Bayesian Spatial Additive Hazard Model

by

Alexander Chernoukhov

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2013

Bayesian Spatial Additive Hazard Model

by

Alexander Chernoukhov

APPROVED BY:

A. Ngom
School of Computer Science

M. Hlynka
Department of Mathematics and Statistics

A. Hussein, Advisor
Department of Mathematics and Statistics

S. Nkurunziza, Advisor
Department of Mathematics and Statistics

September 18, 2013

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

This thesis will be dealing with the problem of Bayesian estimation in additive survival data models accounting for spatial dependencies.

We consider the Aalen's additive hazards model in which baseline hazard function, the regression coefficients as well as the covariates are all allowed to be time varying processes. We incorporate in this model an extra random vector of frailties accounting for spatial variations among the observations.

Consequently, we propose a Bayesian approach to solving the inference problem for such spatial frailty model by assuming piece-wise constant structure on all time-varying functions in the model and hence, imposing appropriately chosen priors on all model parameters.

We then employ some versions of MCMC and Gibbs sampling approaches to carry out the inference about the model parameters and apply the resulting algorithm to Prostate cancer diagnosis data for the state of Louisiana, taken from the Surveillance, Epidemiology, and End Results (SEER) databases (SEER, 2008).

Acknowledgements

I would like to express my gratitude to my supervisors Dr. A. Hussein and Dr. S. Nkurunziza. I want to thank them for valuable comments and suggestions. They supported me both with choosing the research direction and overcoming the technical issues. Also they gave me the opportunities to work independently which allowed me to learn a lot of material from different fields of Mathematics and Statistics and improve the problem solving skills.

Also I want to thank University of Windsor, Department of Mathematics and Statistics and my supervisors for providing me with Research and Graduate Assistantships. This financial support allowed me to concentrate on studying and finish my program at this wonderful university.

I also want to thank my friends in Canada who supported me during the whole year and showed me the Canadian traditions and sights.

And I want to thank my wife very much for her invaluable contribution into my life. She constantly supported and helped me with any difficulties and problems which I encountered since we met. She was always there for me when I needed it and I learned a lot of important things from her. Her support allowed me to come to Canada and successfully finish my study and I believe it will help me to succeed in future.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The era of statistical modeling based on marginal analysis is almost coming to an end in the face of increasing demand to analyze complex, multidimensional and correlated streams of data that are available to investigators in real-time. Among others, methods for spatio-temporal data analysis, which requires conditional specifications taking into account the various spatial and temporal dependencies among observations, are the frontiers of the new era. In this thesis, our main objective is to develop a Bayesian method for the analysis of spatially dependent survival outcomes. Specifically, we consider Aalen's additive hazards model (Aalen, 1980) with a vector of spatial random effects through which the spatial dependencies are to be handled.

Therefore, in this chapter, we will briefly introduce the Aalen's additive model and we will review the existing literature on spatial survival models. We also set out in a more specific fashion, the objectives and organization of the thesis.

1. Literature review

Survival outcomes are particular cases in a more general context of event history outcomes. Data on event histories are usually represented as

$$\{D_i(t) = N_i(t), Y_i(t), \mathbf{z}_i(t);\ 0 \leq t \leq \tau,\ i = 1, ..., N\}, \tag{1.1.1}$$

where $\{N_i(t),\ t \in [0,\tau]\}$, is a counting process for the number of events occurring to the i-th individual in a sample of N individuals, up to time t (inclusive), $Y_i(t) = 1$ if the i-th individual is at risk of having the events of interest and zero otherwise (risk indicator function), while $\mathbf{z}_i(t)$ is time varying, p-dimensional covariate process and $[0,\tau]$ is the time frame during which subjects are observed.

In event history analysis (Andersen et al., 1993), the intensity of the counting process, $\{N_i(t),\ t \in [0,\tau]\}$, is a process defined as

$$I_i(t) = h_i(t)Y_i(t) = E[\mathrm{d}N_i(t)|F(t^-)], \tag{1.1.2}$$

where $h_i(t)$ is the hazard rate, $\mathrm{d}N_i(t) = N_i(t) - N_i(t^-)$ and $F(t^-)$ is the history of the process $\{D(t),\ t \in [0,\tau]\}$. In other words, $F(t^-)$ is a filtration of $\sigma-$algebras generated by the data $\{D(t),\ t \in [0,\tau]\}$, where both $z_i(t)$ and $Y_i(t)$ are assumed to be $F(t)$-measurable $\forall t \in [0,\tau]$.

Most of the currently available event history models are essentially models for the intensity $I_i(t)$. For instance, the celebrated Cox's proportional hazards (PH) model can be expressed as:

$$I_i(t) = \lambda(t)Y_i(t)e^{\boldsymbol{\beta}'\mathbf{z}}, \tag{1.1.3}$$

where in this case, the covariate vector $\mathbf{z}_i$ is independent of time and $\lambda(t)$ is a baseline hazard function. The Cox's PH model has been intensively studied in the literature. We refer the reader to the monograph by Andersen et al. (1993) for detailed treatment of the PH model.

Similarly, the Aalen's additive hazards (AH) model can be specified as:

$$I_i(t) = Y_i(t)(\lambda(t) + \boldsymbol{\alpha}'(t)\mathbf{z}_i(t)), \tag{1.1.4}$$

where $\boldsymbol{\alpha}(t)$ is a p-dimensional vector of time-dependent covariate functions. This was originally proposed in Aalen (1980) as an alternative to the PH model whenever the proportionality assumption is violated. There has been also an extensive literature on the AH model. A detailed account of this model can be found in Martinussen and Scheike (2006), while Hussein et al. (2013) discussed some efficient estimators for the regression coefficients in the AH model.

The spatial modeling of event history data, on the other hand, has just begun to attract attention of the statisticians. For instance, Banerjee et al. (2003) developed a Bayesian method for analysing infant mortality data via Cox's PH model with spatial frailties, while Banerjee and Dey (2005) proposed the same approach for a proportional odds model. Zhang and Lawson (2011) considered an accelerated failure time (AFT) model and proposed a Bayesian version with Gaussian frailties to handle spatial dependencies. Darmofal (2009) applied a Bayesian spatial Cox's PH model to timing of U.S. House members position announcements on the North American Free Trade Agreement (NAFTA). Among the non Bayesian models for handling spatial frailties, we mention the recent work of Lin (2012).

2. Thesis objectives and organization

As mentioned earlier, this thesis will be dealing with the problem of Bayesian estimation in additive survival data models accounting for spatial dependencies. In

general, additive survival models are flexible alternatives to the, better interpretable but more restrictive, proportional hazards models.

In this thesis we consider a very general and flexible model known as the Aalen's additive hazards model in which baseline hazard function, the regression coefficients as well as the covariates are all allowed to be time varying processes. We incorporate in this model an extra random vector $\boldsymbol{\omega}(t)$ (frailties) accounting for spatial variations among the observations. We assume that such frailties are Gausian with covariance structures of either geostatistical or conditional autoregressive (CAR) type, two well-known spatial dependence structures (see for instance Cressie and Wikle, 2011).

We propose a Bayesian approach to solving the inference problem for such additive spatial frailty models by assuming piece-wise constant structure on all time-varying functions in the models and then, imposing appropriately chosen priors on all model parameters.

We employ some variants of MCMC and Gibbs sampling approaches to carry out the inference about the model parameters. We apply the resulting method to Prostate cancer diagnosis data for the state of Louisiana, extracted from the Surveillance, Epidemiology, and End Results (SEER) databases (SEER, 2008).

As far as the author knows, this model and the Bayesian approach taken in this thesis have not been studied in the existing literature on event history analysis.

The thesis will be organized as follows.

In Chapter 2, we will set up the Additive Hazards spatial model (AHS) and obtain the joint likelihood of the data and model parameters for the case when the spatial

frailties have the (CAR) structure. We propose prior distributions for the model parameters, obtain the posteriors, and prescribe an MCMC sampling algorithms to tackle the Bayesian inferences for the model.

In Chapter 3, we will examine the case of geostatistical dependence structure for the spatial components. This case posed a huge computational roadblock, which we could not overcome. Therefore, for this case, will only explain possible priors on the parameters and prescribe a future research avenues that are possible in computing the model parameters.

In Chapter 4 we carry out a small simulation study to verify the performance of the approach and apply it to the SEER data on prostate cancer.

Chapter 5 contains the conclusions of our work.

Finally, the appendix will contain a brief review of the MCMC Gibbs sampling and Metropolis-Hastings methodologies as well as some of the technical proofs of Chapter 2 and the results of simulations.

CHAPTER 2

Additive Hazard Model with Conditional Autoregressive Spatial Structure

1. Additive hazard frailty model

In the current work, we consider an additive hazard model for spatially correlated survival data. We suppose that we have right censored left truncated survival data $\mathbf{D} = \{(N_i(t), Y_i(t), \mathbf{z}_i(t)),\ i = 1 \dots N, 0 \le t \le \tau\}$ from N individuals where $\{N_i(t),\ t \in [0, \tau]\}$ is the counting process of the events happened to the i-th individual, and $\{Y_i(t),\ t \in [0, \tau]\}$ is at-risk process for the i-th individual:

$$Y_i(t) = \begin{cases} 1, & \text{if the } i\text{-th individual is at risk at time } t, \\ 0, & \text{otherwise (dead, censored, truncated, etc).} \end{cases} \tag{2.1.1}$$

The process $\{\mathbf{z}_i(t) = (z_{i1}(t), \dots, z_{ip}(t))^T,\ \mathbf{z}_i(t) \in \mathbf{\Omega},\ i = 1, \dots, N,\ t \in [0, \tau]\}$ representing p time dependent covariates, where $\mathbf{X}^T$ denotes the transpose of $\mathbf{X}$, and $\mathbf{\Omega} \subset \mathbb{R}^p$ is the set of all admissible covariate vectors. Each individual belongs to a certain region $l_i \in \{1, \dots, n\}$ with the total number of regions $n \le N$. The model considered is the extension of the usual Aalen's additive hazard model by including additive, region specific and time dependent, frailty terms $\omega_l(t),\ l = 1 \dots n$.

More specifically, in our model, the hazard $h_i(t)$ of the i-th individual can be expressed as:

$$h_i(t) = \lambda(t) + \sum_{k=1}^{p} \alpha_k(t) z_{ik}(t) + \omega_{l_i}(t), \quad 0 \le t \le \tau, \tag{2.1.2}$$

where τ is the end of study, $\lambda(t)$ is the "baseline hazard", $\alpha_k(t),\ k = 1, \ldots, p$ are time dependent regression coefficients (regression functions), and $\omega_{l_i}(t)$ is a random group specific frailty term for the group l_i to which the i-th individual belongs.

Note that for this model to be correctly specified we should ensure that $h_i(t),\ i = 1, \ldots, N$ are non-negative functions for all $t \in [0, \tau]$. Also it is worth mentioning that the "baseline hazard" $\lambda(t)$ need not be non-negative since it doesn't necessarily represent the hazard of any individual in the population (see Klein and Moeschberger, 2003). Formally, it represents the hazard of a hypothetical "individual" with all covariates $z_{0k}(t)$ set to 0 and null frailty. But for some ways of coding the covariates and frailty, zero values can not make any sense, and therefore the "baseline hazard" can not be interpreted as the hazard of any individual. For example, if the covariate represents the age of the individual plus some value (say, 10 years), then setting this covariate to 0 means that the age of such individual is negative (-10 years). So in this case the "baseline hazard" is not actually the hazard, but only some reference function.

In order for $\lambda(t)$ to be interpretable, one can shift all the covariate values by some number (e.g. by the mean value of the covariate) so that the individual with zero covariates could be really an individual from the population. In this case, $\lambda(t)$ should be always non-negative.

Hereinafter, we suppose that the covariates are coded in such a way that $\lambda(t)$ can be interpreted as the hazard of some individual.

The intensities $\{I_i(t),\ t \in [0,\tau]\}$ of the counting processes $\{N_i(t),\ t \in [0,\tau]\}$ of the individuals can be written as follows (see Silva and Amaral-Turkman, 2004):

$$I_i(t) = Y_i(t)h_i(t) = Y_i(t)\left(\lambda(t) + \sum_{k=1}^{p} \alpha_k(t)^T z_{ik}(t) + \omega_{l_i}(t)\right). \tag{2.1.3}$$

Assuming that all observations are independent, the likelihood of the data $\mathbf{D}$ given baseline hazard $\lambda(t)$, regression function vector $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_p(t))^T$ and frailties vector $\boldsymbol{\omega}(t) = (\omega_1(t), \ldots, \omega_n(t))^T$, is proportional to:

$$L\left(\mathbf{D} \mid \lambda(t), \boldsymbol{\alpha}(t), \boldsymbol{\omega}(t)\right) \propto \prod_{i=1}^{N}\left[\left(\prod_{0<t\leq\tau} I_i(t)^{\mathrm{d}N_i(t)}\right)\exp\left(-\int_0^\tau I_i(u)\mathrm{d}u\right)\right], \tag{2.1.4}$$

where

$$\mathrm{d}N_i(t) = \lim_{\mathrm{d}t\to 0^+}(N_i(t) - N_i(t-\mathrm{d}t)), \tag{2.1.5}$$

is the number of events of the i-th individual at time t, and the product $\prod_{0<t\leq\tau}(\ldots)$ is the product-integral, assuming $0^0 \equiv 1$.

We consider the model where each individual can have only 0 or 1 events. So $\mathrm{d}N_i(t) = 0$ or $\mathrm{d}N_i(t) = 1$ for all individuals and all t. Since $I_i(t)$ is non-zero only when the i-th individual is at risk, (2.1.4) can be rewritten as:

$$L\left(\mathbf{D} \mid \lambda(t), \boldsymbol{\alpha}(t), \boldsymbol{\omega}(t)\right) \propto \prod_{i\in\mathcal{E}} h_i(T_i)\prod_{i=1}^{N}\exp\left(-\int_{\{t:Y_i(t)=1\}} h_i(t)\mathrm{d}t\right), \tag{2.1.6}$$

where $\mathcal{E}$ is the set of all individuals having events during the study period, and T_i is the event time of the i-th individual in $\mathcal{E}$. That is,

$$\mathcal{E} = \{i : N_i(\tau) = 1\}, \tag{2.1.7}$$

$$T_i = \inf\{t : N_i(t) = 1\}, \quad i \in \mathcal{E}. \tag{2.1.8}$$

2. Model specification

2.1. Ensuring non-negativity of the hazard. In Bayesian implementation we put prior distributions on the parameters of the model, i.e. on $\lambda(t)$, $\alpha_k(t)$, $k = 1, \ldots, p$ and $\omega_l(t)$, $l = 1, \ldots, n$. Assuming that there is no prior knowledge about the parameters, we make all the prior distributions vague.

Note that the hazard rate $h(t)$ should be always non-negative.

One approach to ensure non-negativity of $h(t)$ (Silva and Amaral-Turkman, 2004) is to choose prior distributions such that all the parameters of the model are non-negative. This means that baseline, all the covariates, regression functions and frailty terms are not allowed to be negative. This approach assumes that all covariates have positive effect on the hazard, or that the covariates are transformed in a special way. Such assumption is inappropriate if a certain covariate has a negative effect on the survival function.

Another approach given in Cai and Zeng (2011), estimates all the parameters without accounting for the negativity issue, and then modifies the estimator of the survival function in such way that it becomes always non-increasing. Cai and Zeng (2011) mentioned that if non-modified estimator of the survival function is consistent, the modified estimator will also be consistent. While this approach ensures that the survival function estimator is non-increasing function, the actual estimators of the coefficients are not interpretable because the hazard becomes negative.

In this work, we use a more flexible approach. Firstly, we will introduce the prior distributions separately, ignoring the issue of hazard negativity, and only after that

we will constrain the joint distribution of $\lambda(t)$, $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\omega}(t)$ to the region where the cumulative hazard is non-negative for all admissible covariate vectors from $\mathbf{\Omega}$, i.e.

$$h(t) = \lambda(t) + \boldsymbol{\alpha}(t)^T \mathbf{z}(t) + \omega_l(t) \geq 0, \\ \forall \mathbf{z}(t) \in \mathbf{\Omega}, \quad \forall t : 0 < t \leq \tau, \quad \forall l = 1, \ldots, n. \tag{2.2.1}$$

This means that the marginal distributions of the parameters are not exactly the distributions we are introducing but the components of the joint distribution. Since we need only joint distribution and all full conditional distributions of the parameters, we will not consider the marginal distributions at all.

Now, provided that in (2.2.1) the set of admissible covariate vectors $\mathbf{\Omega}$ can be expressed as a Cartesian product of p sets $\mathbf{\Omega} = \Omega_1 \times \cdots \times \Omega_p$ with all $\Omega_i \subset \mathbb{R}$ being the bounded subsets of the real numbers, the conditions above can be rewritten as:

$$\lambda(t) + \sum_{k=1}^{p} \inf_{z \in \Omega_k} \{\alpha_k(t) z\} + \min_{1 \leq l \leq n} \{\omega_l(t)\} \geq 0, \quad \forall t \in (0, \tau]. \tag{2.2.2}$$

Note that depending on the sign of $\alpha_k(t)$ the infimum inside the summation in the expression above is either $\alpha_k(t) \inf \Omega_k$ or $\alpha_k(t) \sup \Omega_k$. Then we can rewrite the constraint in the following form:

$$\lambda(t) + \sum_{k=1}^{p} \min \left\{ \alpha_k(t) \inf \Omega_k, \alpha_k(t) \sup \Omega_k \right\} + \min_{1 \leq l \leq n} \{\omega_l(t)\} \geq 0, \quad \forall t \in (0, \tau]. \tag{2.2.3}$$

This constraint will be included in the joint distribution of the parameters which will be introduced later.

2.2. Partitioning of time. In our model, we estimate all the parameters as piecewise constant functions, i.e. functions constant in the intervals $(t_0, t_1]$, $(t_1, t_2]$, $\ldots$, $(t_{m-1}, t_m]$ where $t_0, \ldots, t_m$ is a finite set of time points such that $0 = t_0 < t_1 < \cdots < t_m = \tau$. The length of the j-th interval is $\Delta t_j = t_j - t_{j-1}$ for $1 \leq j \leq m$.

In this case, each parameter function can be considered as a finite number of scalar parameters. The choice of the points t_i as well as number of these points m is arbitrary. However, one should take into account that the wider the intervals are, the worse is the approximation of the parameter functions, but at the same time if the intervals are very narrow, the data does not provide enough information to accurately estimate the parameters in these intervals. So the width of the intervals and their number should be chosen as a trade-off between the above mentioned problems.

For the case of equidistant time points t_j, the choice of these points reduces to the choice of their number m. This can be done by using the Bayesian model comparison criteria such as DIC or $LCPO$ which will be discussed later.

After time partitioning, we define:

$$\lambda_j \equiv \lambda(t_j), \quad \alpha_{kj} \equiv \alpha_k(t_j), \quad z_{ikj} \equiv z_{ik}(t_j), \quad \omega_{lj} \equiv \omega_l(t_j), \tag{2.2.4}$$

which can be compacted as follows:

$$\boldsymbol{\lambda} \equiv (\lambda_j)_{j=1,\dots,m}, \tag{2.2.5}$$

$$\boldsymbol{\alpha} \equiv (\alpha_{kj})_{k=1,\dots,p,\ j=1,\dots,m}, \tag{2.2.6}$$

$$\mathbf{z} \equiv (z_{ikj})_{i=1,\dots,N,\ k=1,\dots,p,\ j=1,\dots,m}, \tag{2.2.7}$$

$$\boldsymbol{\omega} \equiv (\omega_{lj})_{l=1,\dots,n,\ j=1,\dots,m}. \tag{2.2.8}$$

These parameters fully represent the original time-dependent parameter functions under the assumption of piecewise constancy.

2.3. Conditional autoregressive structure for the frailties. The distribution of the frailty parameters should incorporate the spatial structure of the data.

The one way of doing this is by using the distances between the regions to determine the correlation between frailties of the regions, resulting in the so called geostatistical model. Another approach, conditional autoregression (CAR), uses adjacency structure of the regions instead of the distances.

As was already mentioned, in this chapter we discuss only the CAR model. As regards the geostatistical model, we present the prior, posterior and proposal distributions for it in Chapter 3 without further numerical analysis.

The conditional autoregressive (CAR) structure for frailties allows to take into account the spatial correlation of data based on the adjacency structure of the regions.

To introduce the CAR structure, we assume that $\boldsymbol{\omega}_j$ is independent of all $\boldsymbol{\omega}_{j' \neq j}$. This allows us to introduce the spatial correlation in each time interval independently.

Following Banerjee et al. (2003) and Zhang and Lawson (2011), we consider a conditional autoregressive model (CAR), and particularly the model with the following prior joint distribution of frailties:

$$\pi(\boldsymbol{\omega}_j|\theta_j^2) \propto \frac{1}{\left(\theta_j^2\right)^{n/2}} \exp\left(-\frac{1}{2\theta_j^2} \sum_{\substack{l \sim l' \\ l<l'}} \left(\omega_{lj} - \omega_{l'j}\right)^2\right), \tag{2.2.9}$$

where $l \sim l'$ denotes the adjacency relation between l and l', and condition $l < l'$ ensures that each pair of adjacent regions is included in the summation only once.

This prior was used in Besag et al. (1991) under the name of Gaussian intrinsic autoregression mainly in application to image restoration. It is a particular case of pairwise-difference priors (see Besag et al., 1995). They note that such priors are improper since the corresponding variables are of an arbitrary level, and only their

differences are taken into account. But this impropriety is removed from the posterior distribution by the presence of any informative data.

Although the prior itself is improper, the conditional distribution of any frailty given all others is well defined, and is then proportional to:

$$\pi\left(\omega_{lj} \mid \omega_{l'j\neq lj}, \theta_j^2\right) \propto \frac{1}{\theta_j} \exp\left(-\frac{1}{2\theta_j^2} m_l \left(\omega_{lj} - \overline{\omega}_{lj}\right)^2\right), \tag{2.2.10}$$

where $m_l = \text{card}\left\{l' \mid l' \sim l\right\}$ is the number of regions adjacent to the l-th and $\overline{\omega}_{lj} = \frac{1}{m_l}\sum_{l'\sim l}\omega_{l'j}$ is the average of the frailties adjacent to the l-th, which means that conditionally, the frailties are normally distributed with mean $\overline{\omega}_{lj}$ and variance θ_j^2/m_l:

$$\left(\omega_{lj} \mid \omega_{l'j\neq lj}, \theta_j^2\right) \sim \mathcal{N}\left(\overline{\omega}_{lj}, \frac{\theta_j^2}{m_l}\right). \tag{2.2.11}$$

Further details on the conditional and intrinsic autoregression can be found in Besag and Kooperberg (1995).

Although, in combination with the data likelihood and baseline hazard prior, the joint posterior becomes proper, since the frailties are defined only up to an additive hazard, the data cannot distinguish which part of the hazard ascribe to the baseline and which to the frailties. So this distinguishing relies only on the prior distributions of the baseline and frailties.

However, if the priors are vague as in our case, the estimation of the frailties and baseline can have a very large variance since nothing really prevents the frailties from being greater than the actual ones by some value while keeping baseline less by the same value.

In order to decrease possible variance in estimation, one can make the prior distribution of the frailties proper by including the terms containing the values of frailties themselves in addition to their differences. For instance, one can include the squares of the frailties multiplied by some coefficients, in which case the joint prior distribution of the frailties takes the following form:

$$\pi(\boldsymbol{\omega}_j|\theta_j^2,\varepsilon) \propto \frac{1}{\left(\theta_j^2\right)^{n/2}} \exp\left(-\frac{1}{2\theta_j^2}\sum_{\substack{l\sim l'\\ l<l'}}(\omega_{lj}-\omega_{l'j})^2 - \frac{\varepsilon}{2\theta_j^2}\sum_{l=1}^{n}\omega_{lj}^2\right). \tag{2.2.12}$$

Such prior will shrink the frailties towards 0 because of the presence of the pure square terms. If we suspect that values of the frailties are concentrated not around 0 but around some value μ, then we should include $(\omega_{lj}-\mu)^2$ instead of ω_{lj}^2. This will shrink the frailties towards μ. The parameter ε represents the amount of shrinkage: the greater it is, the more the frailties are shrunk towards μ.

Now, the conditional distributions of the frailties in the case of $\mu = 0$ will take the following form:

$$\pi\left(\omega_{lj} \mid \omega_{l'j\neq lj}, \theta_j^2, \varepsilon\right) \propto \frac{1}{\theta_j} \exp\left(-\frac{m_l+\varepsilon}{2\theta_j^2}\left(\omega_{lj} - \frac{m_l}{m_l+\varepsilon}\overline{\omega}_{lj}\right)^2\right), \tag{2.2.13}$$

which means that frailties are conditionally normally distributed:

$$\left(\omega_{lj} \mid \omega_{l'j\neq lj}, \theta_j^2, \varepsilon\right) \sim \mathcal{N}\left(\frac{m_l}{m_l+\varepsilon}\overline{\omega}_{lj}, \frac{\theta_j^2}{m_l+\varepsilon}\right). \tag{2.2.14}$$

We can see that the less the parameter ε is, the closer this distribution is to the conditional autoregressive model and they become the same if $\varepsilon = 0$.

Another way to deal with impropriety of frailties' prior is to exclude baseline hazard from the model and include its effect in the frailty terms. In this case, we can use (2.2.9) directly without additional parameters, since the frailties are estimable

from the data. This approach is more suitable when one does not know the level of frailties and wants to rely in estimation on the data rather than on the prior distributions.

Although, in this case the baseline and frailties are combined into frailty terms only, hence not distinguishable, this does not affect the estimation of regression functions. Moreover, if the frailties are considered not as random effects but as fixed effects depending on the regions to which the observations belong, exclusion of the baseline does not affect the prediction problem: the hazard of any individual with known covariates and region can be estimated based on the resulting values of the regression functions and frailties.

On hte other hand, if the frailties are considered random effects, this approach does not work since the estimators of the hazard in this case should not depend on the frailties values. So one should use the modified prior distribution like in (2.2.12), or make some additional assumptions about the frailty terms.

In this work, we choose another approach. We assume that the first frailty term is equal to 0, and therefore the first region is the reference level. This means that the baseline is interpreted as the hazard of an individual from the first region with all covariates equal to 0. This assumption eliminates the problem of identifiability of baseline and frailties since observations from the first region have the known value of frailty (zero) and hence allow estimation of the baseline.

The joint prior distribution in this case can be expressed as follows:

$$\pi(\boldsymbol{\omega}_j|\theta_j^2) \propto \frac{1}{\left(\theta_j^2\right)^{(n-1)/2}} \exp\left(-\frac{1}{2\theta_j^2}\sum_{\substack{l\sim l'\\ l<l'}}(\omega_{lj}-\omega_{l'j})^2\right)\delta(\omega_{1j}), \tag{2.2.15}$$

where $\delta(x)$ is the delta-function representing the point-mass at 0 for the first frailty, and the power of the parameter θ_j^2 is changed to $(n-1)/2$ since one frailty is fixed and only $n-1$ of them are included in the CAR distribution.

In order to simplify the notation, we denote the set of all $\theta_j,\ j=1,\ldots,m$ by $\boldsymbol{\theta}$.

2.3.1. *Prior distribution of the CAR model's hyper-parameters.* If the CAR spatial model is used, then we set the prior for the set of hyper-parameters θ_j^2. In order to make the prior conjugate, we take the inverse-gamma prior $\mathcal{IG}(\beta,\gamma)$ for each θ_j with common β and γ. The mean in this case is $\frac{\beta}{\gamma-1}$ and variance is $\frac{\beta^2}{(\gamma-1)^2(\gamma-2)}$. Making $\gamma=2$ we can make this prior vague which is provided by the infinite variance.

Also, we assume that all θ_j^2 are independent, and so their joint distribution is the product of marginal distributions.

2.4. Prior distribution for the baseline hazard. We assume that the prior distributions of the values of baseline hazard in different intervals $\lambda_1,\ldots,\lambda_j$ are independent Gamma distributions with shape and scale parameters $r_0c_0\Delta t_j$ and $1/(c_0\Delta t_j)$, respectively:

$$\lambda_j \sim \mathcal{G}\left(r_0c_0\Delta t_j, \frac{1}{c_0\Delta t_j}\right), \quad 1\le j\le m, \quad r_0>0, \quad c_0>0, \tag{2.2.16}$$

where $\mathcal{G}(a,b)$ denotes the Gamma distribution with shape parameter a and scale parameter b.

The value r_0 represents the "best gues" for the baseline hazard at each interval, and the value c_0 represents the "confidence" in this "guess". Such interpretation of r_0 and c_0 follows from the fact that mean of the above Gamma distribution is r_0 ("best guess") and the variance is $r_0/(c_0\Delta t_j)$ which decreases when c_0 increases ("confidence" in the "best guess"). If the baseline hazard is not known, one should set c_0 very small to make the prior vague.

Such choice of the prior distribution for the baseline hazard is the discretized version of the Gamma-process prior for the cumulative baseline hazard $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$ as in Silva and Amaral-Turkman (2004).

The Gamma process with parameter function $\Lambda^*(t)$ representing the "best guess" for $\Lambda(t)$ and scalar parameter c_0 representing the "confidence" in this "guess" is defined as follows. For any partitioning of the time axis $0 = t_0 < t_1 < \cdots < t_m < \infty$ the increments $\Delta\Lambda(t_j) = \Lambda(t_j) - \Lambda(t_{j-1}),\ j = 1, \ldots, m$ of the cumulative baseline hazard are mutually independent random variables each following Gamma distribution with shape parameter $c_0\Delta\Lambda^*(t_j)$ and scale parameter $1/c_0$. So, the mean of $\Delta\Lambda(t_j)$ is $\Delta\Lambda^*(t_j)$ and the variance is $\frac{\Delta\Lambda^*(t_j)}{c_0}$.

Now if the parameter function for the Gamma process takes the form $\Lambda^*(t) = r_0 t$, where r_0 represents the "best guess" for the $\lambda(t)$ (constant over time), and if we fix the partitioning of time and assume that the baseline hazard is piecewise constant, we obtain the prior distribution given in (2.2.16). Also since the increments are independent, the corresponding values of $\lambda(t_j) = \Delta\Lambda/\Delta t_j$ are also independent. So we obtain the independent Gamma distributions given at the beginning of this subsection.

Since the baseline hazard distributions at different intervals are independent, the joint distribution of them can be found as product of the marginal distributions, and so it is proportional to:

$$\pi(\boldsymbol{\lambda}) \propto \prod_{j=1}^{m} \lambda_j^{r_0 c_0 \Delta t_j - 1} \exp\left(-\lambda_j c_0 \Delta t_j\right), \quad \lambda_j > 0, \quad \forall j = 1, \ldots, m. \tag{2.2.17}$$

As was already mentioned before, this is not exactly the joint prior distribution of the parameters $\lambda_1, \ldots, \lambda_m$, but rather one component of the constrained joint prior distribution of all the parameters in the model which will be introduced later.

2.5. Prior distribution of regression functions. Following Banerjee et al. (2003) we put flat (improper uniform) priors on the regression functions. This is a common practice and we adhere to it. So the joint prior distribution of the regression functions is proportional to 1:

$$\pi(\boldsymbol{\alpha}) \propto 1. \tag{2.2.18}$$

Alternatively, one can consider Gaussian priors. These can be, for example, constructed as a discretized versions of Wiener processes for the cumulative regression functions $A_k(t) = \int_0^t \alpha_k(u)\mathrm{d}u$ similarly to constructing the prior distribution for the baseline hazard. Then, the marginal distributions of α_{kj} in this case are normal with mean 0 and variance $\sigma_k^2/\Delta t_j$.

2.6. Joint prior distribution. The joint constrained prior of the parameters can be obtained by multiplication of all components introduced earlier and an indicator function representing the constraints.

Note that constraining the prior changes the marginal distributions of λ_j and ω_{lj} discussed above and introduces the dependency among baseline λ_j, frailties ω_{lj} and regression functions α_{kj}. So, rigorously speaking, we should have introduced the joint prior of all parameters of the model directly without discussing the marginal prior distributions of the parameters. However we decided to talk about the marginal components first in order to explain the choice of the joint prior.

