{P} 0$  together with the continuous mapping theorem (see Theorem 2) we have that

$$\frac{1}{\frac{R_1(s)}{n}} \xrightarrow{P} \frac{1}{r_1(s)}, \quad \text{as } n \rightarrow \infty. \quad (2.8)$$

Let us reformulate the previous into

$$\begin{aligned} \langle V^{(n)}, V^{(n)} \rangle(t) &= \int_0^t \frac{\alpha_1(s)}{r_1(s)} ds + \int_0^t \left( \frac{1}{\frac{R_1(s)}{n}} - \frac{1}{r_1(s)} \right) \alpha_1(s) ds \\ &\quad + \int_0^t n \left( \frac{x_{i,1} - 1}{R_1^2(s)} + O_p\left(\frac{1}{R_1^3}\right) \right) \alpha_1(s) ds. \end{aligned} \quad (2.9)$$

From (2.8) it is seen, that the second term of (2.9) converges to 0 in probability with  $n \rightarrow \infty$ . For  $E|x_1| < \infty$ , the expression

$$\frac{x_{i,1} - 1}{R_1^2(s)} + O_p\left(\frac{1}{R_1^3(s)}\right)$$

is  $o_p(1/n)$  and the third term of (2.9) converges to 0 in probability with  $n \rightarrow \infty$ . This concludes the convergence of the predictable variance process of  $V^{(n)}$  to the process  $C(t)$  in probability.

Furthermore, we need to show that the predictable variance process of the process

$$V_\epsilon^{(n)} = \int_0^t \sum_{i=1}^n \sqrt{n} \hat{a}_{i,1}(s) I_{\{\sqrt{n} \hat{a}_{i,1}(s) \geq \epsilon\}} dM_i(s), \quad \text{for any } \epsilon > 0,$$

converges to zero. From the same arguments as before

$$\langle V_\epsilon^{(n)}, V_\epsilon^{(n)} \rangle(t) = \int_0^t n \sum_{i=1}^n \hat{a}_{i,1}^2(s) I_{\{\sqrt{n} \hat{a}_{i,1}(s) \geq \epsilon\}} Y_i(s) x_{i,1} \alpha_1(s) ds.$$

We already showed that  $n \sum_{i=1}^n \hat{a}_{i,1}^2(s) Y_i(s) x_{i,1} \alpha_1(s)$  converges in probability to  $\alpha_1(s)/r_1(s)$  and this limit is bounded under the assumption  $\bar{Y}(\tau) \xrightarrow{P} \infty$ . This, together with the fact that  $\hat{a}_{i,1}$  is  $o_p(1/n^{1/2})$ , gives  $\langle V_\epsilon^{(n)}, V_\epsilon^{(n)} \rangle(t) \xrightarrow{P} 0$ .

Hence, we showed that the first term in (2.7) is asymptotically distributed as a zero-mean Gaussian process with covariance process equal to  $C(t)$  and we just need to prove that the second term converges uniformly to 0 in probability. As we already exhibited in Proof of Theorem 4, the expression

$$\begin{aligned} \sup_{0 \leq t \leq \tau} \left| \int_0^t \left( \sum_{i=1}^n \hat{a}_{i,1}(s) Y_i(s) x_{i,1} - 1 \right) \alpha_1(s) ds \right| \\ \leq \frac{1}{2} \frac{\max_{1 \leq k \leq n} x_{k,1}}{R_1(\tau)} A_1(\tau) + \int_0^\tau O_p\left(\frac{1}{R_1^2(s)}\right) \alpha_1(s) ds \end{aligned}$$

and under the assumptions of Theorem 5 the first term is  $o_p(1/n^{1/2})$  while the second is  $O_p(1/n^2)$ . From this we can conclude that the second expression in (2.7) converges to 0 in probability as  $n \rightarrow \infty$ .

The very last is the proof that we can consistently estimate the covariance process  $C(t)$  by

$$\hat{C}(t) = \int_0^t \sum_{i=1}^n \frac{n}{R_1^2(s)} dN_i(s).$$

Similarly as before,

$$\begin{aligned}
\left| C(t) - \hat{C}(t) \right| &= \left| \int_0^t \frac{\alpha_1(s)}{r_1(s)} ds - \int_0^t \sum_{i=1}^n \frac{n}{R_1^2(s)} dN_i(s) \right| \\
&= \left| \int_0^t \frac{\alpha_1(s)}{r_1(s)} ds - \int_0^t \sum_{i=1}^n \frac{n}{R_1^2(s)} Y_i(s) x_{i,1} \alpha_1(s) ds - \int_0^t \sum_{i=1}^n \frac{n}{R_1^2(s)} dM_i(s) \right| \\
&\leq \left| \int_0^t \left( \frac{1}{r_1(s)} - \frac{1}{\frac{R_1(s)}{n}} \right) \alpha_1(s) ds \right| + \left| \int_0^t \sum_{i=1}^n \frac{n}{R_1^2(s)} dM_i(s) \right|.
\end{aligned}$$

We already know that the first term converges to 0 in probability. The convergence of the martingale in the second term to 0 in probability with growing  $n$  is easily proven by applying the Lenglart inequality.

Q.E.D.

The convergence of NPMLE can be very slow if the covariates are close to 0. This is because  $R$  grows very slowly with  $n$  and the approximation of  $\hat{a}_{i,1}$  works well only for  $R$  large.

**Remark 1** A sufficient condition for the existence of a continuous  $r_1(s)$  is that the covariates  $x_{i,1}$  are i.i.d. and distributed according to  $x_1$  with  $E x_1 < \infty$ . This is seen from *the uniform law of large numbers*. Let us have a function  $f(x; s)$ , defined for  $(x, s) \in \mathcal{X} \times S$ , a Cartesian product of a Euclidean set  $\mathcal{X}$  and a compact set  $S$ . The uniform law of large numbers states that for such function, if the following conditions are fulfilled

1.  $f(x; s)$  is continuous at every fixed  $s \in S$  for almost all  $x$ ,
2. there exists a dominating function  $d(x)$  such that  $E d(X) < \infty$  and  $|f(x, s)| \leq d(x)$ ,

there exists a continuous limit in probability of  $1/n \sum_{i=1}^n f(X_i; s)$  uniformly in  $s \in [0, \tau]$ , i.e.

$$\sup_{s \in S} \left| \frac{1}{n} \sum_{i=1}^n f(X_i; s) - E f(X; s) \right| \xrightarrow{P} 0.$$

Here  $f(x; s) = Y_i(s)x$  and it is continuous for all  $x$  for every fixed  $s$ . Particularly, under the conditions in Assumption (\*) and the conditions in Theorem 5 we have that  $x_{i,1}$  are i.i.d. and asymptotically distributed as  $x_1$  with  $E x_1 < \infty$ . Hence  $f(x_{i,1}; s)$  is dominated by  $d(x_{i,1}) = x_{i,1}$  and we have the uniform limit of

$R_1(s)/n$  on  $[0, \tau]$ , in probability, equal to

$$r_1(s) = \mathbb{E} Y(s)x_1 = \mathbb{E} \mathbb{E} [I \{T \geq s\} x_1 | x_1] = \mathbb{E} \exp \{-x_1 A_1(s)\} x_1 \bar{G}(s),$$

where the expectation in the last expression is with respect to the covariate distribution. The uniform law of large numbers stated here gives a stronger variant of convergence, in particular the convergence with probability 1.

**Remark 2** In  $p = 1$  case, we showed that not only Aalen and McKeague-Huffer estimator but also the proposed NPML estimator is consistent and weakly convergent. The asymptotic distributions of these estimators differ though. The asymptotic distribution of Aalen's estimator is a zero-mean Gaussian process with variation process equal to  $\int_0^t r_1^{(3)}(s)/[r_1^{(2)}(s)]^2 \alpha_1(s) ds$  where  $r_1^{(2)}$  and  $r_1^{(3)}$  are uniform limits in probability of  $1/n \sum_i Y_i x_{i,1}^2$  and  $1/n \sum_i Y_i x_{i,1}^3$  respectively. McKeague and Huffer's estimator induces a zero-mean Gaussian process with variation process equal to  $\int_0^t \alpha_1(s)/r_1 ds$ , i.e. both McKeague-Huffer and NPML are asymptotically equivalent.

**Remark 3** A more complicated situation arises if there are ties present in the data. Even though we suppose that the failure times' distributions are absolutely continuous and the  $(N_1, \dots, N_n)$  is a multivariate counting process (i.e. no two  $N_i$ -s jump at the same time), the measurement error or a rounding up can introduce tied observations into the dataset. Let us have two individuals  $i$  and  $k$  failing at the same time point  $s$ . Then the contribution to the likelihood at  $s$  is

$$\{1 - dA_1(s)\}^{R_1(s) - x_{i,1} - x_{k,1}} (1 - \{1 - dA_1(s)\}^{x_{i,1}}) (1 - \{1 - dA_1(s)\}^{x_{k,1}}).$$

Clearly, we need to find a maximum value of following function of  $a$ ,

$$\{1 - a\}^{R_1 - x_{i,1} - x_{k,1}} (1 - \{1 - a\}^{x_{i,1}}) (1 - \{1 - a\}^{x_{k,1}}),$$

what is equivalent to solving the equation

$$\frac{(1 - a)^{x_{i,1}} - 1}{(1 - a)^{x_{k,1}} - 1} ((1 - a)^{x_{k,1}} R_1 - R_1 + x_{k,1}) + x_{i,1} = 0. \quad (2.10)$$

If  $x_{i,1}$  equals  $x_{k,1}$  or their values are close, we have the solution

$$a \approx 1 - \left\{ 1 - \frac{x_{i,1} + x_{k,1}}{R_1} \right\}^{1/x_{k,1}}.$$

In a general case the equation in (2.10) is not trivial to solve and the root needs to be found numerically. We decided not to pursue the topic due to the fact,

that the NPML approach produces inconsistent estimators in general  $p > 1$  case.

We showed that it is all clear and well behaved as long as we stay within the margins of a single covariate case. However, moving onto a general case with  $p$  covariates takes us into a venture of inconsistent results despite using the celebrated and usually reliable maximum likelihood method.

## 2.3 Case $p > 1$

We suppose, that the cumulative hazard function of an  $i$ -th individual equals  $H_i(t) = \sum_{j=1}^p x_{i,j} A_j(t)$ , where  $A_1, \dots, A_p$  are unknown cumulative regression functions. We would like to find estimators for these functions so that they maximize the likelihood in (2.2). Similarly as for the single covariate case we consider the contribution to the likelihood at every time. For  $s$  such that  $\sum_i dN_i(s) = 0$  the contribution is equal to

$$\prod_{i=1}^n \{1 - dH_i\}^{Y_i(s)} = \prod_{j=1}^p \{1 - dA_j\}^{R_j(s)}$$

where we denoted  $R_j(s) = \sum_i Y_i(s) x_{i,j}$ ,  $j = 1, \dots, p$ . Again, this achieves its maximum value when all  $dA_j$ -s are zero, i.e.  $dA_j(s) = 0$ ,  $j = 1, \dots, p$ . Further, let us assume that one and only one individual had failure at time  $s$ , for example  $dN_i(s) = 1$ . Then from (2.2) we can see that the contribution to the likelihood at this time point is equal to

$$\begin{aligned} & \prod_{k \neq i} \{1 - dH_k(s)\}^{Y_k(s)} dH_i(s) \\ &= \prod_{j=1}^p (1 - dA_j(s))^{R_j(s) - x_{i,j}} \left[ 1 - \prod_{j=1}^p (1 - dA_j(s))^{x_{i,j}} \right]. \end{aligned} \quad (2.11)$$

It is not obvious where this function with  $p$  arguments and with the domain of the  $[0, 1]^p$  cube obtains its maximum. Let us denote  $dA_j(s)$  by  $u_j$  for all  $j = 1, \dots, p$ , drop the time notation  $s$  in the expression and study the function's extremes.

**Lemma 1** *Consider the function*

$$f(u_1, \dots, u_p) = \prod_{j=1}^p \{1 - u_j\}^{R_j - x_{i,j}} \left( 1 - \prod_{j=1}^p \{1 - u_j\}^{x_{i,j}} \right) \quad \text{over } [0, 1]^p,$$



where  $R_j \geq x_{i,j} > 0$  for  $j = 1, \dots, p$ . If there is a unique  $j = j_0$  for which  $\pi_j = x_{i,j}/R_j$  is bigger than the others, then the maximum of  $f$  occurs for  $u = (0, \dots, \hat{u}_j, \dots, 0)^\top$ , with  $\hat{u}_j = 1 - (1 - \pi_j)^{1/x_{i,j}}$  and with the maximum value of  $f$  equal to  $\pi_j(1 - \pi_j)^{1/\pi_j - 1}$ .

In case that there are two largest  $\pi_j = \pi_k$ , the solution is not unique with the maximum attained on a subset of  $[0, 1]^p$ .

PROOF Let us say that  $\pi_1$  is the largest and is the only one. To find the global maximum we need to perform a study of function  $f$  and search for critical points in  $(0, 1)^p$ , i.e. we seek the points where all partial derivatives are equal 0 and therefore they might be local extremes or saddle points. However, as it is easily shown there does not exist a single spot inside the  $[0, 1]^p$  cube where all partial derivatives  $df/dx_{i,j} = 0, \forall j$ , and we get the same result when looking on the side-walls of the cube, hence we need to search for a potential global maximum on the vertices. Using partial derivations for each variable while other variables are equal to zero leads to a set of values  $\{\pi_j(1 - \pi_j)^{1/\pi_j - 1}, j = 1, \dots, p\}$  obtained in points  $\{u_j^* = (0, \dots, 1 - (1 - \pi_j)^{1/x_{i,j}}, \dots, 0), j = 1, \dots, p\}$  where the function  $f$  is locally maximal. The largest is the one with the biggest  $\pi_j = \pi_1$ .

When we want to deal with ties it is handy to understand the problem better. Each partial derivative induces a hyperplane in  $[0, 1]^p$  on which this partial derivative equals to 0. When there are no ties among  $\pi_j$ -s then there is no non-empty intersection of all the hyperplanes, therefore no critical points inside  $[0, 1]^p$ , also there cannot be found any subset of hyperplanes such that they intersect – no critical points on the side-walls of the cube. However, let us have a situation where the two largest equal, for instance  $\pi_1 = \pi_2$ . Then we find out that the hyperplanes induced by  $df/dx_{i,1}$  and  $df/dx_{i,2}$  cut through on set  $U^* = \{(1 - [(1 - \pi_1)(1 - u_2)^{-x_{i,2}}]^{1/x_{i,1}}, u_2, 0, \dots, 0), u_2 \in [0, 1]\} \subset [0, 1]^p$ , which lies on one of the side-walls, and  $f$  is the same and maximal on the whole  $U^*$ , equal to  $\pi_1(1 - \pi_1)^{1/\pi_1 - 1}$ . The solution is non-unique and span whole set  $U^*$ . The case of ties among  $\pi_j$ -s smaller than the largest  $\pi_1$  will again impose existence of a set on a side of  $[0, 1]^p$  where the value of  $f$  will be the same and locally maximal but the overall maximum remains located uniquely in  $u_1^* = (1 - (1 - \pi_1)^{1/x_{i,1}}, 0, \dots, 0)$ . If there are multiple ties, e.g.  $\pi_1 = \pi_2 = \pi_3$ , the situation is analogical and the set on which  $f$  gains its maximal value is characterized by points  $(1 - [(1 - \pi_1)(1 - u_2)^{-x_{i,2}}(1 - u_3)^{-x_{i,3}}]^{1/x_{i,1}}, u_2, u_3, 0, \dots, 0) \subset [0, 1]^p$ . The maximal value of  $f$  remains the same with solution being again non-unique.

Q.E.D.

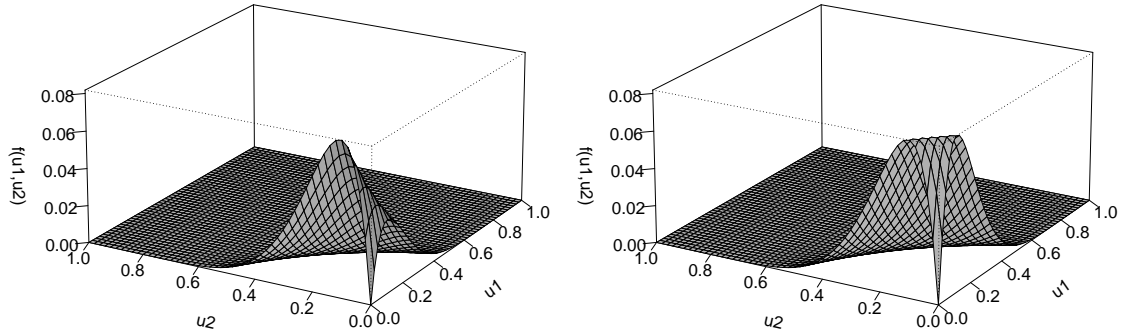


Figure 2.2: Two examples of the function in Lemma 1 with two arguments  $u_1$  and  $u_2$ . The parameters are set  $R_1 = R_2 = 10$ . The covariates in the left hand side and the right hand side of the figure equal  $x_{i,1} = 1, x_{i,2} = 2$  and  $x_{i,1} = x_{i,2} = 2$ , respectively.

**Example 1** There is an illustration of the function in Lemma 1 in the Figure 2.2. The values of the binary function  $f(u_1, u_2)$  are drawn for  $R_1 = R_2 = 10$ . The covariates equal  $x_{i,1} = 1$  and  $x_{i,2} = 2$  in the left hand side of the figure. Obviously,  $\pi_2 = x_{i,2}/R_2 = 0.2$  is greater than  $\pi_1 = x_{i,1}/R_1 = 0.1$  and indeed, the maximal value of the function is obtained at point  $(0, \hat{u}_2)$  with  $\hat{u}_2 = 1 - (1 - \pi_2)^{1/2} \doteq 0.11$ . The achieved maximum equals  $\pi_2(1 - \pi_2)^{1/\pi_2 - 1} \doteq 0.082$ . In the right hand side an example with ties is presented. The values of the covariates are the same and equal to  $x_{i,1} = x_{i,2} = 2$ . Here the maximal value of the function is the same as in the example in the left hand side. Apparently, the maximum is attained on the whole line segment between the points  $(\hat{u}_1, 0)$  and  $(0, \hat{u}_2)$ , where  $\hat{u}_1 = \hat{u}_2 = 1 - (1 - \pi_1)^{1/2} \doteq 0.11$ .

The result of Lemma 1 implies that the nonparametric maximum likelihood estimator for the parameters of monotone Aalen model is

$$\hat{A}_j(t) = \sum_{i=1}^n \int_0^t \left[ 1 - \left\{ 1 - \frac{x_{i,j}}{R_j(s)} \right\}^{1/x_{i,j}} \right] V_{i,j}(s) dN_i(s) \quad \text{for } j = 1, \dots, p, \quad (2.12)$$

where

$$V_{i,j}(s) = I \left\{ \frac{x_{i,j}}{R_j(s)} \geq \frac{x_{i,k}}{R_k(s)}, \forall k = 1, \dots, p \right\}.$$

The random processes  $V_{i,j}$  are left continuous and predictable. Moreover, we have that  $\sum_{j=1}^p V_{i,j}(t) = 1$  for  $t \geq 0$ , hence  $V_{i,j}$  represents weights over the estimators  $\hat{A}_1(t), \dots, \hat{A}_p(t)$  for every  $t$ .

For further use let us us again denote the increments in the estimated processes by

$$\hat{a}_{i,j}(s) = \left[ 1 - \left\{ 1 - \frac{x_{i,j}}{R_j(s)} \right\}^{1/x_{i,j}} \right] V_{i,j}(s)$$

so that  $\hat{A}_j(t) = \sum_i \int_0^t \hat{a}_{i,j}(s) dN_i(s)$ .

**Remark 4** The chance that there will be ties in  $x_{i,j}/R_j$  is rather small, especially when the covariate distribution is continuous, unless a strong multicollinearity is present. If e.g.  $x_{i,2} = cx_{i,1}$ ,  $\forall i$ , for some  $c \in \mathbb{R}$ , then obviously always  $x_{i,1}/R_1 = x_{i,2}/R_2$ .

However, when a situation with ties arises it is necessary to adjust the estimator in (2.12) appropriately. As it is seen in the proof if for some time point  $s$  ties occur, for instance  $x_{i,1}/R_1(s) = x_{i,2}/R_2(s)$  and they are the maximal among  $x_{i,1}/R_1(s), \dots, x_{i,p}/R_p(s)$ , then the solution is obtained on the set

$$\left( u_1 = 1 - [(1 - x_{i,1}/R_1(s))(1 - u_2)^{-x_{i,2}}]^{1/x_{i,1}}, u_2, 0, \dots, 0 \right) \in [0, 1]^p,$$

where  $u_2$  is anywhere between 0 and 1 under the condition that  $1 - [(1 - x_{i,1}/R_1(s))(1 - u_2)^{-x_{i,2}}]^{1/x_{i,1}} \in [0, 1]$ . The simplest would be to take e.g.  $u_2 = 0$  and  $u_1 = 1 - (1 - x_{i,1}/R_1(s))^{1/x_{i,1}}$  or  $u_1 = 0$  and  $u_2 = 1 - (1 - x_{i,2}/R_2(s))^{1/x_{i,2}}$ . That in fact means that we choose whether either process  $\hat{A}_1$  or process  $\hat{A}_2$  jumps at time  $s$ .

**Remark 5** Sending the covariates  $x_{i,j}$  to a constant  $b_j \geq 0$  for all  $i$  and  $j = 1, \dots, p$  gets us into an i.i.d. situation when all individuals are ruled by the same hazard function  $h_i(s) \equiv h(s) = \sum_j b_j \alpha_j(s)$ . In the single covariate case the hazard rate is equal to  $h(s) = b_1 \alpha_1(s) =: \alpha(s)$ . The cumulative version of hazard rate  $\int_0^t \alpha(s) ds$  is in the i.i.d. case well estimated by the Nelson-Aalen estimator

$$A^{N-A}(t) = \sum_{i=1}^n \int_0^t \frac{dN_i(s)}{\sum_{k=1}^n Y_k(s)} = \int_0^t \frac{d\bar{N}(s)}{R(s)}.$$

We denoted the overall risk set by  $R(s) = \sum_{k=1}^n Y_k(s)$ .

Examining our NPML estimator for  $p = 1$  in (2.3) and (2.4) under the i.i.d. case we get that

$$\hat{a}_{i,1}(s) = 1 - \left\{ 1 - \frac{b_1}{b_1 \sum_{k=1}^n Y_k(s)} \right\}^{1/b_1} = \frac{1}{b_1 R(s)} + O_P(R^{-2}(s))$$

and asymptotically the NPML estimator of  $A_1$  multiplied by  $b_1$  is equivalent to the Nelson-Aalen estimator:

$$b_1 \hat{A}_1(s) \approx b_1 \sum_{i=1}^n \int_0^t \frac{1}{b_1} \frac{dN_i(s)}{R(s)} = A^{N-A}(t).$$

The multiple covariate case is analogical, we only have to realize, that if all  $x_{i,j}$  equal to  $b_j$  then  $x_{i,j}/R_j(s) = b_j/\sum_k Y_k(s)b_j = 1/R(s)$  and we deal with ties for all  $j = 0, \dots, p$ . It is impossible to uniquely determine the NPML estimators for all the parameters of Aalen model as in fact we deal with the i.i.d. case and the model with  $p > 1$  is overparametrized. This can be solved without loss of generality by putting  $V_{i,1}(s) = 1$  and  $V_{i,j} = 0$ ,  $j \geq 2$ . Then the NPML estimators are  $\hat{A}_j(s) \equiv 0$ ,  $j \geq 2$  and

$$\hat{A}_1(s) = \sum_{i=1}^n \int_0^s \left[ 1 - \left( 1 - \frac{b_1}{b_1 R(s)} \right)^{1/b_1} \right] dN_i(s).$$

Similarly as for  $p = 1$  we get that  $b_1 \hat{A}_1(s)$  is asymptotically equal to the Nelson-Aalen estimator. The Kaplan-Meier estimator  $S^{K-M}$  for survival function is related to the Nelson-Aalen estimator via the known relationship between the hazard and survival function

$$S^{K-M}(t) = 1 - \prod_{s \in [0,t]} \{1 - dA^{N-A}(s)\}.$$

Here for  $p \geq 1$  we have

$$\hat{S}_i(t) \equiv \hat{S}(t) = 1 - \prod_{s \in [0,t]} \{1 - b_1 d\hat{A}_1(s)\}$$

and due to asymptotic equivalence of the NPMLE and Nelson-Aalen estimator we get that asymptotically the NPML estimator of  $S(t)$  is equivalent to the Kaplan-Meier estimator.

Let us have a closer look at the estimator  $\hat{A}_j$  as it is formulated in (2.12). Using the same representation for  $1 - (1 - x_{i,j}/R_j(s))^{1/x_{i,j}}$  as in the univariate case in (2.5) it is seen, that for great  $n$  and under usual conditions we get

$$\begin{aligned} \hat{A}_j(t) &= \sum_{i=1}^n \int_0^t \left[ \frac{1}{R_j(s)} + \frac{1}{2} \frac{x_{i,j} - 1}{R_j^2(s)} + O_p\left(\frac{1}{R_j^3(s)}\right) \right] V_{i,j}(s) dN_i(s) \\ &\approx \sum_{i=1}^n \int_0^t \frac{1}{R_j(s)} V_{i,j}(s) Y_i(s) z_i^\top \alpha(s) ds. \end{aligned}$$

This approximate representation of  $\hat{A}_j$  serves as an inspiration for the form of its limiting process.

Before moving onto the asymptotic behaviour of the NPML estimator let us remind that we denoted  $z_i = (x_{i,1}, \dots, x_{i,p})^\top$  and that we assume that for every  $i$  the covariate vector  $z_i$  possess the same distribution as the random vector  $z$ .

**Theorem 6 (Inconsistency of NPMLE)** *Let us suppose that the conditions in Assumption (\*) in Section 2.1 are fulfilled and let us denote*

$$Q_j(s) = \sum_{i=1}^n V_{i,j}(s) Y_i(s) z_i^\top \alpha(s), \quad s \in [0, \tau].$$

If  $\bar{Y}(\tau) \xrightarrow{P} \infty$  with  $n \rightarrow \infty$ ,  $\alpha_j, j = 1, \dots, p$ , are continuous and the covariates are i.i.d., nonnegative and bounded, then there exist functions  $r_j(s)$  and  $q_j(s)$ ,  $j = 1, \dots, p$ , on  $[0, \tau]$ , such that

$$\sup_{0 \leq s \leq \tau} \left| \frac{R_j(s)}{n} - r_j(s) \right| \xrightarrow{P} 0,$$

$$\frac{Q_j(s)}{n} \xrightarrow{P} q_j(s), \quad \forall s \in [0, \tau]$$

and

$$\hat{A}_j(t) \xrightarrow{P} B_j(t) = \int_0^t \frac{q_j(s)}{r_j(s)} ds, \quad \text{for } j = 1, \dots, p, \quad t \in [0, \tau]. \quad (2.13)$$

If we denote the limiting process of  $V_{i,j}(s)$  by  $V_j(s) = I\{x_j/r_j(s) \text{ is the biggest}\}$ , then the functions  $q_j(s), s \in [0, \tau], j = 1, \dots, p$ , are equal to

$$q_j(s) = \mathbb{E} [Y(s) z^\top \alpha(s) V_j(s)] = \mathbb{E} \left[ e^{-z^\top A(s)} z^\top \alpha(s) V_j(s) \right] \bar{G}(s). \quad (2.14)$$

**PROOF** The existence and continuity of the functions  $r_j$  is ensured by applying the uniform law of large numbers (see Remark 1). A bit more care is needed when dealing with existence of the limit of  $Q_j(s)/n = 1/n \sum_{i=1}^n Y_i(s) z_i^\top \alpha(s) V_{ij}(s)$  because the function  $V_{ij}$  is not continuous in  $z_i$ . Let us denote  $Q_j^{(n)}(s, (z_1, \dots, z_n)) := Q_j(s)/n$ . Then for all  $m \in 1, \dots, n$ , all  $z_1, \dots, z_n \in \mathbb{R}_+^n$  and  $z'_m \in \mathbb{R}_+^n$  we get

$$\begin{aligned} & \left| Q_j^{(n)}(s, (z_1, \dots, z_n)) - Q_j^{(n)}(s, (z_1, \dots, z'_m, \dots, z_n)) \right| \\ &= \left| \frac{1}{n} Y_m(s) z_m^\top \alpha \, I \left\{ \frac{x_{m,j}}{R_j(s)} > \frac{x_{m,k}}{R_k(s)} \forall k \right\} - \frac{1}{n} Y_m(s) z'^m_\top \alpha \right. \\ & \quad \left. \times I \left\{ \frac{x'_{m,j}}{R_j(s) + (x'_{m,j} - x_{m,j}) Y_m(s)} > \frac{x'_{m,k}}{R_k(s) + (x'_{m,k} - x_{m,k}) Y_m(s)}; \forall k \right\} \right|. \end{aligned} \quad (2.15)$$

If  $Y_m(s) = 0$  then the expression equals zero. Otherwise we have that, for any  $i$ , supremum of the expression in (2.15) over all  $z_1, \dots, z_n \in \mathbb{R}_+^n$  and  $z'_m \in \mathbb{R}_+^n$  is smaller than  $\frac{2}{n} \sup z^\top \alpha(s)$ . For covariates bounded by some  $K \in \mathbb{R}^p$ ,  $z \leq K < \infty$ , it is smaller than  $\frac{2}{n} K^\top \alpha(s)$ . Therefore the so called bounded difference assumption is fulfilled and using the McDiarmid inequality in Theorem 3 we have that for any  $\epsilon > 0$ ,

$$\begin{aligned} P\left(\left|Q_j^{(n)}(s, (z_1, \dots, z_n)) - EQ_j^{(n)}(s, (z_1, \dots, z_n))\right| \geq \epsilon\right) &\leq \\ &\leq 2 \exp\left\{\frac{-2\epsilon^2}{(2K^\top \alpha(s))^2/n}\right\} \xrightarrow{n \rightarrow \infty} 0 \quad \forall \epsilon > 0. \end{aligned}$$

This gives us the convergence of  $Q_j^{(n)}(s, (z_1, \dots, z_n))$  to  $EQ_j^{(n)}(s, (z_1, \dots, z_n))$  in probability, at every  $s \in [0, \tau]$ . Finally, due to the covariates being i.i.d. it is seen that  $EQ_j^{(n)}(s, (z_1, \dots, z_n)) = E \frac{1}{n} \sum_i Y_i(s) z_i^\top \alpha(s) V_{ij}(s) = E Y(s) z^\top \alpha(s) V_j(s)$ .

