Bayesian nonparametric regression for survival and event history data

Andrea Hennerfeind

München, 17.03.2006

Bayesian nonparametric regression for survival and event history data

Andrea Hennerfeind

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Andrea Hennerfeind
aus Freising

München, den 17.03.2006

Erstgutachter: Prof. Dr. Ludwig Fahrmeir

Zweitgutachter: Prof. Dr. Stefan Lang

Drittgutachter: Prof. Dr. Claudia Czado

Rigorosum: 22.6.2006

Vorwort

Diese Arbeit entstand während meiner Tätigkeit als Mitarbeiter im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" am Institut für Statistik an der LMU München und wurde somit durch Mittel der Deutschen Forschungsgemeinschaft gefördert. In diesen Jahren haben mich viele Leute begleitet, die wesentlich zum Gelingen meiner Dissertation und zu einem angenehmen Arbeitsklima beigetragen haben.
Zu allererst möchte ich meinem Doktorvater Ludwig Fahrmeir aufrichtig für seine hervorragende Betreuung danken. Ohne sein Vertrauen, seine Unterstützung und seine sympathische, unkomplizierte und wohlwollende Art, wäre diese Arbeit sicher nie entstanden. Ein ganz besonderer Dank gilt auch meinem Zweitgutachter Stefan Lang, der mir in den vergangenen Jahren in den verschiedensten Hinrichtungen (mein aktueller Lieblings–Versprecher) eine unendlich große Hilfe war.
Des weiteren gebührt mein Dank auch allen übrigen Lehrstuhl– und Institutsmitarbeitern. Insbesondere danke ich Christiane Belitz und Leyre Osuna, die das zweifelhafte Vergnügen hatten, mich als Bürokollegin zu haben. Beide haben entscheidend dazu beigetragen, dass ich mich am Institut so wohl gefühlt habe und standen mir bei allen kleineren und größeren Problemen stets mit Rat und Tat zur Seite. Entgegen anders lautender Gerüchte musste ich im 'Mädchenbüro' zum Glück nie über die neueste Schuhmode diskutieren. Bei Thomas Kneib möchte ich mich für die vielen hilfreichen Diskussionen und Inspirationen bedanken. 'Du tatest etwas großes ohne Geld zu kalkulieren.' Meiner 'Grundausstattung' Alexander Jerak habe ich meine Existenz als 'Ergänzungsausstattung' zu verdanken. Danke auch für die vielen aufbauenden Worte, wenn wir alle mal wieder an unseren Fähigkeiten gezweifelt haben. Auf keinen Fall unerwähnt lassen möchte ich Renata Gebhardt, die durch ihre lebendige Art stets für Aufmunterung gesorgt hat, und Susanne Heim, die für unsere Fortbildungsreise nach Florenz die schönste aller Unterkünfte ausfindig gemacht hat und stets eine kompetente Ansprechpartnerin in Sachen 'fMRI' war. Bei Petra Kragler möchte ich mich herzlich für Ihre Formulierungshilfen und gelegentliche Aufmunterungen bedanken.
Weiterhin möchte ich mich bei meinen Koautoren Leonhard Held und Erik Sauleau für die angenehme und fruchtbare Zusammenarbeit bedanken. Claudia Czado danke ich dafür, dass sie sich meiner Arbeit freundlicherweise als externe Gutachterin angenommen hat.
Nicht zuletzt möchte ich von ganzem Herzen meinen Eltern und meiner Schwester danken,

die immer vollstes Vertrauen in mich gesetzt haben und mir den nötigen familiären Rückhalt gegeben haben. Danke, dass ich mich immer voll auf Euch verlassen kann!
Mein größter Dank gilt meinem Freund Andreas Brezger, der immer für mich da war. Danke für die endlosen fachlichen Diskussionen und Erläuterungen, für wertvolle Ratschläge, für all die Ermutigungen, für den Trost und die Ablenkung und natürlich auch für die vielen schönen gemeinsamen Erlebnisse in den letzten Jahren!

München, Juli 2006 *Andrea Hennerfeind*

Zusammenfassung

Die Überlebenszeitanalyse, oder allgemeiner die Verweildaueranalyse findet in der Praxis zahlreiche Anwendungen vom klassischen Fall der klinischen Studie bis hin zur Modellierung von Kreditrisiken. Oftmals sind die Standard–Modelle jedoch nicht flexibel genug, um der Modellierung komplexer Kovariableninformationen gerecht zu werden. Neben parametrisch und nichtparametrisch modellierten Kovariableneffekten, sowie räumlichen Effekten und zufälligen Effekten zur Berücksichtigung von unbeobachteter Heterogenität, ist bei der Verweildaueranalyse auch häufig eine flexible, nichtparametrische Modellierung von zeitlich variierenden Effekten gefragt.

Diese Arbeit beschäftigt sich mit der Entwicklung von Bayesianischen Verweildauermodellen, die Erweiterungen des klassischen Cox–Modells darstellen. Indem die Hazardrate durch einen strukturierten additiven Prädiktor modelliert wird, entsteht ein flexibles Modell zur Analyse von stetigen Verweildauern unter adäquater Berücksichtigung verschiedenster Arten von Kovariablen. Zeitlich variierende Effekte werden dabei durch P–Splines modelliert. Die Schätzung erfolgt mit Hilfe von Markov Chain Monte Carlo Verfahren.

Weitere Kapitel beschäftigen sich mit der sogenannten relativen Überlebenszeitanalyse und mit Mehrzustandsmodellen. Bei ersterem geht es darum, ein zusätzliches Risiko einer bestimmten Subpopulation zu modellieren, das über das allgemeine Risiko in der gesamten Population hinaus besteht. Mehrzustandsmodelle stellen eine Verallgemeinerung der Verweildauermodelle dar. Anstelle eines bestimmten Übergangs können hier mehrere verschiedene Übergänge simultan analysiert werden.

Die in dieser Arbeit vorgestellten Methoden werden jeweils auf komplexe, reale Problemstellungen angewandt und erweisen sich als wirkungsvolle und flexible Instrumente.

Abstract

Survival analysis, or more generally duration time analysis has a large number of practical applications ranging from the classical field of clinical studies to credit risk analysis. In most cases however, standard survival models do not offer enough flexibility to give appropriate consideration to modelling complex covariate effects. In addition to parametric and

nonparametric effects as well as spatial effects and random effects to capture unobserved heterogeneity, duration time analysis often demands a flexible nonparametric estimation of time–varying effects.

This thesis is concerned with developing Bayesian duration time models representing extensions to the classical Cox model. Modelling the hazard rate through a structured additive predictor leads to a flexible model for the analysis of continuous duration times having regard to the influence of several different types of covariates. Time–varying effects are modelled by P–splines. Inference is accomplished using Markov Chain Monte Carlo simulation techniques.

Further topics are the so–called relative survival analysis and multi–state models. The former topic is concerned with modelling the excess risk of a certain subpopulation relative to the base risk that is present in the whole population. Multi–state models are a generalization of duration time models. Instead of analyzing one particular transition only they allow for the simultaneous analysis of diverse transitions.

The methods presented within this thesis are applied to several complex, real problems and prove to be effective and flexible tools.

Contents

Chapter 1

Introduction

The analysis of survival times is a specific type of regression analysis that has gained considerable attention particularly in the classical field of medical applications, wherefrom the conventional denotation 'survival analysis' arises. The primary interest in medical trials usually is the analysis of the influence of special drugs or therapies on the survival times of patients that are diagnosed with a certain disease. Generally, survival analysis is concerned with analyzing the influence of covariates on the duration time up to any predefined event of interest. As will be illustrated in this work there is a number of further fields of applications including for example the field of credit scoring, where the life of a loan up to a default is analyzed.

In survival analysis a distinction is drawn between discrete time survival analysis, where survival times are only given in certain units of time and continuous time survival analysis. The former can be ascribed to binary response models and may therefore be based on methodology for binary logit, probit or grouped Cox models. For this reason we will only deal with the more challenging case, where survival times are measured on a continuous time scale. Grouping the data for a discrete time survival analysis is possible, but leads to a loss of information and is therefore not recommended. Another idea might be to analyze survival times with generalized linear models for nonnegative continuous responses (like lognormal or gamma regression). However, these methods do not account for censoring and truncation, two specifics of survival data that are due to the fact that survival times can often not be observed completely but only within a specific observation period. For this reasons continuous time survival analysis actually is a separate area of statistical

analysis that is treated extensively in the literature, see e.g. Lawless (1982), Kalbfleisch and Prentice (1980), Blossfeld, Hamerle and Mayer (1989) and Klein and Moeschberger (2003), or Andersen, Borgan, Gill and Keiding (1993) for a counting process representation.

1.1 Basic concepts

Consider survival time to be a nonnegative continuous random variable T, with density function $f(t)$. Then the corresponding distribution function $F(t)$, which is the probability of not surviving until time t is given by

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

In survival analysis however, it is more common to examine the so called survivor function, which is the probability of the complementary event, i.e. the probability of surviving until time t. It given by

$$S(t) = 1 - F(t) = P(T > t).$$

Another quantity that plays a decisive role in survival analysis is the hazard function $\lambda(t)$, which is defined by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

and determines the instantaneous rate of death or failure at time t subject to the condition of survival up to time t.

The following equations, where $\Lambda(t) = \int_0^t \lambda(t)$ denotes the cumulative hazard function, show how the quantities introduced above are related to each other:

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ S(t) &= \exp(-\Lambda(t)) \\ f(t) &= \lambda(t)S(t) = \lambda(t)\exp(-\Lambda(t)) \end{aligned}$$

The distribution of T is completely determined by one of these quantities. As an illustration consider the Weibull distribution, where the hazard rate has the following structure

$$\lambda(t) = \lambda\alpha(\lambda t)^{\alpha-1},$$

with scale parameter $\lambda > 0$ and shape parameter $\alpha > 0$ (see Figure 1.3 below for a graphical representation). The survivor function is thus given by

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(u)du\right) = \exp\left(-\int_0^t \lambda\alpha(\lambda u)^{\alpha-1}du\right) \\ &= \exp\left(-\lambda\alpha\lambda^{\alpha-1}\int_0^t u^{\alpha-1}du\right) = \exp\left(-\lambda^\alpha\alpha[\alpha^{-1}u^\alpha]_0^t\right) = \exp\left(-\lambda^\alpha\alpha(\alpha^{-1}t^\alpha)\right) \\ &= \exp\left(-(\lambda t)^\alpha\right) \end{aligned}$$

leading to

$$f(t) = \lambda(t)S(t) = \lambda\alpha(\lambda t)^{\alpha-1}\exp\left(-(\lambda t)^\alpha\right),$$

which is indeed the density function of a Weibull distribution with parameters λ and α. The exponential model, where the hazard rate is constant over time, i.e. $\lambda(t) = \lambda > 0$ is included as the special case of $\alpha = 1$.

1.2 The Cox model

Consider survival data in conventional form, i.e. assume that each individual i in the study has a lifetime T_i and a censoring time C_i that are independent random variables (random censoring). The observed lifetime is then $t_i = \min(T_i, C_i)$, and δ_i denotes the censoring indicator given by

$$\delta_i = \begin{cases} 1 & T_i \le C_i \\ 0 & \text{else} \end{cases} \tag{1.1}$$

In addition to the lifetime one usually considers some individual–specific covariates that are assumed to have an influence on the lifetime. The data is then given by

$$(t_i, \delta_i; \boldsymbol{v}_i), \quad i = 1, \ldots, n, \tag{1.2}$$

where $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{ir})$ is the vector of the r covariates observed with individual i. Note that covariates may also be time–dependent, but for the moment we restrict discussion to time–constant covariates for simplicity. The benchmark in the area of analyzing the influence of covariates on survival time is the proportional hazards model introduced by

Cox (1972). Here the hazard rate of individual i is modelled as the product

$$\lambda_i(t, \boldsymbol{v}_i) = \lambda_0(t) \cdot \exp(v_{i1}\gamma_1 + \ldots + v_{ir}\gamma_r) = \lambda_0(t) \cdot \exp(\boldsymbol{v}_i'\boldsymbol{\gamma}), \tag{1.3}$$

where $\lambda_0(t)$ is the baseline hazard, that remains unspecified and is independent of the covariates, but only depends on time t. In contrast, the influence of the covariates is independent of time and modelled via a linear predictor $\boldsymbol{v}_i'\boldsymbol{\gamma}$ with a vector of regression coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)$. Through the exponential link function, the covariates act multiplicatively on the hazard rate. In the case of time–constant covariates the time constance of the influence of the covariates implicates that the hazard rates of any two individuals are proportional, which explains why the Cox model is called a proportional hazards model. Let $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ denote the covariate vectors of two individuals i and j, then the ratio of the hazard rates of these individuals is given by

$$\frac{\lambda_i(t, \boldsymbol{v}_i)}{\lambda_j(t, \boldsymbol{v}_j)} = \frac{\lambda_0(t) \cdot \exp(\boldsymbol{v}_i'\boldsymbol{\gamma})}{\lambda_0(t) \cdot \exp(\boldsymbol{v}_j'\boldsymbol{\gamma})} = \exp\left((\boldsymbol{v}_i - \boldsymbol{v}_j)'\boldsymbol{\gamma}\right),$$

which yields

$$\lambda_i(t, \boldsymbol{v}_i) = c \cdot \lambda_j(t, \boldsymbol{v}_j), \quad c = \exp\left((\boldsymbol{v}_i - \boldsymbol{v}_j)'\boldsymbol{\gamma}\right).$$

This implicit assumption of the traditional Cox model is rather restrictive and does often not hold in practice. However, this assumption is crucial for inference based on the partial likelihood proposed by Cox. Supposing that the baseline hazard $\lambda_0(t)$ is arbitrary, the partial likelihood is derived by considering the observed survival times t_i (at which we assume for simplicity that $t_i \neq t_j$ for $i \neq j$) and risk sets

$$R(t_i) = \{j | t_j \geq t_i\}$$

including all individuals j whose survival time is at least t_i, i.e. all individuals that are still at risk shortly before t_i. Given that time t_i is an observed failure time and conditionally on the risk set $R(t_i)$ the probability that the failure is actually observed on individual i (instead of any other individual $j \in R(t_i)$) is given by

$$P(i \text{ fails at } t_i | \text{one failure at } t_i, R(t_i)) = \frac{\exp\left(\boldsymbol{v}_i'\boldsymbol{\gamma}\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{v}_j'\boldsymbol{\gamma}\right)},$$

which is independent of $\lambda_0(t)$. Under the assumption of independence the partial likelihood is hence given by

$$L(\gamma) = \prod_{i=1}^{n} \frac{\exp\left(\boldsymbol{v}_i'\boldsymbol{\gamma}\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{v}_j'\boldsymbol{\gamma}\right)}.$$

The regression parameters $\boldsymbol{\gamma}$ may then be estimated by maximizing the partial likelihood. Based on the estimated parameters $\hat{\boldsymbol{\gamma}}$ the cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ may be estimated in a second step via the plug–in estimator by Breslow, which is defined as follows

$$\hat{\Lambda}_0(t) = \sum_{i:t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp(\boldsymbol{v}_j'\hat{\boldsymbol{\gamma}})}. \tag{1.4}$$

Note that $\hat{\Lambda}_0(t)$ is a step function with jumps at the observed survival times t_i.

1.3 Extensions of the Cox model

To many complex applications the basic Cox model (1.3) is not adequate with respect to several aspects such as

- In applications where predictions are of interest an improved, smooth estimation of the baseline effect is needed.
- Effects of continuous covariates might be of any unknown nonlinear form.
- Some effects might be time–varying, at which the variation is of any unknown (nonlinear) form.
- Survival times might be spatially correlated.
- Unobserved heterogeneity among individuals or units might be present.
- Nonlinear interactions between covariates might exist.

In this thesis, we propose geoadditive survival models as a flexible spatial and spatio–temporal generalization of Cox–type models. Within a unified framework, we extend the common linear predictor of the Cox model to an additive predictor, including a spatial component for geographical effects and nonparametric terms for modelling and exploring

unknown functional forms of the baseline hazard rate, of nonlinear effects of continuous covariates and further time scales, such as calendar time, and of time–varying coefficients. The incorporation of such nonparametric components and their simultaneous estimation with the baseline hazard and the spatial effects motivates the term "geoadditive", originally introduced by Kammann and Wand (2003) in a mixed model approach to semiparametric Gaussian regression. In addition, uncorrelated random effects (also referred to as frailty effects) or nonlinear two–way interactions can be incorporated if appropriate.

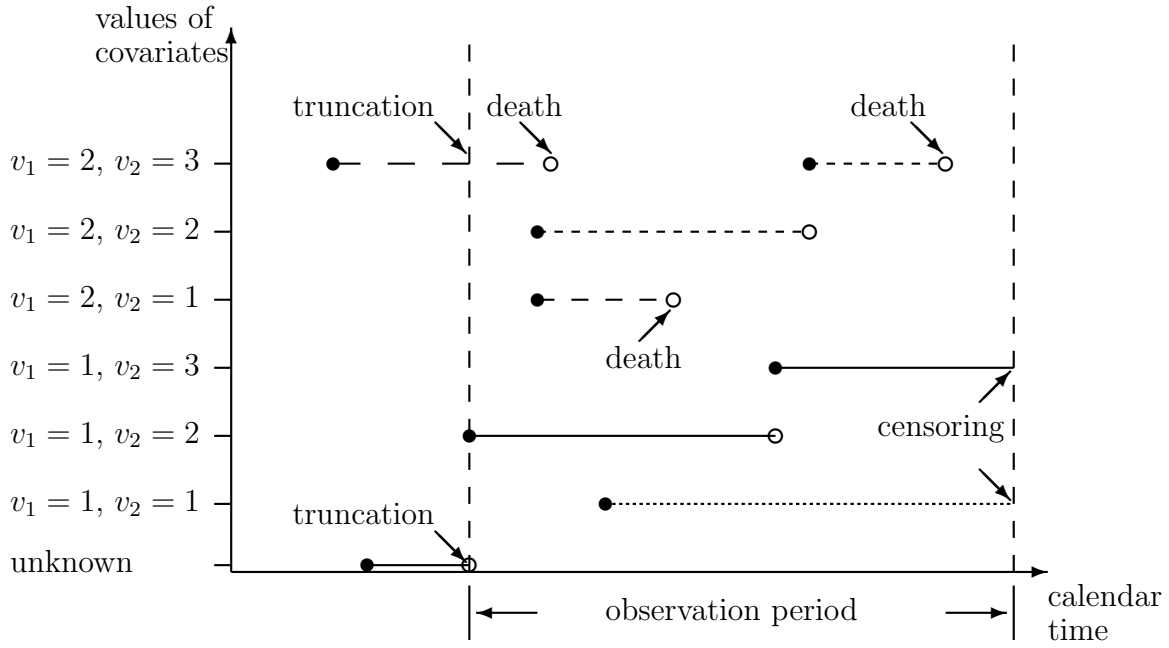
Modelling and inference is developed from a Bayesian perspective, using information from the full likelihood rather than from a partial likelihood.

1.4 Full likelihood

In survival analysis the complexity of the likelihood depends on what kind of censoring and/or truncation is present in the data. Figure 1.1 illustrates some examples of observation structures that we treat within this thesis.

1.4.1 Right censoring

Usually right censored data are considered, where the exact survival time is only observed for some individuals, whereas others are only observed until a certain point of time prior to the event of interest. This involves that it is only known that the survival time is greater than the observed survival time. Right censoring typically appears in studies where individuals enter the study gradually and are only followed within a certain observation period. Observations where the event did not take place until the end of the observation period are right censored. A censoring concept that is often assumed to hold was already presented in Section 1.2 and is called random censoring. Here the survival time T_i and the censoring time C_i of each individual i, $i = 1, \ldots, n$ are assumed to be independent random variables, the observed survival time t_i is the minimum of those two variables and the censoring indicator δ_i is defined as in (1.1). Considering time–constant covariates $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{ip})$ and under the assumption of non–informative censoring, i.e. the assumption that the censoring time is not determined by parameters of interest, the likelihood contribution of individual



– – – – no censoring, no truncation, time–constant covariates $v_1 = 2, v_2 = 1$

- - - - - - - - no censoring, no truncation, $v_1 = 2$, v_2 is changing from 2 to 3

— — no censoring, left truncated, time–constant covariates $v_1 = 2, v_2 = 3$

..... right censored, no truncation, time–constant covariates $v_1 = 1, v_2 = 1$

———— right censored, left truncated, $v_1 = 1$, v_2 is changing from 2 to 3

Figure 1.1: Illustration of 5 different right–censoring and left–truncation schemes each with two covariates $v_1 \in \{1, 2\}$ and $v_2 \in \{1, 2, 3\}$, that may be time–constant or time–varying.

i is given by

$$\begin{aligned} L_i &= \begin{cases} f_i(t_i, \boldsymbol{v}_i) = \lambda_i(t_i, \boldsymbol{v}_i) \cdot S_i(t_i, \boldsymbol{v}_i), & \delta_i = 1 \\ S_i(t_i, \boldsymbol{v}_i), & \delta_i = 0 \end{cases} \\ &= \lambda_i(t_i, \boldsymbol{v}_i)^{\delta_i} \cdot S_i(t_i, \boldsymbol{v}_i) \\ &= \lambda_i(t_i, \boldsymbol{v}_i)^{\delta_i} \cdot \exp\left(- \int_0^{t_i} \lambda_i(u, \boldsymbol{v}_i) du \right). \end{aligned} \tag{1.5}$$

For non–censored observations the likelihood is as usual given by the density f_i at $t_i, \boldsymbol{v}_i$, whilst the likelihood for censored observations, where it is only known that the survival time is at least t_i, is given by the survivor function S_i at $t_i, \boldsymbol{v}_i$. Thus, under the usual assumption of conditional independence the likelihood for the whole sample $(t_i, \delta_i, \boldsymbol{v}_i)$, $i = 1, \ldots, n$ is given by

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \lambda_i(t_i, \boldsymbol{v}_i)^{\delta_i} \cdot S_i(t_i, \boldsymbol{v}_i).$$

1.4.2 Left truncation

Left truncation is the second type of incompletely observed survival data that we deal with in this thesis. Here survival times of certain individuals can only be observed on condition that they exceed a certain, individual–specific truncation time T_i^{tr} involving that some survival times are not to be observed. Left truncation typically occurs in observation studies where individuals that have already been at risk for a known, individual–specific amount of time t_i^{tr} at the beginning of the observation period are included (additionally to individuals that get at risk at a later date within the observation period and thus enter the study gradually). Note that those observations would not be included if their survival time was shorter than t_i^{tr}, i.e. it has to be considered that no shorter survival time than t_i^{tr} may be observed with those individuals. This matter of fact is crucial in data situations, where those individuals being at risk at earlier times differ from individuals being at risk later, regarding values of influential covariates. Considering time–constant covariates, right censoring and left truncation the data are given by

$$(t_i, \delta_i, t_i^{tr}; \boldsymbol{v}_i), \quad i = 1, \ldots, n,$$

where $t_i^{tr} = 0$ if observation i is not left truncated and $t_i^{tr} > 0$ if observation i is left truncated. The individual likelihood contribution of individual i is given by

$$\begin{aligned} L_i &= \begin{cases} S_i(t_i, \boldsymbol{v}_i), & \delta_i = 0, t_i^{tr} = 0 \\ \lambda_i(t_i, \boldsymbol{v}_i) \cdot S_i(t_i, \boldsymbol{v}_i), & \delta_i = 1, t_i^{tr} = 0 \\ \frac{S_i(t_i, \boldsymbol{v}_i)}{S_i(t_i^{tr}, \boldsymbol{v}_i)}, & \delta_i = 0, t_i^{tr} > 0 \\ \lambda_i(t_i, \boldsymbol{v}_i) \frac{S_i(t_i, \boldsymbol{v}_i)}{S_i(t_i^{tr}, \boldsymbol{v}_i)}, & \delta_i = 1, t_i^{tr} > 0 \end{cases} \\ &= \lambda_i(t_i, \boldsymbol{v}_i)^{\delta_i} \cdot \exp\left(- \int_{t_i^{tr}}^{t_i} \lambda_i(u, \boldsymbol{v}_i) du \right), \end{aligned} \tag{1.6}$$

where left truncation is accounted for by conditioning on $T_i > t_i^{tr}$, which results in a division by $S(t_i^{tr}, \boldsymbol{v}_i)$. For a detailed derivation of these likelihoods see e.g. Klein and Moeschberger (2003). Again, under the usual assumption of conditional independence the likelihood for the whole sample $(t_i, \delta_i, t_i^{tr}, \boldsymbol{v}_i)$, $i = 1, \ldots, n$ is given by the product of the individual likelihood contributions.

1.4.3 Time–varying covariates

So far we have only considered time–constant covariates. Now we will illustrate, how the likelihood of survival data with time–varying (piecewise constant) covariates can be rewritten in the form of the likelihood of left truncated survival data with time–constant covariates. For this purpose consider for instance survival data

$$(t_i, \delta_i, t_i^{tr}; v_i(t)), \quad i = 1, \ldots, n,$$

where $v(t)$ is a time–varying covariate that may take two different values $v^{(1)}$ and $v^{(2)}$. The likelihood contribution of an observation i with a trajectory as displayed in Figure 1.2, where $t_i^{(1)}$ and $t_i^{(2)}$ mark the points of time when the covariates change, is given by

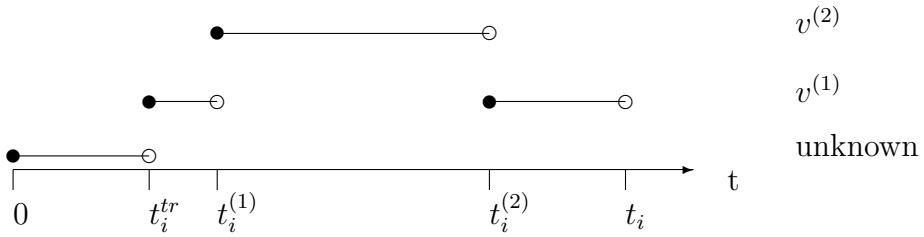


Figure 1.2: Exemplary trajectory for a time–varying covariate $v(t)$ with two different values $v^{(1)}$ and $v^{(2)}$.

$$
\begin{aligned}
L_i &= \lambda_i(t_i, v_i(t_i))^{\delta_i} \cdot \exp\left(- \int_{t_i^{tr}}^{t_i} \lambda_i(u, v_i(u)) du \right) \\
&= \lambda_i(t_i, v^{(1)})^{\delta_i} \cdot \exp\left(- \int_{t_i^{tr}}^{t_i^{(1)}} \lambda_i(u, v^{(1)}) du - \int_{t_i^{(1)}}^{t_i^{(2)}} \lambda_i(u, v^{(2)}) du - \int_{t_i^{(2)}}^{t_i} \lambda_i(u, v^{(1)}) du \right) \\
&= \lambda_i(t_i^{(1)}, v^{(1)})^0 \cdot \exp\left(- \int_{t_i^{tr}}^{t_i^{(1)}} \lambda_i(u, v^{(1)}) du \right) \\
&\quad \lambda_i(t_i^{(2)}, v^{(2)})^0 \cdot \exp\left(- \int_{t_i^{(1)}}^{t_i^{(2)}} \lambda_i(u, v^{(2)}) du \right) \\
&\quad \lambda_i(t_i, v^{(1)})^{\delta_i} \cdot \exp\left(- \int_{t_i^{(2)}}^{t_i} \lambda_i(u, v^{(1)}) du \right).
\end{aligned}
$$

This individual likelihood is identical to the likelihood of three left truncated observations with a time–constant covariate given by

t_i	δ_i	t_i^{tr}	v_i
$t_i^{(1)}$	0	t_i^{tr}	$v^{(1)}$
$t_i^{(2)}$	0	$t_i^{(1)}$	$v^{(2)}$
t_i	δ_i	$t_i^{(2)}$	$v^{(1)}$

For this reason time–varying covariates can be included in the settings described before via data augmentation.

1.5 Modelling the baseline hazard

As can be seen from equations (1.5) and (1.6) the calculation of the likelihood involves solving integrals over the baseline hazard rate, which is the only component that depends on time in case of time–constant covariate effects. Depending on the complexity of the assumed structure of the baseline hazard the integrals may be solved analytically or a numerical integration technique may be required. Starting from the Cox model (1.3) where the baseline hazard is typically unspecified, we present three alternatives to specify the baseline hazard and their implications with calculating the likelihood.

1.5.1 Weibull model

The first alternative is a parametric Weibull model, where the baseline hazard rate is given by

$$\lambda_0(t) = \alpha t^{\alpha-1}$$

with an unknown shape parameter $\alpha > 0$. Note that the exponential model, where the baseline hazard is time–constant is included as the special case of $\alpha = 1$, whereas values of $\alpha < 1$ ($\alpha > 1$) yield a decreasing (increasing) baseline hazard. To give an example, Figure 1.3 displays the shapes of $\lambda_0(t)$ for $\alpha = 0.75$, $\alpha = 1$ and $\alpha = 1.25$. Usually the Weibull distribution is defined by a shape and a scale parameter. With our model the scale parameter is included as an additive intercept term γ_0 in the linear predictor, i.e. the model is given by

$$\lambda_i(t, \boldsymbol{v}_i) \;=\; \alpha t^{\alpha-1} \cdot \exp\left(\gamma_0 + \boldsymbol{v}_i'\boldsymbol{\gamma}\right).$$

The integral in the likelihood given in (1.6) can be calculated analytically as follows

$$\begin{aligned}
\int_{t_i^{tr}}^{t_i} \lambda_i(u, \boldsymbol{v}_i)du &= \int_{t_i^{tr}}^{t_i} \alpha u^{\alpha-1} \cdot \exp\left(\gamma_0 + \boldsymbol{v}_i'\boldsymbol{\gamma}\right) du \\
&= \exp\left(\gamma_0 + \boldsymbol{v}_i'\boldsymbol{\gamma}\right) \cdot \int_{t_i^{tr}}^{t_i} \alpha u^{\alpha-1} du \\
&= \exp\left(\gamma_0 + \boldsymbol{v}_i'\boldsymbol{\gamma}\right) \cdot [u^\alpha]_{t_i^{tr}}^{t_i} \\
&= \exp\left(\gamma_0 + \boldsymbol{v}_i'\boldsymbol{\gamma}\right) \cdot \left((t_i)^\alpha - (t_i^{tr})^\alpha\right).
\end{aligned}$$

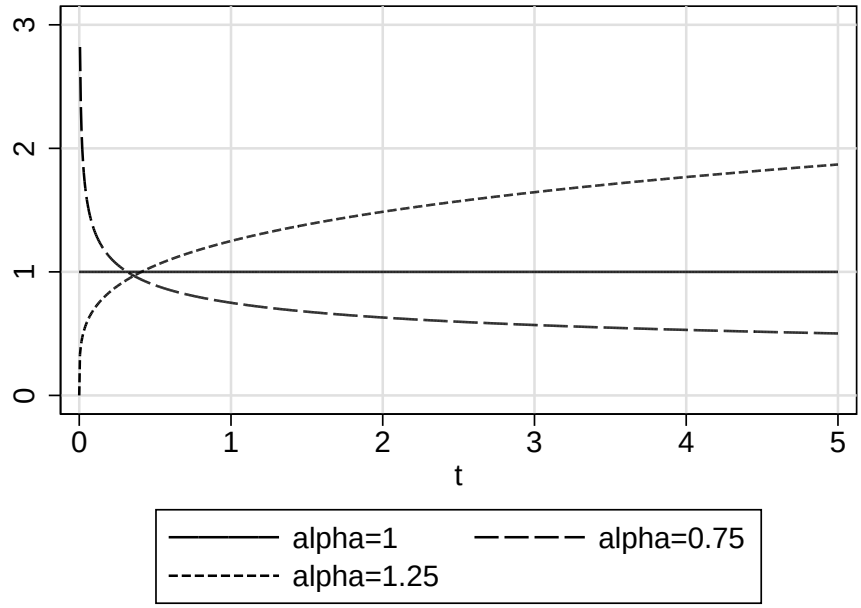


Figure 1.3: Shape of the Weibull baseline hazard $\alpha t^{\alpha-1}$ for different values of α.

Thus a Weibull hazard rate allows for an easy estimation and is a frequently used model assumption. However, the flexibility is limited since the shape of the baseline hazard rate is restricted to monotonic functions as displayed in Figure 1.3.

1.5.2 Piecewise exponential model (p.e.m.)

The basic idea of the p.e.m. is to divide the time axis into a grid that may be equidistant, according to quantiles or of any arbitrary structure given by the intervals

$$(0 = \xi_0, \xi_1], (\xi_1, \xi_2], \ldots, (\xi_{s-1}, \xi_s], \ldots, (\xi_{m-1}, \xi_m], (\xi_m, \infty),$$

where ξ_m is the largest of all observed survival times t_i, $i = 1, \ldots, n$. The baseline hazard rate $\lambda_0(t)$ is assumed to be piecewise constant on that grid, i.e.

$$\lambda_0(t) = \lambda_{0s}, \quad \lambda_{0s} \geq 0$$

for times t within the intervals $(\xi_{s-1}, \xi_s]$, $s = 1, ..., m$. Since estimating the unknown parameters λ_{0s} would involve imposing the restrictions $\lambda_{0s} \geq 0$, $s = 1, \ldots, m$, we prefer to estimate the unrestricted parameters $g_{0s} = \log(\lambda_{0s})$ instead, i.e. we define

$$\gamma_0(t) = \log(\lambda_0(t)) = \gamma_{0s}$$

for times t in the interval $(\xi_{s-1}, \xi_s]$, $s = 1, ..., m$. Furthermore, let $\eta_i(t_i, \boldsymbol{v}_i)$ denote the whole linear predictor of individual i including the log–baseline hazard, i.e.

$$\begin{aligned} \eta_i(t, \boldsymbol{v}_i) &= \gamma_0(t) + \boldsymbol{v}_i'\boldsymbol{\gamma} \quad \text{and hence} \\ \lambda_i(t, \boldsymbol{v}_i) &= \lambda_0(t) \cdot \exp(\boldsymbol{v}_i'\boldsymbol{\gamma}) = \exp(\gamma_0(t) + \boldsymbol{v}_i'\boldsymbol{\gamma}) = \exp(\eta_i(t, \boldsymbol{v}_i)) \end{aligned}$$

Here, $\eta_{is} = \gamma_{0s} + \boldsymbol{v}_i'\boldsymbol{\gamma}$ denotes the piecewise constant linear predictor in the time interval $(\xi_{s-1}, \xi_s]$, $s = 1, ..., m$.