For the CAR model, the joint prior distribution is:

$$\begin{aligned}\pi\left(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta}\right) \propto & \prod_{j=1}^{m} \underbrace{\left(\lambda_j^{c_0 r_0 \Delta t_j - 1} \exp\left(-c_0 \lambda_j \Delta t_j\right)\right)}_{\text{Gamma prior for } \lambda_j} \\ & \times \prod_{j=1}^{m} \underbrace{\left(\frac{1}{\left(\theta_j^2\right)^{(n-1)/2}} \exp\left(-\frac{1}{2\theta_j^2} \sum_{\substack{l \sim l' \\ l<l'}} \left(\omega_{lj} - \omega_{l'j}\right)^2\right) \delta(\omega_{1j})\right)}_{\text{Conditional autoregressive prior for } \boldsymbol{\omega}_j = (\omega_{1j}, \ldots, \omega_{nj})^T} \\ & \times \prod_{j=1}^{m} \underbrace{\left(\mathbb{I}\left\{\lambda_j + \sum_{k=1}^{p} \min\left\{\alpha_{kj} \inf \Omega_k, \alpha_{kj} \sup \Omega_k\right\} + \min_{1 \le l \le n} \omega_{lj} \ge 0\right\}\right)}_{\text{Constraint component}} \\ & \times \prod_{j=1}^{m} \underbrace{\left(\left(\theta_j^2\right)^{-\gamma-1} \exp\left(-\frac{\beta}{\theta_j^2}\right)\right)}_{\text{Inverse-gamma prior for } \theta_j^2}, \end{aligned} \quad (2.2.19)$$

where $\mathbb{I}\{E\}$ denotes the indicator function of the event E. Also, here we used the conditional autoregressive prior for the frailties with additional assumption that the first frailty term is equal to 0.

3. Posterior distribution

In Bayesian analysis, all the information about the parameters of the model is contained in their posterior distribution. So, in order to make inference within the Bayesian framework, the main goal is to find this posterior distribution and compute the necessary quantities using it.

In our proposed model, it is very hard to find the posterior distribution in explicit form. However, we can approximate this distribution by using Markov Chain Monte Carlo (MCMC) method.

Firstly, we need to derive the joint distribution of the data and the parameters which can be easily obtained by simply multiplying the data likelihood and the joint prior distribution of the parameters.

With the assumption of piecewise constant parameters, the likelihood in the formula (2.1.6) becomes:

$$L\left(\mathbf{D} \mid \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}\right) \propto \prod_{j=1}^{m} \left[\left(\prod_{i \in \mathcal{E}_j} h_i(t_j) \right) \exp\left(- \sum_{i=1}^{N} R_{ij} h_i(t_j) \Delta t_j \right) \right], \tag{2.3.1}$$

where $\mathcal{E}_j$ denotes the set of all individuals having events in the interval $(t_{j-1}, t_j]$, and R_{ij} is the proportions of time the i-th individuals is at risk in the interval $(t_{j-1}, t_j]$:

$$\mathcal{E}_j = \{i : N_i(t_j) - N_i(t_{j-1}) = 1\}, \tag{2.3.2}$$

$$R_{ij} = \frac{1}{\Delta t_j} \int_{t_{j-1}}^{t_j} Y_i(t) \mathrm{d}t. \tag{2.3.3}$$

Now, for the CAR model, we can get the joint distribution of the data and parameters (up to normalizing constant) multiplying the expressions given by the formulas

(2.2.19) and (2.3.1):

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta}, \mathbf{D}) \propto L\left(\mathbf{D} \mid \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}\right) \pi\left(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta}\right), \tag{2.3.4}$$

where the collection of all the covariate vectors of individuals $\mathbf{z}$ is considered known.

Note that given the data, the posterior distribution of the parameters is proportional to the joint distribution of the data and parameters:

$$\pi\left(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta} \mid \mathbf{D}\right) = \frac{\pi(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta}, \mathbf{D})}{\pi(\mathbf{D})} \propto \pi(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta}, \mathbf{D}). \tag{2.3.5}$$

Then, the posterior distribution can be obtained by multiplying the joint prior distribution and the likelihood given by equations (2.2.19) and (2.3.1), respectively:

$$\pi\left(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta} \mid \mathbf{D}\right) \propto L\left(\mathbf{D} \mid \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}\right) \pi\left(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\theta}\right). \tag{2.3.6}$$

The distribution given by (2.3.6) is very hard to work with: it is not easy to find the normalizing constant, mean, quantiles and any other quantities of interest. However, we can approximate these quantities by their sample values obtaining a sufficiently large sample from the posterior distribution. So the problem is how to generate this sample. This issue will be discussed in the following section.

4. Obtaining random sample from the posterior distribution

We will use the Markov Chain Monte Carlo (MCMC) algorithm called Metropolis-within-Gibbs to sample from the joint posterior distribution given by the equation (2.3.6). The details regarding this algorithm adopted to our purposes are presented in Appendix A.

The basic idea of all MCMC algorithms is to construct a Markov chain with the limiting distribution equal to the desired posterior distribution. Metropolis-within-Gibbs, particularly, updates the parameters one-by-one using the full conditional distributions. If the full conditional distributions are not standard, the sampling is performed from the so-called proposal distributions instead of real conditionals and the algorithm adjusts for the differences in these distributions by itself in order to obtain the correct limiting distribution.

The parameters of interest for sampling are the baseline hazard λ_j, regression functions α_{kj} and hyper-parameters θ_j for CAR model. The frailties ω_{lj} can be considered either as the parameters of interest along with the previous ones or as nuisance parameters. The nuisance parameters are usually integrated out from the joint distribution. However, in our case, such integration is very difficult to carry out, so we sample frailties along with all other parameters.

The full set of parameters to be sampled consists of m parameters for the baseline hazard $\{\lambda_j\}_{j=1}^m$, m parameters for each of p regression functions $\left\{\{\alpha_{kj}\}_{k=1}^p\right\}_{j=1}^m$, m parameters for each of n frailty terms $\left\{\{\omega_{lj}\}_{l=1}^n\right\}_{j=1}^m$, and m hyper-parameters $\{\theta_j\}_{j=1}^m$ for CAR model, which form $m(1+p+n+1)$ parameters in total.

For MCMC algorithm to work better, on each step we need to find the proposal density close to the conditional density or at least similar in shape (see details in Appendix A). At the same time, this proposal should be simple enough to allow direct sampling from it. In the following subsections we will investigate the conditional densities and offer the proposals satisfying the mentioned requirements.

4.1. Sampling from the baseline hazard's full conditional distribution. The baseline hazard is represented by m components $\lambda_1, \ldots, \lambda_m$.

Let E_j be the number of events in the interval $(t_{j-1}, t_j]$, and $R_j = \sum_{i=1}^{N} R_{ij}$ be the summation of proportions of time all individuals are at risk in this interval. Furthermore, suppose that individuals having events in this interval have the indices $i = i_1, \ldots, i_{E_j}$ in the original dataset.

Denote also $\Gamma(\rho) = \int_0^\infty \xi^{\rho-1} \exp(-\xi) \mathrm{d}\xi$ (gamma-function), $\rho = c_0 r_0 \Delta t_j$, $\varepsilon = \frac{1}{(c_0+R_j)\Delta t_j}$, $c_s = \sum_{k=1}^{p} \alpha_{kj} z_{i_s kj} + \omega_{l_{i_s}}$ and the constraint a computed as:

$$a = -\sum_{k=1}^{p} \min \left\{ \alpha_{kj} \inf \Omega_k, \alpha_{kj} \sup \Omega_k \right\} - \min_{1 \le l \le n} \omega_{lj}. \tag{2.4.1}$$

Then the conditional distribution of λ_j, fixing all other parameters and data, has the form given by the following proposition.

PROPOSITION 4.1. *The distribution of λ_j conditional on all other parameters is a constrained Polynomial-Gamma distribution whose probability density function apart from normalizing constant has the following form:*

$$f_{\lambda_j}(x) \propto \prod_{s=1}^{E_j} (x + c_s) \frac{1}{\varepsilon^\rho \Gamma(\rho)} x^{\rho-1} \exp\left(-\frac{x}{\varepsilon}\right) \mathbb{I}\{x > a\}, \ x > 0. \tag{2.4.2}$$

PROOF. The proof of this proposition is given in Appendix B. □

4.1.1. *Finding the proposal distribution using the mean of the full conditional.* One way of finding suitable proposal distribution is based on locating the mean of the full conditional stated in Proposition 4.1. This requires finding the normalizing constant of this distribution. Fortunately, the normalizing constant and the mean can be found explicitly, as stated in the proposition below.

Let d_f, $f = 0, \ldots, E_j$ be the coefficients of the polynomial $\sum_{f=0}^{E_j} d_f x^f$ obtained by the expansion of the product $\prod_{s=1}^{E_j}(x + c_s)$, and let

$$\mathcal{I}_f = \varepsilon^f \frac{\Gamma(\rho + f)}{\Gamma(\rho)} \left(1 - \frac{\gamma\left(\frac{\max[a,0]}{\varepsilon}, \rho + f\right)}{\Gamma(\rho + f)} \right), \tag{2.4.3}$$

where $\gamma(x, \rho) = \int_0^x \xi^{\rho-1} \exp(-\xi) \mathrm{d}\xi$ is the lower incomplete gamma function.

PROPOSITION 4.2. *The normalizing constant C_{norm} and the mean μ of the distribution stated in Proposition 4.1 can be expressed as follows:*

$$C_{norm} = \sum_{f=0}^{E_j} d_f \mathcal{I}_f, \tag{2.4.4}$$

$$\mu = \frac{\sum_{f=0}^{E_j} d_f \mathcal{I}_{f+1}}{\sum_{f=0}^{E_j} d_f \mathcal{I}_f}. \tag{2.4.5}$$

PROOF. We leave the proof of this proposition to Appendix B. □

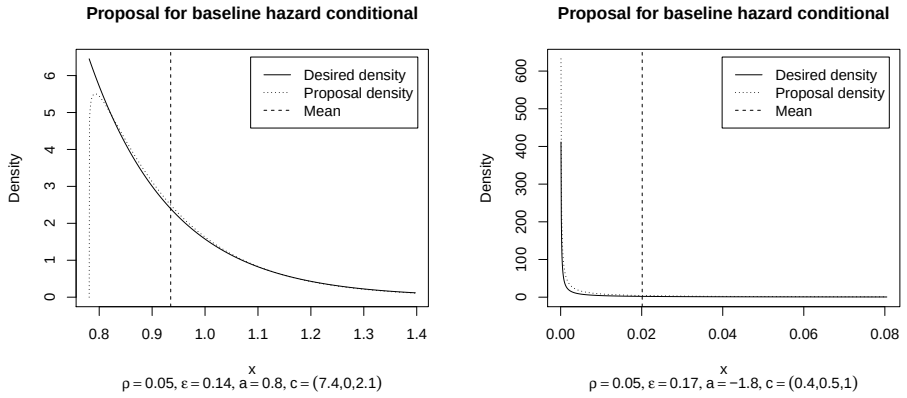
From this point we can introduce the proposal which is close to the distribution discussed above. As a proposal we are going to use the gamma distribution whose origin is moved to $\max[a, 0]$. We leave the scale parameter of this new distribution equal to ε and we adjust the shape parameter in such a way that the mean of the proposal distribution is equal to the mean μ of the actual distribution $f_{\lambda_j}(x)$ derived above. Given the shape parameter ν, the mean of the proposal distribution is $(\max[a, 0] + \nu\varepsilon)$. Then $\nu = \frac{\mu - \max[a,0]}{\varepsilon}$ will provide the desired mean μ.

In the case, when the real distribution $f_{\lambda_j}(x)$ is exactly Gamma, the proposal distribution is equal to it, otherwise they are hopefully close enough to each others.

The Figures 1a–1h illustrate how the proposal distribution is close to the distribution given by (B.1.2). The parameter ρ is fixed at 0.05 by setting $\Delta t_j = 1$, $c_0 = 0.01$

and $r_0 = 5$. The power E_j of the polynomial is chosen randomly from the Poisson distribution, the constraint a is drawn from the normal distribution, and the coefficients c_s are generated from gamma and then shifted by $-a$ which ensures that all $c_s \geq -a$. The scale parameter ε is calculated as $\varepsilon = \frac{1}{(c_0+R_j)\Delta t_j}$ with R_j generated from Poisson distribution.

FIGURE 1. Comparing real and proposal distribution for sampling from the full conditional of the baseline

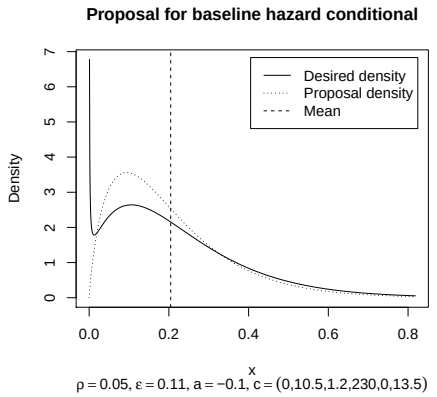


(a) (b)

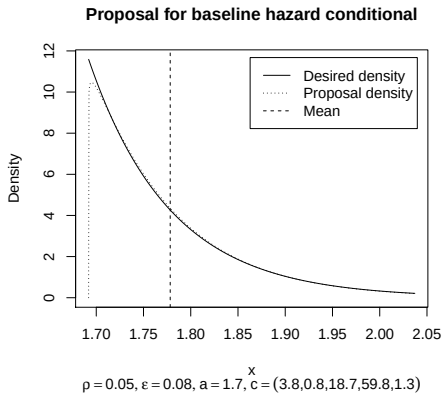
We see that proposal distribution follows the form of a desired distribution well. The main feature to notice is that localization regions of both distributions are the same, which is needed for good convergence of the MCMC algorithm. This shows us that the proposals are satisfactory.

However, this method of finding the proposal distribution has one hidden drawback. The expansion of the polynomial $P(x) = \prod_{s=1}^{E_j}(x + c_s)$ in (B.1.2) to the form of $\sum_{f=0}^{E_j} d_f x^f$ requires the evaluation of 2^{E_j} terms, so the computational complexity of this operation grows extremely fast with the number of events E_j.

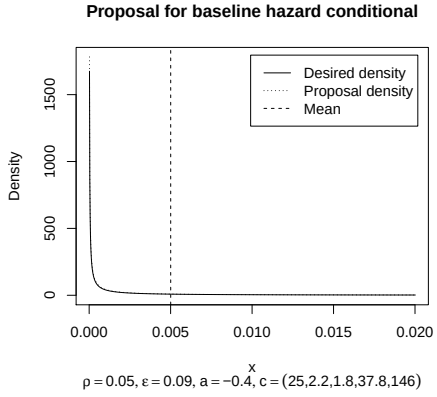
FIGURE 1. Continuation: Comparing real and proposal distribution for sampling from the full conditional of the baseline



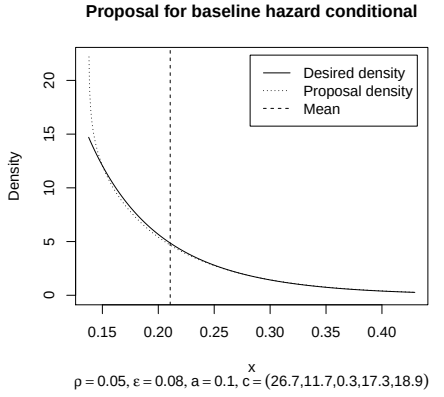
(c)



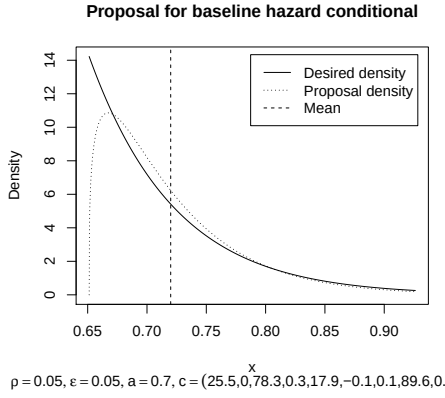
(d)



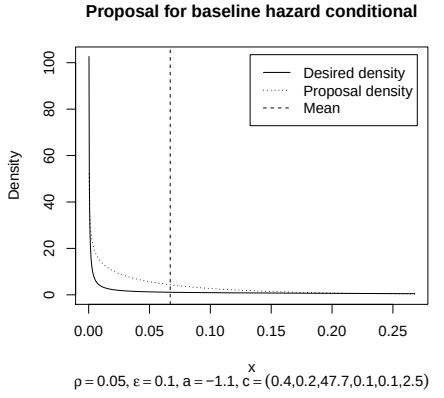
(e)



(f)



(g)



(h)

So, for the case of large E_j instead of direct expansion we recommend using the following method. The coefficient d_0 can be easily obtained by simple multiplication:

$$d_0 = \prod_{s=1}^{E_j} c_s, \tag{2.4.6}$$

and the coefficient d_{E_j} is equal to 1. Regarding the rest $E_j - 1$ coefficients, we can evaluate the polynomial $P(x)$ at $E_j - 1$ different points $x_1, \ldots, x_{E_j-1}$ and find these coefficients by solving the linear system of equations:

$$\begin{cases} \sum_{s=1}^{E_j-1} d_s x_1^s &= P(x_1) - x_1^{E_j} - d_0, \\ & \ldots \\ \sum_{s=1}^{E_j-1} d_s x_{E_j-1}^s &= P(x_{E_j-1}) - x_{E_j-1}^{E_j} - d_0. \end{cases} \tag{2.4.7}$$

Note that value $x = 0$ should not be used for any of the points x_s to avoid the presence of noninformative equation $0 = 0$.

This method of polynomial expansion is much faster than the direct expansion but suffers from numerical instability because of the presence of high powers of x_s. So for large E_j the precision provided by the default floating point variable type in most of the mathematical packages and programming languages is not enough to obtain satisfactorily precise values of d_s. So one should use some non-standard types providing higher precision. For the programming language `C` one can use the type `mpf_t` which can be found in the `GMP` library or the type `mpfr_t` which can be found in `MPFR` library.

4.1.2. *Finding the proposal distribution using the mode of the full conditional.* An alternative to expansion of the polynomial is finding the local maximum $\hat{x}$ of the pdf

$f_{\lambda_j}(x)$ in the region $x > \max\{a, 0\}$ and then derive the proposal distribution based on such local maximum.

The proposition below allows us to find whether such local maximum exists and when it does to find its region of localization.

Let $c_{min} = \min\limits_{1\le s\le E_j} c_s$ be the minimum of the coefficients c_s of $f_{\lambda_j}(x)$ defined in Proposition 4.1, and $c_{max} = \max\limits_{1\le s\le E_j} c_s$ be the maximum of them.

Also define the following two values (if the expressions under the square roots are non-negative):

$$x_L = \frac{1}{2}\left(\varepsilon(\rho+E_j-1) - c_{max} + \sqrt{(\varepsilon(\rho+E_j-1)-c_{max})^2 + 4\varepsilon(\rho-1)c_{max}}\right), \quad (2.4.8)$$

$$x_U = \frac{1}{2}\left(\varepsilon(\rho+E_j-1) - c_{min} + \sqrt{(\varepsilon(\rho+E_j-1)-c_{min})^2 + 4\varepsilon(\rho-1)c_{min}}\right). \quad (2.4.9)$$

PROPOSITION 4.3. *The following statements for the greatest extremum $\hat{x}$ of the pdf $f_{\lambda_j}(x)$ in the region $x > \max\{a, 0\}$ are true:*

(1) If x_U is undefined or $x_U \le \max\{a, 0\}$ then $f_{\lambda_j}(x)$ does not have extrema for $x > \max\{a, 0\}$ and is strictly decreasing in this region.

(2) If there exist extrema of $f_{\lambda_j}(x)$ in $x > \max\{a, 0\}$ then there are only finite number of them and the greatest extremum, $\hat{x}$, is a local maximum. In addition, in this case, x_U is guaranteed to be defined and $\hat{x}$ satisfies the inequality $\max\{a, 0\} < \hat{x} \le x_U$.

(3) If x_L is defined and $x_L > \max\{a, 0\}$, then both $\hat{x}$ and x_U are defined and satisfy the inequality $\max\{a, 0\} < x_L \le \hat{x} \le x_U$.

PROOF. The proof is presented in Appendix B □

Now, for the first case stated in Proposition 4.3, i.e. when x_U is undefined or $x_U \leq \max\{a, 0\}$, we set $\hat{x} = \max\{a, 0\}$ as the maximum. Note that the value of $f_{\lambda_j}(x)$ can be infinite at this point.

If x_U is defined and $x_U > \max\{a, 0\}$, we are sure that if $\hat{x}$ exists it satisfies $\hat{x} \leq x_U$. So we can constrain the search for $\hat{x}$ to the region $\max\{a, 0\} < x \leq x_U$.

Also, if the third condition is satisfied, i.e. x_L is defined and $x_L > \max\{a, 0\}$, the region for the search can be constrained to the region $x_L \leq x \leq x_U$.

We use the modified Newton-Raphson optimization algorithm which searches for extremum in an open interval. The details about this algorithm are presented in Appendix C.

This algorithm attempts to find the extremum in the specified open interval (L, U) and guarantees that the returned value belongs to this interval even if the desired extremum is not found. The ability of the algorithm to return some well-defined value of x for any input is essential for our application. If we find any finite point $\hat{x}$ in the region $x > \max\{a, 0\}$ and construct the proposal distribution with the support in this region and mode at $\hat{x}$, the MCMC algorithm will work with this proposal. However, in order for the proposal distribution to be close to the desired full conditional, we try to use not the arbitrary point but the maximum and only if we fail to do so we rely on the fact that this point can be chosen arbitrarily.

We can run this algorithm with the limits L and U found as follows:

$$L = \begin{cases} \max\{x_L, a, 0\}, & \text{if } x_L \text{ exists,} \\ \max\{a, 0\}, & \text{otherwise,} \end{cases} \tag{2.4.10}$$

$$U = x_U. \tag{2.4.11}$$

Note that, if the maximum is reached at either L or U, the algorithm will not return exactly this value but a value very close to it.

It is worth mentioning that even if the extremum does not exist, the algorithm will still give some value between L and U. This means that if we do not succeed in finding the largest extremum of $f_{\lambda_j}(x)$ we still obtain some value of $\hat{x}$ which can be used for construction of the proposal distribution.

After we find the $\hat{x}$, we can use the Gamma proposal shifted by the value $\max\{a, 0\}$ as before, and we set the scale parameter of it equal to the scale parameter of the original distribution $\varepsilon = 1/((R_j + c_0)\Delta t_j)$. The shape parameter ν of the proposal distribution is chosen such that the mode of this distribution is equal to $\hat{x}$. That is, we set:

$$\nu = \frac{\hat{x} - \max\{a, 0\}}{\varepsilon} + 1. \tag{2.4.12}$$

Note that proposal distribution found by any of the two discussed methods is not an approximation of the desired conditional distribution in any way but it only follows the shape of this distribution. However, the shape similarity is enough for our purposes, since wrapping the sampling from it into a Metropolis-Hastings step adjusts for any differences in these distributions.

4.2. Sampling from the regression functions conditional distribution. The distribution of the component of one regression function α_{kj} conditional on all other parameters is given by the proposition below.

Let q be the number of individuals who had an event in the interval $(t_{j-1}, t_j]$ and whose k-th covariates were not zero at the moment of event. Also suppose that these individuals have indices $i = i_1, \ldots, i_q$ in the original dataset.

In addition, denote the constant

$$C_{constr} = \left(-\lambda_j - \sum_{k' \neq k} \min \left\{ \alpha_{k'j} \inf \Omega_{k'}, \alpha_{k'j} \sup \Omega_{k'} \right\} - \min_{1 \leq l \leq n} \omega_{lj} \right), \tag{2.4.13}$$

coefficients

$$c_s = \frac{1}{z_{ikj}} \left(\lambda_j + \sum_{k' \neq k} \alpha_{k'j} z_{ik'j} + \omega_{l_i s j} \right), \tag{2.4.14}$$

$$\varepsilon = \left(\sum_{i=1}^{N} R_{ij} z_{ikj} \right) \Delta t_j, \tag{2.4.15}$$

and the constraints

$$\begin{aligned} a &= \begin{cases} \frac{C_{constr}}{\sup \boldsymbol{\Omega}_k} & \text{if } \sup \boldsymbol{\Omega}_k > 0, \\ -\infty & \text{otherwise,} \end{cases} \\ b &= \begin{cases} \frac{C_{constr}}{\inf \boldsymbol{\Omega}_k} & \text{if } \inf \boldsymbol{\Omega}_k < 0, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{2.4.16}$$

PROPOSITION 4.4. *The probability density function $f_{\alpha_{kj}}(x)$ of α_{kj} conditional on all other parameters is proportional to:*

$$f_{\alpha_{kj}}(x) \propto \left(\prod_{s=1}^{q} (x + c_s) \right) \exp\left(-\varepsilon x\right) \mathbb{I}\left\{ a \leq x \leq b \right\}, \tag{2.4.17}$$

PROOF. The proof is presented in Appendix B. □

Similarly to the baseline hazard, we propose two methods of constructing the proposal distribution

4.2.1. *Construction of proposal distribution using the mean of the full conditional.* The mean of the distribution given by Proposition 4.4 is obtained as follows.

Let d_f be the coefficients of the polynomial $\sum_{f=0}^{q} d_f x^f$ obtained by expansion of the product $\prod_{s=1}^{q}(x+c_s)$ in the pdf $f_{\alpha_{kj}}(x)$ defined in Proposition 4.4.

Now, if $\varepsilon \neq 0$ denote

$$\mathcal{I}_0 = \frac{1}{\varepsilon}\left(\exp(-\varepsilon a) - \exp(-\varepsilon b)\right), \tag{2.4.18}$$

$$\mathcal{I}_f = \frac{1}{\varepsilon}\left(a^f \exp(-\varepsilon a) - b^f \exp(-\varepsilon b) + f\mathcal{I}_{f-1}\right), \quad f = 1, \ldots, q, \tag{2.4.19}$$

where in case of infinite a or b the values of the corresponding functions are evaluated as limits when argument approaches infinity.

In the case $\varepsilon = 0$ let

$$\mathcal{I}_f = \frac{b^{f+1} - a^{f+1}}{f+1}, \quad f = 0, \ldots, q, \tag{2.4.20}$$

where we suppose that a and b are both finite.

PROPOSITION 4.5. *The normalizing constant C_{norm} and the mean μ of the distribution $f_{\alpha_{kj}}(x)$ can be found as:*

$$C_{norm} = \sum_{f=0}^{q} d_f \mathcal{I}_f, \tag{2.4.21}$$

$$\mu = \frac{\sum_{f=0}^{q} d_f \mathcal{I}_{f+1}}{\sum_{f=0}^{q} d_f \mathcal{I}_f}. \tag{2.4.22}$$

PROOF. The proof is presented in Appendix B □

In the case of $\varepsilon \neq 0$, we can use proposal of the form:

$$S + \operatorname{sign}(\varepsilon)\mathcal{G}\left(\nu, \frac{1}{|\varepsilon|}\right), \tag{2.4.23}$$

where S is equal to a or b depending on the sign of ε:

$$S = \begin{cases} a, & \text{if } \varepsilon > 0, \\ b, & \text{if } \varepsilon < 0, \end{cases} \tag{2.4.24}$$

and ν is calculated in a such way that the mean of the proposal distribution is equal to the mean μ obtained earlier, i.e.:

$$\nu = \varepsilon(\mu - S). \tag{2.4.25}$$

So the proposal density $g(x)$ becomes:

$$g(x) = \frac{|\varepsilon|^{\nu}}{\Gamma(\nu)} \left(\text{sign}(\varepsilon)x - S\right)^{\nu-1} \exp\left(-|\varepsilon| \left(\text{sign}(\varepsilon)x - S\right)\right). \tag{2.4.26}$$

For $\varepsilon = 0$ we can use the Gaussian proposal with mean equal to μ and standard deviation τ which provides the same ratio of Gaussian pdf $g(x)$ at mean μ and one more point y (the choice of which will be discussed later) as that of the original distribution $f_{\alpha_{kj}}(x)$ at the same points:

$$\frac{g(\mu)}{g(y)} = \frac{f_{\alpha_{kj}}(\mu)}{f_{\alpha_{kj}}(y)} \Leftrightarrow \frac{1}{\exp\left(-\frac{(y-\mu)^2}{2\kappa^2}\right)} = \frac{f_{\alpha_{kj}}(\mu)}{f_{\alpha_{kj}}(y)} \Leftrightarrow \kappa = \sqrt{\frac{(y-\mu)^2}{2\ln\left(\frac{f_{\alpha_{kj}}(\mu)}{f_{\alpha_{kj}}(y)}\right)}}. \tag{2.4.27}$$

The point y is chosen to be $0.1\mu + 0.9a$ or $0.1\mu + 0.9b$ whichever produces the greater value of $f_{\alpha_{kj}}(x)$. We do not use the points a and b since $f_{\alpha_{kj}}(x)$ can be 0 at both of them.

4.2.2. *Construction of proposal distribution using the mode of the full conditional.* Similarly to the proposal distribution of baseline hazard, we can construct the proposals for regression functions without expanding the polynomial. Instead, we are trying to find the mode of the distribution $f_{\alpha_{kj}}(x)$ and construct the proposal density $g(x)$ with the same mode.

The following proposition describes the possible behaviour of the function $f_{\alpha_{kj}}(x)$ defined in Proposition 4.4.

PROPOSITION 4.6. *Provided that q and ε are not simultaneously equal to 0, the density $f_{\alpha_{kj}}(x)$ satisfies one of the following conditions:*

(1) $f_{\alpha_{kj}}(x)$ has a unique maximum in the region $a < x < b$.

(2) $f_{\alpha_{kj}}(x)$ is strictly decreasing in the interval $a < x < b$, in which case a is finite.

(3) $f_{\alpha_{kj}}(x)$ is strictly increasing in the interval $a < x < b$, in which case b is finite.

Moreover, if $q \neq 0$, the modified Newton-Raphson algorithm explained in Appendix C applied to the function $f_{\alpha_{kj}}(x)$ with the limits $L = a$ and $U = b$ always converges to a finite value.

PROOF. The proof is presented in Appendix B. □

So in the case of $q \neq 0$, our modified Newton-Raphson algorithm will converge to some satisfactory value, and in the case of $q = 0$, the maximum is $\hat{x} = a$ if $\varepsilon > 0$ or $\hat{x} = b$ if $\varepsilon < 0$ and is always finite.

The obtained point $\hat{x}$ can be used for construction of a proposal distribution. We consider the same form of proposal distribution as in the previous section, i.e. distribution given by formula (2.4.23) in the case of $\varepsilon \neq 0$ and Gaussian proposal in the case of $\varepsilon = 0$.

The shape parameter ν of the Gamma distribution can be found as $\varepsilon(\hat{x} - S) + 1$, the scale parameter remains $\frac{1}{|\varepsilon|}$ and S is the same as before.

If $\varepsilon = 0$ we use the Gaussian proposal with mean at $\hat{x}$ and standard deviation found through the equality of ratios like before.

4.3. Sampling from the frailty's full conditional distribution. The conditional distribution of the frailty ω_{lj} given all other parameters is given by the following proposition.

Let q be the number of individuals from the l-th region who had an event in the interval $(t_{j-1}, t_j]$. Also suppose that these individuals have indices $i = i_1, \ldots, i_q$ in the original dataset.

Let the limits a_1 and b_1 be

$$a_1 = -\lambda_j - \sum_{k=1}^{p} \min\left\{\alpha_{kj} \inf \Omega_k, \alpha_{kj} \sup \Omega_k\right\}, \tag{2.4.28}$$

$$b_1 = +\infty, \tag{2.4.29}$$

and the parameters μ_0 and δ be

$$\mu_0 = \frac{\theta_j^2}{m_l} R_j^{(l)} \Delta t_j + \overline{\omega}_{lj}, \tag{2.4.30}$$

$$\delta^2 = \frac{\theta_j^2}{m_l}. \tag{2.4.31}$$

PROPOSITION 4.7. *The density $f_{\omega_{lj}}(x)$ of the parameter ω_{lj} conditional on all other parameters is proportional to:*

$$f_{\omega_{lj}}(x) \propto \left(\prod_{s=1}^{q}(x + c_s)\right) \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\delta^2}\right) \mathbb{I}\left\{a_1 < x < b_1\right\}. \tag{2.4.32}$$

PROOF. The proof is presented in Appendix B. □

4.3.1. *Construction of proposal distribution using the mean of full conditional.* The mean of the distribution given by Proposition 4.7 is obtained as follows.

Let $x = \frac{y-\mu_0}{\delta}$ and let d_f be the coefficients of the polynomial $\sum_{f=0}^{q} d_f y^f$ obtained by expansion of the product $\prod_{s=1}^{q}(x+c_s)$ in the pdf $f_{\omega_{lj}}(x)$ defined in Proposition 4.4.

Also, let $a = \frac{a_1-\mu_0}{\delta}$ and $b = \frac{b_1-\mu_0}{\delta}$.

Denote:

$$\mathcal{I}_0 = \Phi(b) - \Phi(a), \tag{2.4.33}$$

$$\mathcal{I}_1 = \varphi(a) - \varphi(b), \tag{2.4.34}$$

$$\mathcal{I}_f = a^{f-1}\varphi(a) - b^{f-1}\varphi(b) + (f-1)\mathcal{I}_{f-2}, \tag{2.4.35}$$

where $\Phi(y)$ and $\varphi(y)$ are the CDF and pdf of the standard normal distribution, respectively.