To proceed further with the proof we will rely on the well-known martingale theory on counting processes, see Section 1.2. It is clear that the counting process  $N_i$  can be decomposed into sum of a martingale and an integrated compensator for  $N_i$ , i.e.  $N_i(t) = M_i(t) + \int_0^t Y_i(s) z_i^\top \alpha(s) ds$ ,  $i = 1, \dots, n$ . It is seen that the processes  $\hat{a}_{i,j}(s)$  are bounded and predictable, and using the aforementioned decomposition we get

$$\hat{A}_j(t) = \sum_{i=1}^n \int_0^t \hat{a}_{i,j}(s) dM_i(s) + \sum_{i=1}^n \int_0^t \hat{a}_{i,j}(s) Y_i(s) z_i^\top \alpha(s) ds. \quad (2.16)$$

The first term is a martingale, denoted e.g.  $W_j(t)$ , and applying the theory in Section 1.2 we get that  $W_j(t)$  has the predictable variation process

$$\langle W_j, W_j \rangle(t) = \sum_{i=1}^n \int_0^t \hat{a}_{i,j}^2 Y_i(s) z_i^\top \alpha(s) ds.$$

Now our goal is to show that  $W_j(t)$  goes to zero in probability while the second term of the decomposition (2.16), denoted e.g.  $A_j^*(t)$ , tends in probability to the  $B_j(t)$  from (2.13). Once we show that  $A_j^*(t) \xrightarrow{P} B_j(t)$ , then the proof is finished by application of the Slutsky theorem.

Using the approximation in (2.5), which is equally valid in the multiple covariates case, it can be shown that

$$\langle W_j, W_j \rangle(t) = \sum_{i=1}^n \int_0^t \frac{1}{R_j^2(s)} V_{i,j}(s) Y_i(s) z_i^\top \alpha(s) ds + o_p(R_j(t)^{-2})$$

therefore  $\langle W_j, W_j \rangle(t)$  is of order  $O_p(1/R_j(t))$  and clearly going to zero in probability. Applying the same corollary of the Lengart inequality as in the proof for Theorem 4 we have the uniform convergence of  $W_j$  to zero on  $[0, \tau]$ . Now let us show that  $A_j^*(t)$  goes to the aforementioned  $B_j(t)$  in probability. Again, we have that

$$A_j^*(t) = \sum_{i=1}^n \int_0^t \frac{1}{R_j(s)} V_{i,j}(s) Y_i(s) z_i^\top \alpha(s) ds + o_p(n^{-1/2}) = \int_0^t \frac{Q_j(s)}{R_j(s)} + o_p(n^{-1/2}).$$

We already showed in the beginning of the proof that  $Q_j(s)/n$  has a limit in probability equal to  $q_j(s)$ , while for  $\bar{Y}(\tau) \xrightarrow{P} \infty$  implies that both  $R_j(s)$  and the limiting function  $r_j(s)$  are greater than zero for any  $s \in [0, \tau]$ . Hence we have

$$\left| \frac{Q_j(s)}{R_j(s)} - \frac{q_j(s)}{r_j(s)} \right| \xrightarrow{P} 0$$

and this concludes the proof.

Q.E.D.

The limit of  $R_j(s)/n$  on  $[0, \tau]$  is of a similar form as in the single covariate case (see Remark 1):

$$r_j(s) = \mathbb{E} Y(s) x_j = \mathbb{E} \mathbb{E} [I \{T \geq s\} x_j | z] = \mathbb{E} \exp \{-z^\top A(s)\} x_j \bar{G}(s) \quad (2.17)$$

and the limiting functions  $B_j$  are

$$\begin{aligned} B_j(s) &= \int_0^t \frac{q_j(s)}{r_j(s)} ds = \int_0^t \frac{\mathbb{E} [Y(s) z^\top \alpha(s) V_j(s)]}{\mathbb{E} Y(s) x_j} ds \\ &= \int_0^t \frac{\mathbb{E} \left( e^{-z^\top A(s)} z^\top \alpha(s) I \{x_j/r_j(s) \text{ is the biggest}\} \right)}{\mathbb{E} e^{-z^\top A(s)} x_j} ds \end{aligned}$$

which is generally equal to  $A_j(t) = \int_0^t \alpha_j(s) ds$  only when  $p = 1$ . As  $B_j$  is expressed as a Lebesgue integral,  $B_j$  is absolutely continuous on  $[0, \tau]$  and the derivative equals  $b_j(s) = q_j(s)/r_j(s)$ . This would not be true if we did not suppose that  $A_j$  was absolutely continuous.

**Remark 6** Even though the NPML estimators for  $A_j, j = 1, \dots, p$ , are inconsistent, it is possible to determine their asymptotic features. Using similar techniques as in Theorem 5 it can be shown that asymptotically the process  $\sqrt{n}(\hat{A}_j(t) - B_j(t))$  is a zero-mean Gaussian process with the predictive variation

process equal to

$$C_j(t) = \int_0^t \frac{q_j(s)}{r_j^2(s)} ds.$$

The predictive variation process can be consistently estimated by

$$\hat{C}_j(t) = \int_0^t \sum_{i=1}^n \frac{n V_{i,j}(s) dN_i(s)}{R_j^2(s)},$$

or for data with less observations it is more accurate to use exact expression for jumps  $\hat{a}_{i,j}$ , i.e. we estimate the predictive variation process by

$$\hat{C}_j(t) = \int_0^t \sum_{i=1}^n n \hat{a}_{i,j}^2(s) dN_i(s) ds.$$

Using the estimator  $\hat{C}_j(t)$  we can construct naive pointwise  $(1-\alpha)100\%$  confidence bands around the estimator  $\hat{A}_j(t)$ , i.e.

$$\hat{A}_j(t) \pm u_\alpha \sqrt{\frac{\hat{C}_j(t)}{n}}, \quad t \in [0, \tau],$$

where  $u_\alpha$  is a  $\alpha$ th quantile of standard normal distribution.

## 2.4 Average consistency

In the previous section we proved that the proposed NPML estimator  $\hat{A}_j$  of the integrated regression functions  $A_j$  is for general  $p$  inconsistent and it converges in probability to  $B_j \neq A_j$ . There is, however, an interesting twist to this result when we look at the average intensity, i.e. the intensity of a randomly picked subject. This average intensity under the model defined in Assumption (\*) in Section 2.1 equals

$$\sum_{j=1}^p r_j(s) \alpha_j(s), \quad s \in [0, \tau]. \quad (2.18)$$

In next we show that this feature is intact even when we substitute  $b_j$ -s instead of  $\alpha_j$ -s into (2.18).

**Corollary 1 (Melda corollary)** *Let us suppose, that the assumptions of Theorem 6 are fulfilled. Then*

$$\sum_{j=1}^p r_j(s) b_j(s) = \sum_{j=1}^p r_j(s) \alpha_j(s). \quad (2.19)$$



PROOF Using the fact, that the limiting processes  $B_j$  in Theorem 6 equal  $b_j(t) = q_j(s)/r_j(s)$  we have that the average sum is

$$\sum_{j=1}^p r_j(s) b_j(s) = \sum_{j=1}^p q_j(s).$$

Further using the expressions for  $q_j$  in (2.14) and similarly for  $r_j$  in (2.17) we get that

$$\begin{aligned} \sum_{j=1}^p q_j(s) &= \sum_{j=1}^p \mathbb{E} \left[ V_j(s) e^{-z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s) \\ &= \mathbb{E} \left[ e^{-z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s) = \sum_{j=1}^p \mathbb{E} \left[ e^{-z^\top A(s)} z_j \right] \bar{G}(s) \alpha_j(s) \\ &= \sum_{j=1}^p r_j(s) \alpha_j(s). \end{aligned}$$

Q.E.D.

This interesting result says that even though the processes  $\hat{A}_j$  are inconsistent, with  $\hat{A}_j \xrightarrow{prob} B_j \neq A_j$ , we still have an 'average-consistency' effect.

An estimator of  $\sum_{j=1}^p r_j(s) dB_j(s)$  can be obtained by inserting  $d\hat{A}_j$  and  $R_j/n$  instead of  $dB_j$  and  $r_j$  into formula and using the approximation from before

$$\sum_{j=1}^p \frac{R_j(s)}{n} d\hat{A}_j(s) \approx \sum_{j=1}^p \frac{R_j(s)}{n} \sum_{i=1}^n \frac{V_{i,j}(s)}{R_j(s)} dN_i(s) = \sum_{i=1}^n \frac{1}{n} dN_i(s), \quad n \gg 0.$$

Let us look at possible interpretations of the obtained feature. For continuous distribution we have

$$r_j(s) = \mathbb{E} [Y(s) z_j] = \mathbb{E} \left[ e^{-z^\top A(s)} z_j \right] \bar{G}(s)$$

where expectation is w.r. to covariates distribution and  $\bar{G}(s)$  is the censoring survival function. Under independent censoring and i.i.d. covariates we get

$$\begin{aligned} \sum_{j=1}^p r_j(s) \alpha_j(s) &= \sum_{j=1}^p \mathbb{E} [Y(s) z_j \alpha_j(s)] = \mathbb{E} \left[ e^{-z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s) \\ &= \mathbb{E} f(s, z) \bar{G}(s) \end{aligned}$$

where by  $f(\cdot, z)$  we denoted the probability density of the distribution of failure time under the covariate vector equal to  $z$ . In the case of weak censoring the

last term gets close to  $E f(s, z)$ . From this we can see the motivation of the form of the estimator, which when is integrated and under no censoring is for large  $n$  close to Glivenko-Cantelli statistics  $1/n \sum_{i=1}^n N_i(s)$ , i.e. the consistent estimator for distribution function.

The average consistency can be understood in the way that if we consider a random individual, on average we are able to guess correctly the probability of their survival.

## 2.5 An example for case with 3 exponentially distributed covariates

This section is devoted to assessing the performance of the proposed estimators on a simulated example with the main interest in how big the bias from the real values is as well as in the average consistency effect.

Let us have a triple of covariates  $(x_1, x_2, x_3) \sim Exp(\lambda_1, \lambda_2, \lambda_3)$ . In next we derive the expressions for the limiting functions  $B_j$  from Theorem 6. First, we have that

$$r_1(s) = \overline{G}(s) \frac{\lambda_1 \lambda_2 \lambda_3}{(A_1(s) + \lambda_1)^2 (A_2(s) + \lambda_2) (A_3(s) + \lambda_3)} \quad (2.20)$$

and the same is for  $r_2$  and  $r_3$  again after swapping 1 and 2, or 1 and 3, respectively. To calculate the numerator  $q_j$  we use the representation from (2.14)

$$q_j(s) = E e^{-z^\top A(s)} z^\top a(s) V_j(s) \overline{G}(s) = \int_{D_j(s)} e^{-z^\top A(s)} z^\top a(s) f(z) dz \overline{G}(s)$$

where  $\overline{G}(s) = 1 - G(s)$  is censoring survival function,  $D_j(s) = \{z : \frac{x_j}{r_j(s)} > \frac{x_k}{r_k(s)}, \forall k \neq j\}$ . By a simple integration we get that the limit of  $\hat{A}_1$  equals  $B_1(t) = \int_0^t q_1(s)/r_1(s) ds = \int_0^t b_1(s) ds$  where

$$b_1(s) = \text{term}_1(s) \alpha_1(s) + \text{term}_2(s) \alpha_2(s) + \text{term}_3(s) \alpha_3(s).$$

The exact expressions for  $\text{term}_1$ ,  $\text{term}_2$  and  $\text{term}_3$  follow. The first term has the simplest form and it equals

$$\begin{aligned} \text{term}_1(s) &= \frac{\lambda_1 \lambda_2 \lambda_3}{(A_2(s) + \lambda_2) (A_3(s) + \lambda_3)} \\ &\times \left( \frac{1}{C(s)} + \frac{1}{(A_1(s) + \lambda_1)^2} - \frac{1}{\left\{ \frac{r_2(s)}{r_1(s)} (A_2(s) + \lambda_2) + A_1(s) + \lambda_1 \right\}^2} \right) \end{aligned}$$

$$- \frac{1}{\left\{ \frac{r_3(s)}{r_1(s)}(A_3(s) + \lambda_3) + A_1(s) + \lambda_1 \right\}^2}$$

where we denoted

$$C(s) = \left\{ \frac{r_2(s)}{r_1(s)}(A_2(s) + \lambda_2) + \frac{r_3(s)}{r_1(s)}(A_3(s) + \lambda_3) + A_0(s) + \lambda_1 \right\}^2.$$

The second term is

$$\begin{aligned} \text{term}_2(s) &= \lambda_1 \lambda_2 \lambda_3 \\ &\times \left[ \frac{r_1(s)}{r_2(s)} \frac{1}{C(s)(A_2(s) + \lambda_2)(A_3(s) + \lambda_3)} \right. \\ &- \frac{r_2(s)}{r_1(s)} \frac{1}{(A_2(s) + \lambda_2)(A_3(s) + \lambda_3) \left\{ \frac{r_2(s)}{r_1(s)}(A_2(s) + \lambda_2) + A_1(s) + \lambda_1 \right\}^2} \\ &+ \frac{1}{C(s)(A_2(s) + \lambda_2)^2(A_3(s) + \lambda_3)} \\ &- \frac{1}{(A_2(s) + \lambda_2)^2(A_3(s) + \lambda_3) \left\{ \frac{r_2(s)}{r_1(s)}(A_2(s) + \lambda_2) + A_1(s) + \lambda_1 \right\}^2} \\ &- \frac{1}{(A_2(s) + \lambda_2)^2(A_3(s) + \lambda_3) \left\{ \frac{r_3(s)}{r_1(s)}(A_3(s) + \lambda_3) + A_1(s) + \lambda_1 \right\}^2} \\ &\left. + \frac{1}{(A_1(s) + \lambda_1)(A_2(s) + \lambda_2)^2(A_3(s) + \lambda_3)} \right] \end{aligned}$$

and  $\text{term}_3$  is the same as  $\text{term}_2$  only with swapping the suffices  $_1$  and  $_2$  in the whole expression including  $C(s)$ . Expressions for  $B_2$  and  $B_3$  can be found in similar fashion.

The theoretical average intensity from Section 2.4 equals to

$$\sum_{j=1}^p r_j(s) b_j(s)$$

and if we insert the expressions for  $r_j$  and  $b_j$  from above, we will indeed get that  $\sum_{j=1}^p r_j(s) b_j(s) = \sum_{j=1}^p r_j(s) \alpha_j(s)$ . The sample counterpart of the cumulative average intensity can be obtained as an approximation of the integral

$$\int_0^t \sum_{j=1}^p \frac{R_j(s)}{n} \hat{a}_{i,j}(s) dN_i(s).$$

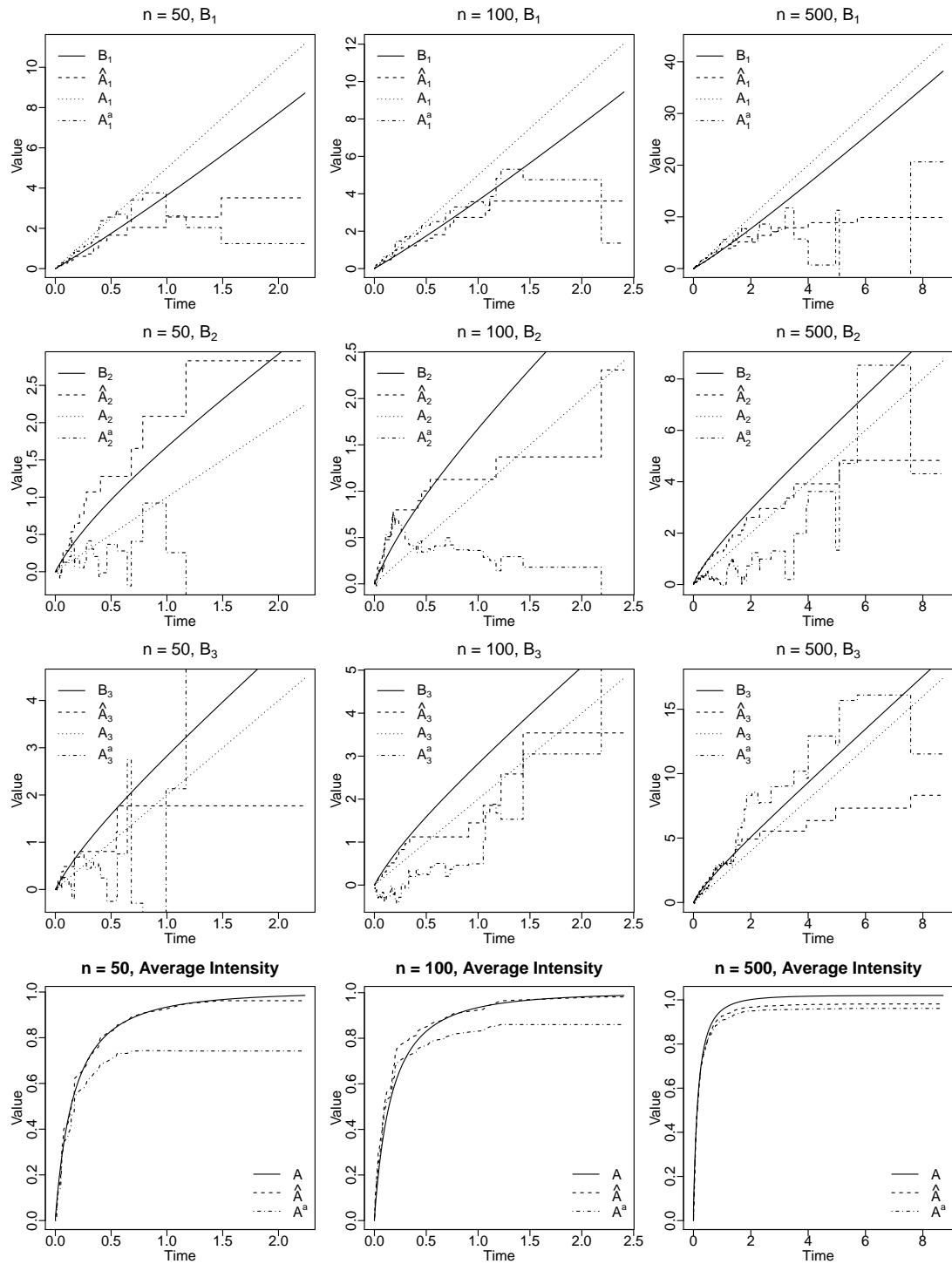


Figure 2.3: Cumulative regression functions in a simulated Aalen model with hazard rate  $h_i(t) = 5x_{i,1} + x_{i,2} + 2x_{i,3}$  and  $z_i$  i.i.d. and exponentially distributed. Left column is from the simulation with 50 observations, in the middle 100 simulated observations and in right column 500 simulated observations. In *dotted lines* the real cumulative parameter processes are plotted, the *dashed lines* are the NPML estimators  $\hat{A}_j$  and the *solid lines* are the asymptotic processes  $B_j$  of the NPML estimators. The Aalen estimators  $A_j^a$  are included in *dash-dotted lines*. Graphs in the bottom row shows the average intensities.

We ran simulations from a model with the hazard rate equal to  $h_i(t) = 5x_{i,1} + x_{i,2} + 2x_{i,3}$ , where covariates were independent and exponentially distributed,  $x_{i,1} \sim \text{Exp}(1)$ ,  $x_{i,2} \sim \text{Exp}(2)$  and  $x_{i,3} \sim \text{Exp}(3)$ . We considered three different sample sizes:  $n = 50$ ,  $n = 100$  and  $n = 500$ . We calculated the NPML estimators  $\hat{A}_j$  based on (2.12) and compared them to the true asymptotic functions  $B_j$  obtained from the calculations above. Out of curiosity we also estimated the cumulative regression functions using the classic Aalen OLS estimator.

In Figure 2.3 the results of the simulations are presented. The estimators of the cumulative regression functions are plotted in the upper three rows with sample size growing from left to right. The bottom row shows the results on the average intensities. The true cumulative regression functions  $A_j$  are plotted in dotted lines, the Aalen estimators  $A_j^a$  are in dash-dotted lines and the proposed NPML estimators  $\hat{A}_j$  are in dashed lines. The limiting functions  $B_j$  of the NPML estimators are in solid lines. In the bottom graphs, only the average intensities calculated from the true  $A_j$ -s are plotted as the average intensities obtained the limiting  $B_j$ -s are the same.

As can be seen from the figures, the deviance of the NPML estimators  $\hat{A}_j$  from the true regression functions  $A_j$  is apparent with growing sample size. It seems, however, that the NPML estimators are more stable than the classic Aalen estimators  $A_j^a$  and are able to extract the average intensity very well.

## Chapter 3

# Bayesian analysis with Beta process prior

The Bayesian approach to nonparametric problems has been a rather overlooked topic and become more popular only in last decades. The delay in development of these methods can be mainly attributed to the lack of the computing power needed in solving many Bayesian problems as well as the complexity of dealing with probabilities on infinite dimensional spaces.

In general, constructing a nonparametric Bayesian estimator for a cumulative distribution function  $F$  means that we assume that  $F$  is a stochastic process ruled by a certain probability distribution. Lévy processes have proven to be a natural choice for a prior process when one conducts Bayesian modelling of a process with trajectories possibly containing jumps. A first extensive class of priors for distribution functions called Dirichlet processes was introduced by Ferguson in [13]. These processes were further generalized to the neutral to the right processes by Doksum in [12]. Kalbfleisch, [24], and others utilized the well-known relationship  $F(t) = 1 - \exp\{-H(t)\}$  between cumulative distribution function  $F$  and cumulative hazard rate  $H$  and proposed a process to model the prior of a cumulative hazard rate  $H$  instead of  $F$ . They suggested a process with Gamma distributed independent increments in disjoint intervals, hence this process was named a Gamma process. Similarly, Hjort in [19] introduced Beta processes which again had independent increments and their increments were "almost" Beta distributed. The Beta processes are the processes of our particular interest in this chapter. Among other popular priors within this field we can list Pólya trees priors, correlated prior processes and various finite-mixture models of the previous. For an extensive overview see [41].

In our situation we have unknown regression functions  $A_j, j = 1, \dots, p$  that in fact are assumed to be continuous, however, we would like to continue with

the work we have done in Chapter 2. Thus we will view these functions as random processes which are discontinuous at time points of countable amount on a bounded time interval, in particular they are assumed to be jump processes with positive jumps which occur always in failure times and with infinitely many tiny positive jumps at random time points (this is induced by the Lévy processes imposed as prior distribution on the processes as we will see in next).

Let us suppose  $\mathcal{D}$  is a set of all distribution functions on  $[0, \infty)$  such that if  $F \in \mathcal{D}$  then  $F(0) = 0$ . Applying a relationship between  $F$  and  $H$  such as in (1.2) we get a corresponding set of cumulative hazard rates, denoted e.g.  $\mathcal{A}$ , induced by  $\mathcal{D}$ . For any function  $H$  on  $[0, \infty)$  to be a valid member of  $\mathcal{A}$ ,  $H$  must be 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  $\lim_{t \rightarrow \infty} H(t) = \infty$ . These condition must be met otherwise  $F$  corresponding to  $H$  would not be a distribution function. A process  $H$  that possesses the conditions above is also called a *nondecreasing independent increment process (NII process)* or *subordinator* and the corresponding distribution function  $F$  is a *neutral to the right process*.

Following the theory of the NII processes as it was summarized e.g. by Kim in [27] we state that for any NII process  $H$  there exists a unique random measure  $\mu$  on  $[0, \infty) \times [0, 1]$  such that it is a Poisson random measure and it uniquely determines the process  $H$  by

$$H(t) = \int_0^t \int_0^1 x \mu(ds, dx).$$

Furthermore, there exists a unique  $\sigma$ -finite measure  $\nu$  on  $[0, \infty) \times [0, 1]$  which is a compensator of the process  $\mu$ ,

$$\nu([0, t] \times B) = E \mu([0, t] \times B) = E \left( \sum_{s \in [0, t]} I \{ \Delta H(s) \in B \setminus \{0\} \} \right),$$

where  $t \geq 0$ ,  $B$  is a Borel subset of  $[0, 1]$ .

Let us assume a NII process  $H$  with fixed discontinuities at time points  $t_1, \dots, t_n$ . Then it admits a Lévy representation

$$E \exp\{-\theta H(t)\} = \left[ \prod_{i: t_i \leq t} E \exp\{-\theta \Delta H(t_i)\} \right] \exp \left\{ - \int_0^1 (1 - e^{\theta u}) dL_t(u) \right\}$$

for  $\theta \geq 0$ ,  $t \geq 0$  where  $L_t, t \geq 0$  is a continuous Lévy measure. It can be seen that

$$\nu([0, t] \times B) = \int_{u \in B} dL_t(u) + \sum_{t_i \leq t} \int_{u \in B} dF_{t_i}^H(u) \quad (3.1)$$

where  $F_{t_i}^H$  is a distribution function of  $\Delta H(t_i)$ , the size of the jump of the process  $H$  at the fixed discontinuity time point  $t_i$ . Consequently,  $\nu$  can be seen as an extension of the continuous Lévy measure  $dL_t$  to a measure which incorporates the fixed discontinuities as well. As we will see later, the posterior distribution of a hazard process will always have fixed discontinuities located precisely at the failure times. For the sake of convenience, from now on we will call also the extended measure  $\nu$  in (3.1) the Lévy measure of the process  $H$ .

Similarly as in Kim and Lee's work ([27], [31]) we will assume following set of processes to be our candidates for prior distribution of the cumulative regression functions: NII processes with continuous Lévy measure of the following form

$$\nu(ds, du) = f_s(u) ds du, \quad s \geq 0, \quad u \in [0, 1], \quad (3.2)$$

where  $f_s$  is such that  $\lim_{t \rightarrow \infty} \int_0^t \int_0^1 u f_s(u) ds du = \infty$ . Let us follow the notation from the previous chapter, viz.  $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$  and  $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ . In an i.i.d. case, when no covariates are present and all uncensored observations are ruled by an unknown distribution function  $F$  with corresponding cumulative hazard function  $H(t) = \int_0^t dF(s)/(1 - F(s_-))$ , it was shown in [19] and [31] that if  $H$  is a priori a stochastically continuous subordinator with Lévy measure as in (3.2) then a-posteriori  $H$  is again a subordinator with Lévy measure equal to

$$\nu^{POST}(ds, du) = (1 - u)^{\bar{Y}(s)} f_s(u) ds du + dF_s^H(u) \frac{1}{\Delta \bar{N}(s)} d\bar{N}(s), \quad (3.3)$$

where  $F_s^H(u)$  is a distribution function on  $[0, 1]$  such that

$$dF_s^H(u) \propto u^{\Delta \bar{N}(s)} (1 - u)^{\bar{Y}(s) - \Delta \bar{N}(s)} f_s(u) du.$$

Apparently,  $F_s^H$  characterizes the distribution of a jump size of  $H$  at a fixed discontinuity point induced by the observed data set, i.e. at a time point  $s$  such that for some  $i = 1, \dots, n$ , the process  $\Delta N_i(s) = 1$  jumped,  $\delta_i = 1$  and  $s = T_i$  is a failure time. Notice that a-posteriori the cumulative hazard function is again a subordinator and the distribution function is a neutral to the right process. Consequently, the NII processes display conjugacy with right-censored i.i.d. survival data.

Let us return to the Aalen model scenario. We assume that the  $i$ -th individual failure time is ruled by distribution function  $F_i$  and the corresponding cumulative hazard function is  $H_i$ . Then using (1.2) and (2.1) the following relations are true if the  $i$ -th individual hazard function abides monotone Aalen model (under time-



independent covariate vector  $z_i = (x_{i,1}, \dots, x_{i,p})^\top$ :

$$1 - F_i(t) = \prod_{s \in [0,t]} \{1 - dH_i(s)\} = \prod_{s \in [0,t]} \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j}} = \prod_{j=1}^p \{1 - dG_j(t)\}^{x_{i,j}}.$$

By  $G_j$ -s we denote the distribution functions corresponding to cumulative regression functions  $A_j$ -s. Using NII processes in Bayesian analysis of the monotone Aalen model means that we assume that  $G_1, \dots, G_p$  are a priori distributed as a set of independent neutral to the right processes, i.e. a priori  $G_1, \dots, G_p \in \mathcal{D}$ . Then there exists a corresponding set of subordinators  $A_1, \dots, A_p \in \mathcal{A}$  induced by  $G_1, \dots, G_p$  and every  $A_j$  is a priori a process with a Lévy measure as in (3.2). In particular we will focus on one special case when the Lévy measure of  $A_j$  is

$$\nu_j(ds, du) = f_s^j(u) ds du = c_j(s)u^{-1}(1-u)^{c_j(s)-1} dA_j^0(s) du, \quad j = 1, \dots, p,$$

where  $c_j$  is a piecewise constant nonnegative function,  $A_j^0$  is a continuous function and  $A_j^0 \in \mathcal{A}$ . This type of NII process is called a Beta process with parameters  $c_j$  and  $A_j^0$ .

Next section gives more details on the Beta processes and explains the usage of these processes as prior processes for the cumulative hazard functions. In Section 3.2 we derive the posterior distribution of the cumulative regression functions under the Beta process priors. The Bayesian estimators defined as the expectation of the posterior distribution (as modus would be hard to obtain) are introduced in Section 3.3 and their small sample features are investigated. The section is concluded by an algorithm for generating the Bayesian estimators using MCMC. A special case of  $p = 1$  is considered in Section 3.4 and the obtained asymptotic result in form of a Bernstein-von Mises theorem is in agreement with the results in the NPML estimation. We move on to the general case of  $p > 1$  in Section 3.5 and we show that the proposed Bayesian estimators converge to set of functions  $D_j, j = 1, \dots, p$ , that are in general not equal to the sought  $A_j, j = 1, \dots, p$ . Similarly as in the NPML case, the proposed Bayesian estimators exhibit the average consistency feature. This is dealt with in Section 3.6. The results derived for Beta process priors are extended to a case when  $A_j, j = 1, \dots, p$ , are a priori distributed as general NII processes in Section 3.7. In Section 3.8 we revisit the simulated example with three exponentially distributed covariates from Section 2.5. The chapter is concluded with a discussion of the results obtained in both Chapter 2 and Chapter 3.