In the case of a p.e.m., the integral reduces to a sum, and, after some calculations, the likelihood contribution of observation i in each time interval $(\xi_{s-1}, \xi_s]$ can be expressed as

$$L_{is} = \exp(y_{is}\eta_{is} - \exp(\Delta_{is} + \eta_{is}))$$

where

$$y_{is} = \begin{cases} 1 & t_i \in (\xi_{s-1}, \xi_s], \delta_i = 1 \\ 0 & \text{else.} \end{cases}$$

$$\Delta_{is}^* = \begin{cases} 0, & \xi_s < t_i^{tr} \\ \xi_s - t_i^{tr}, & \xi_{s-1} < t_i^{tr} \leq \xi_s < t_i \\ t_i - t_i^{tr}, & \xi_{s-1} < t_i^{tr} < t_i \leq \xi_s \\ \xi_s - \xi_{s-1}, & t_i^{tr} \leq \xi_{s-1} < \xi_s < t_i \\ t_i - \xi_{s-1}, & t_i^{tr} < \xi_{s-1} < t_i \leq \xi_s \\ 0, & t_i \leq \xi_{s-1} \end{cases}$$

$$\Delta_{is} = \log \Delta_{is}^* \quad (\Delta_{is} = -\infty \text{ if } \Delta_{is}^* = 0).$$

That is to say that the likelihood of a p.e.m. is proportional to a Poisson–likelihood with responses y_{is} and with the predictor η_{is} containing an additional offset term Δ_{is}, see Fahrmeir and Tutz (2001, Section 9.1) or Ibrahim et al. (2001, Section 3.1) for details. This result yields that a p.e.m. may be estimated based on methodology for Poisson regression models, i.e. within the context of generalized linear models (GLMs) via data augmentation. In practise this means that the data set has to be modified in such a way that for every individual i there is an observation row for each interval $(\xi_{s-1}, \xi_s]$ beginning with the interval that includes the left truncation time t_i^{tr} up to the interval in that observation time t_i ends. Instead of the indicator of non–censoring δ_i the modified data set contains the

indicator y_{is}, instead of survival time t_i the variable ξ_s as well as the offset Δ_{is} (covariates are duplicated). To give a short example, if we have an equidistant grid with length 0.1, the observations

i	t	δ	t^{tr}	v_1	v_2
1	0.35	1	0.16	0	3
2	0.12	0	0	1	5
⋮	⋮	⋮	⋮		

have to be modified to

i	y	ξ	Δ	v_1	v_2
1	0	0.2	log(0.04)	0	3
1	0	0.3	log(0.10)	0	3
1	1	0.4	log(0.05)	0	3
2	0	0.1	log(0.10)	1	5
2	0	0.2	log(0.02)	1	5
⋮	⋮	⋮	⋮	⋮	⋮

and a Poisson regression with response y, covariates ξ, v_1 and v_2 and offset Δ may be accomplished. Note that time–varying covariates can be accounted for by varying the covariates adequately from line to line in the table above. Hence, the assumption of a p.e.m. is quite convenient, however, due to the assumption of a piecewise constant hazard rate the estimated (log–)baseline effect is a step function on the defined grid, which may not be adequate with continuous survival times. Furthermore it is a moot question how to choose an "ideal" grid. While a small grid size might lead to intervals with sparse data and hence unreliable estimates of the according parameters, a large grid size might not allow for enough flexibility. Within our Bayesian analysis we will use a rather small grid size and specify random walk priors (as described in Fahrmeir and Lang (2001a)) for the parameters γ_{0s} to penalize too abrupt jumps between neighboring parameters $\gamma_{0,s-1}$ and γ_{0s} yielding a flexible "smooth step function".

1.5.3 P–spline model

Modelling the baseline hazard by a (Bayesian) P–spline is the most flexible alternative, that will be primarily discussed within this thesis. While the log–baseline hazard is assumed to be a piecewise constant function with the p.e.m., i.e. a polynomial of degree zero within each predefined interval, we now consider extensions to piecewise polynomials of an arbitrary degree l. Depending on the degree l this leads to more or less smooth functions instead of step functions. Again, the time axis is divided into a grid

$$(0 = \xi_0, \xi_1], (\xi_1, \xi_2], \ldots, (\xi_{s-1}, \xi_s], \ldots, (\xi_{m-1}, \xi_m], (\xi_m, \infty),$$

where ξ_s are usually called (inner) knots within the context of spline regression. Then a polynomial spline has the following smoothness properties:

- A spline is a polynomial of degree l within each interval ξ_{s-1}, ξ_s, $s = 1, \ldots, m$.
- A spline is $l - 1$ times differentiable at the knots ξ_s.

As shown in De Boor (1978) a spline with those properties may for example be written as a linear combination of $M = m + l$ B–spline basis functions B_{sl} of degree l. Hence the function $g_0(t) = \log(\lambda_0(t))$ that is denoting the (smooth) function that describes the log–baseline effect can be written as

$$g_0(t) = \sum_{s=1}^{M} \beta_s B_{sl}(t),$$

where β_s, $s = 1, \ldots, M$ are unknown parameters. Note, that we are again estimating the log–baseline hazard instead of the baseline hazard to avoid implying the restriction $\lambda_0(t) \geq 0$. Figure 1.4 shows B–spline basis functions for degrees $l = 0$, $l = 1$ and $l = 2$, respectively, with only several basis functions being displayed for reasons of clarity. B–spline basis functions of degree zero are piecewise constant and do not overlap (in this respect that each basis function B_{s0} is nonzero only within the interval $(\xi_{s-1}, \xi_s]$), which again illustrates that the p.e.m. is included as the special case of $l = 0$. B–spline basis functions of degree one are nonzero within the range of two subsequent intervals $(\xi_{s-1}, \xi_s]$ and $(\xi_s, \xi_{s+1}]$ and are linear functions within each interval, whereas basis functions of degree two are nonzero within the range of three subsequent intervals and are quadratic functions

within each interval. B–spline basis functions of degree $l = 3$ (not shown) would be nonzero within the range of four subsequent intervals and be cubic functions within each interval etc. Figure 1.5 exemplarily shows the construction of a B–spline of degree $l = 2$ with $m = 5$ inner knots. Panel (a) displays the $M = l + m = 6$ basis functions B_{s2} that cover the considered range of $(0, 1]$ in a way that for each point within this interval $l + 1 = 3$ basis functions take (positive) values different from zero. The weighted basis functions $\beta_s B_{s2}$ are displayed in panel (b) and panel (c) shows the resulting spline function, which is the sum of the weighted basis functions $\sum_{s=1}^{M} \beta_s B_{s2}$. For more details on B–splines see De Boor (2001).

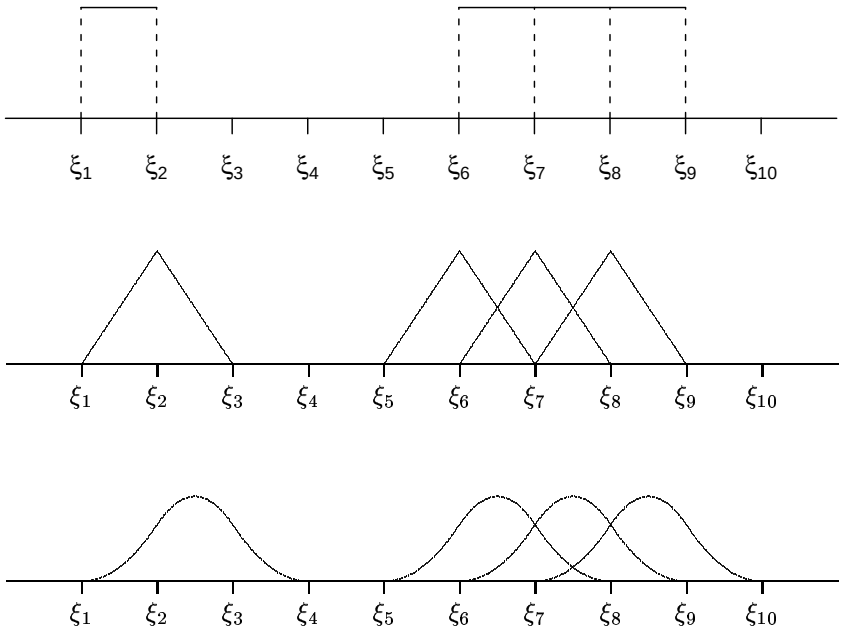
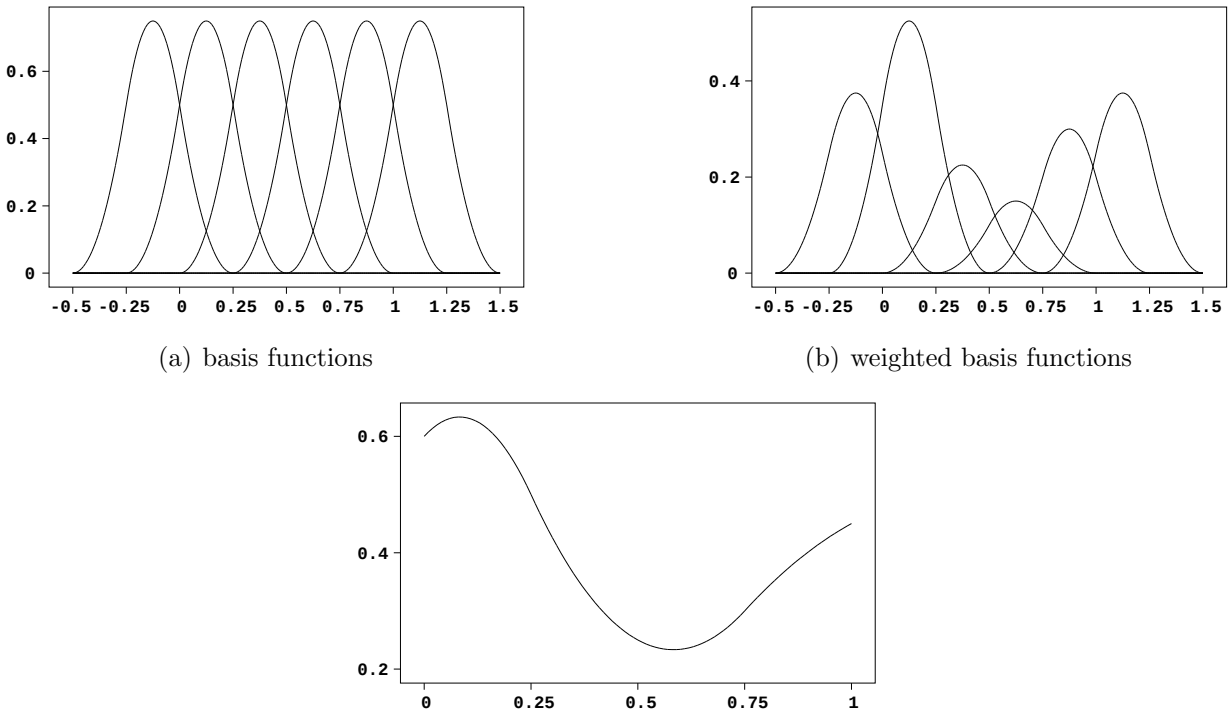


Figure 1.4: Some B–spline Basis functions for degrees $l = 0$, $l = 1$, and $l = 2$, respectively.

Besides the degree l the structure of the resulting spline considerably depends on the number and the position of the knots. While a small number of knots might not guarantee enough flexibility, a very large number of knots might lead to over–fitting and thus deliver unreliable results. An attractive solution to this problem are penalized splines (P–splines), that are based on roughness penalties and presented by Eilers and Marx (1996). The basic idea is to use a rather large number of equidistant knots, but penalize too rough functions



(a) basis functions

(b) weighted basis functions

(c) resulting spline

Figure 1.5: Construction of B–splines: 6 B–spline basis functions of degree $l = 2$ with 5 inner knots at 0.0, 0.25, 0.5, 0.75 and 1.0 (a), weighted basis functions (b), and the resulting spline (c).

by imposing difference penalties on neighboring parameters β_{s-1} and β_s. In this thesis we will use Bayesian versions of P–splines as developed in Lang and Brezger (2004).

While the integrals in the likelihood (1.5) and (1.6), respectively, can be solved analytically with the two previously presented approaches (Weibull model and p.e.m.), this is not in general true for survival models where the baseline hazard is modelled by a P–spline. Apart from B–splines of degree $l = 0$ and $l = 1$ these integrals can only be solved numerically. For this we use numerical integration in form of the trapezoidal rule. Here the basic idea is to approximate the function $\lambda_0(t)$ by a piecewise linear function $\tilde{\lambda}_0(t)$ as displayed in Figure 1.6, where the area under $\tilde{\lambda}_0$ is trapezoidal within each interval. In order to guarantee that the approximation is also fairly accurate in time slices where observed survival times are sparse, equidistant time points are used as additional knots besides the observed life times t_i. Now an integral $\int_0^t \lambda_0(u)du$ is approximated by $\int_0^t \tilde{\lambda}_0(u)du$, which is the sum of the areas of the corresponding trapezoids. For an interval $(t_{i-1}, t_i]$ the area of

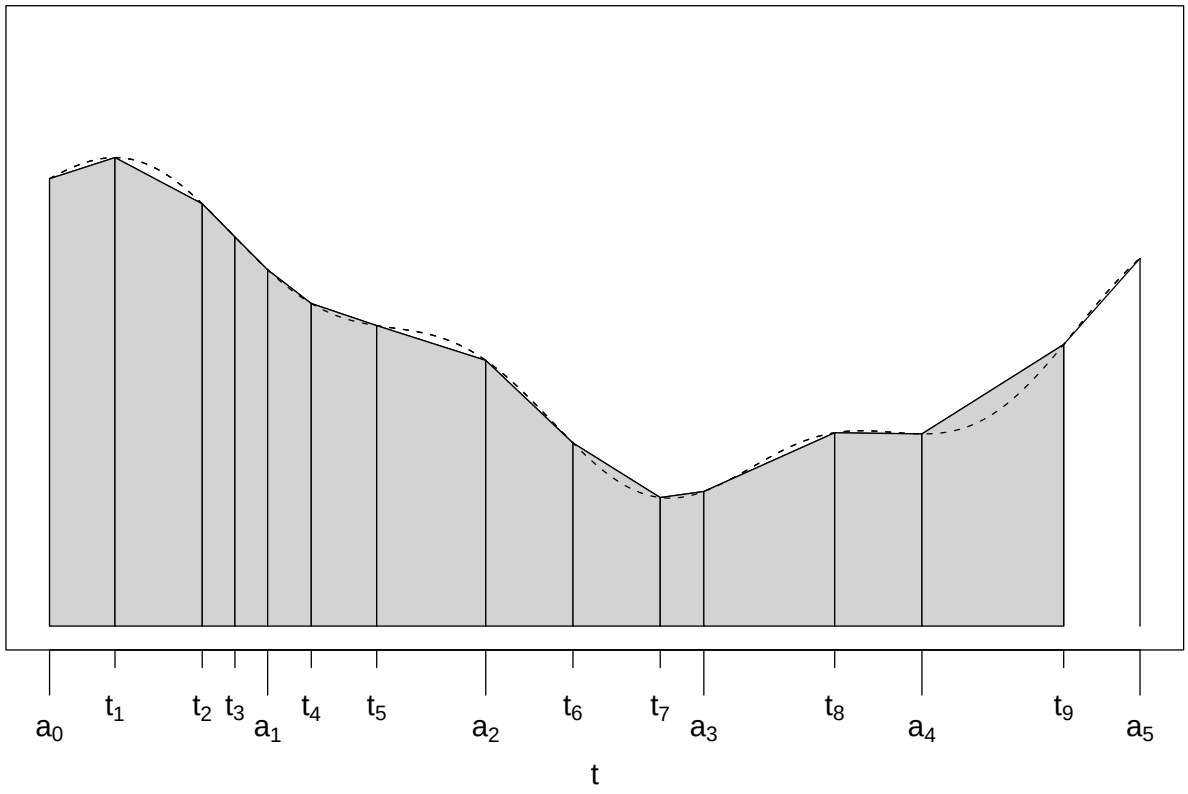


Figure 1.6: Trapezoidal rule: The function $\lambda_0(t)$ (dashed line) is approximated by a piecewise linear function through the points $(t_i, \lambda_0(t_i))$, where t_i are (ordered) observed survival times, as well as the additional points $(a_s, \lambda_0(a_s))$ with $a_0 = 0$. Hence the integral $\int_0^{t_9} \lambda_0(u)du$ is approximated by the sum of the areas of the gray shaded trapezoids.

the corresponding trapezoid would for example be given by

$$\frac{1}{2} \cdot (t_i - t_{i-1}) \cdot (\lambda_0(t_i) + \lambda_0(t_{i-1})).$$

Note, that we discussed the case of data where no left truncation is present, but the trapezoidal rule may be applied to left truncated observations as well. Here we use the truncation times t_i^{tr} as additional knots in Figure 1.6 and approximate the integrals $\int_{t_i^{tr}}^{t_i} \lambda_0(u)du$ by $\int_{t_i^{tr}}^{t_i} \tilde{\lambda}_0(u)du$.

1.6 Relations to other survival models

The scope of this section is to clarify in what way the extended continuous–time Cox model that is presented in this thesis is related to some other common survival models, namely to discrete time survival models and to location–scale models for $\log(T)$, which are also called accelerated failure time (AFT) models.

1.6.1 Discrete time survival analysis

Discrete time survival models are basically used in two different situations. Firstly they are used in cases where failures actually only occur at discrete time points. This is for example the case with durations of unemployment, since employments usually end and begin with monthly allowance. The second situation is given where failures may occur at any arbitrary point of time, however survival times can not be observed continuously, but are only known to lie between two successive follow ups. This case is known as interval censoring and typically occurs in medical studies where data can only be observed at regular consultations.

Now consider discrete time $D \in \{1, 2, \ldots\}$, then the discrete hazard function is given by

$$\lambda^{discr}(s, \boldsymbol{v}) = P(D = s | D \geq s, \boldsymbol{v}), \quad s = 1, 2, \ldots$$

which is the probability of failure at time point s, given that the failure time is at least s and given the covariates $\boldsymbol{v}$.

The grouped proportional hazards model

Consider the case of interval censoring where survival times T are continuous but are only observed at k follow up times ξ_s, $s = 1, \ldots, k$. Let time be divided in intervals

$$[\xi_0, \xi_1), [\xi_1, \xi_2), \ldots, [\xi_{s-1}, \xi_s), \ldots, [\xi_{q-1}, \xi_q), [\xi_k, \infty)$$

where $\xi_0 = 0$ and $q = k - 1$. Then $D = s$, $s = 1, \ldots, q$ denotes failure within the according interval, i.e. $T \in [\xi_{s-1}, \xi_s)$ and the discrete hazard function is given by

$$\begin{aligned}
\lambda^{discr}(s,\boldsymbol{v}) &= P(D=s|D\geq s,\boldsymbol{v}) = P(T<\xi_s|T\geq\xi_{s-1},\boldsymbol{v}) \\
&= \frac{P(\xi_{s-1}\leq T<\xi_s|\boldsymbol{v})}{P(T\geq\xi_{s-1}|\boldsymbol{v})} = \frac{F(\xi_s,\boldsymbol{v})-F(\xi_{s-1},\boldsymbol{v})}{S(\xi_{s-1},\boldsymbol{v})} \\
&= \frac{S(\xi_{s-1},\boldsymbol{v})-S(\xi_s,\boldsymbol{v})}{S(\xi_{s-1},\boldsymbol{v})} \\
&= 1-\frac{S(\xi_s,\boldsymbol{v})}{S(\xi_{s-1},\boldsymbol{v})} \\
&= 1-\frac{\exp\left(-\int_0^{\xi_s}\lambda(t,\boldsymbol{v})dt\right)}{\exp\left(-\int_0^{\xi_{s-1}}\lambda(t,\boldsymbol{v})dt\right)} \\
&= 1-\exp\left(-\int_{\xi_{s-1}}^{\xi_s}\lambda(t,\boldsymbol{v})dt\right), \qquad (1.7)
\end{aligned}$$

i.e. the discrete hazard function λ^{discr} may be written as a function of the continuous survivor function S and the continuous hazard rate λ, respectively. Inserting the formula of the proportional hazards model or Cox model for continuous time (1.3) given by

$$\lambda(t,\boldsymbol{v}) = \lambda_0(t)\cdot\exp\left(\boldsymbol{v}'\boldsymbol{\gamma}\right)$$

into (1.7) yields the grouped proportional hazards model given by

$$\begin{aligned}
\lambda^{discr}(s,\boldsymbol{v}) &= 1-\exp\left(-\exp(\boldsymbol{v}'\boldsymbol{\gamma})\int_{\xi_{s-1}}^{\xi_s}\lambda_0(t)dt\right) \\
&= 1-\exp\left(-\exp(\gamma_{0s}+\boldsymbol{v}'\boldsymbol{\gamma})\right)
\end{aligned}$$

with $\gamma_{0s} = \log\left(\int_{\xi_{s-1}}^{\xi_s}\lambda_0(t)dt\right)$. An alternative formulation is given by

$$\log(-\log(1-\lambda^{discr}(s,v))) = \gamma_{0s}+\boldsymbol{v}'\boldsymbol{\gamma}$$

and hence the grouped proportional hazards model is a sequential complementary log–log model. Though in general grouping implies a loss of information (see e.g. Gould and Lawless (1988)), it should be annotated that the parameter vector $\boldsymbol{\gamma}$ remains unchanged by the transition between the continuous and the discrete model.

Models for binary response

While sequential models for ordinal responses fit for the estimation of discrete time survival models without right–censoring, binary models for the indicators

$$y_{is} = \begin{cases} 1 & d_i = s \text{ and } \delta_i = 1 \\ 0 & \text{else} \end{cases} \qquad i = 1, \ldots, n, \quad s = 1, \ldots, d_i \tag{1.8}$$

that indicate whether or not a failure occurred with individual i at time s or in interval $[\xi_{s-1}, \xi_s)$, respectively, may be used instead for discrete survival analysis in cases where right–censoring is present. For the purpose of fitting a binary model, the survival data have to be augmented in a similar way as described in Subsection 1.5.2 for the p.e.m. Note however, that the p.e.m. is a continuous time survival model, where the information on the exact survival time (within each interval) is retained and enters the model via an offset term. To give a short example, right–censored discrete time survival data with two covariates v_1 and v_2 given by

i	d	δ	v_1	v_2
1	3	1	0	3
2	2	0	1	5
⋮	⋮	⋮	⋮	⋮

have to be augmented to

i	s	y	v_1	v_2
1	1	0	0	3
1	2	0	0	3
1	3	1	0	3
2	1	0	1	5
2	2	0	1	5
⋮	⋮	⋮	⋮	⋮

where for every individual i one line has to be created for each discrete point in time s (every time interval, respectively) at which the individual is observed. Covariates are duplicated and an indicator variable y_{is} is created according to (1.8), that takes the

value 0 in every line related to a right–censored observation i and takes the value 1 in the last line related to an uncensored observation i and 0 in the preceding lines. In practice such a data augmentation for right censored data with time–constant covariates might for example be accomplished with the STATA command `prsnperd` as illustrated on `http://www.ats.ucla.edu/stat/stata/library/survival2.htm`. Note that left–truncation and time–varying covariates can be easily accounted for by omitting the accordant first lines in the tabular above and varying the covariates adequately from line to line, respectively. The discrete hazard function may now equivalently be written as

$$\lambda_i^{discr}(s, \boldsymbol{v}_i) = P(y_{is} = 1 | \boldsymbol{v}_i)$$

and a binary regression with response y_{is} and covariates s, v_1 and v_2 may be accomplished. Thompson (1977) for example considers the logistic model

$$\lambda^{discr}(s, \boldsymbol{v}) = \frac{\exp(\gamma_{0s} + \boldsymbol{v}'\boldsymbol{\gamma})}{1 + \exp(\gamma_{0s} + \boldsymbol{v}'\boldsymbol{\gamma})} \tag{1.9}$$

and shows that this model is very similar to the proportional hazards model if grouping intervals become short.

1.6.2 Log–location–scale models

Log–location–scale models are an alternative model class for continuous survival data, where, in contrast to the classical Cox model (with time–constant covariates and time–constant effects of covariates), proportional hazards are not presumed in general. To account for the nonnegativity of survival time, $\log(T)$ instead of T is related to a linear predictor given by

$$\log(T) = \gamma_0 + \boldsymbol{v}'\boldsymbol{\gamma} + \sigma\varepsilon, \tag{1.10}$$

where σ is a constant scale parameter and ε is an error term independent of $\boldsymbol{v}$. For the special case where ε follows the standard extreme value distribution, we retain a Weibull model with $\lambda = \exp\left(-(\gamma_0 + \boldsymbol{v}'\boldsymbol{\gamma})\right)$ and $\alpha = 1/\sigma$, which is a proportional hazards model. However, other distributions of ε do not yield proportional hazards models. Assuming a normal distribution for ε, for example, results in lognormal distributed survival times and a nonproportional hazards model. In the case of parametric models the parameters can be estimated easily via maximum likelihood techniques. Since such parametric approaches

are quite restrictive, semi–parametric procedures that leave the distribution of the error term unspecified and only provide the estimation of the parameter vector $\boldsymbol{\gamma}$ may be used alternatively, see e.g. Kalbfleisch and Prentice (2002). However, there is no pendant to the Breslow estimator (1.4) (which is used with Cox models to estimate the cumulative baseline hazard on the basis of the estimated parameter vector $\hat{\boldsymbol{\gamma}}$) and hence semi–parametric procedures may not be used in cases where prediction is of interest. A method that offers a joint estimation of the parameter vector $\boldsymbol{\gamma}$ and the baseline hazard with a flexible, smooth error distribution is presented by Komárek, Lesaffre and Hilton (2005), who propose to use a mixture of normals for the error distribution, with mixture weights being smoothed.
The model assumption in (1.10) may be rewritten as

$$T = \exp(\gamma_0 + \boldsymbol{v}'\boldsymbol{\gamma}) \exp(\sigma\varepsilon),$$

which, according to Lesaffre, Komárek and Declerck (2004), leads to a hazard rate of the following structure

$$\lambda(t, \boldsymbol{v}) = \exp\left(-(\gamma_0 + \boldsymbol{v}'\boldsymbol{\gamma})\right) \cdot \lambda_0(\exp(-(\gamma_0 + \boldsymbol{v}'\boldsymbol{\gamma})) \cdot t),$$

i.e. as with the Cox model the covariates act multiplicatively on the hazard rate, but here the effect of a covariate additionally acts as an acceleration (deceleration) of the event time, which explains why such models are also referred to as accelerated failure time models (AFT). In this way classical AFT models seem to be more general than classical Cox models, however, in contrast to Cox models, AFT models do not allow for the inclusion of time–varying covariates and time–varying effects. For this reason we focus on extensions of the Cox model, where the inclusion of time–varying covariates and time–varying effects yields nonproportional hazards models.

1.7 Competing risks and multi–state models

So far we have only considered one type of failure. However, with a number of applications one may distinguish between several types of failures or events. In clinical studies, for example, the events may stand for several causes of death. Models for this type of data are referred to as competing risks models. Let $h = 1, \ldots, H$ denote the distinct events.

Corresponding to the definition of the hazard rate in (1.3) event–specific individual hazard rates λ_{hi} are given by

$$\lambda_{hi}(t, v_{hi}) = \lambda_{h0}(t) \exp\left(v_{hi}' \gamma_h\right), \quad h = 1, \ldots, H. \tag{1.11}$$

Here, $\lambda_{h0}(t)$ denotes the event–specific baseline hazard, v_{hi} denotes the vector of covariates having an influence on the accordant hazard rate and γ_h is the related event–specific vector of parameters. As mentioned in the context of Cox models this basic model is not adequate to many complex applications and needs to be extended with respect to the aspects mentioned in Section 1.3.

Note, that in a discrete time setting competing risks models may be analyzed by multicategorical regression models via data augmentation (in a similar way as described above for models for binary response). Fahrmeir and Lang (2001b), for example, present flexible, Bayesian multicategorical regression models to analyze discrete unemployment durations, with finding a full–time employment and finding a part–time employment as competing events.

Multi–state models present a further extension to survival models. The models described so far only consider one initial state and one or a number of terminating events. Multi–state models on the other hand may be applied to analyze general event history data. Here the various events are considered as transitions from one state to another. This type of data is for example given in clinical studies where the interest lies in analyzing transitions between different states of health.

Event–specific or transition–specific hazard rates are defined as before in (1.11), but the likelihood of multi–state models is slightly more complex than the likelihood of competing risks models. While (survival models and) competing risks models assume that each individual is at risk to experience any event during the whole observation time, this is not necessarily true with multi–state models. Here, a state structure specifies the diverse states (that might be absorbing or transient) and defines which transitions are possible and which ones are not. Thus it is important to consider that some individuals might not be at risk to experience certain events over periods of time being in certain states. For this reason the application of multicategorical regression models to discrete time event history data is not completely straightforward, but demands some additional consideration.

1.8 Overview

The thesis is organized as follows. The second chapter, which forms the core of this thesis, is based on the manuscript "Geoadditive Survival Models" by Hennerfeind, Brezger and Fahrmeir that is accepted for publication in the *Journal of the American Statistical Association (JASA) Theory and Methods Section*. Here we will present our nonparametric Bayesian survival model approach to extend the basic Cox model with respect to the aspects listed in Section 1.3. We will describe models, likelihood and priors for unknown functions and parameters, discuss the inference via MCMC and present some simulations and applications to different data sets.

In the third chapter of this thesis we deal with so called relative survival analysis, that is used to model the excess risk of a certain subpopulation relative to the base risk that is present in the whole population. Such models are typically used in the area of clinical studies, that aim at identifying prognostic factors for disease specific mortality with data on specific causes of death being not available. Our work has been motivated by real data on breast cancer where causes of death are not known. This chapter forms an extension of the analyses presented in the manuscript "Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geoadditive models" by Sauleau, Hennerfeind, Buemi and Held which is accepted for publication in *Statistics in Medicine*. The usefulness of our relative survival approach is supported by means of a simulated data set.

The fourth chapter is concerned with extensions to more general event history models. Embedded in the counting process framework (Andersen, Borgan, Gill and Keiding 1993) we present flexible multi–state models that are used to model transitions between a finite number of different states and include the survival model as well as the competing risks model as special cases. Applications to medical data on structural valve degeneration of biological prostheses and to sleep–electroencephalography data with multiple recurrent states of sleep illustrate our methods.

All approaches presented within this thesis are implemented in the statistical software package *B*ayesX. In Chapter 5 we present a tutorial to exemplify how Bayesian survival and multi–state models may be analyzed using *B*ayesX.

Chapter 2

Nonparametric regression for survival data

2.1 Introduction

In epidemiological, economic or social science applications, survival data often contain geographical or spatial information such as the district or postal code of the residence of individuals in the study. Analyzing and modelling geographical patterns for survival or waiting times, in addition to the impact of other covariates, is of obvious interest in many studies. For example, Henderson, Shimakura and Gorst (2002) model spatial variation in survival of acute myeloid leukemia patients in northwest England, Banerjee, Wall and Carlin (2003) apply a spatial frailty model to infant mortality in Minnesota, and Li and Ryan (2002) analyze the effect of risk factors on the onset of childhood asthma with spatial data from the East Boston Asthma Study. In Subsection 2.5.3 of this thesis, we will apply our approach to data on waiting times to coronary artery bypass graft (CABG). Within a discrete–time setting, spatial survival data from this study are analyzed by Crook, Knorr–Held and Hemingway (2003), and Fahrmeir, Lang, Wolff and Bender (2003) investigate the impact of small area labor market regions and other covariates, such as calendar time, age and unemployment benefits, on unemployment duration with discrete–time models.

A particular advantage of our approach is that all unknown functions and parameters are treated within a unified general framework by assigning appropriate priors with the same structure but different forms and degrees of smoothness. Based on previous work

(Fahrmeir and Lang, 2001a; Lang and Brezger, 2004) on semiparametric regression, nonlinear effects of unknown functions of time, in particular the log–baseline hazard rate, and of continuous covariates or further time scales are modelled through Bayesian versions of penalized splines (P–splines) introduced by Eilers and Marx (1996), Marx and Eilers (1998) for generalized additive models in a frequentist setting. Basically, time is treated in the same way as a continuous covariate, but the degree and amount of smoothness may be different. For example, simple random walk priors for the log–baseline effect in a piecewise exponential model are P–splines of degree zero. The spatial component is modelled by Gaussian Markov random field (MRF) priors, as common in disease mapping, by two–dimensional penalized tensor–product splines, or by a geostatistical (kriging) stationary Gaussian random field (GRF) model. From a computational point of view, MRF's and P–splines are clearly preferable to GRF's because their posterior precision matrices are band matrices or can be transformed into a band matrix–like structure. This special structure considerably speeds up computations and enhances numerical stability compared to the full precision matrices arising from the GRF approach.

For data observed on a irregular discrete lattice, MRF's seem to be most appropriate. If exact locations are available, P–spline or GRF surface smoothers seem to be more natural, but they can also be applied to discrete lattices after computing centroids of regions.

Our unified general framework also has theoretical and computational advantages for posterior analysis. Extending previous results for mixed models in Sun, Tsutakawa and Speckman (1999), we can show propriety of posteriors under regularity conditions. This is important, because some of our priors are diffuse or partially improper. From the computational point of view, full conditionals of blocks of parameters have similar structure, and lead to efficient MCMC techniques. Smoothing parameters are an integral part of the model and can be estimated jointly with unknown functions and other parameters. Inferential procedures have been implemented in C++ as part of *BayesX* (Brezger, Kneib and Lang 2005).

Non– and semiparametric Bayesian survival models have become quite popular in recent years, and some previous work deals with special or related cases of our approach. For fully Bayesian models without a spatial component Ibrahim, Chen and Sinha (2001) provide a good introduction and overview. Joint estimation of the baseline hazard and usual linear covariate effects in the Cox model has been considered by several authors.

Gamerman (1991) proposes a Gaussian random walk model for the log–baseline hazard in the piecewise exponential model, and Sinha (1993) suggests a joint Gaussian smoothness prior, and Cai, Hyndman and Wand (2002) and Cai and Betensky (2003) use a mixed model representation of linear regression splines to estimate the baseline hazard. In all these approaches, however, effects of continuous covariates are assumed to be of the usual linear parametric form, and no spatial component is present.

Survival models with a spatial component have recently been suggested in several publications. The approaches differ in the specification of the baseline hazard rate and in the model chosen for the spatial component, but the remaining part of the predictor is still of linear parametric form. Thus, non–parametric terms for flexible modelling and estimation of the effects of continuous covariates, further time scales and time–varying coefficients are not considered in these approaches. Li and Ryan (2002) add a spatial component in form of a stationary Gaussian process to the linear predictor of the Cox model. Treating the baseline hazard as a nuisance parameter, inference for the linear predictor and for correlation function parameters is based on a marginal rank likelihood. No procedure for estimating the spatial (random) effects is provided. Henderson et al. (2002) propose a Cox model with conditionally independent spatial gamma frailties, with means following either a geostatistical model or a Markov random field. For inference they use MCMC methods, except the baseline hazard estimate. For this they plug in the Breslow estimator at each iteration of the chain. Banerjee et al. (2003) assume a parametric Weibull baseline hazard and geostatistical or MRF priors for the spatial component. In comparison they prefer MRF priors, since computing times for geostatistical GRF models are much larger. This is in agreement with our own findings. Banerjee and Carlin (2003) develop Bayesian spatio–temporal survival models, modelling baseline hazard functions nonparametrically through a beta mixture approach and assuming MRF or CAR (conditionally autoregressive) priors for spatial effects, and Carlin and Banerjee (2002) extend this approach to multivariate MRF models, with applications to cancer survival data from Iowa. A good overview is given in Banerjee et al. (2004).

An empirical Bayes pendant to our semiparametric fully Bayesian approach has been developed in Kneib and Fahrmeir (2004) (see also Kneib, 2006).

The rest of the chapter is organized as follows. In Section 2.2 we describe models, likelihood, and priors for unknown functions and parameters. MCMC inference and model

choice are outlined in Subsection 2.3.1 and Subsection 2.3.2, respectively, and Subsection 2.3.3 provides results on the propriety of posteriors in geoadditive survival models under regularity assumptions. Performance is studied in Section 2.4 through simulation studies. Applications in Section 2.5 illustrate the method.

2.2 Models, likelihood and priors

2.2.1 Observation model and likelihood

Consider survival data in usual form, i.e., it is assumed that each individual i in the study has a lifetime T_i and a censoring time C_i that are independent random variables. The observed lifetime is then $t_i = \min(T_i, C_i)$, and δ_i denotes the censoring indicator. The data are then given by

$$(t_i, \delta_i; \boldsymbol{v}_i), \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{v}_i$ is the vector of covariates. Covariates may also be time–dependent, but we restrict discussion to time–constant covariates for simplicity. The same applies to left truncation (see Subsection 1.4.2), which might easily be included, but it is not discussed here for facility of inspection.