PROPOSITION 4.8. *Then the normalizing constant C_{norm} and the mean μ of the distribution $f_{\omega_{lj}}(x)$ can be found as:*

$$C_{norm} = C^{Y}_{norm}\delta^q, \tag{2.4.36}$$

$$\mu = \delta\mu_Y + \mu_0, \tag{2.4.37}$$

where

$$C^{Y}_{norm} = \sum_{f=0}^{q} d_f \mathcal{I}_f, \tag{2.4.38}$$

$$\mu_Y = \frac{\sum_{f=0}^{q} d_f \mathcal{I}_{f+1}}{\sum_{f=0}^{q} d_f \mathcal{I}_f}. \tag{2.4.39}$$

PROOF. The proof is presented in Appendix B. □

Therefore the proposal $g(x)$ will have μ as its mean parameter. The standard deviation parameter κ of the proposal should be chosen such that the proposal distribution $g(x)$ would be close to the distribution $f_{\omega_{lj}}(x)$ given by the Proposition 4.7. A convenient way of doing this is to make the proposal density equal to $f_{\omega_{lj}}(x)$ at the mean μ. This ensures that both densities have approximately the same scale. Also in the case when the density $f_{\omega_{lj}}(x)$ is exactly normal, the proposal density $g(x)$ will be equal to $f_{\omega_{lj}}(x)$ exactly.

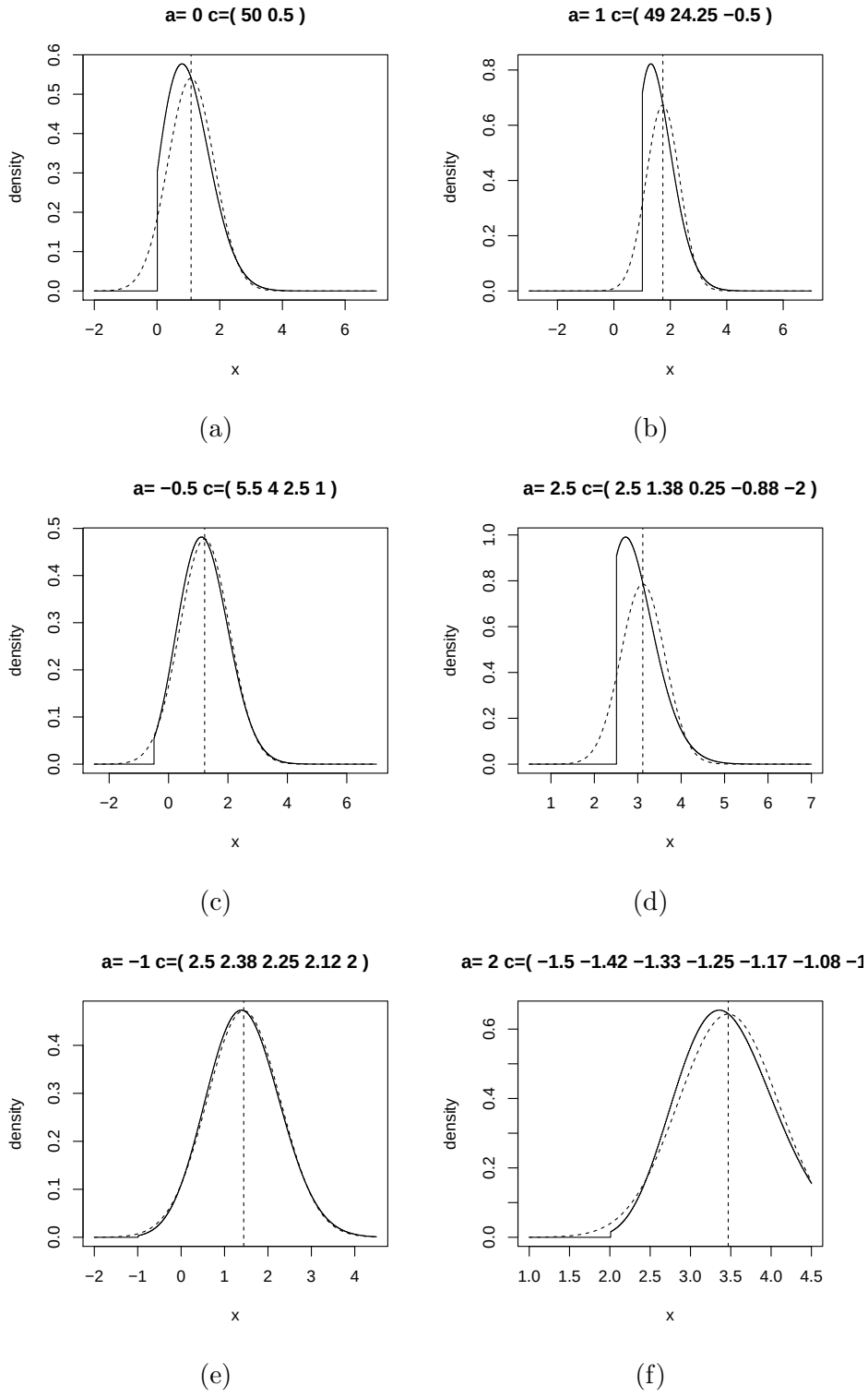
Thus, the parameter κ can be found from the following equation:

$$f_{\omega_{lj}}(\mu) = g(\mu) \quad \Leftrightarrow \quad f_{\omega_{lj}}(\mu) = \frac{1}{\sqrt{2\pi}\kappa} \quad \Leftrightarrow \quad \kappa = \frac{1}{\sqrt{2\pi} f_{\omega_{lj}}(\mu)}. \tag{2.4.40}$$

The examples showing how the normal proposal with the parameters obtained above is close to a distribution constrained from the left side only, can be seen in the Figures 2a–2f. The solid line shows desired constrained distribution, the dashed line shows proposed normal distribution and the vertical line indicates the mean of both distributions.

The parameter a shown in the graph is the lower constraint. The upper bound b is set to infinity in all the graphs presented. The polynomials are expressed in the form $(x + c_1) \times \cdots \times (x + c_q)$, and the values of the coefficients $c_1, \ldots, c_q$ are shown on the graph. We see that for different polynomials (the power of the polynomials increases from graph to graph) and different constraint parameter values (they are chosen randomly) the proposed density follows the shape of the desired density satisfactory close. The constrained distribution is especially close to normal if the power of the polynomial is high.

FIGURE 2. Comparing real and proposal distribution for sampling from frailty's full conditional



4.3.2. *Construction of proposal distribution using the mode of the full conditional.* As before, along with the polynomial expansion method, we also offer an alternative method based on finding the extremum.

The following proposition describes the possible behaviour of the function $f_{\omega_{lj}}(x)$ defined in Proposition 4.7.

PROPOSITION 4.9. *The density $f_{\omega_{lj}}(x)$ satisfies one of the following conditions:*

(1) $f_{\omega_{lj}}(x)$ has a unique maximum in the region $x > a_1$.

(2) $f_{\omega_{lj}}(x)$ is strictly decreasing in the interval $x > a_1$.

The modified Newton-Raphson algorithm explained in Appendix C applied to the function $f_{\omega_{lj}}(x)$ with the limits $L = a_1$ and $U = +\infty$ always converges to the local maximum or to a_1.

PROOF. The proof is presented in Appendix B. □

The obtained $\hat{x}$ can be used as the mean of the proposal Gaussian distribution. The standard deviation parameter can not be found like in previous case because we do not know the normalizing constant, and hence we can not find the exact value of $f_{\omega_{lj}}(x)$. However, we can use the method of equal ratios as we did before. Since there is a possibility that extremum is outside the region $x > a_1$, the obtained $\hat{x}$ cannot be the actual extremum. And since we do not allow $\hat{x}$ to reach the limit a_1, the value of the function $f_{\omega_{lj}}(x)$ at the points between a_1 and $\hat{x}$ can be greater than that at $\hat{x}$. Further, for the equal ratio method we assume that $\hat{x}$ is the maximum. Therefore we

cannot always use the points near the limit a_1 and we need to use some point $y > \hat{x}$. We can use $y = \hat{x} + \max\{|a_1|, |\hat{x}|\}$.

The standard deviation κ can be then obtained using formula (2.4.27).

4.4. Sampling from spatial hyper-parameters' full conditional distributions. The full conditional distribution of a frailty hyper-parameter θ_j^2 is proportional to:

$$\pi\left(\theta_j^2 \mid \boldsymbol{\omega}_j\right) \propto \frac{1}{\left(\theta_j^2\right)^{(n-1)/2}} \exp\left(-\frac{1}{2\theta_j^2} \sum_{\substack{l \sim l' \\ l<l'}} \left(\omega_{lj} - \omega_{l'j}\right)^2\right) \left(\theta_j^2\right)^{-\gamma-1} \exp\left(-\frac{\beta}{\theta_j^2}\right)$$

$$= \left(\theta_j^2\right)^{-\left(\gamma+\frac{n-1}{2}\right)-1} \exp\left(-\frac{1}{\theta_j^2}\left(\beta + \frac{1}{2} \sum_{\substack{l \sim l' \\ l<l'}} \left(\omega_{lj} - \omega_{l'j}\right)^2\right)\right), \quad (2.4.41)$$

This means that the posterior distribution is also inverse-gamma:

$$\left(\theta_j^2 \mid \boldsymbol{\omega}_j\right) \sim \mathcal{IG}\left(\gamma + \frac{n-1}{2}, \beta + \frac{1}{2} \sum_{\substack{l \sim l' \\ l<l'}} \left(\omega_{lj} - \omega_{l'j}\right)^2\right). \quad (2.4.42)$$

So the prior for θ_j^2 appears to be conjugate which explains the inverse-gamma choice for it. This distribution is one of the standard distributions, and thus it is straightforward to sample directly from it without even using Metropolis-Hastings step.

CHAPTER 3

Geostatistical spatial model

In this chapter we present the prior, posterior and proposal distributions for the model with geostatistical spatial structure. However, the complex posterior distribution of the frailty hyper-parameters does not allow us to obtain the proposal distribution as we did it for the CAR model.

One possible solution of this issue could be to use the Metropolis Random Walk, for which the proposals are chosen to be Gaussian distributions centered at the previously sampled point. However, this method has a slow convergence, so it requires a large number of iterations. For the big data which we analyse and the large number of parameters of the model, performing so many iterations takes a huge amount of time.

So for now we do not have a solution for this problem and leave it for future research.

1. Prior distribution for the geostatistical model

The joint prior distribution of all parameters of the model for the geostatistical spatial structure is obtained the same way it was obtained for the CAR structure with the only difference in the frailties' distributions and distributions of their hyper-parameters.

1.1. Geostatistical model for frailties. If the correlation between the observations from different regions is expected to depend on the distances between the regions, we can use the geostatistical model for which the frailty terms $\boldsymbol{\omega}_j = (\omega_{1j}, \ldots, \omega_{nj})^T$ are considered jointly multivariate normal with the variance-covariance matrix depending on the distances between locations (see Banerjee et al., 2003), i.e.:

$$\boldsymbol{\omega}_j \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\big(\sigma_j^2, \phi_j\big)\right), \tag{3.1.1}$$

where $\mathbf{0}$ is the $n \times 1$ vector of zeros and $\boldsymbol{\Sigma}\big(\sigma_j^2, \phi_j\big)$ is the $n \times n$ matrix with the components:

$$\boldsymbol{\Sigma}_{ll'}\big(\sigma_j^2, \phi_j\big) = \sigma_j^2 \exp\left(-\phi_j d_{ll'}\right), \quad l, l' \in \{1, \ldots, n\}, \tag{3.1.2}$$

where $d_{ll'}$ denotes the distance between the locations of the groups l and l'. The parameter σ_j^2 represents the variance of each of the frailty terms, and parameter ϕ_j determines the correlation between the frailties (the less this parameter is, the more is the correlation). In this model the correlation between the groups decreases exponentially with the distance.

Note that the introduced covariance structure assumes the isotropic correlation which means that it decreases in the same manner in any direction. In order to introduce anisotropic correlation, one should use different kind of variance-covariance matrix.

The frailties in different intervals are assumed independent. So the joint prior distribution of all frailties can be obtained by multiplication of the priors in different intervals.

To simplify the notation, we denote the set of all $\sigma_j,\ j = 1, \ldots, m$ as $\boldsymbol{\sigma}$, and the set of all ϕ_j as $\boldsymbol{\phi}$.

1.1.1. *Prior distributions of the hyper-parameters.* For the geostatistical spatial model, we put the prior distributions on the variance parameters σ_j^2 and correlation parameters ϕ_j. The prior distributions for ϕ_j and σ_j^2 are taken according to Banerjee et al. (2003).

We take the gamma prior $\mathcal{G}(a, 1/a)$ for ϕ_j with the shape and scale parameters a and $1/a$ respectively, so that the mean is 1 and variance is $1/a$. For a vague prior, the value of a should be taken small.

The prior for σ_j^2 is assumed to be inverse-gamma $\mathcal{IG}(\beta, \gamma)$ with β and γ as scale and shape parameters, respectively. The mean in this case is $\frac{\beta}{\gamma-1}$ and variance is $\frac{\beta^2}{(\gamma-1)^2(\gamma-2)}$. By setting $\gamma = 2$ we can make this prior vague which is provided by the infinite variance. The inverse-gamma prior is chosen in order to make it conjugate.

1.1.2. *Joint prior distribution for the geostatistical model.* For geostatistical model the joint constrained prior distribution takes the following form:

$$\begin{aligned}\pi\left(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\phi}\right) \propto & \prod_{j=1}^{m}\underbrace{\left(\lambda_j^{c_0 r_0 \Delta t_j - 1}\exp\left(-c_0\lambda_j\Delta t_j\right)\right)}_{\text{Gamma prior for } \lambda_j} \\ & \times \prod_{j=1}^{m}\underbrace{\left(\frac{1}{\sqrt{\det\left(\boldsymbol{\Sigma}\left(\sigma_j^2,\phi_j\right)\right)}}\exp\left(-\frac{1}{2}\boldsymbol{\omega}_j^T\left(\boldsymbol{\Sigma}\left(\sigma_j^2,\phi_j\right)\right)^{-1}\boldsymbol{\omega}_j\right)\right)}_{\text{Joint multivariate normal prior for } \boldsymbol{\omega}_j=(\omega_{1j},\ldots,\omega_{nj})^T} \\ & \times \prod_{j=1}^{m}\underbrace{\left(\mathbb{I}\left\{\lambda_j + \sum_{k=1}^{p}\min\left\{\alpha_{kj}\inf\Omega_k, \alpha_{kj}\sup\Omega_k\right\} + \min_{1\le l\le n}\omega_{lj} \ge 0\right\}\right)}_{\text{Constraint component}} \\ & \times \prod_{j=1}^{m}\underbrace{\left(\phi_j^{a-1}\exp\left(-a\phi_j\right)\right)}_{\text{Gamma prior for } \phi_j}\prod_{j=1}^{m}\underbrace{\left(\left(\sigma_j^2\right)^{-\gamma-1}\exp\left(-\frac{\beta}{\sigma_j^2}\right)\right)}_{\text{Inverse-gamma prior for } \sigma_j^2}, \quad (3.1.3)\end{aligned}$$

where $\det(\mathbf{X})$ denotes the determinant of the matrix $\mathbf{X}$ and all other notation is the same as in the CAR model.

This likelihood is similar to that of the CAR model with the exception that the parameters ϕ_j are included in the distribution in a very complex way.

2. Obtaining a random sample from the posterior distribution

2.1. Posterior distribution. For the posterior distribution of the parameters we have:

$$\pi\left(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\phi}\mid\mathbf{D}\right) \propto L\left(\mathbf{D}\mid\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\omega}\right)\pi\left(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\phi}\right), \quad (3.2.1)$$

where the likelihood and the prior distribution are given by the equations (2.3.1) and (3.1.3) respectively.

This is obtained similarly to the case of CAR model.

2.2. Sampling from the frailty full conditional distribution. The full conditional distribution of the l-th frailty for the geostatistical model it is proportional to:

$$\begin{aligned}\pi\left(\omega_{lj} \mid \lambda_j, \boldsymbol{\alpha}_j, \omega_{l'j\neq lj}, \sigma_j^2, \phi_j, \mathbf{D}\right) \\ \propto \prod_{i\in\mathcal{E}_j\cap\mathcal{S}_l}\left(\lambda_j + \sum_{k=1}^{p}\alpha_{kj}z_{ikj} + \omega_{lj}\right)\exp\left(-R_j^{(l)}\omega_{lj}\Delta t_j\right) \\ \times \exp\left[-\frac{1}{2\sigma_j^2}\left(\omega_{lj}^2 + 2\sum_{l'\neq l}\omega_{lj}\omega_{l'j}\left(\mathbf{H}(\phi_j)\right)_{ll'}^{-1}\right)\right] \\ \times \mathbb{I}\left\{\omega_{lj} \geq -\lambda_j - \sum_{k=1}^{p}\min\left\{\alpha_{kj}\inf\Omega_k, \alpha_{kj}\sup\Omega_k\right\}\right\}, \quad (3.2.2)\end{aligned}$$

where $\left(\mathbf{H}(\phi_j)\right)_{ll'}^{-1}$ denotes the ll'-th component of the matrix $\left(\mathbf{H}(\phi_j)\right)^{-1}$ with $\mathbf{H}(\phi_j) = \frac{1}{\sigma_j^2}\boldsymbol{\Sigma}\left(\sigma_j^2, \phi_j\right)$ being the correlation matrix of the frailties and all other notation as before.

The same way it was done for the CAR model, it can be shown that the distribution given by equation (3.2.2) has the following form:

$$f_{\omega_{lj}}(x) = \left(\prod_{s=1}^{q}(x + c_s)\right)\frac{1}{\sqrt{2\pi\delta^2}}\exp\left(-\frac{(x-\mu_0)^2}{2\delta^2}\right)\mathbb{I}\left\{a_1 < x < b_1\right\}, \quad (3.2.3)$$

where the power of the polynomial is $q = \text{card}(\mathcal{E}_j \cap \mathcal{S}_l)$, and the limits a_1 and b_1 are the following:

$$\begin{aligned} a_1 &= -\lambda_j - \sum_{k=1}^{p}\min\left\{\alpha_{kj}\inf\Omega_k, \alpha_{kj}\sup\Omega_k\right\}, & (3.2.4)\\ b_1 &= +\infty. & (3.2.5)\end{aligned}$$

The parameters μ_0 and δ of this distribution are:

$$\mu_0 = \sigma_j^2 R_j^{(l)} \Delta t_j - \sum_{l' \neq l} \left(\mathbf{H}(\phi_j) \right)^{-1}_{ll'}, \tag{3.2.6}$$

$$\delta^2 = \sigma_j^2. \tag{3.2.7}$$

So the algorithm for sampling from geostatistical frailty full conditional distribution is exactly the same as for the CAR model discussed before.

2.3. Sampling from the full conditionals of geostatistical model variance hyper-parameters. The full conditional distribution of a frailty variance parameter σ_j^2 is proportional to:

$$\begin{aligned} \pi\left(\sigma_j^2 \mid \boldsymbol{\omega}_j, \phi_j\right) &\propto \frac{1}{\left(\sigma_j^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma_j^2} \boldsymbol{\omega}_j^T \left(\mathbf{H}(\phi_j) \right)^{-1} \boldsymbol{\omega}_j \right) \left(\sigma_j^2\right)^{-\gamma-1} \exp\left(-\frac{\beta}{\sigma_j^2} \right) \\ &= \left(\sigma_j^2\right)^{-\left(\gamma+\frac{n}{2}\right)-1} \exp\left(-\frac{1}{\sigma_j^2} \left(\beta + \frac{1}{2} \boldsymbol{\omega}_j^T \left(\mathbf{H}(\phi_j) \right)^{-1} \boldsymbol{\omega}_j \right) \right), \end{aligned} \tag{3.2.8}$$

which means that it has the from of inverse-gamma distribution:

$$\left(\sigma_j^2 \mid \boldsymbol{\omega}_j, \phi_j\right) \sim \mathcal{IG}\left(\gamma + \frac{n}{2}, \beta + \frac{1}{2} \boldsymbol{\omega}_j^T \left(\mathbf{H}(\phi_j) \right)^{-1} \boldsymbol{\omega}_j \right). \tag{3.2.9}$$

2.4. Sampling from the full conditionals of geostatistical model correlation hyper-parameters. The full conditional distribution of the correlation hyper-parameter ϕ_j has the following form:

$$\pi\left(\phi_j \mid \boldsymbol{\omega}_j, \sigma_j^2\right) \propto \frac{1}{\sqrt{\det \mathbf{H}(\phi_j)}} \exp\left(-\frac{1}{2\sigma_j^2} \boldsymbol{\omega}_j^T \left(\mathbf{H}(\phi_j) \right)^{-1} \boldsymbol{\omega}_j \right) \phi_j^{a-1} \exp\left(-a\phi_j\right) \tag{3.2.10}$$

The variable ϕ_j is included in this distribution in a very complex way (through the determinant and inverse of matrix $\mathbf{H}(\phi_j)$. So we did not succeed in obtaining a proposal distribution which would follow the form of this distribution. This is the

only problematic parameter in the geostatistical model. All other parameters have the same procedures of sampling like in CAR model and thus can be implemented similarly.

We leave the procedure of obtaining the suitable proposal distribution for the parameter ϕ_j for the future research. After it is found, the geostatistical model can be easily implemented based on the implementation of the CAR model.

CHAPTER 4

Application of the Method

1. Model Implementation

We implement the whole specified model in a program using the combination of `R` and `C` programming languages.

The routine MCMC sampling part is implemented in `C` programming language. The necessary functions are written in `C` and compiled into a shared library. In order to allow analysing data with a very large number of observations (greater than 50,000), all the routine functions use the multiple-precision floating point types from the open source `C` libraries `GMP` and `MPFR`. These libraries allow to operate with numbers which are much closer to zero than it is allowed by the standard floating-point types of `R` and `C` languages.

This is essential for the data with large number of observations since the expressions for the posterior full conditional distributions become extremely small in this case. As a result, the computed expressions become zero due to rounding and all the analyses become impossible. The mentioned libraries allow to overcome this problem which greatly increases the applicability of the method.

For the method of constructing proposal distributions which uses the means of full conditionals (which requires computation of normalizing constants), the usage

of these libraries is necessary even for comparatively small number of observations, particularly for data in which the number of events during one time interval exceeds 50.

Also, we require reimplementation of CDF of gamma distribution with the use of multiple precision types and reimplementation of an algorithm for solving linear systems of equations, for which we implement a Householder method of solving them. The implementation of these functions is based on a code from another open source library called `GSL`. It contains most of the frequently used mathematical functions and methods but with the use of only standard double precision floating point type.

The rest of the program is written using the `R` language which calls the routine MCMC-sampling functions from the compiled `C` shared library.

The resulting program allows to incorporate the strong analytical capabilities of `R` language while utilizing the fast execution of `C` code for the computationally hard routine sampling procedures. Also, the `C` implementation allows to use the methods of the mentioned `GMP` and `MPFR` libraries.

While we implement both of the discussed methods of constructing the proposal distributions, we use only the one based on the mode of the full conditional since it is much faster and numerically stable especially when we analyse the data with around 50,000 observations later on. So in future we discuss only the results of the mode-based method.

2. Simulation Study

We conduct the simulations in order to study the performance of the proposed method of estimating the parameters. These simulations are intended to study how the input values of the algorithm affect the model parameters estimation.

2.1. Data Generation.

2.1.1. *Generating parameters.* We choose the number of covariates $p = 2$, the number of regions $n = 5$ and the study period $\tau = 1$. The baseline hazard is set to be a linear function of time $\lambda(t) = 0.3 + t$. The first regression function is set to be a linear function, and the second one a quadratic function of time: $\alpha_1(t) = -0.3t$, $\alpha_2(t) = -0.4t^2$. The sets of admissible covariate values are $\Omega_1 = \Omega_2 = [-1, 1]$.

The adjacency structure of the regions is represented by the tridiagonal matrix:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \tag{4.2.1}$$

where $A_{ll'} = 1$ denotes that the regions l and l' are adjacent.

The values of the frailties are assumed to be constant over time and are generated from the CAR distribution with the first frailty set to 0 and parameter $\theta = 0.1$. If the value of any frailty appears to be less than -0.3, all frailties are resampled.

The covariate values are also assumed to be constant and generated from uniform distribution with the limits specified in Ω_1 and Ω_2.

The generated parameters satisfy the condition of non-negative hazard given by (2.2.3), so the specified model is well defined.

2.1.2. *Generating Event Times.* In order to generate the event times for the individuals, we propose using the piecewise constant hazard approximation of survival function for each individual. For sufficiently large number of break points, the data generated from a distribution with such survival function is very close to the desired distribution.

We choose $M = 1,000$ equidistant break points $\tilde{t}_s = hs$, $h = 1/M$ from 0 to 1 and approximate the above mentioned functions by step functions $\tilde{\lambda}(t)$, $\tilde{\alpha}_1(t)$ and $\tilde{\alpha}_2(t)$ that are constant during the defined intervals and equal to the values of the approximated functions at the midpoints of the intervals. We will write the formula only for the baseline $\tilde{\lambda}(t)$, and the formulae for the regression functions approximations will be similar to it:

$$\tilde{\lambda}(t) = \lambda\left(\frac{\tilde{t}_{s(t)} + \tilde{t}_{s(t)-1}}{2}\right), \quad s(t) = \min\{r : t_r \geq t\}. \tag{4.2.2}$$

Note that this time partitioning is not the same partitioning which was discussed in the previous chapters. The breakpoints $\tilde{t}_s$ are used only for data generation.

Now, we compute the hazard for each individual at all breakpoints $\tilde{t}_s$ from which we can obtain the approximations of survival functions of these individuals. Then in order to generate the event time of one individual, we generate a random number between 0 and 1 and then apply the inverse survival function to this number. The obtained value is considered as the event time of that individual. The inverse survival function can be found based on the approximations $\tilde{\lambda}(t)$, $\tilde{\alpha}_1(t)$ and $\tilde{\alpha}_2(t)$ introduced above. Since all these three functions and frailties are piecewise constant functions any linear combination of them is also a piecewise constant function. So the problem

of finding the inverse survival function reduces to finding the inverse survival functions for the piecewise constant hazard.

The survival function has the following relation with the hazard:

$$S(t) = \exp\left(-\int_0^t h(u)\mathrm{d}u\right) \Leftrightarrow \ln S(t) = -\int_0^t h(u)\mathrm{d}u. \tag{4.2.3}$$

For the piecewise constant hazard, the integral reduces to the summation:

$$\ln S(t) = -\sum_{\substack{s\geq 1:\\ \tilde{t}_s \leq t}} h(\tilde{t}_s)\Delta\tilde{t}_s - h\left(\tilde{t}_{s(t)+1}\right)(t - \tilde{t}_{s(t)}), \tag{4.2.4}$$

where $s(t) = \max\{s \geq 0 :\ \tilde{t}_s \leq t\}$.

So if we want to find t given the value $S(t) = \hat{S}$ we can firstly find the index $\hat{s}$ such that the corresponding time point $t_{\hat{s}}$ is the beginning of the interval to which the desired t belongs. This $\hat{s}$ can be found from the following equation:

$$\hat{s} = \max\{s :\ 0 \leq s \leq M \text{ and } S(\tilde{t}_s) \geq \hat{s}\}. \tag{4.2.5}$$

Now when we know $\hat{s}$, we can find t from the equation:

$$-h\left(\tilde{t}_{\hat{s}+1}\right)(t - \tilde{t}_{\hat{s}}) = \ln\hat{S} + \sum_{s=1}^{\hat{s}} h(\tilde{t}_s)\Delta\tilde{t}_s \tag{4.2.6}$$

$$\Leftrightarrow t = -\frac{\ln\hat{S} + \sum_{s=1}^{\hat{s}} h(\tilde{t}_s)\Delta\tilde{t}_s}{h\left(\tilde{t}_{\hat{s}+1}\right)} + \tilde{t}_{\hat{s}}, \tag{4.2.7}$$

where we assume that if $\hat{s} = M$ then $t = +\infty$.

After all the event times are generated, we generate random censoring times from exponential distribution with the rate $\nu = -\ln(0.5)/\tau \approx 0.69$. This rate is chosen so that the probability of censoring time to exceed τ is equal to 0.5.

Now, if an individual has censoring time greater than the event time, then the event time is recorded and this individual is marked as having the event, otherwise the individual is marked as censored.

2.2. Studying the performance of the method. We run simulations for different values of N (number of observations) and m (number of time break points) and construct the plots representing the results of estimation which can be found in Appendix D. For every N and m we run the algorithm several times with different numbers of iterations. The first quarter of iterations in all cases is considered as burn-in period and is excluded from the estimation.

The thick line on each plot shows the real value of the parameter, the solid thin line shows the median value of parameter the among the simulated values, and dashed lines show the 2.5% and 97.5% quantiles. These quantiles take the place of confidence intervals in the frequentist's estimation.

One can notice that the number of iterations almost do not play any role in estimation. Thus, the estimators with the number of iterations equal 100 are almost the same as ones with the number of iterations equal to $5,000$. This can be explained by the very fast convergence of the designed MCMC algorithm.

The number of observations, on the other hand, has a very important role. One can notice that the confidence limits become significantly narrower when the number of observations increase.

Regarding the number of break points, we can notice the tendency of the estimators to be closer to the actual values when the number of time points increases.

This seems reasonable since we are trying to approximate time-varying functions. However, for large numbers of intervals compared to number of observations, the confidence limits become very wide. This can be explained by the fact that the accuracy of estimation depends on the number of observations carrying the information about the parameters. But when we increase the number of intervals, there are less subjects having events during each interval, thus increasing the variability.

So the number of intervals should be chosen as a trade-off between accuracy of the piecewise constant approximation and variability of estimators.

From all above mentioned we conclude that for data analysis there is no need to perform huge number of MCMC iterations and 500 seems to be quite reasonable. Also the method is sensitive to the number of observations, so it works better for big data. Regarding the number of time points, this choice should be made depending on the particular application. It is worth mentioning, that it is better not to take this number so big that some of the intervals do not have events at all. In this case the hazard for such intervals is estimated as zero or almost zero. While theoretically it may be correct estimation of hazard, practically such intervals appear not due to zero hazard, but due to lack of observations or discrete time recording. So, for reasonable estimation, such intervals should be joined with adjacent intervals to ensure that each interval has at least one event.

TABLE 1. Variables Used in Data Analysis

Variable Name in Analysis	Variable Values in Analysis	Variable Name in SEER Database
age	0–106	Age at diagnosis
stage	Distant, Localized/regional	SEER historic stage A
race	Black, White	Race recode (White, Black, Other)
marriage	Single, Married, Other	Marital status at diagnosis
region	1–64	County
time	1–72	Survival time recode (total # of months)
event	0, 1	Vital status recode (study cutoff used)

3. Application to the Prostate Cancer Data

3.1. Data Description. We apply the proposed method to the Surveillance, Epidemiology, and End Results prostate cancer data (SEER, 2008).

The data analysed is similar to that analysed in Zhang and Lawson (2011).

We extract the data set from the SEER 17 Registries Incidence database based on the November 2007 submission. As Zhang and Lawson (2011) we use only the Louisiana cases.

The variables considered are presented in the Table 1.

The observations with unknown age were excluded, and according to the stage of cancer, all observations were divided into two groups like in Zhang and Lawson (2011):

'Localized/regional' and 'Distant'. The cases with stage other than 'Localized', 'Regional', 'Localized/regional (Prostate cases)' or 'Distant' were also excluded from considerations. The first three groups were joined into one group, 'Localized/regional'.

The observations with 'Other' or 'Unknown' race were ignored. So only individuals with the race 'Black' or 'White' were included in the analysis.

The marital status was recoded into 'Married', 'Single' and 'Other'. The 'Other' category included all categories other than 'Married' and 'Single'. The observations with unknown marital status were ignored.

The county numbers presented in the SEER database include only odd numbers from 1 to 127. In the variable 'region' we recoded them into the numbers from 1 to 64. The numbers of the counties were in alphabetical order with respect to their names.

The survival times in the database represent the total number of months the patients survived. In order to avoid zero survival times, in our analysis we add 1 to all survival times so the survival times used in our analysis have different interpretation. In the original data, the value n of survival means that the patient's survival time is between n and $n+1$ months. In our analysis the value n means that survival time is between $n-1$ and n months. This changes allow us to avoid zero survival times by only slightly changing the interpretation. So the original survival times taking values 0–71 in our analysis are recoded to 1–72.

The event indicator variable 'event' contains the vital status recoded as 0 (Alive) and 1 (Dead).

Further, we constructed dummy variables for the categorical variables 'race', 'stage' and 'marriage'.

Therefore, the total number of variables is 5: 1 for 'age', 1 for 'race', 1 for 'stage' and 2 for 'marriage'. In addition, we have 64 counties the effects of which are analysed through introducing 63 spatial frailty terms (recall that the first frailty term is set to be always 0), and the intercept term is represented by the baseline hazard.