### 3.1 Beta process prior

Let us suppose that  $c$  is a piecewise constant nonnegative function on  $[0, \infty)$  and  $A^0 \in \mathcal{A}$  has jumps at points  $t_1, \dots, t_n$ , i.e.  $A^0$  is a cumulative hazard with a finite number of jumps. Let us denote the continuous part of  $A^0$  by  $A_c^0(t) = A^0(t) - \sum_{t_i \leq t} \Delta A^0(t_i)$ , where  $\Delta A^0(t_i) = A^0(t_i) - A^0(t_{i-})$  are the jumps.

Following the definition of Beta process in work of Hjort [19] we say that  $A$  is a Beta process on  $[0, \infty)$  with parameters  $c(t)$  and  $A^0(t)$  with fixed discontinuities at time points  $t_1, \dots, t_n$  if  $A$  has paths in  $\mathcal{A}$ , has independent increments, and if  $A$  admits a Lévy representation

$$\mathbb{E} \exp\{-\theta A(t)\} = \left[ \prod_{i:t_i \leq t} \mathbb{E} \exp\{-\theta \Delta A(t_i)\} \right] \exp \left\{ - \int_0^1 (1 - e^{-\theta u}) dL_t(u) \right\} \quad (3.4)$$

for  $\theta \geq 0$ ,  $t \geq 0$ , where  $L_t, t \geq 0$  is a continuous Lévy measure of form

$$dL_t(u) = \left\{ \int_0^t c(s) u^{-1} (1-u)^{c(s)-1} dA_c^0(s) \right\} du, \quad 0 < u < 1. \quad (3.5)$$

The jump sizes of process  $A$ , denoted as  $\Delta A(t_i) = A(t_i) - A(t_{i-})$  at times  $t_i, i = 1, \dots, n$  are beta distributed with parameters

$$\Delta A(t_i) \sim \text{beta}(c(t_i) \Delta A^0(t_i), c(t_i) [1 - \Delta A^0(t_i)]).$$

The first part of the Lévy representation in (3.4) is attributed to the fixed discontinuities while the second belongs to the continuous part. If there are no fixed discontinuities present, the Lévy measure simplifies down to

$$\nu(ds, du) = \frac{d}{ds} dL_s(u) = c_j(s) u^{-1} (1-u)^{c_j(s)-1} dA^0(s) du. \quad (3.6)$$

A sample path of a Beta process with Lévy measure equal to  $\nu$  as in (3.6) will have random jumps (i.e. discontinuities) in various time points. When fixed discontinuities  $t_1, \dots, t_n$  are present then the Lévy measure equals

$$\begin{aligned} \nu(ds, du) &= c_j(s) u^{-1} (1-u)^{c_j(s)-1} dA_c^0(s) du \\ &+ \sum_{i=1}^n \frac{u^{c(s) \Delta A^0(s)-1} (1-u)^{c(s)[1-\Delta A^0(s)]-1}}{B(c(s) \Delta A^0(s), c(s)[1-\Delta A^0(s)])} (u) du \delta_{t_i}(ds), \end{aligned} \quad (3.7)$$

where the ratio in the second term is the Beta density of the jump size at  $s$  and  $\delta_{t_i}$  is the Dirac measure with unit mass at point  $t_i$ . There is an obvious resemblance between the expression for the Lévy measure of a process with discontinuities in

(3.7) and the Lévy measure of a posterior process in (3.3). In a homogeneous case under the Beta process prior with continuous Lévy measure, the posterior process is again a Beta process with discontinuities that are fixed in failure time points. In the upcoming work we will assume that the cumulative regression functions are a priori distributed as Beta processes with continuous Lévy measures as in (3.6). A-posteriori we get that these processes are Lévy with discontinuous Lévy measures as in (3.7) and even more, they are Beta processes outside the failure times.

The existence and features of the Beta process was derived by Hjort in [19] who came up with the idea within the time-discrete framework where it is true that the increments of the Beta process are exactly beta distributed. Then he extended this work to a time-continuous case showing that even though some desirable features (like convolution) of beta distribution stay preserved, the distribution of the increments of the Beta process is not exactly beta distributed. We can only say that

$$dA(t) \sim \text{beta}(c(t)dA^0(t), c(t)[1 - dA^0(t)])$$

in infinitesimal way.

Now, let us have an observed dataset on  $n$  homogeneous objects with possibly censored failure times in  $T_1, \dots, T_n$ . Let us suppose, that the corresponding cumulative hazard rate  $H$  is a priori a Beta process with parameters  $c$  and  $A^0$ , symbolically written as

$$H \sim \text{Beta}(c(\cdot), A^0(\cdot)).$$

It is natural, that we do not know in advance, in which sites the cumulative hazard rate will have jumps, so we assume a continuous  $A^0$ . The expectation of the process  $H$  a priori is  $A^0$  and it can be derived analogically as in p. 1272 in [19] from the Lévy representation of the Beta process in (3.4). We have that

$$\mathbb{E} \exp\{-\theta H(t)\} = \exp\left\{-\int_0^1 (1 - e^{-\theta u}) dL_t(u)\right\}, \quad (3.8)$$

and by differentiating both sides of the expression w.r. to  $\theta$  and then putting  $\theta$  equal zero leads us to

$$\begin{aligned} \mathbb{E} H(t) &= \int_0^1 u dL_t(u) = \int_0^t \int_0^1 u c(s) u^{-1} (1 - u)^{c(s)-1} du dA^0(s) \\ &= \int_0^t c(s) \frac{1}{c(s)} dA^0(s) = A^0(t). \end{aligned} \quad (3.9)$$

The prior variance is derived in similar fashion. By differentiating the equality in (3.8) w.r. to  $\theta$  twice and setting  $\theta$  to zero we get

$$\mathbb{E} H^2(t) = \left( \int_0^1 u \, dL_t(u) \right)^2 + \int_0^1 u^2 \, dL_t(u) = [A^0(t)]^2 + \int_0^t \frac{dA_0(s)}{c(s) + 1}.$$

Hence, we have  $\text{Var} H(t) = \mathbb{E} H^2(t) - [\mathbb{E} H(t)]^2 = \int_0^t dA^0(s)/(c(s) + 1)$ .

As we already advertised, subordinators are conjugate with right-censored i.i.d. survival data and the same is true when we choose a Beta process prior. The posterior distribution of the cumulative hazard rate is again a Beta process with parameters

$$H | \text{Data} \sim \text{Beta} \left( c(\cdot) + \bar{Y}(\cdot), \int_0^{(\cdot)} \frac{c(s)dA^0(s) + d\bar{N}(s)}{c(s) + \bar{Y}(s)} \right). \quad (3.10)$$

The derivation of this can be seen in [19], see especially Theorem 4.1 and Corollary 4.1. In particular, at the sites of failures the jumps in the cumulative hazard rate  $H$  are beta distributed with parameters

$$\Delta H(T_i) | \text{Data} \sim \text{beta} (d\bar{N}(T_i), c(T_i) + \bar{Y}(T_i) - d\bar{N}(T_i)).$$

We can take the posterior mean to be a Bayesian estimator. We already know, that a Beta process with fixed discontinuities can be written as a summation of the jump sizes at fixed discontinuities and a corresponding Beta process freed of these fixed jumps. Let us denote  $H_c = H - \Delta H$ . Then

$$\mathbb{E}(H(t) | \text{Data}) = \sum_{\substack{i: T_i \leq t \\ \delta_i = 1}} \mathbb{E} \Delta H(T_i) + \mathbb{E} H_c(t),$$

where the expectations on the right side are with regards to the posterior distribution. As  $H_c$  is a-posteriori a Beta process with parameters  $c(\cdot) + \bar{Y}(\cdot)$  and  $\int_0^{(\cdot)} c(s)dA^0(s)/(c(s) + \bar{Y}(s))$ , applying similar steps as when deriving the prior expectation in (3.9) together with the features of beta distribution we get the Bayesian estimator equal to (Theorem 4.3 in [19])

$$\begin{aligned} \mathbb{E}(H(t) | \text{Data}) &= \sum_{\substack{i: T_i \leq t \\ \delta_i = 1}} \frac{d\bar{N}(T_i)}{c(T_i) + \bar{Y}(T_i)} + \int_0^t \frac{c(s)dA^0(s)}{c(s) + \bar{Y}(s)} \\ &= \int_0^t \frac{c(s)dA^0(s) + d\bar{N}(s)}{c(s) + \bar{Y}(s)}. \end{aligned} \quad (3.11)$$

When looking at the estimator above, it becomes clear, that sending the parameter  $c$  to zero, i.e. imposing a vague prior, gives the Nelson-Aalen estimator. On the contrary, when  $c$  approaches infinity the estimator lands in the prior guess  $A^0$ .

Apparently, Beta processes can be a good choice for a prior distribution of a cumulative hazard rate and when taking the posterior mean to be the Bayesian estimator, it is easy to calculate it without need to use aid of MCMC procedures. Furthermore, as it was mentioned in Remark 3A in [19],  $L_t$  from (3.5) has a full support on the interval  $[0, 1]$ , meaning that the corresponding Beta process is a member of  $\mathcal{A}$  and hence a proper cumulative hazard rate. We will see, that this feature is preserved when we consider a cumulative hazard rate to be a sum of Beta processes, as it is in our case of Aalen model. Moreover, as the jumps of the sample paths are contained in  $[0, 1]$ , none of the cumulative regression functions can become negative at any point within  $[0, \infty]$ . This agrees well with the monotonicity we require.

Finally, let us formulate the prior distribution for the cumulative regression functions  $A_1, \dots, A_p$  of a monotone Aalen model. For each  $j = 1, \dots, p$ , we consider a non-negative piecewise constant function  $c_j$  and a continuous  $A_j^0$  such that  $A_j^0 \in \mathcal{A}$ . We will assume that a priori the cumulative regression functions  $A_j, j = 1, \dots, p$ , are distributed as a set of  $p$  independent Beta processes with parameters  $c_j$  and  $A_j^0$ , i.e.

$$A_j \sim \text{Beta}(c_j(\cdot), A_j^0(\cdot)), \quad j = 1, \dots, p. \quad (3.12)$$

Let us recall, that the expectation of the process  $A_j(t)$  a priori is  $A_j^0(t)$  while the prior variance equals  $\int_0^t dA_j^0(s)/(c_j(s) + 1)$ . Hence, the parameter  $c_j$  can be viewed as amount of trust of a statistician towards the prior guess  $A_j^0$ .

**Remark 7** Notice, that to make the calculations tractable we assumed that a priori the processes  $A_1, \dots, A_p$  are independent. This condition does not have to be fulfilled when working with a real data set.

## 3.2 Posterior distribution

Again, we will start from the likelihood of the model adjusted in the way, that it accommodates the jumps, as it was specified in Section 2.1. Let us recall that

the likelihood contribution of  $i$ th individual at the time point  $T_i$  is either

$$\prod_{s \in [0, T_i]} \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j} Y_i(s)} \quad (3.13)$$

in case of a censored observation, i.e. when  $T_i = C_i$  and  $\delta_i = 0$ , or

$$\prod_{s \in [0, T_i]} \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j} Y_i(s)} \left[ 1 - \prod_{j=1}^p \{1 - dA_j(T_i)\}^{x_{i,j}} \right]$$

if there has been observed a failure,  $T_i = T_i^0$  and  $\delta_i = 1$ . When we combine these with the prior information, which is of the following form

$$\prod_{s \in [0, \infty)} \prod_{j=1}^p \left[ dA_j(s)^{c_j(s) dA_j^0(s) - 1} \right] \{1 - dA_j(s)\}^{c_j(s)(1 - dA_j^0(s)) - 1},$$

we gather the form of the posterior distribution of the  $A_j$  processes.

Outside the failure times all the information from the observed data comes in form of (3.13) and a-posteriori the regression functions  $A_j$  in these times will be again distributed as Beta processes, with the increments approximately distributed as

$$dA_j(s) | \text{Data} \sim \text{beta}(c_j(s) dA_j^0(s), c_j(s)(1 - dA_j^0(s)) + R_j(s)),$$

where  $R_j(s) = \sum_{i=1}^n x_{i,j} Y_i(s)$ ,  $j = 1, \dots, p$ . This is equivalent with saying that outside the failure times the process  $A_j$  is a-posteriori a Beta process with parameters

$$A_j(t) | \text{Data} \sim \text{Beta} \left( c_j(t) + R_j(t), \int_0^t \frac{c_j(s) dA_j^0(s)}{c_j(s) + R_j(s)} \right). \quad (3.14)$$

This outcome is based on the work of Hjort in [19], and can be proven in detail using similar techniques as in the proof of Theorem 4.1 (i) there. The obtained result in (3.14) parallels the findings contained in Theorem 4.2 and Corollary 4.1 in [19]. The only difference is that the terms  $f_s^j(u)$  are multiplied by  $(1 - u)^{\sum_{i=1}^n x_{i,j} Y_i(s)} = (1 - u)^{R_j(s)}$  instead of  $(1 - u)^{\bar{Y}(s)}$ . This is given by a different likelihood of the observed data based on the Aalen model assumption.

Determining the posterior behaviour of  $A_j$ -s is more difficult in failure times. Let us consider the time point  $T_i$ , where  $i$ -th individual has experienced a failure and all other individuals have not. Then combining the likelihood and the prior

process leaves as with a following function:

$$\prod_{j=1}^p \left[ dA_j(s)^{c_j(s)dA_j^0(s)-1} \right] \{1 - dA_j(s)\}^{c_j(s)(1-dA_j^0(s))+R_j(s)-x_{i,j}-1} \\ \times \left[ 1 - \prod_{j=1}^p \{1 - dA_j(s)\}^{x_{i,j}} \right].$$

To get a better understanding of this function let us consider a time-discrete framework so that  $s \in S = \{0, \eta, 2\eta, \dots, \tau\}$ , for some  $\eta > 0$ . For the regression functions we have that they equal to  $A_j(s) = \sum_{i:i\eta \leq s} \alpha_j(i\eta)$ , where  $\alpha_j(i\eta) = \Delta A_j(i\eta) = A_j((i-1)\eta, i\eta]$  corresponds to the increment gained in the interval  $((i-1)\eta, i\eta]$ . The model can be reformulated from (2.1) to

$$1 - h_i(s) = \prod_{j=1}^p \{1 - \alpha_j(s)\}^{x_{i,j}}, \quad s \in S, \quad (3.15)$$

where  $\alpha_j(s)$  and  $h_i(s)$  are proportional to  $\eta$ , i.e. the  $\alpha_j(s)/\eta$  and  $h_i/\eta, \forall j, \forall i$ , have limits when  $\eta \rightarrow 0$ .

In this time-discrete setting, let us assume that the cumulative regression function  $A_j$  is a priori distributed as a time-discrete Beta process defined on  $S$  and with parameters equal to  $c_j(s)$  and  $A_j^0(s) = \sum_{i:i\eta \leq s} \alpha_j^0(i\eta)$ . Again,  $\alpha_j^0$  are assumed to be proportional to  $\eta$ . The increments of such a Beta process on  $S$  are *exactly* beta distributed with parameters

$$\alpha_j(s) \sim \text{beta}(c_j(s)\alpha_j^0(s), c_j(s)(1 - \alpha_j^0(s))), \quad s \in S.$$

Let us again have  $n$  individuals and  $n$  possibly censored times  $T_i \in S, i = 1, \dots, n$  ruled by the hazard rate  $h_i(s) = \sum_{j=1}^p x_{i,j}\alpha_j(s), s \in S$ . Then the posterior distribution of  $A_j$  outside the jump times is again a Beta process with

$$\alpha_j(s) | \text{Data} \sim \text{Beta} \left( c_j(s) + R_j(s), \frac{c_j(s)\alpha_j^0(s)}{c_j(s) + R_j(s)} \right),$$

and the random jumps in the process trajectory outside the failure times are beta distributed with parameters

$$\alpha_j(s) | \text{Data} \sim \text{beta}(c_j(s)\alpha_j^0(s), c_j(s)(1 - \alpha_j^0(s)) + R_j(s)).$$

This result is an analogy with the nonparametric analysis of time-discrete homogeneous data using a Beta process in [19]. For more details on time-discrete Beta processes see in particular Section 2 in the last cited paper.

The compound posterior distribution of the increments  $(\alpha_1(s), \dots, \alpha_p(s))$  at a time point when one of the individuals fails and others do not, e.g. at  $s = T_i \in S$ , is defined on  $[0, 1]^p$  and equals to

$$\frac{1}{k} \prod_{j=1}^p \alpha_j(s)^{c_j(s)\alpha_j^0(s)-1} (1 - \alpha_j(s))^{c_j(s)(1-\alpha_j^0(s))+R_j(s)-x_{i,j}-1} \left\{ 1 - \prod_{j=1}^p (1 - \alpha_j(s))^{x_{i,j}} \right\}.$$

The normalizing constant for this distribution is

$$k = \prod_{j=1}^p \Gamma(c_j(s)\alpha_j^0(s)) \left[ \prod_{j=1}^p \frac{\Gamma(c_j(s) - c_j(s)\alpha_j^0(s) + R_j(s) - x_{i,j})}{\Gamma(c_j(s) + R_j(s) - x_{i,j})} - \prod_{j=1}^p \frac{\Gamma(c_j(s) - c_j(s)\alpha_j^0(s) + R_j(s))}{\Gamma(c_j(s) + R_j(s))} \right]. \quad (3.16)$$

The expression for the constant  $k$  was determined by utilizing the fact that  $\Gamma(x)\Gamma(y)/\Gamma(x+y) = \int_0^1 u^{x-1}(1-u)^{y-1}du$ . Sending  $\eta$  to zero takes us back to the time-continuous case. When  $\eta \approx 0$ , the posterior density at the jump point  $s = T_i$  becomes close to

$$\frac{1}{k} \prod_{j=1}^p \alpha_j(s)^{-1} (1 - \alpha_j(s))^{c_j(s)+R_j(s)-x_{i,j}-1} \left\{ 1 - \prod_{j=1}^p (1 - \alpha_j(s))^{x_{i,j}} \right\}$$

with  $k$  approaching

$$k \approx \prod_{j=1}^p \frac{\Gamma(1 + c_j(s)\alpha_j^0(s))}{c_j(s)\alpha_j^0(s)} \times \sum_{j=1}^p [\psi(c_j(s) + R_j(s)) - \psi(c_j(s) + R_j(s) - x_{i,j})] c_j(s)\alpha_j^0(s) \quad (3.17)$$

where  $\psi(x) = \Gamma'(x)/\Gamma(x)$  is a digamma function. The equality  $\Gamma(c_j(s)\alpha_j^0(s)) = \Gamma(1 + c_j(s)\alpha_j^0(s))/c_j(s)\alpha_j^0(s)$  is a known feature of the Gamma function. The second part of the expression was reached by applying the following approximation

$$\frac{\Gamma(c_j(s) + R_j(s) - c_j(s)\alpha_j^0(s))}{\Gamma(c_j(s) + R_j(s))} \approx -\psi(c_j(s) + R_j(s))c_j(s)\alpha_j^0(s) + 1,$$

which can be obtained from a basic derivative rule  $F(x+dx) \approx F(x) + F'(x)dx$ .

By looking closer at the expression for  $k$  in (3.17) it is understood that the product is of order  $1/\Delta^p$  while the terms in summation are of order  $\Delta$  and the whole expression is asymptotically proportional to  $1/\Delta^{p-1}$ . Hence the derived posterior for continuous distribution via time-discretization makes sense only if



$p = 1$ . This complication is overcome by letting only one of the  $\alpha_j$ -s to be positive while keeping the others equal to zero.

Let us return to the continuous setting. In next, we will state the form of the posterior distribution of jumps  $(\Delta A_1(T_i), \dots, \Delta A_p(T_i))$  in failure time point  $T_i$ . Following lemma will be a cornerstone for establishing the Bayesian estimators later in this chapter. Before we move onto the lemma, let us remind that in time-continuous case we assume that the functions  $A_j^0$  are continuous, hence there exist  $\alpha_j^0, j = 1, \dots, p$  such that  $A_j^0(t) = \int_0^t \alpha_j^0(s) ds$ .

**Lemma 2** *Assume that  $(T_i, \delta_i) = (t, 1)$  and that there are no other failures at this time point. Then only one of the  $A_j$  processes jumps at  $t$ ; with probability*

$$p_{ij}(t) = \frac{[\psi(c_j(t) + R_j(t)) - \psi(c_j(t) + R_j(t) - x_{i,j})] c_j(t) \alpha_j^0(t)}{\sum_{k=1}^p [\psi(c_k(t) + R_k(t)) - \psi(c_k(t) + R_k(t) - x_{i,k})] c_k(t) \alpha_k^0(t)} \quad (3.18)$$

the jump  $\Delta A_j(t)$  is positive and coming from the density

$$g_{ij}(u) = \frac{u^{-1} (1-u)^{c_j(t)+R_j(t)-x_{i,j}-1} \{1 - (1-u)^{x_{i,j}}\}}{\psi(c_j(t) + R_j(t)) - \psi(c_j(t) + R_j(t) - x_{i,j})} \quad (3.19)$$

while all other jumps  $\Delta A_k(t)$  are equal to zero.

**PROOF** To prove this result we will resort to the time-discretization. Similarly as in the motivation on previous pages, we assume that the time variables are observed on intervals of length  $\eta$ . Then we are able to derive the explicit simultaneous density for jumps  $(\Delta A_1(s), \dots, \Delta A_p(s)) = (A_1(s - \eta, s], \dots, A_p(s - \eta, s])$  given that  $T_i \in (s - \eta, s]$  and given the rest of the data. It follows from the results on previous page that this density is

$$g_\eta(u_1, \dots, u_p) = \frac{1}{k} \prod_{j=1}^p u_j^{c_j(s) A_j^0(s-\eta, s]-1} (1-u_j)^{c_j(s)-c_j(s) A_j^0(s-\eta, s]+R_j(s)-x_{i,j}-1} \\ \times \left\{ 1 - \prod_{j=1}^p (1-u)^{x_{i,j}} \right\}. \quad (3.20)$$

The normalizing constant  $k$  is like in (3.16), in detail

$$k = \prod_{j=1}^p \Gamma(c_j(s) A_j^0(s - \eta, s]) \left[ \prod_{j=1}^p \frac{\Gamma(c_j(s) - c_j(s) A_j^0(s - \eta, s] + R_j(s) - x_{i,j})}{\Gamma(c_j(s) + R_j(s) - x_{i,j})} \right. \\ \left. - \prod_{j=1}^p \frac{\Gamma(c_j(s) - c_j(s) A_j^0(s - \eta, s] + R_j(s))}{\Gamma(c_j(s) + R_j(s))} \right].$$

We need to prove that the limit of the distribution  $g_\eta(u_1, \dots, u_p)$  when  $\eta \rightarrow 0$  is as in (3.19). This can be done showing that all product moments converge to their respective counterparts, as all product moments uniquely characterize the distribution (see e.g. [37]). Let us denote the random vector distributed accordingly to (3.20), by  $(U_1, \dots, U_p)$ . Then for the  $q$ -th moment  $\mathbb{E} U_l^q$  we have

$$\begin{aligned} & \int_0^1 \dots \int_0^1 \frac{1}{k} \prod_{j=1}^p u_j^{c_j(s)A_j^0(s-\eta, s]-1} (1-u_j)^{c_j(s)(1-A_j^0(s-\eta, s])+R_j(s)-x_{i,j}-1} du_1 \dots du_p \\ &= \frac{1}{k} \left\{ \prod_{\substack{j=1 \\ j \neq l}}^p \frac{\Gamma(c_j(s)A_j^0(s-\eta, s])\Gamma(c_j(s)(1-A_j^0(s-\eta, s])+R_j(s)-x_{i,j})}{\Gamma(c_j(s)+R_j(s)-x_{i,j})} \right. \\ & \quad \times \frac{\Gamma(q+c_l(s)A_l^0(s-\eta, s])\Gamma(c_l(s)(1-A_l^0(s-\eta, s])+R_l(s)-x_{i,l})}{\Gamma(q+c_l(s)+R_l(s)-x_{i,l})} \\ & \quad - \prod_{\substack{j=1 \\ j \neq l}}^p \frac{\Gamma(c_j(s)A_j^0(s-\eta, s])\Gamma(c_j(s)(1-A_j^0(s-\eta, s])+R_j(s))}{\Gamma(c_j(s)+R_j(s))} \\ & \quad \left. \times \frac{\Gamma(q+c_l(s)A_l^0(s-\eta, s])\Gamma(c_l(s)(1-A_l^0(s-\eta, s])+R_l(s))}{\Gamma(q+c_l(s)+R_l(s))} \right\}, \end{aligned}$$

and for  $\eta$  close to zero we have

$$\begin{aligned} \mathbb{E} U_l^q(s) &\approx \frac{1}{\Gamma(c_l A_l^0(s-\eta, s] + 1)} \frac{p_{il}(s)\Gamma(q+c_l(s)A_l^0(s-\eta, s])}{\psi(c_l(s)+R_l(s)) - \psi(c_l(s)+R_l(s)-x_{i,l})} \\ &\quad \times \left\{ \frac{\Gamma(c_l(s)(1-A_l^0(s-\eta, s])+R_l(s)-x_{i,l})}{\Gamma(q+c_l(s)+R_l(s)-x_{i,l})} \right. \\ &\quad \left. - \frac{\Gamma(c_l(s)(1-A_l^0(s-\eta, s])+R_l(s))}{\Gamma(q+c_l(s)+R_l(s))} \right\}. \end{aligned}$$

The limit is equal to

$$\begin{aligned} \mathbb{E} U_l^q(s) &\xrightarrow{\eta \rightarrow 0} \frac{p_{il}(s)\Gamma(q)}{\psi(c_l(s)+R_l(s)) - \psi(c_l(s)+R_l(s)-x_{i,l})} \\ &\quad \times \left\{ \frac{\Gamma(c_l(s)+R_l(s)-x_{i,l})}{\Gamma(q+c_l(s)+R_l(s)-x_{i,l})} - \frac{\Gamma(c_l(s)+R_l(s))}{\Gamma(q+c_l(s)+R_l(s))} \right\}. \end{aligned} \quad (3.21)$$

The  $q$ -th moment  $\mathbb{E} U_l^q(s)$  of the limiting distribution stated in the lemma equals to

$$p_{il}(s) \int_0^1 u^q g_{il}(u) du$$

and by using the similar techniques as before we get that it is exactly the limit in (3.21). Hence, for all  $q \geq 1$  we have that the  $q$ -th moment of the time-discrete distribution converges to the  $q$ -th moment of the distribution stated in the lemma

with  $\eta \rightarrow 0$ . Hence, the limiting distribution of jumps  $(\Delta A_1(t), \dots, \Delta A_p(t))$  at  $t = T_i$  is given by the limits of all the product moments and it equals the distribution given in the lemma.

Q.E.D.

### 3.3 Bayesian estimators

The expectation of a posterior distribution is often taken to be a Bayesian estimator. Similarly as in the i.i.d. case in (3.11) the posterior distribution of a cumulative regression function consists of two independent components. The first component is the stochastically continuous Beta process outside the failure times while the second is comprised solely of jumps placed at fixed discontinuities at failure sites.

From Lemma 2 it is seen that the posterior distribution of a jump occurrence in the trajectory of process  $A_j$  at every failure time  $T_i$  is of Bernoulli distribution with probability  $p_{ij}$  in (3.18). Let us denote by  $U_{ij}$  the size of the jump of the process  $A_j$  at the failure time  $T_i$  in case it occurs.  $U_{ij}$  is a random variable ruled by the density  $g_{ij}$  in (3.19). The expectation and variance of  $U_{ij}$ , denoted e.g.  $\xi_{ij}$  and  $\sigma_{ij}^2$  respectively, are easily calculated. The expectation  $\xi_{ij}$  equals to

$$\begin{aligned} \xi_{ij}(T_i) &= \mathbb{E} U_{ij} = [\psi(c_j(T_i) + R_j(T_i)) - \psi(c_j(T_i) + R_j(T_i) - x_{i,j})]^{-1} \\ &\quad \times \left( \int_0^1 (1-u)^{c_j(T_i)+R_j(T_i)-x_{i,j}-1} du - \int_0^1 (1-u)^{c_j(T_i)+R_j(T_i)-1} du \right) \\ &= \frac{(c_j(T_i) + R_j(T_i) - x_{i,j})^{-1} - (c_j(T_i) + R_j(T_i))^{-1}}{\psi(c_j(T_i) + R_j(T_i)) - \psi(c_j(T_i) + R_j(T_i) - x_{i,j})}, \end{aligned} \quad (3.22)$$

and similarly, for the variance  $\sigma_{ij}^2$  we have

$$\begin{aligned} \sigma_{ij}^2(T_i) &= \text{Var } U_{ij} = \frac{(c_j(T_i) + R_j(T_i) - x_{i,j})^{-1} (c_j(T_i) + R_j(T_i) - x_{i,j} + 1)^{-1}}{\psi(c_j(T_i) + R_j(T_i)) - \psi(c_j(T_i) + R_j(T_i) - x_{i,j})} \\ &\quad - \frac{(c_j(T_i) + R_j(T_i))^{-1} (c_j(T_i) + R_j(T_i) + 1)^{-1}}{\psi(c_j(T_i) + R_j(T_i)) - \psi(c_j(T_i) + R_j(T_i) - x_{i,j})} - \xi_{ij}(T_i)^2. \end{aligned}$$

Combining the expectation of the stochastically continuous part of the posterior and the expectations of the jump sizes at the points of fixed discontinuities located at the failure times we arrive at Bayes estimators of  $A_j$ -s written as

$$\tilde{A}_j(t) = \mathbb{E}(A_j(t) | \text{Data}) = \int_0^t \frac{c_j(s) dA_j^0(s)}{c_j(s) + R_j(s)} + \sum_{T_i \leq t, \delta_i=1} p_{ij}(T_i) \xi_{ij}(T_i). \quad (3.23)$$

The first part of the expression is the expectation of the Beta process with parameters  $c_j + R_j$  and  $\int c_j(s) dA_j^0(s)/(c(s) + R_j(s))$ , obtained by similar techniques as in (3.9).