In Cox's proportional model the hazard rate for individual i is assumed as the product

$$\lambda_i(t; \boldsymbol{v}_i) = \lambda_0(t) \exp(\gamma_1 v_{i1} + \ldots + \gamma_r v_{ir}) = \lambda_0(t) \exp(\boldsymbol{v}_i' \boldsymbol{\gamma}). \tag{2.2}$$

The baseline hazard rate is unspecified, and, through the exponential link function, the covariates $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ act multiplicatively on the hazard rate. As pointed out in the introduction, in a number of applications there is a need for extending this basic model with respect to several aspects. We propose novel nonparametric Bayesian survival models that can deal with these issues in a flexible and unified framework. Reparametrizing the baseline hazard rate through $\exp\{g_0(t)\}$, $g_0(t) = \log\{\lambda_0(t)\}$ and partitioning the vector of covariates into groups of covariates $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s}$ and $\boldsymbol{v}$, we extend model (2.2) to the nonparametric multiplicative observation model

$$\lambda_i(t) := \lambda_i(t; \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{s}_i, \boldsymbol{v}_i) = \exp\{\eta_i(t)\} \tag{2.3}$$

with geoadditive predictor

$$\eta_i(t) = g_0(t) + \sum_{j=1}^{p} g_j(t) z_{ij} + \sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i'\boldsymbol{\gamma} + b_{g_i}. \tag{2.4}$$

Here $g_0(t) = \log\{\lambda_0(t)\}$ is the log–baseline effect, $g_j(t)$ is a time–varying effect of the covariate z_j, and $f_j(x_j)$ is the nonlinear effect of a continuous covariate x_j. The function $f_{spat}(s)$ is a (structured) spatial effect, where s, $s = 1, \ldots, S$ is either a spatial index, with $s_i = s$ if subject i is from area s, or an exact spatial coordinate $s = (x_s, y_s)$, e.g. for centroids of regions or if exact locations of individuals are known. The vector $\boldsymbol{\gamma}$ is the vector of usual linear fixed effects, and b_g is a subject– or group–specific frailty or random effect, with $b_{gi} = b_g$ if individual i is in group g, $g = 1, ..., G$. For $G = n$, we obtain individual–specific frailties, for $G < n$, b_g might be the effect of center g in a multicenter study or the unstructured (uncorrelated random) spatial effect of an area (i.e. $b_g = b_s$), for example. Random slopes could also be introduced, but we omit this here. Several other extensions of the model, such as choice of other link functions, inclusion of interactions and competing risks, are possible. We discuss this in the concluding section. For identifiability reasons, we center all unknown functions about zero, and include an intercept term in the parametric linear term.

Under the assumption about noninformative censoring, the likelihood is given by

$$\begin{aligned} L &= \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot \exp\left(- \int_0^{t_i} \lambda_i(u) du \right) \\ &= \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot S_i(t_i)\,, \end{aligned} \tag{2.5}$$

inserting (2.3) and (2.4).

To obtain a unified and generic notation, we rewrite the observation model in general matrix notation. This is useful for defining priors in the next subsection and for developing posterior analysis in Section 2.3 as well as for describing results on propriety of posteriors for mixed models in Subsection 2.3.3.

Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_i, \ldots, \eta_n)'$ denote the predictor vector, where $\eta_i := \eta_i(t_i)$ is the value of predictor (2.4) at the observed lifetime $t_i, i = 1, \ldots, n$. Correspondingly, let $\boldsymbol{g}_j = (g_j(t_1), \ldots, g_j(t_n))'$ denote the vector of evaluations of the functions $g_j(t), j = 0, \ldots, p$, $\boldsymbol{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ the vector of evaluations of the functions $f_j(x_j), j = 1, \ldots, q$,

$\boldsymbol{f}_{spat} = (f_{spat}(s_1), \ldots, f_{spat}(s_n))'$ the vector of spatial effects, and $\boldsymbol{b} = (b_{g_1}, \ldots, b_{g_n})'$ the vector of uncorrelated random effects. Furthermore, let $\tilde{\boldsymbol{g}}_j = (g_j(t_1)z_{1j}, \ldots, g_j(t_n)z_{nj})', j = 1, \ldots, p$.

In the following, we can always express vectors $\boldsymbol{g}_0$, $\tilde{\boldsymbol{g}}_j$, $\boldsymbol{f}_j$, $\boldsymbol{f}_{spat}$ and $\boldsymbol{b}$ as the matrix product of an appropriately defined design matrix $\boldsymbol{Z}$, say, and a (possibly high–dimensional) vector $\boldsymbol{\beta}$ of parameters, e.g. $\tilde{\boldsymbol{g}}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, etc. Then, after reindexing, we can represent the predictor vector $\boldsymbol{\eta}$ in generic notation as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_0\boldsymbol{\beta}_0 + \ldots + \boldsymbol{Z}_m\boldsymbol{\beta}_m. \tag{2.6}$$

2.2.2 Priors for parameters and functions

The Bayesian model formulation is completed by assumptions about priors for parameters and functions. For fixed effect parameters $\boldsymbol{\gamma}$ in (2.6) we assume diffuse priors $p(\boldsymbol{\gamma}) \propto const.$ A weakly informative normal prior would be another choice. Uncorrelated random effects are assumed to be i.i.d. Gaussian, $b_g \sim N(0, \tau_b^2)$.

Priors for functions and spatial components are defined by a suitable design matrix $\boldsymbol{Z}_j$, $j = 0, \ldots, m$, and a prior for the parameter vector $\boldsymbol{\beta}_j$. The general form of a prior for $\boldsymbol{\beta}_j$ in (2.6) is

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right), \tag{2.7}$$

where $\boldsymbol{K}_j$ is a precision or penalty matrix of $\text{rank}(\boldsymbol{K}_j) = r_j$, shrinking parameters towards zero or penalizing too abrupt jumps between neighboring parameters. For P–splines and MRF priors, $\boldsymbol{K}_j$ will be rank deficient, i.e., $r_j < d_j = \dim(\boldsymbol{\beta}_j)$, and the prior is partially improper.

For *unknown functions* $f_j(x_j)$ or $g_j(t)$, we assume Bayesian P–spline priors as in Lang and Brezger (2004). Random walk priors, suggested in Fahrmeir and Lang (2001a), may be used as smoothness priors for the baseline effect and time–varying covariate effects in a piecewise exponential model, correspond to the special case of P–splines with degree zero. The basic idea of P–spline regression (Eilers and Marx 1996) is to approximate a function $f_j(x_j)$ as a linear combination of B–spline basis functions B_m, i.e.

$$f_j(x_j) = \sum_{m=1}^{d_j} \beta_{jm}B_m(x_j). \tag{2.8}$$

The basis functions B_m are B–splines of degree l defined over a grid of equally spaced knots $x_{min} = \xi_0 < \xi_1 < \ldots < \xi_s = x_{max}$, $d_j = l + s$. The number of knots is moderate, but not too small, to maintain flexibility, but smoothness of the function is encouraged by difference penalties for neighboring coefficients in the sequence $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_j})'$. The Bayesian analogue are first or second order random walk smoothness priors

$$\beta_{jm} = \beta_{j,m-1} + u_{jm} \qquad \text{or} \qquad \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm} \tag{2.9}$$

with i.i.d. Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $p(\beta_{j1}) \propto const$, or $p(\beta_{j1})$ and $p(\beta_{j2}) \propto const$, for initial values. A first order random walk penalizes abrupt jumps $\beta_{jm} - \beta_{j,m-1}$, and a second order random walk penalizes deviations from a linear trend. The amount of smoothness or penalization is controlled by the variance τ_j^2, which acts as a smoothness (hyper–)parameter, with hyperprior defined by (2.13). The joint prior of the regression parameters $\boldsymbol{\beta}_j$ is Gaussian and can be easily computed as a product of conditional densities defined by (2.9) as

$$\boldsymbol{\beta_j} \mid \tau_j^2 \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta_j}'\mathbf{K_j}\boldsymbol{\beta_j}\right), \tag{2.10}$$

which is the generic form (2.7).

The penalty matrix $\boldsymbol{K}_j$ is of the form $\boldsymbol{K}_j = \boldsymbol{D'D}$, where $\boldsymbol{D}$ is a first or second order difference matrix. For second order random walks, for example, $\boldsymbol{D}$ is given by

$$\boldsymbol{D}_{d_j-2\times d_j} = \begin{pmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{pmatrix}.$$

The matrix $\boldsymbol{K}_j$ has band structure which is very useful for computationally efficient MCMC updating schemes (compare Section 2.3). It has rank $r_j = d_j - 1$ and $r_j = d_j - 2$ for first and second order random walk priors, respectively. The $n \times d_j$ design matrix $\boldsymbol{Z}_j$ consists of the basis functions evaluated at the observations x_{ij}, i.e., $\boldsymbol{Z}_j(i, m) = B_m(x_{ij})$. Priors for the unknown functions $g_j(t)$ are defined in complete analogy as in (2.8) and (2.9). The design matrix for time–varying effect terms $\tilde{\boldsymbol{g}}_j$, $j = 1, \ldots, p$ is derived as $\boldsymbol{Z}_j(i, m) = z_{ij}B_m(x_{ij})$.

A common choice for approximating smooth curves are quadratic or cubic B–splines and a second order penalty. This specification is also preferred by Eilers and Marx (1996) and Lang and Brezger (2004) in order to obtain sufficiently smooth results. Computationally,

linear splines are simpler. The simplest choice are B–splines of degree zero, i.e. $B_m(x) \equiv 1$ over the m–th interval, and $B_m(x) \equiv 0$ elsewhere. Then the effect is approximated by a piecewise constant function, and the function values follow a random walk model as in Fahrmeir and Lang (2001a). This special choice, with time t as covariate, is the easiest way to smooth the baseline in the piecewise exponential model; moreover the integral in the likelihood (2.5) reduces to a sum, see the next section. With P–splines of higher degree, however, estimation of smooth baseline effects is improved in terms of MSEs, see Section 2.4. Another nice feature of cubic B–splines is that the well known smoothing splines appear as a special case with knots at every observation point.

For comparison we also consider an alternative parametric form. In the parametric case we choose a Weibull form for the baseline hazard (Banerjee, Wall and Carlin (2003)):

$$\lambda_0(t_i) = \alpha t_i^{\alpha-1}. \tag{2.11}$$

A $GA(0.01, 0.01)$ prior is assumed for α, so that α has a prior mean of 1 (corresponding to a constant hazard over time) and a large variance of 100.

For the *structured spatial effect* $f_{spat}(s)$ we assume either Markov random field (MRF) priors, two–dimensional tensor product P–spline priors, or Gaussian random field (GRF) priors, common in geostatistics (kriging).

In the case of *MRF priors* we define areas as neighbors if they share a common boundary and assume that the effect of an area s is conditionally Gaussian, with the mean of the effects of neighboring areas as expectation and a variance that is inverse proportional to the number of neighbors of area s. Setting $f_{spat}(s) := \beta_s^{spat}$ we have

$$\beta_s^{spat} \,|\, \beta_{s'}^{spat}, s' \neq s \sim N\left(\frac{1}{N_s}\sum_{s' \in \delta_s} \beta_{s'}^{spat}, \frac{\tau_{spat}^2}{N_s}\right),$$

where N_s is the number of neighbors of area s, and $s' \in \delta_s$ denotes that area s' is a neighbor of area s. The $n \times S$ design matrix $\boldsymbol{Z}_{spat}$ is now a 0/1 incidence matrix. Its value in the i–th row and s–th column is 1 if observation i is located in site or region s, and zero otherwise. The $S \times S$ penalty matrix $\boldsymbol{K}_{spat}$ has the form of an adjacency matrix with $\text{rank}(\boldsymbol{K}_{spat}) = r_{spat} = S - 1$. As for one–dimensional functions the amount of spatial smoothness is controlled by the variance τ_{spat}^2. A generalization to weighted means of neighboring areas is possible but not considered here.

Our second approach is based on *two–dimensional P–splines*, a rather parsimonious, but flexible method for modelling interactions between continuous covariates described in Lang and Brezger (2004) for Gaussian regression. Considering the x– and y–coordinates of the geographical center of each area, the spatial effect can be seen as an interaction between two continuous covariates x_s and y_s. The assumption is that the unknown structured spatial effect $f_{spat}(s)$ can be approximated by the tensor product of one–dimensional B–splines, i.e.

$$f_{spat}(s) = f_{spat}(x_s, y_s) = \sum_{m_1=1}^{d_{spat}} \sum_{m_2=1}^{d_{spat}} \beta_{m_1 m_2}^{spat} B_{spat,m_1}(x_s) B_{spat,m_2}(y_s).$$

Now the B–splines of degree l are defined over a regular two–dimensional grid of a moderate, but not too small number of equally spaced knots $\xi_{\rho\nu}$, $\rho, \nu = 1, \ldots, d_{spat} - 1$. We restrict ourselves to an equal number of knots for each direction. Knots are equally spaced within each direction, but the distance may differ between direction x_s and y_s. Priors for $\boldsymbol{\beta}^{spat} = (\beta_{11}^{spat}, \ldots, \beta_{1d_{spat}}^{spat}, \ldots, \beta_{d_{spat}1}^{spat}, \ldots, \beta_{d_{spat}d_{spat}}^{spat})'$ are based on MRF priors for spatial data on a regular lattice (see e.g. Besag and Kooperberg, 1995). Since there is no natural ordering of parameters, priors have to be defined by specifying the conditional distributions of $\beta_{m_1 m_2}^{spat}$ given neighboring parameters and the variance component τ_{spat}^2. The most commonly used prior specification based on the four nearest neighbors can be defined by

$$\beta_{m_1 m_2}^{spat} | \cdot \sim N\left(\frac{1}{4}(\beta_{m_1-1,m_2}^{spat} + \beta_{m_1+1,m_2}^{spat} + \beta_{m_1,m_2-1}^{spat} + \beta_{m_1,m_2+1}^{spat}), \frac{\tau_{spat}^2}{4}\right) \tag{2.12}$$

for $m_1, m_2 = 2, \ldots, d_{spat} - 1$ and appropriate changes for corners and edges. For example, for the upper left corner we obtain $\beta_{11}^{spat} | \cdot \sim N(\frac{1}{2}(\beta_{12}^{spat} + \beta_{21}^{spat}), \frac{\tau_{spat}^2}{2})$. For the left edge, we get $\beta_{1m_2}^{spat} | \cdot \sim N(\frac{1}{3}(\beta_{1,m_2+1}^{spat} + \beta_{1,m_2-1}^{spat} + \beta_{2,m_2}^{spat}), \frac{\tau_{spat}^2}{3})$.

The prior (2.12) is a direct generalization of a first order random walk in one dimension. Its conditional mean can be interpreted as a least squares locally linear fit at knot position $\xi_{\rho\nu}$ given the neighboring parameters. More details can be found in Lang and Brezger (2004). Defining $\boldsymbol{K}_{spat} = \boldsymbol{D}_1' \boldsymbol{D}_1 + \boldsymbol{D}_2' \boldsymbol{D}_2$, where $\boldsymbol{D}_1 = \boldsymbol{I} \otimes \boldsymbol{D}$ and $\boldsymbol{D}_2 = \boldsymbol{D} \otimes \boldsymbol{I}$, the prior can again be expressed in the general form (2.7). Here, $\boldsymbol{D}$ is the first order difference matrix known from the one–dimensional case, and $\boldsymbol{D}_1' \boldsymbol{D}_1$ corresponds to the penalization in the direction of x and $\boldsymbol{D}_2' \boldsymbol{D}_2$ corresponds to the penalization in the direction of y.

Our third option are *stationary Gaussian random field* (GRF) priors, which can be seen as two–dimensional surface smoothers based on special basis functions, e.g. radial

basis functions, and have been used by Kammann and Wand (2003) for modelling the spatial component in Gaussian regression models. The spatial component $f_{spat}(s) = \beta_s^{spat}$ is assumed to follow a zero mean stationary Gaussian random field $\{\beta_s^{spat} : s \in \mathbb{R}^2\}$ with variance τ_{spat}^2 and use an isotropic covariance function $\text{cov}(\beta_s^{spat}, \beta_{s'}^{spat}) = C(\|s - s'\|)$ as proposed by Stein (1999). For a finite array $s \in \{1, \ldots, S\}$ of sites as in our application the prior can be brought in the general form

$$\boldsymbol{\beta}^{spat} \mid \tau_{spat}^2 \propto \exp\left(-\frac{1}{2\tau_{spat}^2}(\boldsymbol{\beta}^{spat})'\boldsymbol{K}_{spat}\boldsymbol{\beta}^{spat}\right)$$

with penalty matrix $\boldsymbol{K}_{spat} = \boldsymbol{C}^{-1}$, where $C[k,l] = C(\|s_k - s_l\|), 1 \leq k, l \leq n$, and design matrix $\boldsymbol{Z}_{spat} = \boldsymbol{C}$.

For the covariance function $C(r)$ we follow again recommendations of Stein (1999) and use the Matérn family of covariance functions $C(r; \rho, \nu)$. For the special case $\nu = 1.5$ for the smoothness parameter the covariance functions simplify to

$$C(r; \rho, \nu) = \tau_{spat}^2(1 + |r|/\rho)e^{-|r|/\rho},$$

which is the simplest member of the Matérn family that results in differentiable surface estimates as Kammann and Wand (2003) point out. The parameter ρ controls how fast covariances die out with increasing distance r. We choose ρ according to the rule

$$\hat{\rho} = \max_{k,l} \|s_k - s_l\|/c$$

to ensure scale invariability. This rule proved to work well in practice. The constant c is chosen in such a way that $C(c)$ is small, e.g. $C(c) = 0.001$.

While the dimension of the penalty matrix in a MRF equals the number of different regions S, in a GRF the dimension corresponds to the number of distinct locations which is likely to be close to or equal to the sample size. To reduce this computational burden Kammann and Wand (2003) propose low–rank kriging to approximate stationary Gaussian random fields. Therefore they define a 'representative' subset of knots $\mathcal{D} = \{\kappa_1, \ldots, \kappa_M\}$ of the set of distinct locations by applying a space filling algorithm (compare Johnson et al. (1990) and Nychka and Saltzman (1998) for details). Based on these knots, we obtain the approximation $f_{spat}(s) = \boldsymbol{z}'_{spat}(s)\boldsymbol{\beta}^{spat}$ with the M–dimensional design vector $\boldsymbol{z}_{spat}(s) = (C(\|s - \kappa_1\|), \ldots, C(\|s - \kappa_M\|))'$ and penalty matrix $\boldsymbol{K}_{spat} = \tilde{\boldsymbol{C}}$ and $\tilde{C}[k,l] = C(\|\kappa_k -$

$\kappa_l\|$). The number of knots controls the trade–off between accuracy of the approximation and numerical simplification. Details on GRF and (low–rank) kriging can be found in Kammann and Wand (2003), Kneib and Fahrmeir (2005) or Kneib (2006).

Still a serious drawback of this approach is the computational effort involved. Since the penalty matrix $\boldsymbol{K}_{spat}$ has no longer band structure it is not possible to employ efficient matrix algorithms for sparse matrices like the Cholesky decomposition in order to draw samples from our multivariate normal proposal density and to compute the determinant of the precision matrix, which is needed to calculate the acceptance probability of the MH–step in every iteration (compare Section 2.3). For the application in Section 2.5, e.g., this means that the required CPU time multiplies approximately by the factor 20, even if we use low–rank kriging with a moderate number of 100 knots. It depends mainly on the data at hand, which of the different approaches leads to the best fit. For data observed on a discrete lattice or on the level of geographical regions as in our application, MRFs seem to be most adequate, while surface smoothers as 2d P–splines or kriging may be more natural in situations where exact locations are available. In general, MRFs exhibit more rough results, while 2d P–splines produce the smoothest estimates. GRFs also tend to give quite smooth curves.

A decision between MRFs and 2d P–splines or GRFs, respectively, may depend on ones beliefs about the characteristics of the corresponding effect. If an effect is supposed to vary smoothly (e.g. in case it is influenced by temperature or atmospheric pressure) surface estimators can be expected to be the better choice. If, on the other hand, an effect is likely to be induced, for example, by characteristics of geographical or political units, which may depend on neighbors, but may quite as well be rather heterogenic, then a MRF should be preferred. However, in applications sometimes surface estimators outperform MRFs even for discrete data (and vice versa). This may be due to some regions having few neighbors or observations, since a more smooth surface estimator is able to reduce the bias for such regions.

In real data applications we do not know how much of the spatial variation is explained by structured, spatially correlated effects and how much by unstructured, uncorrelated effects. Therefore we may fit an additional (unstructured) area–specific random effect. We recommend to interpret only the sum of the two effects, since identifiability is weak in that case.

We routinely assign inverse Gamma priors $IG(a_j; b_j)$

$$p(\tau_j^2) \propto \frac{1}{(\tau_j^2)^{a_j+1}} \exp\left(-\frac{b_j}{\tau_j^2}\right) \tag{2.13}$$

to all variances. They are proper for $a_j > 0$, $b_j > 0$, and we use $a_j = b_j = 0.001$ as a standard choice for a weakly informative prior. From our experience results are rather insensitive to the choice of $a_j > 0$ and $b_j > 0$ for moderate to large data sets and the posterior distribution is proper in any case under some regularity assumptions (see Subsection 2.3.3 and Hennerfeind, Brezger and Fahrmeir (2005) for a proof). However, since the limiting case, when a_j and b_j are zero, leads to an improper posterior distribution, we present a sensitivity analysis in Section 2.4 and compare the results to those we obtained with a uniform prior for the standard deviation τ_j, as proposed in Gelman (2004). Note that uniform priors are a special (improper) case of the prior (2.13) with $a_j = -0.5$, $b_j = 0$, still leading to proper posteriors under regularity assumptions.

The Bayesian model specification is completed by assuming that all priors for parameters are conditionally independent, and that all priors are mutually independent.

2.3 Markov chain Monte Carlo inference

In what follows, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', ..., \boldsymbol{\beta}_m')'$ denote the vector of all regression coefficients in the generic notation (2.6), $\boldsymbol{\gamma}$ the vector of fixed effects, and $\boldsymbol{\tau}^2 = (\tau_0^2, ..., \tau_m^2)$ the vector of all variance components. Full Bayesian inference is based on the entire posterior distribution

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2 \mid data) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)\, p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2).$$

Due to the (conditional) independence assumptions, the joint prior factorizes into

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2) = \left\{ \prod_{j=0}^{m} p(\boldsymbol{\beta}_j \mid \tau_j^2) p(\tau_j^2) \right\} p(\boldsymbol{\gamma}),$$

where the last factor can be omitted for diffuse fixed effect priors.

The likelihood $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)$ is given by inserting (2.3), (2.4) into (2.5), but the integral requires integration over all terms depending on survival time t, i.e. terms of the form

$$I_i = \int_0^{t_i} \exp\left(g_0(u) + \sum_{j=1}^{p} g_j(u) z_{ij} \right) du, \tag{2.14}$$

where $g_j(t) = \sum \beta_{jm} B_m(t)$. Apart from B–splines $B_m(t)$ of degree zero, i.e. random walk models, and linear B–splines, these integrals are not available in closed form. The first case leads to the piecewise exponential model: The time axis is divided into a grid

$$0 = \xi_0 < \xi_1 < ... < \xi_{t-1} < \xi_t < ... < \xi_s = t_{max},$$

and $g_j(t)$ is assumed to be a piecewise constant function, i.e.

$$g_j(t) = \beta_{jt}$$

in time interval $(\xi_{t-1}, \xi_t]$, $t = 1, ..., s$. In this case, the integral reduces to a sum, and, after some calculations, the log–likelihood contribution of observation i in the interval $(\xi_{t-1}, \xi_t]$ can be expressed as

$$l_{it} = y_{it}\eta_{it} - \exp\left(\Delta_{it} + \eta_{it}\right)$$

where

$$y_{it} = \begin{cases} 1 & t_i \in (\xi_{t-1}, \xi_t], \delta_i = 1 \\ 0 & \text{else.} \end{cases}$$

$$\Delta'_{it} = \begin{cases} \xi_t - \xi_{t-1}, & \xi_t < t_i \\ t_i - \xi_{t-1}, & \xi_{t-1} < t_i \leq \xi_t \\ 0, & \xi_{t-1} \geq t_i \end{cases}$$

$$\Delta_{it} = \log \Delta'_{it} \quad (\Delta_{it} = -\infty \text{ if } \Delta'_{it} = 0).$$

This likelihood is proportional to a Poisson–likelihood, with the predictor η_{it} containing an additional offset term Δ_{it}, see Fahrmeir and Tutz (2001, Section 9.1) or Ibrahim et al. (2001, Section 3.1) for details.

For linear B–splines, the integrals can still be solved analytically, but expressions are rather messy and the computational effort is quite high, see Cai et al. (2002, Appendix). Following their suggestion, we use simple numerical integration in form of the trapezoidal rule for linear B–splines as well as for the commonly used cubic B–splines, where analytical integration is not possible anyway.

2.3.1 Updating full conditionals

Full Bayesian inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters, given the rest of the data. For updating the parameter vectors $\boldsymbol{\beta}_j$, which correspond to the time–independent functions $f_j(x_j)$, as well as spatial effects $\boldsymbol{\beta}^{spat}$, fixed effects $\boldsymbol{\gamma}$ and random effects $\boldsymbol{b}$, we use a slightly modified version of an MH–algorithm based on iteratively weighted least squares (IWLS) proposals, developed for fixed and random effects by Gamerman (1997) and adapted to generalized additive mixed models in Brezger and Lang (2006).

Suppose we want to update $\boldsymbol{\beta}_j$, with current value $\boldsymbol{\beta}_j^c$ of the chain. Then a new value $\boldsymbol{\beta}_j^p$ is proposed by drawing a random vector from a (high–dimensional) multivariate Gaussian proposal distribution $q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)$, which is obtained from a quadratic approximation of the log–likelihood by a second order Taylor expansion with respect to $\boldsymbol{\beta}_j^c$, in analogy to IWLS iterations in generalized linear models. More precisely, the goal is to approximate the posterior by a Gaussian distribution, obtained by accomplishing *one* IWLS step in every iteration of the sampler. Then, random samples have to be drawn from a high dimensional multivariate Gaussian distribution with precision matrix and mean

$$\boldsymbol{P}_j = \boldsymbol{Z}_j' \boldsymbol{W}(\boldsymbol{\beta}_j^c) \boldsymbol{Z}_j + \frac{1}{\tau_j^2} \boldsymbol{K}_j, \quad \boldsymbol{m}_j = \boldsymbol{P}_j^{-1} \boldsymbol{Z}_j' \boldsymbol{W}(\boldsymbol{\beta}_j^c)(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\eta}}).$$

Here, $\tilde{\eta}_i = \eta_i(t_i) - f_j(x_{ij})$, $\boldsymbol{W}(\boldsymbol{\beta}_j^c) = diag(w_1, \ldots, w_n)$ is the weight matrix for IWLS with weights calculated from the current state $\boldsymbol{\beta}_j^c$ as follows

$$w_i = \exp\left(\sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i' \boldsymbol{\gamma} + b_{g_i} \right) \cdot I_i.$$

Concisely written we get $w_i = \int_0^{t_i} \lambda_i(u) du = \Lambda_i(t_i)$, which is the cumulative hazard rate. The working observations $\tilde{y}_i$ are given by

$$\tilde{y}_i = \eta_i(t_i) + \frac{\delta_i}{w_i} - 1.$$

See Appendix A1 for a detailed derivation of those quantities. The proposed vector $\boldsymbol{\beta}_j^p$ is accepted as the new state of the chain with probability

$$\alpha(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p) = \min\left(1, \frac{p(\boldsymbol{\beta}_j^p \mid \cdot) q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^c)}{p(\boldsymbol{\beta}_j^c \mid \cdot) q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)}\right)$$

where $p(\boldsymbol{\beta}_j \mid \cdot)$ is the full conditional for $\boldsymbol{\beta}_j$ (i.e. the conditional distribution of $\boldsymbol{\beta}_j$ given all other parameters and the data).

For a fast implementation, we use the fact that the precision matrices of the Gaussian proposal distributions are banded for MRS and 2d P–spline models, so that Cholesky decompositions can be performed efficiently. Now, random numbers from the high dimensional proposal distributions can be efficiently drawn using an algorithm by Rue (2001). The acceptance probability $\alpha(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)$ involves the determinant $det(\boldsymbol{P}_j)$ of the proposal density, since $\boldsymbol{P}_j$ depends on $\boldsymbol{\beta}_j^c$. Fortunately, this quantity is obtained as a simple by–product of the Cholesky decomposition with negligible computational effort.

Note, however, that this is not the case for GRF models. Here, $\boldsymbol{K}_j$ is not banded, and thus $\boldsymbol{P}_j$ is not banded, either. Therefore, drawing from the proposal and evaluating its determinant is much more demanding in terms of CPU time.

For the parameters $\boldsymbol{\beta}_j$ corresponding to the functions $g_0(t), ..., g_p(t)$ depending on time t, the IWLS–MH algorithm requires considerably more computational effort, because the integrals in the log–likelihood as well as first and second derivatives are involved now. Therefore, we adopt a computationally faster MH–algorithm based on conditional prior proposals, although IWLS–MH has better mixing properties. This algorithm was first developed by Knorr–Held (1999) for state space models and extended for generalized additive mixed models in Fahrmeir and Lang (2001a). It requires only evaluation of the log–likelihood, not of derivatives. However, draws are not performed for the entire vector $\boldsymbol{\beta}_j$, but iteratively for blocks of subvectors, see Fahrmeir and Lang (2001a) for details. In the case of the parametric Weibull prior (2.11) a new value α^p is proposed by drawing from a Gamma distribution $GA(\alpha^c \cdot b_w, b_w)$, with α^c denoting the current value of the chain and b_w being tuned automatically during the burn–in period.

The full conditionals for the variance parameters τ_j^2 are (proper) inverse Gamma with parameters

$$a_j' = a_j + \frac{1}{2} r_j \quad \text{and} \quad b_j' = b_j + \frac{1}{2} \boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j,$$

including the case $a_j = -0.5$, $b_j = 0$ of uniform priors on τ_j. Updating can be done by simple Gibbs steps, drawing random numbers directly from the inverse Gamma densities. In complete analogy, the full conditional for a variance component τ_{spat}^2 of the spatial effect and τ_b^2 of a random intercept or slope is again an inverse gamma distribution, and updating

is straightforward.

2.3.2 Model choice

Bayesian model choice is an area of ongoing research with several competing proposals ranging from (modified) Bayes factors to posterior predictive loss approaches (Gelfand and Gosh 1998). We routinely use the Deviance Information Criterion (DIC) developed in Spiegelhalter, Best, Carlin and van der Linde (2002). It is given as

$$DIC = D(\overline{\boldsymbol{\theta}}) + 2p_D = \overline{D(\boldsymbol{\theta})} + p_D,$$

where $\boldsymbol{\theta}$ is the vector of parameters, $D(\overline{\boldsymbol{\theta}})$ is the deviance of the model evaluated at the posterior mean estimate $\overline{\boldsymbol{\theta}}$, $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance and $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is the effective number of parameters. Since it is at least unclear, how the saturated model should be defined in the case of survival data when the baseline hazard and other nonparametric functions are parameters of interest, we use the unstandardized deviance $D(\boldsymbol{\theta}) = -2\cdot$log–likelihood instead of the saturated deviance. Banerjee and Carlin (2004, Section 4) provide good arguments why the DIC is a reasonable criterion in connection with censored survival data.

2.3.3 Propriety of posteriors in geoadditive survival models

Consider a geoadditive survival model with predictor in generic form (2.6), where $\boldsymbol{Z}_0\boldsymbol{\beta}_0$ corresponds to an effect with prior (2.7) for $\boldsymbol{\beta}_0$ such that $\dim(\boldsymbol{\beta}_0) = d_0 \geq d_j$, $\text{rank}(\boldsymbol{K}_0) = r_0 \geq r_j$, $j = 1, \ldots, m$. This assumption is usually fulfilled for the spatial component or for a high–dimensional vector of group–specific uncorrelated random effects.

Denote by $\boldsymbol{\eta}_u$, $\boldsymbol{V}_u$, $\boldsymbol{Z}_u = (\boldsymbol{Z}_{1u}, ..., \boldsymbol{Z}_{mu})$, $\boldsymbol{Z}_{0u}$ the (sub–)predictor and sub–design matrices corresponding to uncensored observations. Assume that the following conditions hold:

(C1) $\text{rank}(\boldsymbol{V}_u) = \text{rank}(\boldsymbol{V}) = p = \dim(\boldsymbol{\gamma})$,
$\text{rank}(\boldsymbol{Z}_{ju}) = \text{rank}(\boldsymbol{Z}_j) = d_j = \dim(\boldsymbol{\beta}_j), \quad j = 0, ..., m$
$\text{rank}(\boldsymbol{Z}_u'\boldsymbol{R}\boldsymbol{Z}_u + \boldsymbol{K}) = d$
where $d = d_1 + ... + d_m, \quad \boldsymbol{K} = \text{diag}(\boldsymbol{K}_1, ..., \boldsymbol{K}_m), \quad \boldsymbol{R} = \boldsymbol{I} - \boldsymbol{V}_u(\boldsymbol{V}_u'\boldsymbol{V}_u)^{-1}\boldsymbol{V}_u'$

(C2) The priors $p(\tau_j^2)$, $j = 1, ..., m$, are proper, and $\int p(\tau_0^2)\tau_0^{-(r_0-p-(d-r)-(d_0-r_0))}d\tau_0^2 < \infty$, where $r = r_1 + ... + r_m$.

Theorem 1: If conditions (C1), (C2) hold then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$, where $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_m^2)'$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)'$, is proper.

The following corollary is easier to check.

Corollary 1: Assume proper inverse Gamma priors $Ga(a_j, b_j)$ for τ_j^2 with $j = 0, ..., m$ and $r_0 + 2a_0 - p - (d - r) - (d_0 - r_0) > 0$.
If condition (C1) holds, then the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\tau}^2, \tau_0^2 \mid \boldsymbol{y})$ is proper.

Proofs are based on Sun et al. (1999), and are outlined in Hennerfeind, Brezger and Fahrmeir (2005).

Remark 1: Condition (C1) is equivalent to rank($\boldsymbol{Z}_{0u}$)=d_0 and

$$\text{rank}\begin{pmatrix} \boldsymbol{V}_u'\boldsymbol{V}_u & \boldsymbol{V}_u'\boldsymbol{Z}_u \\ \boldsymbol{Z}_u'\boldsymbol{V}_u & \boldsymbol{Z}_u'\boldsymbol{Z}_u + \boldsymbol{K} \end{pmatrix} = p + d$$

Remark 2: Under additional assumptions, proper posteriors may also be obtained for $a_j < 0$, $b_j = 0$, e.g. for the uniform prior on τ_j. A rigorous proof could be based on a generalization of Sun, Tsutakawa and He (2001).

Remark 3: Informally expressed condition (C1) is fulfilled if the information provided by uncensored observations is sufficient to support the estimation of each single parameter in $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0)$. Considering the cases where $\boldsymbol{\beta}_0$ denotes a spatial effect or a group–specific random effect, rank($\boldsymbol{Z}_{0u}$)=d_0 is fulfilled if the data set comprises at least one uncensored observation per area or group, respectively. Condition (C2) is fulfilled if the inverse Gamma priors for τ_j^2 are proper and $r_0 + 2a_0$ is greater than the number of improper priors for parameters in $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\beta}_0)$.
Note that these conditions are sufficient conditions and may be weakened in some places.