3.2. Choice of the number of intervals. As was already mentioned before, our method depends on the partitioning of time. We use the equidistant time break points, so the choice of the break points reduces to the number of these breakpoints.

We choose the optimal number of intervals based on the Deviance Information Criterion (DIC) as in Banerjee et al. (2003); Zhang and Lawson (2011) and the summary Log-Conditional Predictive Ordinate ($LCPO$) as in Silva and Amaral-Turkman (2004).

The DIC is a Bayesian analog of Akaike's Information Criterion (AIC). The value of DIC incorporates the information about the model fit and about the model complexity. The better the fit and the simpler the model is, the less is the DIC value. DIC can be found based on the deviance statistic (see Banerjee et al., 2003):

$$\mathrm{Dev}(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}) = -2 \ln L(\mathbf{D}|\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}), \tag{4.3.1}$$

where $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ are the baseline, regression functions and frailties values introduced before, $\mathbf{D}$ is the data and $L(\mathbf{D}|\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega})$ is the likelihood function.

Now, the fit of the model is represented by the posterior expectation of the deviance:

$$\overline{D} = E\left[\text{Dev}(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega})|\mathbf{D}\right]. \tag{4.3.2}$$

The complexity of the model is captured by the effective number of parameters p_D which can be defined as the posterior expectation of deviance minus deviance evaluated at the posterior expectation of the parameters:

$$p_D = E\left[\text{Dev}(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}) \mid \mathbf{D}\right] - \text{Dev}\left(E\left[(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}) \mid \mathbf{D}\right]\right). \tag{4.3.3}$$

Then the DIC can be computed as the sum of the obtained values $\overline{D}$ and p_D:

$$DIC = \overline{D} + p_D. \tag{4.3.4}$$

The smallest value of DIC among the models compared indicates the preferred model. Banerjee et al. (2003) mention that DIC can not be used for the identification of the correct model, but can only be used to compare the alternative formulations all of which can be incorrect. Also they note that the value of DIC itself has no meaning and only the differences on the DIC for the compared models are meaningful.

The posterior expectations in the formulas above can be obtained by the Monte-Carlo integration, i.e. using the fact that the posterior expectation of any measurable function $T(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega})$ of the sampled parameters can be estimated by the mean value of this function among all sampled values of parameters (see, e.g. Gelfand and Smith, 1990):

$$E\left[T(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}) \mid \mathbf{D}\right] \approx \frac{1}{I}\sum_{s=1}^{M} T\left(\boldsymbol{\lambda}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\omega}^{(s)}\right), \tag{4.3.5}$$

where I is the number of sampled values of parameters, and the superscript (s) indicates that we use the s-th sampled value of the parameter.

Another measure which can be used for model comparison is the summary Log-Conditional Predictive Ordinate $LCPO$ measure. This measure is based on the cross-validating predictive densities of observations given all other observations (see Silva and Amaral-Turkman, 2004), i.e.:

$$CPO_i = \pi(y_i|\mathbf{D}_{-i}) = E\left[L(y_i|\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\omega}) \mid \mathbf{D}_{-i}\right], \tag{4.3.6}$$

where y_i denotes the i-th observation and $\mathbf{D}_{-i}$ denotes the rest of the data after deleting the i-th observation from it, the expectation is computed with respect to the model parameters (with y_i fixed) and we assume that the observations are conditionally independent given the model parameters. The larger the value of CPO_i is, the better the i-th observation agrees with the model obtained using the rest of the data.

The CPO_i can be computed by the MCMC algorithm using:

$$CPO_i \approx \frac{1}{\frac{1}{I}\sum_{s=1}^{I} \frac{1}{L\left(y_i|\boldsymbol{\lambda}^{(s)}, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\omega}^{(s)}\right)}}, \tag{4.3.7}$$

where the denominator represents the Monte-Carlo integration of the reciprocal of the marginal likelihood of the i-th observation.

Comparison of two models can be made using a summary measure $LCPO$ defined as:

$$LCPO = \sum_{i=1}^{N} \ln CPO_i. \tag{4.3.8}$$

The large values of $LCPO$ imply a better model adequacy.

TABLE 2. Values of DIC, p_D and $LCPO$ for different numbers of intervals

Number of Breakpoints m	Actual Number of Parameters n_{par}	Effective Number of Parameters p_D	Deviance Information Criterion DIC	Conditional Predictive Ordinate $LCPO$
1	70	63	189,244	-94,630
12	840	463	-692,284	327,009
24	1,680	740	-1,780,091	827,536
36	2,520	953	-3,098,172	1,403,102
72	5,040	1,273	-7,702,954	3,393,691

From the simulation study we found that a reasonable choice of the number of MCMC algorithm iterations is 500. So we run this number of iterations for different numbers of breakpoints in order to obtain the optimal value of m. However we prefer to run $5,000$ iterations for the data analysis since the number of model parameters here is much more than that used in simulations and we want to decrease the estimation errors due to small sample.

We use both DIC and $LCPO$ for choosing the optimal number of breakpoints in the model. The values of DIC, p_D and $LCPO$ for different numbers of breakpoints m are presented in Table 2.

One can notice that the effective number of parameters is much less than the actual number of parameters in the model. Also it increases even slower that the actual number of parameters. For example, the effective number of parameters changes from 953 to $1,273$ for the case of $m = 36$ and $m = 72$ while the actual number of parameters increases from $2,520$ to $5,040$. This means that the penalty for introducing more parameters in the model is very small, and at the same time, the fit of the model

significantly improves when we use more time breakpoints which is indicated by the rapid decrease of DIC and increase of $LCPO$. So clearly, the more intervals we take the better a model becomes.

However, the further increase of number of breakpoints is unreasonable. The survival is presented in integer number of months. So it is impossible for the data to have any intermediate values between consequent integers. Therefore, the smallest reasonable partitioning is partitioning into the intervals of the length 1. This case is represented by the number of intervals equal to 72.

So the optimal number of intervals for this data is 72. Thus we select the model with $m = 72$ for further data analysis.

3.3. Analysis of the data. The estimated values of the baseline hazard, regression functions, and spatial frailty terms are presented in the Figures 3–6.

The solid line indicates the median values of parameters while the dashed lines show the 2.5% and 97.5% quantiles thus representing the confidence limits.

The Regression function 1 stands for the 'age', Regression function 2 for the 'race' (0 - Black, 1 - White), Regression function 3 for the 'stage' (0 - Distant, 1 - Localized/regional) and Regression functions 4 and 5 together stand for 'marriage' ((0,0) - Married, (0,1) - Single, (1,0) - Other).

The value of the first frailty term (corresponding to Acadia parish) is forced to be 0, so it is not shown in the figures.

The baseline hazard corresponds to the person with 'age=0', 'race=Black', 'stage=Distant' and 'marriage=Married' from the first county (Acadia).

FIGURE 3. Estimated Parameters of The Model

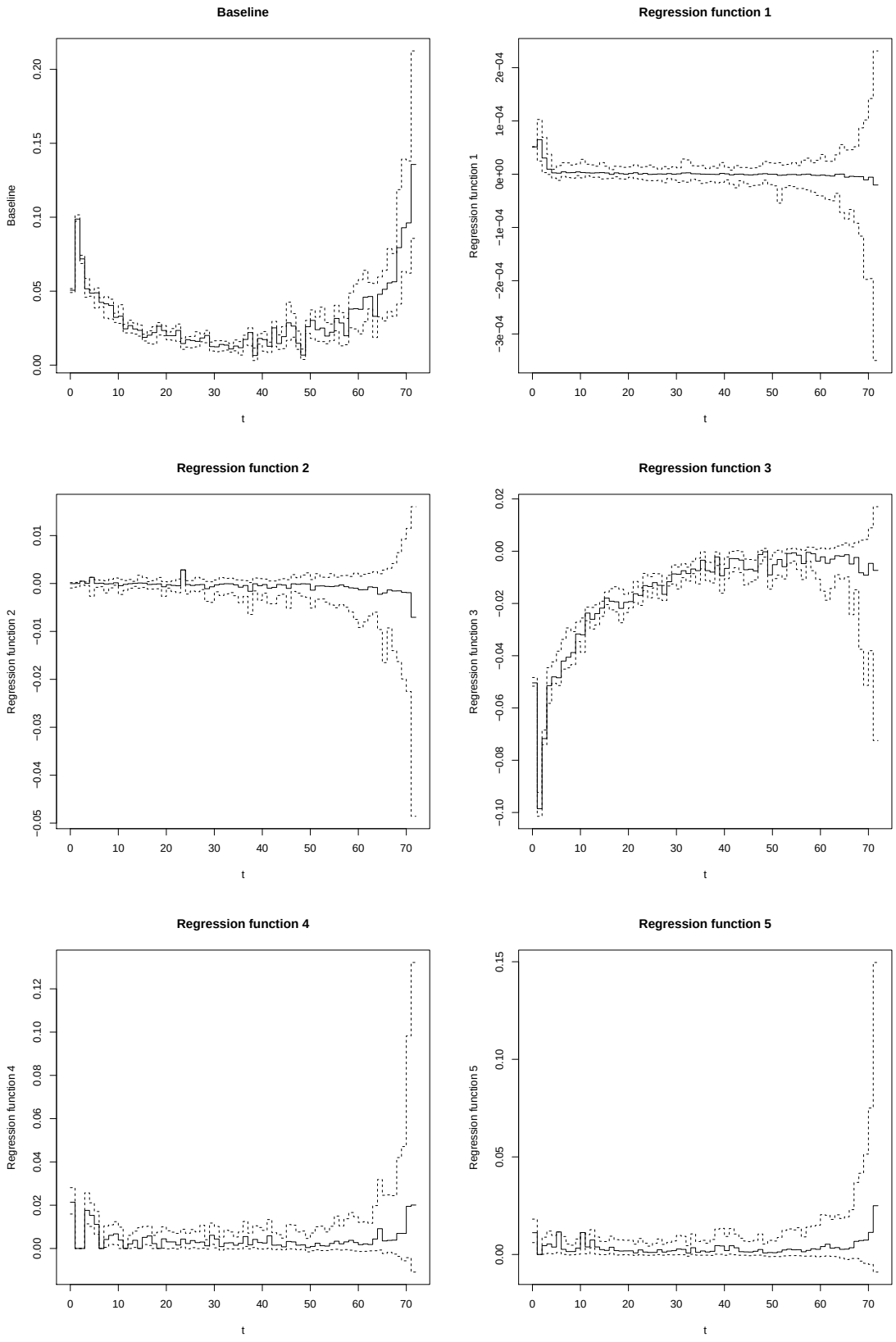


FIGURE 4. Estimated Parameters of The Model (Continuation)

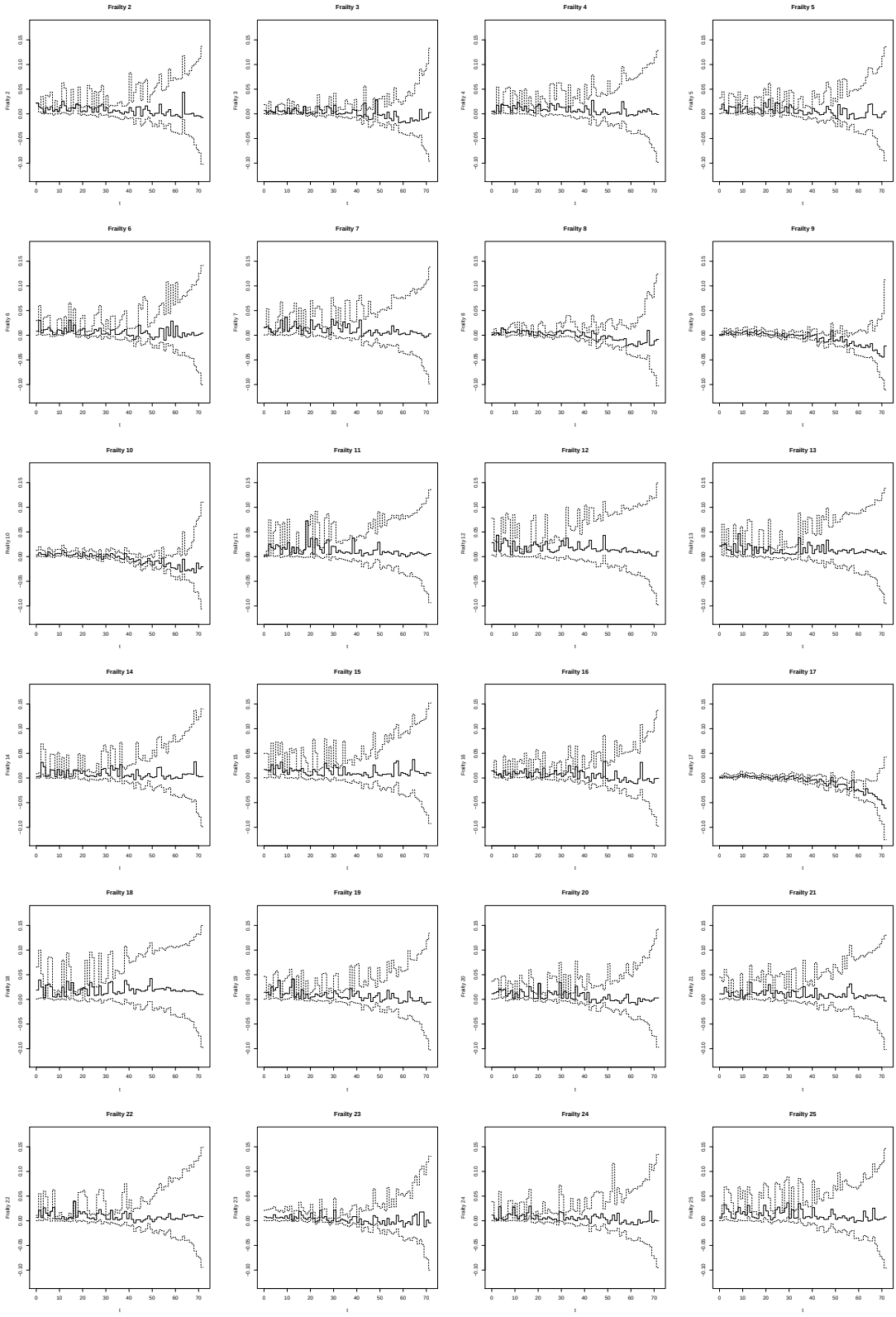


FIGURE 5. Estimated Parameters of The Model (Continuation)

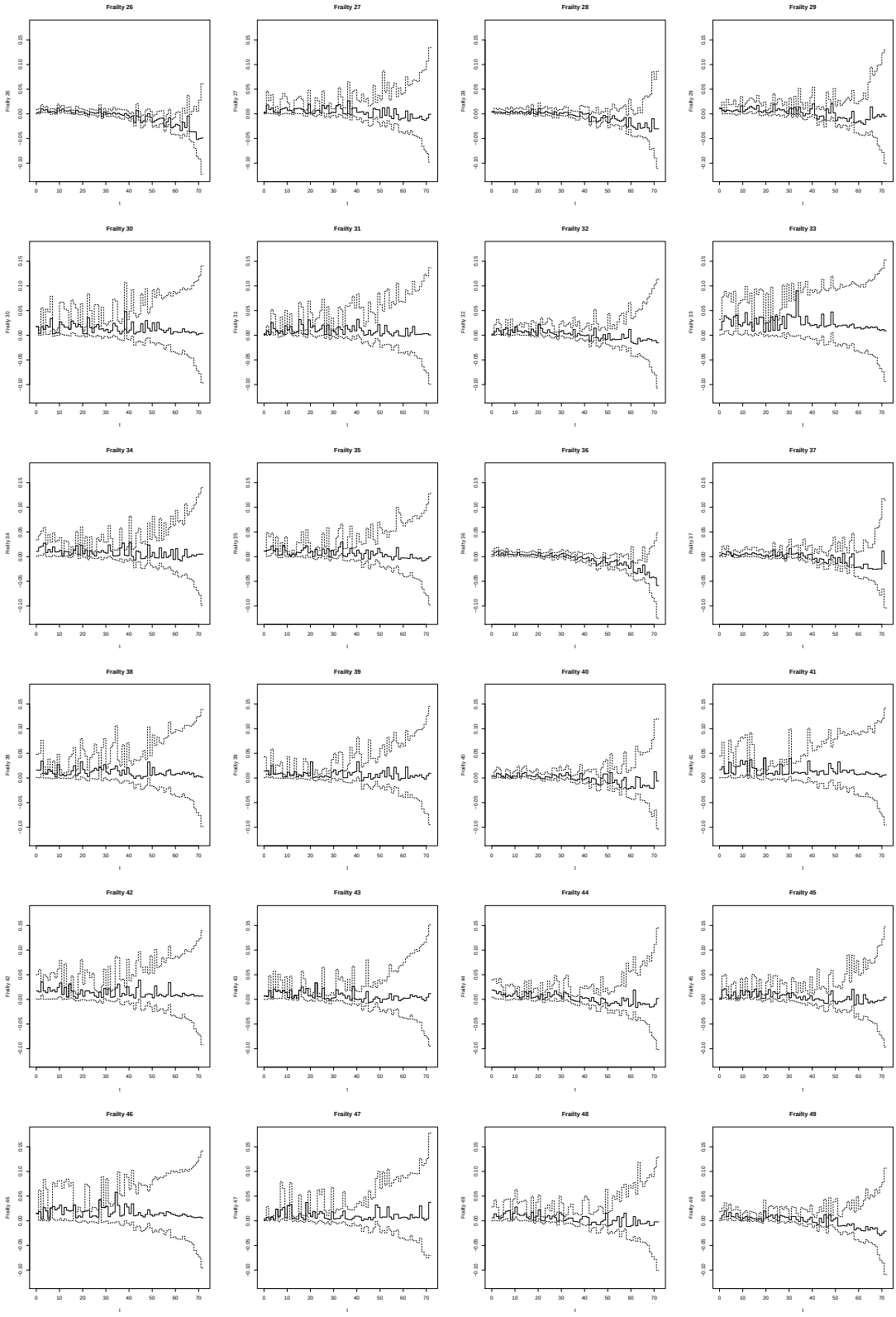
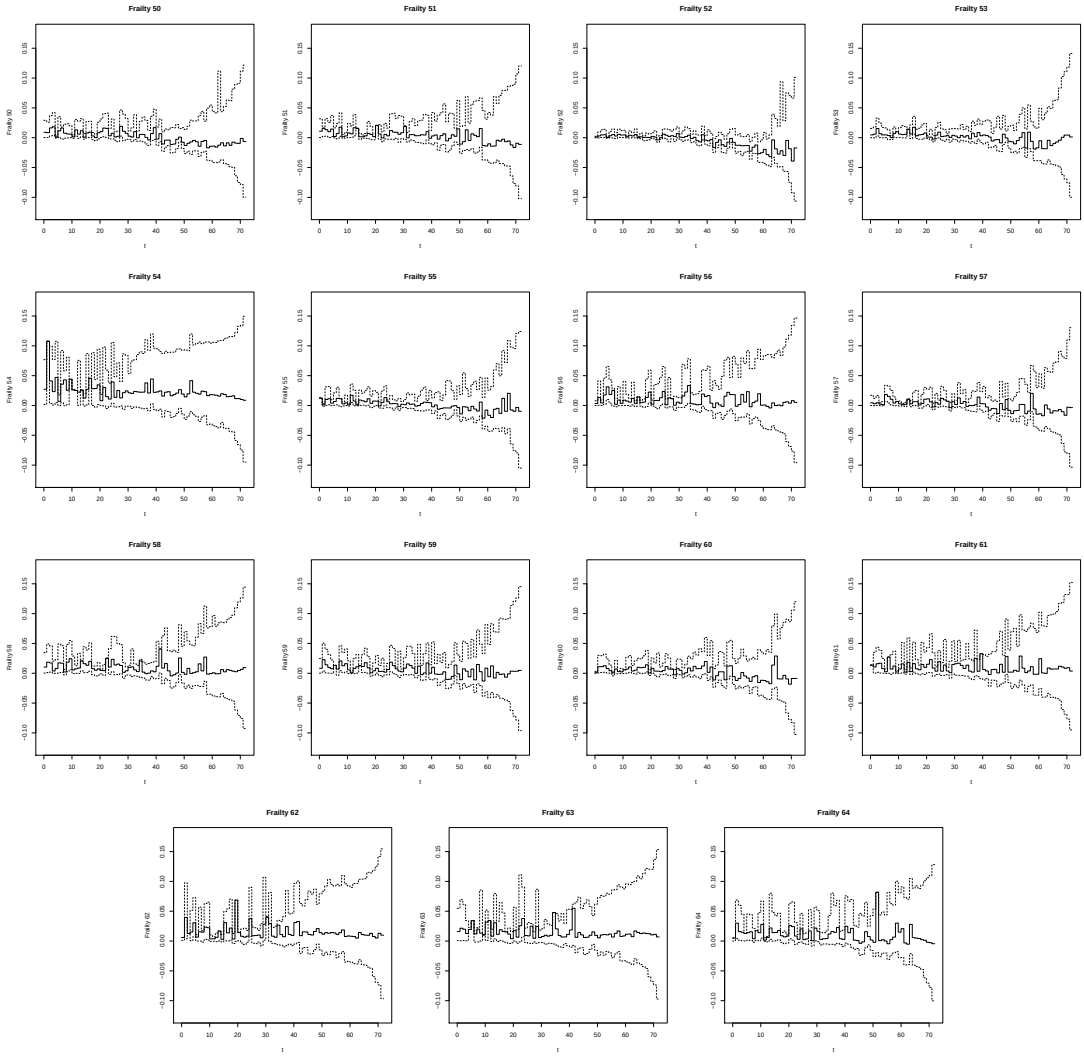


FIGURE 6. Estimated Parameters of The Model (Continuation)



We see that the hazard for such person starts from the small value before 1 month, takes the maximum value between 1 and 2 months and rapidly decreases to the initial value at 5 months. After that it slowly decreases until 30 months and remains approximately constant until 50 months and then increases. The confidence limits for the baseline hazard starting from 65 months are very wide which means that the estimated values are not reliable in this region. The confidence limits for the first 4

month, on the other hand, are almost equal to the median value which means that the estimated value of the baseline hazard is very accurate. Such a difference in accuracy is explained by the number of events (deaths) at these regions. The total number of events in the region $t \geq 65$ is 57 with only 2 events corresponding to the interval 71–72 months and a little more for the preceding intervals. For the interval 0–1 months, on the other hand, there are 2091 events which allows to estimate the hazard with a very good accuracy.

Note that the values of the Regression function 3 (stage) are approximately equal to the negative values of the baseline hazard. This means that for the persons with the 'stage=Localized/regional' the hazard is almost 0. So the stage effect is obviously very significant during the first 4 months. After that the stage is much less important while the hazard is always less for the person with Localized/regional stage.

The age effect (Regression function 1) is significant during the first 5 months and after that it is not significant since the confidence intervals always contain zero.

Regarding the race (Regression function 2), it seems insignificant. However the confidence intervals are skewed towards negative values which indicates that there is a tendency for the black persons to have greater hazard than white.

Now, according to the Regression function 4 and 5 plots, the marital status is not significant after ten months within which the hazard for both Single and Other is greater than that of the Married, and the hazard corresponding to the persons with 'marriage=Other' which includes Divorced, Separated and Widowed, is a little greater than that of Single. While the effect after 10 months does not seem to be

significant, there is still a tendency for it to increase the hazard compared to the Married persons.

Regarding spatial effects, while some frailty effects seem to be insignificant, we can notice the tendency of other ones to be negative or positive. For instance, the values of the 36-th (Orleans) and 52-nd (Saint Tammany) frailty tend to be negative. This means that the hazard of the patients in these counties is less than that of patients in the first county (Acadia parish). For the 33-rd (Madison parish) and 54-th (Tensas parish), however, the frailties are always positive which indicates larger hazards in these counties.

These tendencies agree with the results obtained by Zhang and Lawson (2011). Note that the AFT model used by them considers the effect of the covariates and frailties on the survival time while our model considers the effects on the hazard. So, the positive effect obtained in our model corresponds to the negative effect in theirs and visa versa.

The important thing that need to be mentioned is that the data indicates significant time-dependency of the effects which cannot be caught by the AFT model. This suggests the usage of our model.

CHAPTER 5

Conclusions

We developed a Bayesian Spatial Additive Hazard Survival Model in which all covariates and spatial effects have additive form with respect to the hazard rate and the spatial dependency assumed to have a conditional autoregressive form. All the included effects are allowed to be time-varying which makes the model more general than the models existing in literature.

The estimation of the parameters is made through Markov Chain Monte Carlo sampling from the posterior distribution which is carefully designed to have good convergence properties. In order to provide good convergence, the appropriate proposal distributions were constructed.

The model is implemented in a program which uses the combination of `R` and `C` programming languages and utilizes the multiple precision floating point data types allowing to apply model to the big data.

Using the mentioned computer program, we conducted the simulations to study the performance of the algorithm for different sets of parameters.

Finally, we applied the proposed model to the prostate cancer data from the SEER database, and presented the results of the analysis in the form of plots which were discussed in detail and supplied with necessary comments.

In addition, we presented (but not implemented) the model with the geostatistical spatial structure leaving the implementation as a future research direction.

APPENDIX A

Introduction to Markov Chain Monte Carlo

We use the Markov Chain Monte Carlo (MCMC) method for sampling from the posterior distribution. As discussed for example in Gelfand and Smith (1990), Casella and George (1992) or Tierney (1994a,b), this method allows to obtain samples from any distribution the pdf (or pmf) of which is known up to a multiplicative constant. The basic idea of the method is to generate a Markov Chain which limiting distribution is the same as the desired distribution. Then the states of this Markov Chain can be considered as a sample from this distribution.

1. Gibbs sampler

We will use a particular MCMC algorithm which is called Gibbs sampler (see Gelfand and Smith, 1990; Casella and George, 1992). This method can be expressed as follows. Suppose we need to obtain a sample of I random vectors $\mathbf{U}^{(i)} = \left(U_1^{(i)}, \ldots, U_K^{(i)}\right)^T, i = 1, \ldots, I$ from the multivariate distribution with the pdf $f(\mathbf{U}) = f(U_1, \ldots, U_K)$. Then Gibbs sampler algorithm consists of the following steps:

(1) Begin with some starting set of values $\mathbf{U}^{(0)} = \left(U_1^{(0)}, \ldots, U_K^{(0)}\right)^T$.

(2) On the i-th iteration ($i \in \{1, \ldots, I\}$) we use the following Markovian updating scheme:

(a) draw $U_1^{(i)}$ from the conditional distribution $f\left(U_1 \left| U_2^{(i-1)}, \ldots, U_K^{(i-1)} \right.\right)$;

(b) draw $U_2^{(i)}$ from $f\left(U_2 \left| U_1^{(i)}, U_3^{(i-1)}, \ldots, U_K^{(i-1)}\right.\right)$;

...

(c) draw $U_j^{(i)}$ from $f\left(U_j \left| U_1^{(i)}, \ldots, U_{j-1}^{(i)}, U_{j+1}^{(i-1)}, \ldots, U_K^{(i-1)}\right.\right)$;

...

(d) draw $U_K^{(i)}$ from $f\left(U_K \left| U_1^{(i)}, U_2^{(i)}, \ldots, U_{K-1}^{(i)}\right.\right)$;

(e) set $\mathbf{U}^{(i)} = \left(U_1^{(i)}, \ldots, U_K^{(i)}\right)^T$.

(3) Repeat step 2 until all I desired vectors are obtained.

It is proved that under mild regularity conditions the distribution of the vectors obtained by the algorithm above, converges to the desired distribution $f(U_1, \ldots, U_K)$ (see Gelfand and Smith, 1990, and their references).

However, Gibbs sampler requires *availability* of full conditional distributions, i.e. existence of the efficient algorithm to generate samples directly from these distributions. This is not provided in our case since the full conditionals appear to be non-standard constrained distributions.

Application of Gibbs sampler for constrained distributions is discussed in Gelfand et al. (1992). They offer several methods of sampling from a constrained distribution which belongs to one of the standard ones, and also for some cases of non-standard.

Unfortunately, our conditional distributions don't belong to the class of distributions discussed by them. So we need to use some other methods for sampling from conditionals.

2. Metropolis-Hastings step

If the conditional distribution is complex non-standard distribution as in our case, one should use the Metropolis-Hastings sampling algorithm for sampling from conditional distribution.

The basic idea fo the method is explained in Hastings (1970). He considers the discrete distributions but he mentions that continuous distributions can be more than adequately approximated by the discrete distributions. This is actually what happens in computations because the generated by computer random variables can not be continuous, but they are discrete with very small steps though.

The continuous version of Metropolis-Hastings algorithm is well described in Tierney (1994a). Adopted to our purposes, it can be expressed as follows. Let $p(U_j|U_s, \forall s \neq j)$ be a known function proportional to the desired conditional distribution $f(U_j|U_s, \forall s \neq j)$ which is unknown. Also suppose that we have proposal density function $g(U_j|U_s, \forall s \neq j)$ which is a pdf of some standard distribution which is easy to sample from. Then, the Metropolis-Hastings algorithm for the j-th vector component on the i-th iteration of the Gibbs sampler is the following:

(1) Draw a proposal state U_j' from the distribution with density:

$$g\left(U_j \left| U_1^{(i)}, \ldots, U_{j-1}^{(i)}, U_{j+1}^{(i-1)}, \ldots, U_K^{(i-1)} \right.\right). \tag{A.2.1}$$

(2) Calculate the acceptance ratio AR:

$$AR = \frac{p\left(U_j' \middle| U_1^{(i)}, \ldots, U_{j-1}^{(i)}, U_{j+1}^{(i-1)}, \ldots, U_K^{(i-1)}\right)}{p\left(U_j^{(i-1)} \middle| U_1^{(i)}, \ldots, U_{j-1}^{(i)}, U_{j+1}^{(i-1)}, \ldots, U_K^{(i-1)}\right)} \times \frac{g\left(U_j^{(i-1)} \middle| U_1^{(i)}, \ldots, U_{j-1}^{(i)}, U_{j+1}^{(i-1)}, \ldots, U_K^{(i-1)}\right)}{g\left(U_j' \middle| U_1^{(i)}, \ldots, U_{j-1}^{(i)}, U_{j+1}^{(i-1)}, \ldots, U_K^{(i-1)}\right)}. \quad \text{(A.2.2)}$$

Here we assume that denominators of both fractions are positive. Since all generated values of $\mathbf{U}$ belong to the region with positive density by construction of the MCMC algorithm, then it suffices for the initial state to be within this region to ensure that the denominator of the first fraction is always positive. The second denominator is positive because the value U_j' is generated from the proposal density which provides its positivity at this point.

Therefore, provided that the initial value $\mathbf{U}^{(0)}$ is chosen properly, the acceptance ratio will be always possible to calculate.

(3) Calculate the acceptance probability a based on the acceptance ratio:

$$a = \min\{AR, 1\}. \quad \text{(A.2.3)}$$

(4) Accept the proposal state U_j' with acceptance probability a, i.e. set

$$U_j^{(i)} = \begin{cases} U_j' & \text{with probability } a, \\ U_j^{(i-1)} & \text{with probability } 1-a. \end{cases} \quad \text{(A.2.4)}$$

Theoretically, this algorithm eventually converges to the desired distribution for any proposal density $g(U_j|U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_K)$ which is almost everywhere positive wherever the desired conditional density $f(U_j|U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_K)$ is positive (see Tierney, 1994a, subsection 2.3.3). However, it works best if the proposal is close to the desired conditional density.

Note that actually Metropolis-Hastings algorithm can be used directly for sampling from the multivariate posterior distribution rather than from univariate full conditionals. But in this case it will be very hard to find the appropriate multivariate proposal distribution which would provide good convergence. So we only use one step of Metropolis-Hastings algorithm on each iteration of the Gibbs sampler.

This combined algorithm having one iteration of Metropolis-Hastings algorithm inside the Gibbs sampler, is known as Metropolis-within-Gibbs. Combined Metropolis algorithms are discussed in Tierney (1994a,b). They provide the necessary theoretical background and prove the convergence of such methods. They mention that in the case of being unable to sample from the conditional distribution in Gibbs sampler directly, one can use the approximate algorithms like rejection sampling or grid-based sampling. But in order to ensure that stationary distribution doesn't change, one should embed such algorithm in a Metropolis chain. This guarantees that the equilibrium distribution is exactly the desired distribution no matter how good the approximation is.

Often the proposal contains some parameters which need to be tuned in order to improve the convergence properties. This can be hard to do sometimes. The adaptive algorithms can be used to automate the tuning process during the sampling. Roberts and Rosenthal (2007) discuss the adaptive MCMC algorithms, showing that one should be careful with adopting since it can destroy the convergence to the desired distribution. In particular they discuss the adaptive Metropolis-within-Gibbs algorithm on the example ("Stairway to Heaven"). The adaptive algorithms are very

useful if one uses, for example, Metropolis random walk, i.e. the Metropolis algorithm with proposal having the form of normal distribution centred at the previously generated point. In this case the variance parameter should be tuned to ensure that the rejection doesn't occur too often and at the same time random walk explores the sample space good enough.