The posterior variance is again a summation of the variance of the stochastically continuous part and the variance in the fixed discontinuities. This is true thanks to the fact that the continuous component and the fixed discontinuities of the posterior distribution are independent. The sought expression for variance is

$$\begin{aligned} \text{Var}(A_j(t) | \text{Data}) &= \int_0^t \frac{c_j(s) dA_j^0(s)}{(c_j(s) + R_j(s))(c_j(s) + R_j(s) + 1)} \\ &+ \sum_{T_i \leq t, \delta_i=1} [p_{ij}(T_i) \sigma_{ij}^2(T_1) + p_{ij}(T_i)(1 - p_{ij}(T_i)) \xi_{ij}(T_i)^2]. \end{aligned} \quad (3.24)$$

We obtained the posterior variance of the Beta process outside the failure times similarly as the prior variance of a Beta process in the i.i.d. case in Section 3.1. The last term in the variance formula was derived by using the knowledge on  $\xi_{ij}$  and  $\sigma_{ij}^2$ , the first and second moments of size of the jumps at fixed discontinuities. In detail, the variance in jump times equals

$$\begin{aligned} \text{Var}(\Delta A_j(T_i)) &= \text{E}(\Delta A_j(T_i))^2 - (\text{E} \Delta A_j(T_i))^2 = p_{ij}(T_i) \text{E} U_{ij}^2 - p_{ij}^2(T_i) \xi_{ij}^2(T_i) \\ &= p_{ij}(T_i) \{\text{Var} U_{ij} + (\text{E} U_{ij})^2\} - p_{ij}^2(T_i) \xi_{ij}^2(T_i) \\ &= p_{ij}(T_i) \{\sigma_{ij}^2(T_i) + \xi_{ij}^2(T_i)\} - p_{ij}^2(T_i) \xi_{ij}^2(T_i). \end{aligned}$$

Knowing the variance of  $\tilde{A}_j$  is useful for setting the pointwise credibility bands for the estimators. Furthermore, the covariance between two processes  $A_j$  and  $A_k$  is equal to

$$\text{cov}(A_j(t), A_k(t) | \text{Data}) = - \sum_{T_i \leq t, \delta_i=1} p_{ij}(T_i) p_{ik}(T_i) \xi_{ij}(T_i) \xi_{ik}(T_i).$$

The digamma function in the expressions (3.18), (3.19) and (3.22) can be for large  $R_j$ -s and small  $c_j$ -s (relatively to  $R_j$ -s) approximated by an asymptotical expansion  $\psi(x) = \log(x) + O(x^{-1})$ , see [7]. Applying this gives simplified asymptotically equivalent versions of formulas for  $p_{ij}$  and  $\xi_{ij}$ . This approximations can be further applied to gain approximated versions of the Bayesian estimators  $\tilde{A}_j$ . We have

$$\begin{aligned} \xi_{ij}(T_i) &\approx \frac{(c_j(T_i) + R_j(T_i) - x_{i,j})^{-1} - (c_j(T_i) + R_j(T_i))^{-1}}{\log(c_j(T_i) + R_j(T_i)) - \log(c_j(T_i) + R_j(T_i) - x_{i,j})} \\ &\approx \frac{(c_j(T_i) + R_j(T_i) - x_{i,j})^{-1} - (c_j(T_i) + R_j(T_i))^{-1}}{(c_j(T_i) + R_j(T_i))/(c_j(T_i) + R_j(T_i) - x_{i,j}) - 1} \end{aligned}$$

hence

$$\xi_{ij}(T_i) \approx \frac{1}{c_j(T_i) + R_j(T_i)} \approx \frac{1}{R_j(T_i)}$$

and similarly

$$\begin{aligned} p_{ij}(T_i) &\approx \frac{[\log(c_j(T_i) + R_j(T_i)) - \log(c_j(T_i) + R_j(T_i) - x_{i,j})] c_j(T_i) \alpha_j^0(T_i)}{\sum_{k=1}^p [\log(c_k(T_i) + R_k(T_i)) - \log(c_k(T_i) + R_k(T_i) - x_{i,k})] c_k(T_i) \alpha_k^0(T_i)} \\ &\approx \frac{x_{i,j}/(c_j(T_i) + R_j(T_i) - x_{i,j}) c_j(T_i) \alpha_j^0(T_i)}{\sum_{k=1}^p x_{i,k}/(c_k(T_i) + R_k(T_i) - x_{i,k}) c_k(T_i) \alpha_k^0(T_i)} \\ &\approx \frac{x_{i,j}/R_j(T_i) c_j(T_i) \alpha_j^0(T_i)}{\sum_{k=1}^p x_{i,k}/R_k(T_i) c_k(T_i) \alpha_k^0(T_i)}. \end{aligned}$$

Inserting these approximations into (3.23) yields

$$\begin{aligned} \tilde{A}_j(t) &\approx \int_0^t \frac{c_j(s) dA_j^0(s)}{R_j(s)} + \sum_{T_i \leq t, \delta_i = 1} \frac{1}{R_j(T_i)} \frac{x_{i,j}/R_j(T_i) c_j(T_i) \alpha_j^0(T_i)}{\sum_{k=1}^p x_{i,k}/R_k(T_i) c_k(T_i) \alpha_k^0(T_i)} \\ &\approx \sum_{i=1}^n \int_0^t V_{ij}(s) \frac{dN_i(s)}{R_j(s)} \end{aligned} \quad (3.25)$$

where

$$V_{ij}(s) = \frac{x_{i,j}/R_j(s) c_j(s) \alpha_j^0(s)}{\sum_{k=1}^p x_{i,k}/R_k(s) c_k(s) \alpha_k^0(s)}. \quad (3.26)$$

Note, that in expression (3.25) we dropped the first term  $\int c_j/R_j dA_j^0$  coming from the stochastically continuous part of the solution as for  $c_j$  small relatively to  $R_j$  it is close to zero. Using similar techniques leads us to an approximated expression for the variance of the Bayesian estimator, viz.

$$\begin{aligned} \text{Var}(\tilde{A}_j(t)) &\approx \\ &\sum_{i=1}^n \int_0^t V_{i,j}(s) \left[ \frac{2R_j(s) - x_{i,j} + 1}{R_j(s)(R_j(s) + 1)(R_j(s) - x_{i,j} + 1)} + \frac{1 - V_{i,j}(s)}{R_j^2(s)} \right] dN_i(s). \end{aligned} \quad (3.27)$$

Finally, it can be seen that the Bayesian estimator in (3.25) is for large  $n$  of a similar form as the NPML estimator. The difference is, however, present in the "weighting" processes  $V_{i,j}$ . While in NPML the values of  $V_{i,j}$  were either 0 or 1, here the weights are contained in the interval  $[0, 1]$ . Similarly as in NPML they sum up to  $\sum_{j=1}^p V_{i,j} = 1$ . It means that all Bayesian estimators  $\tilde{A}_1, \dots, \tilde{A}_p$  jump at every site of the observed failure while in NPML only one of the processes  $\hat{A}_1, \dots, \hat{A}_p$  jumps.

The validity of the approximations of the Bayesian estimators  $\tilde{A}_j$ -s is conditioned on the assumptions on  $c_j$ -s and  $R_j$ -s. Determining sensible functions for  $c_j$ -s is a question of choosing between functions with small enough values to

give us reasonably good approximations and greater valued functions that produce a less restrictive prior (we know, that small  $c_j$ -s imply small variance of the prior Beta process). However, as long as the covariates are not very close to zero with growing size we have growing  $R_j$ -s and asymptotically the exact and approximated formulas for  $\tilde{A}_j$ -s become equivalent.

The estimators  $\tilde{A}_j$  can be calculated directly due to the fact that the posterior mean has an explicit (although complicated) form. When the number of observations is great, we can use the approximated versions of the estimator. Another way to obtain the estimators is to use the aid of MCMC algorithms. The essential part of MCMC procedure is an effective generation of the processes  $A_j, \forall j$ , from the posterior distributions. Here the problem separates into two steps: generation of the stochastically continuous part (a Beta process) and simulation of a jump size in the fixed discontinuities.

Let a Beta process with parameters  $c_j$  and  $A_j^0$  be a prior process for  $A_j, \forall j$ . Then under monotone Aalen model, from (3.14) we have that the posterior distributions for  $A_j$ -s are Beta processes in the zone outside of jump times,

$$A_j(t) | \text{Data} \sim \text{Beta} \left( c_j(t) + R_j(t), \int_0^t \frac{c_j(s) dA_j^0(s)}{c_j(s) + R_j(s)} \right).$$

Recall, that  $R_j(t) = \sum_{i=1}^n x_{i,j} Y_i(t), j = 1, \dots, p$ . As shown in [33], a Beta process can be well enough approximated with a compound Poisson process. As we know from (3.6), the Lévy measure of a Beta process  $A$  with parameters  $c$  and  $A_0$  and with no fixed discontinuities equals to

$$\nu(ds, du) = c(s)u^{-1}(1-u)^{c(s)-1} dA^0(s) du, \quad u \in (0, 1).$$

Following the work in [33], let us consider a Lévy process  $A_\epsilon$  with Lévy measure equal to

$$\nu_\epsilon(ds, du) = \frac{\Gamma(\epsilon + c(t))}{\Gamma(\epsilon)\Gamma(c(t))} c(s)u^{\epsilon-1}(1-u)^{c(s)-1} dA^0(s) du, \quad u \in (0, 1).$$

It was shown in Theorem 2 in [33] that under the fulfilled assumptions of  $0 < \inf_{0 \leq s \leq \tau} c(s) \leq \sup_{0 \leq s \leq \tau} c(s) < \infty$  and  $A(\tau) < \infty$  for some  $\tau > 0$ , the process  $A_\epsilon \xrightarrow{\mathcal{D}} A$  on  $D[0, \tau]$  with  $\epsilon \rightarrow 0$ .

Thus generating a sample path of the Lévy process  $A_\epsilon$  for  $\epsilon$  small enough should provide us with a reasonable approximation of a sample path of process  $A$ . The total mass of the Lévy measure  $\nu_\epsilon$  on  $[0, \tau]$  equals to  $\lambda = \int_0^\tau c(t) dA_0(t)/\epsilon$  and it is finite for  $\epsilon > 0$ . Hence we can simulate a random path of  $A_\epsilon$  as a compound Poisson process with the number of jumps following  $Poisson(\lambda)$ , sites of jumps

distributed according to distribution with the probability density proportional to  $c(t)dA_0(t)I\{t \in [0, \tau]\}$  and jumps size distribution equal to  $beta(\epsilon, c(t))$ .

The difficulties lie in the observed death times where according to Lemma 2 we get fixed discontinuity in the process  $A_j$  at time  $T_i$ , such that  $\delta_i = 1$ , with probability  $p_{i,j}(t)$  in (3.18) and the jump size  $\Delta A_j(T_i)$  comes from the density  $g_{i,j}(u)$  in (3.19), while other jumps  $\Delta A_k(T_i)$  are zero. The choice, which of the processes jumps, is done randomly by drawing a variable from corresponding multinomial distribution. Sampling a value of the jump size can be done by using two latent variables, as it was done for Cox model in [32]. This reduces the problem to a simple generation from exponential, truncated exponential and gamma distribution for each fixed discontinuity.

Firstly, transform  $u \rightarrow v = -\log(1 - u)$ . It is seen that  $v$  spans the whole positive part of real line,  $v \in (0, \infty)$ . The posterior distribution of the jump size becomes

$$g_{i,j}^v(v) \propto (1 - \exp\{-v\})^{-1} \exp\{-v[c_j(T_i) + R_j(T_i) - x_{i,j} - 1]\}(1 - \exp\{-vx_{i,j}\}).$$

Following the two latent variable approach from [32], we assume to have a geometric variable  $y$  and a truncated exponential variable  $w$  with distributions given in

$$[y|v] \sim geom(1 - \exp\{-v\})$$

and

$$[w|v] \sim vx_{i,j} \exp\{-vx_{i,j}w\}(1 - \exp\{-vx_{i,j}\})^{-1}I_{(0,1)}(w).$$

In the last we stated the density of the truncated exponential distribution. Now it can be noticed, that under the fixed  $y$  and  $w$  the distribution of the sought parameter  $v$  is gamma with parameters

$$[v|y, w] \sim gamma(2, c_j(T_i) + R_j(T_i) - x_{i,j} + y + wx_{i,j}).$$

Once the value of  $v$  is generated, the actual jump size is easily found as  $u = 1 - e^{-v}$ .

Let us have  $T_1, \dots, T_n$  survival times collected from  $n$  individuals ruled by a monotone Aalen model. Suppose, that  $l$  of these times are recorded deaths and denote  $t_1, \dots, t_l$  the uncensored failure times. Let  $M$  be a total number of iterations of the Gibbs sampler and let us denote  $A_j^{(m)}$ ,  $m = 1, \dots, M$ ,  $j = 1, \dots, p$ , the sample paths of the posterior Beta processes of the cumulative regression functions  $A_j$ ,  $j = 1, \dots, p$ . Let us have a small enough  $\epsilon > 0$ . Set  $\tau < \infty$ . Before we move onto the algorithm, notice that starting values of the jump sizes in fixed

discontinuities are needed. These can be drawn from the prior distribution, i.e.

$$\Delta A_j^{(0)}(t_i) \sim \text{beta}(c(t_i)\Delta A_j^0(t_i), c(t_i)(1 - \Delta A_j^0(t_i))).$$

We will summarize all the steps of the MCMC algorithm in next:

### Sampling algorithm for $A_1, \dots, A_p$ :

1. Let  $m \leftarrow 1$ .
2. For every  $i = 1, \dots, l$  draw  $U \sim U[0, 1]$ , if  $\sum_{k=1}^{j-1} p_{i,k} \leq U < \sum_{k=1}^j p_{i,k}$  (taking  $\sum_{k=1}^0 p_{i,k} = 0$ ), then  $A_j^{(m)}$  jumps at  $t_i$ , i.e.  $\Delta A_j^{(m)} > 0$ , while  $\Delta A_k^{(m)} = 0, \forall k \neq j$ .
3. Let  $j \leftarrow 1$ .
4. **Stochastically continuous part:**
  - Sample  $K$ , the number of random jumps from  $Poisson(\lambda)$ , where  $\lambda = \frac{1}{\epsilon} \int_0^\tau c_j(t) dA_j^0(t)$ .
  - Sample the random jump sites  $s_1, \dots, s_K$  from the probability density proportional to  $c_j(t) dA_j^0(t) I\{t \in [0, \tau]\}$  and order them.
  - Sample the jump sizes  $z_1, \dots, z_K$  at sites  $s_1, \dots, s_K$  from the beta distribution with  $[z_i | s_i] \sim \text{beta}(\epsilon, R_j(s_i) + c_j(s_i))$ .
5. **Fixed discontinuities at  $t_1, \dots, t_l$ :**
  - Set  $i \leftarrow 1$ .
  - If the process  $A_j^{(m)}$  does not jump at  $t_i$ , then set  $\Delta A_j^{(m)}(t_i) = 0$ , else
    - ◊ Set  $v = -\log(1 - \Delta(A_j^{(m-1)}(t_i)))$ .
    - ◊ Sample  $[y|v] \sim \text{geom}(1 - \exp\{-v\})$ .
    - ◊ Sample  $x$  from a truncated exponential distribution with probability density  $v x_{i,j} \exp\{-v x_{i,j} w\} (1 - \exp\{-v x_{i,j}\})^{-1} I_{(0,1)}(w)$ .
    - ◊ Sample  $[v|y, w] \sim \text{gamma}(2, c_j(T_i) + R_j(T_i) - x_{i,j} + y + w x_{i,j})$ .
    - ◊ Set  $\Delta A_j^{(m)}(t_i) = 1 - e^{-v}$ .
  - Set  $i \leftarrow i + 1$ .
6. Set  $j \leftarrow j + 1$  and do steps 4 and 5. Repeat until  $j = p$ .
7. Set  $m \leftarrow m + 1$  (till  $m$  reach a large  $M$ ) and return back at step 2.



The output of the algorithm is an MCMC chain whose members are sample paths of  $A_j, j = 1, \dots, p$ . After discarding of several starting members (burn-in part), we can produce pointwise estimators of  $A_j, j = 1, \dots, p$  and  $(1-\alpha)100\%$  pointwise credibility bands for the estimators by taking a 50% quantile,  $\alpha/2 \times 100\%$  and  $(1 - \alpha/2) \times 100\%$  quantile respectively.

### 3.4 Case $p = 1$

As we already specified in Section 2.2, if there is only one covariate present, the hazard rate for  $i$ th individual equals  $h_i(s) = x_{i,1}\alpha_1(s)$  and transforming the covariate into  $x_{i,1} = \exp\{\beta w_i\}$  for some arbitrarily chosen  $\beta \in \mathbb{R} \setminus 0$ , we get the familiar expression of the hazard rate  $h_i(s) = \alpha_1(s)e^{\beta w_i}$  for one-covariate Cox model with an unknown baseline hazard  $\alpha_1$  and known  $\beta$ . Also the Bayesian estimator for the cumulative regression function  $A_1 = \int \alpha_1(s)ds$  of a monotone Aalen model under a Beta process prior agrees with the estimator of baseline hazard rate of a Cox model with known  $\beta$ , see p. 1284 in [19].

As it can be seen from [19], the Bayesian estimator of  $A_1$  equals

$$\tilde{A}_1(t) = E(A_1(t) | \text{Data}) = \int_0^t \frac{c_1(s)dA_1^0(s) + J(s)d\bar{N}(s)}{c_1(s) + R_1(s)},$$

in which  $J(s)$  is defined via

$$\frac{(c_1(s) + R_1(s) - x_{i,1})^{-1} - (c_1(s) + R_1(s))^{-1}}{\psi(c_1(s) + R_1(s)) - \psi(c_1(s) + R_1(s) - x_{i,1})} = \frac{J(s)}{c_1(s) + R_1(s)}.$$

Applying the approximation  $\psi(x) = \log(x) + O(x^{-1})$  shows that, for large  $n$ , term  $J(s)$  is close to 1 and therefore the estimator  $\tilde{A}_1(t)$  is close to

$$\int_0^t \left\{ \frac{c(s)dA_1^0(s)}{c_1(s) + R_1(s)} + \frac{d\bar{N}(s)}{c_1(s) + R_1(s)} \right\} \approx \int_0^t \frac{d\bar{N}(s)}{R_1(s)}.$$

From the last expression it is seen that with growing  $n$  the impact of the prior information vanishes and  $\tilde{A}_1(t)$  behaves similarly as the NPML estimator - and we already know that in case of  $p = 1$  the NPML estimator proved to be consistent. This naturally implies the consistency of the Bayesian estimator  $\tilde{A}_1$  of  $A_1$ .

A study of the asymptotic distribution reveals that a statement known under the name Bernstein - von Mises theorem is in force here. This theorem states that the posterior distribution centered around NPML estimator  $\hat{A}_1 = \int_0^t \sum_{i=1}^n \left\{ 1 - (1 - x_i/R_1(s))^{1/x_{i,1}} \right\} dN_i(s)$  is asymptotically equal to the asymp-

otic distribution of NPMLE. Even though asymptotic features are rarely of main interest of Bayesians, the Bernstein-von Mises assertion is not an unusual outcome and for parametric problems is valid under fairly mild conditions (see e.g. [47]).

**Theorem 7 (Bernstein - von Mises theorem for  $p = 1$ )** *Let us suppose that the conditions in Assumption (\*) in Section 2.1 are fulfilled. Let  $\hat{A}_1$  be the NPML estimator of the cumulative regression function  $A_1 = \int \alpha_1(s)ds$ . Then, for  $\alpha_1$  continuous and  $\bar{Y}(\tau) \rightarrow \infty$ , the process  $\sqrt{n}(A_1(t) - \hat{A}_1(t))$  a-posteriori converges weakly on  $D[0, \tau]$ , w.p. 1, to a zero-mean Gaussian process  $W$  with independent increments and variance function equal to  $C(t) = \int_0^t \alpha_1(s)/r_1(s)ds$ ,*

$$\sqrt{n}(A_1(t) - \hat{A}_1(t) \mid (T_i, \delta_i, z_i); i = 1, \dots, n) \xrightarrow{\mathcal{D}} W(C(t)), \quad \text{w.p. 1.}$$

PROOF A similar problem was already solved by Kim in [28] where he deals with a general version of Cox model with  $p > 1$  covariates and unknown  $\beta$ . As we already know, Aalen model with only one covariate can be viewed as a Cox model with the covariate  $w_i = \log x_{i,1}/\beta$ , for some  $\beta \in \mathbb{R} \setminus 0$ . The proof would further go along the lines of the proof of Theorem 3.2. in [28], starting with decomposition

$$\begin{aligned} \sqrt{n}(A_1(t) - \hat{A}_1(t) \mid \text{Data}) &= \sqrt{n}(A_1(t) - A_1^D(t) \mid \text{Data}) \\ &\quad + \sqrt{n}(A_1^D(t) - \mathbb{E} A_1^D(t) \mid \text{Data}) \\ &\quad + \sqrt{n}(\mathbb{E} A_1^D(t) - \hat{A}_1(t) \mid \text{Data}), \end{aligned}$$

where we denoted the part with fixed discontinuities in the failure times by  $A_1^D(t) = \sum_{T_i \leq t} \Delta A_1(T_i)$ . The main task is to show that the first term (i.e. the stochastically continuous part) and the third term converge weakly to 0, w.p. 1, while the second term converges weakly to the process  $W(C(t))$ , w.p. 1. These quests are done in same fashion as in [28] as the convergence either to zero or to the Gaussian process is proved conditionally on  $\beta$ . The only difference is that we do not have a term  $xe_0$  (where  $x = \sqrt{n}(\beta - \hat{\beta})$ ,  $\hat{\beta}$  is the MLE of  $\beta$  and  $e_0 = \int \mathbb{E} w e^{\beta w} Y / \mathbb{E} e^{\beta w} Y$ ) included in the asymptotic process. This term is in Cox model with an unknown  $\beta$  related to uncertainty introduced by this unknown parameter.

Q.E.D.

The posterior distribution of  $A_1$  and Hadamard differentiability hand in hand with the functional delta method (see for instance Section 20 in [47]) gives a way

to establish analogical result for any smooth functional of  $A_1$ , e.g. the distribution function.

### 3.5 Case $p > 1$

Similarly as in NPML estimation, the situation changes dramatically once we have more than one covariate in the model. Let us go back to the approximation of the Bayesian estimators in (3.25) which under the usual conditions gives us a clue about the tendencies of the estimators with growing  $n$ , viz.

$$\tilde{A}_j(t) \approx \sum_{i=1}^n \int_0^t V_{ij}(s) \frac{dN_i(s)}{R_j(s)} \approx \sum_{i=1}^n \int_0^t \frac{1}{R_j(s)} V_{ij}(s) Y_i(s) z_i^\top \alpha(s) ds.$$

Looking closer at the weighting factor  $V_{ij}$  in (3.26), we see that with growing  $n$  it is similar to

$$V_{ij}(s) = \frac{x_{i,j}/R_j(s) c_j(s) \alpha_j^0(s)}{\sum_{k=1}^p x_{i,k}/R_k(s) c_k(s) \alpha_k^0(s)} \approx V_j(s) = \frac{x_j/r_j(s) c_j(s) \alpha_j^0(s)}{\sum_{k=1}^p x_k/r_k(s) c_k(s) \alpha_k^0(s)}.$$

Now we can proceed towards the inconsistency theorem analogical to its counterpart Theorem 6 in NPML estimation.

**Theorem 8 (Inconsistency of  $\tilde{A}_j(t)$ )** *Let us suppose that the conditions in Assumption (\*) in Section 2.1 are fulfilled and let us denote*

$$E_j(s) = \sum_{i=1}^n V_{i,j}(s) Y_i(s) z_i^\top \alpha(s), \quad s \in [0, \tau].$$

*If  $\bar{Y}(\tau) \rightarrow \infty$  with  $n \rightarrow \infty$ ,  $\alpha_j, j = 1, \dots, p$ , are continuous and  $E z_j^2 < \infty, \forall j$ , then there exist functions  $r_j(s)$  and  $e_j(s)$ ,  $j = 1, \dots, p$ , on  $[0, \tau]$ , such that*

$$\sup_{0 \leq s \leq \tau} \left| \frac{R_j(s)}{n} - r_j(s) \right| \xrightarrow{P} 0 \quad \text{and} \quad \sup_{0 \leq s \leq \tau} \left| \frac{E_j(s)}{n} - e_j(s) \right| \xrightarrow{P} 0,$$

*and the processes  $\tilde{A}_j = E(A_j | (T_i, \delta_i, z_i); i = 1, \dots, n)$  converge in probability to  $D_j$ , w. p. 1,  $\forall j$ ,*

$$\tilde{A}_j(t) \xrightarrow{P} D_j(t) = \int_0^t \frac{e_j(s)}{r_j(s)} ds, \quad \text{w.p. 1, } \forall j.$$

**PROOF** First, let us see that using the approximation  $\psi(x) = \log(x) + O(x^{-1})$ , for  $n \gg 0$  and small  $c_j$ -s, the mean of stochastically continuous part of  $\tilde{A}_j(t)$  is

$O_p(1/n)$  while the second part

$$\sum_{T_i \leq t, \delta_i=1} p_{ij}(T_i) \xi_{ij}(T_i) = \int_0^t \sum_{i=1}^n \frac{(x_{i,j}/R_j(s)) c_j(s) \alpha_j^0(s)}{\sum_{k=1}^p (x_{i,k}/R_k(s)) c_k(s) \alpha_k^0(s)} \frac{dN_i(s)}{R_j(s)} + o_p\left(\frac{1}{R_j^3(t)}\right).$$

It is sufficient to show that  $\tilde{A}_j^*(t) = \int_0^t \sum_{i=1}^n V_{ij}(s) dN_i(s)/R_j(s)$  converges in probability to desired  $D_j$ . We will start with Doob-Meyer decomposition of the  $N_i$  processes that gives

$$\tilde{A}_j^*(t) = \int_0^t \sum_{i=1}^n V_{ij}(s) \frac{dN_i(s)}{R_j(s)} = \int_0^t \sum_{i=1}^n \frac{V_{ij}(s)}{R_j(s)} Y_i(s) z_i^\top \alpha(s) ds + \int_0^t \sum_{i=1}^n V_{ij}(s) \frac{dM_i(s)}{R_j(s)}. \quad (3.28)$$

The latter is a zero-mean martingale with a covariate process equal to

$$\int_0^t \sum_{i=1}^n \left[ \frac{V_{ij}(s)}{R_j(s)} \right]^2 Y_i(s) z_i^\top \alpha(s) ds \leq \int_0^t \sum_{i=1}^n \frac{1}{R_j^2(s)} Y_i(s) z_i^\top \alpha(s) ds = o_p(n^{-1})$$

and therefore it converges to 0 in probability, w.p. 1. Now we only need to show that the first term in (3.28) converges in probability to  $\int e_j(s)/r_j(s) ds$ , with probability 1. First, the existence of a limit  $e_j(s) = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n V_{ij}(s) Y_i(s) z_i^\top \alpha(s)$ , uniformly in  $[0, \tau]$ , is ensured by uniform law of large number as long as  $E z^2 < \infty$ , as well as the existence of an uniform limit  $r_j(s) = \lim_{n \rightarrow \infty} n^{-1} R_j(s)$ , see Remark 1 in Section 2.2. The proof is finished by application of the Slutsky theorem.

Q.E.D.

Similarly as in NPML, the limiting function of  $E_j/n$  is equal to

$$e_j(t) = E [V_j(s) Y(s) z^\top a(s)] = E \left[ e^{-z^\top A(s)} z^\top \alpha(s) V_j(s) \right] \bar{G}(s). \quad (3.29)$$

**Remark 8** It is interesting to assess how the estimators for  $A_j$ -s behave when the factors  $c_j$ -s are sent to their extremes. When all  $c_j$ -s go to infinity, the estimators land in prior guesses  $A_j^0$ . On the other hand, for  $c_j \rightarrow 0, \forall j$ , we expect to get the noninformative case (as for a noninformative prior). However, it turns out that when all  $c_j$ -s go to 0 at the same speed, the estimators become close to

$$\tilde{A}_j(t) \approx \int_0^t \sum_{i=1}^n \frac{\alpha_j^0(s) x_{i,j}/R_j(s)}{\sum_{k=1}^p \alpha_k^0(s) x_{i,k}/R_k(s)} \frac{dN_i(s)}{R_j(s)},$$

still carrying along the prior guesses  $\alpha_1^0, \dots, \alpha_p^0$ .

With growing sample size the estimators converge to

$$\tilde{A}_j(t) \xrightarrow{P} \int_0^t \sum_{i=1}^n \mathbb{E} \left[ \frac{\alpha_j^0(s) x_j / R_j(s)}{\sum_{k=1}^p \alpha_k^0(s) x_k / R_k(s)} \frac{n Y z^\top}{R_j(s)} \right] \alpha(s) ds, \text{ w.p. } 1,$$

and for  $\alpha_1^0 = \dots = \alpha_p^0$ ,

$$\tilde{A}_j(t) \xrightarrow{P} \int_0^t \sum_{i=1}^n \mathbb{E} \left[ \frac{x_j / R_j(s)}{\sum_{k=1}^p x_k / R_k(s)} \frac{n Y z^\top}{R_j(s)} \right] \alpha(s) ds, \text{ w.p. } 1,$$

gaining "average" weights. The estimators remain, however, inconsistent.

### 3.6 Average consistency

In the previous section we proved that the proposed Bayesian estimator  $\tilde{A}_j$  of the cumulative regression functions  $A_j$  is for general  $p$  inconsistent and it converges in probability to  $D_j \neq A_j$ . There is, however, noticeable, that the estimators have the same feature as we observed in the NPML case. We call it the average consistency and it means that when we look at the average intensity, i.e. the intensity of a randomly picked subject

$$\sum_{j=1}^p r_j(s) \alpha_j(s), \quad s \in [0, \tau],$$

it remains the same when we plug  $d_j(s) = e_j(s)/r_j(s)$  into the expression instead of  $\alpha_j$ .

**Corollary 2 (Melda corollary II)** *Let us denote  $d_j(s) = e_j(s)/r_j(s)$ . Suppose, that the assumptions of Theorem 6 are fulfilled. Then*

$$\sum_{j=1}^p r_j(s) d_j(s) = \sum_{j=1}^p r_j(s) \alpha_j(s).$$

**PROOF** We proceed identically as in proof of Corollary 1. Since the derivative of the limiting processes  $D_j$  in Theorem 6 equals  $d_j(s) = e_j(s)/r_j(s)$ , we have that the average intensity is

$$\begin{aligned} \sum_{j=1}^p r_j(s) d_j(s) &= \sum_{j=1}^p e_j(s) = \sum_{j=1}^p \mathbb{E} \left[ V_j(s) e^{-z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s) \\ &= \mathbb{E} \left[ e^{-z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s) = \sum_{j=1}^p \mathbb{E} \left[ e^{-z^\top A(s)} z_j \right] \bar{G}(s) \alpha_j(s) \end{aligned}$$

and again  $\sum_{j=1}^p r_j(s) d_j(s) = \sum_{j=1}^p r_j(s) \alpha_j(s)$ . We used the expressions for  $e_j$  in (3.29) and for  $r_j$  in (2.17) in the derivation of the result above.