2.4 Simulation Study

Performance was investigated through simulation studies. In particular we were interested in the following questions: How influential is the choice of MRF versus smoother spatial priors and the choice of a piecewise exponential model (P–spline of degree zero) versus a cubic P–spline model for the baseline hazard rate? How sensitive are the results with respect to the hyperparameters for the variance parameters? And how does a P–spline model perform compared to a Weibull–model in cases where the Weibull assumption is indeed true and in cases where it is not true, respectively.

Simulation Setup I

Life times T_i, $i = 1, ..., 1236$, were generated from Weibull distributions according to the hazard model

$$\lambda_i(t) = \lambda_0(t) \exp(f_1(x_i) + f_{spat}(s_i) + \gamma v_i) = \exp(\log(3t^2) + \sin(x_i) + \sin(x_{s_i} \cdot y_{s_i}) - 0.3v_i), \tag{2.15}$$

with Weibull baseline hazard rate $\lambda_0(t) = 3t^2$, a binary covariate v, with the v_is randomly drawn from a Bernoulli $B(1; 0.5)$ distribution, and a continuous covariate x, with the x_is randomly drawn from a uniform $U[-3, 3]$ distribution. The spatial covariate s_i denotes one of the $s = 1, \ldots, S = 309$ counties of the former Federal Republic of Germany and x_{s_i} and y_{s_i} are the centered coordinates of the geographic center of county s_i. We simulated four observations per county, resulting in $309 \times 4 = 1236$ observations in total. The censoring was done as follows: We randomly selected a certain proportion of observations ($\approx 17\%$ and $\approx 50\%$, respectively) that were to be censored. Censoring variables C_i for these selected observations were then generated as i.i.d. draws from corresponding uniform $U[0, T_i]$ distributions.

Keeping the predictor fixed, 100 replications $\{T_i^{(r)}, C_i^{(r)}, i = 1, ..., 1236\}$ respectively $\{(t_i^{(r)}, \delta_i^{(r)}), i = 1, ..., 1236\}$, $r = 1, ..., 100$ of censored survival times were generated.

To investigate the first question, the log–baseline hazard $g_0(t)$ was modelled by second order random walk priors, corresponding to a piecewise exponential model, and alternatively as a cubic P–spline with 20 knots. The spatial effect was modelled as a MRF and alternatively as a two–dimensional cubic P–spline with 12×12 knots. Simulations with

GRF priors are not feasible due to much higher computation times, but the general message will be the same. A cubic P–spline prior with 20 knots was chosen for $f_1(x) = \sin(x)$ in each case. Hyperparameters of inverse Gamma priors for variance components were set to $a = 0.001$, $b = 0.001$, the standard choice.

For each replication $r = 1, ..., 100$, we computed the mean square errors

$$MSE_r(g_0) = \frac{1}{1236} \sum_{i=1}^{1236} (\widehat{g}_0^{(r)}(t_i^{(r)}) - g_0(t_i^{(r)}))^2,$$

for the log–baseline hazard $g_0(t)$,

$$MSE_r(f_1) = \frac{1}{1236} \sum_{i=1}^{1236} (\widehat{f}_1^{(r)}(x_i) - f_1(x_i))^2$$

for $f_1(x) = \sin(x)$, and

$$MSE_r(f_{spat}) = \frac{1}{1236} \sum_{i=1}^{1236} (\widehat{f}_{spat}^{(r)}(s_i) - f_{spat}(s_i))^2$$

for the spatial effect $f_{spat}(s) = \sin(x_c \cdot y_c)$, where $\widehat{g}_0^{(r)}$ and $\widehat{f}_k^{(r)}$, $k = 1, spat$, are posterior mean estimates for simulation run r. The $MSE(\gamma)$ was computed in the usual way.

Results: MRF versus 2d P–spline, p.e.m. versus P–spline model

Figures 2.1 and 2.2 display boxplots of the logarithmic MSEs ($\log(MSE_r)$, $r = 1 \ldots, 100$). As was to be expected, the P–spline model has smaller MSEs for g_0 when compared to the piecewise exponential model. Interestingly, the MSEs for $\gamma = -0.3$, $f_1(x)$ and $f_{spat}(s)$ are more or less unaffected by the choice of the smoothness prior for the log–baseline $g_0(t)$. Estimated functions of replication r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$, for $g_0(t)$, $f_1(x)$ and $f_{spat}(s)$ are displayed in Figures 2.3–2.5 (for the censoring level of 17%). Regarding the two different levels of censoring Figures 2.1 and 2.2 show that the estimation of the log–baseline effect is the effect that is strongest influenced by the level of censoring. While increasing the censoring level from 17% to 50% leads to an approximately 2.75 times larger MSE for $g_0(t)$ the MSE for $f_{spat}(s)$ is only increased by a factor of ca. 1.35. Due to the simulation scheme, where the spatial effect is defined as a smooth 2–dimensional function of the spatial coordinates ($f_{spat}(s_i) = \sin(x_{s_i} \cdot y_{s_i})$), the MSEs for f_{spat} are smaller when a 2–dimensional P–spline prior is assumed instead of a MRF prior.

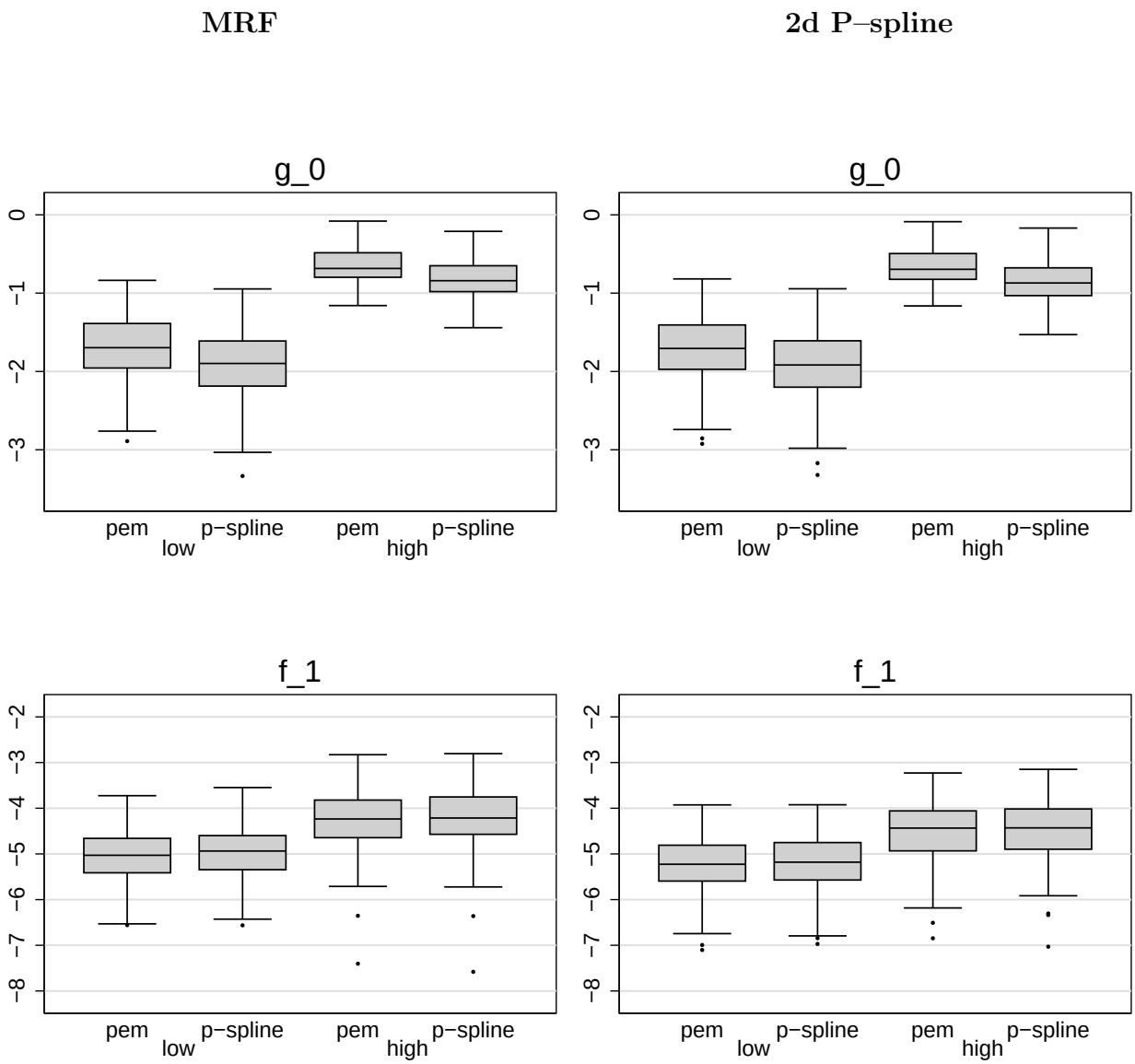


Figure 2.1: Simulation: model comparison via boxplots of log–MSEs for data sets with low (ca. 17%) and high censoring level (ca. 50%), for estimations with MRF priors (left panel) and 2–d P–spline priors (right panel) for the spatial effect each with cubic P–spline priors for the log–baseline and p.e.m.s, respectively.

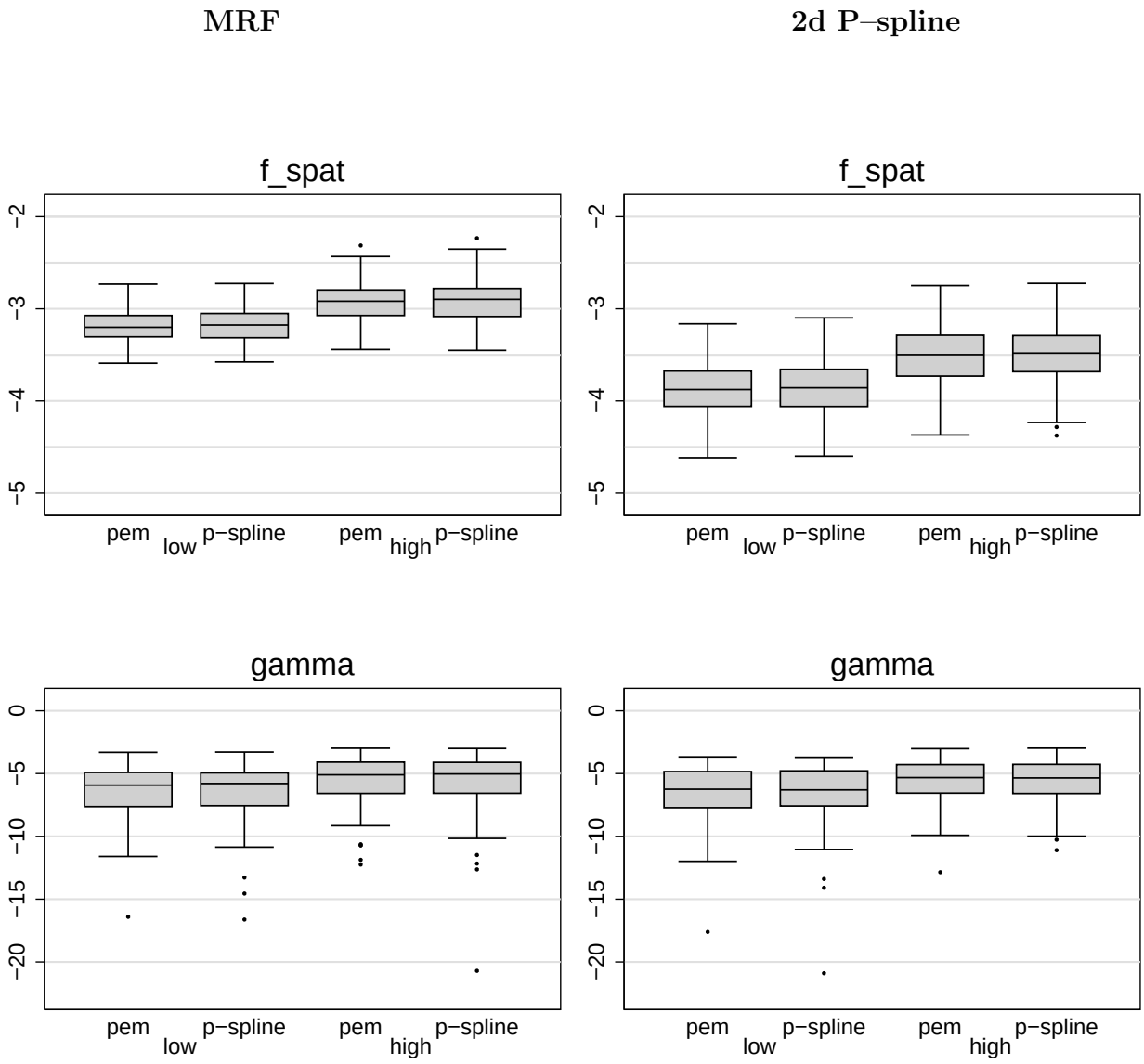
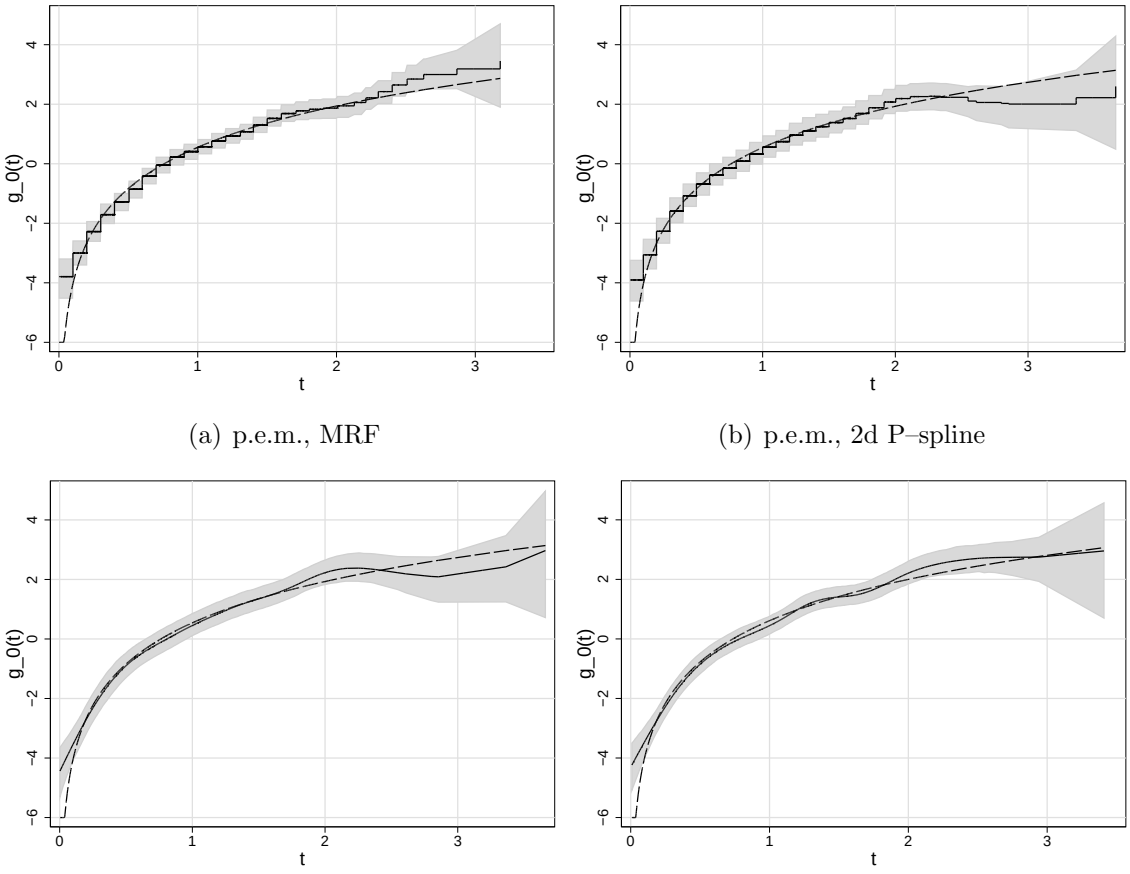


Figure 2.2: Simulation: model comparison via boxplots of log–MSEs for data sets with low (ca. 17%) and high censoring level (ca. 50%), for estimations with MRF priors (left panel) and 2–d P–spline priors (right panel) for the spatial effect each with cubic P–spline priors for the log–baseline and p.e.m.s, respectively.



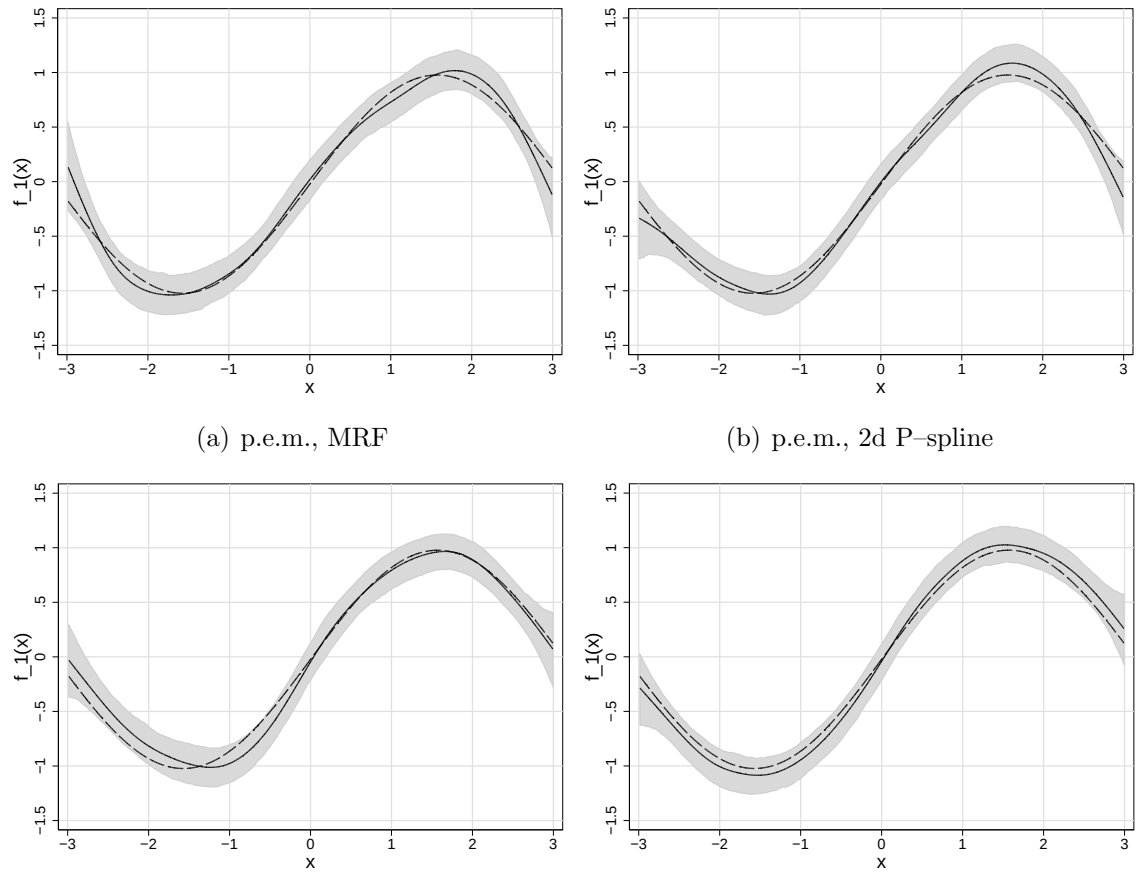
(a) p.e.m., MRF

(b) p.e.m., 2d P–spline

(c) P–spline model, MRF

(d) P–spline model, 2d P–spline

Figure 2.3: (log–)Baseline effects $g_0(t)$ for the various model specifications; displayed are posterior mean estimates and 95% credible intervals of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ (solid line and grey shaded area), and the true (log–)baseline effect (dashed line). a) p.e.m., MRF, r=11, MSE=0.183 b) p.e.m., 2d P–spline, r=51, MSE=0.181 c) P–spline model, MRF, r=51, MSE=0.148 d) P–spline model, 2d P–spline, r=7, MSE=0.145



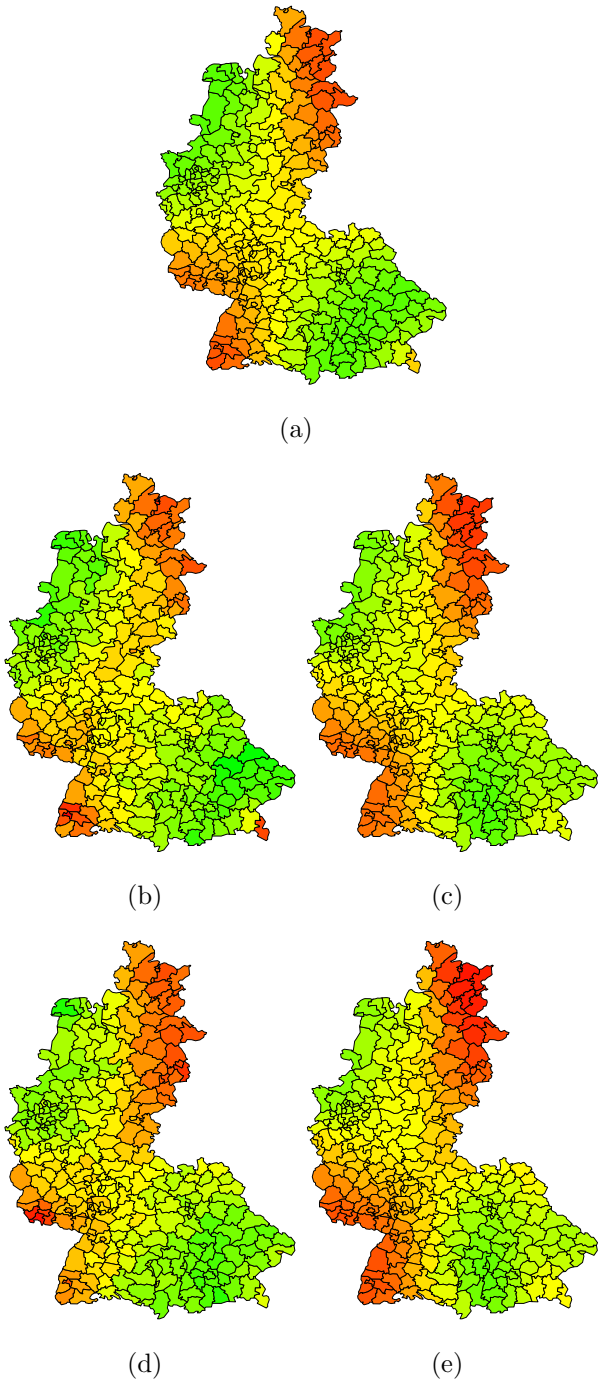
(a) p.e.m., MRF

(b) p.e.m., 2d P–spline

(c) P–spline model, MRF

(d) P–spline model, 2d P–spline

Figure 2.4: Nonparametric effects $f_1(x)$ for the various model specifications; displayed are posterior mean estimates and 95% credible intervals of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ (solid line and grey shaded area), and the true function (dashed line). a) p.e.m., MRF, r=53, MSE=0.0064 b) p.e.m., 2d P–spline, r=36, MSE=0.0053 c) P–spline model, MRF, r=67, MSE=0.0068 d) P–spline model, 2d P–spline, r=19, MSE=0.0056



(a)

(b) (c)

(d) (e)

Figure 2.5: Spatial effects for the various model specifications with effects ranging from -1.3 to 1.65; displayed are posterior mean estimates of run r, with r chosen such that MSE_r is the median of $MSE_1, \ldots, MSE_{100}$ a) true function b) p.e.m., MRF, r=41, MSE=0.041 c) p.e.m., 2d P–spline, r=13, MSE=0.021 d) P–spline model, MRF, r=12, MSE=0.042 e) P–spline model, 2d P–spline, r=13, MSE=0.021

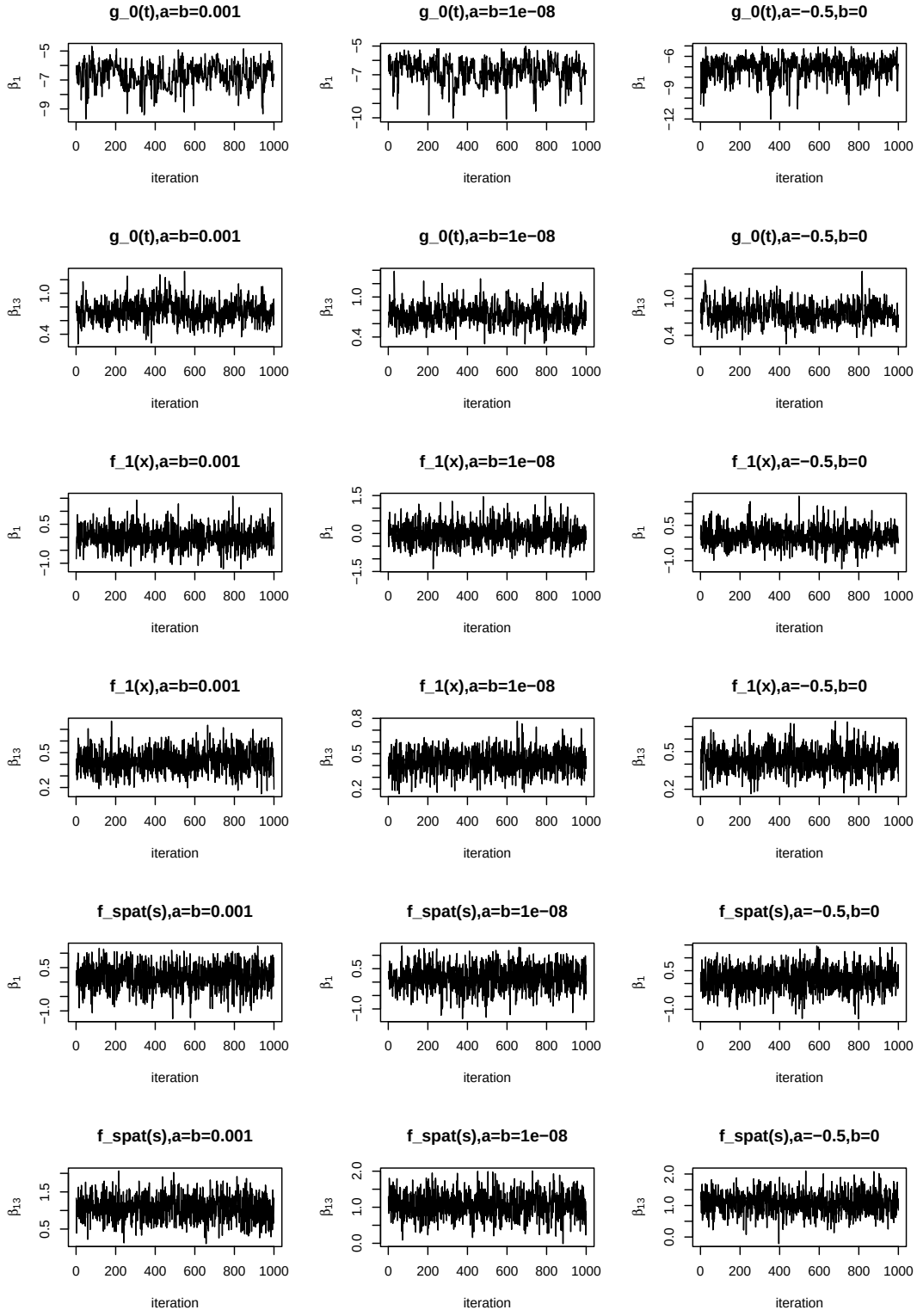


Figure 2.6: Selected sampling paths of run $r = 1$ for parameters $\beta_{j,1}$ and $\beta_{j,13}$, $j = 0, 1, spat$ and different choices for the parameters a and b of the $IG(a; b)$ hyperpriors.

Results: influence of hyperparameters

To investigate the second question, in particular to analyze the behavior of the Markov chains when a and b approach zero (and the prior for the hyperparameters thus approaches the $IG(0;0)$ distribution, that leads to an improper posterior), we focus on the P–spline model with MRF–prior and a censoring level of 17% and alternatively set $a = b = 0.0001$, $a = b = 0.00001$ and $a = b = 0.00000001$. We additionally run the simulation study with $a = -0.5$, $b = 0$, i.e. uniform priors on the standard deviations τ_0, τ_1 and τ_{spat} that act as smoothing parameters for the log–baseline, the nonlinear effect of x and the spatial effect, respectively. Selected sampling paths of run $r = 1$ are exemplarily shown in Figure 2.6. We did not face problems with mixing or convergence of Markov chains with any of these prior distributions. An exception are the first one or two parameters of the baseline effect, i.e. $\beta_{0,1}$ and $\beta_{0,2}$, corresponding to the effect of small times t, where the mixing properties are not always optimal. This can be explained by the very steep increase of the 'true' log–baseline, reaching to minus infinity as t approaches zero whereas it is quite flat elsewhere. In this situation a global variance might not be an ideal choice. Another point may be the usage of conditional prior proposals that usually lead to poorer mixing properties than IWLS–proposals do. Figure 2.7 displays kernel density estimators of the posterior mean of the variance parameters based on $\widehat{\tau}_j^{2\,(r)}, r = 1, \ldots, 100$ for $j = 0, 1, spat$. Obviously the different choices of the hyperparameters a and b of the inverse Gamma prior do not seem to have much effect, whereas the uniform prior on the standard deviations tends to result in somewhat larger estimates for the variance parameters and thus in less smooth effects. The posterior distribution of the variance parameter of the spatial effect is quite robust, as the full conditional is dominated by the values of r_j =rank($\boldsymbol{K}_j$) and $\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j$ at this. Figure 2.8 displays boxplots of the logarithmic MSEs ($\log(MSE_r), r = 1, \ldots, 100$), that are computed as before. While the MSEs are quite unaffected by the choice of the hyperparameters $a = b$ of the inverse Gamma prior, the uniform prior results in a slightly smaller MSE for $g_0(t)$, but a slightly bigger MSE for $f_1(x)$. Altogether we come to the conclusion that (at least with this model) it does not seem to be crucial, which one of these weakly informative priors is assumed for the variance parameters.

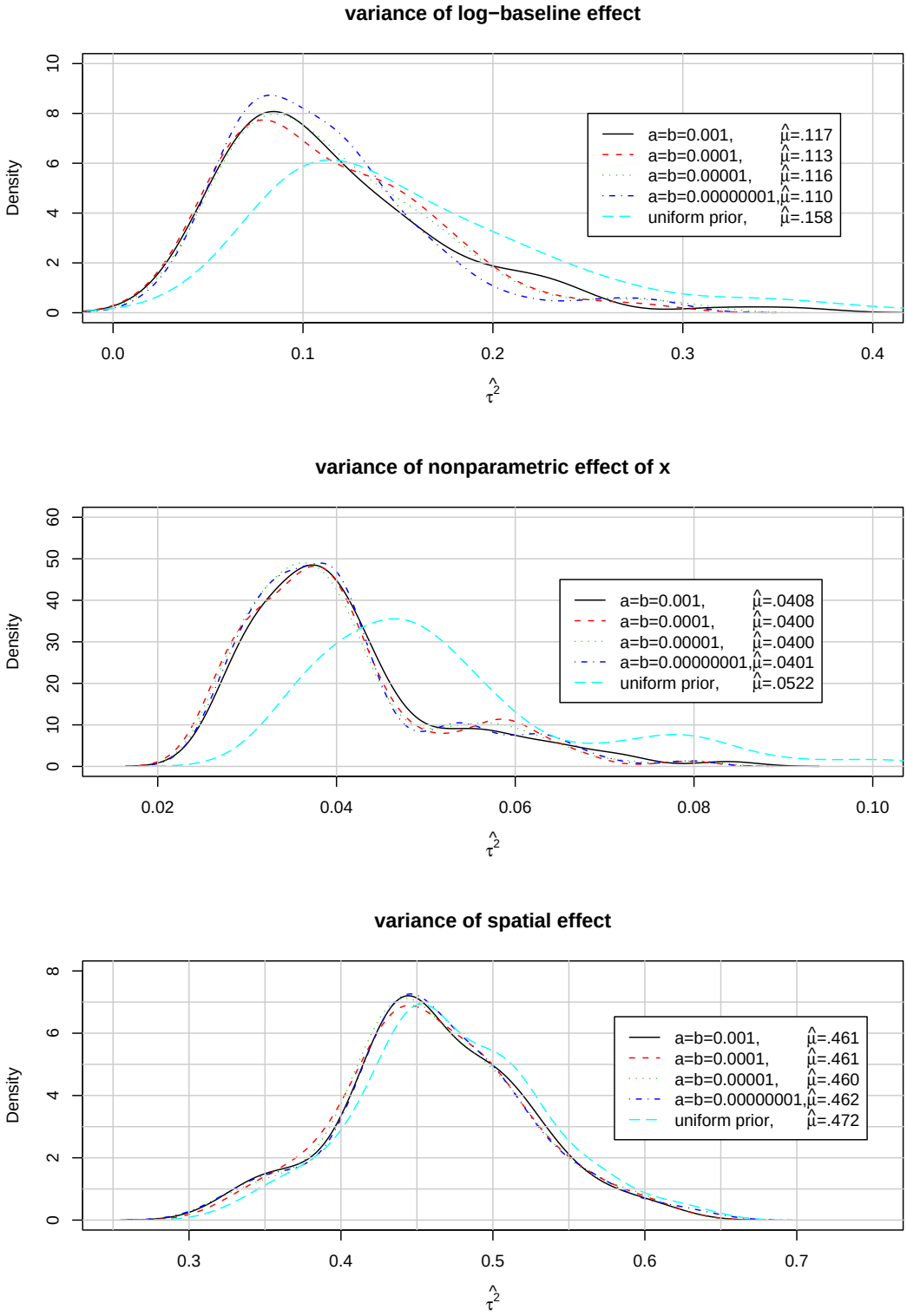


Figure 2.7: Kernel density estimates based on $\widehat{\tau}_j^{2(r)}$, $r = 1, \ldots, 100$ for $j = 0, 1$ and *spat*, respectively. $\hat{\mu}$ denotes the mean estimated smoothing parameter.