In our work we choose to derive the proposals which are close to the desired full conditionals instead of using the methods like random walk which are very simple in implementation but require tuning procedures and suffering from slow convergence.

APPENDIX B

Proofs of Propositions Concerning Full Conditional Distributions

In this Appendix we present the proofs of the propositions stated in Chapter 2.

1. Baseline full conditional distribution

PROOF OF PROPOSITION 4.1. The full conditional distribution of λ_j given all other parameters and data can be obtained by extracting all the terms containing λ_j from the posterior distribution given by the formula (2.3.6). Thus it is proportional to:

$$\pi\left(\lambda_j \mid \boldsymbol{\alpha}_j, \boldsymbol{\omega}_j, \mathbf{D}\right) \propto \prod_{i \in \mathcal{E}_j} \left(\lambda_j + \sum_{k=1}^{p} \alpha_{kj} z_{ikj} + \omega_{l_i}\right) \times \lambda_j^{c_0 r_0 \Delta t_j - 1} \exp\left(-(R_j + c_0)\lambda_j \Delta t_j\right) \times \mathbb{I}\left\{\lambda_j + \sum_{k=1}^{p} \min\left\{\alpha_{kj} \inf \Omega_k, \alpha_{kj} \sup \Omega_k\right\} + \min_{1 \le l \le n} \omega_{lj} > 0\right\}, \ \lambda_j > 0, \quad \text{(B.1.1)}$$

where $\mathcal{E}_j$ is the set of all individuals for which the event occurred in the interval $(t_{j-1}, t_j]$ and $R_j = \sum_{i=1}^{N} R_{ij}$ with R_{ij} being the proportions the individuals are at risk in this interval which are defined in (2.3.3).

So, if there are no events in the interval $(t_{j-1}, t_j]$ then the distribution of λ_j given by (B.1.1) reduces to constrained Gamma. If, in addition, the inequality inside the indicator function in this equation holds for all $\lambda_j > 0$, this distribution becomes

Gamma with the shape and scale parameters $c_0 r_0 \Delta t_j$ and $\frac{1}{(c_0+R_j)\Delta t_j}$ respectively. If however there are events in the interval $(t_{j-1}, t_j]$ the distribution becomes more complex.

Note that with regard to λ_j the first term of the equation (B.1.1) is a polynomial of the power $E_j = \text{card}(\mathcal{E}_j)$ equal to the number of events in the interval $(t_{j-1}, t_j]$. Then the distribution of λ_j has the form of:

$$f_{\lambda_j}(x) \propto \prod_{s=1}^{E_j} (x + c_s) \frac{1}{\varepsilon^\rho \Gamma(\rho)} x^{\rho-1} \exp\left(-\frac{x}{\varepsilon}\right) \mathbb{I}\left\{x > a\right\}, \; x > 0, \tag{B.1.2}$$

where $c_s = \sum_{k=1}^{p} \alpha_{kj} z_{i_s kj} + \omega_{l_{i_s}}$ with i_s being the indexes of the individuals from cE_j; $\rho = c_0 r_0 \Delta t_j$, $\varepsilon = \frac{1}{(c_0+R_j)\Delta t_j}$ and $\Gamma(\rho) = \int_0^\infty \xi^{\rho-1} \exp(-\xi) \mathrm{d}\xi$ is a gamma-function. The constraint a inside the indicator function is computed according to (B.1.1) as:

$$a = -\sum_{k=1}^{p} \min\left\{\alpha_{kj} \inf \Omega_k, \alpha_{kj} \sup \Omega_k\right\} - \min_{1 \le l \le n} \omega_{lj}. \tag{B.1.3}$$

Also we make a convention that in the case of $E_j = 0$ the product $\prod_{s=1}^{0}(x + c_s) \equiv 1$. Note that $c_s \geq -a, \; \forall s = 1 \dots E_j$, which follows directly from the definition of constraint a and coefficients c_s.

This gives us the distribution of exactly the same form as stated. □

Proof of Proposition 4.2. The polynomial $\prod_{s=1}^{E_j}(x + c_s)$ can be expanded into $\sum_{f=0}^{E_j} d_f x^f$ which leads to the following expression for the density:

$$f_{\lambda_j}(x) = \frac{1}{C_{norm} \varepsilon^\rho \Gamma(\rho)} \sum_{f=0}^{E_j} d_f x^{\rho+f-1} \exp\left(-\frac{x}{\varepsilon}\right) \mathbb{I}\left\{x > \max[a, 0]\right\}. \tag{B.1.4}$$

This allows us to find the normalizing constant explicitly:

$$\begin{aligned} C_{norm} &= \int_{\max[a,0]}^{\infty} C_{norm} f_{\lambda_j}(x)\mathrm{d}x \\ &= \sum_{f=0}^{E_j}\left(d_f\varepsilon^f\frac{\Gamma(\rho+f)}{\Gamma(\rho)}\int_{\max[a,0]}^{\infty}\frac{1}{\varepsilon^{\rho+f}\Gamma(\rho+f)}x^{\rho+f-1}\exp\left(-\frac{x}{\varepsilon}\right)\mathrm{d}x\right) \\ &= \sum_{f=0}^{E_j}\left(d_f\varepsilon^f\frac{\Gamma(\rho+f)}{\Gamma(\rho)}\left(1-\frac{\gamma\left(\frac{\max[a,0]}{\varepsilon},\rho+f\right)}{\Gamma(\rho+f)}\right)\right) = \sum_{f=0}^{E_j} d_f\mathcal{I}_f, \end{aligned} \tag{B.1.5}$$

with $\frac{\gamma\left(\frac{x}{\varepsilon},\rho\right)}{\Gamma(\rho)}$ being a CDF of gamma distribution $\mathcal{G}(\rho,\varepsilon)$, where:

$$\gamma(x,\rho) = \int_0^x \xi^{\rho-1}\exp(-\xi)\mathrm{d}\xi \tag{B.1.6}$$

is the lower incomplete gamma function, and we denote:

$$\mathcal{I}_f = \varepsilon^f\frac{\Gamma(\rho+f)}{\Gamma(\rho)}\left(1-\frac{\gamma\left(\frac{\max[a,0]}{\varepsilon},\rho+f\right)}{\Gamma(\rho+f)}\right). \tag{B.1.7}$$

Similarly to normalizing constant, we can now find the mean of this distribution:

$$\mu = \frac{\sum_{f=0}^{E_j} d_f\mathcal{I}_{f+1}}{\sum_{f=0}^{E_j} d_f\mathcal{I}_f}, \tag{B.1.8}$$

which is exactly the formula stated. □

Proof of Proposition 4.3. We will use the log-transformation of the pdf for finding the maximum:

$$\psi(x) = \ln f_{\lambda_j}(x) = \sum_{s=1}^{E_j}\ln(x+c_s) + (\rho-1)\ln x - \frac{x}{\varepsilon} + C, \tag{B.1.9}$$

$$\psi'(x) = \sum_{s=1}^{E_j}\frac{1}{x+c_s} + \frac{\rho-1}{x} - \frac{1}{\varepsilon}, \tag{B.1.10}$$

$$\psi''(x) = -\sum_{s=1}^{E_j}\frac{1}{(x+c_s)^2} - \frac{\rho-1}{x^2}, \tag{B.1.11}$$

where C is a constant not depending on x. Recall that the terms $(x + c_s)$ are all positive provided that $x > a$ which makes the usage of the logarithm possible.

So we search for local maximum of $\psi(x)$ in the region $x > \max\{a, 0\}$. To avoid singularity at the ending point, the optimization algorithm should be designed in such way that it does not allow x to reach the bound.

Now consider the following functions:

$$\psi_{min}(x) = E_j \ln(x + c_{min}) + (\rho - 1)\ln(x) - \frac{x}{\varepsilon} + C, \tag{B.1.12}$$

$$\psi_{max}(x) = E_j \ln(x + c_{max}) + (\rho - 1)\ln(x) - \frac{x}{\varepsilon} + C, \tag{B.1.13}$$

where the function $\psi_{min}(x)$ is obtained from the function $\psi(x)$ by replacing all the coefficients c_s by the minimum of them $c_{min} = \min\limits_{1 \le s \le E_j} c_s$, and $\psi_{max}(x)$ is obtained from $\psi(x)$ by replacing all the coefficients c_s by the maximum $c_{max} = \max\limits_{1 \le s \le E_j} c_s$.

The corresponding first derivatives will then be the following:

$$\psi'_{min}(x) = \frac{E_j}{x + c_{min}} + \frac{\rho - 1}{x} - \frac{1}{\varepsilon}, \tag{B.1.14}$$

$$\psi'_{max}(x) = \frac{E_j}{x + c_{max}} + \frac{\rho - 1}{x} - \frac{1}{\varepsilon}. \tag{B.1.15}$$

The simple expressions for $\psi_{max}(x)$ and $\psi_{min}(x)$ allow us to find the largest extremum for each of them analytically (if it exists):

$$x_L = \frac{1}{2}\left(\varepsilon(\rho + E_j - 1) - c_{max} + \sqrt{(\varepsilon(\rho + E_j - 1) - c_{max})^2 + 4\varepsilon(\rho - 1)c_{max}}\right) \tag{B.1.16}$$

$$x_U = \frac{1}{2}\left(\varepsilon(\rho + E_j - 1) - c_{min} + \sqrt{(\varepsilon(\rho + E_j - 1) - c_{min})^2 + 4\varepsilon(\rho - 1)c_{min}}\right) \tag{B.1.17}$$

where x_L and x_U are the maximums of $\psi_{max}(x)$ and $\psi_{min}(x)$ respectively provided that the expressions under the square roots are non-negative.

Note that since $c_{min} \leq c_s \leq c_{max}$, $\forall s = 1, \ldots, E_j$, then $\psi'_{max}(x) \leq \psi'(x) \leq \psi'_{min}(x)$, $\forall x > \max\{a, 0\}$.

This implies that if $\psi_{min}(x)$ does not have extrema then $\psi'_{min}(x) < 0$, $\forall x > \max\{a, 0\}$, and thus, $\psi'(x) < 0$ in this region from which we conclude that $\psi(x)$ does not have extrema as well and is strictly decreasing.

Note that when $x \to \infty$ the function $\psi(x)$ tends to negative infinity and its derivative approaches a negative value:

$$\lim_{x\to\infty} \psi(x) = -\infty, \quad \lim_{x\to\infty} \psi'(x) = -\frac{1}{\varepsilon} < 0. \tag{B.1.18}$$

The form of the function $\psi'(x)$ tells us that there can be only finite number of solutions for the equation $\psi'(x) = 0$. So if there exist extrema of $\psi(x)$, there are only finite number of them. Since $\psi'(x)$ eventually becomes negative, the greatest extremum can not be local minimum. So it is local maximum. This proves the first statement.

Now since $\hat{x}$ is local maximum, $\psi'(\hat{x}) = 0$ and $\psi'(x) > 0$ for some $x < \hat{x}$. Then $\psi'_{min}(x) > 0$ and so since $\psi'_{min}(x)$ eventually becomes negative and is continuous, there exists an extremum x_U of $\psi_{min}(x)$ which is greater or equal to the extremum of $\psi(x)$, i.e. $x_U \geq \hat{x}$.

Similarly, the existence of the largest extremum x_L of $\psi_{max}(x)$ implies the existence of extrema for $\psi(x)$ and holding of inequality $x_L \leq \hat{x}$. Then as was shown before x_U is defined and $\hat{x} \leq x_U$.

This proves all the statements of the proposition. □

2. Regression function full conditional distribution

Proof of Proposition 4.4. The full conditional distribution of the component of one regression function α_{kj} given all other parameters can be obtained by extracting all the terms containing α_{kj} from the posterior distribution given by the formula (2.3.6). Thus it is proportional to:

$$\begin{aligned}\pi\left(\alpha_{kj} \mid \lambda_j, \alpha_{kj'\neq kj}, \boldsymbol{\omega}_j, \mathbf{D}\right) \propto & \prod_{i\in\mathcal{E}_j\cap\mathcal{C}_{kj}}\left(\lambda_j + \alpha_{kj}z_{ikj} + \sum_{k'\neq k}\alpha_{k'j}z_{ik'j} + \omega_{l_ij}\right) \\ & \times \exp\left(-\left(\sum_{i=1}^{N} R_{ij}z_{ikj}\right)\alpha_{kj}\Delta t_j\right) \\ & \times \mathbb{I}\left\{\alpha_{kj}\inf\Omega_k \geq -\lambda_j - \sum_{k'\neq k}\min\left\{\alpha_{k'j}\inf\Omega_{k'}, \alpha_{k'j}\sup\Omega_{k'}\right\} - \min_{1\leq l\leq n}\omega_{lj}\right\} \\ & \times \mathbb{I}\left\{\alpha_{kj}\sup\Omega_k \geq -\lambda_j - \sum_{k'\neq k}\min\left\{\alpha_{k'j}\inf\Omega_{k'}, \alpha_{k'j}\sup\Omega_{k'}\right\} - \min_{1\leq l\leq n}\omega_{lj}\right\} \\ & \times \mathbb{I}\left\{0\in\Omega_k \Rightarrow 0 \geq -\lambda_j - \sum_{k'\neq k}\min\left\{\alpha_{k'j}\inf\Omega_{k'}, \alpha_{k'j}\sup\Omega_{k'}\right\} - \min_{1\leq l\leq n}\omega_{lj}\right\},\end{aligned} \tag{B.2.1}$$

where $\mathcal{C}_{kj}$ is the set of individuals whose k-th covariate takes non-zero value at time t_j, i.e.

$$\mathcal{C}_{kj} = \left\{i : \ z_{ikj} \neq 0\right\}. \tag{B.2.2}$$

The first term in (B.2.1) is the polynomial of the power $q = \mathrm{card}(\mathcal{E}_j \cap \mathcal{C}_{kj})$ which is equal to the number of individuals who had an event in the interval $(t_{j-1}, t_j]$ and whose k-th covariate was not zero at the moment of event.

The last three terms (indicator functions) represent constraints. If $\boldsymbol{\Omega}_k$ contains only non-negative values then the distribution is constrained from the left, if it contains only non-positive values then the distribution is constrained from the right, and

if it contains both positive and negative values then the distribution is constrained from both sides. The last constraint is introduced for mathematical completeness and is actually a constraint on the variables in the condition rather than on α_{kj}. Since every set $\mathbf{\Omega}_k$ contains 0 (which makes $\lambda(t)$ to be interpretable as the baseline hazard), this constraint basically tells us that the right hand side of the first two restricting inequalities is always non-positive.

Since $z_{ikj} \neq 0$, $\forall i \in \mathcal{E}_j \cap \mathcal{C}_{kj}$, then we can divide each term in the product by corresponding z_{ikj} and obtain the distribution of the following form:

$$f_{\alpha_{kj}}(x) \propto \left(\prod_{s=1}^{q} (x + c_s) \right) \exp\left(-\varepsilon x\right) \mathbb{I}\left\{a < x < b\right\}, \tag{B.2.3}$$

where the coefficients of the polynomial c_s and the parameter ε are defined as:

$$c_s = \frac{1}{z_{ikj}} \left(\lambda_j + \sum_{k' \neq k} \alpha_{k'j} z_{ik'j} + \omega_{l_{i_s} j} \right), \tag{B.2.4}$$

$$\varepsilon = \left(\sum_{i=1}^{N} R_{ij} z_{ikj} \right) \Delta t_j, \tag{B.2.5}$$

with the indices i_s being the indices of the individuals from $\mathcal{E}_j \cap \mathcal{C}_{kj}$.

In the case of $q = 0$ the convention is made that $\prod_{s=1}^{0}(x + c_s) \equiv 1$. Also note that in the case when all individuals which are at risk in the interval $(t_{j-1}, t_j]$ have the k-th covariate equal to 0, it is impossible to estimate α_{kj} from the data. In this case one should merge the interval $(t_{j-1}, t_j]$ with the adjacent intervals in order to obtain at least one individual at risk with non-zero k-th covariate.

The constraints $-\infty \leq a < b \leq \infty$ are constants calculated according to (B.2.1) as follows:

$$\begin{aligned} a &= \begin{cases} \frac{C_{constr}}{\sup \mathbf{\Omega}_k} & \text{if } \sup \mathbf{\Omega}_k > 0, \\ -\infty & \text{otherwise,} \end{cases} \\ b &= \begin{cases} \frac{C_{constr}}{\inf \mathbf{\Omega}_k} & \text{if } \inf \mathbf{\Omega}_k < 0, \\ +\infty & \text{otherwise,} \end{cases} \end{aligned} \tag{B.2.6}$$

where C_{constr} is the right side of the restricting inequalities inside the indicator functions of (B.2.1):

$$C_{constr} = \left(-\lambda_j - \sum_{k' \neq k} \min \left\{ \alpha_{k'j} \inf \Omega_{k'}, \alpha_{k'j} \sup \Omega_{k'} \right\} - \min_{1 \leq l \leq n} \omega_{lj} \right), \tag{B.2.7}$$

which is non-positive as was already told before.

Note that all the coefficients c_s of the polynomial in (B.2.3) satisfy exactly one of the conditions $-c_s \leq a \leq 0$ or $-c_s \geq b \geq 0$ which implies that $\operatorname{sign}(x + c_s) = const$, $\forall x : a < x < b$. □

Proof of Proposition 4.5. Expanding the product $\prod_{s=1}^{q}(x + c_s)$ to the form $\sum_{f=0}^{q} d_f x^f$ we can rewrite the pdf $f_{\alpha_{kj}}(x)$ as:

$$f_{\alpha_{kj}}(x) = \frac{1}{C_{norm}} \sum_{f=0}^{q} d_f x^f \exp(-\varepsilon x) \mathbb{I}\{a \leq x \leq b\}. \tag{B.2.8}$$

As before, the normalizing constant C_{norm} can be obtained by integration of each term in the summation(B.2.8). The recurrence relation between the integrals $\mathcal{I}_f = \int_a^b x_f \exp(-\varepsilon x) \mathrm{d}x$ used for integrating the summation (B.2.8) in the case of $\varepsilon \neq 0$ is the following:

$$\mathcal{I}_0 = \frac{1}{\varepsilon} \left(\exp(-\varepsilon a) - \exp(-\varepsilon b) \right), \tag{B.2.9}$$

$$\mathcal{I}_f = \frac{1}{\varepsilon} \left(a^f \exp(-\varepsilon a) - b^f \exp(-\varepsilon b) + f \mathcal{I}_{f-1} \right), \quad f = 1, \ldots, q, \tag{B.2.10}$$

and in the case of $\varepsilon = 0$ these integrals become:

$$\mathcal{I}_f = \frac{x^{f+1}}{f+1}, \quad f = 0, \dots, q. \tag{B.2.11}$$

Then the normalizing constant is equal to:

$$C_{norm} = \sum_{f=0}^{q} d_f \mathcal{I}_f, \tag{B.2.12}$$

and the mean of this distribution can be obtained as:

$$\mu = \frac{\sum_{f=0}^{q} d_f \mathcal{I}_{f+1}}{\sum_{f=0}^{q} d_f \mathcal{I}_f}. \tag{B.2.13}$$

□

Proof of Proposition4.6. Firstly, consider the relation between the sign of ε and the infiniteness of a or b.

If both a and b are finite, it means that $\inf \Omega_k < 0$ and $\sup \Omega_k > 0$. So the covariates z_{ikj} of individuals can have any sign and therefore ε can have any sign as well.

If a is finite and b is infinite then it means that $\inf \Omega_k = 0$ and so all $z_{ikj} \geq 0$. Then $\varepsilon \geq 0$ and in case when there is at least one individual at risk with non-zero k-th covariate, it becomes strictly positive: $\varepsilon > 0$. The case when there are no such individuals is not considered since in this case it is impossible to estimate α_{kj} from the data.

Similarly, if a is infinite and b is finite, $\varepsilon < 0$.

The case of both a and b infinite is impossible since Ω_k contains non-zero values and so ether $\inf \Omega_k < 0$ or $\sup \Omega_k > 0$ or both this conditions are satisfied. This implies that at least one of a and b is finite.

In the case of $q = 0$, the function $f_{\alpha_{kj}}(x)$ is just the exponent and so the correctness of the stated in this proposition statements is obvious.

In the case of $q \neq 0$ we study the behaviour of function $f_{\alpha_{kj}}(x)$ using its logarithm transformation:

$$\psi(x) = \ln f_{\alpha_{kj}}(x) = \sum_{s=1}^{q} \ln|x + c_s| - \varepsilon x, \tag{B.2.14}$$

$$\psi'(x) = \sum_{s=1}^{q} \frac{1}{x + c_s} - \varepsilon, \tag{B.2.15}$$

$$\psi''(x) = -\sum_{s=1}^{q} \frac{1}{(x + c_s)^2}. \tag{B.2.16}$$

The second derivative is always negative, so the function $\psi(x)$ is strictly concave and therefore it can have at most one extremum which should be maximum. Then the stated cases of function behaviour are obvious.

Now, consider the behaviour of modified Newton-Raphson algorithm.

If both a and b are finite, then setting $L = a$ and $U = b$ will provide that the algorithm returns a value between a and b.

Consider now the case when a is finite and b is infinite.

The coefficients c_s all satisfy $c_s \geq -a \geq 0$ and so $x + c_s \geq 0$. Then the function $\psi'(x)$ is positive near $x = -c_{min}$ and negative at infinity:

$$\lim_{x \to -c_{min}^{+}} \psi'(x) = +\infty > 0, \tag{B.2.17}$$

$$\lim_{x \to +\infty} \psi'(x) = -\varepsilon < 0. \tag{B.2.18}$$

Since $\psi'(x)$ is continuous for $x > -c_{min}$ there exists a point where $\psi'(x) = 0$. Since the function is concave, the Newton-Raphson algorithm will converge to this point,

and so our modified Newton-Raphson algorithm will converge to this point or will approach a if this point is less than a.

The case of infinite a and finite b is symmetric.

Since the case when both a and b are infinite is impossible, we conclude that the statement of the proposition about the Newton-Raphson algorithm is true. □

3. Frailty full conditional distribution

Proof of Proposition 4.7. The full conditional distribution of the frailty ω_{lj} given all other parameters can be obtained by extracting all the terms containing ω_{lj} from the posterior distribution given by the formula (2.3.6). Thus it is proportional to:

$$\begin{aligned}\pi\left(\omega_{lj} \mid \lambda_j, \boldsymbol{\alpha}_j, \omega_{l'j\neq lj}, \theta_j^2, \mathbf{D}\right) &\propto \prod_{i\in\mathcal{E}_j\cap\mathcal{S}_l}\left(\lambda_j + \sum_{k=1}^{p}\alpha_{kj}z_{ikj} + \omega_{lj}\right)\exp\left(-R_j^{(l)}\omega_{lj}\Delta t_j\right)\\ &\quad\times\exp\left(-\frac{1}{2\theta_j^2}m_l(\omega_{lj}-\overline{\omega}_{lj})^2\right)\\ &\quad\times\mathbb{I}\left\{\omega_{lj}\geq -\lambda_j - \sum_{k=1}^{p}\min\left\{\alpha_{kj}\inf\Omega_k, \alpha_{kj}\sup\Omega_k\right\}\right\}, \quad \text{(B.3.1)}\end{aligned}$$

where $\mathcal{S}_l$ is the set of all individuals which belong to the l-th group, $R_j^{(l)} = \sum_{i\in\mathcal{S}_l} R_{ij}$ is the proportion of time the individuals in l-th group are at risk in the interval $(t_{j-1}, t_j]$.

This distribution has the following form:

$$f_{\omega_{lj}}(x) = \left(\prod_{s=1}^{q}(x+c_s)\right)\frac{1}{\sqrt{2\pi\delta^2}}\exp\left(-\frac{(x-\mu_0)^2}{2\delta^2}\right)\mathbb{I}\left\{a_1 < x < b_1\right\}, \qquad \text{(B.3.2)}$$

where the power of the polynomial is $q = \text{card}(\mathcal{E}_j \cap \mathcal{S}_l)$, and the limits a_1 and b_1 are the following:

$$a_1 = -\lambda_j - \sum_{k=1}^{p} \min \left\{ \alpha_{kj} \inf \Omega_k, \alpha_{kj} \sup \Omega_k \right\}, \tag{B.3.3}$$

$$b_1 = +\infty. \tag{B.3.4}$$

The parameters μ_0 and δ of this distribution are computed as:

$$\mu_0 = \frac{\theta_j^2}{m_l} R_j^{(l)} \Delta t_j + \overline{\omega}_{lj}, \tag{B.3.5}$$

$$\delta^2 = \frac{\theta_j^2}{m_l}. \tag{B.3.6}$$

□

Proof of Proposition 4.8. Expanding the product and applying the transformation $x = \frac{y-\mu_0}{\delta}$ we obtain the following distribution:

$$f_Y(y) \propto \left(\sum_{f=0}^{q} d_f y^f \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right) \mathbb{I}\left\{ a < y < b \right\}, \tag{B.3.7}$$

where the coefficients of the polynomial are changed in accordance with expansion and transformation, and the constraints a_1 and b_1 are changed to $a = \frac{a_1-\mu_0}{\delta}$ and $b = \frac{b_1-\mu_0}{\delta}$ respectively.

This distribution is very similar to constrained normal. So normal distribution can be used as proposal. We take normal proposal with mean equal to the mean of the distribution in equation (B.3.7) and variance adjusted to fit the shape of this distribution.

In order to obtain the mean, we need to find the normalizing constant first. It can be obtained by integrating the function given by formula (B.3.7). It is straightforward

if we know the integrals of the form:

$$\mathcal{I}_f = \int_a^b \frac{1}{\sqrt{2\pi}} y^f \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y, \quad f = 0, 1, 2, \ldots \tag{B.3.8}$$

For $f = 0$ and 1 this integral will be:

$$\mathcal{I}_0 = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = \Phi(b) - \Phi(a), \tag{B.3.9}$$

$$\begin{aligned}\mathcal{I}_1 = \int_a^b \frac{1}{\sqrt{2\pi}} y \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = -\int_{y=a}^b \mathrm{d}\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\right) \\ = -\varphi(y)\Big|_a^b = \varphi(a) - \varphi(b),\end{aligned} \tag{B.3.10}$$

where $\Phi(y)$ and $\varphi(y)$ are the CDF and pdf of the standard normal distribution, respectively.

Now $\mathcal{I}_f$ can be found recursively for all $f = 2, 3, \ldots$

$$\begin{aligned}\mathcal{I}_f = \int_a^b \frac{1}{\sqrt{2\pi}} y^f \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = -\int_{y=a}^b y^{f-1} \mathrm{d}\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\right) \\ = -y^{f-1}\varphi(y)\Big|_a^b + (f-1)\int_a^b \frac{1}{\sqrt{2\pi}} y^{f-2} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y \\ = a^{f-1}\varphi(a) - b^{f-1}\varphi(b) + (f-1)\mathcal{I}_{f-2}.\end{aligned} \tag{B.3.11}$$

Then the normalizing constant for the function given by equation (B.3.7) is simply:

$$C_{norm}^Y = \sum_{f=0}^q d_f \mathcal{I}_f. \tag{B.3.12}$$

So, the exact expression for the $f_Y(y)$ in (B.3.7) is the following:

$$f_Y(y) = \frac{\left(\sum_{f=0}^q d_f y^f\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \mathbb{I}\{a < y < b\}}{\sum_{f=0}^q d_f \mathcal{I}_f}. \tag{B.3.13}$$

The mean of this distribution can be found similarly:

$$\mu_Y = \frac{\sum_{f=0}^q d_f \mathcal{I}_{f+1}}{\sum_{f=0}^q d_f \mathcal{I}_f}. \tag{B.3.14}$$

The mean of the $f_{\omega_{lj}}(x)$ can be found by back-transformation to x as:

$$\mu = \delta\mu_Y + \mu_0 \tag{B.3.15}$$

The normalizing constant for $f_{\omega_{lj}}(x)$ can be found by back transformation as well:

$$C_{norm} = C^Y_{norm}\delta^q \tag{B.3.16}$$

□

Proof of Proposition 4.9. The log-transformed pdf and its derivatives have the following form here:

$$\psi(x) = \ln f_{\omega_{lj}}(x) = \sum_{s=1}^{q} \ln(x + c_s) - \frac{(x-\mu_0)^2}{2\delta^2}, \tag{B.3.17}$$

$$\psi'(x) = \sum_{s=1}^{q} \frac{1}{x + c_s} - \frac{x-\mu_0}{\delta^2}, \tag{B.3.18}$$

$$\psi''(x) = -\sum_{s=1}^{q} \frac{1}{(x + c_s)^2} - \frac{1}{\delta^2}. \tag{B.3.19}$$

The second derivative is always strictly negative, so if the extremum exists, it is the only extremum which function $\psi(x)$ can have and it is maximum.

Using the fact that:

$$\lim_{x \to -c^+_{min}} \psi'(x) = +\infty > 0, \tag{B.3.20}$$

$$\lim_{x \to +\infty} \psi'(x) = -\infty < 0, \tag{B.3.21}$$

and that $\psi'(x)$ is continuous in $x > -c_{min}$, where $c_{min} = \min_{1 \le s \le q}\{c_s\}$ and we assume that $c_{min} = +\infty$ if $q = 0$, we get that extremum always exists (if $q = 0$ this is the usual maximum of normal distribution at μ_0). So our modified Newton-Raphson algorithm will converge to this extremum or to the limit a_1 if extremum is less than this limit.

So the algorithm will not tend to infinity and we do not need the upper limit and can simply use $U = +\infty$. □

APPENDIX C

Modified Newton-Raphson Algorithm for Finding the Extremum in an Open Interval

The original Newton-Raphson algorithm consists in producing the sequence of points x_r which is expected to converge to the extremum:

$$x_r = x_{r-1} - \frac{\psi'(x_{r-1})}{\psi''(x_{r-1})}, \quad r = 1, \ldots, M, \tag{C.0.1}$$

where the initial value x_0 is chosen arbitrary but such that it provides the convergence of the algorithm, and the number M is chosen such that the consecutive points or the values of function at these points become sufficiently close to each other. We will use the closeness of points as the criterium of choosing M:

$$M = \min\{r \geq 1 : |x_r - x_{r-1}| < \delta\}, \tag{C.0.2}$$

where δ is a prespecified value representing the desired accuracy.

This algorithm attempts to solve the equation $\psi'(x) = 0$ the solution of which is expected to be the desired extremum.

Note that this algorithm does not suppose any constraints on x. So in order to solve the optimization problem in a particular region we need to modify this algorithm not allowing x_r to take values outside that region. Also since the ending points can contain singularities we do not allow x_r to reach them as well.

So, we propose the following modified algorithm of finding the extremum in the opened interval (L, U):

$$x_r = \begin{cases} x_r^{(0)}, & \text{if } L < x_r^{(0)} < U, \\ \zeta x_{r-1} + (1-\zeta)L, & \text{if } x_r^{(0)} \leq L, \\ \zeta x_{r-1} + (1-\zeta)U, & \text{if } x_r^{(0)} \geq U, \end{cases} \tag{C.0.3}$$

where $x_r^{(0)}$ is the value obtained using the regular Newton-Raphson formula (C.0.1), ζ is some value from the interval $(0, 1)$ and the limits L and U satisfy the condition $-\infty \leq L < U \leq \infty$, i.e. the upper bound is strictly greater than the lower and any or both of them can be infinite. We choose the value ζ to be 0.01 which does not allow x_r to reach L or U but allows it to come very close to them if the Newton-Raphson method attempts to take the next value outside the interval (L, U).

Also we limit the maximum number of iterations by some value M_{max} (which is chosen to be 500) to avoid the possible infinite cycle.

In the case of zero second derivative or undefined first or second derivative at some iteration, the algorithm stops and returns the last found value. This applies also to the case when the second derivative is zero at the initial point x_0.

Note that our modified Newton-Raphson algorithm does not guarantee that the extremum is found. Particularly in the case of non-existence of extremum in the interval or existence of several extrema. But it is worth mentioning, that the algorithm produces some well-defined value of x in a finite number of steps for any input, i.e. for any function $\psi(x)$ and any initial value $x_0 \in (L, U)$, and ensures that this value belongs to the interval (L, U). This is a very important property of the algorithm in our application.