Q.E.D.

### 3.7 Neutral to the right processes

As we already mentioned in the beginning of this chapter, Beta processes belong to a larger group of priors called subordinators or independent increment processes (NII). This general class of priors was widely used especially by Kim and Lee in [27], [29], [30] and [28]. They showed that if the cumulative hazard rate in homogeneous model or cumulative baseline hazard rate in Cox model is a priori distributed as a NII process, then a-posteriori it is again a NII process. Furthermore, the posterior expectation tends to the real value of the estimated characteristic and the posterior distribution centred around the NPML estimate correspond to the distribution of the NPML estimate (Bernstein-von Mises theorems). The aim of this section is to guess how a Bayesian estimator based on a general NII process prior performs under the monotone Aalen model.

Let us recall, that a NII process is induced by a so called neutral to the right process that in our situation serves (latently) as a prior process for the distribution function. Let the  $i$ -th individual possessing the time-independent covariate vector  $z_i = (x_{i,1}, \dots, x_{i,p})^\top$  be ruled by distribution function  $F_i$  and the corresponding cumulative hazard function is  $H_i$ . Then, as already advertised, the following equation is true if the  $i$ th individual is ruled by monotone Aalen model with cumulative hazard rate  $H_i(t) = \sum_{j=1}^p x_{i,j} A_j(t)$ :

$$1 - F_i(t) = \prod_{j=1}^p \{1 - dG_j(t)\}^{x_{i,j}}.$$

By  $G_j$ -s we denote the distribution functions corresponding to cumulative regression functions  $A_j$ -s. Using neutral to the right processes in the monotone Aalen model means that we assume that  $G_1, \dots, G_p$  are a priori distributed as a set of neutral to the right processes. Let these prior processes  $G_j$ -s be such that the NII processes  $A_j$  induced by  $G_j$  have Lévy measure of the following form

$$\nu_j(dt, du) = f_t^j(u) dt du, \quad t \in [0, \tau], \quad u \in [0, 1], \quad j = 1, \dots, p,$$

where  $f_t^j$  are such that  $\lim_{t \rightarrow \infty} \int_0^t \int_0^1 u f_s^j(u) ds du = \infty$ .

Utilizing the knowledge we gained in Section 3.2 and the theory on NII processes in [29] and [30] we arrive at the posterior distribution of  $A_j$  being again a NII process with the Lévy measure equal to

$$\nu_j(dt, du | \text{Data}) = (1 - u)^{R_j(t)} f_t^j(u) dt du + \sum_{T_i \in \text{death times}} dH_{ij}(u) \delta_{T_i}(dt) p_j(T_i)$$

where  $H_{ij}$  is a distribution function on  $[0, 1]$  with density

$$h_{ij}(u) = \frac{1}{k_{ij}} (1 - (1 - u)^{x_{i,j}}) (1 - u)^{R_j(T_i) - x_{i,j}} f_{T_i}^j(u),$$

$k_{ij}$  is the integration factor (and also a function of  $x_{i,j}$ ,  $R_j$  and  $T_i$ ), and  $p_j(t_i)$  is the probability that a-posteriori the process  $A_j$  has a jump at  $T_i$ ,

$$p_j(t_i) = \frac{k_{ij}}{\sum_{l=1}^p k_{il}}.$$

We introduce this result without proof but it is a corollary of slightly adjusted version of Lemma 5.1 in [30] combined with generalization of Lemma 2 from Beta processes to NII processes. The Bayes estimator of  $A_j$  is then equal to the posterior mean

$$\mathbb{E}\{A_j(t) | \text{Data}\} = \int_0^t \int_0^1 u(1 - u)^{R_j(s)} f_s^j(u) du ds + \sum_{T_i \leq t, \delta_i=1} p_j(T_i) \xi_{ij}(T_i)$$

where

$$\xi_{ij}(T_i) = \int_0^1 u h_{ij}(u) du.$$

The mean of the stochastically continuous part vanishes with growing sample size (under some assumptions). To understand how the other part of the estimator behaves asymptotically we follow the Kim and Lee's approach, [29], [30] and [28], and suppose that the function  $f_t^j$  in the prior distribution is of the form

$$f_t^j(u) = \frac{1}{u} g_t^j(u) \lambda^j(t),$$

where  $\int_0^1 g_t^j(u) du = 1$  for all  $t$  and  $\lambda^j(t)$  is bounded and positive on  $[0, \tau]$ . Under certain conditions on  $g_s^j$  functions parallel to the conditions in Kim and Lee's work and using similar laborious techniques as in proof of Theorem 4.1 in [29] we expect to arrive to an asymptotic process of following form

$$\mathbb{E}\{A_j(t) | \text{Data}\} \xrightarrow{P} \int_0^t \frac{1}{r_j} \mathbb{E}\{p_j(s) Y(s) z^\top \alpha(s)\} ds, \quad \text{w.p. 1.}$$

And particularly for the Beta process prior,  $p_j(s) = V_j(s)$ . Consequently, the Bayesian machine using a general NII process as prior still produces inconsistent estimators.

### 3.8 An example with 3 exponentially distributed covariates II

In this section we revisit the example from Section 2.5 with three exponentially distributed covariates and provide a similar analysis as in the NPML case. Afterwards, we will devote ourselves to a study of features of the approximations (3.25) and (3.27) in Section 3.3 in comparison to the exact forms in (3.23) and (3.24). Furthermore, we run the MCMC algorithm proposed at the end of Section 3.3 and compare the obtained estimators with the aforementioned exact estimators.

In the example in Section 2.5 we supposed that with every individual we collected the covariate vector  $(x_{i,1}, x_{i,2}, x_{i,3}) \sim \text{Exp}(\lambda_1, \lambda_2, \lambda_3)$ . Asymptotically, the Bayesian estimators  $\tilde{A}_j$  of the integrated regression functions equal to  $D_j = \int e_j(s)/r_j(s)ds$ , where

$$e_j(s) = \mathbb{E} \left[ V_j e^{-z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s)$$

and

$$r_j(s) = \mathbb{E} \left[ e^{z^\top A(s)} z^\top \alpha(s) \right] \bar{G}(s).$$

The expectation is with respect to the covariate distribution. We will derive the form of  $D_j$  only for  $j = 1$  case, as the results are analogical for  $j = 2, 3$ . We have already found the expression for  $r_1$  in (2.20) so let us focus on  $e_1$ . The asymptotic weight process  $V_1$  is equal to

$$V_1(s) = \frac{x_1/r_1(s)c_1(s)\alpha_1^0(s)}{\sum_{k=1}^3 x_k/r_k(s)c_k(s)\alpha_k^0(s)},$$

and the function  $e_1$  is a triple integral

$$\begin{aligned} & \int_0^\infty \int_0^\infty \int_0^\infty V_1(s) \exp\{-x_1(A_1(s) + \lambda_1) - x_2(A_2(s) + \lambda_2) - x_3(A_3(s) + \lambda_3)\} \\ & \quad \times (x_1\alpha_1(s) + x_2\alpha_2(s) + x_3\alpha_3(s))\lambda_1\lambda_2\lambda_3 dx_1 dx_2 dx_3 \bar{G}(s). \end{aligned}$$

This is an integral of following type  $\int x/(x+a)e^{-xb}dx$  and there is no explicit form of corresponding primitive function. Hence, the resulting function  $D_1$  is an



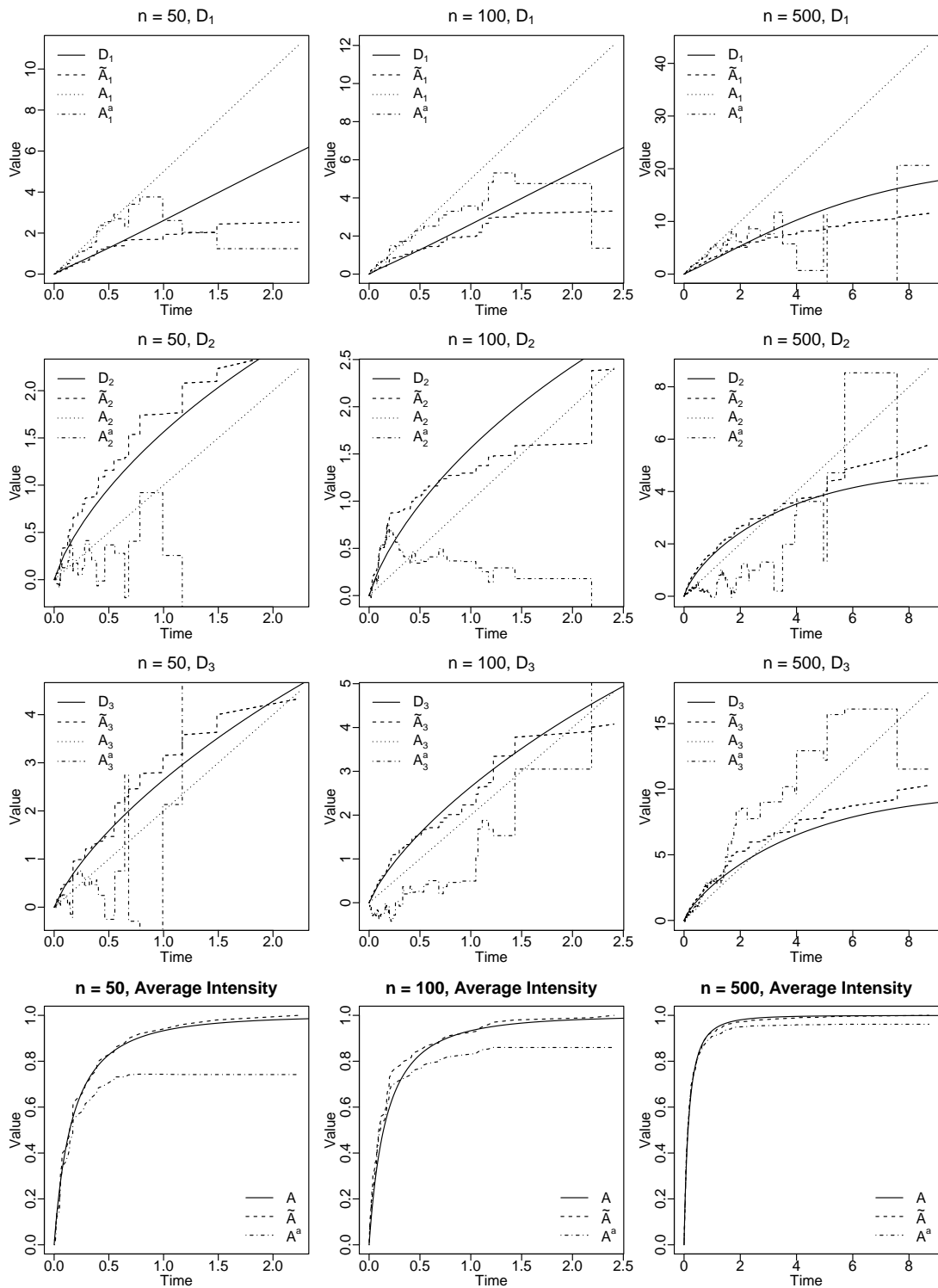


Figure 3.1: *Revisited example from Figure 2.3:* Cumulative regression functions in a simulated Aalen model with hazard rate  $h_i(t) = 5x_{i,1} + x_{i,2} + 2x_{i,3}$  and  $z_i$  i.i.d. and exponentially distributed. The number of observations is  $n = 50$  in left column,  $n = 100$  in the middle and  $n = 500$  in right column. In *dotted lines* are plotted the real cumulative parameter processes, the *dashed lines* are the exact Bayesian estimators  $\tilde{A}_j$  and the *solid lines* are the asymptotic processes  $D_j$  of the Bayesian estimators from Theorem 8. The Aalen estimators  $A_j^a$  are included in *dash-dotted lines*. The bottom row shows average intensities.

integral of ratio of the expression with the triple integral from above in numerator and the function  $r_1$  from (2.20) in denominator.

Again, we present the results for the same datasets generated from the model  $h_i(t) = 5x_{i,1} + x_{i,2} + 2x_{i,3}$ , where covariates were independent and exponentially distributed,  $x_{i,1} \sim \text{Exp}(1)$ ,  $x_{i,2} \sim \text{Exp}(2)$  and  $x_{i,3} \sim \text{Exp}(3)$  as in Section 2.5. We assumed the same and the most simple prior Beta process for every  $A_j$ , with  $c_j(t) \equiv c_0 = 0.1$  and  $A_j^0(t) = \alpha_j^0 t \equiv t$ ,  $j = 1, \dots, 3$ .

The results are presented in Figure 3.1 in similar layout as in the NPML case in Figure 2.3. The true cumulative regression functions  $A_j$  are plotted in dotted lines, the Aalen estimators  $A_j^a$  are in dash-dotted lines and the exact Bayesian estimator  $\tilde{A}_j$  from 3.23 are in the dashed lines. The exact estimators were calculated on a thin grid in  $[0, \tau]$  with 1000 equidistant points. The rough limiting functions  $D_j$  calculated by approximating the triple integral in  $d_j$  by a triple sum are included in solid lines. There is a clear difference between the limiting functions  $A_j$  and  $D_j$  and the affinity of the estimators  $\tilde{A}_j$  and  $\hat{A}_j^a$  towards their respective asymptotic counterparts is apparent from the graphs. The average intensities are plotted in the graphs in the bottom row. The average intensities calculated from the true  $A_j$ -s are in solid lines, while the ones derived from the Bayes estimator and standard Aalen estimators are plotted in dashed and dash-dotted lines, respectively.

The other task of this section is to explore the performance of the proposed approximations in (3.25) and (3.27) for the exact estimators of the posterior expectation and posterior variance in (3.23) and (3.24). We also want to compare them to estimators based on the output of the MCMC algorithm stated at the end of Section 3.3. The approximated estimators and their variances are concentrated only at the failure times points and are the easiest to obtain, therefore they are our candidates for practical usage (in the case, that we would like to use these inconsistent methods). The exact estimator sums the expectation from the Beta process in between the failure times and the exact expectation of the beta distributed jump sizes at fixed discontinuities. Let us denote the ordered set of the observed lifetimes by  $T_{(1)}, \dots, T_{(n)}$ . Under the chosen prior the expectation of the stochastically continuous part simplifies down to

$$\int_0^t \frac{c_j(s) dA_j^0(s)}{c_j(s) + R_j(s)} = \sum_{T_{(i)} \leq t} \left[ \frac{c_0 a_0}{c_0 + R_j(T_{(i)})} (T_{(i)} - T_{(i-1)}) \right. \\ \left. + \frac{c_0 a_0}{c_0 + R_j(T_{(i)})} (t - T_{(i)}) I\{T_{(i)} = \sup_{k=1, \dots, n} \{T_{(k)} : T_{(k)} \leq t\}\} \right],$$

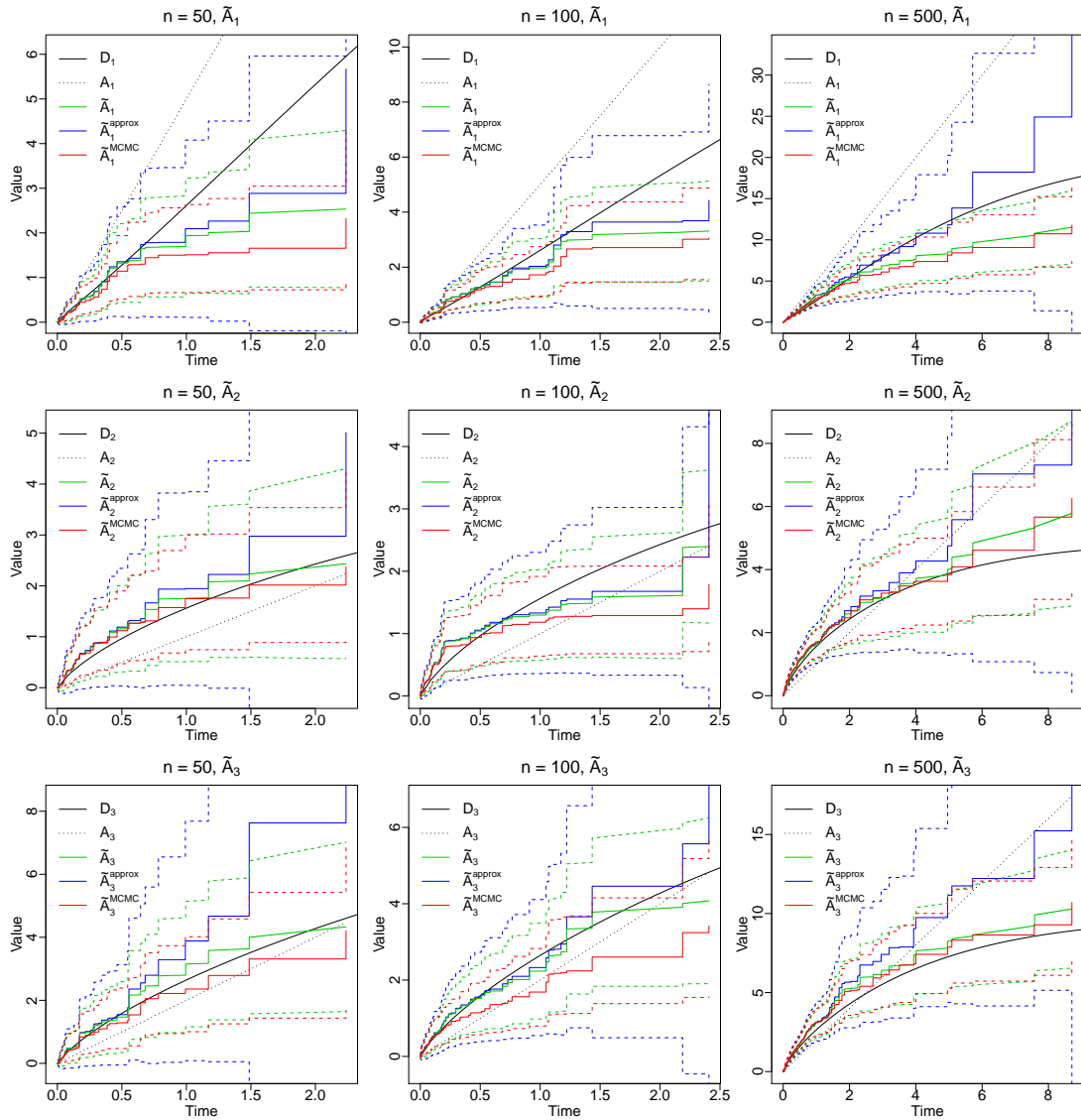


Figure 3.2: *Revisited example from Figure 2.3:* Cumulative regression functions in a simulated monotone Aalen model with hazard rate  $h_i(t) = 5x_{i,1} + x_{i,2} + 2x_{i,3}$  and  $z_i$  i.i.d. and exponentially distributed. The number of observations is  $n = 50$  in left column,  $n = 100$  in the middle and  $n = 500$  in right column. In *black dotted lines* the real cumulative parameter processes are plotted, the *solid green lines* are the exact Bayesian estimators  $\tilde{A}_j$  and the *solid blue lines* are the approximated versions  $\tilde{A}_j^{approx}$  of the estimators from (3.25). The MCMC based estimators  $\tilde{A}_j^{MCMC}$  are included in *solid red lines*. The 95% pointwise credibility intervals are included in dashed lines and respective colors.

where we assumed  $T_{(0)} = 0$ . As already mentioned, these estimators were calculated on a thin grid on  $[0, \tau]$ . Additionally, in similar fashion we obtained the exact variances from (3.24). The pointwise credibility bands using either exact or asymptotic posterior mean and variance are based on a normal approximation, hence they are of following form

$$\text{mean}(A_j(t) | \text{data}) \pm u_\alpha \sqrt{\text{var}(A_j(t) | \text{data})}, \quad t \in [0, \tau],$$

where  $u_\alpha$  is a  $\alpha$ th quantile of the standard normal distribution. The finite sample distribution of the increments of Bayesian estimator is not normal and it is of rather complicated structure. From martingale theory we can, however, expect that the limiting distribution of  $\sqrt{n}(\tilde{A}_j(t) - D_j(t))$  is equivalent to a Gaussian martingale, hence the distribution of the process  $\tilde{A}_j$  at every time point will be closer and closer to a normal distribution with growing  $n$ .

Next, we calculated Bayesian estimators based on the Markov chain generated using the algorithm in Section 3.3. We ran 1500 repetitions and discarded first 500 members of the chains. The estimators were created in pointwise fashion on the same grid. We always took the 50% quantile of the values of the members of the chain at the particular time point of the thin net to obtain the Bayesian estimator. The posterior median instead of mean was used to offer a comparison to the estimators based on the exact and approximated posterior expectation. The 95% pointwise credibility intervals were created by taking the 2.5% and 97.5% quantile at every point of the net. The results obtained from the same datasets as used before are presented in Figure 3.2.

It is seen, that the three variants of estimators are quite similar, in particular at the beginning of the observation windows, where lots of data are available. The 95% pointwise credibility intervals are the thinnest when calculated from the exact formulas and get a lot wider when the approximations are used. The MCMC credibility bands give very similar results to the exact credibility bands.

### 3.9 Discussion

The Bayesian approach to the monotone Aalen model using the Beta process as a prior process for  $A_j$ -s produces a remarkably similar outcome to the one gained via the NPML estimation in Chapter 2.

When looking closer at the set of the NPML and Bayesian estimators  $\hat{A}_1, \dots, \hat{A}_p$  and  $\tilde{A}_1, \dots, \tilde{A}_p$ , it is seen that they both are of the following form

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

where  $\sum_{j=1}^p V_{ij}(s) = 1$  and they both converge to a function of the form

$$\int_0^t \frac{f_j(s)}{r_j(s)} ds,$$

with  $f_j(s) = E\{V_{ij}(s)Y(s)z^\top \alpha(s)\}$ . The condition  $\sum_{j=1}^p V_{ij}(s) = 1$  is sufficient for the estimators to have the average consistency feature. The Bayesian estimator can also be viewed as a smoother version of the NPML with "weights" that are not strictly 0 or 1.

Let us suppose that we have a class of estimators for the cumulative regression functions  $A_j, j = 1, \dots, p$ , denoted e.g.  $\mathcal{C}$ , such that a set of the estimators for Aalen model  $C_j^*$  are members of  $\mathcal{C}$  if

$$C_j^*(t) = \int_0^t \sum_{i=1}^n \frac{V_{ij}(s)}{R_j(s)} dN_i(s), \quad j = 1, \dots, p, \quad (3.30)$$

for predictable processes  $V_{ij}, j = 1, \dots, p$ , such that the condition  $\sum_{j=1}^p V_{ij}(s) = 1$  is fulfilled. Then

- the estimators  $C_j^*$  converge to  $C_j = \int_0^{(\cdot)} f_j(s)/r_j(s)ds$  with  $f_j(s) = E\{V_{ij}(s)Y(s)z^\top \alpha(s)\}$ , for all  $j = 1, \dots, p$ , and
- the average consistency effect is present with  $\sum_{j=1}^p r_j(s)dC_j(s) = \sum_{j=1}^p r_j(s)dA_j(s)$ .

In general, the limiting function  $C_j \neq A_j$ . Now the question is, under which conditions  $C_j = A_j, j = 1, \dots, p$ , i.e. the set of estimators of class  $\mathcal{C}$  are consistent, and if there are any consistent estimators contained in  $\mathcal{C}$ . Clearly, if

$$\sum_{i=1}^n \frac{V_{ij}(s)}{R_j(s)} Y_i(s) z_i^\top \xrightarrow{P} \underbrace{(0, \dots, 0, 1, 0, \dots, 0)}_{\substack{1 \text{ is at } j\text{-th position, } 0 \text{ elsewhere}}},$$

then the estimator  $C_j^*$  given in (3.30) is consistent. If we write  $V = (V_{ij})_{i,j=1}^{n,p}$  the  $n \times p$  matrix then the condition above can be reformulated as  $\text{diag}\{\mathbf{1}_n^\top Z\}^{-1} V^\top Z \xrightarrow{P} I_p$ , where  $I_p$  is a  $p \times p$  matrix with ones on the diagonal and zeros elsewhere. If we can find a matrix  $V$  which would oblige to the condition  $\text{diag}\{\mathbf{1}_n^\top Z\}^{-1} V^\top Z \xrightarrow{P} I_p$ , we would gain a consistent estimator. This can

be seen as a lead how to construct a set of consistent estimators for  $A_1, \dots, A_p$ , which would be of similar structure as the NPML and Bayesian estimators.

The condition  $\sum_{j=1}^p V_{ij}(s) = 1$  is not necessary for an estimator to be consistent. Let us have a look at the traditional least squares estimator of Aalen which is equal to

$$A^a(t) = \int_0^t (Z(s)^\top Z(s))^{-1} Z(s)^\top dN(s).$$

The rows of the  $p \times n$  matrix  $(Z(s)^\top Z(s))^{-1} Z(s)^\top$  define the estimators for each  $A_j$ . If we denote  $(Z(s)^\top Z(s))^{-1} Z(s)^\top = \Psi(s)^\top = (\psi_1(s), \dots, \psi_p(s))^\top$ , with the  $n$ -dimensional vectors  $\psi_j(s) = (\psi_{1j}(s), \dots, \psi_{nj}(s))^\top$ , then the estimators for the regression functions can be written out as follows

$$A_j^a(t) = \int_0^t \psi_j(s)^\top dN(s) = \int_0^t \sum_{i=1}^n \psi_{ij}(s) dN_i(s), \quad j = 1, \dots, p.$$

For  $A_j^a, j = 1, \dots, p$ , to be the members of the class  $\mathcal{C}$ , there must exist random processes  $V_{ij}$  such that  $\psi_{ij} = V_{ij}/R_j$  and the condition  $\sum_{j=1}^p \psi_{ij}(s) R_j(s) = 1, \forall i, \forall s$ , must hold. If we denote  $R(s) = (R_1(s), \dots, R_p(s))^\top$ , then we in fact require that  $\Psi(s)R(s) = \mathbf{1}_n$ , where  $\mathbf{1}_n = (1, \dots, 1)^\top$  with  $n$  components equal to 1. Notice, that according to the notation in Section 1.3,  $R(s) = Z(s)^\top \mathbf{1}_n$  and therefore

$$\Psi(s)R(s) = Z(s)(Z(s)^\top Z(s))^{-1} Z(s)^\top \mathbf{1}_n.$$

We can see, that the expression above resembles a hat matrix from an ordinary least squares estimation multiplied by  $\mathbf{1}_n$ . Thus it is seen, that  $\Psi(s)R(s)$  equals to  $\mathbf{1}_n$  if the intercept is included in the model and only for  $s \leq \min_{i=1, \dots, n} \{T_i\}$  or if we allow multiple events on every subject, hence leaving the subject in risk-group even after encountering a failure. In general, however, the Aalen least squares estimators  $A_j^a, j = 1, \dots, p$ , are not members of the class  $\mathcal{C}$ . Still, given that the Aalen estimators are consistent, they possess the average consistency feature. For the finite sample, we have that the matrix  $Z(s)(Z(s)^\top Z(s))^{-1} Z(s)^\top$  has zeros in all rows and columns which correspond to the entries from the individuals with observed time smaller than  $s$ . If the intercept is present in the model, the multiplication  $Z(s)(Z(s)^\top Z(s))^{-1} Z(s)^\top$  by  $\mathbf{1}_n$  gives the vector with zeros again at the positions corresponding to the smaller observed times and ones elsewhere. Hence,

$$\frac{1}{n} \sum_{j=1}^p R_j^\top(s) \Psi^\top(s) dN(s) = \frac{1}{n} \sum_{i=1}^n dN_i(s),$$

matching the same result as in Section 2.4. Using the similar derivation we get the same conclusion for the weighted least squares estimator of Huffer and McKeague.

Another option would be to generalize our NPML and Bayesian estimators by introducing weights, i.e. we would have a weighted NPML estimator or a weighted Bayesian estimator

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

To achieve the consistency the weights would have to be asymptotically equal to

$$w_{ij}(s) \approx \alpha_j(s) \frac{r_j(s)}{q_j(s)}$$

for NPML estimator or with  $e_j$  instead of  $q_j$  for Bayesian estimator. It is not entirely clear how to find weights  $w_{ij}$  which would fit this condition, though.

## Chapter 4

# Bayesian analysis with correlated piecewise constant prior

In previous chapters we explored the possibilities of estimating the cumulative regression functions in the monotone Aalen model by applying two solid approaches: the nonparametric maximum likelihood method and the Bayesian analysis with Beta processes. We proved that the derived estimators were inconsistent for general  $p > 1$ . This result left us empty handed and in need for a valid method if we want to utilize the monotone Aalen model in practice.

There are, however, other options of using the Bayesian approach to solve the statistical problems. Arjas and Gasbarra, [6], suggested Bayesian inference of homogeneous lifetime data using a simple piecewise constant process with dependent increments for prior for hazard function. The hazard function was a priori a random function which was constant on some intervals and the level of the function in an interval was dependent on the value in the previous one. Number of the intervals and variation of the function from one interval to another was controlled by four hyperparameters. This setting included 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. Arjas and Gasbarra did not discuss the consistency issue neither they conducted a simulation study to explore how this prior performs in estimation of the hazard rate. Their concept is, however, natural and elegant and offers great variability within the choice of hyperparameters.



Hence, the main objective of this chapter is to conduct Bayesian modelling in monotone Aalen model based on the Arjas and Gasbarra prior. We assume that the regression functions are continuous, i.e.  $\alpha_j$ -s exist, and the piecewise constant process is applied as a prior for the regression functions. The advantage of this Bayesian approach as opposed to the analysis with Beta processes is that the method estimates the regression functions directly.

On the following pages such type of modelling is demonstrated. In the next section the process used as prior to regression function is explained. In Section 4.2 the posterior distribution under monotone Aalen model is derived and followed by the explanation of the MCMC algorithm used for estimation. Section 4.3 is devoted to simulation study conducted to explore the performance of the method. We discuss the obtained outcome in Section 4.4.