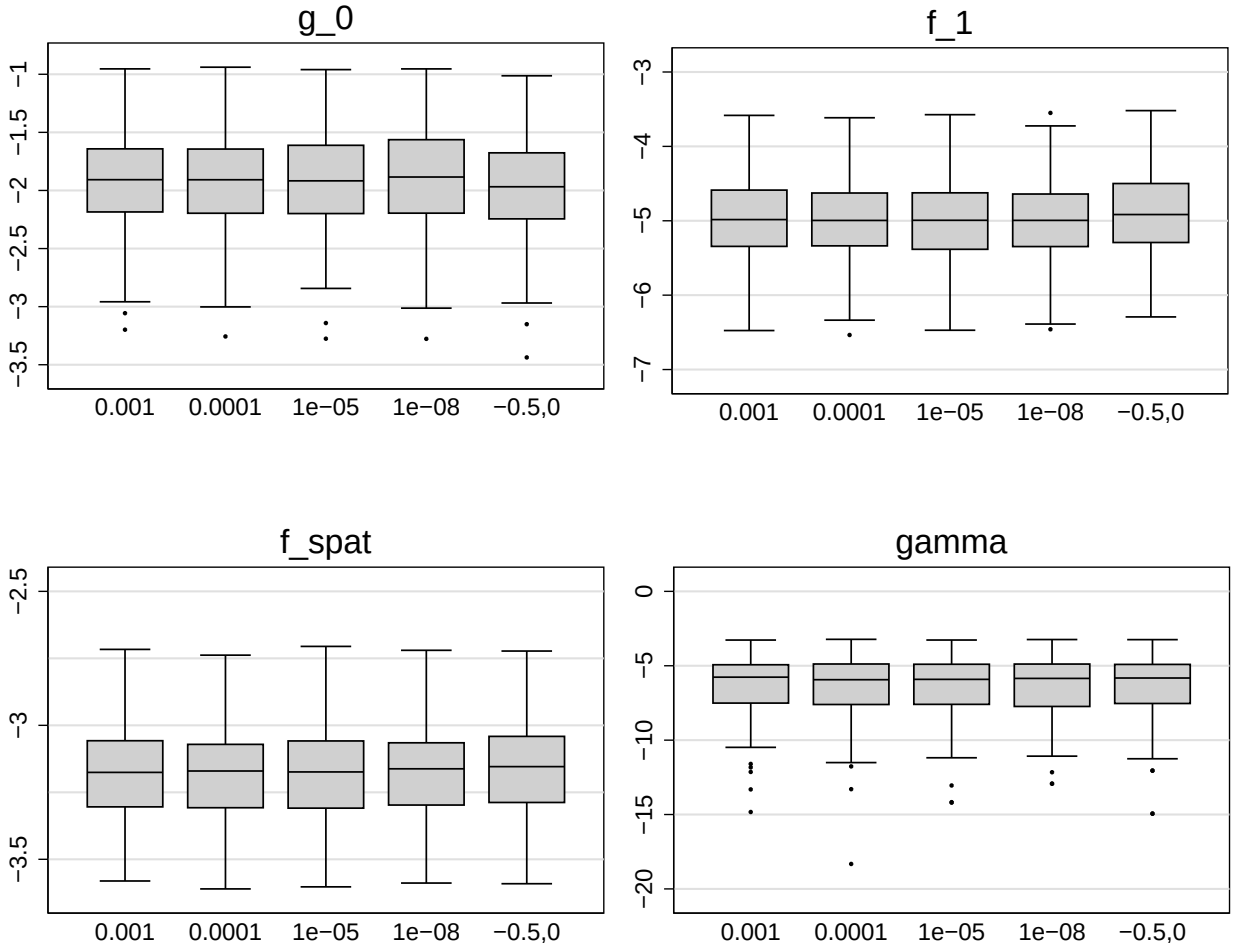
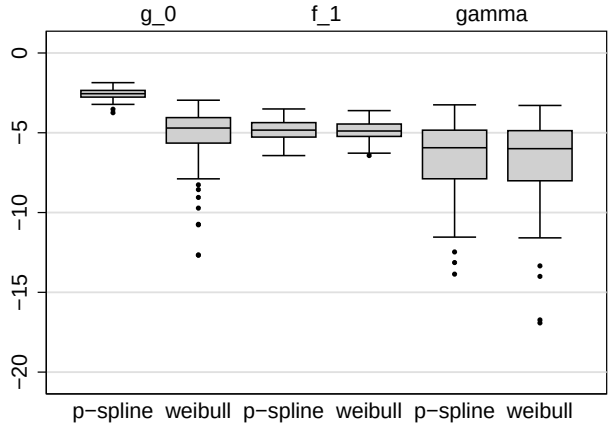
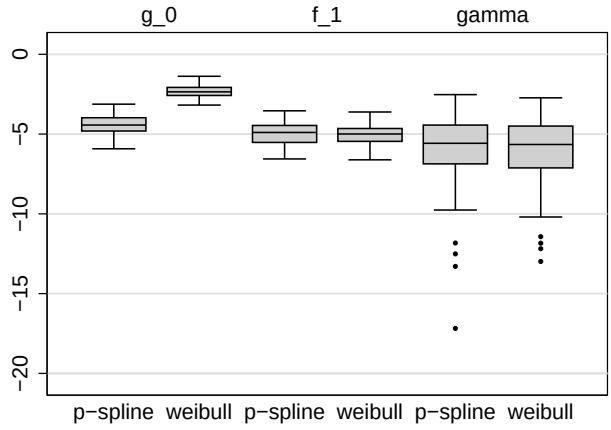


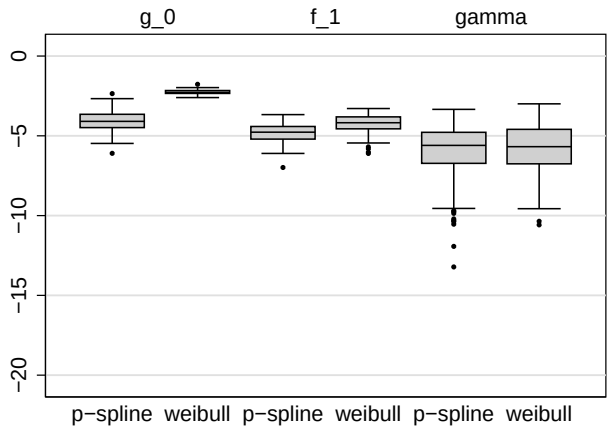
Figure 2.8: Simulation: Comparison via boxplots of $\log MSE$s for estimations with $IG(a;b)$ hyperpriors for $a = b = 0.001$, $a = b = 0.0001$, $a = b = 1e-05$, $a = b = 1e-08$ and $a = -0.5, b = 0$ respectively.



(a) Weibull hazard with $\alpha = 2$



(b) linear baseline hazard



(c) bathtub–shaped baseline hazard

Figure 2.9: Simulation: model comparison via boxplots of log–MSEs.

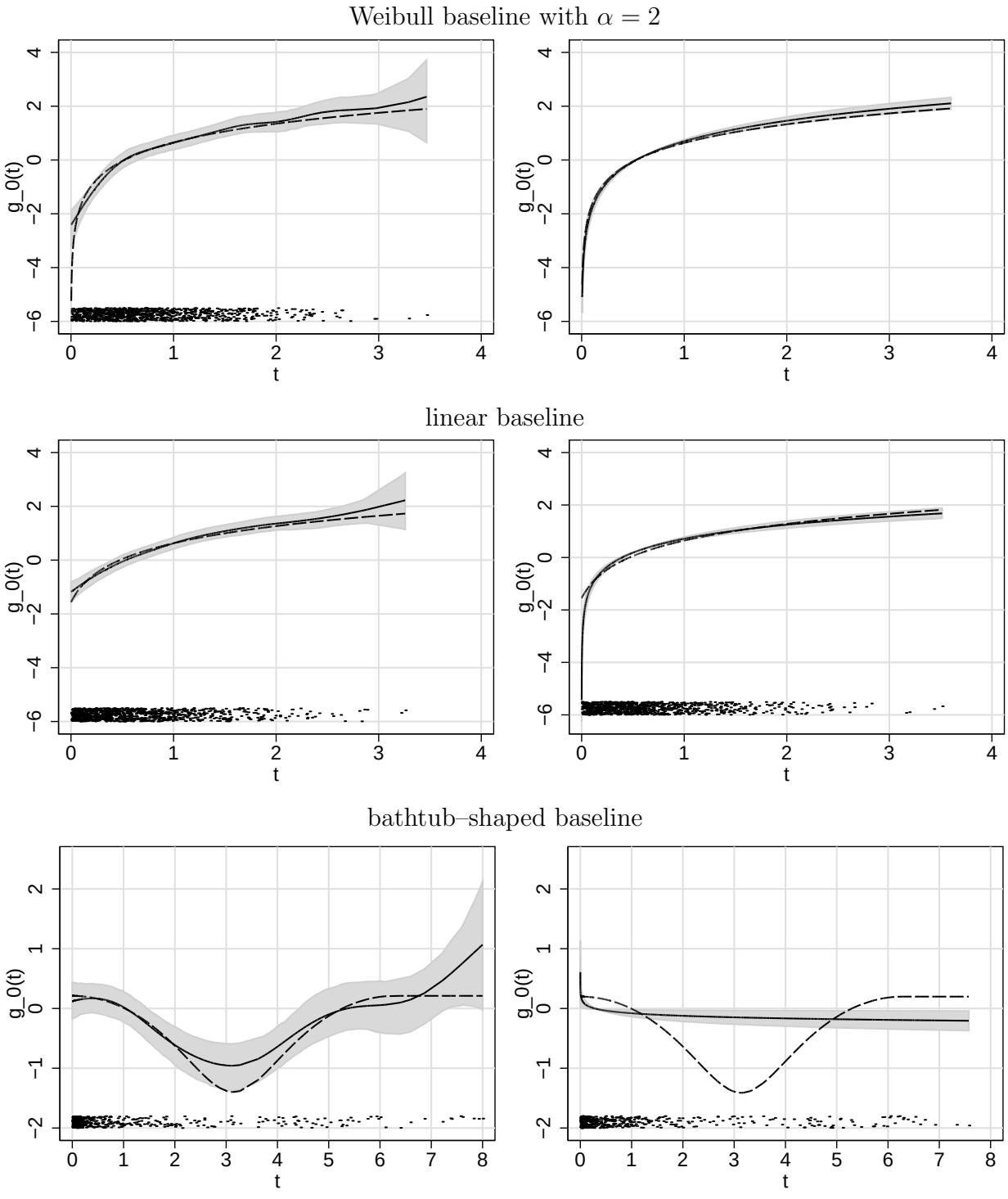


Figure 2.10: Simulation: posterior mean estimates together with 95% credible intervals for the log–baseline effects (solid line and grey shaded area) of run r, with r chosen such that $MSE_r(g_0(t))$ is the median of $MSE_1(g_0(t)), \ldots, MSE_{100}(g_0(t))$ and the true log–baseline (dashed line).

Simulation setup II

To shed some light on the question what is lost and gained by applying a P–spline model instead of a Weibull–model in cases where the true baseline hazard is Weibull shaped and not Weibull shaped, respectively, lifetimes $T_i, i = 1, \ldots, 1000$ were generated according to the hazard models

$$\lambda_i(t) = \lambda_0(t) \cdot \exp(\gamma v_i + f_1(x_i)) = \lambda_0(t) \cdot \exp\left(0.3v_i + \sin(x_i)\right),$$

with v and x denoting a binary and a continuous covariate as in (2.15). The baseline hazard $\lambda_0(t)$ is once chosen to be a Weibull baseline with shape parameter $\alpha = 2$, i.e.

$$\lambda_0(t) = \alpha \cdot t^{\alpha-1} = 2 \cdot t$$

once to be a likewise monotonic, linear, but non–Weibull hazard rate given by

$$\lambda_0(t) = 0.25 + 2 \cdot t$$

and once to be bathtub–shaped according to the following equation

$$\lambda_0(t) = \begin{cases} 0.75 \cdot (\cos(t) + 1.5), & t \leq 2\pi \\ 0.75 \cdot (1 + 1.5), & t > 2\pi \end{cases}$$

i.e. the baseline hazard is assumed to be initially high, to decrease after some time and increase again later on until the time $t = 2\pi$, from where on the hazard stays constant. Such bathtub–shaped hazard rates appear quite frequently in survival time studies and can not be comprehended with parametric approaches like the Weibull model.

While lifetimes T_i may be generated straightforward by drawing random numbers from according Weibull distributions in the former case, i.e. Weibull distributions with shape parameter α and scale parameter $(1/\exp(0.3v_i + \sin(x_i)))^{\frac{1}{\alpha}}$, a more elaborate simulation technique is required for the second and third choice for the baseline hazard. Two possibilities are for example given by the *thinning method*, a kind of rejection algorithm, where a dominating hazard rate is required (see Lewis and Shedler, 1979) and the *inversion method*, that is applicable in cases where the cumulative baseline hazard $\Lambda_0(t)$ and its inverse $\Lambda_0^{-1}(t)$ may at least be evaluated numerically (see e.g. Devroye, 1986 and Bender et al., 2005).

With our simulation we used the inversion method, where lifetimes T_i are generated as follows

$$T_i = \Lambda_0^{-1}\left(-\log(U_i)\exp(-0.3v_i - \sin(x_i))\right),$$

with U_i randomly drawn from a standard uniform distribution, i.e. $U_i \sim U[0,1]$. As before the censoring was done in a second step: We randomly selected a proportion of ca. 30% of observations that were to be censored. Censoring variables C_i for these selected observations were then generated as i.i.d. draws from corresponding uniform $U[0, T_i]$ distributions.

Again, keeping the predictor fixed, 100 replications $\{T_i^{(r)}, C_i^{(r)}, i = 1, ..., 1000\}$ respectively $\{(t_i^{(r)}, \delta_i^{(r)}),\ i = 1, ..., 1000\}$, $r = 1, ..., 100$ of censored survival times were generated with each of the three baseline hazards. Estimation was done with a cubic P–spline prior with 20 equidistant knots for the log–baseline effect $g_0(t)$ and with a Weibull model with a $GA(0.01; 0.01)$–prior on α as described in (2.11), respectively. A cubic P–spline prior with 20 knots was also assumed for the nonparametric effect of x and a diffuse prior was assumed for the fixed effect of v.

Results: P–spline model versus Weibull model

The MSEs were calculated as described above and Figure 2.9 displays boxplots of the logarithmic MSEs. As before with the comparison between P–spline models and p.e.m.s the MSEs of f_1 and γ corresponding to the nonparametric effect of the continuous covariate x and the fixed effect of the binary covariate v are barely affected by the choice of the prior for the baseline hazard. However, as was to be expected the MSE of the log–baseline effect $g_0(t)$ is smaller with the Weibull model in the case where the true baseline hazard has a Weibull structure (Figure 2.9a)), but the MSE is smaller with the P–spline model in case of $\lambda_0(t) = 0.25 + 2 \cdot t$ and the bathtub–shaped baseline hazard (Figure 2.9 b) and c)). Estimated log–baseline hazards of replication r, with r chosen such that $MSE_r(g_0(t))$ is the median of $MSE_1(g_0(t)), \ldots, MSE_{100}(g_0(t))$ are displayed in Figure 2.10. Concerning the simulation where the true baseline hazard has an exact Weibull structure, Figure 2.10 reveals that the Weibull model yields very good results for $\hat{g_0}(t)$. The cubic P–spline model also yields quite satisfactory results for the most part, but does not reflect the steep increase at the beginning (note that $g_0(0) = -\infty$ with Weibull hazard rates), which is the main reason for the discrepancy in MSE when compared to the Weibull model. Bayesian P–splines

with locally adaptive variances as developed in the context of generalized additive models in Lang and Brezger (2004) for functions with changing curvature (and highly oscillating functions) might provide a solution. Here the global variance parameters τ_j^2 in equation (2.9) are replaced by local variances τ_j^2/δ_{jm}, where the weights δ_{jm} are additional hyperparameters. Regarding the linear non–Weibull shaped baseline hazard ($\lambda_0(t) = 0.25 + 2 \cdot t$), results are contrary to those the simulation with Weibull structure yields. While the cubic P–spline model reflects the shape of the log–baseline satisfactorily, the true shape can by definition not be reflected correctly by the Weibull model, which heavily underestimates the log–baseline for very small values of t. The bathtub–shaped baseline is again reflected rather sufficiently by the P–spline model, but also this structure can not at all be retrieved with a Weibull model, which suggests a largely flat baseline hazard with an increased risk for very small values of t.

Conclusion

Altogether we come to the conclusion that the choice of the prior for the baseline hazard does not seem to be very important in cases where the only interest is to gain information on time–constant effects of covariates. However, in cases where the baseline hazard is of interest, we do not recommend to use a Weibull model since it is quite restrictive and only outperforms the P–spline model in cases where the baseline hazard actually is Weibull shaped. A flexible estimation becomes even more important in cases where time–varying effects of covariates are to be examined.

2.5 Application

To illustrate our methods we present three applications to complex data sets with slightly different requirements, with spatial information being given for the last data set only. The examples arise from different fields, namely from the fields of credit risk, insurance and biometrics.
Unless otherwise noted:

- time–varying as well as nonparametric effects of continuous covariates are modelled

by cubic P–splines with 20 knots,

- diffuse priors are assumed for fixed effect parameters,
- MRF priors are assumed for structured spatial effects,
- unstructured (uncorrelated) random effects are assumed to be i.i.d. Gaussian with mean zero,
- the parameters of $IG(a;b)$ hyperpriors for variance parameters are set to $a = b = 0.001$.

2.5.1 Overdraft credit risk

Our first application is on overdraft credit data from a Swiss bank. The data comprises information on the monthly account movements of 2891 debtors (companies) with date of first borrowing within the observation period, i.e. between June, 1999 and June, 2004. Besides the date when the credit was first granted (`date`), the observed credit duration t_i and the external covariate "rate of unemployment" (`unempl`), the following continuous, monthly varying covariates are given:

- `tvb`: transaction volume (receipts of payments, credit items etc.)/borrowings (averaged over 5 months, restricted to values between 0 and 20)
- `ndtl`: number of days with transgressed credit limit

The question of interest is to analyze the influence of these covariates on the risk of default, i.e. the risk that a debtor cannot repay his credit. There are only 69 (ca. 2.4%) defaults observed, whereas 2109 credits (ca. 73%) are still existing at the end of the observation period in June 2004. The remaining 713 credits (ca. 25%) are either repayed between June 99 and June 2004 or sold to another bank. It is unknown which case is true and we consider these 713 observations as well as the 2109 observations, where credits were still existing in June 2004 as right censored and define the indicator of non–censoring of debtor $i, i = 1, \ldots, 3068$ as follows:

$$\delta_i = \begin{cases} 1 & \text{default of debtor } i \\ 0 & \text{else} \end{cases}$$

Since all covariates are continuous, the hazard rate $\lambda_i(\mathtt{t})$ is modelled as follows

$$\lambda_i(\mathtt{t}) = \exp(g_0(\mathtt{t}) + f_{tvb}(\mathtt{tvb}_i) + f_{unempl}(\mathtt{unempl}_i) + f_{ndtl}(\mathtt{ndtl}_i) + f_{date}(\mathtt{date}_i)),$$

where the log–baseline effect $g_0(\mathtt{t})$ as well as all other effects are modelled by P–splines. Since the distribution of the number of days with transgressed credit limits is quite left–skew, the position of the 20 knots was chosen according to quantiles in the case of f_{ndtl}. The estimated effects are shown in Figure 2.11. It can be concluded that the log–baseline hazard rate is highest in the first months after a credit is first granted and is decreasing almost linearly with time. The effect of the covariate transaction volume/borrowings is u–shaped, meaning that low values near zero and high values near 20 lead to an increased default risk. While it is quite perspicuous that debtors with a transaction volume that is very low compared to their borrowings and a consequently low value of `tvb` are at a high risk, it is less clear why the risk is increased with debtors with high values of `tvb`. A possible explanation is that this effect is caused by debtors that already have a bad credit history and thus a reduced credit limit resulting in low borrowings and a therefor high value of `tvb`. The effect of the rate of unemployment is roughly linearly increasing, meaning that the risk of default is rising with an increasing rate of unemployment. Furthermore we observe that credits that were first granted at the beginning of the observation period are at a higher risk than credits granted in later years. This might either be due to an improved credit risk management and a more restrictive credit policy or to the economic cycle. As was to be expected the risk is rising with the number of days with transgressed credit limit, where the increase is steepest in the beginning.

2.5.2 Long term care insurance

As a further illustration, we analyze data on survival time after entering long term care insurance (LTC) from a German private insurance company. The data was recorded between April 1, 1995, when compulsory LTC insurance was introduced by the German government, and December 31, 1998. It contains information on 5603 recipients of benefits from LTC insurance. This data set has already been analyzed by Czado and Rudolph (2002), and more details on the data set are given there. In a first step, they analyzed the data with a conventional Cox model with fixed effects of covariates and products of covariates. After

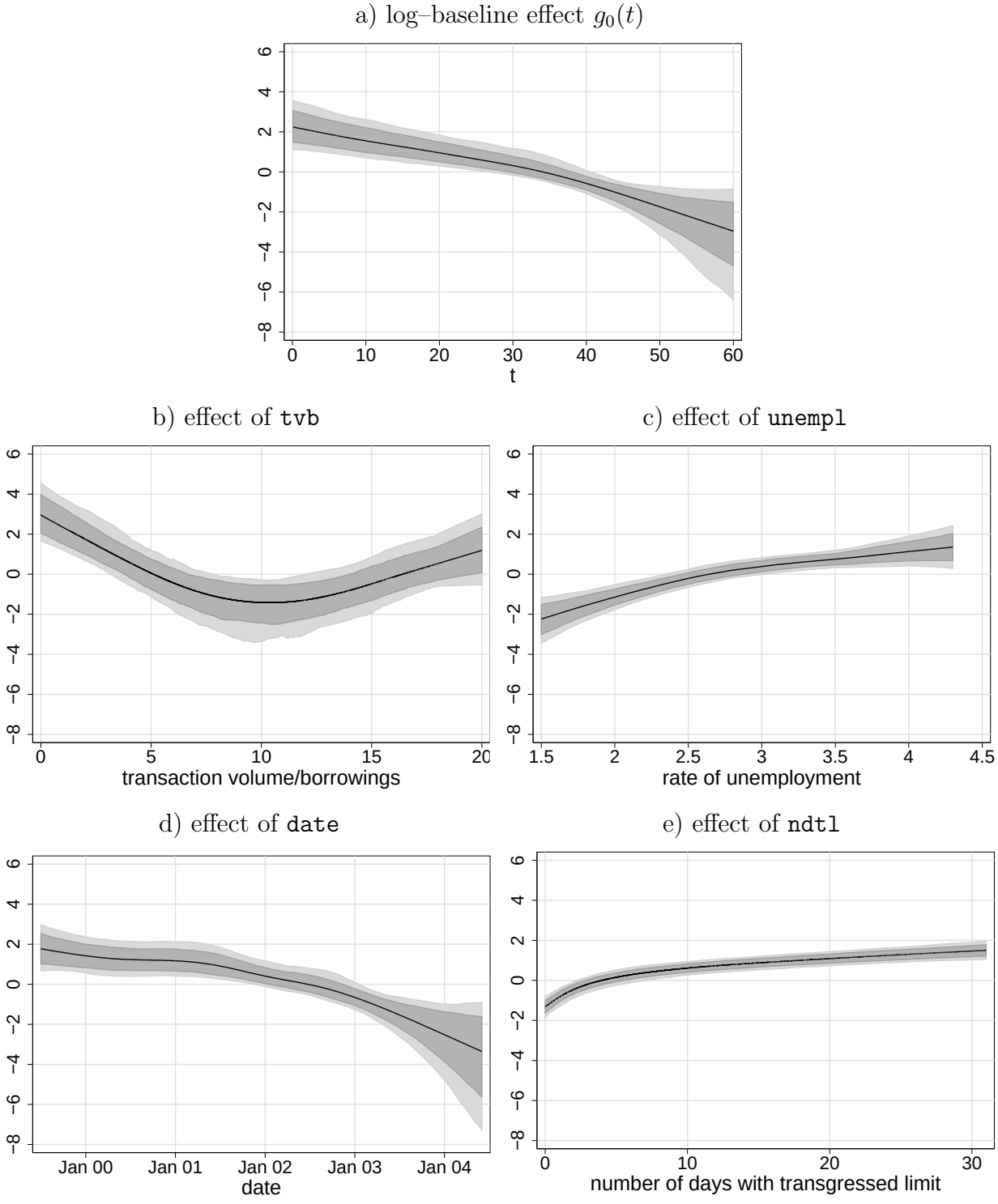


Figure 2.11: Overdraft credit risk: posterior means together with pointwise 80% and 95% credible intervals

careful model diagnosis (and the inclusion of two dummy–coded variables for modelling deviations from linear effects) they extended their Cox model to a model with time–varying effects, which were modelled through 0–1 step functions. Our analysis is based on their final model – CR model for short. The covariates that are included are `sex` of claimant (1=female, 0=male) and the following time–dependent covariates:

`age(t)` = age of claimant when a state transition occurs at time `t`,

$$\texttt{nh(t)} = \begin{cases} 1 & \text{care in a nursing home at time } \texttt{t} \\ 0 & \text{care at home at time } \texttt{t} \end{cases}$$

$$\texttt{leveli(t)} = \begin{cases} 1 & \text{care at level i at time } \texttt{t},\ i = 2,3 \\ 0 & \text{else,} \end{cases}$$

with `level1(t)` as the reference category. Transition times between care levels and care required (at home or in a nursing home) and dates of death or right–censoring are given in day units. The three levels of care (and benefits) are defined as follows:

- Level 1: Care level 1 is reserved to persons in considerable need of LTC. They would at least once a day require help for at least two activities in areas of personal hygiene, nutrition or mobility. They would also need help several times a week with household chores. Care level 1 can only be granted if the applicant needs help for at least 90 minutes a day, including 45 minutes of basic care.

- Level 2: Care level 2 is ear–marked for persons in severe need of LTC. They need help at least three times a day with personal hygiene, eating or getting around. In addition, they need help several days a week in housekeeping. In care level 2 the time of help required must be at least 3 hours a day, of which 2 hours must be needed for basic care.

- Level 3: Care level 3 is reserved to persons in extreme need of care. They require round–the–clock help every day, as well as household help several times a week. Care level 3 demands that the applicant needs at least 5 hours help a day, including a minimum of 4 hours of basic care.

	BPH		CR	
	mean	std. dev.	mean	std. dev.
`level2`	0.81	0.07	0.81	0.07
`level3`	1.71	0.07	1.75	0.07
`sex` · `nh`	-0.42	0.10	-0.44	0.10
`level2` · `nh`	-0.33	0.15	-0.33	0.14
`level3` · `nh`	-0.66	0.14	-0.67	0.14

Table 2.1: LTC data: posterior means and standard deviations for fixed effects

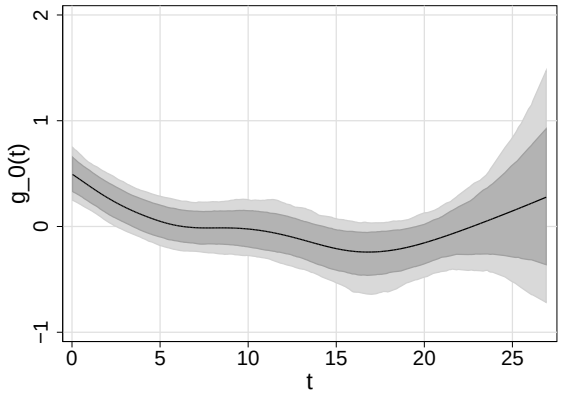
The exact date of first receipt of benefits is given for every claimant. In most cases this date is prior to April 1, 1995, i.e. most of the observations are left truncated (ca. 70%). Furthermore, about 60% of the observations are right censored.

As a start we apply a Bayesian multiplicative proportional hazard (BPH for short) model $\lambda(t) = \exp(\eta(t))$ with predictor

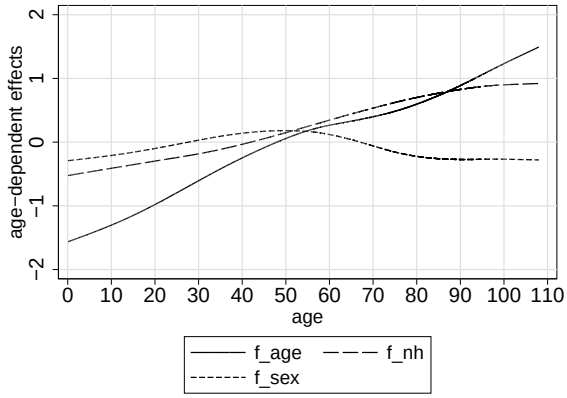
$$\begin{aligned}\eta(t) \;=\; & g_0(t) + \gamma_{l2} \cdot \texttt{level2(t)} + \gamma_{l3} \cdot \texttt{level3(t)} + \\ & f_{age}(\texttt{age}) + f_{sex}(\texttt{age}) \cdot \texttt{sex} + f_{nh}(\texttt{age}) \cdot \texttt{nh(t)} + \\ & \gamma_{snh} \cdot (\texttt{sex} \cdot \texttt{nh(t)}) + \gamma_{l2nh} \cdot (\texttt{level2(t)} \cdot \texttt{nh(t)}) + \gamma_{l3nh} \cdot (\texttt{level3(t)} \cdot \texttt{nh(t)})\end{aligned}$$

Results for fixed effects are given in Table 2.1. The log–baseline $g_0(\texttt{t})$ and the main effect of `age` as well as the age–dependent effects $f_{sex}(\texttt{age})$ of `sex` and $f_{nh}(\texttt{age})$ of `nh` are displayed in Figure 2.12. Our results for the age–independent effects are highly comparable to those of the CR model. Although differences concerning modelling of the age–dependent functions lead to slightly differing results, the age–dependent effects are as well quite similar for the most part.

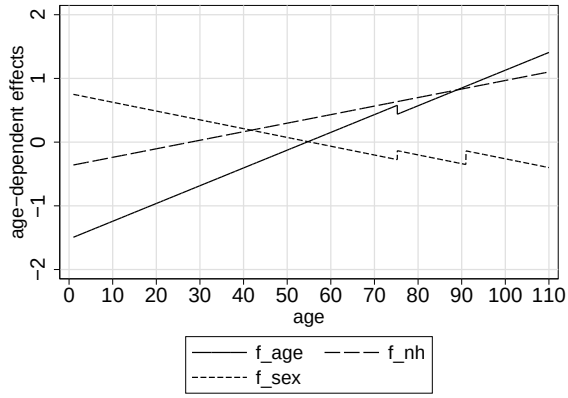
We conclude from our results that the hazard rate is increased in the first three to four years of receiving benefits. The hazard is also increased for claimants receiving more care (Level 2 and Level 3), but these main effects decrease by ca. 40% for claimants living in a nursing home. Furthermore, care in a nursing home seems to decrease the hazard rate with young people, but increase the hazard risk in the case of elderly claimants. Compared to male claimants, female claimants living in a nursing home have a lower hazard. The main effect of `age` increases almost linearly, while the interaction with `sex` is quite small.



(a) BPH



(b) BPH



(c) CR

Figure 2.12: LTC data: posterior mean together with 80% and 95% credible intervals of the centered log–baseline effect (a), main effect of `age` (centered) and age–dependent effects of `nh` and `sex` for BPH (b) and CR (c)

	mean	std. dev.
sex · nh	-0.42	0.10
level2 · nh	-0.33	0.14
level3 · nh	-0.59	0.14

Table 2.2: LTC data, BNPH model: posterior means and standard deviations for fixed effects

The latter is comparable with the fixed effect interaction (including two steps) of the CR model only for age over 50 years.

Since Czado and Rudolph found out that the effect of care level is time–dependent, we modify our PH–model to a nonproportional hazard (BNPH for short) model with predictor

$$\begin{aligned}\eta(t) &= g_0(\mathtt{t}) + g_{l2}(\mathtt{t}) \cdot \mathtt{level2(t)} + g_{l3}(\mathtt{t}) \cdot \mathtt{level3(t)} + \\ &\quad f_{age}(\mathtt{age}) + f_{sex}(\mathtt{age}) \cdot \mathtt{sex} + f_{nh}(\mathtt{age}) \cdot \mathtt{nh(t)} + \\ &\quad \gamma_{snh} \cdot (\mathtt{sex} \cdot \mathtt{nh(t)}) + \gamma_{l2nh} \cdot (\mathtt{level2(t)} \cdot \mathtt{nh(t)}) + \gamma_{l3nh} \cdot (\mathtt{level3(t)} \cdot \mathtt{nh(t)}).\end{aligned}$$

In contrast to the global log–baseline hazard rate estimated with the BPH model the log–baseline hazard rate $g_0(\mathtt{t})$ and the effect $g_{l2}(\mathtt{t})$ of level2 are now more or less time–constant (compare Figure 2.13 a) and b)). Note that $g_0(\mathtt{t})$ is centered about zero, while $g_{l2}(\mathtt{t}) \approx const. = 0.74$. This means that the time–variation in the effect of level2(t) of the CR model cannot be detected. The increased hazard for level3(t) for smaller t (Figure 2.13 c)) corresponds to a similar finding of the initial CR. A possible interpretation is that this effect is caused by individuals which are already in a bad health state and therefore need level 3 care immediately at the beginning of LTC. The BNPH model with time–varying effects of level2 and level3 can be interpreted as a model with three separate baseline effects $g_0(\mathtt{t})$, $g_0(\mathtt{t}) + g_{l2}(\mathtt{t})$, $g_0(\mathtt{t}) + g_{l3}(\mathtt{t})$ for claimants needing care of level 1,2 or 3, respectively. The corresponding estimated curves are displayed in Figure 2.13 d). As can be gathered form Figure 2.13 e) and Table 2.2 the remaining effects are quite similar to those the BPH model yields.

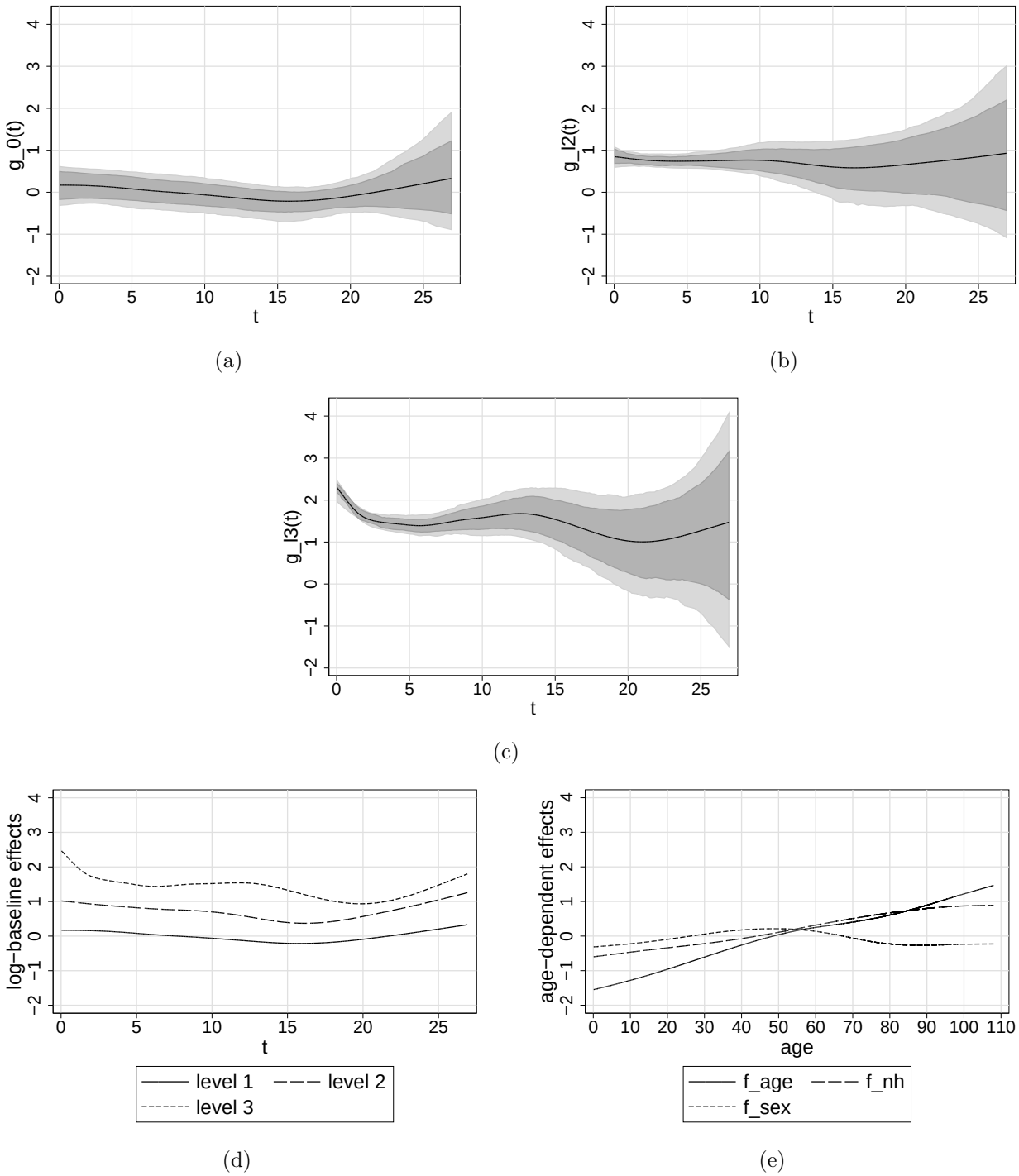


Figure 2.13: LTC data, BNPH model: a)–c) posterior means together with 80% and 95% credible intervals of time–dependent effects, d) posterior means of log–baseline effects for claimants needing care level 1, 2 and 3, respectively, d) posterior means of age–dependent effects

2.5.3 Waiting times to CABG

As a third illustration we apply our methods to data from a study in London and Essex that aims to analyze the effects of area of residence and further individual specific covariates on waiting times to coronary artery bypass graft (CABG). The data comprise observations for 3015 patients with definite coronary artery disease who were referred to one cardiothoracic unit from five contiguous health authorities. Waiting times from angiography to CABG are given in days. Covariates are, among others, sex, age (in years), number of diseased vessels (1, 2, 3), and the area of residence (one of 488 electoral wards).