APPENDIX D

Results of Simulations

FIGURE 7. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 1$ and number of iterations $I = 100$

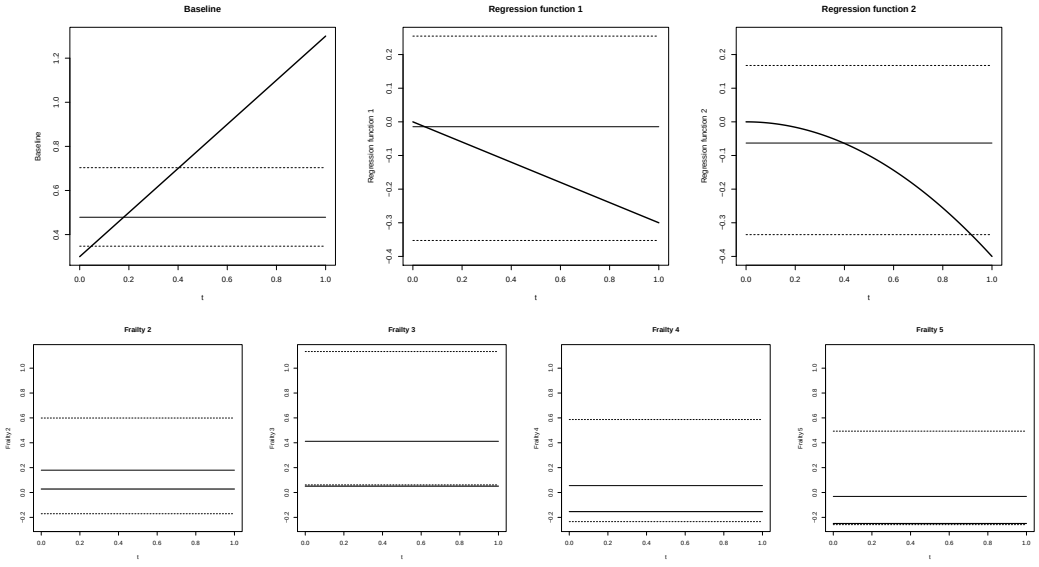


FIGURE 8. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 1$ and number of iterations $I = 500$

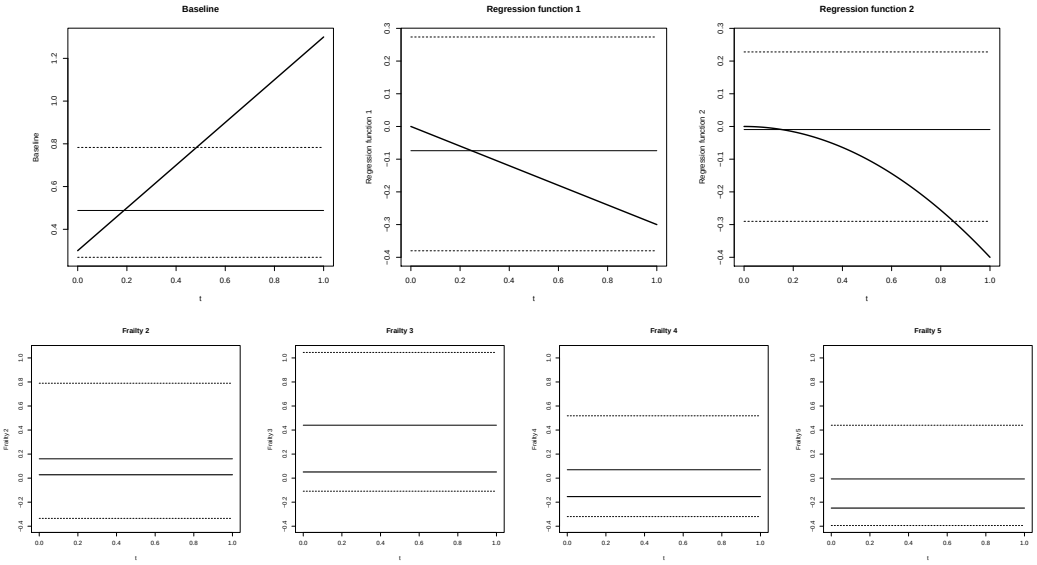


FIGURE 9. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 1$ and number of iterations $I = 1000$

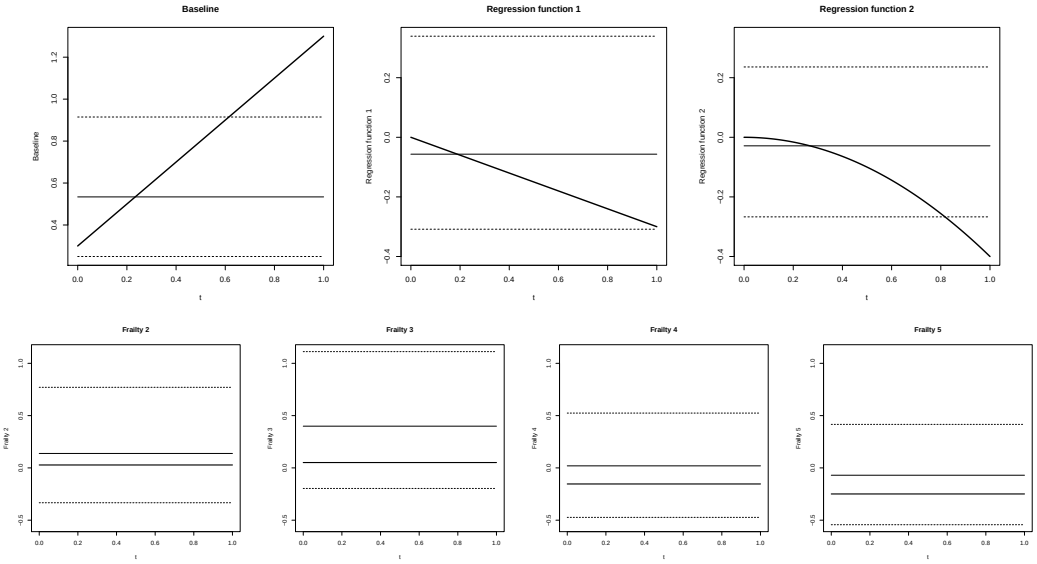


FIGURE 10. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 1$ and number of iterations $I = 5000$

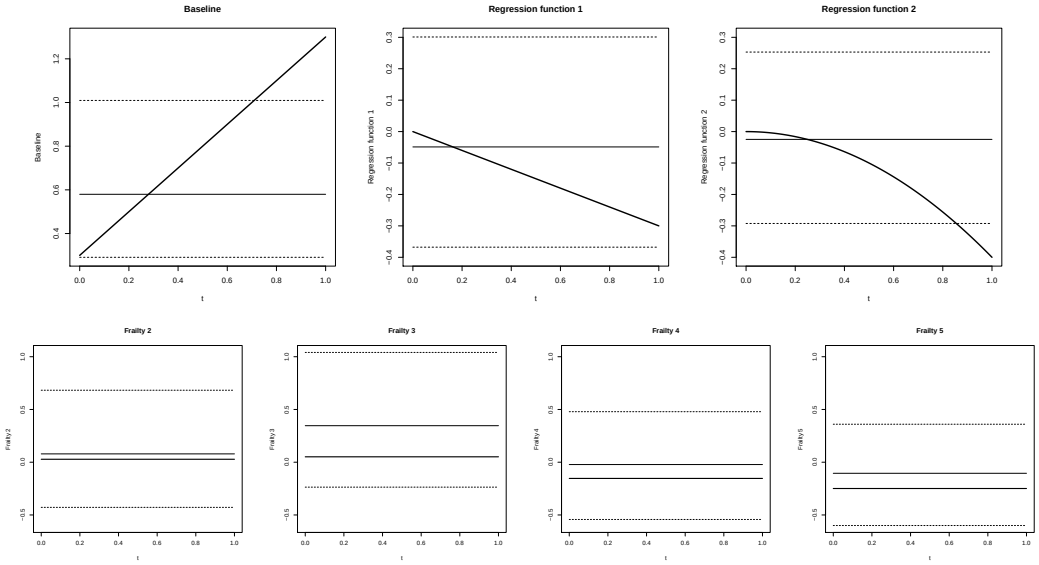


FIGURE 11. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 5$ and number of iterations $I = 100$

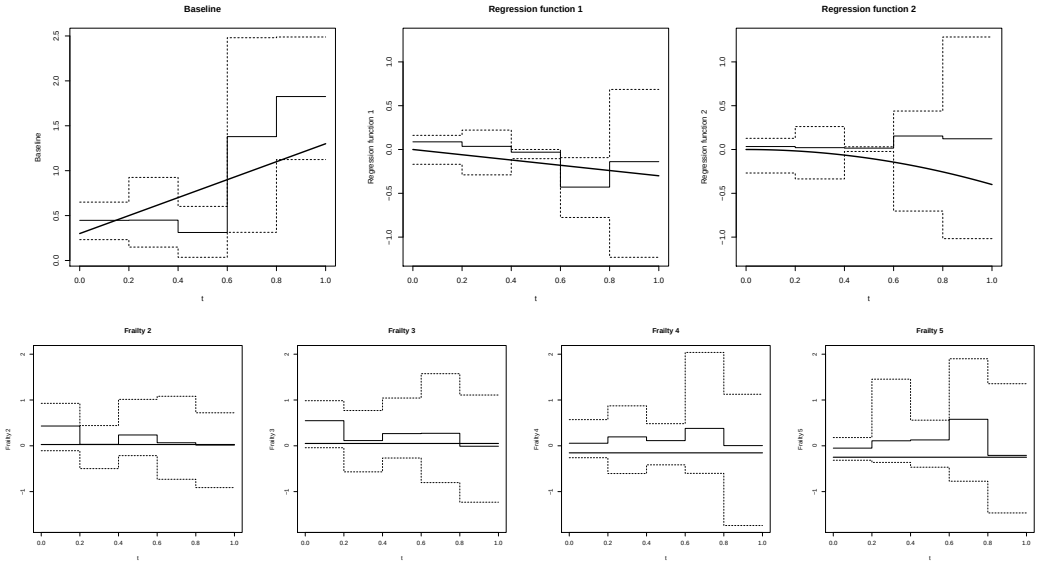


FIGURE 12. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 5$ and number of iterations $I = 500$

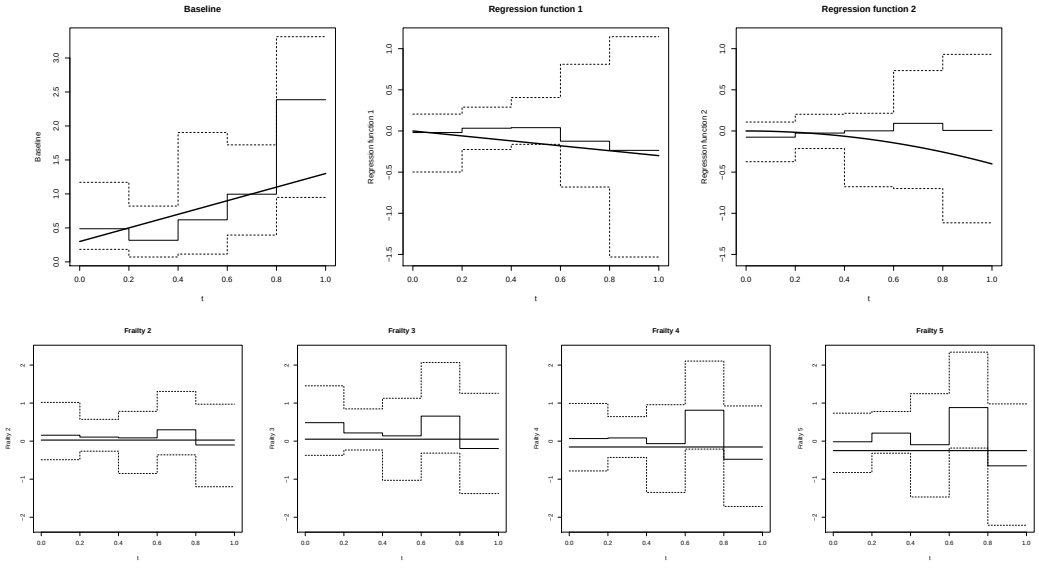


FIGURE 13. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 5$ and number of iterations $I = 1000$

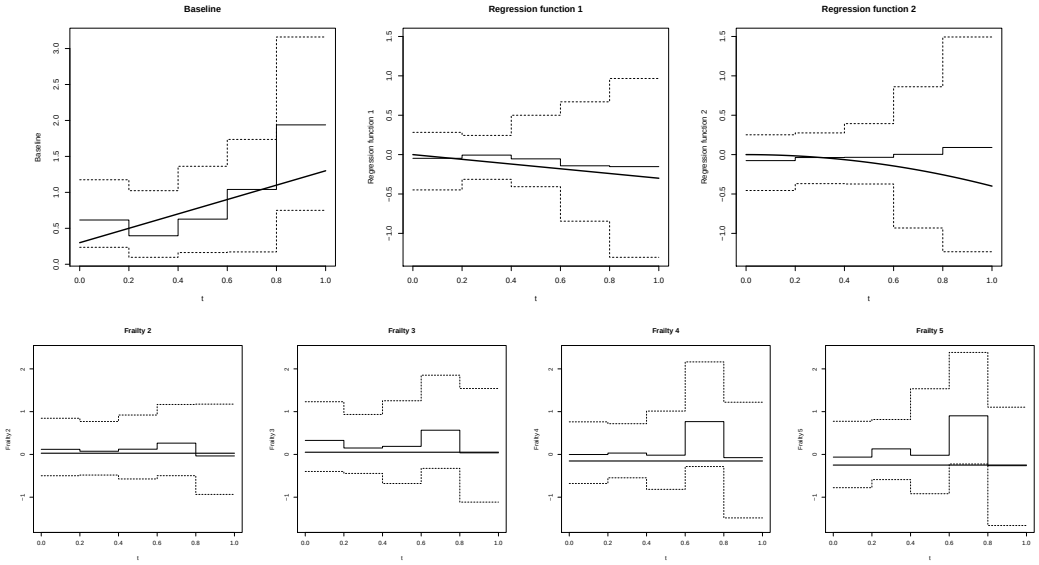


FIGURE 14. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 5$ and number of iterations $I = 5000$

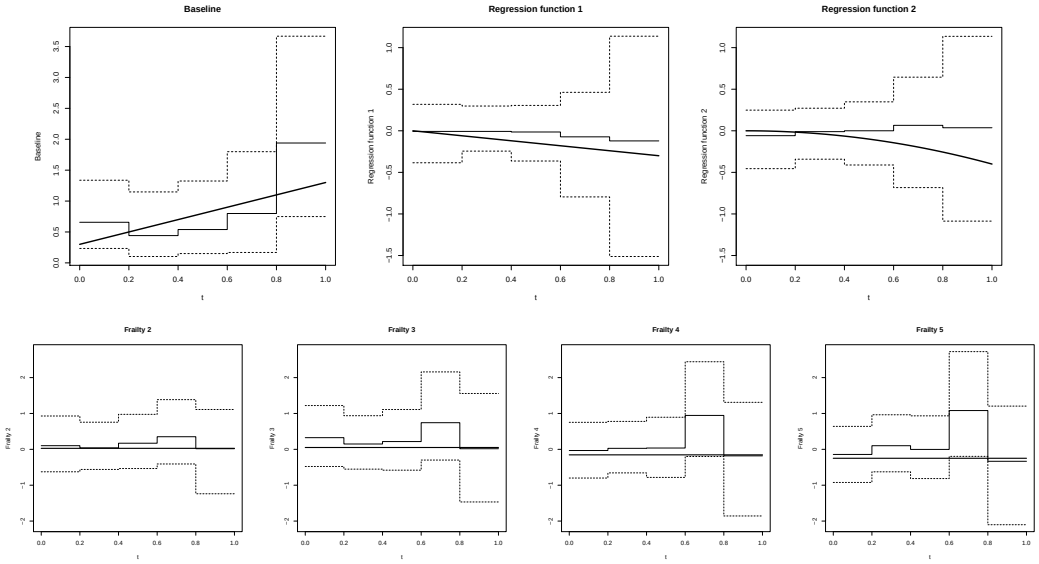


FIGURE 15. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 10$ and number of iterations $I = 100$

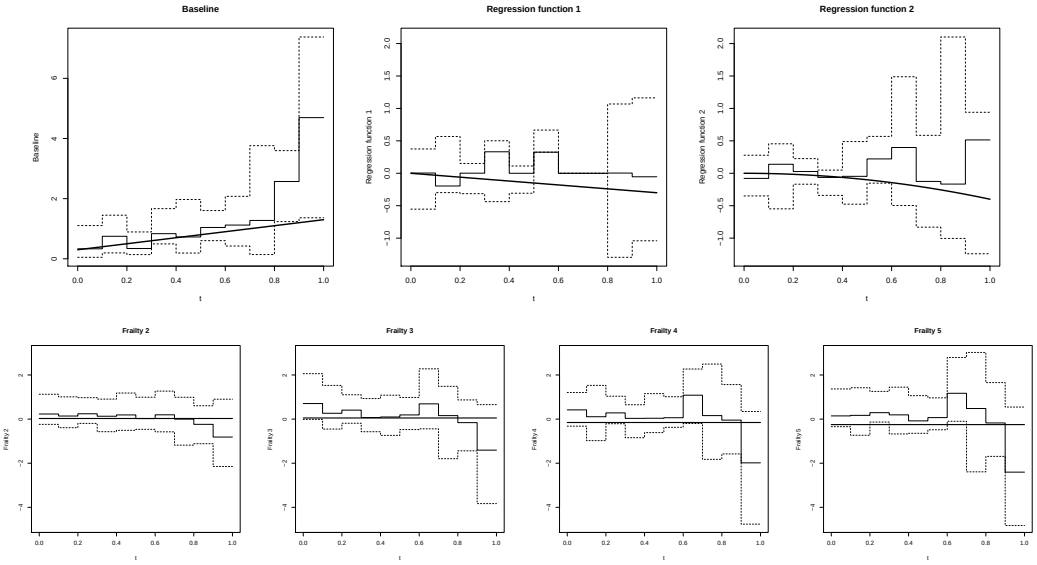


FIGURE 16. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 10$ and number of iterations $I = 500$

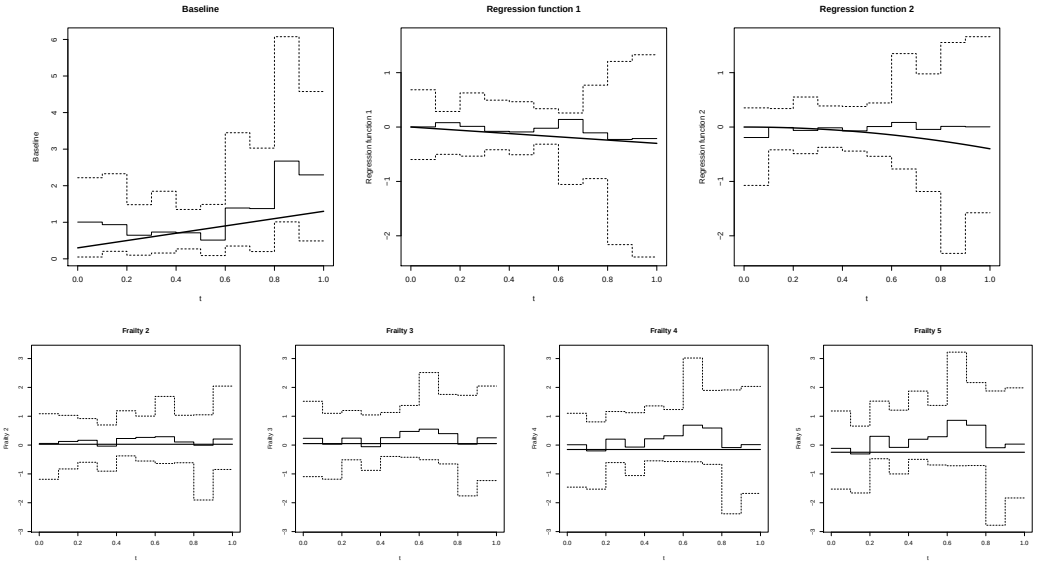


FIGURE 17. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 10$ and number of iterations $I = 1000$

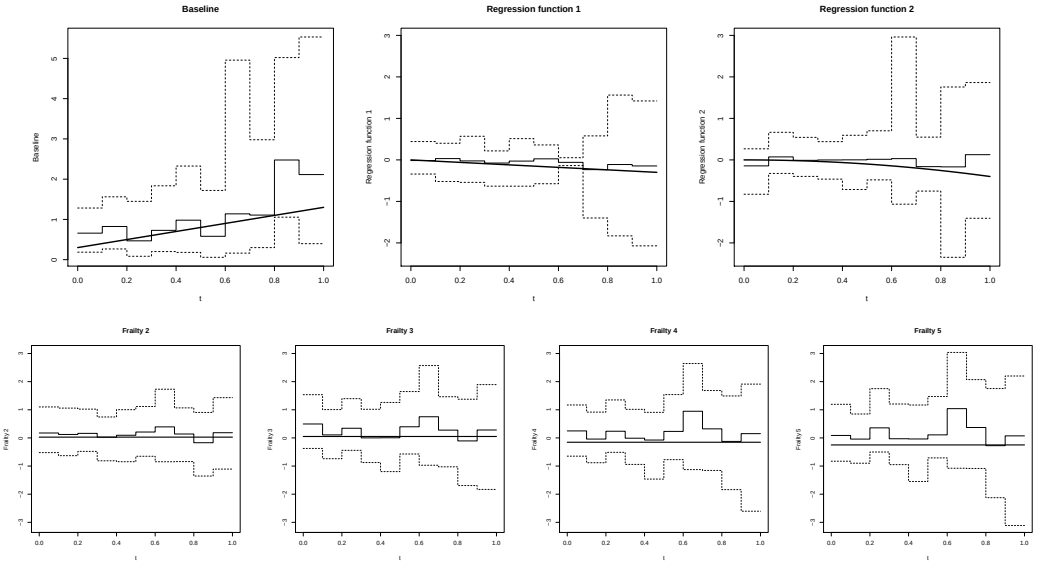


FIGURE 18. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 10$ and number of iterations $I = 5000$

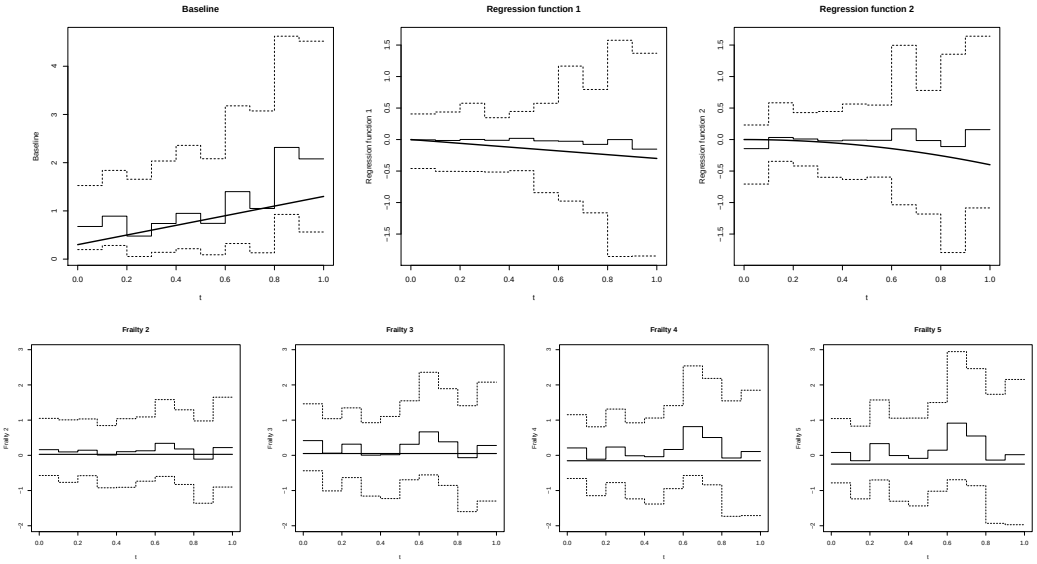


FIGURE 19. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 50$ and number of iterations $I = 100$

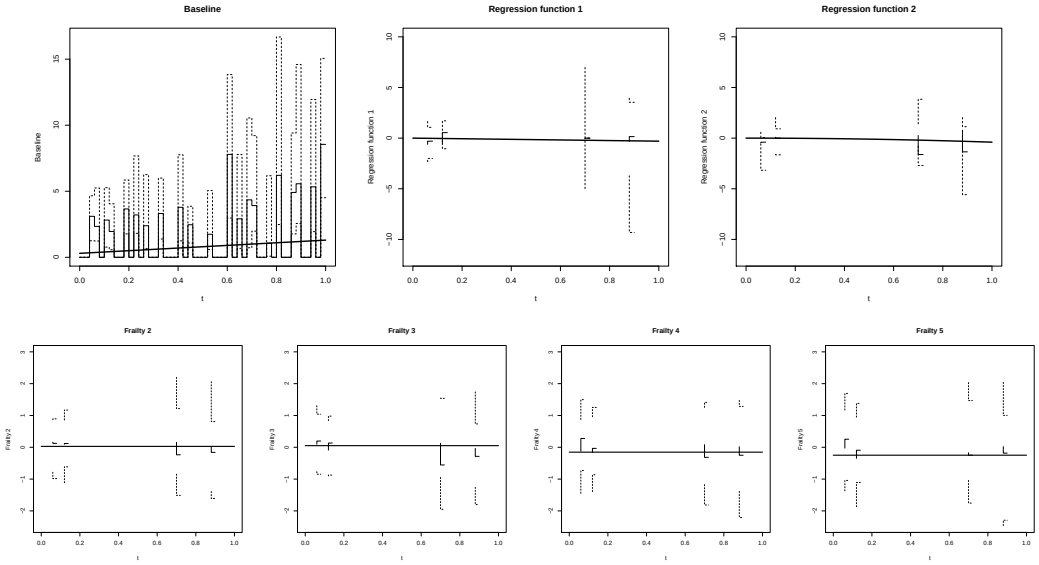


FIGURE 20. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 50$ and number of iterations $I = 500$

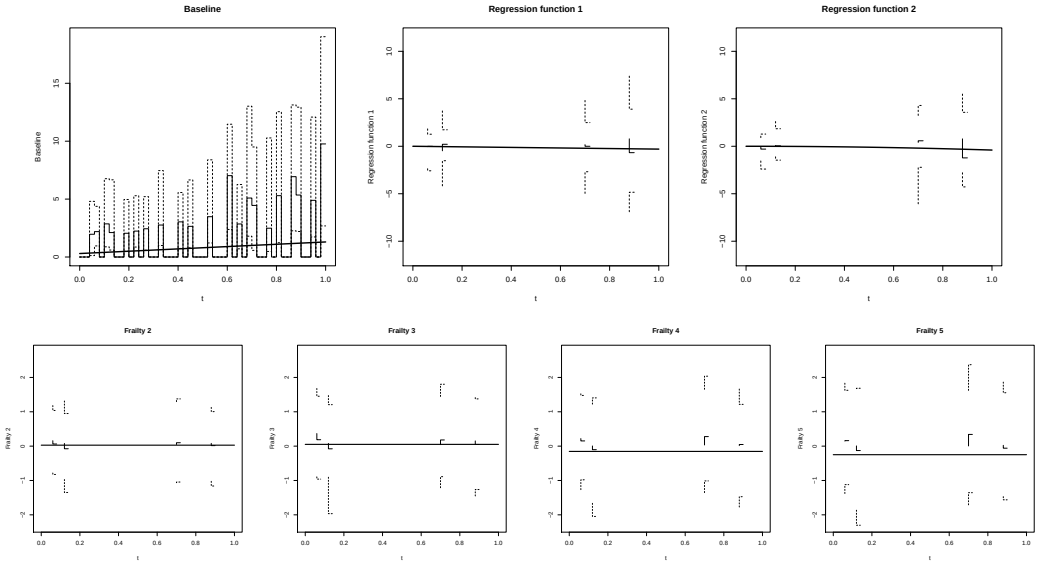


FIGURE 21. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 50$ and number of iterations $I = 1000$

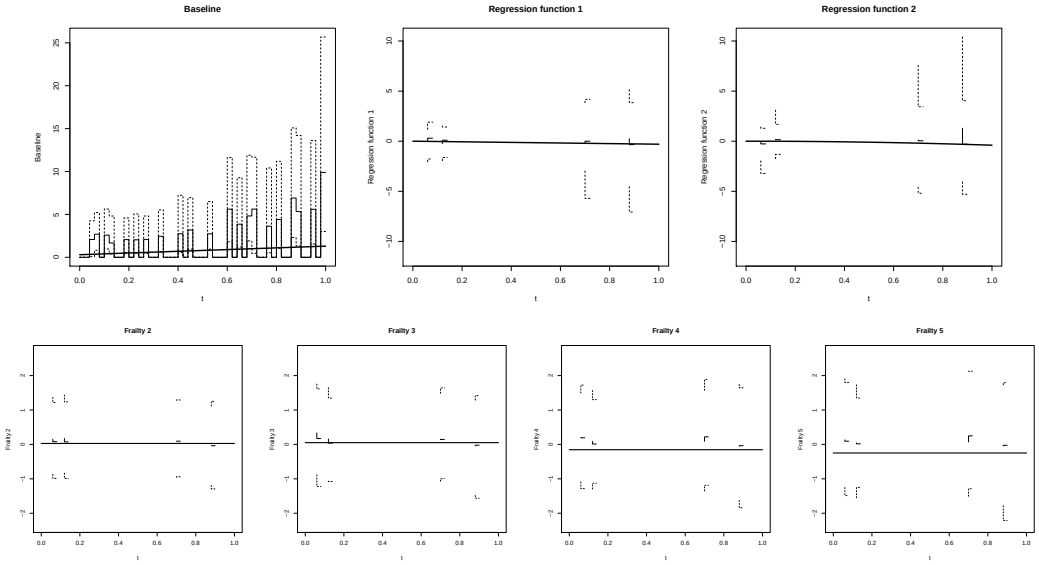


FIGURE 22. Estimated parameters for number of observations $N = 100$, number of breakpoints $m = 50$ and number of iterations $I = 5000$

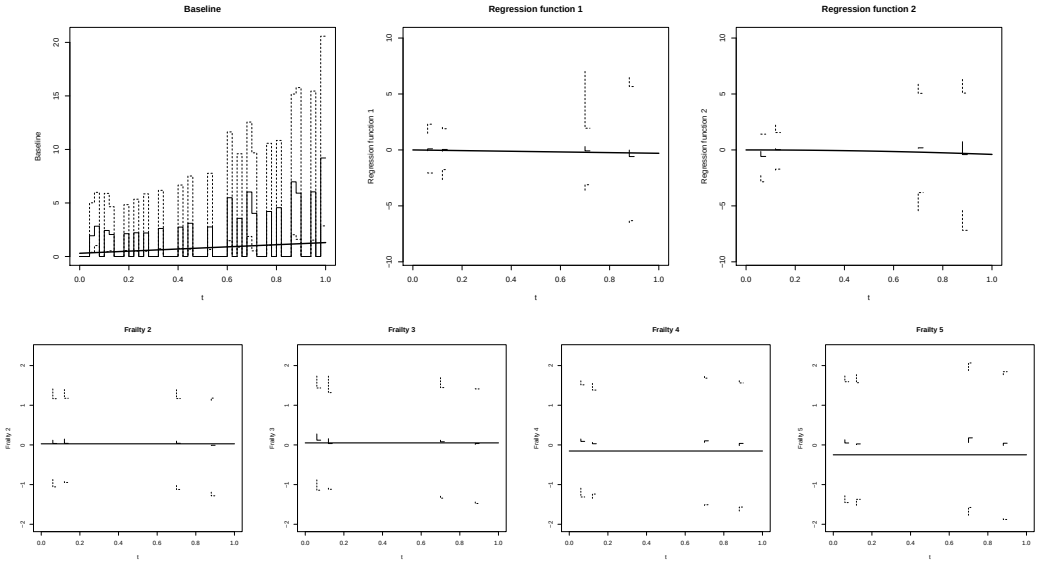


FIGURE 23. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 1$ and number of iterations $I = 100$

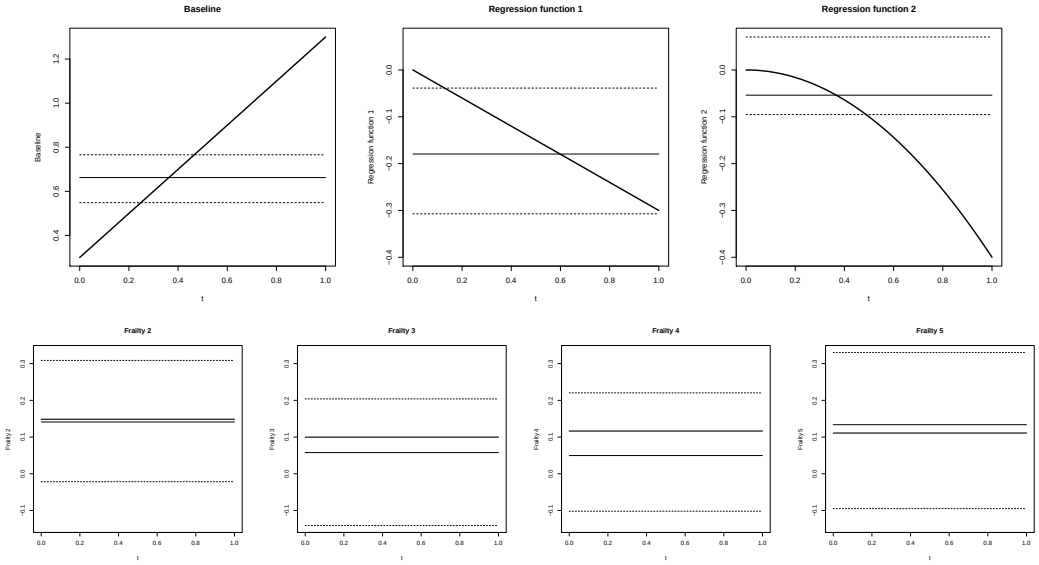


FIGURE 24. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 1$ and number of iterations $I = 500$