## 4.1 Prior distribution

Based on [6], we model the unknown regression functions  $\alpha_1(t), \dots, \alpha_p(t)$  in observed time window  $[0, \tau]$ , where  $\tau = \max\{T_i\}$ , as a correlated piecewise constant function. The values of regression function  $\alpha_j$  are assumed to be constant within  $m^{(j)} + 1$  intervals which emerge from dividing the time window  $[0, \tau]$  by  $m^{(j)}$  jump times  $W_1^{(j)}, \dots, W_{m^{(j)}}^{(j)}$ . The value of regression function  $\alpha_j$  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  $\alpha_j$  can be expressed as a simple jump process

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

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

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

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

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

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  $E X = \frac{a}{b}$ ,  $\text{Var } X = \frac{a}{b^2}$ . The parameters of the prior distribution for  $\lambda_k$ -s 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  $\lambda_{k-1}^{(j)}$ . 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 (thus we incorporate a martingale structure into the model)

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

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

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

In case the hyperparameter  $a$  is small, the regression function  $\alpha_j$  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, [6], a homogeneous Poisson process was utilized as the prior process for jump times that split the observation window into disjoint intervals. Here, as it will be clear from derivation in Section 4.2, the computational evaluation of the posterior distribution of  $\lambda_k$ -s 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  $(W_1^{(j)}, \dots, W_{m^{(j)}}^{(j)})$  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

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

The number of the jumps in 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  $j$ th regression functions  $\alpha_j$  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 (1.6) is proportional to the following formula:

$$\begin{aligned} L((T_i, z_i, \delta_i), i = 1, \dots, n) &\propto \prod_{i=1}^n h_i(T_i)^{\delta_i} \exp \left\{ - \int_0^{T_i} h_i(t) dt \right\} \\ &= \prod_{i=1}^n \left[ \sum_{j=1}^p \alpha_j(T_i) x_{i,j} \right]^{\delta_i} \exp \left\{ - \int_0^{T_i} \sum_{j=1}^p \alpha_j(t) x_{i,j} dt \right\} \\ &= \prod_{i=1}^n \left[ \sum_{j=1}^p \sum_{k=1}^{m^{(j)}+1} I_{\{W_{k-1}^{(j)} \leq T_i < W_k^{(j)}\}} \lambda_k^{(j)} x_{i,j} \right]^{\delta_i} \\ &\quad \times \exp \left\{ - \sum_{j=1}^p \sum_{k=1}^{m^{(j)}+1} I_{\{T_i \geq W_{k-1}^{(j)}\}} \lambda_k^{(j)} x_{i,j} \left( \min\{W_k^{(j)}, T_i\} - W_{k-1}^{(j)} \right) \right\}. \end{aligned}$$

In next section the derivation of the posterior distribution is explained.

## 4.2 The posterior distribution and the Gibbs sampler

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

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

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  $\alpha_j$  we proceed sequentially by updating the pairs  $(\lambda_k^{(j)}, W_k^{(j)})$ ,  $k = 1, \dots, m^{(j)}$  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 the number of added intervals by  $\eta$ , then altogether we have  $2(m^{(j)} + \eta) + 1$  steps within every iteration for  $\alpha_j$ . 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  $H^{(1)}, \dots, H^{(p)}$  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 on next pages with overall summary of the algorithm at the end of the section.

### 4.2.1 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 every regres-

sion function, i.e.  $\alpha_j$ , 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  $\alpha_j$  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  $\alpha_j$  and all the characteristics of all other  $\alpha_l$ -s. 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  $\lambda_k^{(j)}$ .

Before we move to the derivation of the posterior distribution, let us remind that

$$h_i(t) = \sum_{j=1}^p x_{i,j} \alpha_j(t) = \sum_{j=1}^p \sum_{k=1}^{m^{(j)}+1} x_{i,j} I_{\{W_{k-1}^{(j)} \leq t < W_k^{(j)}\}} \lambda_k^{(j)}, \quad i = 1, \dots, n.$$

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

$$\begin{aligned} p(\lambda_k^{(j)} | \lambda_1^{(j)}, \dots, \lambda_{k-1}^{(j)}, \lambda_{k+1}^{(j)}, \dots, \lambda_{m^{(j)}+1}^{(j)}, W_1^{(j)}, \dots, W_{m^{(j)}}^{(j)}, \{h_i\}_{i=1}^n, a_0, b_0, a, c, d, \text{data}) \\ = p(\lambda_k^{(j)} | \lambda_{k-1}^{(j)}, \lambda_{k+1}^{(j)}, W_{k-1}^{(j)}, W_k^{(j)}, \{(T_i, \delta_i, z_i, h_{i(-j)}) : T_i \geq W_{k-1}^{(j)}\}, a_0, b_0, a) \\ \propto p(\lambda_k^{(j)}, \lambda_{k+1}^{(j)}, \{(T_i, \delta_i, z_i, h_{i(-j)}) : T_i \geq W_{k-1}^{(j)}\} | \lambda_{k-1}^{(j)}, W_{k-1}^{(j)}, W_k^{(j)}, a_0, b_0, a) \\ = \gamma(\lambda_k^{(j)}; a, a/\lambda_{k-1}^{(j)}) \gamma(\lambda_{k+1}^{(j)}; a, a/\lambda_k^{(j)}) \\ \times \prod_{i: T_i \in I_k^{(j)}} \left( \lambda_k^{(j)} x_{i,j} + h_{i(-j)}(T_i) \right)^{\delta_i} \\ \times \exp \left\{ - \sum_{i: T_i \in I_k^{(j)}} \lambda_k^{(j)} x_{i,j} (T_i - W_{k-1}^{(j)}) - \sum_{i: T_i \geq W_k^{(j)}} \lambda_k^{(j)} x_{i,j} (W_k^{(j)} - W_{k-1}^{(j)}) \right\} \end{aligned} \quad (4.1)$$

where we denoted by  $h_{i(-j)}(t) = h_i(t) - \alpha_j(t)x_{i,j}$  the complement of the hazard function for  $i$ th 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 (4.1) we get a sum of functions  $\sum_{r=0}^R \beta_r f_r(\lambda_k^{(j)})$ , where  $R = \sum_{i: T_i \in I_k^{(j)}} \delta_i$ . The terms in the sum are

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

and

$$\beta_r = \sum_{l_1=1}^R \sum_{l_2=l_1+1}^R \cdots \sum_{l_{R-r}=l_{R-r-1}+1}^R \left[ \prod_{\substack{i=1 \\ i \notin \{l_1, \dots, l_{R-r}\}}}^R x_{i,j}^* \right] h_{l_1(-j)}(T_{l_1}^*) \cdots h_{l_{R-r}(-j)}(T_{l_{R-r}}^*). \quad (4.2)$$

Here we used  $T_1^*, \dots, T_R^*$  as an auxiliary notation for the set of the failure times in the interval  $I_k^{(j)}$  corresponding to the subjects with  $j$ th covariates equal to  $x_{1,j}^*, \dots, x_{R,j}^*$ .

This case of distribution can be viewed as a mixture of distributions proportional to  $f_r$  weighted by factors  $\beta_r$ . 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  $\alpha_1$  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  $f_r$  (here put  $\sum_{s=0}^{-1} w(s) = 0$ ).

The simulation from the distribution proportional to function  $f_r$  is done similarly as in Arjas and Gasbarra's work in [6]. 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

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

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 the 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  $\xi$  to let the modes of both  $d_{r,\xi}$  and  $g_{r,\xi}$  equal. This is guaranteed when  $\xi$  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  $\lambda_{m^{(j)+1}^{(j)}}$  does not influence any subsequent level of the hazard function and therefore the posterior distribution for  $\lambda_{m^{(j)+1}^{(j)}}$  simplifies to a mixture of gamma distributions, symbolically written as

$$\sum_{r=0}^R \beta_r \gamma\left(a + r, a/\lambda_{m^{(j)}}^{(j)} + \sum_{i:T_i \in I_{m^{(j)+1}^{(j)}}} x_{i,j}(T_i - W_{m^{(j)}}^{(j)})\right) \quad (4.3)$$

where  $\beta_r$  is as in (4.2) and again  $R$  being total of observed deaths in  $I_{m^{(j)+1}^{(j)}}$ .

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  $\beta_r, r = 0, \dots, R$  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,  $r = 0, \dots, R$ , 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  $\beta_r$ -s is in need. One of the options is for every  $r$  such that it produces larger number of combinations than a fixed number  $L$  (i.e. if  $\binom{R}{r} > L$ ) then instead of using all the combinations in the evaluation of  $\beta_r$  we would randomly choose only  $L$  combinations. 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. The various choices of  $L$  and its impact on the posterior distribution is discussed in Section 4.3 along with the simulations.

### 4.2.2 Posterior distribution of jump times

The posterior distribution for the particular jump time  $W_k^{(j)}$  in the regression function  $\alpha_j$  is again determined only by the parts of the likelihood and prior information that are affected by  $W_k^{(j)}$ . The posterior probability of jump time  $W_k^{(j)}$  can be written as

$$\begin{aligned}
& p(W_k^{(j)} | W_1^{(j)}, \dots, W_{k-1}^{(j)}, W_{k+1}^{(j)}, \dots, W_{m^{(j)}}^{(j)}, \lambda_1^{(j)}, \dots, \lambda_{m^{(j)+1}^{(j)}}, \{h_i\}_{i=1}^n, \\
& \quad a_0, b_0, a, c, d, \text{data}) \\
& = p(W_k^{(j)} | W_{k-1}^{(j)}, W_{k+1}^{(j)}, \lambda_k^{(j)}, \lambda_{k+1}^{(j)}, \{(T_i, \delta_i, z_i, h_i) : T_i \geq W_{k-1}^{(j)}\}, c, d) \\
& \propto p(W_k^{(j)}, W_{k-1}^{(j)}, W_{k+1}^{(j)}, \lambda_k^{(j)}, \lambda_{k+1}^{(j)}, \{(T_i, \delta_i, z_i, h_i) : T_i \geq W_{k-1}^{(j)}\}, c, d) \\
& \propto \exp\left\{-\frac{d}{c}(1 - e^{-c\tau})\right\} d \exp\{-cW_k^{(j)}\} \prod_{i:T_i \in I_k^{(j)}} h_i(T_i)^{\delta_i} \prod_{l:T_l \in I_{k+1}^{(j)}} h_l(T_l)^{\delta_l} \\
& \quad \times \exp\left\{-\sum_{i:T_i \geq W_{k+1}^{(j)}} \left[\lambda_k^{(j)} x_{i,j}(W_k^{(j)} - W_{k-1}^{(j)}) + \lambda_{k+1}^{(j)} x_{i,j}(W_{k+1}^{(j)} - W_k^{(j)})\right] \right. \\
& \quad \quad - \sum_{i:T_i \in I_{k+1}^{(j)}} \left[\lambda_k^{(j)} x_{i,j}(W_k^{(j)} - W_{k-1}^{(j)}) + \lambda_{k+1}^{(j)} x_{i,j}(T_i - W_k^{(j)})\right] \\
& \quad \quad \left. - \sum_{i:T_i \in I_k^{(j)}} \lambda_k^{(j)} x_{i,j}(T_i - W_{k-1}^{(j)})\right\}. \tag{4.4}
\end{aligned}$$

The expression is in core similar to the result of Arjas and Gasbarra, [6]. In the examined interval the posterior distribution is between the observation times proportional to  $u \exp(vW_k^{(j)})$ . A new jump position can be generated from this piecewise continuous distribution for example by using rejection 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

$$\begin{aligned}
& \exp\left\{-\frac{d}{c}\right\} \prod_{i:T_i \in [W_{m^{(j)-1}^{(j)}}, \tau]} h_i(T_i)^{\delta_i} \exp\left\{-\sum_{i:T_i = \tau} \lambda_{m^{(j)}}^{(j)} x_{i,j}(\tau - W_{m^{(j)-1}^{(j)}})\right. \\
& \quad \left. - \sum_{i:T_i \in [W_{m^{(j)-1}^{(j)}}, \tau)} \lambda_{m^{(j)}}^{(j)} x_{i,j}(T_i - W_{m^{(j)-1}^{(j)}})\right\}.
\end{aligned}$$



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  $[W_{m^{(j)}}^{(j)}, W_{m^{(j)+1}^{(j)}}^{(j)})$  and  $[W_{m^{(j)+1}^{(j)}}^{(j)}, \tau)$  into the sets of the intervals. We set  $m^{(j)} \leftarrow m^{(j)} + 1$  and sample value  $\lambda_{m^{(j)+1}^{(j)}}^{(j)}$  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, [6].

Another option is to use the Metropolis-Hasting algorithm. Let us denote the conditional posterior distribution of  $W_k^{(j)}$  from (4.4) 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. [18]). 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. The disadvantage of the Metropolis-Hastings algorithm in comparison to the Gibbs sampler is the necessity to repeat the sampling from the proposal density until we reach the acceptance, what might be time consuming.

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

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

the  $u$ th member of the Markov chain  $\{H^{(j)}(u)\}_{u=0}^U$  generated in  $u$ th iteration of the Gibbs sampler. The chain  $\{H^{(j)}(u)\}_{u=0}^U$  corresponds to  $j$ th regression function  $\alpha_j, j = 1, \dots, p$ . The steps of the algorithm can be summarized as follows:

**Sampling Algorithm:**

- generate a starting trajectory  $H^{(j)}(0)$  for  $j = 1, \dots, p$  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  $u$ th iteration, where  $u \in \{1, \dots, U\}$ , do
  - for  $j = 1, \dots, p$  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 (4.1) (sampling from the mixture distribution),
    4. sample  $W_k^{(j)}(u)$  from posterior distribution in (4.4),  $k \leftarrow k + 1$
    5. repeat steps 3. and 4. until  $k = m^{(j)}(u)$ ,
    6. sample  $\lambda_{m^{(j)}(u)}^{(j)}(u)$  from posterior distribution in (4.1),
    7. sample  $W_{m^{(j)}(u)}^{(j)}(u)$  from posterior distribution in (4.4), 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 (4.3).

The problem of ergodicity of every complement  $H^{(j)}(u)$  of the resulting Markov chain is similar to the original Arjas nad Gasbarra's method as long as the other complements  $H^{(k)}(u), k \neq j$  are held fixed. If the birth-death Metropolis-Hastings algorithm is used for simulation of new intervals, the proposal density and the acceptance probability 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 [18].

**4.3 Simulations**

The posterior distribution of the method proposed in this chapter 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 rely on the aid of simulation techniques. The method was tested on 300 datasets sampled from a model  $h_i(t) = \alpha_1(t) + \alpha_2(t)x_{i,2} + \alpha_3(t)x_{i,3}$  of the hazard rate on interval  $[0, 1]$  with regression functions equal to

$$\begin{aligned}\alpha_2(t) &= \sin(\pi t) + 1.5 \\ \alpha_3(t) &= \exp(-3t) + 1\end{aligned}$$

and the baseline hazard rate  $\alpha_1(t)$  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  $x_{i,2}$  and  $x_{i,3}$ , 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

$$\text{PRIOR 1: } a_0 = 0.1, b_0 = 0.1, a = 0.5, c = 1, d = 25,$$

$$\text{PRIOR 2: } a_0 = 0.1, b_0 = 0.1, a = 0.2, c = 0.5, d = 35.$$

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.

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  $\approx 0.8$ . 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 on a thin grid 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. Alongside classic Aalen estimators were calculated with 95% pointwise confidence bands. These bands were created in pointwise fashion on a thin grid by taking 2.5% and 97.5% sample quantiles of the members of Markov chains without the burn-in.

|                | $n$ |         | $A_1$ Bayes | $A_1$ Aalen | $A_2$ Bayes | $A_2$ Aalen | $A_3$ Bayes | $A_3$ Aalen |
|----------------|-----|---------|-------------|-------------|-------------|-------------|-------------|-------------|
| <b>PRIOR 1</b> | 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       |
| <b>PRIOR 2</b> | 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       |

Table 4.1: 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.

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

$$BIAS(\hat{A}_j) = \int_0^{\tau^*} (\hat{A}_j(t) - A_j(t)) dt,$$

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 grid of 100 time points within the interval  $[0, \tau^*]$ . We chose the right end  $\tau^*$  so that the interval  $[0, \tau^*]$  represented the part of the whole interval  $[0, 1]$  where in all instances the Aalen estimators were calculated. The minimal value of  $\tau^*$  for all datasets was equal to 0.17. The

|                |        | $A_1$ |      |       | $A_2$ |      |       | $A_3$ |      |       |
|----------------|--------|-------|------|-------|-------|------|-------|-------|------|-------|
|                |        | Bayes |      | Aalen | Bayes |      | Aalen | Bayes |      | Aalen |
|                |        | 95%   | 99%  | 95%   | 95%   | 99%  | 95%   | 95%   | 99%  | 95%   |
| <b>PRIOR 1</b> | 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  |
| <b>PRIOR 2</b> | 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  |

Table 4.2: Results of simulation study: average values of simultaneous 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  $80$ .

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 Tab. 4.1.

Further we examined the coverage of the pointwise credibility/confidence bands for Bayesian estimation and Aalen estimators on  $[0, \tau^*]$ . The coverage was calculated as the proportion of the datasets where the true cumulative regression functions were contained within the pointwise credibility/confidence bands (again, evaluated on the thin grid on  $[0, \tau^*]$ ). See Tab. 4.2 for the results.

From the results in Tab. 4.1 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  $A_1$ . Overall, the average of the functional BIAS of the Bayesian estimators does not suggest any discrepancy from the consistency, as it has a decreasing tendency for both priors. Interesting is that, 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 sought parameter is fairly small. Further, from the coverage results in Tab. 4.2 we see that the incidences when the real regression function is contained in the 95% pointwise credibility bands on the shortened interval  $[0, \tau^*]$  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 interesting result is that the surface of the estimated credibility bands is for smaller datasets ( $n = 25$ ) about

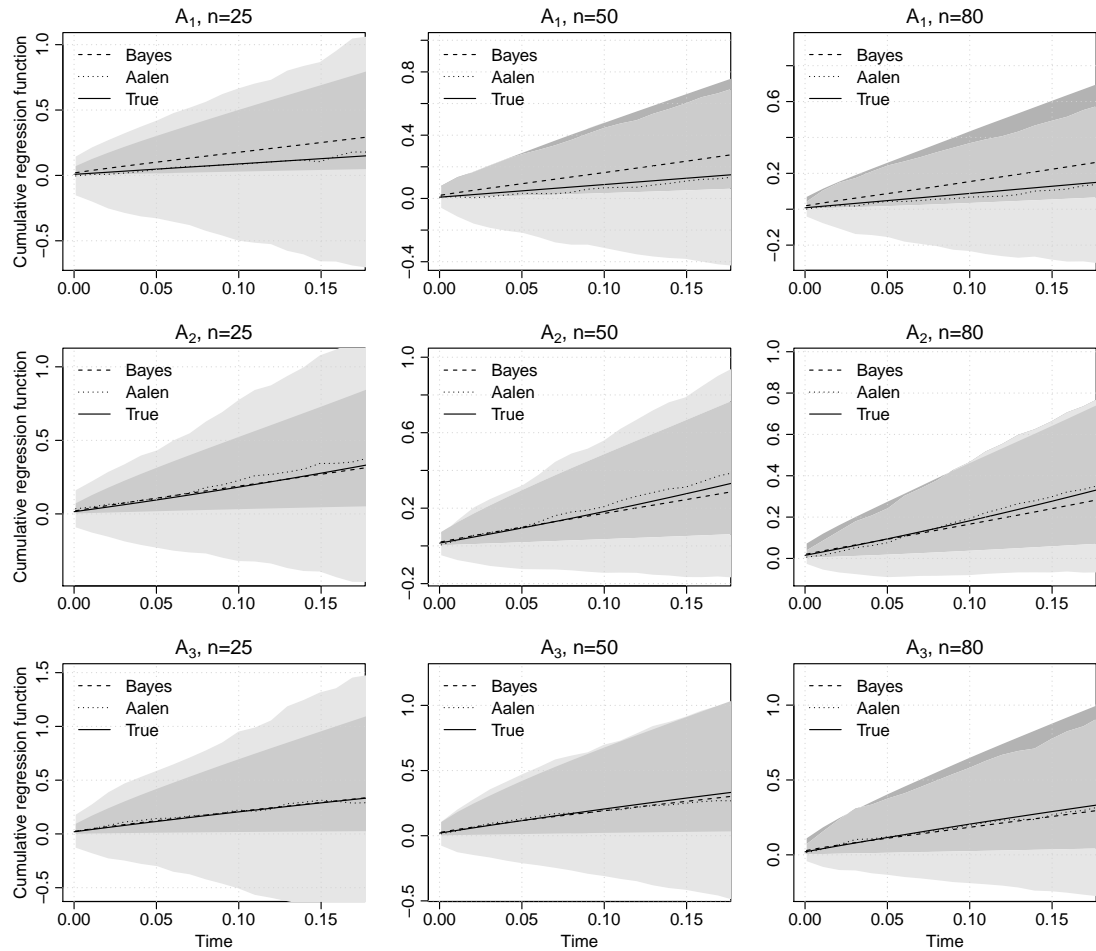


Figure 4.1: Graphs of the pointwise averages of the estimators obtained from 100 repetitions for prior 1 and numbers of observations  $n = 25, 50$  and  $80$ . The true regression function is plotted in solid line, average of the Bayesian estimators in dashed line and average of the Aalen's estimators 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.

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 be 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 from 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. The Bayesian estimators work with more information from

the very start as they are restricted to the positive values, while Aalen estimators span the whole real line at every time point.

Furthermore, 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  $\tau^* = 0.17$  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.

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 Fig. 4.1. 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. 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.

### 4.3.1 Computational aspects of the estimation

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 3.20GHz  $\times$  4 and 3.8 GiB RAM. An 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  $n = 25, 50$  and  $80$ ,

respectively. The number of observations and choice of hyperparameters influence the computational time in a great deal. The most crucial is the approximation of the mixture weights  $\beta_r$  suggested in Section 4.2. To assess the performance of the proposed approximations with different choice of  $L$  we decided to resort to an easier setting.

The most simple situation arises when we suppose that the regression functions are constant over the whole time window. Here, in particular, let us have a model  $h_i(t) = 1 + x_{i,2} + x_{i,3}$ , where  $x_{i,2} \sim \Gamma(2, 2)$  and  $x_{i,3} \sim \Gamma(1, 2)$ . We suppose that the regression function  $\alpha_j(s) \equiv \lambda_j, \forall s$ . The posterior distribution of the level  $\lambda_j$  is the mixture of the gamma distributions similarly as in the very last interval of the general model with jumps, see (4.3). The restriction to no jumps in the path of the regression functions reduces the complexity of the Bayesian solution tremendously. The approximation problem however remains present as all non-censored observations fall in one considered interval, the whole time window. We tested the performance of the approximation proposed at the end of Section 4.2.1 on 1000 small datasets with  $n = 10$  observations. The chosen hyperparameters of the prior distribution were  $a_0 = 0.1$  and  $b_0 = 0.1$ . Here we were able to evaluate the posterior distribution of the  $\lambda_j$  with exact weights. We calculated the approximated posterior distributions with  $L = 20$  and  $L = 50$  alongside. It turned out that the output of the Gibbs samples based on the approximated posterior distributions are very close to the estimators obtained from the exact distributions, see Table 4.3. In Table 4.3 there are shown the averaged characteristics calculated across all the simulated datasets. In the first column we have the averages of the mean of all but burn-in trajectories  $\lambda_j, \forall j$ , and the averages of the median, 2.5% quantile and 97.5% quantile of all trajectories follow in next columns. The estimators based on the exact posterior distribution and the  $L = 20$  approximated posterior distributions for every simulated dataset are plotted in Figure 4.2. As it can be seen, the pairs of the two estimators exhibit very good correspondence.

|                   |         | mean( $\lambda_j$ ) | median( $\lambda_j$ ) | 0.025 quantile( $\lambda_j$ ) | 0.975 quantile( $\lambda_j$ ) |
|-------------------|---------|---------------------|-----------------------|-------------------------------|-------------------------------|
| EXACT:            | $j = 1$ | 1.178               | 1.043                 | 0.092                         | 2.975                         |
|                   | $j = 2$ | 1.216               | 1.050                 | 0.055                         | 3.329                         |
|                   | $j = 3$ | 1.190               | 0.844                 | 0.001                         | 4.304                         |
| APPROX $L = 20$ : | $j = 1$ | 1.174               | 1.041                 | 0.092                         | 2.981                         |
|                   | $j = 2$ | 1.216               | 1.047                 | 0.056                         | 3.327                         |
|                   | $j = 3$ | 1.194               | 0.848                 | 0.001                         | 4.297                         |
| APPROX $L = 50$ : | $j = 1$ | 1.174               | 1.045                 | 0.095                         | 2.980                         |
|                   | $j = 2$ | 1.214               | 1.045                 | 0.054                         | 3.322                         |
|                   | $j = 3$ | 1.194               | 0.847                 | 0.001                         | 4.318                         |

Table 4.3: The statistics of estimated regression functions based on the evaluation of the posterior distribution, exact on the top, approximated on the bottom.



Hence, it seems that the approximation with  $L = 20$  should be sufficient enough for models with about 10 observations within intervals.

The results in Table 4.3 suggest that the median of the trajectories might be more accurate than the mean value. This can be caused by the possible outliers produced in some iterations of the Gibbs sampler due to the fact, that the estimator is sampled from the Gamma distribution, hence a nonsymmetrical distribution with heavier right tail.

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 and 48.3 seconds for the approximated posteriors with  $L = 20$  and  $L = 50$ , respectively. There is apparent a slight decrease in the computation time when  $L = 20$  is used. The increase of the computational time when  $L = 50$  is used is because of the random generation of the permutations which is more time consuming than the exact calculations. Once the number of observations within the intervals is greater than 15 the little cost of time due to random samples generation will be negligible in comparison to evaluation of all permutations.

The simulations conducted in the simpler setting should give us certain ease when using the proposed approximation of  $\beta_r$ . Obviously, there are no means to test the accuracy of the approximation in larger datasets as obtaining the exact values of  $\beta_r$  is computationally intractable. Still, the results gained in the simple scenario suggest that the approximation with sensibly chosen  $L$  should be satisfactory. The choice of  $L$  should be considered in connection with the mean number of observations expected within the intervals.

The number of iterations 500 and burn-in value 100 in the simple case of constant regression functions was based on the MCMC trace of the parameters

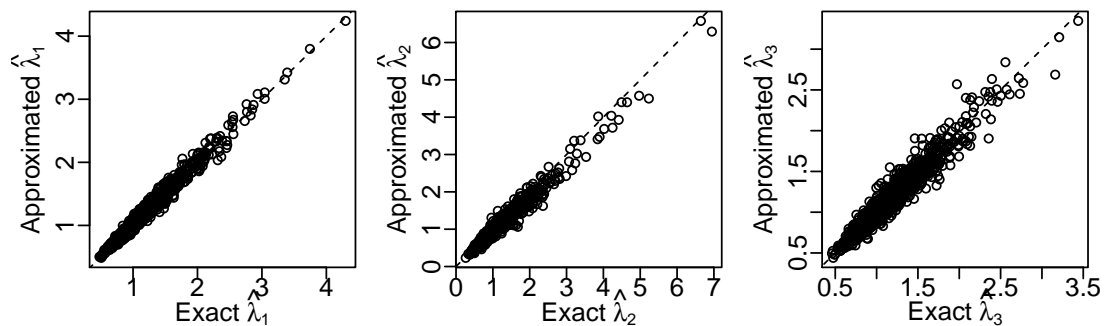


Figure 4.2: The values of the estimators obtained via exact posterior distribution plotted against the estimators from the  $L = 20$  approximated posterior distribution for a dataset with observations  $n = 10$ , the hazard rate  $h_i = 1 + x_{i,2} + x_{i,3}$  and a constant estimation with no jumps.

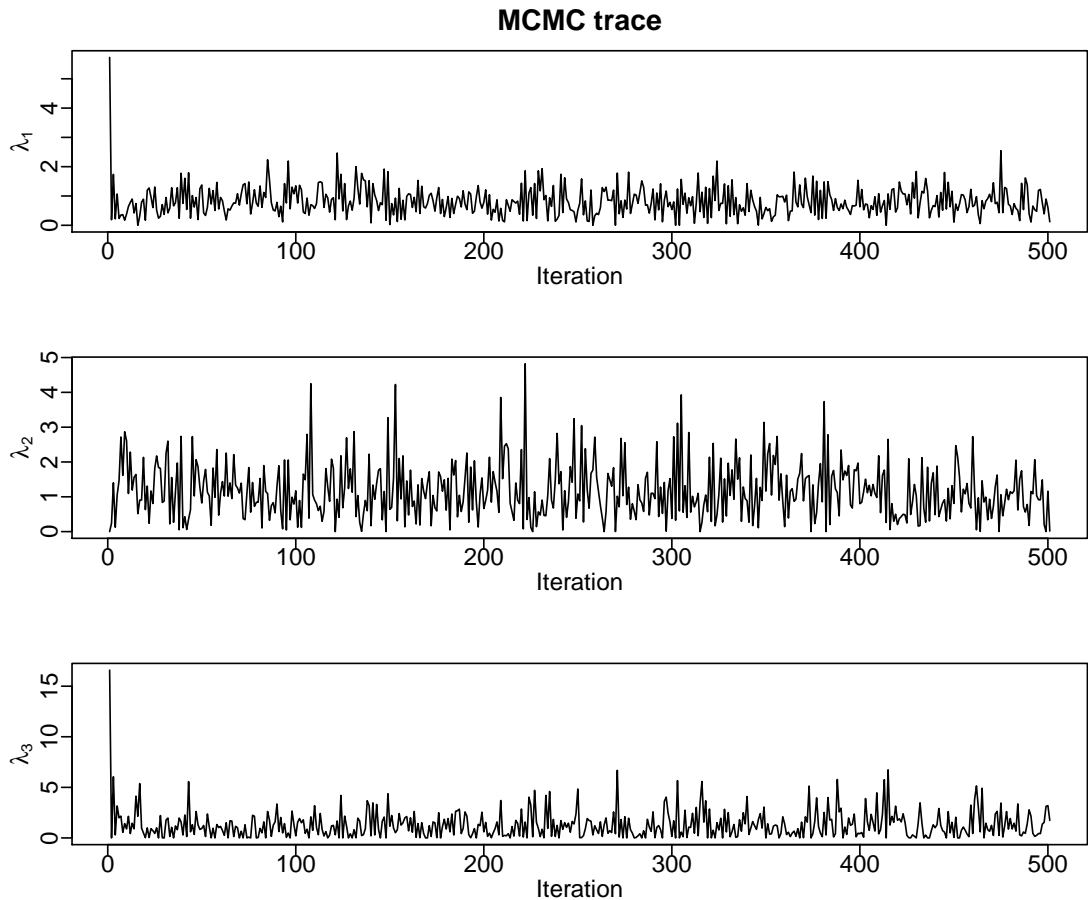


Figure 4.3: The MCMC trace of the regression functions for a dataset with observations  $n = 10$ , the hazard rate  $h_i = 1 + x_{i,2} + x_{i,3}$  and a constant estimation with no jumps.