The data were previously analyzed by Crook et al. (2003) who classified waiting times in months and applied discrete–time survival methodology as described for example in Fahrmeir and Tutz (2001, chap. 9). Here we apply continuous–time geoadditive survival models, with waiting times given in days as in the original data set. We analyzed and compared a hierarchy of models, with model comparison based on the deviance information criterion (DIC), developed in Spiegelhalter et al. (2002). Whilst a log–baseline effect is included in any model, covariate effects are only added gradually. The (log–)baseline prior was assumed as a (log–)piecewise exponential model with grid length $\triangle = 50$ days and, alternatively, as a cubic P–spline model with 20 knots. Table 2.3 gives values for fit (deviance) and complexity (effective number of parameters pD) for a selected number of models. A comparison between the two baseline specifications shows that the complexity is quite alike, whereas the fit is essentially better with the P–spline model than with the p.e.m. The ranking of the models is the same as in Crook et al. (2003).

In the following we will present detailed results for the best models in terms of DIC. Models 7 and 8 correspond to a continuous–time model with hazard rate

$$\lambda(\mathtt{t}) = \exp(g_0(\mathtt{t}) + f_{age}(\mathtt{age}) + f_{spat}(\mathtt{ward}) + \gamma_1 \mathtt{sex} + \gamma_2 \mathtt{dv2} + \gamma_3 \mathtt{dv3}),$$

where $g_0(\mathtt{t})$ is the log–baseline rate, $f_{age}(\mathtt{age})$ is the nonlinear effect of `age` and $f_{spat}(\mathtt{ward})$ is the structured spatial effect. The remaining covariates are dummy–coded: $\mathtt{sex} = 1$ for female, and $\mathtt{sex} = 0$ for male, $\mathtt{dv2} = 1$ if the number of diseased vessels equals 2, $\mathtt{dv2} = 0$ else, and $\mathtt{dv3} = 1$ if the number of diseased vessels equals 3, $\mathtt{dv3} = 0$ else. For comparison we also estimated model 8 with a Weibull prior (2.11) for the baseline hazard. The DIC of this Weibull model is 15273, composed of a deviance of 15190 and pD=42. Accordingly the p.e.m. and the P–spline–model yield a lower DIC in spite of being less parsimonious.

Table 2.3: Model comparison based on the DIC

		P–spline–model			p.e.m.		
	model specification	dev.	pD	DIC	dev.	pD	DIC
1	$g_0(t)$	15607	13	15632	15777	12	15800
2	$g_0(t)$+R	15553	38	15630	15722	38	15798
3	$g_0(t)$+MRF+R	15523	47	15616	15693	44	15782
4	$g_0(t)$+MRF	15532	40	15611	15705	38	15780
5	$g_0(t)$+f_{age}+sex+dv2+dv3	15069	20	15108	15234	19	15273
6	$g_0(t)$+f_{age}+sex+dv2+dv3+R	14934	79	15092	15085	83	15251
7	$g_0(t)$+GRF+f_{age}+sex+dv2+dv3	15017	33	15082	–	–	–
8	$g_0(t)$+MRF+f_{age}+sex+dv2+dv3	14967	56	15079	15125	58	15241
9	$g_0(t)$+MRF+f_{age}+sex+dv2+dv3+R	14943	68	15078	15097	71	15240
10	$g_0(t)$+MRF+f_{age}+sex+$g_1(t)$dv2+$g_2(t)$dv3	14945	64	15073	15107	65	15237

Since the DIC is not improved substantially by adding a random (unstructured) spatial effect (indicated by the letter R in Table 2.3) we do not discuss model 9 in detail.

Since the distribution of the values of `age` is quite skew, it would be an interesting alternative to choose a P–spline prior with knot positions according to quantiles, but we used equidistant knots here, which is our standard choice. The spatial effect $f_{spat}(ward)$ is modelled through a MRF prior. In the case of the P–spline model we alternatively modelled the spatial effect through a GRF prior with 100 knots (model 7). Although this model is more parsimonious, the DIC is greater than with a MRF prior due to a greater deviance. Since the data augmentation that has to be accomplished for the p.e.m. results in an "observation number" of more than 30000, a GRF prior would lead to a computation time of several days, which is not very viable.

The nonproportional hazard model 10, which has the lowest DIC of all models we compared, is a modification of the geoadditive proportional hazard rate model 8, where the fixed effects γ_2 and γ_3 of $dv2$ and $dv3$ are replaced by time varying effects.

Inverse Gamma priors $IG(0.001; 0.001)$ were routinely assumed for the variances, but we also specified uniform priors on standard deviations for comparison. The results were quite alike (similar values of DIC and estimated effects), but uniform priors tend to lead to a

Table 2.4: Posterior mean estimates and standard deviations for the fixed effects on time to CABG

effect	P–spline m., GRF		P–spline m., MRF		p.e.m., MRF		Weibull m., MRF	
sex	-0.04	(0.08)	-0.05	(0.08)	-0.04	(0.08)	-0.04	(0.08)
dv2	1.48	(0.10)	1.49	(0.10)	1.50	(0.10)	1.48	(0.10)
dv3	1.79	(0.09)	1.81	(0.09)	1.82	(0.09)	1.79	(0.10)

slightly better fit coming along with a somewhat larger number of effective parameters pD. However, in contrast to our simulation study, we sometimes faced problems with mixing of Markov chains with $IG(\varepsilon, \varepsilon)$ priors with very small ε's (like $\varepsilon = 0.00000001$) in the case of the age effect, which is presumably due to the skew distribution of the values of `age` (i.e. the small number of young patients).

Table 2.4 contains estimation results for the fixed effects in models 7 and 8. While the effect of `sex` is nonsignificant, the effects of two or three diseased vessels are clearly significant and show that waiting times are decreasing with increasing number of vessels. These results correspond to the findings of Crook et al. (2003). The nonparametric baseline effects in Figure 2.14 show an initially high, but strongly decreasing chance of CABG immediately after diagnosis, followed by a slow increase between 150 and 450 days. Later, the chance of being operated decreases. The overall pattern is similar to the results in Crook et al. (2003), obtained with a discrete–time model. However, with the P–spline prior we get a distinctly smoother curve. The Weibull model also yields a sharp decline in the first days after diagnosis, however, due to the monotonicity of a Weibull baseline, the slow increase between 150 and 450 days can not be detected. The effect of age (Figure 2.14) is almost constant between 40 and 80 years and does not have significant influence on the waiting time. Also, the estimates under a piecewise exponential, a cubic P–spline and a Weibull baseline prior are visually indistinguishable – regardless of which prior is chosen for the structured spatial effect.

The maps in Figure 2.15 show the estimates for the structured spatial effects and give an impression of the spatially varying chance of CABG with green (red) areas indicating an increased (decreased) effect. Again, the estimates under a piecewise exponential and a cubic P–spline baseline prior are visually nearly indistinguishable in the case of a MRF

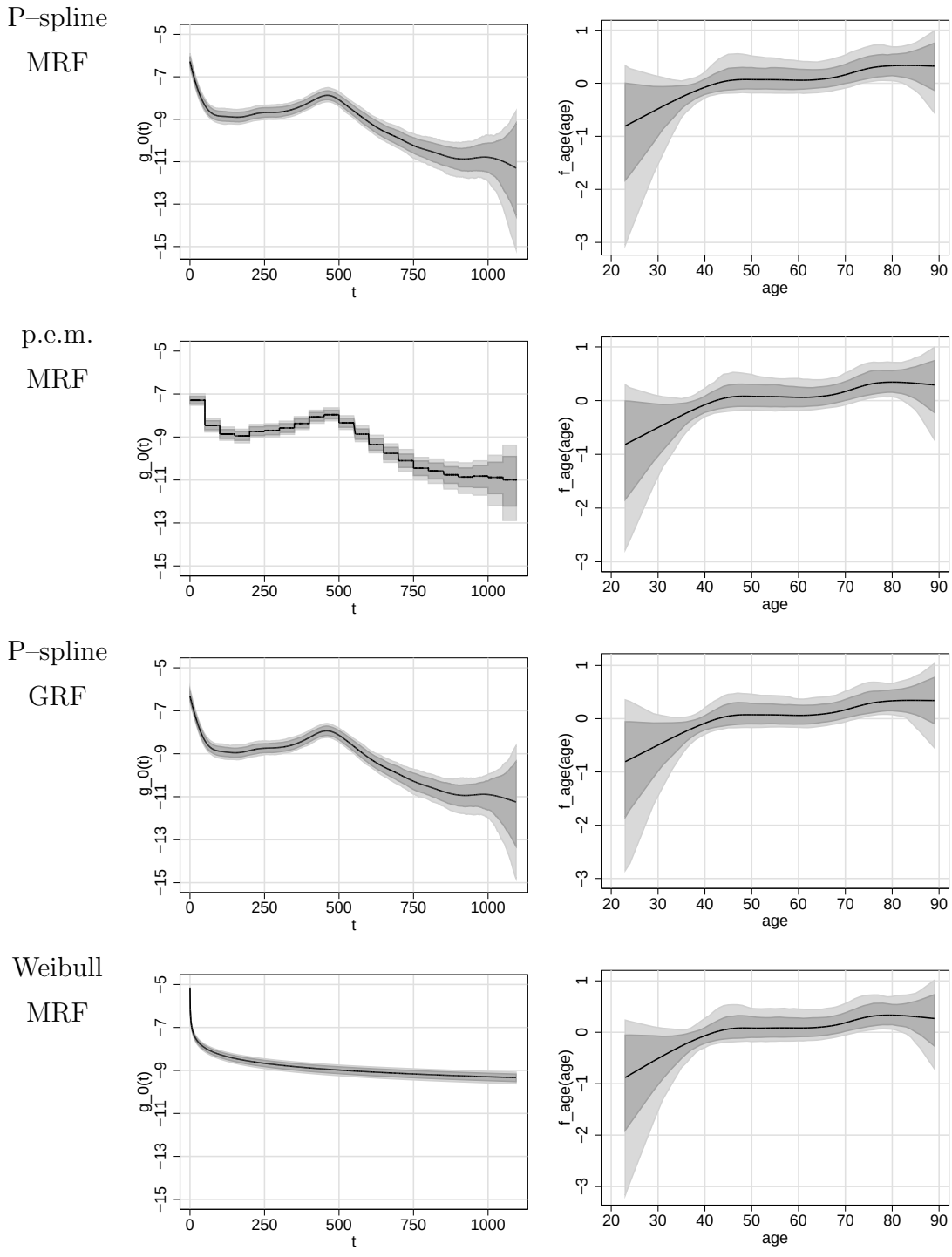


Figure 2.14: Posterior mean estimate for the (log–)baseline effect including the intercept term (left panel) and the (centered) effect of age on time to CABG (right panel) together with 80% and 95% credible intervals

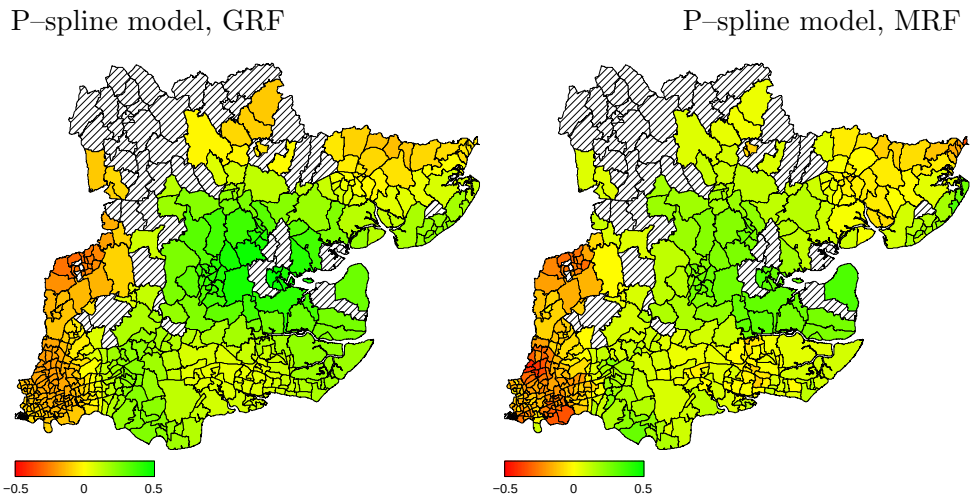


Figure 2.15: Posterior mean estimates of the structured spatial effect on time to CABG; the estimates under the p.e.m. with a MRF prior are visually indistinguishable from those of the P–spline model with MRF prior, and are therefore not shown here

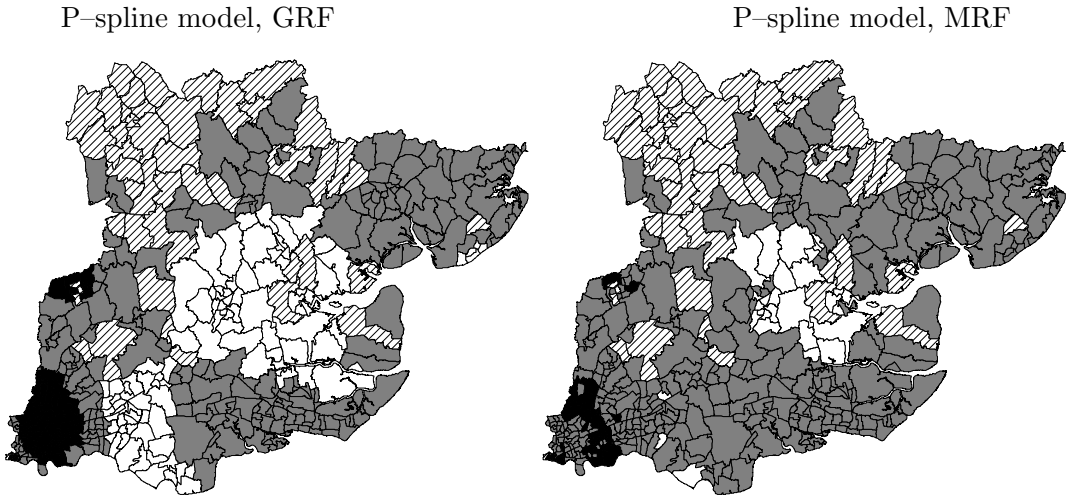


Figure 2.16: Posterior probabilities of the structured spatial effects, with white (black) areas indicating that at least 80% of the sample estimates were positive (negative)

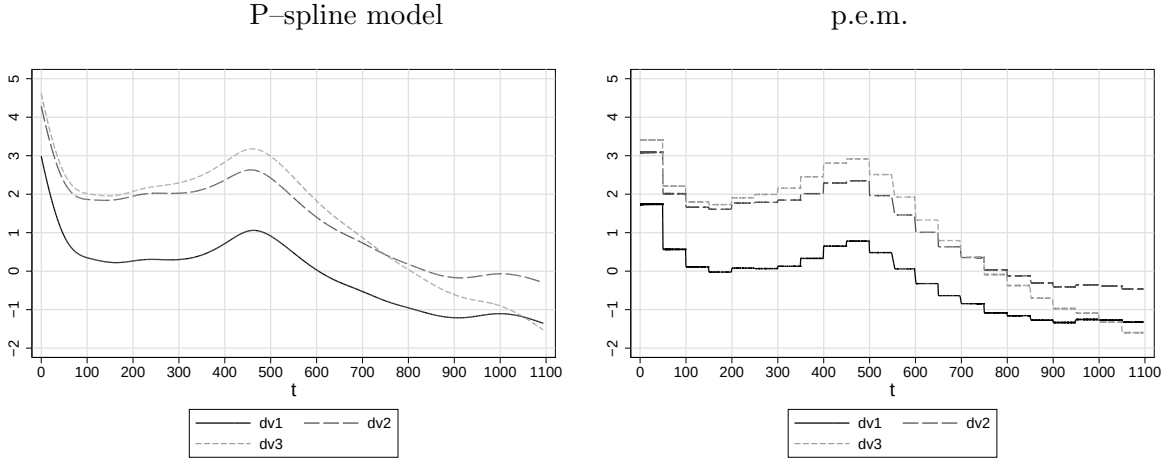


Figure 2.17: (log–)baseline effects on time to CABG: posterior mean estimates for 1 diseased vessel ($dv1$), 2 diseased vessels ($dv2$) and 3 diseased vessels ($dv3$)

prior. Also a Weibull prior yields virtually the same result (not shown). Predictably, the GRF prior results in a smoother estimated spatial effect than the MRF prior does, but besides that the results are quite alike. Areas with increased chances are Chelmsford and Malden in North Essex, while in areas around Harlow in North Essex and Walthamstow and Chingford in North East London chances are lower, that means patients have to wait longer for surgery. The maps in Figure 2.16 show posterior probabilities of these spatial effects. White (black) areas indicate that at least 80 % of the sample estimates were positive (negative). Remaining grey areas are considered as 'nonsignificant'. Striped areas denote wards, where no patient was observed.

Model 10 with time–varying effects $g_1(t)$ and $g_2(t)$ of $dv2$ and $dv3$ can be interpreted as a model with three separate baseline effects $g_0(t)$, $g_0(t)+g_1(t)$, $g_0(t)+g_2(t)$ for patients with one, two or three diseased vessels, respectively. The corresponding estimated curves are displayed in Figure 2.17 and indicate that the proportional hazards assumption is violated, because the baseline effect of patients with three diseased vessels crosses the two other curves.

Different choices for hyperpriors

In the following we exemplary present some additional results of model 8 that were obtained with other choices of $IG(a;b)$ priors. In addition to our standard choice $a = b = 0.001$ we set $a = b = 1e-08$ and $a = -0.5$, $b = 0$ (i.e. uniform prior on the standard deviation).

Figure 2.18 exemplarily shows sampling paths of the first and 19th parameter of each vector $\boldsymbol{\beta}_j$, $j = 0, age, spat$ corresponding to the log–baseline effect, the effect of `age` and the spatial effect, respectively. Independently of the choice of the prior for the hyperparameters the mixing is not optimal for the first parameters of the parameter–vector $\boldsymbol{\beta}_0$ corresponding to the log–baseline effect. In accordance with our simulation study this might be due to the usage of conditional prior proposals and the assumption of a global variance, since the effect is steeply dropping in the first 100 days, but comparatively flat elsewhere. Apart from that we did not face problems with mixing or convergence in the case of $IG(0.001; 0.001)$ and $IG(-0.5; 0)$ priors. However, in the case of an $IG(1e-08; 1e-08)$ prior mixing properties are poor for the first parameters of the effect of `age`, where we have sparse data since there is only a very small number of young patients that suffer from coronary artery diseases. As shown in Figure 2.19 a) the estimated log–baseline effects $g_0(t)$ are not influenced by the choice of the hyperprior. The same applies to the fixed effects as well as the spatial effect. Figure 2.19 b) however reveals a much smoother effect with the $IG(1e-08; 1e-08)$ prior compared to the effects the other two choices for the hyperpriors yield. But since credible intervals are quite large, each estimated effect is within the 95% credible interval of each other estimated effect of `age`.

We conclude that the results are in general quite insensitive regarding the choice of non–informative hyperpriors. However, in situations where data are sparse $IG(a;b)$ priors with a and b close to zero might lead to poor mixing and are therefore not recommended.

2.6 Conclusion

Spatial extensions of statistical models for analyzing survival data will be of increasing relevance because spatial small–area information is often available. Assessment of spatial effects on hazard or survivor functions is not only of interest in its own but can be quite useful for detecting unobserved covariates which carry spatial information. In this chapter,

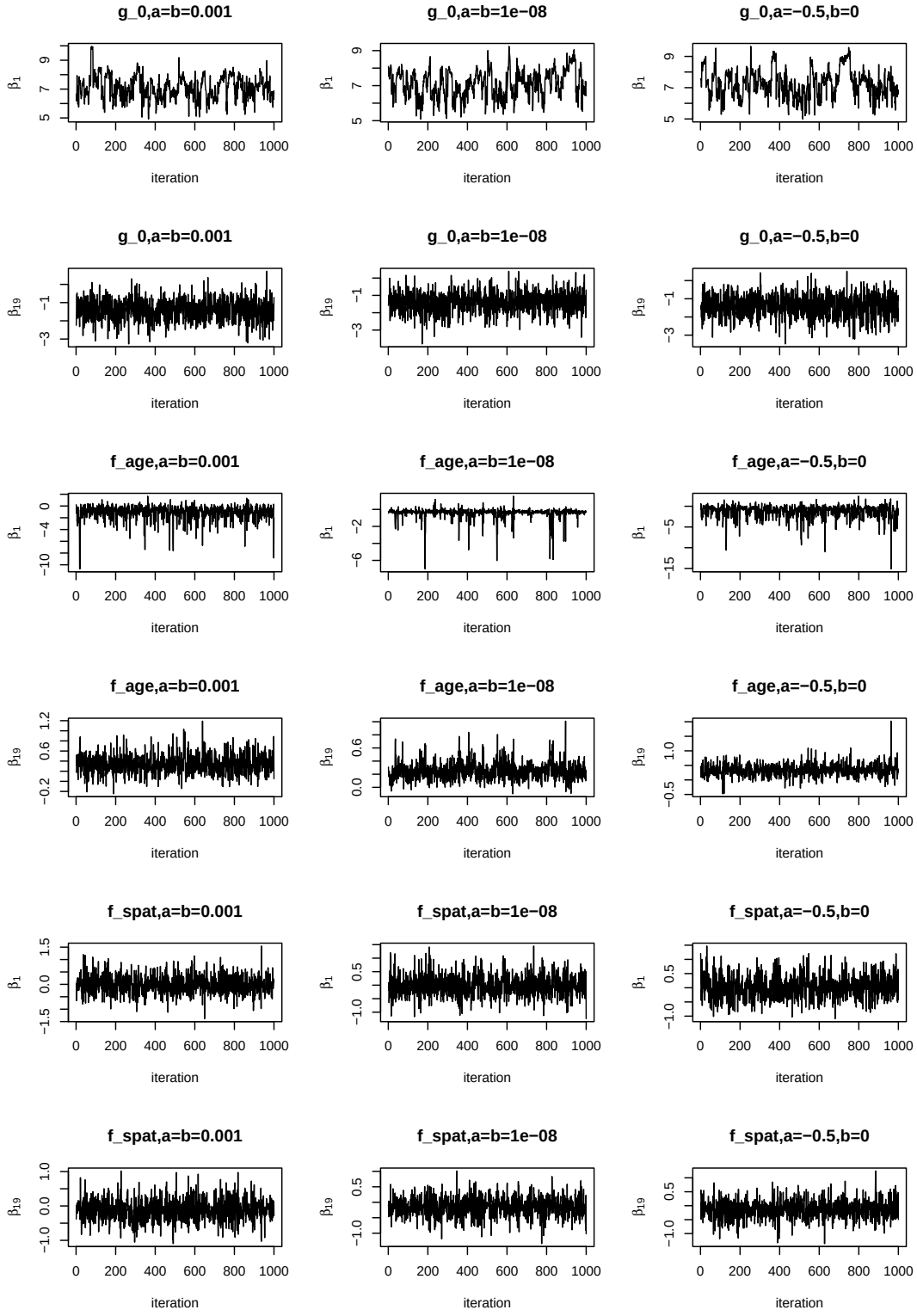
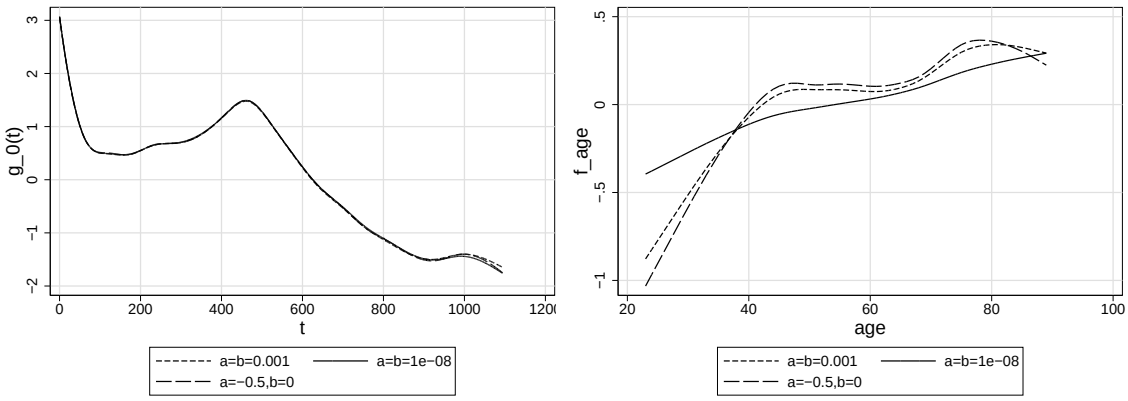


Figure 2.18: Selected sampling paths for parameters $\beta_{j,1}$ and $\beta_{j,19}$, $j = 0, age, spat$ and different choices for the parameters a and b of the $IG(a; b)$ hyperpriors.



(a) (b)

Figure 2.19: Estimated log–baseline effects $g_0(t)$ and effects of age f_{age} with different specifications of $IG(a;b)$ hyperpriors.

we have developed a flexible class of nonparametric geoadditive survival models within a unified Bayesian framework for modelling and inference. Model choice is an important area of ongoing research. A comparison of competing proposals in the context of flexible Bayesian event history models will be of considerable importance.

Chapter 3

Relative Survival Analysis

3.1 Introduction

Many clinical studies aim at identifying prognostic factors for disease specific mortality. However, data on specific causes of death is often not available or not reliable (Percy *et al.* (1981)) and thus it is not possible to differentiate between cases of death that are actually related to the disease of interest and those cases of death that are related to other causes that are independent of this disease. Since the composition of patients in a clinical study usually is quite heterogenous concerning covariates like age (which is the main influencing factor for natural mortality), the natural mortality risk may differ heavily between patients. Thus it might very well be the case that a higher number of deaths is observed with older people although a disease is more likely to be lethal with younger people. In such situations the Cox model is not suitable since therewith it is not possible to distinguish whether a variable like sex or age has an effect on disease specific mortality, on natural mortality or on both. Consequently this model will deliver effects that represent some mixture of the effects on natural and disease related mortality and may therefore be misleading regarding the identification of prognostic factors. Moreover, comparisons of the results from different population–based prognostic studies are difficult due to differences in the natural mortality of the populations. A remedy to this problem is provided by a relative survival analysis which allows for a correction for the effect of other independent causes of death by using the natural mortality in the underlying population as a reference.

Several models for relative survival analysis in a frequentist setting have been discussed

in the literature. Esteve *et al.* (1990) assume that the observed hazard for total mortality is the sum of two hazards, namely the expected, natural mortality hazard and a disease related mortality hazard. Whereas the first component is obtained from external sources the disease related hazard is estimated parametrically assuming a piecewise constant baseline effect and time–constant fixed effects of covariates. This approach was extended by Bolard *et al.* (2001) and Giorgi *et al.* (2003) by allowing for time–varying effects, i.e. dropping the proportional hazards assumption. Bolard *et al.* (2001) consider time–by–covariate interactions originally proposed by Cox (1972) as well as piecewise proportional hazards, developed by Moreau *et al.* (1985) for crude survival analysis. The drawbacks of these methods are that temporal variations in the effects of covariates are limited to pre–specified parametric forms of interaction functions and step–functions on pre–specified time intervals, respectively. A more flexible method is proposed by Giorgi *et al.* (2003) who assume quadratic B–splines with two inner knots for the baseline effect as well as for time–varying effects of covariates. In the Bayesian approach we present here we extend the model of Esteve *et al.* (1990) by modelling the disease related hazard with a flexible geoadditive predictor that may include a log–baseline effect, nonlinear effects of continuous covariates and time–varying effects modelled by P–splines, as well as a spatial effect, random effects and the usual fixed effects.

The rest of this chapter is organized as follows. In Section 3.2 we describe models, likelihood and priors for unknown functions and parameters. Some comments on the inference via MCMC are given in Section 3.3. To illustrate our approach we present an application to data on the survival of women suffering from breast cancer in Section 3.4. Reliability of our approach is verified in Section 3.5 by means of a simulated data set with known risk profile.

3.2 Model, likelihood and priors

Consider right–censored survival data as described in Subsection 2.2.1 in which T_i now denotes survival time of observation i until death of any cause, t_i denotes the observed survival time and δ_i denotes the censoring indicator. Following Esteve *et al.* (1990) we assume that the hazard rate for total mortality $\lambda_i(t, a_i, cov_i) := \lambda_i(t)$ at time t after diagnosis of an individual i with age a_i at diagnosis and a vector of covariates $cov_i = (z_i, x_i, s_i, v_i)$

(possibly including age) is defined as the following sum of two hazards:

$$\begin{aligned} \lambda_i(t, a_i, cov_i) := \lambda_i(t) &= \lambda_i^e(a_i + t, cov_i^{sub}) + \lambda_i^c(t, cov_i) \\ &= \lambda_i^e(a_i + t, cov_i^{sub}) + \exp\left(\eta_i(t, cov_i)\right) \end{aligned} \tag{3.1}$$

The first summand $\lambda_i^e(a_i + t, cov_i^{sub})$ represents the expected hazard for natural mortality in a population and is obtained from mortality tables using external sources, i.e. there are no unknown parameters involved here. This component depends only on age at time t after diagnosis (i.e. $a_i + t$) and cov_i^{sub}, a subvector including those covariates in cov_i mortality tables account for (usually sex and period). The second summand $\lambda_i^c(t, cov_i)$ is the disease related mortality hazard rate which is estimated from the data at hand. This component is modelled by a flexible, possibly geoadditive predictor as in (2.4). To simplify notation the dependence on cov_i^{sub} and cov_i, respectively will be suppressed in the following, i.e. we define $\lambda_i^e(a_i + t, cov_i^{sub}) := \lambda_i^e(a_i + t)$ and $\lambda_i^c(t, cov_i) := \lambda_i^c(t)$. Depending on what kind of covariates are given in cov_i, the predictor may be composed of the following summands:

$$\eta_i(t, cov_i) := \eta_i(t) = g_0(t) + \sum_{j=1}^{p} g_j(t) z_{ij} + \sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i'\boldsymbol{\gamma} + b_{g_i}, \tag{3.2}$$

where $g_0(t) = \log\{\lambda_0(t)\}$ is the (disease related) log–baseline hazard, $g_j(t)$ are time–varying effects of covariates z_j, $f_j(x_j)$ is the nonlinear effect of a continuous covariate x_j, $f_{spat}(s_i)$ is the (structured) effect of a spatial covariate s, γ is the vector of linear effects and b_g is a unit– or group–specific frailty or random effect (see Subsection 2.2.1 for a more detailed description).

Once more, for a interpretation of equation (3.1) one may say that the natural mortality hazard λ_i^e covers the basic mortality risk a population is exposed to and the disease related hazard λ_i^c models the excess mortality risk that patients are exposed to beyond the basic risk due to the disease they suffer from. From a statistical point of view λ_i^e is an additive offset.

Under the assumption about noninformative censoring the likelihood is given by

$$L = \prod_{i=1}^{n} \left(\lambda_i(t_i)\right)^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \lambda_i(u)du\right)$$

Inserting (3.1) results in

$$\begin{aligned} L &= \prod_{i=1}^{n} (\lambda_i^e(a_i + t_i) + \lambda_i^c(t_i))^{\delta_i} \exp\left(- \int_0^{t_i} (\lambda_i^e(a_i + u) + \lambda_i^c(u))\, du \right) \\ &= \prod_{i=1}^{n} (\lambda_i^e(a_i + t_i) + \lambda_i^c(t_i))^{\delta_i} \exp\left(- \int_0^{t_i} \lambda_i^c(u) du \right) \exp\left(- \int_0^{t_i} \lambda_i^e(a_i + u) du \right), \end{aligned} \tag{3.3}$$

where the last factor does not depend on the parameters to be estimated. Hence the following proportionality holds

$$L \propto \prod_{i=1}^{n} (\lambda_i^e(a_i + t_i) + \lambda_i^c(t_i))^{\delta_i} \exp\left(- \int_0^{t_i} \lambda_i^c(u) du \right). \tag{3.4}$$

This formula only differs from the likelihood of a crude survival model given in (2.5) by the term $\lambda_i^e(a_i + t_i)$.

Again, for defining priors and developing posterior analysis we can rewrite the observation model in generic matrix notation and represent the predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_i, \ldots, \eta_n)'$, where $\eta_i := \eta_i(t_i)$, as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_0\boldsymbol{\beta}_0 + \ldots + \boldsymbol{Z}_m\boldsymbol{\beta}_m. \tag{3.5}$$

See Subsection 2.2.1 for details. Then, to complete the Bayesian model formulation priors for parameters and functions are assumed as described for the crude survival analysis in Subsection 2.2.2, i.e. we assume diffuse priors for fixed effect parameters, i.i.d. Gaussian priors for uncorrelated random effects, P–splines for the baseline effect, nonparametric effects of continuous covariates and time–varying effects, Markov random field (MRF) priors, two–dimensional tensor product P–spline priors, or Gaussian random field (GRF) priors for structured spatial effects and inverse Gamma hyperpriors for all variance components.

3.3 Markov chain Monte Carlo inference

As before in Section 2.3 let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', ..., \boldsymbol{\beta}_m')'$ denote the vector of all regression coefficients in the generic notation (3.5), $\boldsymbol{\gamma}$ the vector of fixed effects, and $\boldsymbol{\tau}^2 = (\tau_0^2, ..., \tau_m^2)$ the vector of all variance components. Full Bayesian inference is based on the entire posterior distribution

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2 \mid data) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)\, p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2).$$

Due to the (conditional) independence assumptions, the joint prior factorizes into

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2) = \left\{ \prod_{j=0}^{m} p(\boldsymbol{\beta}_j \mid \tau_j^2) p(\tau_j^2) \right\} p(\boldsymbol{\gamma}),$$

where the last factor can be omitted for diffuse fixed effect priors. The likelihood $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)$ is given by inserting (3.2) into (3.4). Note that the integral does not require integration over the natural mortality hazard $\lambda_i^e(a_i + t)$ (which is fix anyway), but just over the same terms as before with the crude survival analysis, i.e. terms of the form

$$I_i = \int_0^{t_i} \exp\left(g_0(u) + \sum_{j=1}^{p} g_j(u) z_{ij} \right) du,$$

where $g_j(t) = \sum \beta_{jm} B_m(t)$. As described in Section 2.3 we usually use the trapezoidal rule to solve these integrals numerically.