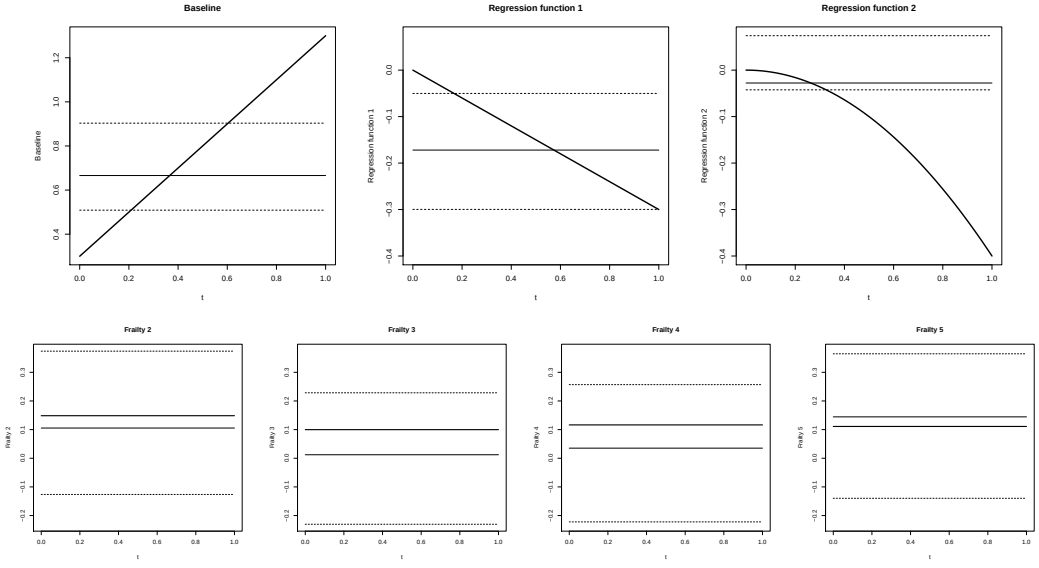


FIGURE 25. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 1$ and number of iterations $I = 1000$

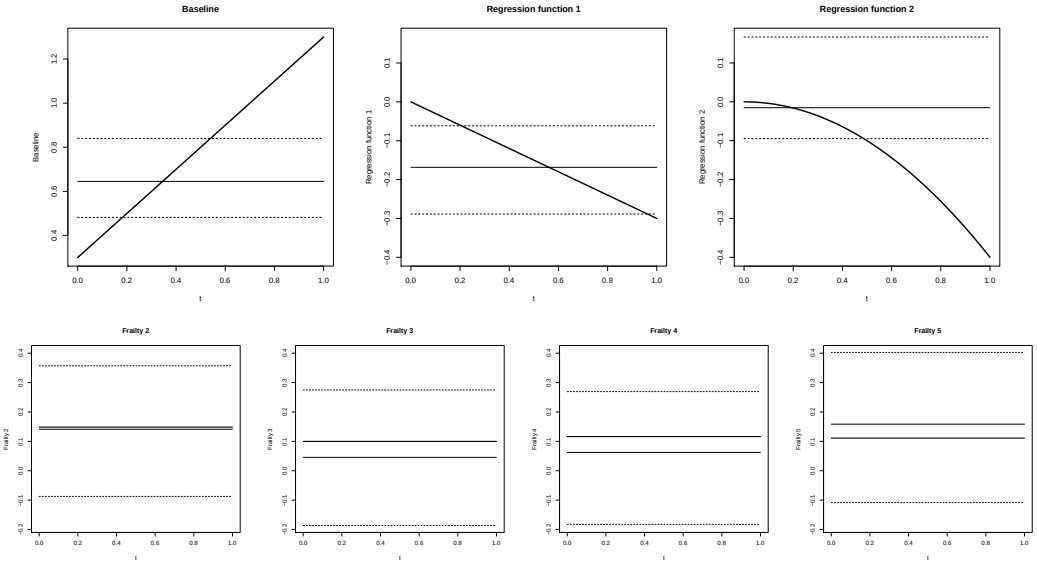


FIGURE 26. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 1$ and number of iterations $I = 5000$

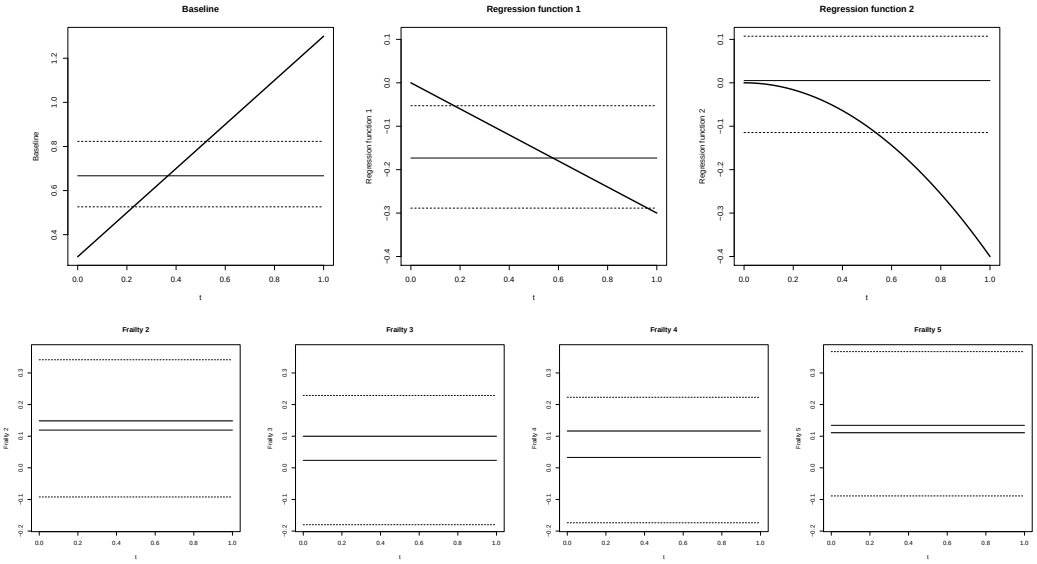


FIGURE 27. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 5$ and number of iterations $I = 100$

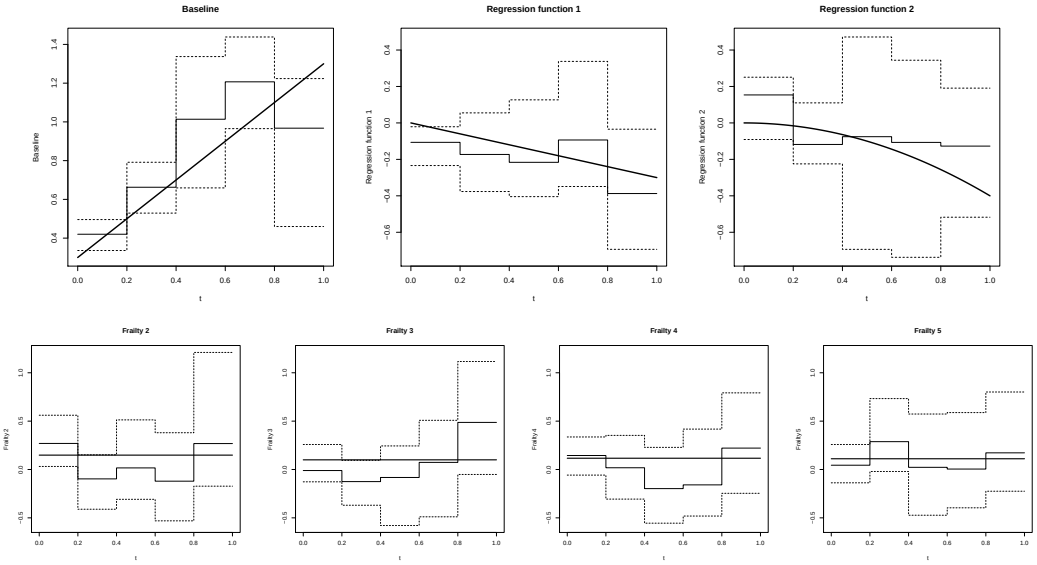


FIGURE 28. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 5$ and number of iterations $I = 500$

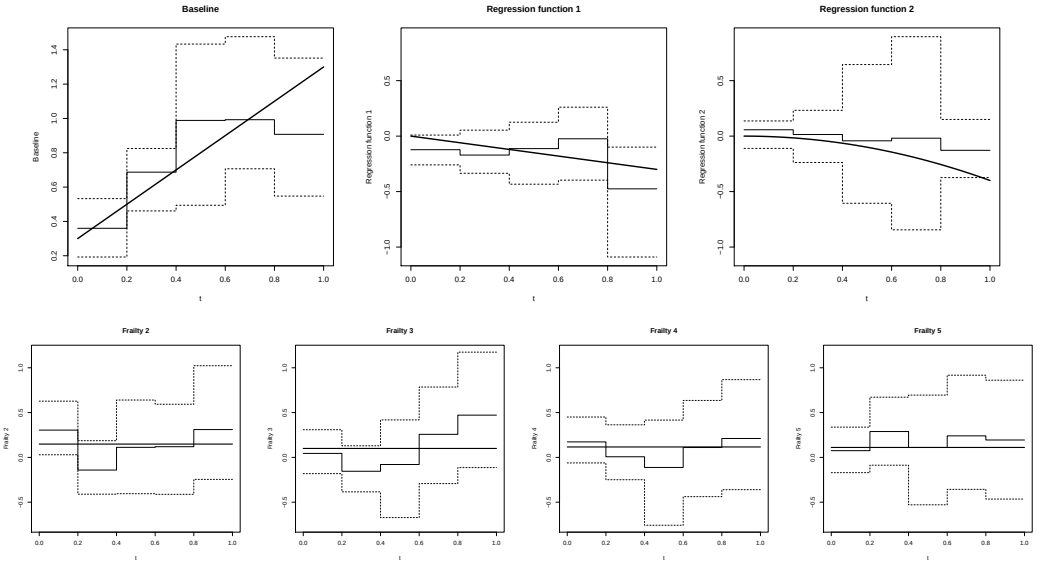


FIGURE 29. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 5$ and number of iterations $I = 1000$

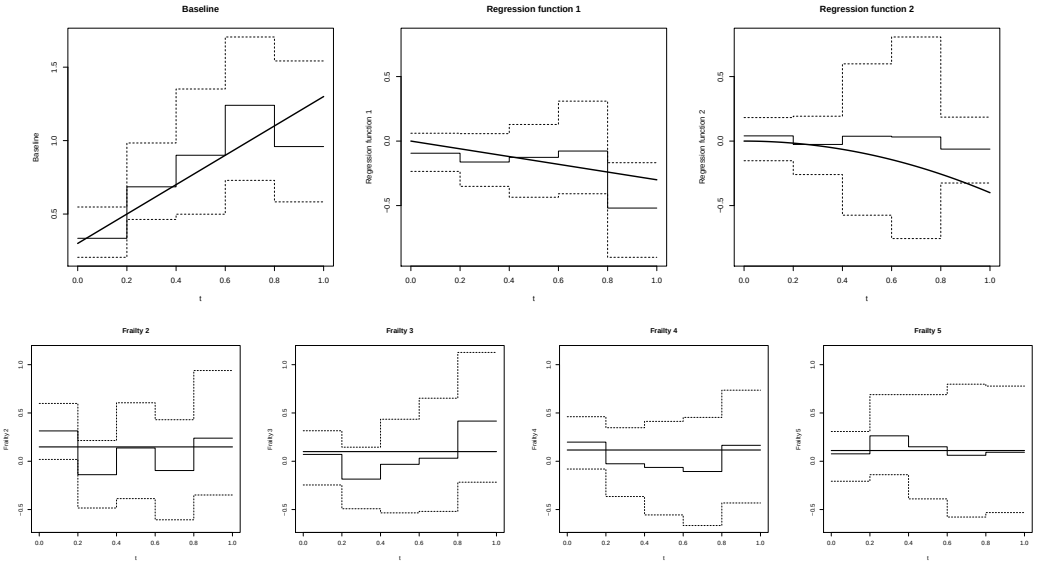


FIGURE 30. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 5$ and number of iterations $I = 5000$

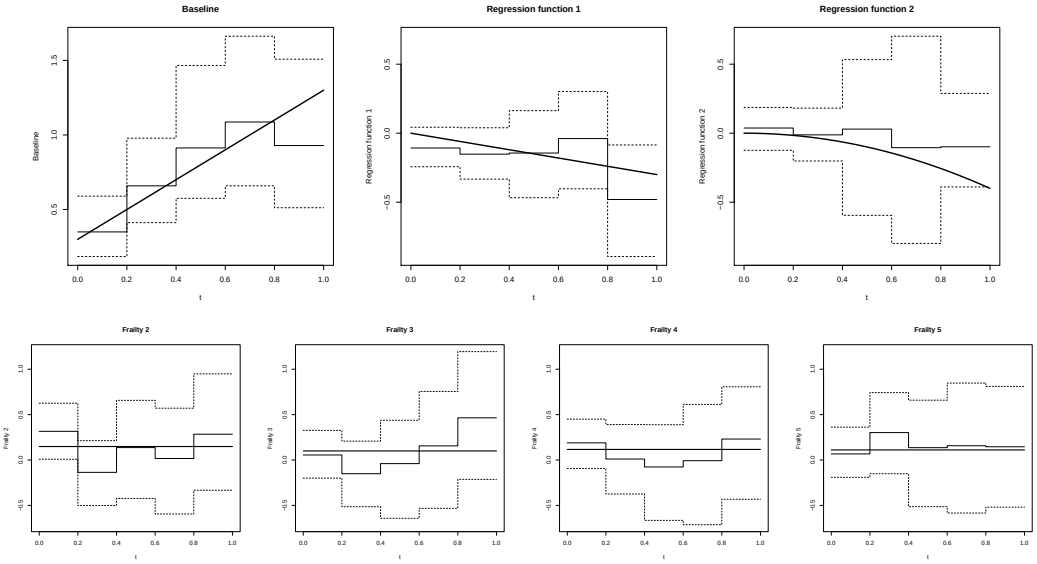


FIGURE 31. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 10$ and number of iterations $I = 100$

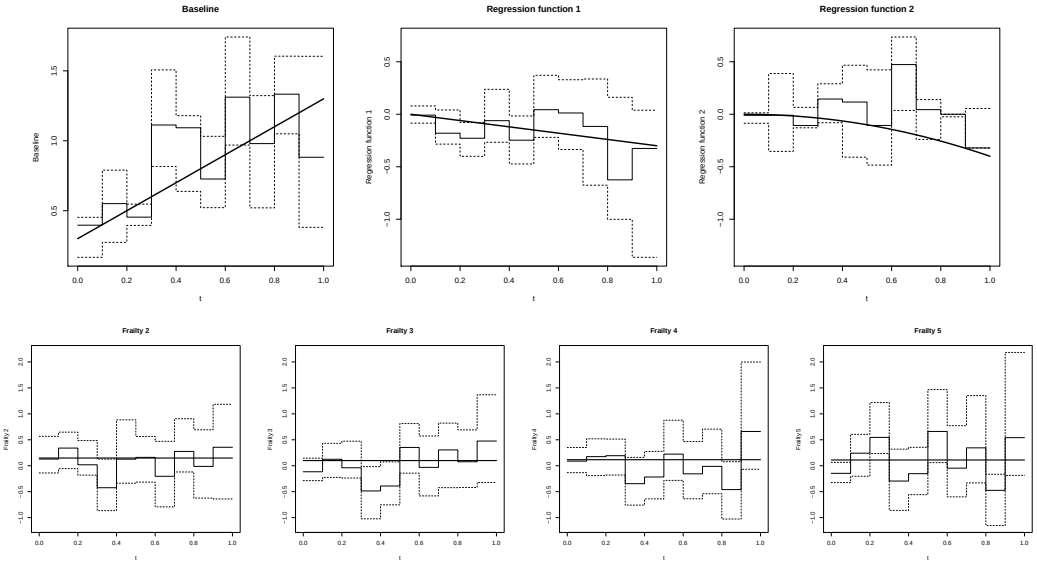


FIGURE 32. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 10$ and number of iterations $I = 500$

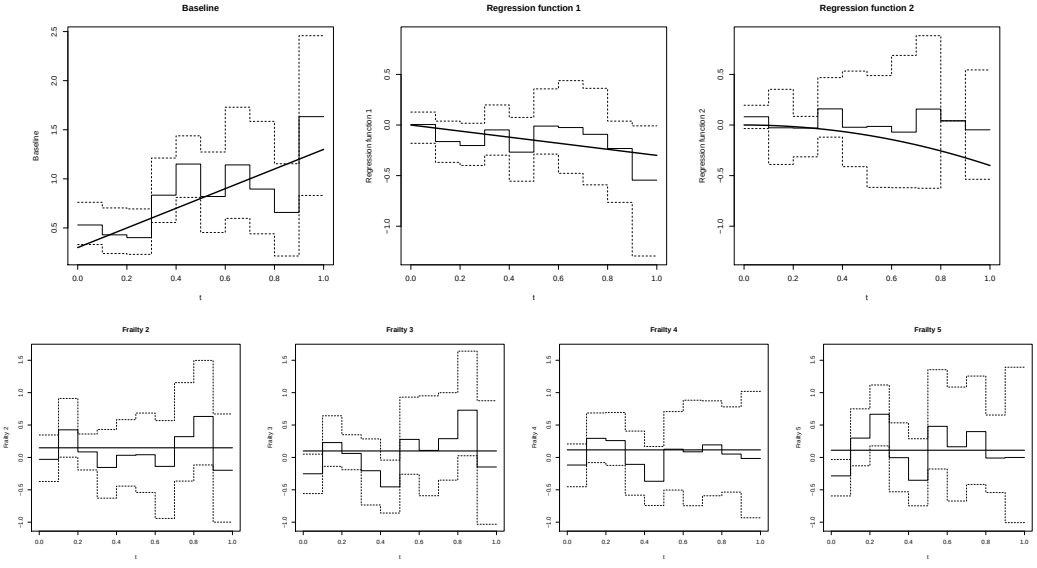


FIGURE 33. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 10$ and number of iterations $I = 1000$

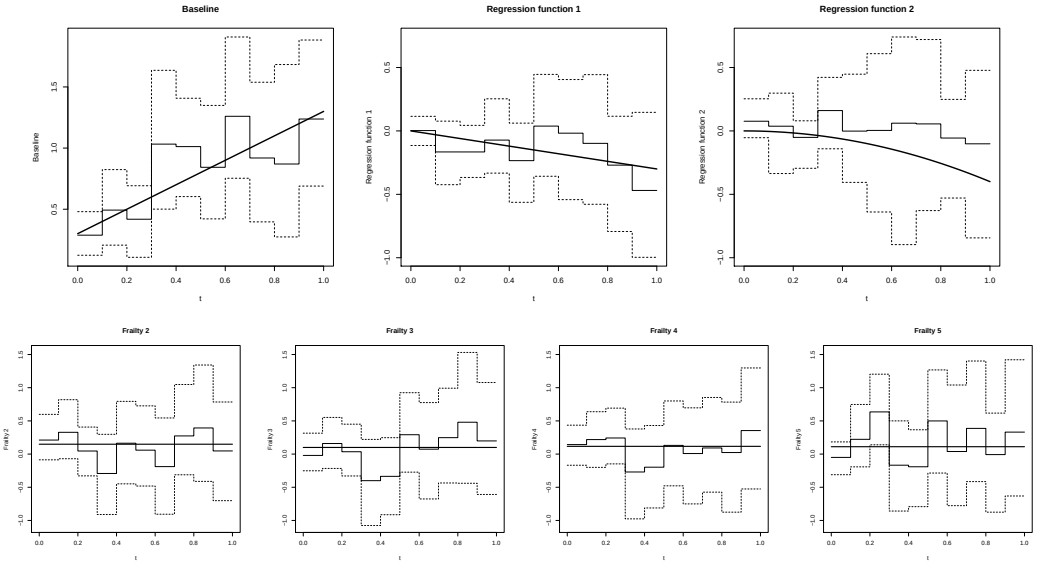


FIGURE 34. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 10$ and number of iterations $I = 5000$

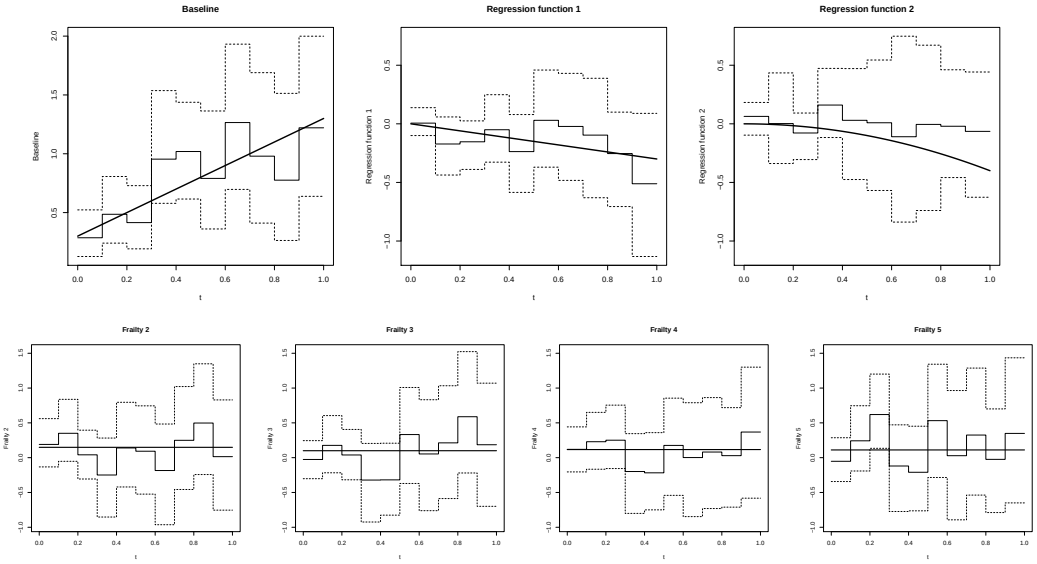


FIGURE 35. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 50$ and number of iterations $I = 100$

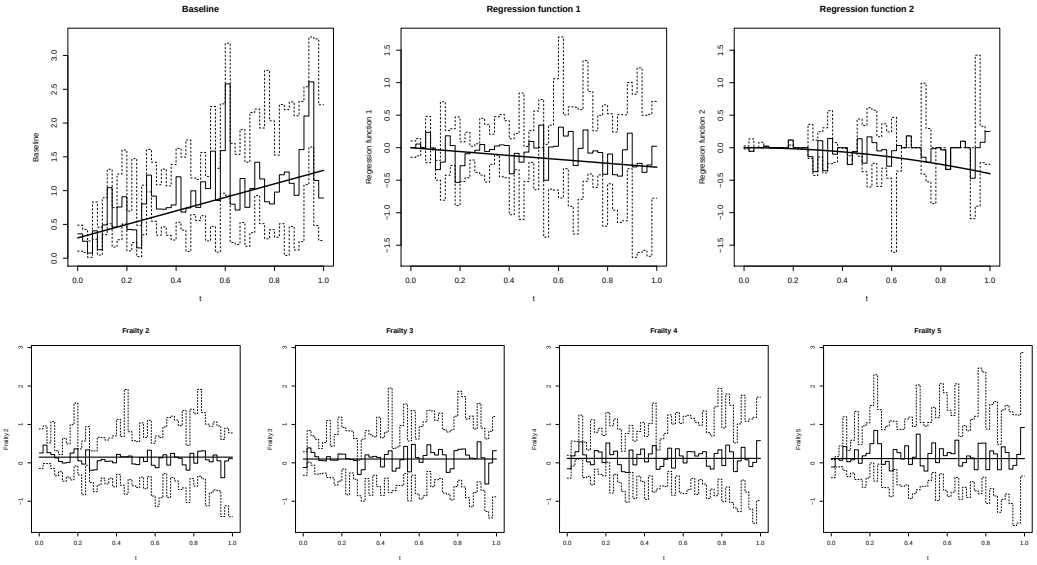


FIGURE 36. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 50$ and number of iterations $I = 500$

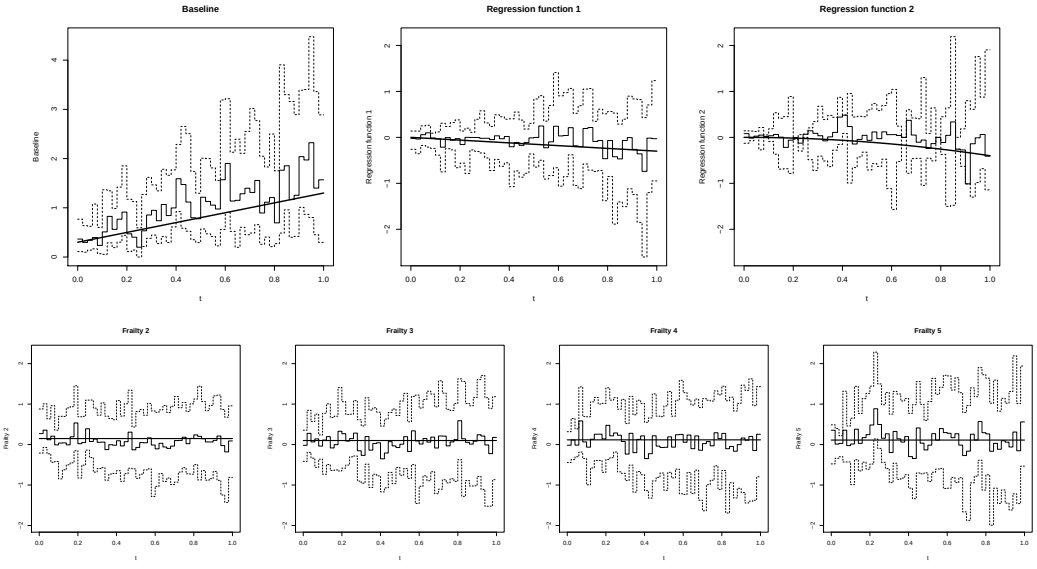


FIGURE 37. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 50$ and number of iterations $I = 1000$

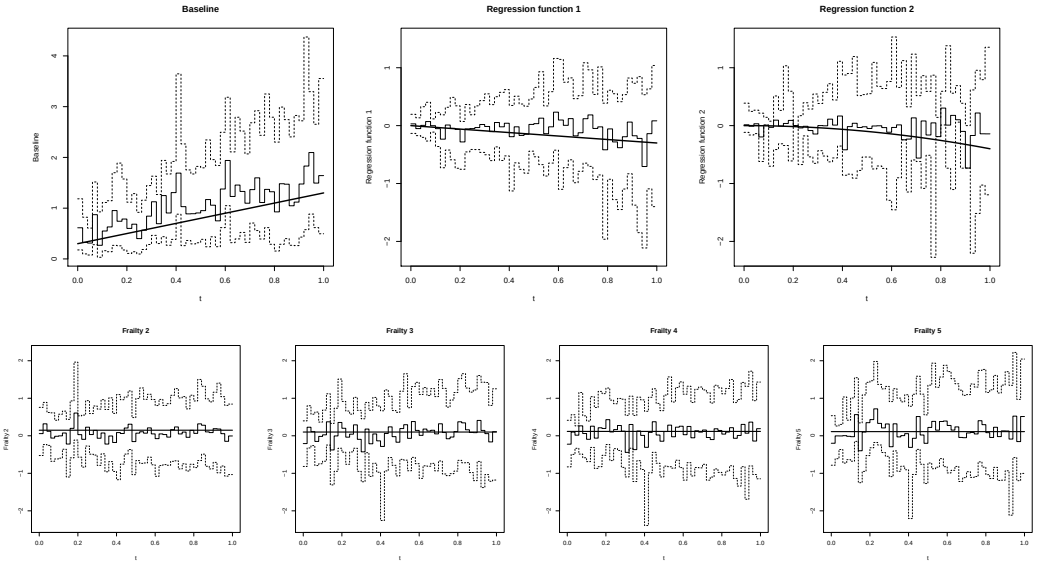


FIGURE 38. Estimated parameters for number of observations $N = 1000$, number of breakpoints $m = 50$ and number of iterations $I = 5000$

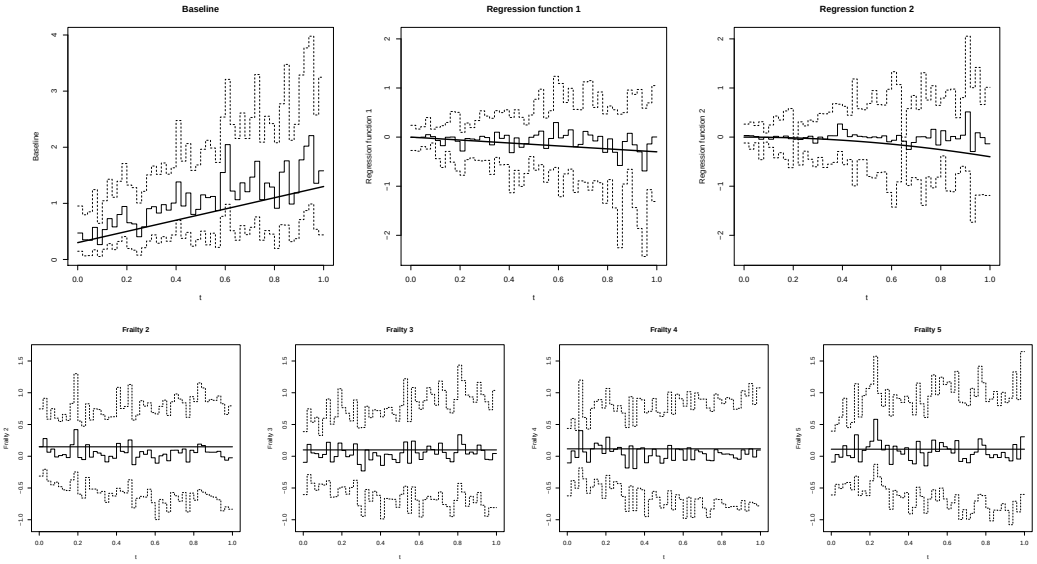


FIGURE 39. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 1$ and number of iterations $I = 100$

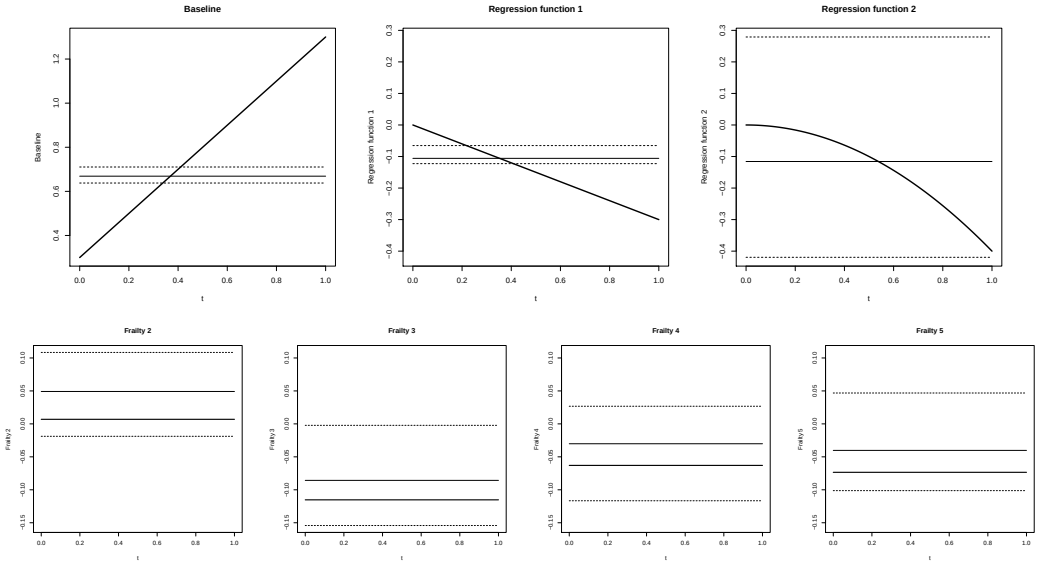


FIGURE 40. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 1$ and number of iterations $I = 500$