$\lambda_j, j = 1, 2, 3$ . An example of MCMC traces of the estimated parameters for a randomly chosen dataset is in Figure 4.3.

Assessing the convergence when the original situation is considered is problematic due to moving jumps and varying number of intervals. To justify the chosen number of iterations and burn-in we plotted MCMC traces of regression functions for number of datasets at random time points. Most of the traces evinced signs of stabilization around a mean value after 50 to 100 iterations. An example of the MCMC traces of estimated regression functions at a randomly chosen time point can be seen in next section.

### 4.3.2 Choice of the hyperparameters

In this section we deal with the choice of hyperparameters for one of the datasets from simulation study with  $n = 50$ . Let us recall, that the MCMC settings was  $L = 20$ ,  $U = 500$  and burn-in = 100. We have tried several hyperparameters for

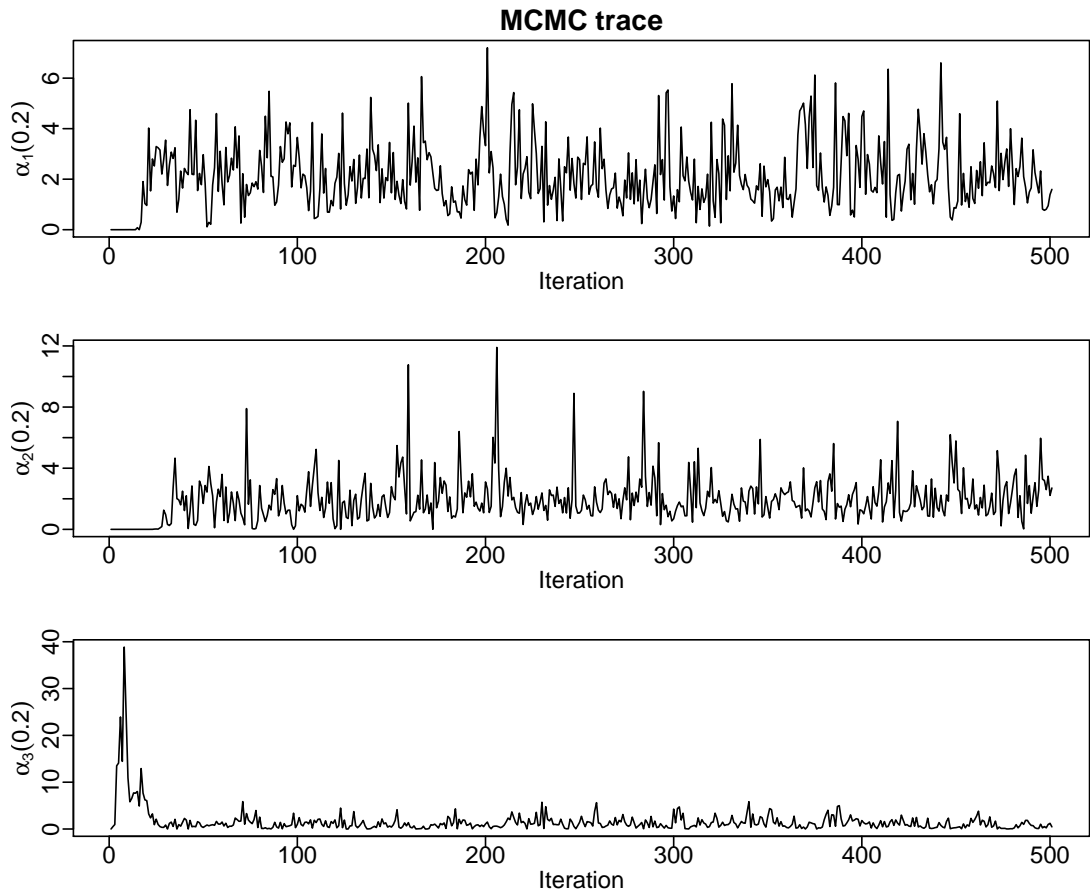


Figure 4.4: The MCMC trace of the trajectories obtained with the estimation with PRIOR 2. The plotted values are taken at time point  $t = 0.2$ .

prior distributions. In particular, we had

$$\text{PRIOR 1: } a_0 = 0.1, b_0 = 0.1, a = 1, c = 1, d = 20,$$

$$\text{PRIOR 2: } a_0 = 0.1, b_0 = 0.1, a = 0.5, c = 1, d = 50,$$

$$\text{PRIOR 3: } a_0 = 0.1, b_0 = 0.1, a = 10, c = 1, d = 20.$$

The prior was always the same for each regression function. We chose low values for  $a_0$  and  $b_0$  to set a rather uninformative start of the MCMC trajectory. The value of the regression functions in the first interval is on average a priori equal to 1, but the prior variance equal to 10 allowed for great variability. The level of variability throughout the whole trajectories was controlled by choice of parameters  $a$  and  $d$ . Smaller value of parameter  $a$  and greater values of parameter  $d$  implied greater variation from one interval to another. Hence, PRIOR 2 would produce the most variable estimators while PRIOR 3 should be the flattest. PRIOR 1 is a middle way between these PRIOR 2 and 3.

The estimated regression functions and their cumulative versions are plotted in Figure 4.5, Figure 4.6 and Figure 4.7. For comparison we added classic Aalen estimators to graphs with the cumulative regression functions. Apparently, the obtained estimators exhibit the behaviour expected from the choice of their respective prior distributions. The greatest variation is indeed to see in Figure 4.6 with PRIOR 2. The estimators tend to copy the real regression functions in particular at the beginning of the time window. The great variation at the start could be possibly attributed to the values of parameter  $a_0$  and  $b_0$ . The effect of different prior distribution is less obvious when looking at the cumulative regression functions on the right hand side of figures, apart from the very flat estimators with PRIOR 3. We also plotted MCMC traces from the estimation with PRIOR 2 at time point 0.2 in Figure 4.4.

## 4.4 Discussion

The estimation proposed in this chapter was taken down a slightly different path than in Chapter 3 where we estimated the cumulative regression function with Beta process priors. On contrary, here we worked with likelihood based on continuous regression functions. The performance of the method based on Arjas nad Gasbarra's correlated prior was tested in the simulation study. The focus was in particular on the consistency of the estimators, which was very lacking in the previous chapters. The method provides one with the estimators of the regression functions  $\alpha_j$  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 than the noncumulative ones, hence, they might be preferred. On average the estimators of  $\alpha_j$  are 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 intervals. 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

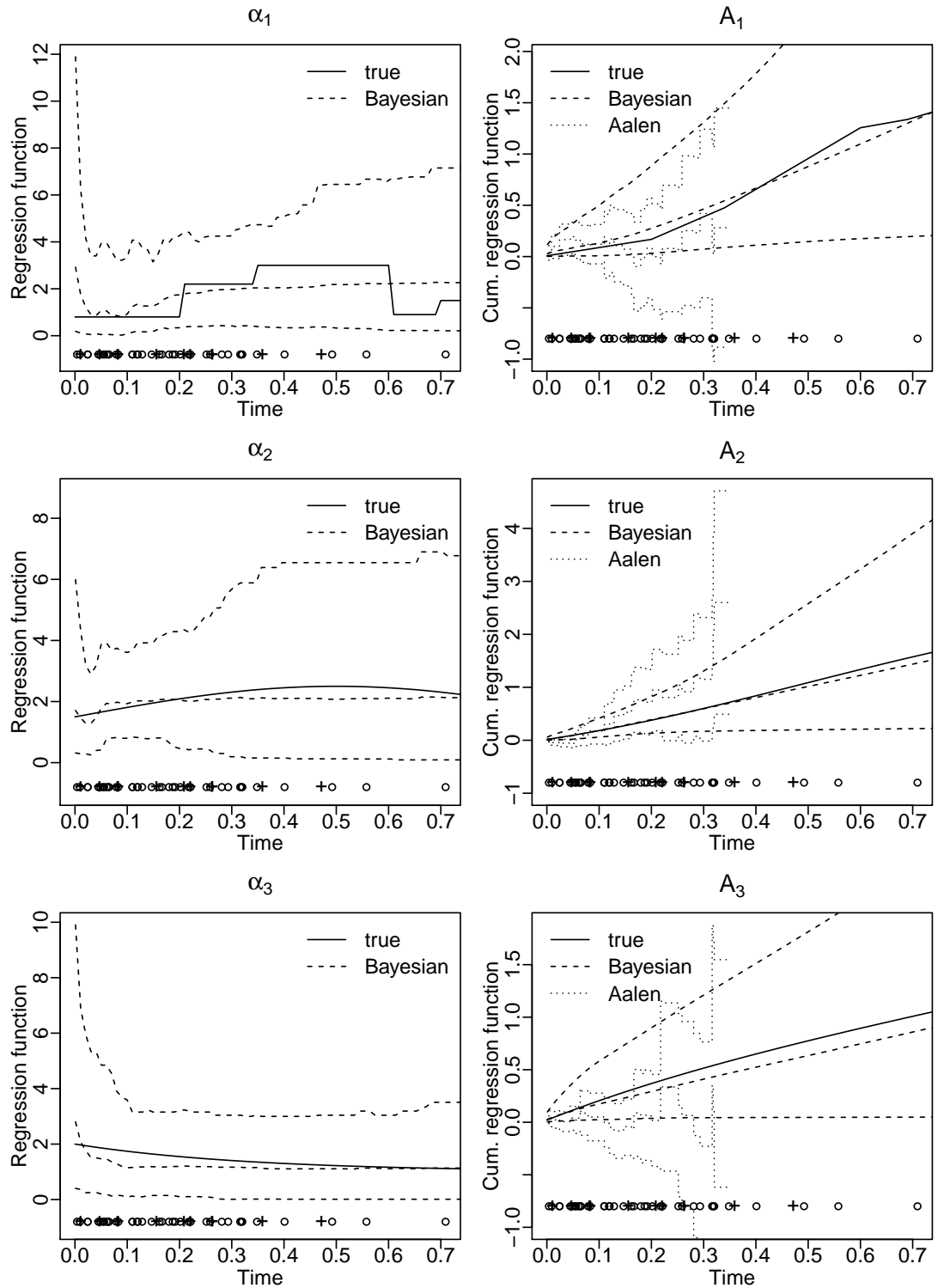


Figure 4.5: PRIOR 1 with parameters  $a_0 = 0.1$ ,  $b_0 = 0.1$ ,  $a = 1$ ,  $c = 1$  and  $d = 20$ . The left hand graphs show the real regression functions in solid lines and estimated regression functions in dashed lines. The true cumulative regression functions are plotted on the right hand figures in solid lines. The estimated versions are in dashed lines and Aalen estimators are in dotted lines.

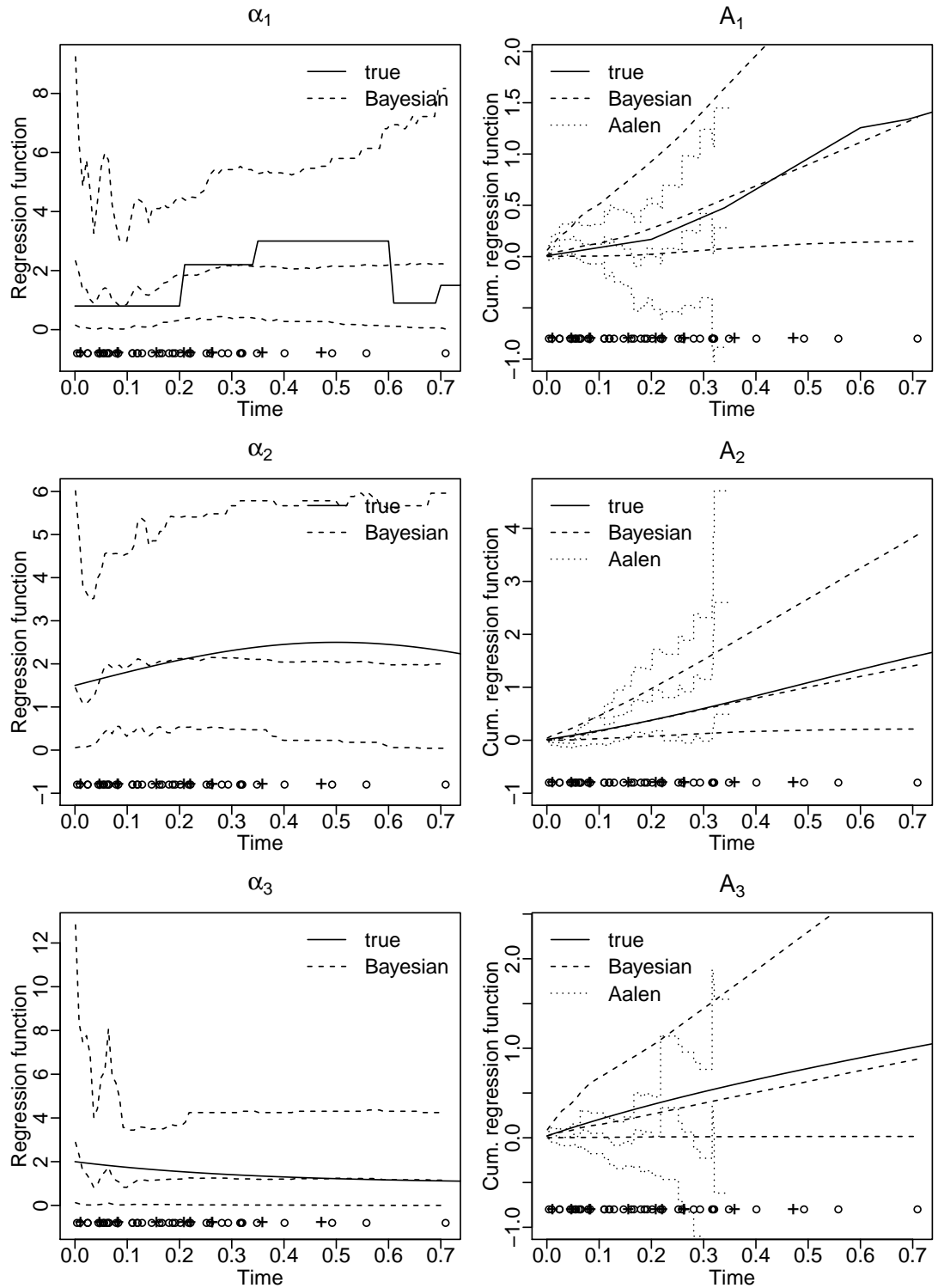


Figure 4.6: PRIOR 2 with parameters  $a_0 = 0.1$ ,  $b_0 = 0.1$ ,  $a = 0.5$ ,  $c = 1$  and  $d = 50$ . The left hand graphs show the real regression functions in solid lines and estimated regression functions in dashed lines. The true cumulative regression functions are plotted on the right hand figures in solid lines. The estimated versions are in dashed lines and Aalen estimators are in dotted lines.

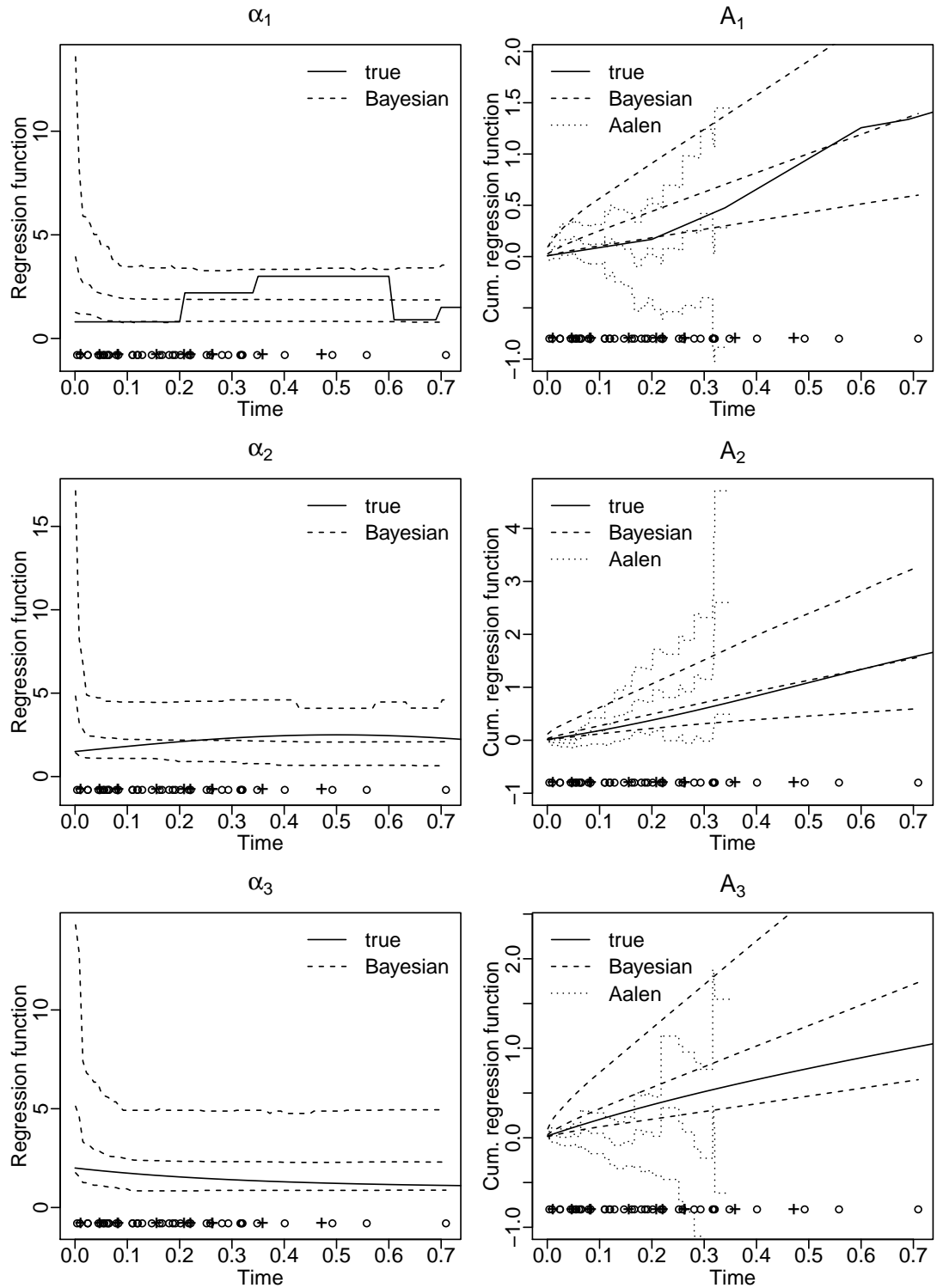


Figure 4.7: PRIOR 3 with parameters  $a_0 = 0.1$ ,  $b_0 = 0.1$ ,  $a = 10$ ,  $c = 1$  and  $d = 20$ . The left hand graphs show the real regression functions in solid lines and estimated regression functions in dashed lines. The true cumulative regression functions are plotted on the right hand figures in solid lines. The estimated versions are in dashed lines and Aalen estimators are in dotted lines.

to abide by the monotonicity condition. Another advantage is that we can obtain the estimators of the  $\alpha_j$ -s 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 chapter by suggesting a few possibilities for a future work in this direction. It is clear that a greater simulation study is needed to draw a steady conclusion on the consistency of the proposed estimators. Furthermore, it is feasible to extend the proposed method into several more complicated scenarios. It could be applied on a 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 to a Bayesian approach to 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 [43].



# Chapter 5

## The real data example revisited

In this chapter we return to the real dataset introduced in Section 1.5. We apply the estimators developed in the previous chapters and look at their performance in real data setting in Section 5.1. Apart from the real dataset we also take a look at the well-known Danish malignant melanoma data in Section 5.2 as this dataset has served as a benchmark dataset in the survival analysis for years.

### 5.1 Time-delay dataset

Let us remind that the dataset consists of records of the time-delay, i.e. the duration from the onset of myocardial infarction until the reperfusion surgery takes place. The data were collected at the Royal Vinohrady Teaching Hospital in Prague in the Czech Republic and contains 622 entries. We already did a preliminary analysis of four factors which are believed to increase the hazard function (i.e. shorten the time-delay) in Section 1.5. Based on the beforehand knowledge on the time-delay issue we concluded that the monotone Aalen model is an appropriate choice to model the relationship between the hazard rate and the four chosen factors, namely gender, presence of first contact delay (yes/no), which part of the day the myocardial infarction set in (day/night) and working status of the patient (employed/unemployed or retired).

We estimated the cumulative regression functions using the nonparametric maximum likelihood estimators developed in Chapter 2 and the Bayesian estimators based on Beta process priors from Chapter 3. The parameters of Beta process prior were set to  $c_j(t) \equiv c = 0.001$  and  $A_j^0(t) = \alpha_j^0 t \equiv 0.001t$ ,  $j = 1, \dots, 5$ , i.e. a priori the cumulative regression functions increase on average by 1 every 1000 minutes but by setting  $c = 0.001$  we allow for a great variability. As we already know, when the prior is the same for all covariates then asymptotically the effect of the parameters of the prior processes disappears. In particular, at

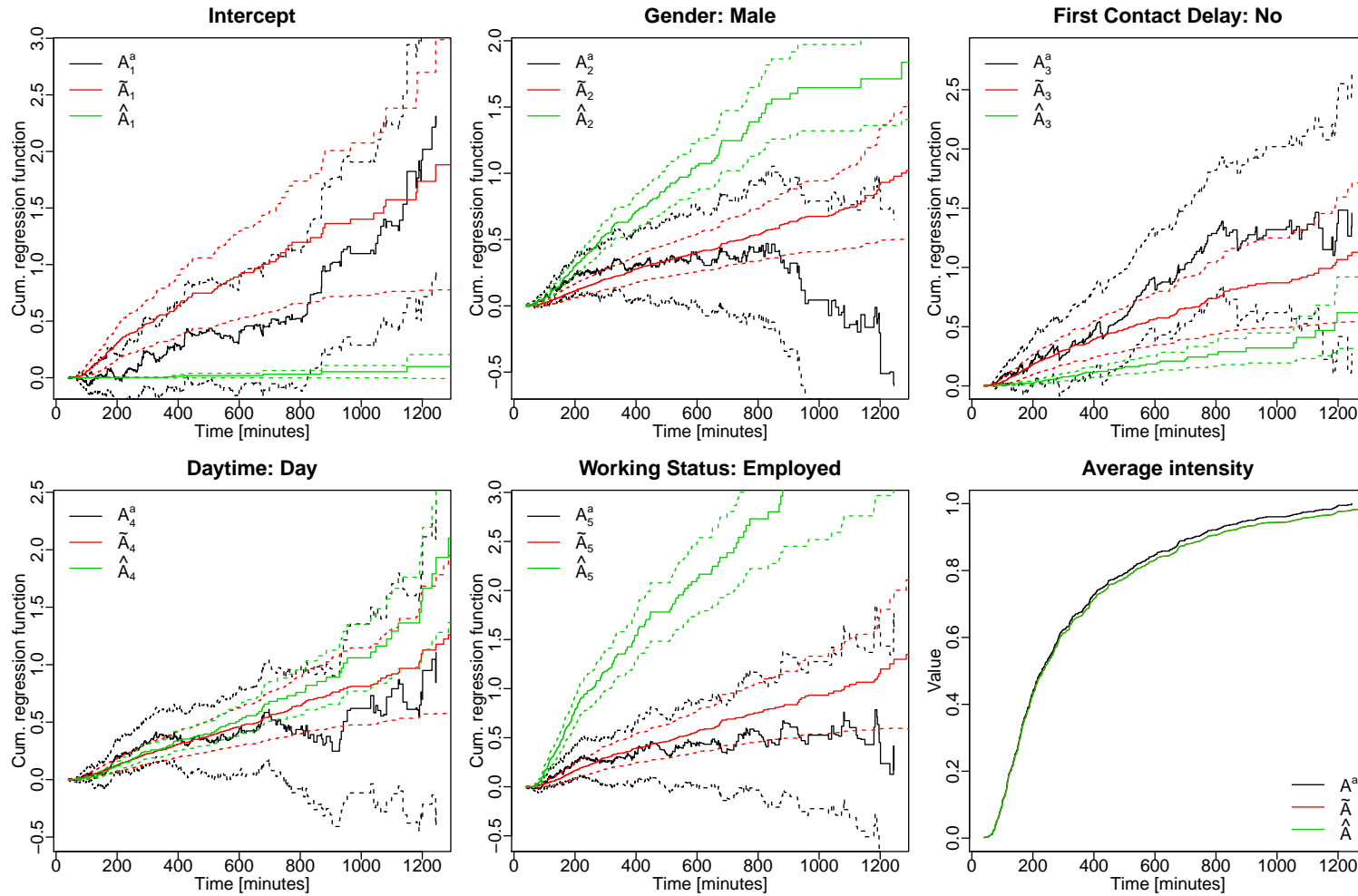


Figure 5.1: The estimated cumulative regression functions from the classic Aalen model for the time-delay of the myocardial infarction patients, NPML estimators and Bayesian estimators with Beta process priors. Estimates of the average intensity are included.

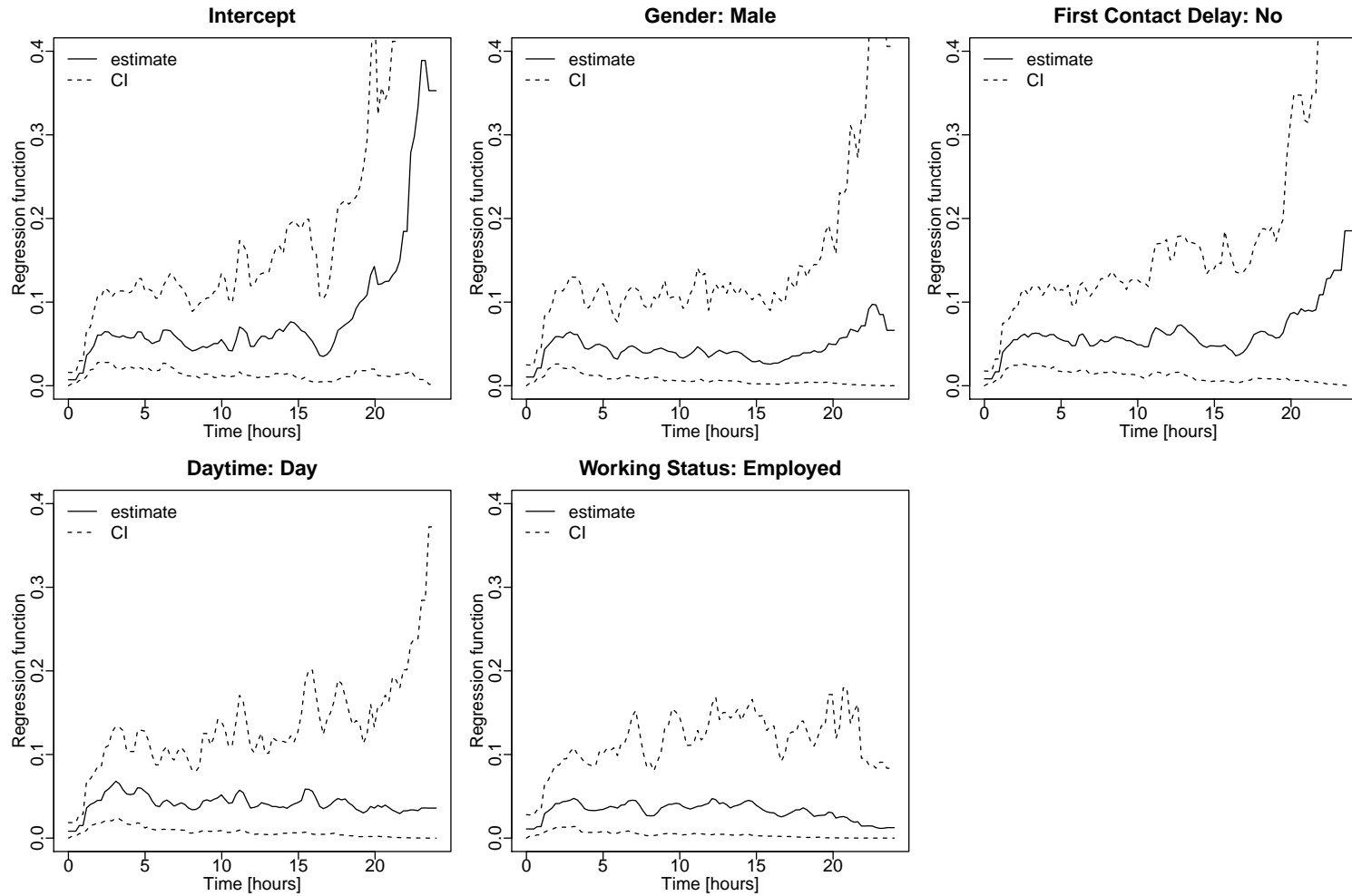


Figure 5.2: The Bayesian estimation with correlated prior for the regression functions for the data on time-delay of the myocardial infarction patients. The pointwise 95% credibility bands are included in dashed lines.