Again, full Bayesian inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters, given the rest of the data. Basically all parameters are updated as described in Section 2.3. However, the calculation of the means and precision matrices of the multivariate Gaussian distributions, that are used within the IWLS–MH algorithm to approximate the posterior of the parameter vectors $\boldsymbol{\beta}_j$, which correspond to the time–independent functions $f_j(x_j)$, as well as spatial effects $\boldsymbol{\beta}^{spat}$, fixed effects $\boldsymbol{\gamma}$ and random effects $\mathbf{b}$, is slightly more complex. Suppose we want to update $\boldsymbol{\beta}_j$, with current value $\boldsymbol{\beta}_j^c$ of the chain. Then a new value $\boldsymbol{\beta}_j^p$ is proposed by drawing a random vector from a multivariate Gaussian distribution with precision matrix and mean

$$\boldsymbol{P}_j = \boldsymbol{Z}_j' \boldsymbol{W}(\boldsymbol{\beta}_j^c) \boldsymbol{Z}_j + \frac{1}{\tau_j^2} \boldsymbol{K}_j, \quad \boldsymbol{m}_j = \boldsymbol{P}_j^{-1} \boldsymbol{Z}_j' \boldsymbol{W}(\boldsymbol{\beta}_j^c)(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\eta}}).$$

where $\tilde{\eta}_i = \eta_i(t_i) - f_j(x_{ij})$, $\boldsymbol{W}(\boldsymbol{\beta}_j^c) = diag(w_1, \ldots, w_n)$ is the weight matrix for IWLS with weights

$$w_i \;\; = \;\; \Lambda_i^c(t_i) - \frac{\lambda_i^e(a_i + t_i) \lambda_i^c(t_i) \delta_i}{\lambda_i(t_i)^2}$$

obtained form the current state $\boldsymbol{\beta}_j^c$ and with $\Lambda_i^c(t_i) = \int_0^{t_i} \lambda_i^c(u) du$. The working observations $\tilde{y}_i$ are given by

$$\tilde{y}_i = \eta_i(t_i) + \frac{\delta_i \lambda_i^c(t_i)/\lambda_i(t_i) - \Lambda_i^c(t_i)}{w_i}.$$

See Appendix A2 for a detailed derivation of those quantities.

3.4 Application

We illustrate our method by an application to data on breast cancer that was gathered in the years from 1988 to 2002 by a cancer registry that covers the Haut–Rhin 'department' which is located in the north-east of France, adjacent to Germany and Switzerland. This department has 3525 km^2 and 707555 inhabitants (in 1999) and is partitioned into 377 municipalities. The largest distance between the centroids of two municipalities is about 95 kms. The data set contains 3726 cases of breast cancer diagnosed between January the 1st 1988 and January the 1st 1998. There were 1235 ($\approx 33\%$) deaths observed whereas the causes of death are unknown. Observed lifetimes are given in days and range from 0 to 14 years, with a median of 6.4 years. Covariates are age at time of diagnosis (ranging from 20.6 years to 87.1 years), date of diagnosis (ranging from 1988.0 (i.e. 01.01.1988) to 1998.0), area of residence (one of 377 municipalities) and number of metastases at the date of diagnosis (no metastasis, one metastasis or more than one metastasis). This is part of a data set that has been analyzed via crude survival analysis by Sauleau et al. (2006).

For comparison only we analyze the data with the crude survival model (2.3) although this model does not account for natural mortality and is thus not appropriate to the data at hand where causes of death are not available. Generally the specification of the hazard rate is given by

$$\begin{aligned} \lambda_i(t, cov_i) &= \exp(\eta_i(t, cov_i)) \\ cov_i &= (a_i, p_i, s_i, \texttt{meta1}_i, \texttt{meta2}_i), \end{aligned} \tag{3.6}$$

where t is time since diagnosis and cov_i is the vector of covariates with a_i denoting the age of patient i at date of diagnosis p_i (period), s_i denoting the municipality patient i resides in and the dummy–coded covariates $\texttt{meta1}_i$ and $\texttt{meta2}_i$ denoting, whether patient i has one metastasis and more than one metastasis, respectively.

A relative survival analysis should be more suitable and deliver better results. Therefore we alternatively assume a composed hazard rate of the following structure

$$\begin{aligned} \lambda_i(t, cov_i) &= \lambda_i^e(a_i + t, p_i + t) + \exp(\eta_i(t, cov_i)) \\ cov_i &= (a_i, p_i, s_i, \texttt{meta1}_i, \texttt{meta2}_i), \end{aligned} \tag{3.7}$$

where $\lambda_i^e(a_i + t, p_i + t)$ is the natural mortality rate of women of age $a_i + t$ at date $p_i + t$ as recorded in mortality tables for the Haut–Rhin department. The second summand

$\lambda_i^c = \exp(\eta_i(t, cov_i)$ represents the disease related hazard rate and is modelled in the same way as the hazard rate in (3.6).

A hierarchy of models is analyzed with both approaches and compared via the deviance information criterion (DIC). Whilst a log–baseline effect $g_0(t)$ modelled by a cubic P–spline prior with 20 knots is included in any model, covariate effects are only included gradually. Effects $f_a(a_i)$ and $f_p(p_i)$ of continuous covariates are modelled by cubic P–splines with 20 knots. Diffuse priors are assigned to the fixed effects γ_1 and γ_2 of the dummy–coded covariates `meta1` and `meta2`. The structured spatial effect $f_{spat}(s_i)$ is modelled by a MRF prior which is our standard choice with area–level data. An unstructured (random) spatial effect b_{s_i} is included additionally or alternatively in some of the models. Table 3.1 gives values for fit and complexity of a selected number of models according to the two components of the deviance information criterion. Model I, which contains a structured spatial effect modelled by a MRF–prior, the effect of the number of metastases and the effect of age, yields a DIC of 9308 for the crude survival model with hazard rate (3.6) and 9249 for the relative survival model. Leaving out one or more of these effects leads to a larger DIC. As Table 3.1 shows the DIC is slightly reduced by the additional inclusion of a period effect. Models III and IV are versions of model II where the spatial effect is modelled by an unstructured (random) effect b_s and the sum of a structured and an unstructured effect, respectively. However, those models will not be discussed here since they do not lead to an improvement in terms of DIC. Figure 3.1 displays the estimated nonparametric effects of model II with predictor

$$\eta_i = g_0(t) + f_a(a_i) + f_p(p_i) + f_{spat}(s_i) + \gamma_1 \texttt{meta}_1 + \gamma_2 \texttt{meta}_2.$$

With the software *BayesX* all unknown functions are centered about zero, and an intercept term is included in the parametric linear term for identifiability reasons. For plotting, the estimated effects of age a_i and period p_i are all centered at the observed values, i.e.

$$\sum_{i=1}^{3726} \hat{f}_a(a_i) = \sum_{i=1}^{3726} \hat{f}_p(p_i) = 0,$$

while the intercept is added to the log–baseline effects. Hence it can be derived from Figure 3.1(a) and (b) that the estimated global risk level is higher with the crude survival model (since the log–baseline effect resulting from a crude survival analysis exceeds the

log–baseline effect resulting from a relative survival analysis). This results from the fact that the crude survival analysis delivers an estimation of the risk of dying of any cause, whereas only the disease related excess mortality risk of breast cancer patients is estimated by means of a relative survival analysis, where the natural mortality risk is accounted for separately. Panels (a) and (b) further reveal that the crude survival analysis yields a fairly constant log–baseline effect $g_0(t)$, whereas a relative survival analysis results in an effect, that is increasing in the first two years and decreasing in the time between the third and the 11th year after diagnosis. Presumably the decrease in risk is not reflected in panel (a) as not accounting for natural mortality that is increasing with time after diagnosis (since patients are aging) might lead to a neutralization. The estimated effects of age at time of diagnosis exhibit an u–shaped risk profile and are displayed in panels (c) and (d). While a crude survival analysis yields an increased risk for patients diagnosed with breast cancer in their younger days, but a still much higher risk for those women diagnosed at an age of more than 70 years, a relative survival analysis suggests that women diseased in early life have the greatest risk. This result is in accordance with the fact that cancers are often more aggressive with younger people. The differences between the two approaches were to be expected since older women have a higher natural mortality risk that is not accounted for separately with the crude, but only with the relative survival analysis. As displayed in panels (e) and (f) both approaches yield a higher risk for patients that were diagnosed with breast cancer in earlier periods. This effect might be explained by medical progress. Figures 3.2 (a) and (b) display the values of the structured spatial effect in each municipality. The two approaches yield a similar spatial pattern, but it is more pronounced with the relative survival analysis. The risk seems to be higher in the south of the region. None of these effects is significant on a level of 95%, but a couple of regions exhibit effects that are significant on a level of 80%, such as some regions in the north–east that have a lower risk (Figures 3.2(c) and (d)). The estimated parameters $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for the fixed effects of `meta1` and `meta2` are greater with the relative survival approach. In detail the results are as follows:

	crude	relative
$\hat{\gamma_1}$	0.66	0.96
$\hat{\gamma_2}$	2.23	2.74

meaning that compared to patients with no metastases the hazard rate is about 1.9 (9.3)

		crude survival			relative survival		
	Model	$D(\bar{\theta})$	p_D	DIC	$D(\bar{\theta})$	p_D	DIC
I	$g_0(t) + f(a) + f_{spat}(s) + \texttt{meta}$	9268	20	9308	9208	20	9249
II	$g_0(t) + f(a) + f(p) + f_{spat}(s) + \texttt{meta}$	9259	24	9307	9200	23	9246
III	$g_0(t) + f(a) + f(p) + b_s + \texttt{meta}$	9264	24	9312	9205	24	9253
IV	$g_0(t) + f(a) + f(p) + f_{spat}(s) + b_s + \texttt{meta}$	9250	29	9308	9192	28	9248
V	$g_0(t) + f(a) + f(p) + f_{spat}(s) + g(t) * \texttt{meta}$	9239	28	9296	9187	27	9241

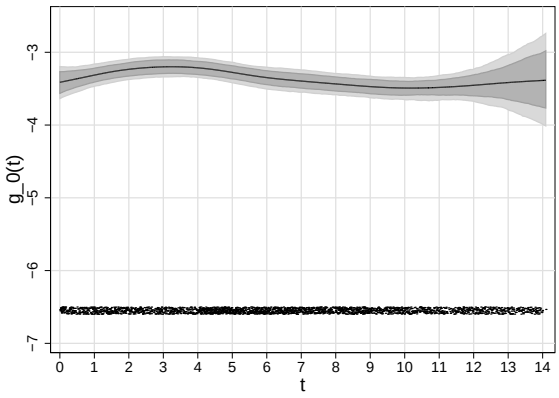
Table 3.1: Deviance, effective number of parameters p_D and DIC for some of the models we compare.

and 2.6 (15.5) times higher for patients with one (more than one) metastasis, respectively.

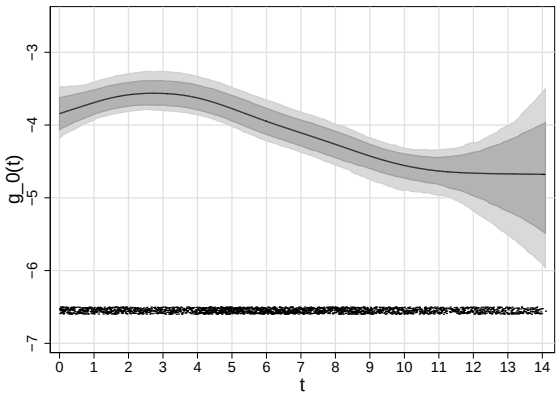
To investigate if the proportional hazards assumption is appropriate, the number of metastases is included as a covariate with time–varying effect in model II, i.e. the disease–related log–hazard of model V is

$$\lambda_i^c = \exp\left(g_0(t) + \texttt{meta1}_i \cdot g_1(t) + \texttt{meta2}_i \cdot g_2(t) + f_{age}(a_i) + f_p(p_i) + f_{spat}(s_i)\right).$$

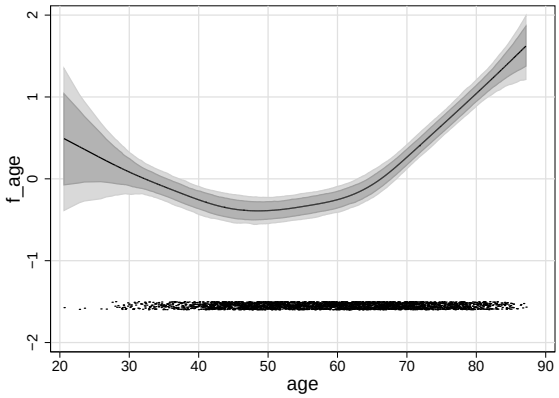
Here $g_0(t)$ is the log–baseline effect for patients without metastases, $g_0(t)+g_1(t)$ corresponds to the log–baseline for patients with one metastasis and $g_0(t)+g_2(t)$ for patients with more than one metastasis. The time–dependent functions $g_k(t), k = 0, 1, 2$ are modelled with cubic P–spline priors with 20 knots. As displayed in Table 3.1 the DIC is reduced by allowing for a temporal variation in the effect of the number of metastases. The three log–baseline effects are plotted in Figure 3.3 and reveal that the differences in risk between the patient groups seem to diminish with time after diagnosis. The log–baseline effect for patients with more than one metastasis even crosses the other curves, but this result must not be over–interpreted since there are only 70 patients with more than one metastasis in the study. The remaining estimated effects of model V resemble the results of model II and are not shown for this reason.



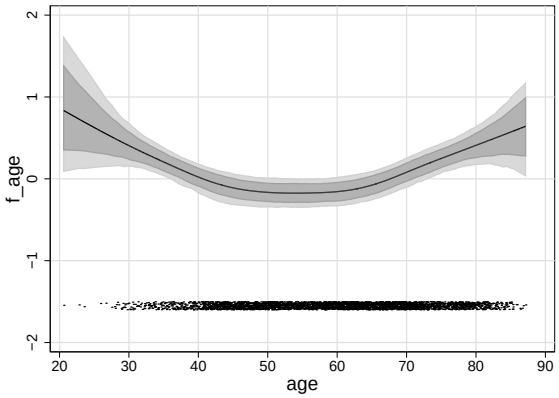
(a) log–baseline effect $g_0(t)$



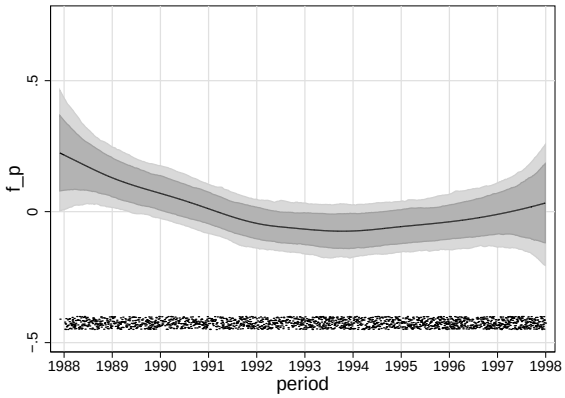
(b) log–baseline effect $g_0(t)$



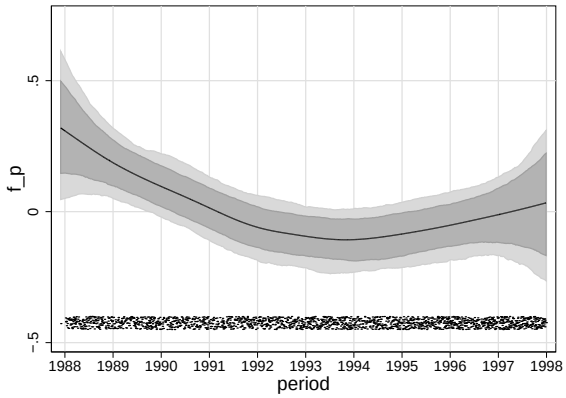
(c) centered effect of age



(d) centered effect of age



(e) centered effect of period



(f) centered effect of period

Figure 3.1: Model II: Posterior means and pointwise 80% and 95% confidence intervals for the baseline effect including the intercept term (a,b), the centered effect of age (c,d) and the centered effect of period (e,f). Figures b,d and f result from a relative survival analysis.

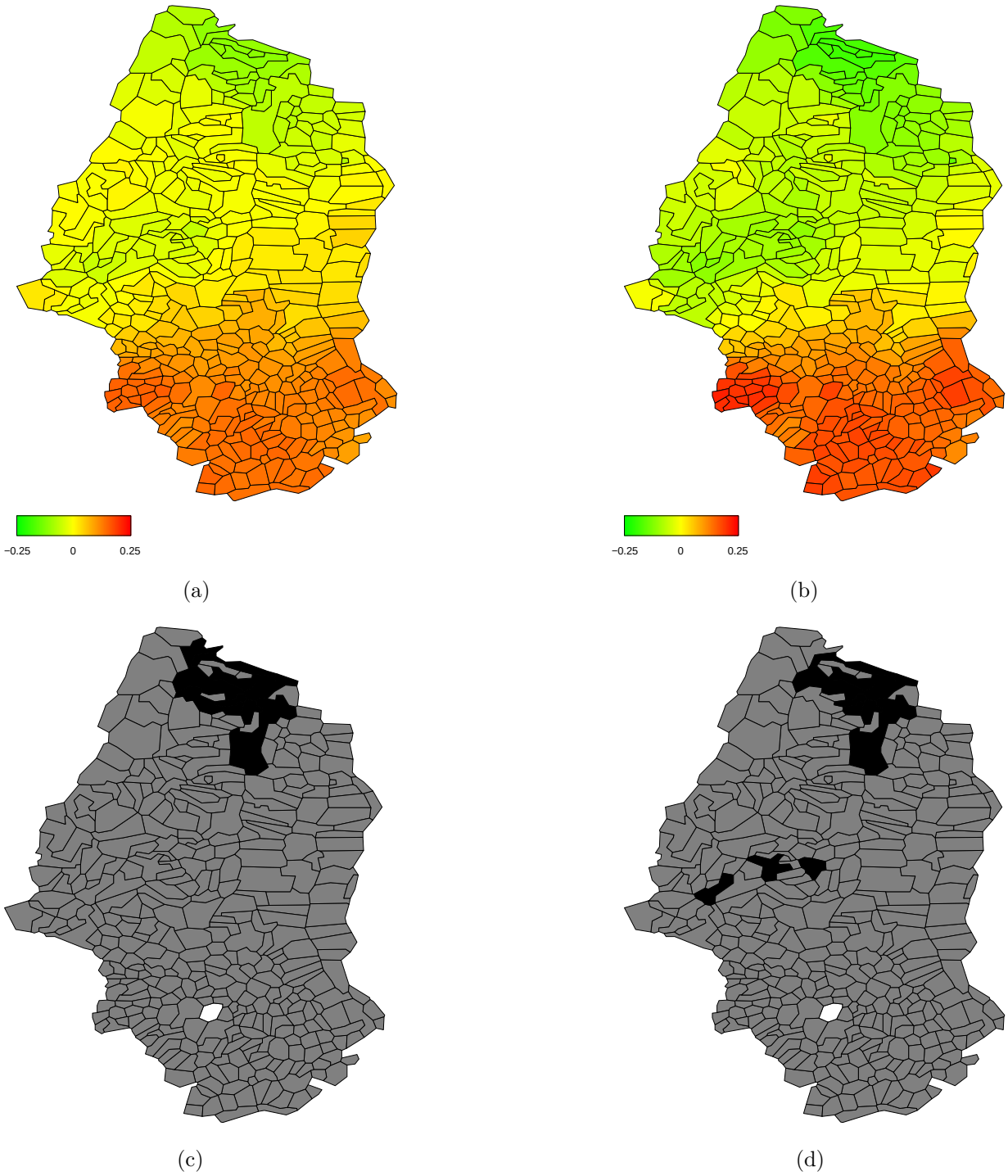
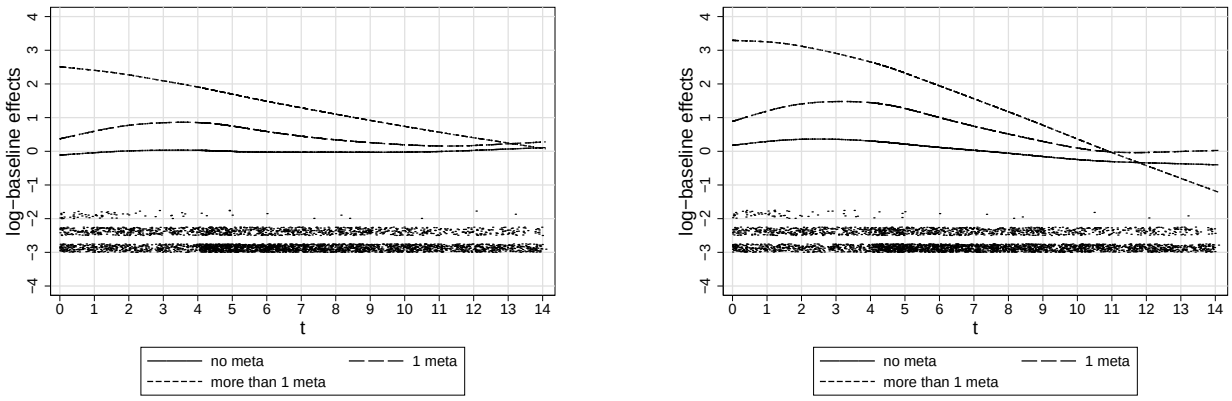


Figure 3.2: Model II: posterior means of the structured spatial effect (MRF) and posterior probabilities for a nominal level of 80%, where black denotes regions with strictly negative credible intervals and white denotes regions with strictly positive credible intervals. Panels b) and d) result from a relative survival analysis.



(a) crude survival

(b) relative survival

Figure 3.3: Model V: posterior means of the log–baseline effects for patients with no metastases, one metastasis and more than one metastasis (dots in the lowest, middle and highest row mark observed lifetimes of patients with no metastases, one metastasis and more than one metastasis, respectively)

3.5 Simulation

To verify the reliability of our relative survival model and to show that a model that does not account for natural mortality can indeed be misleading concerning the effects of covariates in such cases where data on specific causes of death is not available, we simulate an appropriate data set with known risk profile. Survival times are generated according to a hazard rate that is the sum of a natural hazard rate and a disease related hazard rate. This data set is then analyzed with a crude survival model like in (2.3) and with a relative survival model like in (3.1) and the results are compared subsequently.

As for the data generation we simulate survival times based on the covariates of our real breast cancer data set, using known specifications for the baseline and the covariate effects that resemble the effects estimated by the relative survival analysis of the real data set. However, for the sake of simplification we neither consider a spatial effect nor a period effect. We first simulate survival times for each subject and the censoring is done in a second step. In detail, survival times $T_i, i = 1, \ldots, 3726$, are generated according to the following hazard rate model

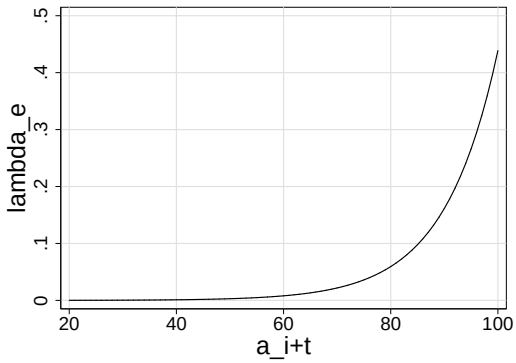
$$\begin{aligned}\lambda_i(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) &= \lambda_i^e(a_i + t) + \lambda_i^c(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) \\ &= \lambda_i^e(a_i + t) + \exp(g_0(t) + f_{age}(a_i) + \gamma_1 \texttt{meta1}_i + \gamma_2 \texttt{meta2}_i),\end{aligned}$$

where the natural hazard rate λ_i^e is chosen in order to resemble the natural mortality rates used with the application, but only depends on $a_i + t$, which is the age of individual i at time t after diagnosis. In our application natural mortality also depends on calendar time, but we did not consider this here. As illustrated in Figure 3.4(a) the natural hazard rate is increasing exponentially with age at time t after diagnosis. The disease related hazard rate λ_i^c depends on time t after diagnosis, the age at time of diagnosis a_i, and the two binary covariates $\texttt{meta1}_i$ and $\texttt{meta2}_i$, which indicate whether an individual i has one and more than one metastasis, respectively. As displayed in Figure 3.4(b) the disease related log–baseline $g_0(t)$ is increasing in the first 2.5 years after diagnosis, decreasing in the time span between 2.5 and 12 years and staying constant afterwards. In contrast to the natural mortality risk, the effect of age on the disease related risk is u–shaped and highest with patients diseased in early life, whereas it is less increased with the initially oldest patients in the study, who are diagnosed with breast cancer at the age of 87 (Figure 3.4(c)). Finally the disease related log–hazard is increased by $\gamma_1 = 0.95$ and $\gamma_2 = 2.75$ for individuals with one metastasis ($\texttt{meta1}_i = 1$) and more than one metastasis ($\texttt{meta2}_i = 1$), respectively. Since the data used in our application were only gathered until the year 2002 we consider all survival times exceeding the year 2002 as censored, i.e. observed survival times are given by $t_i = \min(T_i, 2002.0 - p_i)$ with p_i denoting the exact date of diagnosis observed in the real data set. This mechanism results in a censoring rate of approximately 60% (compared to approximately 67% with the real data set).

The data set generated in this way is initially analyzed with a crude survival model like in (2.3) that does not distinguish between natural mortality and disease related mortality. More precisely we wrongly assume a hazard rate as follows:

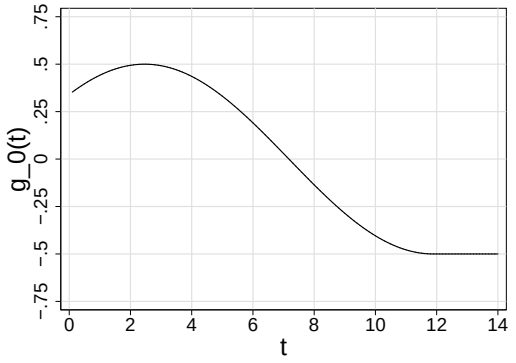
$$\lambda_i(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) = \exp(g_0(t) + f_{age}(a_i) + \gamma_1 \texttt{meta1}_i + \gamma_2 \texttt{meta2}_i),$$

where the log–baseline $g_0(t)$ and the age–effect f_{age} are modelled as cubic P–splines with 20 knots (with second order random walk smoothness priors and $IG(0.001, 0.001)$ priors for



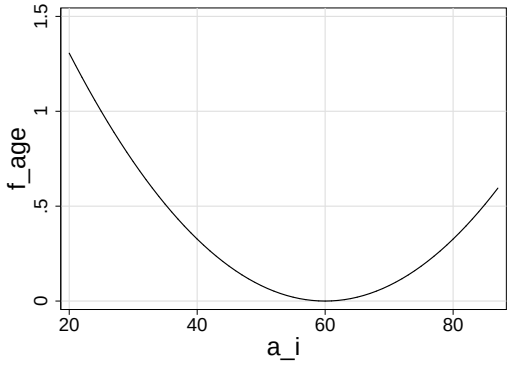
a) natural hazard rate against age

$$\lambda_i^e(a_i + t) = \exp\left((a_i + t - 30)/10\right)/2500$$



b) disease related log–baseline effect

$$g_0(t) = \begin{cases} 0.5 \cdot \sin(t/3 + 0.75) - 4 & , t \leq 4.5\pi - 2.25 \\ -4.5 & , t > 4.5\pi - 2.25 \end{cases}$$



c) disease related effect of age at time of diagnosis

$$f_{age}(a_i) = ((a_i - 60)/35)^2$$

Figure 3.4: Simulation: specifications for the natural hazard rate, the disease related log–baseline effect and the disease related effect of age at time of diagnosis

the variance components) and γ_1 and γ_2 are fixed effects with diffuse priors. Expectedly the estimated log–baseline and the effect of age do not reflect the true disease related effects but rather present a mixture of the two effects on natural mortality and disease related mortality. The estimated log–baseline effect is increasing in the first years after diagnosis, but the subsequent decline is less steep than with the true log–baseline effect (Figure 3.5(a)). While the disease related log–baseline is decreasing between the 2.5th and 12th year after diagnosis, the natural mortality risk of each single patient is increasing with time (since people are getting older) and these two effects seem to kind of balance. As can be seen from Figure 3.5(c) the crude survival model underestimates the risk for women diagnosed with breast cancer in early years and overestimates the risk of women diseased at an old age. Again, this high risk for older people results from the increasing natural mortality risk that is not accounted for separately. Finally also the fixed effects of the covariates `meta1` and `meta2` are not estimated correctly, but are rather underestimated by $\hat{\gamma_1} = 0.68$ and $\hat{\gamma_2} = 2.33$ (with standard deviations of 0.05 and 0.13, respectively). This underestimation is due to the fact that only a part of the cases of death (namely those cases that are related to the disease) are in association with the number of metastases, whereas the crude survival analysis estimates the average influence based on all cases of death.

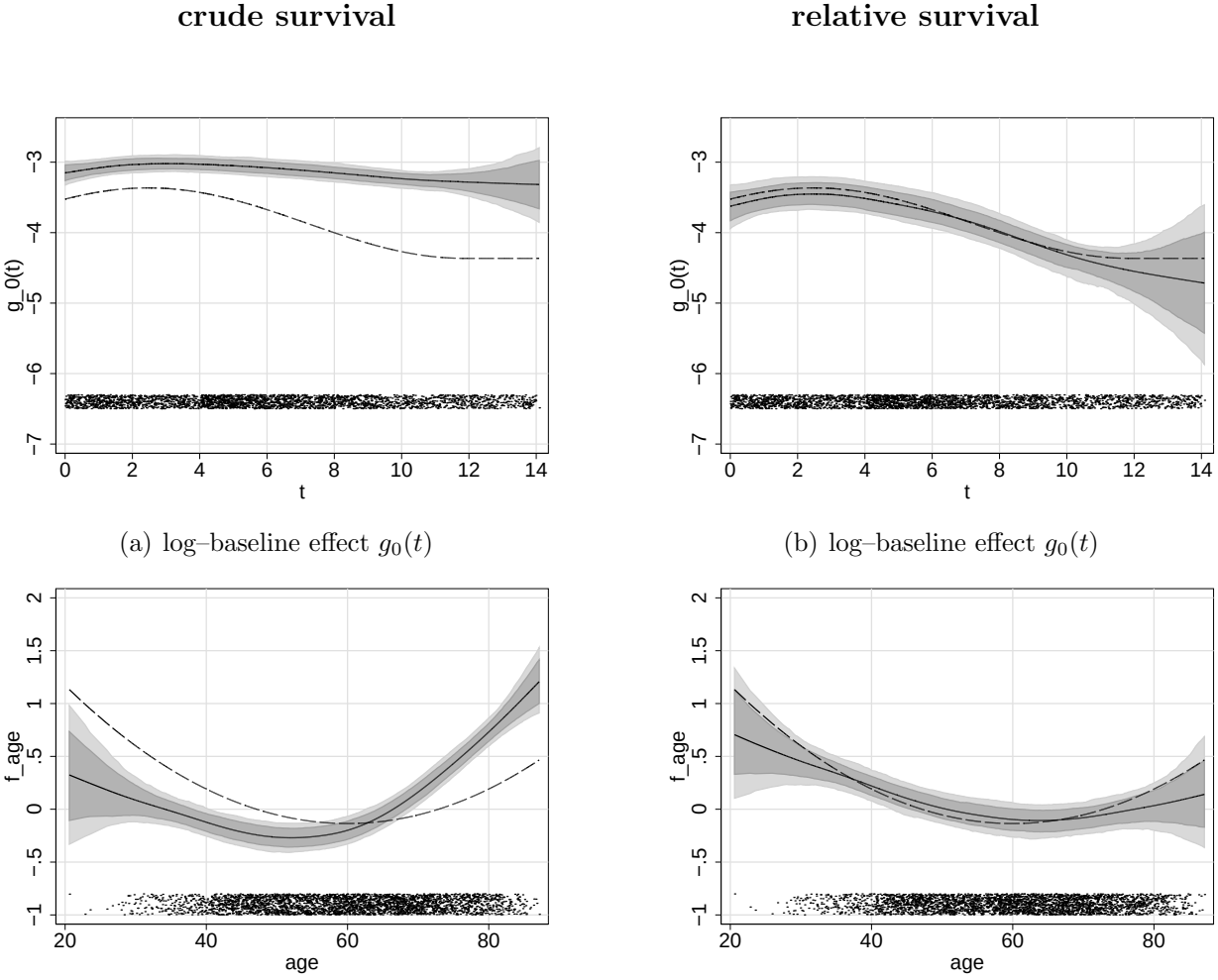
Now we re–analyze the generated data set with a relative survival model as described in (3.1). That is we assume a hazard rate as follows:

$$\begin{aligned}\lambda_i(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) &= \lambda_i^e(a_i + t) + \lambda_i^c(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) \\ &= \frac{\exp\left(\frac{a_i+t-30}{10}\right)}{2500} + \exp(g_0(t) + f_{age}(a_i) + \gamma_1 \texttt{meta1}_i + \gamma_2 \texttt{meta2}_i),\end{aligned}$$

where the disease related hazard rate λ_i^c is modelled as the total hazard rate λ_i was modelled before. However, the total hazard is now amended by the known natural mortality rate λ_i^e in order to account for cases of death that are not related to the disease of interest. As displayed in Figures 3.5(b) and (d) the true disease related log–baseline and the effect of age are now estimated quite satisfactorily, even though the effect of age is a bit too flat which might be due to the very small number of young patients. Also the fixed effects of $\texttt{meta1}_i$ and $\texttt{meta2}_i$ are estimated quite well with $\hat{\gamma_1} = 0.98$ and $\hat{\gamma_2} = 2.79$ (with standard deviations of 0.07 and 0.15, respectively).

3.6 Conclusion

In summary it can be ascertained that the simulation supports the usefulness of the relative survival approach since it yields results that are highly comparable to those of our application. As the simulation has shown, a model that does not account for natural mortality is not suitable for the identification of prognostic factors for disease specific mortality in cases where data on causes of death is not available since effects of covariates on natural mortality and effects on disease specific mortality intermix and can not be separated easily ex post.



(a) log–baseline effect $g_0(t)$

(b) log–baseline effect $g_0(t)$

(c) centered effect of age

(d) centered effect of age

Figure 3.5: Simulation: posterior means (solid line) together with pointwise 80% and 95% confidence intervals and true disease related effects (dashed lines) for the log–baseline effect including the intercept term (a,b) and the centered effect of age (c,d). Figures b and d result from a relative survival analysis.

Chapter 4

Multi–state models

4.1 Introduction

In the previous chapters we described methods for analyzing data, where only one type of event is considered. This chapter is concerned with extensions to more general event history data, that is ascertained by observing individuals over time and contains information on the times of occurrence of certain events and the types of events that occur.

In the simplest case one may distinguish between several distinct types of terminating events, i.e. from a statistical point of view each event represents a transition from a transient state to a certain absorbing state. Here just one transient state, but an arbitrary, finite number of absorbing states may be considered. Models for this type of data are referred to as competing risks models. In clinical studies, for example, the competing risks might be the diverse causes of death.

The most general case that we discuss is given by continuous–time multi–state models. Here the various events are considered as transitions from one state to another. A state structure specifies the diverse states (that might be absorbing or transient) and defines which transitions are possible. Each individual may experience a certain number of events over time, i.e. pass through the considered, possibly recurrent states, with transition times being arbitrary and measured on a continuous time–scale. Hence we consider individual counting processes instead of individual survival times. This type of data is for example given in clinical studies where the interest lies in analyzing transitions between different states of health. Note that survival data represent a special type of event history data

with just one type of event that is a transition from the only transient state to the only absorbing state.