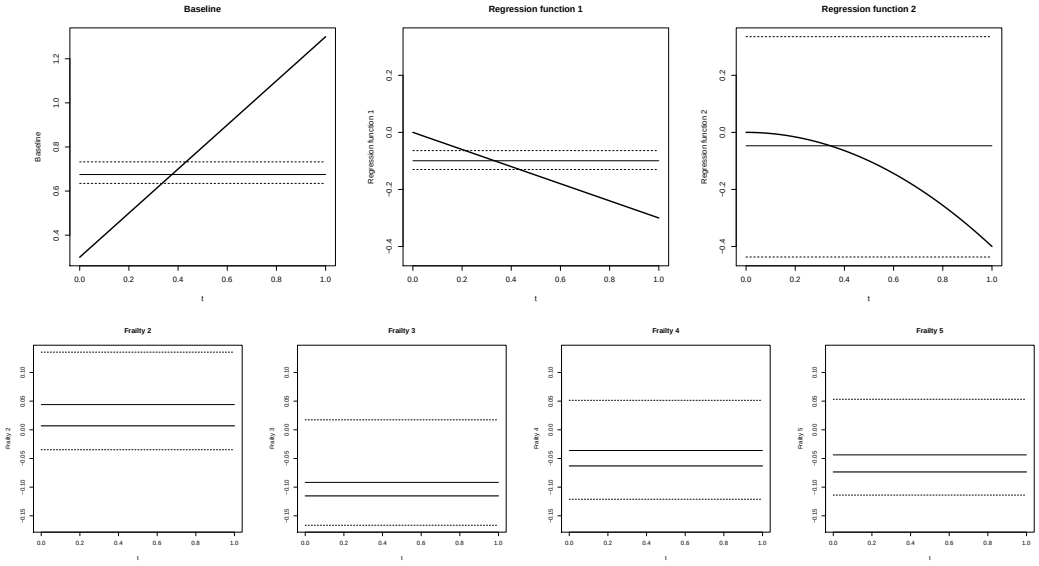


FIGURE 41. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 1$ and number of iterations $I = 1000$

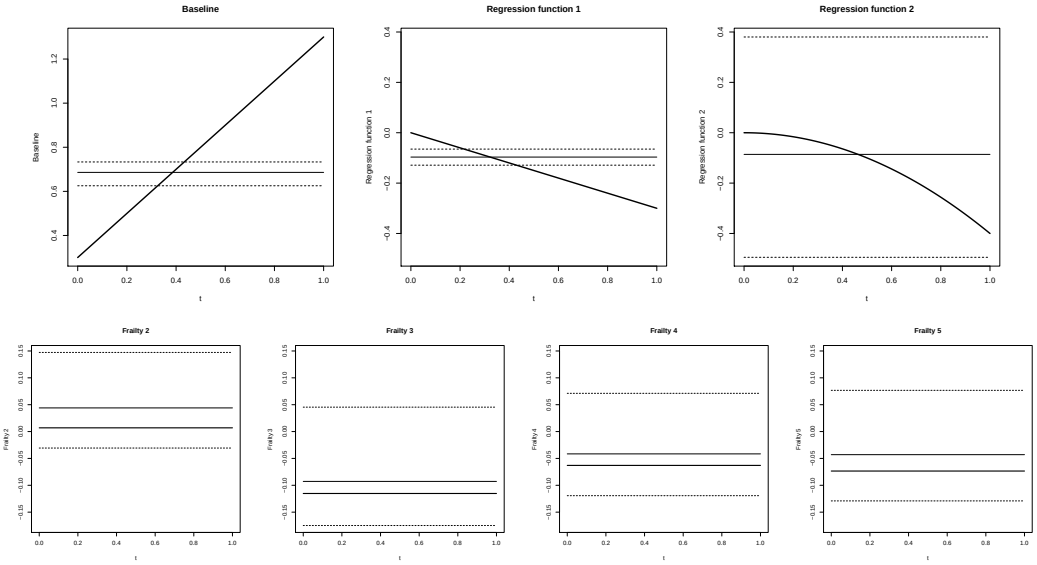


FIGURE 42. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 1$ and number of iterations $I = 5000$

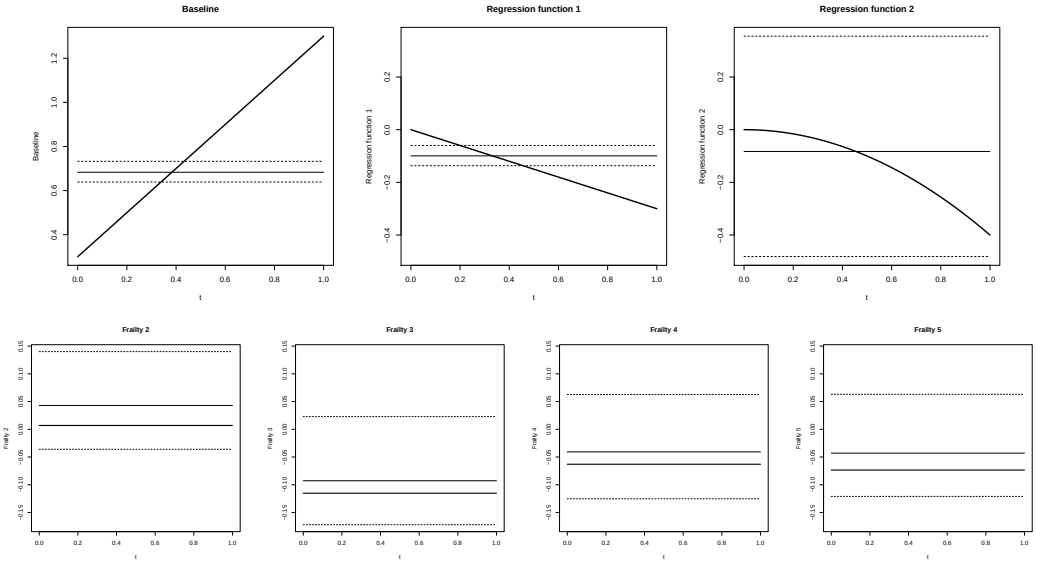


FIGURE 43. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 5$ and number of iterations $I = 100$

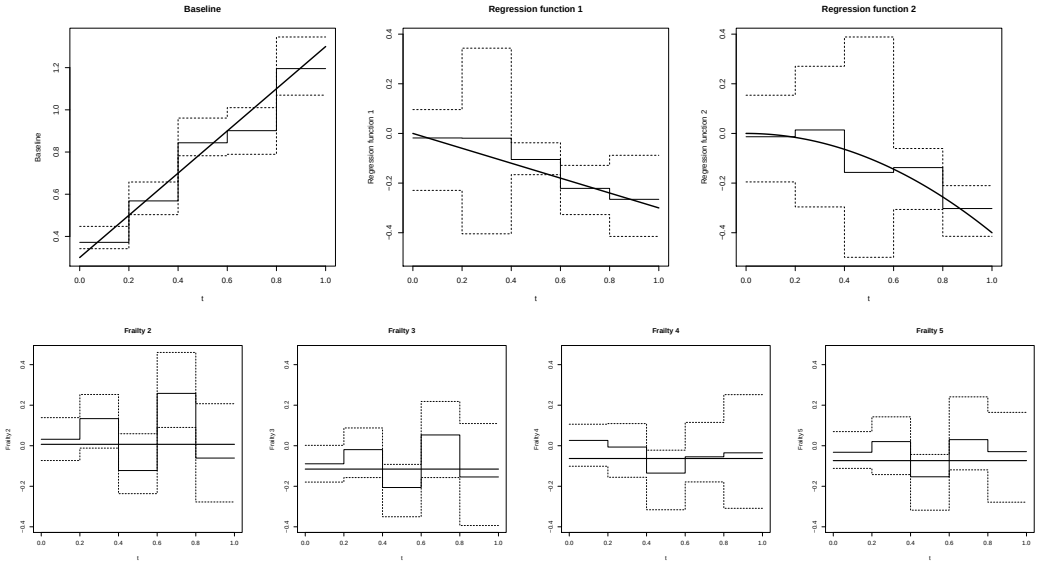


FIGURE 44. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 5$ and number of iterations $I = 500$

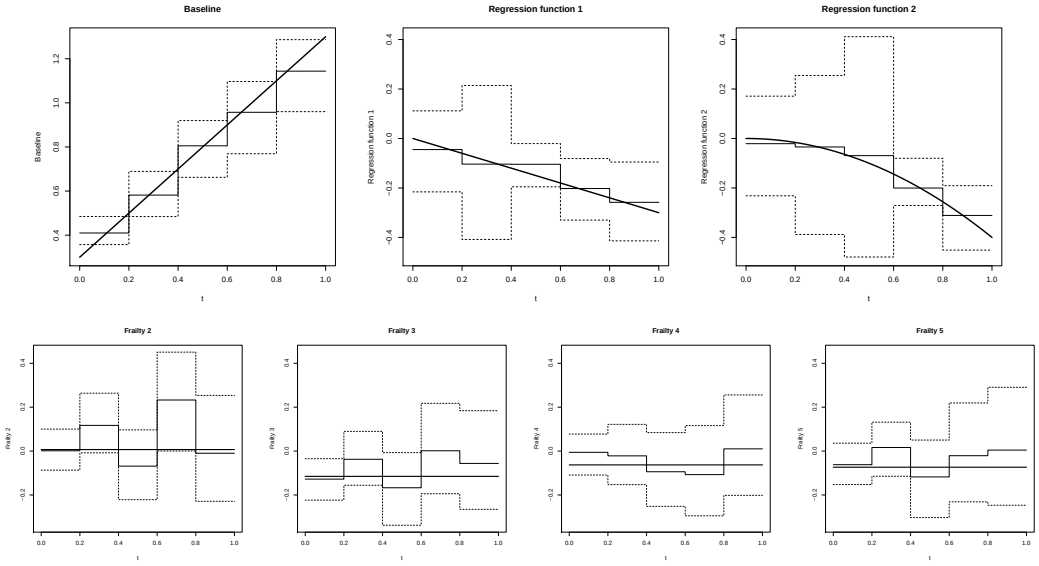


FIGURE 45. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 5$ and number of iterations $I = 1000$

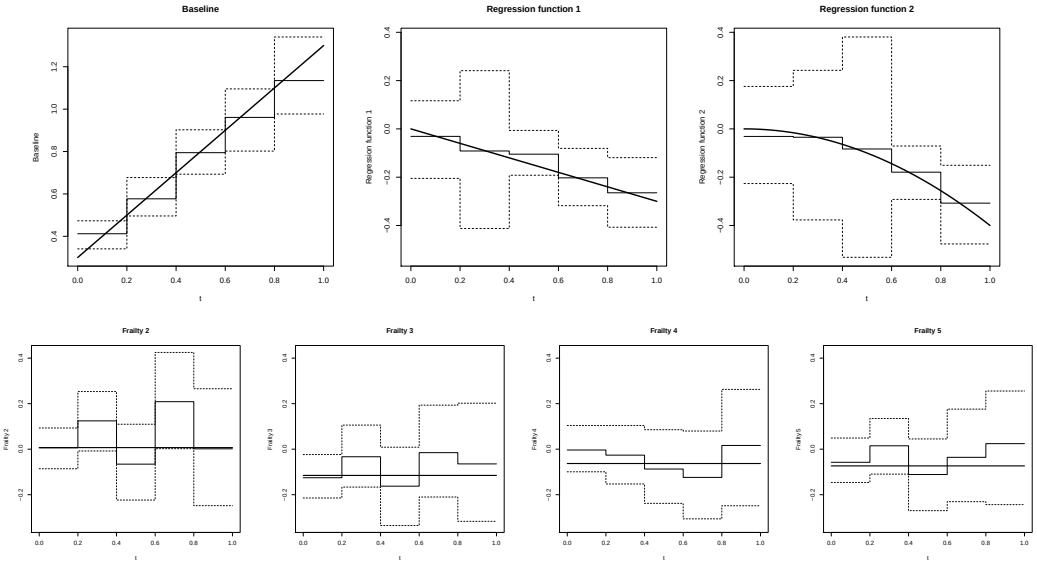


FIGURE 46. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 5$ and number of iterations $I = 5000$

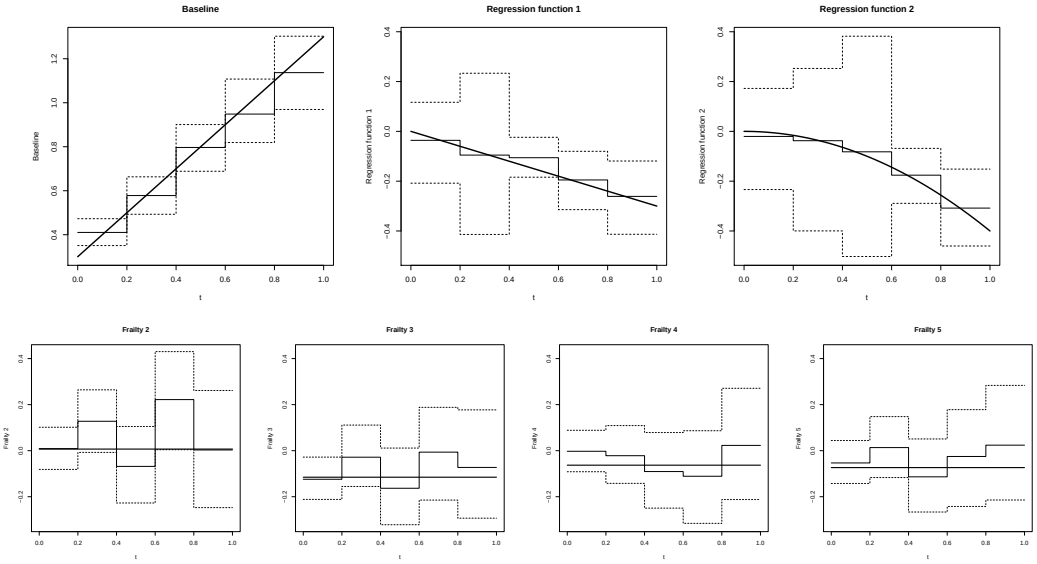


FIGURE 47. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 10$ and number of iterations $I = 100$

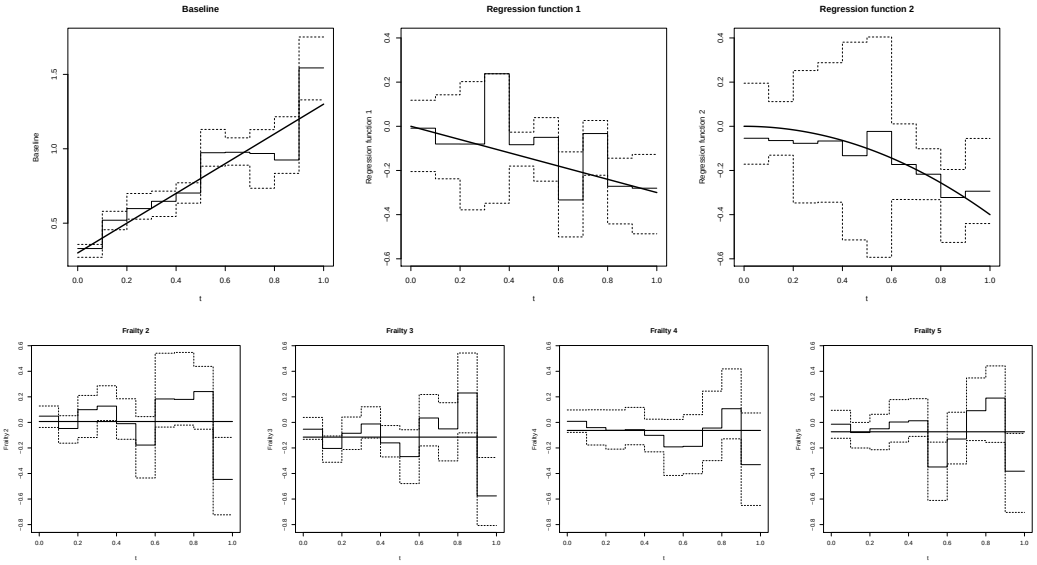


FIGURE 48. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 10$ and number of iterations $I = 500$

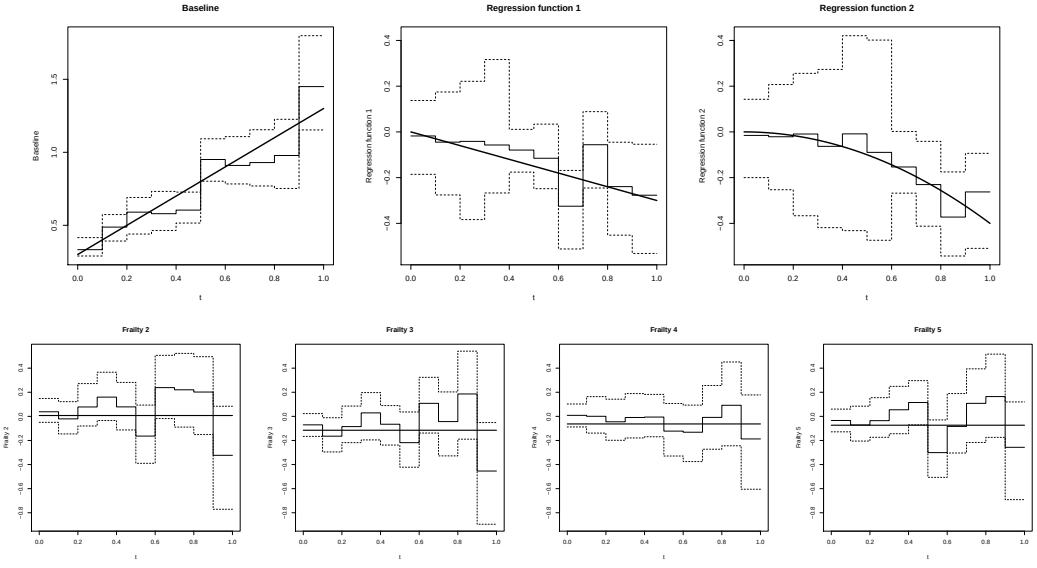


FIGURE 49. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 10$ and number of iterations $I = 1000$

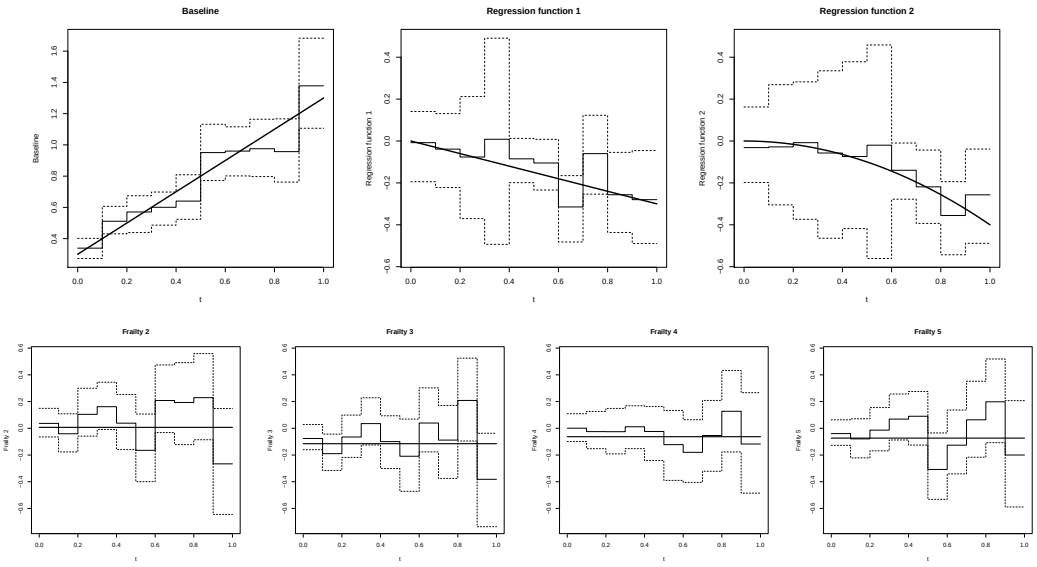


FIGURE 50. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 10$ and number of iterations $I = 5000$

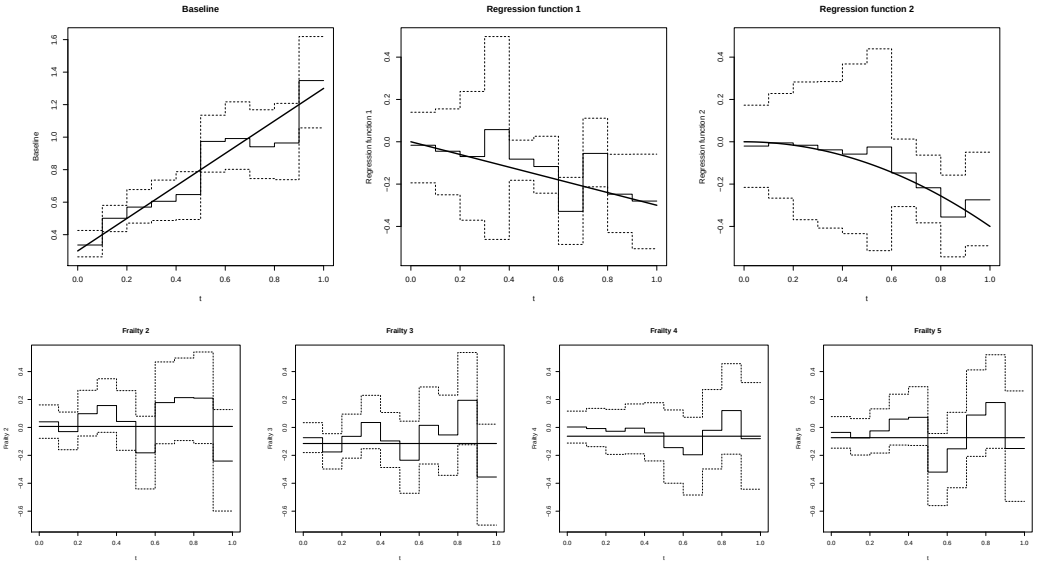


FIGURE 51. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 50$ and number of iterations $I = 100$

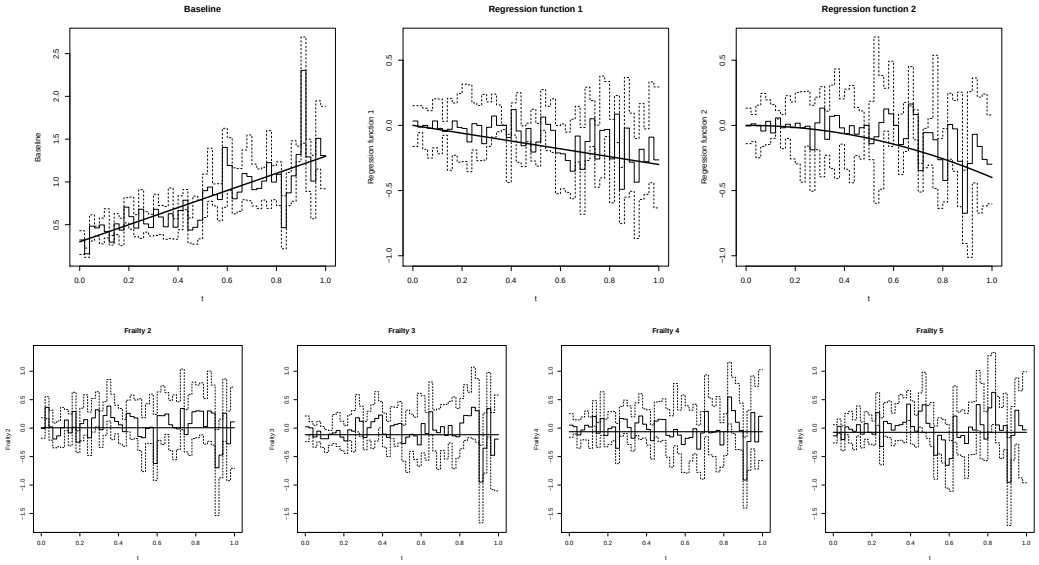


FIGURE 52. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 50$ and number of iterations $I = 500$

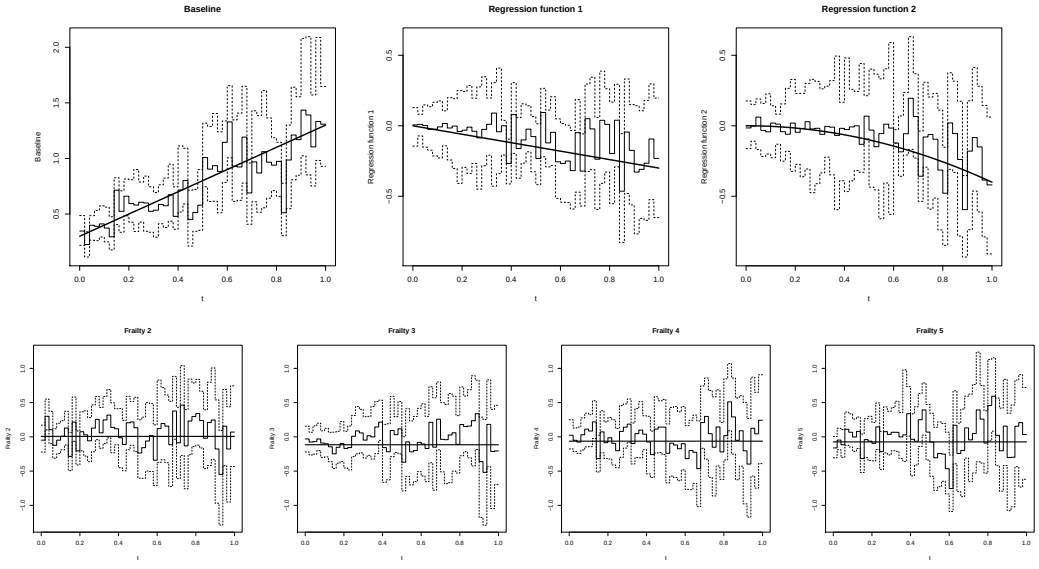


FIGURE 53. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 50$ and number of iterations $I = 1000$

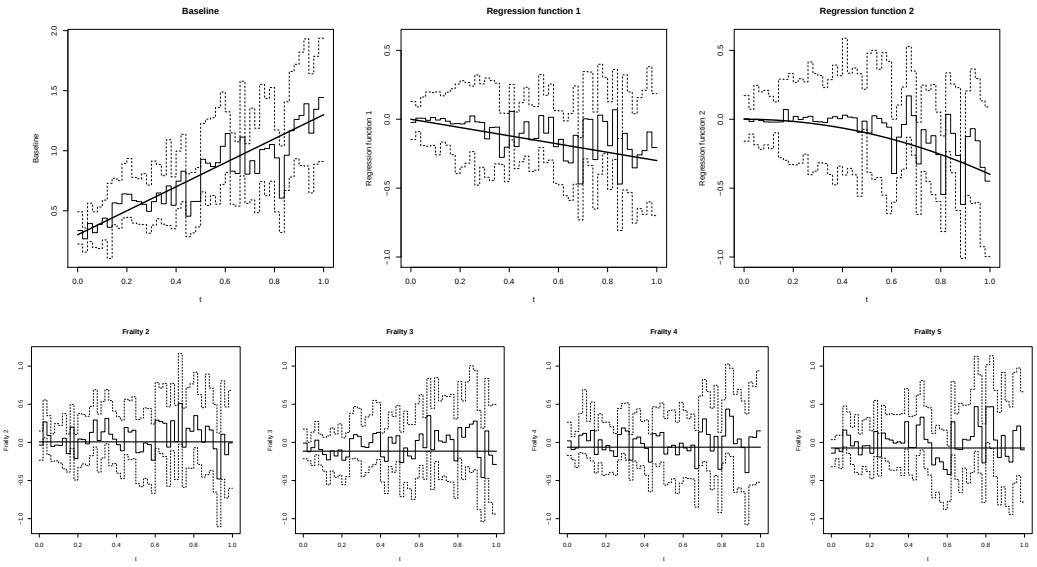
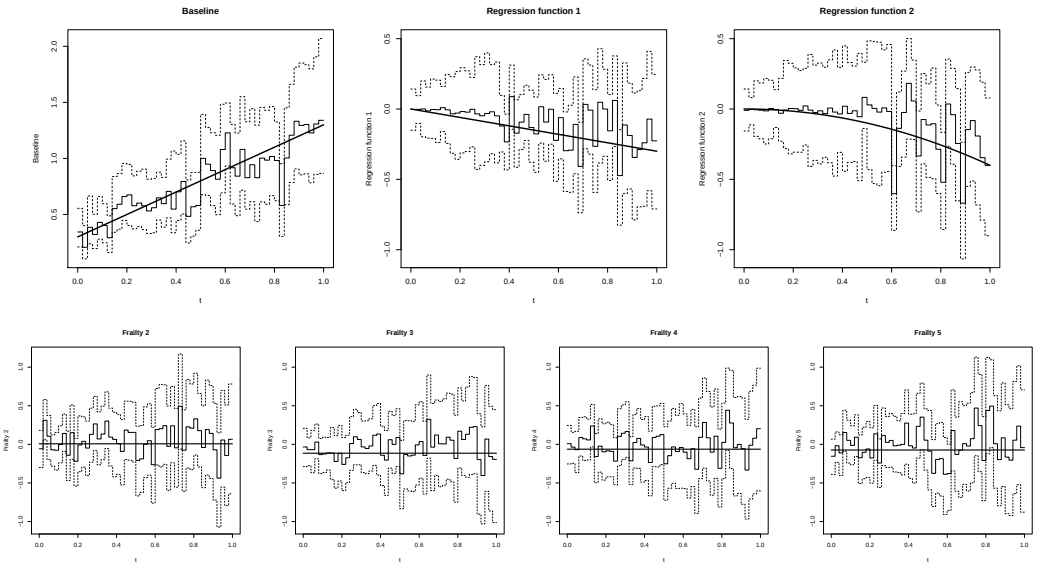


FIGURE 54. Estimated parameters for number of observations $N = 10000$, number of breakpoints $m = 50$ and number of iterations $I = 5000$



Bibliography

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In Klonecki, W., Kozek, A., and Rosiski, J., editors, *Mathematical Statistics and Probability Theory*, volume 2 of *Lecture Notes in Statistics*, pages 1–25. Springer New York.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer US.

Banerjee, S. and Dey, D. (2005). Semiparametric proportional odds models for spatially correlated survival data. *Lifetime Data Analysis*, 11(2):175–191.

Banerjee, S., Wall, M. M., and Carlin, B. P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4(1):pp. 123–142.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 10(1):pp. 3–41.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregression. *Biometrika*, 82(4):pp. 733–746.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*,

43(1):pp. 1–20.

Cai, J. and Zeng, D. (2011). Additive mixed effect model for clustered failure time data. *Biometrics*, 67(4):1340–1351.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):pp. 167–174.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley series in probability and statistics. Wiley, New Jersey.

Darmofal, D. (2009). Bayesian spatial survival models for political event processes. *American Journal of Political Science*, 53(1):241–257.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):pp. 398–409.

Gelfand, A. E., Smith, A. F. M., and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, 87(418):pp. 523–532.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):pp. 97–109.

Hussein, A., Nkurunziza, S., and Tomanelli, K. (2013). Efficient estimation for Aalen's additive hazards model. *Australian and New Zealand journal of statistics*. Under review.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Statistics for Biology and Health. Springer, 2nd edition.

Lin, P.-S. (2012). Analysis of spatial frailty models by a weighted estimating equation. *Journal of Statistical Planning and Inference*, 142(6):pp. 1436–1444.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Statistics for Biology and Health. Springer New York.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of Applied Probability*, 44(2):pp. 458–475.

SEER (2008). Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 17 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2007 Sub (1973-2005 varying) - Linked To County Attributes - Total U.S., 1969-2005 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2008, based on the November 2007 submission.

Silva, G. L. and Amaral-Turkman, M. A. (2004). Bayesian analysis of an additive survival model with frailty. *Communications in Statistics – Theory and Methods*, 33(10):pp. 2517–2533.

Tierney, L. (1994a). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):pp. 1701–1728.

Tierney, L. (1994b). Rejoinder: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):pp. 1758–1762.

Zhang, J. and Lawson, A. B. (2011). Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. *Journal of Applied Statistics*,

38(3):pp. 591–603.

Vita Auctoris

NAME:	Alexander Chernoukhov
PLACE OF BIRTH:	Gaijunai, Lithuania, USSR
YEAR OF BIRTH:	1989
EDUCATION:	Moscow Institute of Physics and Technology, Moscow, Russia 2006–2012 M.Sc.
	University of Windsor, Windsor, Ontario 2012–2013 M.Sc.