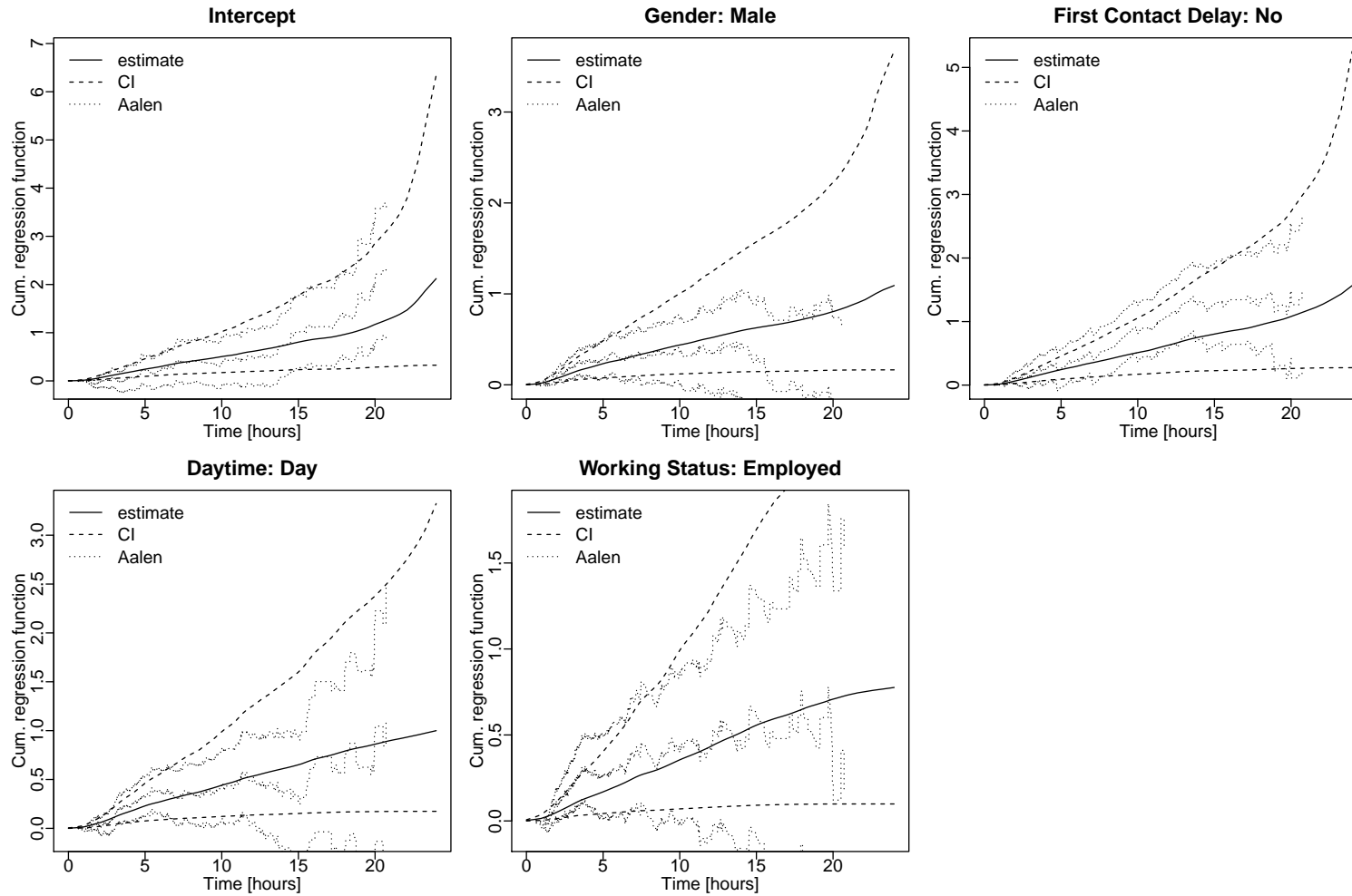


Figure 5.3: The Bayesian estimation with correlated prior: cumulative versions of the estimated regression functions with pointwise 95% credibility bands in dashed lines. Aalen least squares estimators are included in dotted lines.

fixed discontinuities for great  $n$  the parameters crosses out while the stochastically continuous part vanishes. Hence, with great amount of data and equal priors there is no or little impact of the choice of the prior parameters on the estimators.

The fitted cumulative regression functions can be seen at Figure 5.1. The nonparametric maximum likelihood estimators  $\hat{A}_j$  with 95% pointwise confidence bands based on the derivations in Remark 6 are in green lines while in red lines we have the approximated Bayesian estimators and 95% pointwise credibility bands calculated from the approximated variances in (3.27). The pointwise credibility bands were calculated using a normal approximation, for details see Section 3.8. The least squares estimators are included in black lines. Obviously, both the NPML and Bayesian estimators are very different from the Aalen least squares estimators and even though they are of similar structure, they are different from each other. The last subplot, however, shows that averaged cumulative intensities coincide.

As a next step we estimated the regression functions using the Bayesian approach with correlated prior introduced in Chapter 4. As the increase in the hazard rate associated with one minute is very small, the estimation based on sampling from gamma distribution can run into a computational difficulties for gamma distributions with mean close to zero. Therefore the data were transformed into hours with maximal observed value equal to 24 hours. We assumed a flexible prior with great variation and with following set of parameters  $a_0 = 0.01$ ,  $b_0 = 0.1$ ,  $a = 1$ ,  $c = 0.05$  and  $d = 3$  for every regression function. The mean number of jumps in the trajectories was a priori around 40. The estimated regression functions based on the averages of the MCMC trajectories on a grid and their cumulative versions are plotted in Figure 5.2 and Figure 5.3. The pointwise 95% credibility bands made of 2.5% and 97.5% quantiles are included in dashed lines. The cumulative versions are plotted on the right hand side of the figure together with the Aalen least squares estimators. Both Bayesian and least squares estimators show quite similar behaviour. The shape of the Bayesian estimators of the cumulative regression functions suggests that the regression functions are close to linear. The pointwise credibility bands are almost the same or slightly wider than the Aalen least squares pointwise confidence bands.

## 5.2 Danish malignant melanoma

The famous Danish dataset contains survival times of patients with malignant melanoma who had their tumour completely removed. The 225 patients were

followed during the study from 1962 to 1977 and the information whether the patient survived or died during the follow-up up to 1977 was recorded. If the person died from other reason than the melanoma, their observation was marked as censored. The variable of main interest was the time since the operation. Furthermore, gender, age of the patient and characteristics of the removed tumour were collected. In our analysis we will focus on gender and two characteristics of the tumour: thickness of the removed tumour and the ulceration, i.e. dichotomous variable describing if ulcers were present on the surface of the tumour. Let us ponder for a while if the monotone Aalen model is a suitable choice to analyse the relationship between the survival times and the three variables. The greater values of thickness and presence of ulcers is known to be related to later stages of the skin cancer, hence with growing thickness and present ulceration the survival time will shorten (i.e. the two characteristics can be viewed as risk factors). This goes well with the assumption of the monotone Aalen model. A questionable aspect is the sex variable as there is no obvious reason why the mortality should be greater for one gender than for another. However, according to several sources there is a statistical evidence that men are more likely to develop and die of melanoma than women and this risk is linked to the fact, that men have higher annual exposure to UV. If we accept this information as reliable enough, we may proceed to the analysis of the dataset using the monotone Aalen model. We will also compute the classic Aalen estimators to allow for comparison with our estimators.

For total of 205 patients we have complete information on their gender, thickness of the tumour and ulceration. Most of the survival times are censored and there is precisely 57 cases of death and 148 censored observations. The maximal observed time equals to 5565 days while the greatest noncensored observation is 3338 days since the surgery.

We calculated the estimators of the cumulative regression functions related to gender, thickness of the tumour and presence of ulceration using the nonparametric maximum likelihood estimator developed in Chapter 2, Bayesian estimators with Beta processes in Chapter 3 and Bayesian estimators based on the correlated prior in Chapter 4.

The estimators based on the nonparametric maximum likelihood method and Bayesian with Beta processes can be seen at Figure 5.4. Similarly as for the dataset on time-delays from previous section, we set the parameters of Beta process prior to  $c_j(t) \equiv c = 0.001$  and  $A_j^0(t) = \alpha_j^0 t \equiv 0.001t$ ,  $j = 1, \dots, 4$ , i.e. we think that the cumulative regression functions increase by 1 every 1000 days but by setting  $c = 0.001$  we state that we do not have a great belief in the chosen prior

processes. The NPML estimators  $\hat{A}_j$  together with the 95% pointwise confidence bands are plotted in green lines while the approximated Bayesian  $\tilde{A}_j$  estimators with 95% pointwise confidence bands based on approximated variances and normal approximation are in red lines. The black lines represent the Aalen least squares estimator. Obviously, our estimators differ from the Aalen estimators greatly, what is no surprise considering their asymptotic features. The estimated cumulative regression functions still average each other out, as it can be seen at the plot of the average intensity in the right hand bottom corner of the Figure 5.4.

Similarly as for the time-delay dataset, the estimation based on Bayesian analysis with correlated prior required that the survival time since operation was considered in years and not days. This is done to avoid the computational difficulties with sampling from gamma distribution with mean close to zero, which happens when the risk associated with a day is very small. In Figure 5.5 there are displayed results for the Bayesian analysis with the correlated prior where we chose the parameters  $a_0 = 0.01$ ,  $b_0 = 1$ ,  $a = 0.1$ ,  $c = 0.1$  and  $d = 3$ . Again, the chosen parameters allowed for great variation with average prior number of jumps equal to approximately 20 jumps. The number of iteration was 500 and first 100 was discarded as burn-in part. The posterior pointwise mean and credibility bands based on 2.5% and 97.5% quantiles can be seen in the graphs. The estimated cumulative regression functions are not equal to Aalen's least squares estimators and for the cumulative regression functions related to thickness and ulceration we can see a certain deviation. In the simulation study which is described in Section 4.3 we observed many datasets for which the estimators based on the correlated prior and on the least squares differed, see for example Figure 4.5. Also the differences seem to average each other out, hence overall the estimated hazard rate should be quite similar based on both approaches. What is of greater concern here is the size of the pointwise credibility bands which are considerably greater. This is partly caused by the flexibility of the chosen prior and it is present especially for greater time hence possibly it is implied by lack of the failures after 10 years of follow-up. It is also a characteristic trait of the method (discussed in Section 4.4). The other disadvantage is the computational time required to run the MCMC which was about 1 hour.

Let us return to the interpretation of the model. In particular, in our setting, the baseline hazard rate contained in the intercept represents a female patient with the tumour thickness equal to zero and with no ulceration, hence, it is a healthy woman without melanoma. Of course, the estimation is an extrapolation from the data which contain solely patients with diagnosed melanoma. It is still

expected that this healthy woman has a certain risk of eventually dying of skin cancer over the course of years, hence this cumulative baseline hazard rate is expected to be either close to zero or slightly growing. This is not fulfilled with the Aalen least squares estimator of the intercept, which is decreasing and have therefore difficult interpretation. The problems with a decreasing least squares estimator of intercept can be overcome by shifting the quantitative covariates to have their mean in zero, i.e. in our case we would subtract 2.92 from the thickness variable. This is however not possible when using the monotone Aalen model as all covariates must be nonnegative. The restriction to using only nonnegative variables can be seen as a limitation of usage of the monotone Aalen model, as is for example in case of the malignant melanoma where shifting variables might propose better interpretation of the estimated cumulative baseline hazard rate.



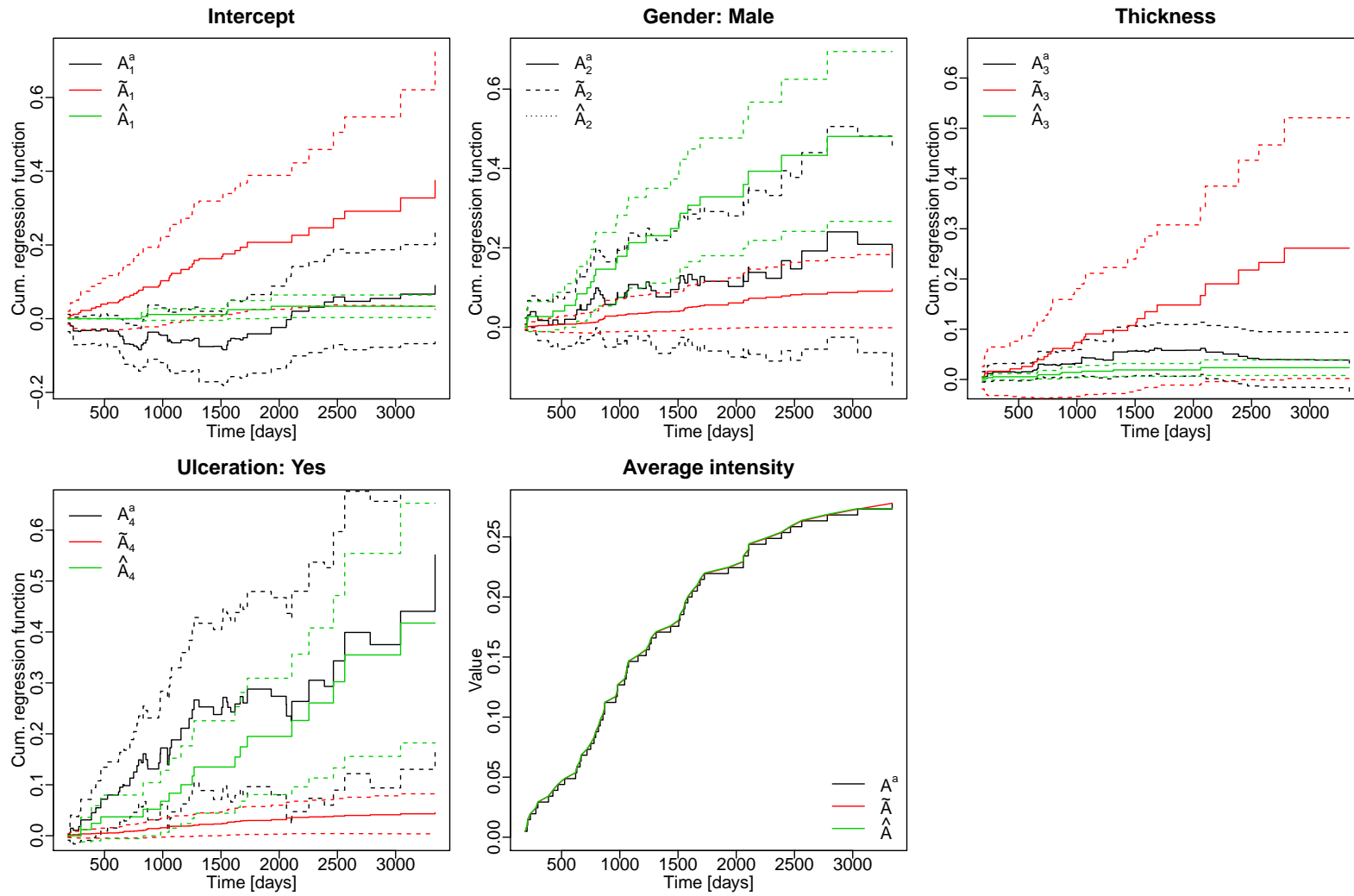


Figure 5.4: The estimated cumulative regression functions from the classic Aalen model for the Danish melanoma data, NPML estimators and Bayesian estimators with Beta process prior. Estimates of the average intensity are included.

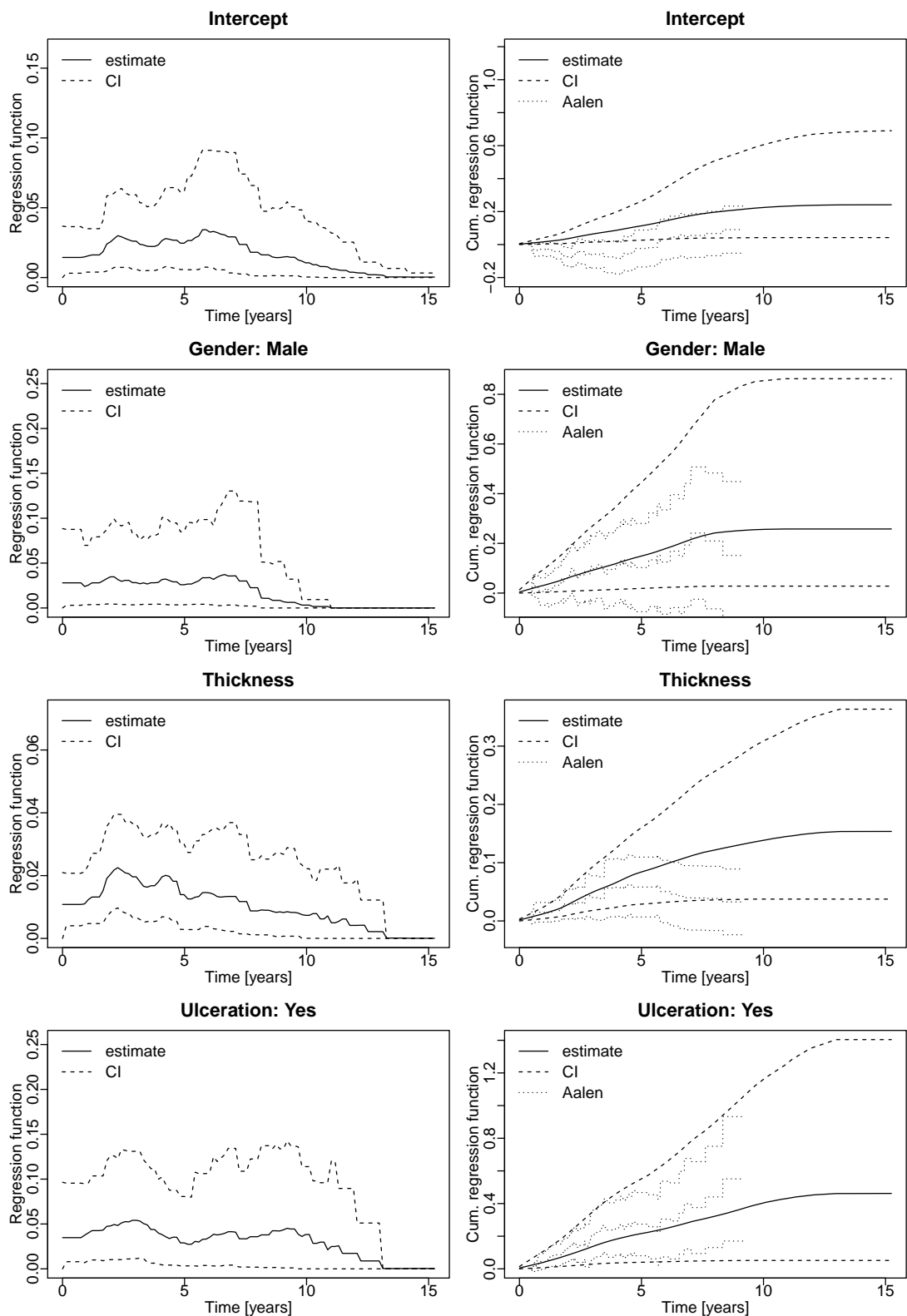


Figure 5.5: *Left:* The Bayesian estimation with correlated prior for the regression functions for the Danish melanoma data. The pointwise 95% credibility bands are included in dashed lines. *Right:* The cumulative versions of the estimated regression functions with pointwise 95% credibility bands in dashed lines. Aalen least squares estimators are included in dotted lines.

# Chapter 6

## Discussion

The main objective of the thesis was to find estimators for the (cumulative) regression functions in the Aalen model under the monotonicity restriction. We supposed that the cumulative regression functions  $A_j(t) = \int_0^t \alpha_j(s) ds$  were non-decreasing and the covariates were nonnegative. With this restraint we had that the regression functions  $\alpha_j$  were nonnegative and it allowed for a more natural interpretation of the regression functions. If the baseline hazard is included in the model, it represents a level of hazard of failure for a normal healthy individual. Every additional covariate then can be viewed as an excess additive risk imposed on the individual. This is a frequent concept in many epidemiological and health studies where often an impact of a particular hazardous behaviour or environment on health is investigated. This could be for instance a study of a time to manifestation of lung cancer of workers in uranium mines in relation to the average daily dose of radon measured in the respective sites of the mines or an effect of reported daily dose of cigarettes on time to developing a cardiovascular disease. The monotone diversion from the classical Aalen model carries along its nice features, i.e. that the effect of the covariates on the hazard function can vary over time and it does not require proportional hazards as in Cox model.

In search for an estimator of regression functions we resorted to the maximum likelihood methods. First, we developed well defined nonparametric maximum likelihood estimators for the cumulative regression functions. We explored its features including the large sample behaviour. It revealed itself that the nonparametric maximum likelihood estimation works well solely under the model with hazard rate equal to  $h(t) = \alpha_1(t)x_1$  where no baseline hazard and only a single covariate is included. For more covariates the estimators become inconsistent. These results have been proven and the limiting functions and weak convergence were investigated. A consistency was, however, found for the aver-

age intensity (i.e. a concrete linear combination of the nonparametric maximum likelihood estimators of the cumulative regression functions).

As the next step we explored a Bayesian approach by assuming Beta processes as prior processes for the cumulative regression functions. We explained in detail the features of these step processes, which turns out to be perfectly suited to be cumulative regression functions with regards to existence of the well defined cumulative hazard function. We derived the posterior distribution, investigated its traits and took the posterior expectation to serve as the Bayesian estimators for the cumulative regression functions. Unlike with many Bayesian machines nowadays the estimators have explicit form and can be calculated without aid of the MCMC procedures. Similarly as in the nonparametric maximum likelihood estimation, the Bayesian estimators proved to be inconsistent if more than one covariate was involved. The same consistency result for the average intensity as of the nonparametric maximum likelihood estimators was found.

The third path of the research in the thesis was again taken towards the Bayesian methods, however with continuity assumption on the cumulative hazard rate. We proposed a sensible prior distribution with a martingale structure based on Arjas and Gasbarra's work, [6]. The method approximates the baseline hazard rate and the regression functions using piecewise constant functions with a random number and locations of jump times. We derived the posterior distribution for the parameters of the model and proposed a sampling algorithm for generation of the estimators via Gibbs sampling. The performance of the method was tested in the simulation study. The results of the simulations suggest a 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 intervals. 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. All in all, the method looks promising, however, the computational demands and lack of the knowledge on large sample behaviour makes its use questionable.

The performance of all estimators was displayed under particular settings in examples, on a data on time-delay of patients with myocardial infarction and on the "benchmark" Danish malignant melanoma dataset.

The inconsistency we have revealed with the nonparametric maximum likelihood method and the Bayesian analysis with Beta processes is rather curious. There is not an obvious cause of this discrepancy. We are left to consider this

to be one of the infinitely-dimensional cases where the consistency is not always reached also with trustworthy maximum likelihood based methods. The additive character of the hazard rate could possibly inflict the difficulties with grasping the truth underlying the data.

## 6.1 Future research directions

The lack of easy-to-obtain and consistent estimators which would abide the monotonicity restriction is obvious. There are several possibilities which has not been pursued in this work and that look fairly promising.

We could abandon the nonparametric setting and specify the form of the regression functions in advance. A fairly non-restrictive set-up is to assume that  $\alpha_j$  are piecewise constant functions with fixed numbers of jumps and estimate the values of these functions via maximum likelihood method. The most simple situation arises when that the cumulative regression functions are linear, i.e.  $A_j(t) = \alpha_j t$ . The estimator would be given by the following estimation equation

$$\sum_{i=1}^n \int_0^\tau \frac{z_i}{z_i^\top \alpha} dN_i(s) - \sum_{i=1}^n \int_0^\tau Y_i(s) z_i ds = 0.$$

This is derived from the log-likelihood of the data under the assumption of constant regression functions. The generalization to piecewise constant regression functions is straightforward. In general we can suppose that the regression functions are given a parametric form,  $A_j(t) = \alpha_{j,\theta}(t)$ . The vector of unknown parameters  $\theta$  is again found by the maximum likelihood method.

The problem with estimation of several functional parameters could be overcome by another simplification. Let us suppose we have an additive model of the hazard rate with the following form

$$h(t) = h_0(t) + \alpha^\top z,$$

where  $h_0$  is a baseline hazard,  $\alpha = (\alpha_1, \dots, \alpha_p)^\top$  is a  $p$ -dimensional vector of unknown parameters and  $z$  is a covariate vector. Then there is only one infinitely-dimensional parameter to estimate and either the nonparametric maximum likelihood method or Bayesian methods could be applied. This type of model was already analysed Sinha et al. in [42] via integrated likelihood with Gamma process priors. Their work could possibly be extended by incorporating a parametric form for the regression functions  $\alpha_j = \alpha_{j,\theta}(t)$  to allow for the variation of the effect of covariates on the hazard rate.

Another option is to lean on the least squares method which gave the consistent estimators under the classical Aalen model. The goal would be to derive a monotone alternative to Aalen's and Huffer-McKeague's estimators by using restricted least squares estimation with constraint that the increments are positive, i.e.  $\Delta A_j \geq 0$ .

Furthermore, other Bayesian machines utilizing popular priors like Polya trees and mixture Dirichlet priors could be explored. In frequentist framework a possibility would be to transform the actual problem into a model with relative and excess risk and use related methods.

# Bibliography

- [1] O. O. Aalen. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, 6(4):701–726, 1978.
- [2] O. O. Aalen. A Model for Nonparametric Regression Analysis of Counting Processes. *Springer Lecture Notes in Statistic*, 2:1–25, 1980.
- [3] O. O. Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925, 1989.
- [4] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Kieding. *Statistical models based on counting processes*. Springer Series in Statistics Series. Springer-Verlag GmbH, 1993.
- [5] P. K. Andersen and R. D. Gill. Cox’s regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.
- [6] E. Arjas and D. Gasbarra. Nonparametric bayesian inference from right censored survival data, using Gibbs sampler. *Statistica Sinica*, 4:505–524, 1994.
- [7] J. M. Bernardo. Algorithm AS 103 psi(digamma function) computation. *Applied Statistics*, 25:315–317, 1976.
- [8] R. A. Boyles, A. W. Marshall, and F. Proschan. Inconsistency of the maximum likelihood estimator of a distribution having increasing failure rate average. *The Annals of Statistics*, 13:413–417, 1985.
- [9] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.
- [10] P. De Blasi and N. L. Hjort. The Bernstein–von Mises theorem in semiparametric competing risks models. *Journal of Statistical Planning and Inference*, 139(7):2316–2328, 2009.
- [11] P. Diaconis and D. Freedman. On the Consistency of Bayes Estimates. *The Annals of Statistics*, 14(1):1–26, 1986.
- [12] K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2(2):183–201, 04 1974.
- [13] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 03 1973.

- [14] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. John Wiley & Sons, Ltd., 1991.
- [15] S. Ghosal. A review of consistency and convergence rates of posterior distribution. In *Proceedings of National Conference in Bayesian Analysis*, Benaras Hindu University, Varanashi, India, 1996.
- [16] R. D. Gill, J. A. Wellner, and Jens Præstgaard. Non- and Semi-Parametric Maximum Likelihood Estimators and the Von Mises Method (Part 1) [with Discussion and Reply]. *Scandinavian Journal of Statistics*, 16(2):97–128, 1989.
- [17] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.
- [18] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- [19] N. L. Hjort. Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data. *The Annals of Statistics*, 18(3):1259–1294, 1990.
- [20] J. Huang and J. A. Wellner. Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis. Lecture notes in Statistics*, pages 16–9. Springer Verlag, 1996.
- [21] F. W. Huffer and I. W. McKeague. Weighted least squares estimation for Aalen’s additive risk model. *Journal of the American Statistical Association*, 86:114–129, 1991.
- [22] M. Jacobsen. Maximum likelihood estimation in the multiplicative intensity model: A survey. *International Statistical Review*, 52:193–207, 1984.
- [23] S. Johansen. An extension of Cox’s regression model. *International Statistical Review*, 51:165–174, 1983.
- [24] J. D. Kalbfleisch. Non-Parametric Bayesian Analysis of Survival Time Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40:214–221, 1978.
- [25] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, Inc., New York, 1980.
- [26] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.
- [27] Y. Kim. Nonparametric Bayesian Estimators for Counting Processes. *The Annals of Statistics*, 27(2):562–588, 1999.
- [28] Y. Kim. The Bernstein–von Mises theorem for the proportional hazard model. *The Annals of Statistics*, 34(4):1678–1700, 2006.



- [29] Y. Kim and J. Lee. On Posterior Consistency of Survival Models. *The Annals of Statistics*, 29:666–686, 2001.
- [30] Y. Kim and J. Lee. Bayesian Analysis of Proportional Hazard Models. *The Annals of Statistics*, 31(2):493–511, 2003.
- [31] Y. Kim and J. Lee. A Bernstein-Von Mises Theorem in the Nonparametric Right-Censoring Model. *The Annals of Statistics*, 32(4):1492–1512, 2004.
- [32] P. W. Laud, P. Damien, and A. F. M. Smith. Bayesian nonparametric and covariate analysis of failure time data. Dey, D. (ed.) et al., *Practical nonparametric and semiparametric Bayesian statistics*. New York, NY: Springer. Lect. Notes Stat., Springer-Verlag. 133, 213-225 (1998)., 1998.
- [33] J. Lee and Y. Kim. A new algorithm to generate beta processes. *Computational Statistics & Data Analysis*, 47(3):441–453, 2004.
- [34] M. H. Maathuis and J. A. Wellner. Inconsistency of the MLE for the Joint Distribution of Interval-Censored Survival Times and Continuous Marks. *Scandinavian Journal of Statistics*, 35(1):83–103, 2008.
- [35] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, number 141 in London Mathematical Society Lecture Note Series, pages 148–188. Cambridge University Press, 1989.
- [36] W. Pan and R. Chappell. A note on inconsistency of NPMLE of the distribution function from left truncated and case I interval censored data. *Lifetime Data Anal*, 5(3):281–91, 1999.
- [37] C. R. Rao. *Linear Statistical Inference and Its Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1973.
- [38] R. Rebolledo. Central limit theorems for local martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 51(3):269–286, 1980.
- [39] J. Rojo and F. J. Samaniego. On nonparametric maximum likelihood estimation of a distribution uniformly stochastically smaller than a standard. *Statistics & Probability Letters*, 11(3):267–271, 1991.
- [40] G. L. Silva and M. A. Amaral-Turkman. Bayesian analysis of an additive survival model with frailty. *Communications in Statistics – Theory and Methods*, 33:2517–2533, 2004.
- [41] D. Sinha and K. D. Dipak. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, 92(439):1195–1212, 1997.
- [42] D. Sinha, M. B. McHenry, S. R. Lipsitz, and M. Ghosh. Empirical bayes estimation for additive hazards regression models. *Biometrika*, 96(3):545–558, 2009.

- [43] J. Timková. Bayesian nonparametric estimation of hazard rate in survival analysis using Gibbs sampler. In *WDS 2008 Proceedings of contributed papers, Part I: Mathematics and Computer Sciences, Charles University, Prague*, pages 80–87, 2008.
- [44] J. Timková. Bayesian nonparametric estimation of hazard rate in monotone aalen model. Accepted in *Kybernetika*.
- [45] A. A. Tsiatis. A large sample study for Cox’s regression model. *The Annals of Statistics*, 9:93–108, 1981.
- [46] B. W. Turnbull. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*, 69(345):169–173, 1974.
- [47] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
- [48] M. Woodroffe. Estimating a Distribution Function with Truncated Data. *The Annals of Statistics*, 13(1):163–177, 1985.
- [49] Y. Wu and S. Ghosal. Posterior consistency for some semiparametric problems. *The Indian Journal of Statistics*, 86:114–129, 2008.
- [50] D. Zeng and D.Y. Lin. Efficient Estimation for the Accelerated Failure Time Model. *Journal of the American Statistical Association*, 102:1387–1396, 2007.
- [51] M. Zhou. Nonparametric Bayes estimator of survival functions for doubly/interval censored data. *Statistica Sinica*, 14:533–546, 2004.