Multi–state models are discussed widely in the literature. Andersen and Keiding (2002) provide a good overview over multi–state models with linear predictors in a frequentist setting. Fahrmeir and Klinger (1998) propose a nonparametric multiplicative multi–state hazard model that allows to model nonlinear functional forms of covariates and time–varying effects, with estimation being based on penalized likelihoods and smoothing splines. While several models for the analysis of spatially correlated survival data have been proposed in recent publications, spatial models have received far less attention in the more general setting of multi–state models.

Within this chapter we will illustrate how the Bayesian methods presented in Chapter 2 for analyzing extended Cox models are carried forward to continuous–time multi–state models. We present an approach where the hazard or transition rates for the particular events are modelled via independent structured additive predictors each including a nonparametrically modelled log–baseline effect as well as transition–specific effects of (time–independent or time–dependent) covariates with possibly linear, nonlinear, spatially correlated, time–varying or random effects. In principle different time scales, like e.g. time since an individual–specific initial point of time and duration in the current state could be considered as basic times for the various transition rates of a multi–state model. For simplification however, in what follows we consider time t since an individual–specific initial point of time as the basic time scale with every transition rate, while other time scales might be treated as time–dependent, but piecewise constant covariates.

The rest of this chapter is organized as follows. In Section 4.2 we will describe models, likelihood and priors for unknown functions and parameters. In Section 4.3 we comment on how the MCMC inference described in Section 2.3 for extended Cox models may be utilized with multi–state models. In Section 4.4 we illustrate our methods by applications to medical data on structural valve degeneration (SVD) of biological prostheses where reoperation and death without previous reoperation act as competing risks, and to sleep–electroencephalography data with multiple recurrent states of sleep.

4.2 Models, likelihood, priors and MCMC inference

Within this chapter we will present two alternative representations of multi–state data. We start off with a notation embedded in the counting process framework (Andersen, Borgan, Gill and Keiding 1993), which is quite common with multi–state data. The second notation is more closely related to the representation of survival data as introduced in Subsection 2.2.1.

Consider n individuals and let N_{hi}, for $h = 1, \ldots, H, i = 1, \ldots, n$, denote the counting processes for events of type h, where $N_{hi}(t)$ is the number of observed type h events experienced by the ith individual up to time t. We assume that individual intensity processes exist and have multiplicative structure

$$\alpha_{hi}(t) = Y_{hi}(t)\lambda_{hi}\{t; \boldsymbol{z}_{hi}(t), \boldsymbol{x}_{hi}(t), s_{hi}(t), \boldsymbol{v}_{hi}(t)\}, \tag{4.1}$$

where $Y_{hi}(t)$ are left–continuous 1–0 processes indicating whether or not individual i is at risk of experiencing a type h event just before time t. The individual type h hazard or transition rate λ_{hi} in (4.1) depends on t and on possibly transition–specific and time–dependent covariates. As in (2.4), the covariate vector $\boldsymbol{z}_{hi}(t)$ is assumed to have time–varying effects, $\boldsymbol{x}_{hi}(t)$ consists of continuous covariates with possibly nonlinear effects, s_{hi} denotes a spatial location and $\boldsymbol{v}_{hi}(t)$ comprises covariates with linear effects. Note that right censored survival data with lifetimes T_i, independent censoring times $C_i, i = 1, \ldots, n$, observed lifetimes $t_i = \min(T_i, C_i)$, and censoring indicators δ_i are a special case with $h = 1$, $N_i(t) = I(T_i \leq t, \delta_i = 1)$, $Y_i(t) = I(t_i \geq t)$ and $\lambda_i(t)$ as in (2.3) and (2.4).

The transition rate $\lambda_{hi}(t)$ for individual i is assumed to follow a multiplicative model

$$\lambda_{hi}(t) := \lambda_{hi}(t; \boldsymbol{z}_{hi}(t), \boldsymbol{x}_{hi}(t), s_{hi}(t), \boldsymbol{v}_{hi}(t)) = \exp(\eta_{hi}(t)), \tag{4.2}$$

with the general form of the predictor given by

$$\eta_{hi}(t) = g_{h0}(t) + \sum_{j=1}^{p} g_{hj}(t) z_{hij} + \sum_{j=p+1}^{p+q} f_{hj}(x_{hij}(t)) + f_{h,spat}(s_{hi}) + \boldsymbol{v}'_{hi}(t)\boldsymbol{\gamma}_h. \tag{4.3}$$

Here $g_{h0}(t) = \log(\lambda_{h0}(t))$ is the log–baseline effect for transition h, $g_{hj}(t)$ are time–varying effects of covariates $z_{hj}(t)$, $f_{hj}(x_{hj}(t))$ is the nonlinear effect of $x_{hj}(t)$, $f_{h,spat}$ is the spatially correlated effect of s_h, and $\boldsymbol{\gamma}_h$ is the vector of usual linear fixed effects.

As a further extension, i.i.d. random effects (also referred to as frailty effects) and random slopes could be introduced in (4.3), but we omit this here. For given predictors $\eta = \{\eta_{hi}, h = 1, \ldots, H, i = 1, \ldots, n\}$, the individual likelihood $L_i(\eta)$ and the likelihood $L(\eta)$ are given by

$$\begin{aligned} L_i(\eta) &= \prod_{h=1}^{H} \int_0^\infty \lambda_{hi}(s) dN_{hi}(s) \cdot \exp\left\{ - \int_0^\infty Y_{hi}(s) \lambda_{hi}(s) ds \right\} \\ L(\eta) &= \prod_{i=1}^{n} L_i(\eta). \end{aligned} \tag{4.4}$$

Note that the first integral in (4.4) always reduces to a sum because $N_{hi}(s)$ is a step function. Numerical problems arise in the evaluation of the second integral in (4.4). Again, only if time–varying functions in the predictor are step functions, this integral also reduces to a sum (compare Section 2.3). Otherwise numerical integration in form of the trapezoidal rule is employed as illustrated in Figure 1.6 in Chapter 1.

An alternative formulation of the likelihood, that shows the close connection to survival models more clearly, arises from considering multi–state data where for each individual i times of occurrences of certain events $t_{i1}, t_{i2}, \ldots, t_{in_i}$ (as well as possibly a left truncation time t_{i0}) and H–dimensional event–type indicators $\boldsymbol{\delta}_{i1}, \boldsymbol{\delta}_{i2}, \ldots, \boldsymbol{\delta}_{in_i}$ are given, that indicate which type of event occurred and are defined as follows

$$\delta_{ikh} = \begin{cases} 1 & \text{individual } i \text{ experienced a type } h \text{ event at time } t_{ik} \\ 0 & \text{else} \end{cases}$$

for $k = 1, \ldots, n_i$ and $h = 1, \ldots, H$. Note that $\boldsymbol{\delta}_{in_i} = (0, \ldots, 0)$ if an observation is right censored, i.e. if individual i is in a transient state at time t_{in_i} but for some reason the observation is discontinued at that point of time. Furthermore a state structure has to be given that defines which state transitions are possible and hence defines the risk processes $Y_{hi}(t)$. Via this kind of notation the individual likelihood may be alternatively written as

$$L_i(\eta) = \prod_{h=1}^{H} \prod_{k=1}^{n_i} \left[\lambda_{hi}(t_{ik})^{\delta_{ikh}} \cdot \exp\left\{ - \int_{t_{i,k-1}}^{t_{ik}} Y_{hi}(s) \lambda_{hi}(s) ds \right\} \right]. \tag{4.5}$$

From this equation it can be seen how the likelihood of such a multi–state model is multiplicatively composed of likelihood contributions of according survival models (for left truncated data) where one of the H events is modelled at each.

As regards assumptions about priors for parameters and functions and hyperpriors for variance components we may refer to Subsection 2.2.1 since we do not assume correlations of any kind between transition rates and the priors do not depend on the type h of transition (see Conclusion for a short discussion of this assumption). Hence with each single transition rate may be proceeded as described in Chapter 2 for the hazard rate of a survival model, i.e. each transition–specific log–baseline effect g_{h0} as well as every unknown function f_{hj} and g_{hj} might for example be modelled via a Bayesian P–spline, while diffuse priors are assumed for fixed effects parameters $\boldsymbol{\gamma}_h$ and MRF priors are our standard choice for structured spatial effects $f_{h,spat}$. In the framework of the generic notation as described for survival models in (2.6), after reindexing we can represent the predictor vectors $\boldsymbol{\eta}_h$=$(\eta_{h1}(t_{1,1}), \ldots, \eta_{h1}(t_{1,n_1}), \ldots, \eta_{hn}(t_{n,1}, \ldots, \eta_{hn}(t_{n,n_n}))'$ as

$$\boldsymbol{\eta}_h = \mathbf{V}_h \boldsymbol{\gamma}_h + \mathbf{Z}_{h0} \boldsymbol{\beta}_{h0} + \ldots + \mathbf{Z}_{hm_h} \boldsymbol{\beta}_{hm_h}. \tag{4.6}$$

Priors for functions and spatial components are then defined by suitable design matrices $\mathbf{Z}_{hj}$, $h = 1, \ldots, H$, $j = 0, \ldots, m_h$, and a prior for each corresponding parameter vector $\boldsymbol{\beta}_{hj}$. The general form of a prior for $\boldsymbol{\beta}_{hj}$ is given by

$$p(\boldsymbol{\beta}_{hj}|\tau^2_{hj}) \quad \propto \quad \tau_{hj}^{-r_{hj}} \exp\left(-\frac{1}{2\tau^2_{hj}} \boldsymbol{\beta}'_{hj} \boldsymbol{K}_{hj} \boldsymbol{\beta}_{hj}\right), \tag{4.7}$$

where $\boldsymbol{K}_{hj}$ is an adequate precision or penalty matrix of $\text{rank}(\boldsymbol{K}_{hj}) = r_{hj}$, shrinking parameters towards zero or penalizing too abrupt jumps between neighboring parameters. We assign inverse Gamma priors $IG(a_{hj}; b_{hj})$

$$p(\tau^2_{hj}) \quad \propto \quad \frac{1}{(\tau^2_{hj})^{a_{hj}+1}} \exp\left(-\frac{b_{hj}}{\tau^2_{hj}}\right) \tag{4.8}$$

to all variances, with $a_{hj} = b_{hj} = 0.001$ being our standard choice.

4.3 Markov Chain Monte Carlo inference

As with survival models, full Bayesian inference via MCMC simulation is again based on updating full conditionals of single parameters or blocks of parameters (each with parameters corresponding to the same transition rate λ_{hi}), given the rest of the data. For

updating the parameter vectors $\boldsymbol{\beta}_{hj}$, which correspond to the time–independent functions f_{hj}, as well as spatial effects $\boldsymbol{\beta}_h^{spat}$, which correspond to spatial functions $f_{h,spat}$, fixed effects $\boldsymbol{\gamma}_h$ and random effects $\boldsymbol{b}_h$, we use the slightly modified version of the MH–algorithm based on iteratively weighted least squares (IWLS) proposals, which is described in Section 2.3 and the Appendix, respectively, for survival models. The full conditional of a parameter vector $\boldsymbol{\beta}_{hj}$ with prior $p\left(\boldsymbol{\beta}_{hj}|\tau_{hj}^2\right)$ is for example given by

$$\begin{aligned}
p(\boldsymbol{\beta}_{hj}|\cdot) &\propto L(\boldsymbol{\beta}_{hj}) \cdot p(\boldsymbol{\beta}_{hj}|\tau_{hj}^2) \\
&= \prod_{i=1}^{n} \prod_{\tilde{h}=1}^{H} \prod_{k=1}^{n_i} \left[\lambda_{\tilde{h}i}(t_{ik})^{\delta_{ik\tilde{h}}} \cdot \exp\left\{ - \int_{t_{i,k-1}}^{t_{ik}} Y_{\tilde{h}i}(s) \lambda_{\tilde{h}i}(s) ds \right\} \right] \cdot p\left(\boldsymbol{\beta}_{hj}|\tau_{hj}^2\right) \\
&\propto \prod_{i=1}^{n} \prod_{k=1}^{n_i} \left[\lambda_{hi}(t_{ik})^{\delta_{ikh}} \cdot \exp\left\{ - \int_{t_{i,k-1}}^{t_{ik}} Y_{hi}(s) \lambda_{hi}(s) ds \right\} \right] \cdot p\left(\boldsymbol{\beta}_{hj}|\tau_{hj}^2\right) \\
&= \prod_{i=1}^{n} \prod_{k=1}^{n_i} \left[L_{hik}\left(\boldsymbol{\beta}_{hj}\right) \right] \cdot p\left(\boldsymbol{\beta}_{hj}|\tau_{hj}^2\right),
\end{aligned}$$

at which the second proportionality holds because the transition rates $\lambda_{\tilde{h}i}, \tilde{h} \neq h$ do not depend on $\boldsymbol{\beta}_{hj}$. Note that for $Y_{hi} = 1$ the likelihood contribution L_{hik} has the same structure as an individual likelihood contribution L_i for a left–truncated survival time (compare equations (1.6) and (2.5)). For $Y_{hi} = 0$ it follows that $\delta_{ikh} = 0$, since a type h transition can only be observed if the individual is at risk for a type h transition, i.e. if $Y_{hi} = 1$. Hence $Y_{hi} = 0$ implies that $L_{hik} = 1$. As a consequence of these insights IWLS proposals for the parameter vectors $\boldsymbol{\beta}_{hj}$ may be derived in the same manner as described in Section 2.3 and the Appendix, respectively. Thus, a new value $\boldsymbol{\beta}_{hj}^p$ is proposed by drawing a random sample from a high dimensional multivariate Gaussian distribution $q(\boldsymbol{\beta}_{hj}^c, \boldsymbol{\beta}_{hj}^p)$ which is obtained from a quadratic approximation of the log–likelihood by a second order Taylor expansion with respect to the current value of the chain $\boldsymbol{\beta}_{hj}^c$. The precision matrix and mean of this proposal distribution are given by

$$\boldsymbol{P}_{hj} = \boldsymbol{Z}'_{hj} \boldsymbol{W}_h(\boldsymbol{\beta}_{hj}^c) \boldsymbol{Z}_{hj} + \frac{1}{\tau_{hj}^2} \boldsymbol{K}_{hj}, \quad \boldsymbol{m}_{hj} = \boldsymbol{P}_{hj}^{-1} \boldsymbol{Z}'_{hj} \boldsymbol{W}_h(\boldsymbol{\beta}_{hj}^c)(\tilde{\mathbf{y}}_h - \tilde{\boldsymbol{\eta}}_h).$$

Here, $\tilde{\boldsymbol{\eta}}_h = \boldsymbol{\eta}_h - \mathbf{Z}_{hj}\boldsymbol{\beta}_{hj}$, $\boldsymbol{W}_h(\boldsymbol{\beta}_{hj}^c) = diag(w_{h,1,1}, \ldots, w_{h,n,n_n})$ is the weight matrix for IWLS with weights calculated from the current state $\boldsymbol{\beta}_{hj}^c$ as follows

$$w_{hik} = \int_{t_{i,k-1}}^{t_{ik}} Y_{hi}(s)\lambda_{hi}(s)ds, \quad i = 1, \ldots, n, \quad k = 1, \ldots, n_i.$$

The vector of working observations $\tilde{\mathbf{y}}_h$ is given by

$$\tilde{\mathbf{y}}_h = \mathbf{W}_h^{-1}(\boldsymbol{\beta}_{hj}^c)\Delta_h - \mathbb{1} + \boldsymbol{\eta}_h$$

with $\Delta_h = (\delta_{1,1,h}, \ldots, \delta_{n,n_n,h})'$. The proposed vector $\boldsymbol{\beta}_{hj}^p$ is accepted as the new state of the chain with probability

$$\alpha(\boldsymbol{\beta}_{hj}^c, \boldsymbol{\beta}_{hj}^p) = \min\left(1, \frac{p(\boldsymbol{\beta}_{hj}^p \mid \cdot)q(\boldsymbol{\beta}_{hj}^p, \boldsymbol{\beta}_{hj}^c)}{p(\boldsymbol{\beta}_{hj}^c \mid \cdot)q(\boldsymbol{\beta}_{hj}^c, \boldsymbol{\beta}_{hj}^p)}\right).$$

For the parameters $\boldsymbol{\beta}_{hj}$ corresponding to the functions $g_{h0}(t), ..., g_{hp}(t)$ depending on time t, we again adopt the computationally faster MH–algorithm based on conditional prior proposals, that only requires evaluation of the log–likelihood, not of derivatives (see Fahrmeir and Lang (2001a) for details).

The full conditionals for the variance parameters τ_{hj}^2 are (proper) inverse Gamma with parameters

$$a'_{hj} = a_{hj} + \frac{1}{2}r_{hj} \quad \text{and} \quad b'_{hj} = b_{hj} + \frac{1}{2}\boldsymbol{\beta}'_{hj}\boldsymbol{K}_{hj}\boldsymbol{\beta}_{hj},$$

Updating can be done by simple Gibbs steps, drawing random numbers directly from the inverse Gamma densities.

4.4 Application

4.4.1 Biological valve prostheses

Our first application is on data from 455 patients who underwent biological mitral valve replacement (MVR) at the German Heart Center in Munich between 1974 and 2000. This data has been analyzed before by Kaempchen et al. (2003) and more details about the medical background may be found therein. The aim of our analysis is to assess the influence of several covariates on reoperation free survival, at which death and reoperation due to a failure of the biological valve are considered as competing risks. The state structure of

this competing risks model is illustrated in Figure 4.1. Note that death after reoperation is not of interest with this analysis and is therefore not considered. Within the observation period 212 patients died without a previous reoperation and 125 patients had to undergo a reoperation; the remaining 118 observations are right–censored. Covariates that are given include the `sex` and the `age` of patients at valve replacement as well as the diagnosis (insufficiency, narrowness or malformation of the mitral valve) and information on the initial valve replacement, namely the `date` of implantation and whether or not an additional aortocoronary venous bypass (ACVB) was accomplished. Including all the covariates, the transition rates λ_{ri} (reoperation) and λ_{di} (death without previous reoperation) are modelled as follows

$$\begin{aligned}
\lambda_{ri}(t) &= \exp\left[g_{r0}(t) + f_{r,age}(\texttt{age}_i) + f_{r,date}(\texttt{date}_i)\right. \\
&\quad \left. +\gamma_{r1} \cdot \texttt{sex}_i + \gamma_{r2} \cdot \texttt{diag1}_i + \gamma_{r3} \cdot \texttt{diag2}_i + \gamma_{r4} \cdot \texttt{acvb}_i\right] \\
\lambda_{di}(t) &= \exp\left[g_{d0}(t) + f_{d,age}(\texttt{age}_i) + f_{d,date}(\texttt{date}_i)\right. \\
&\quad \left. +\gamma_{d1} \cdot \texttt{sex}_i + \gamma_{d2} \cdot \texttt{diag1}_i + \gamma_{d3} \cdot \texttt{diag2}_i + \gamma_{d4} \cdot \texttt{acvb}_i\right],
\end{aligned}$$

where t is time since valve replacement, g_{h0}, $h = r, d$ are the log–baseline effects, $f_{h,age}$ and $f_{h,date}$ are nonlinear effects of the `age` of a patient at valve replacement and of the `date` of valve replacement, respectively. All of these possibly nonlinear effects are modelled via cubic P–splines with 20 knots. The remaining covariates are dummy–coded: `sex` $= 1$ for female, and `sex` $= 0$ for male, `diag1` $= 1$ if the patient was diagnosed with an insufficiency of the mitral valve, `diag1` $= 0$ else, `diag2` $= 1$ if the patient was diagnosed with a narrowness of the mitral valve, `diag2` $= 0$ else and `acvb` $= 1$ if an additional ACVB was accomplished, `acvb` $= 0$ else. Diffuse priors were assumed for the parameters γ.

Figure 4.2 displays the estimated nonlinear effects. Concerning the risk of a reoperation we observe that the risk is highest between ca. 9.5 and 16.5 years after the initial valve replacement, while it is lower in the first 9.5 years and after 16.5 years of reoperation free survival, at which confidence intervals become quite broad for $t > 20$ years. The risk of death without previous reoperation on the other hand is very high directly after the valve replacement and is steeply decreasing within the first 2 years and slowly increasing from that time on. The initially high risk might arise from consequences of the operation or incompatibilities, while the slow increase for $t > 2$ is due to aging. The effect of age at

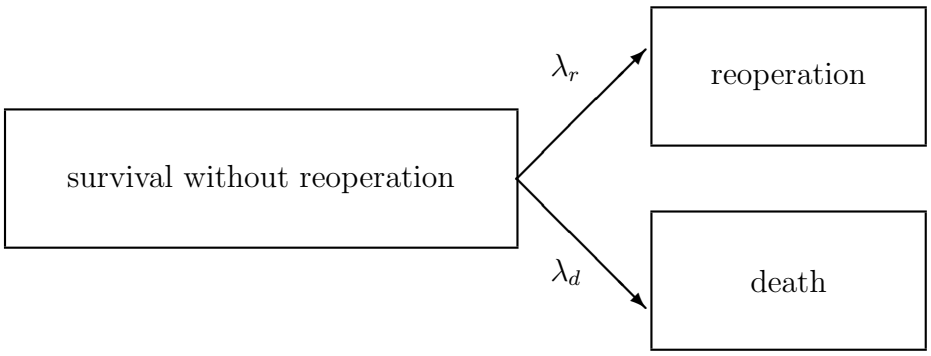
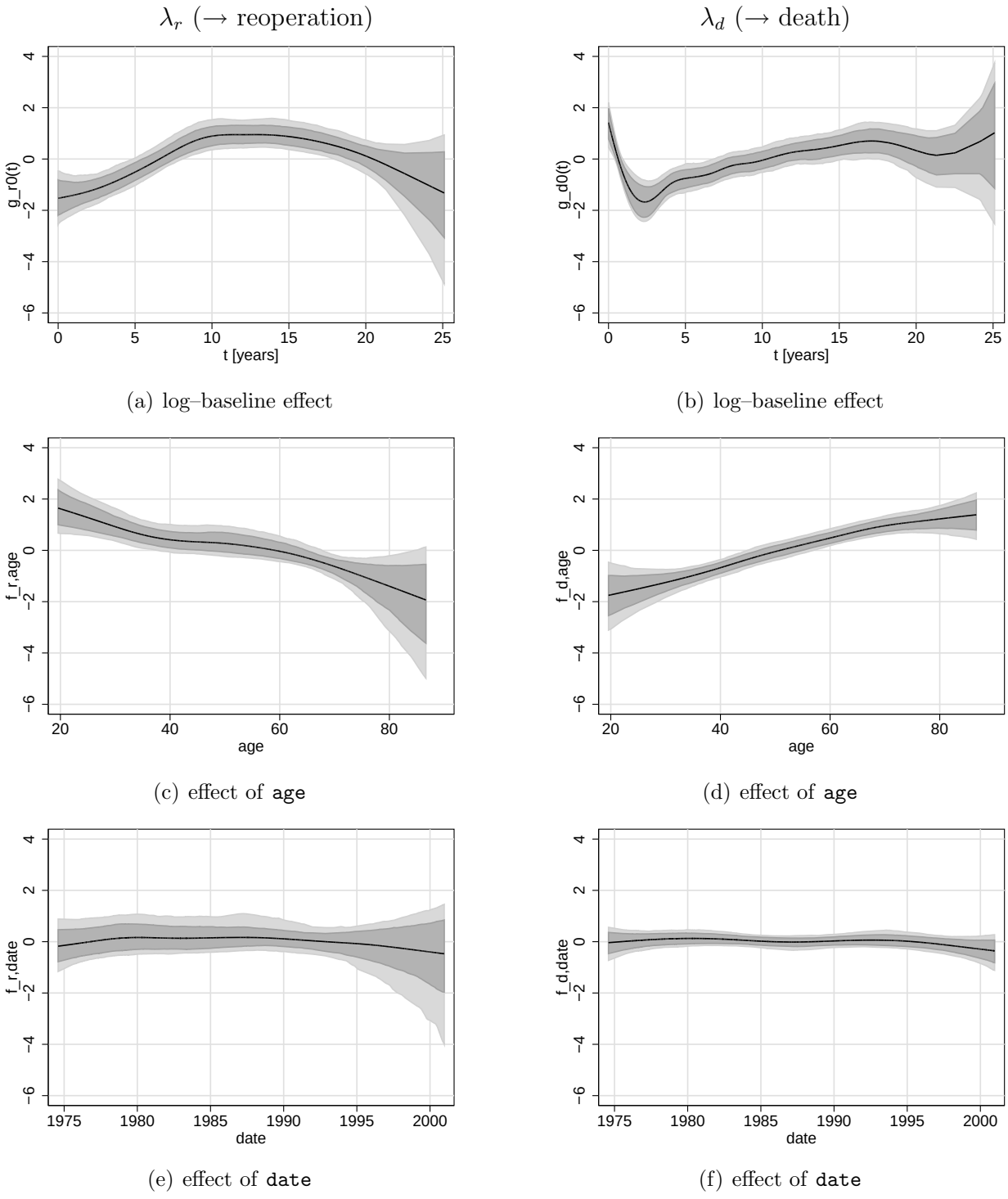


Figure 4.1: State structure of the competing risks model for the analysis of reoperation free survival after biological mitral valve replacement.

valve replacement turns out to be rather linear with both predictors. While the risk of a reoperation is decreasing with increasing age it is vice versa with the risk of death without previous reoperation. This result seems quite perspicuous. The lifespan of patients that got a biological valve prostheses at the age of 80 or more, for example, is likely to be shorter than the endurance of the valve prostheses. The date of the valve replacement does not seem to have an influence on any of the two risks. In the first instance this appears to be very disappointing since it would mean there has been no medical progress within 26 years. However, medical progress involved that over the years more and more patients with very severe illnesses could be operated, that would not have been operated in earlier years since there would have been no chances of success. Consequently the composition of patient groups with respect to the severity of the illness is heterogeneous over the years. Hence our result is likely to be due to the fact that the severity of the illness is not considered with our analysis as it is only recorded with 116 out of those 455 patients. Concerning the fixed effects we observe that an additional ACVB reduces the risk of a reoperation significantly on the basis of a 80% significance level ($\hat{\gamma}_{r4} = -1.32$, with a standard deviation of 0.79), while the remaining fixed effects are not significant (compare Table 4.1).

4.4.2 Human sleep processes

Our second application is about analyzing human sleep processes. The data set arises from recordings of electroencephalographic (EEG) data during one night taken for a group of 27



(a) log–baseline effect

(b) log–baseline effect

(c) effect of `age`

(d) effect of `age`

(e) effect of `date`

(f) effect of `date`

Figure 4.2: MVR data: posterior mean together with 80% and 95% credible intervals of the centered log–baseline effects (a) and (b), the effects of `age` (c) and (d), and the effects of `date` (e) and (f) on the competing risks reoperation (left panel) and death (right panel).

	λ_r		λ_d	
	mean	std. dev.	mean	std. dev.
`sex`	0.17	0.22	-0.01	0.15
`diag1`	-0.29	0.25	0.14	0.18
`diag2`	0.07	0.23	-0.13	0.21
`acvb`	-1.32	0.79	0.11	0.20

Table 4.1: MVR data: posterior mean estimations of fixed effects γ together with standard deviations.

patients at the Max–Planck–Institut für Psychiatrie in Munich. Sleep–EEG data describe the nocturnal sleep rhythm, usually classified in several stages such as awake, non–rapid eye movement (NREM) and rapid eye movement (REM). Such sleep states indicating the depth of sleep are recorded every 30 seconds. In addition, secretion of several hormones is measured every 10, 20 or 30 minutes. The hormone cortisol is for example supposed to be interrelated with the sleep structure. Figure 4.3 exemplarily displays the processes of sleep states and nocturnal cortisol secretion for two patients. Without any kind of smoothing it is difficult to identify typical sleep patterns. Furthermore individual–specific sleeping customs must be considered in order to detect population effects.

Besides a dynamic analysis of the transition intensities between the distinct states, a main concern is to investigate the question whether high cortisol concentrations have a positive effect on the propensity to REM sleep, which has been hypothesized in simple correlation and variance analyses. It is also of interest to allow this effect to vary over night. Due to the very low number of direct transitions from AWAKE to REM, we consider only a somewhat reduced state structure, which is illustrated in Figure 4.4 and comprises the following four types of events

$h = 1$ transition from AWAKE to SLEEP, (AS)
$h = 2$ transition from SLEEP to AWAKE, (SA)
$h = 3$ transition from REM to NREM, (RN)
$h = 4$ transition from NREM to REM, (NR)

where SLEEP implies REM and NREM sleep states. In principle it might be of interest

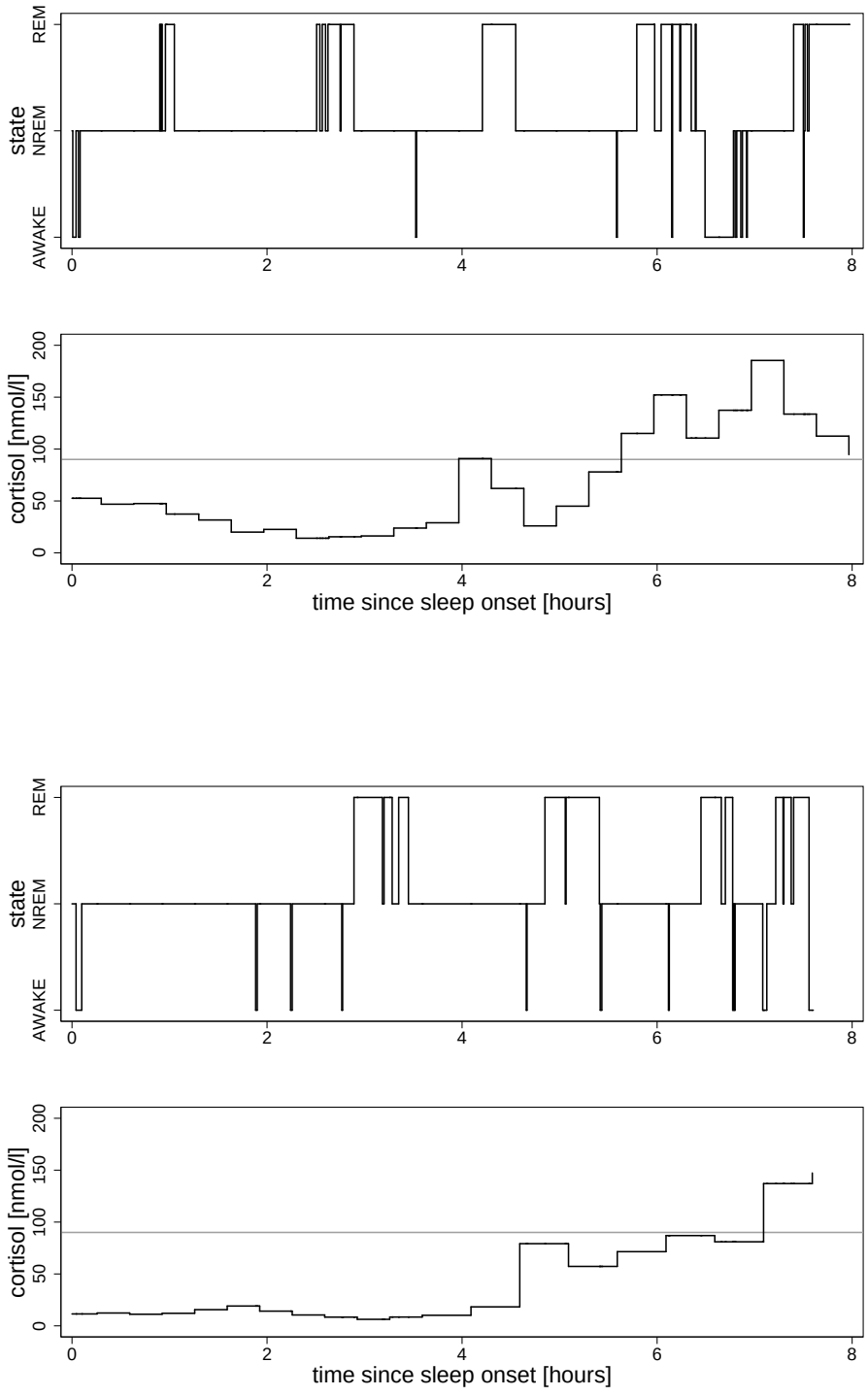


Figure 4.3: Individual sleep processes for two patients ($i = 1$ and $i = 21$, respectively) together with the corresponding cortisol secretion.

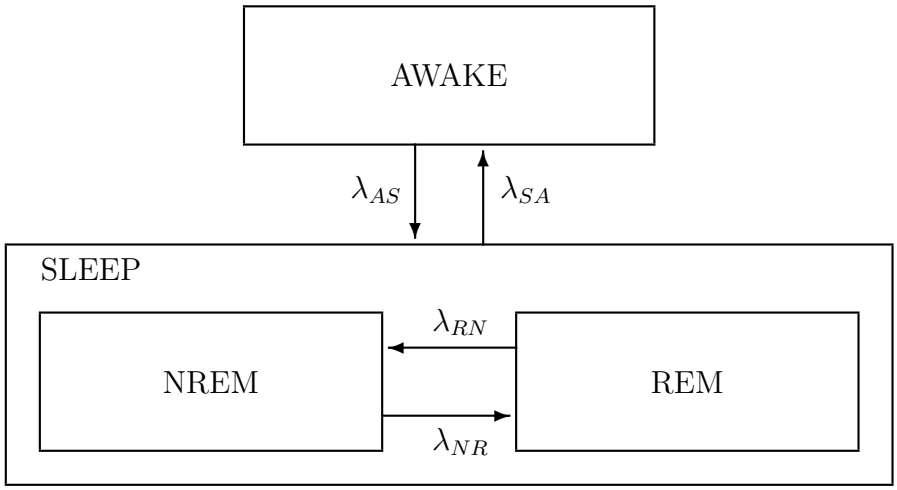


Figure 4.4: State structure for the analysis of human sleep processes.

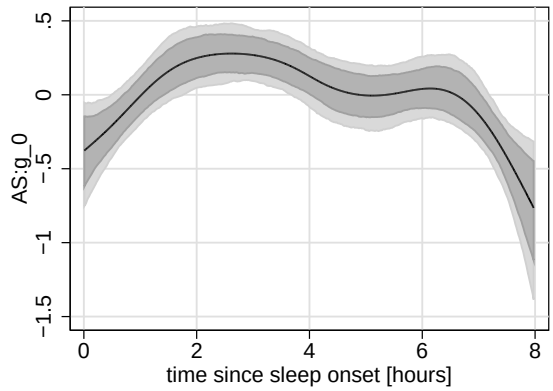
to separately analyze the transitions AWAKE → REM, AWAKE → NREM and NREM → AWAKE, REM → AWAKE, respectively, but our data pool is not sufficient for such a detailed analysis. In order to achieve some synchronization we take time t since sleep onset as basic time scale. For the analysis of the possibly time–varying effect of high cortisol secretion on the transition intensity from NREM to REM, we generate the time–dependent dummy coded covariate $\mathsf{c}_i(t)$, which takes the value one if the concentration of cortisol is higher than 90 nmol/l at time t with patient i and the value zero otherwise. Observed concentrations of cortisol range from 1 to 450 nmol/l, with 90 nmol/l being the 70% quantile. Based on the previous considerations, we analyze a multi–state model with the following four transition rates

$$\begin{aligned} \lambda_{hi} &= \exp\left(g_{h0}(t) + b_{hi}\right), \quad h = AS, SA, RN \\ \lambda_{hi} &= \exp\left(g_{h0}(t) + \mathsf{c}_i(t) \cdot g_{h1}(t) + b_{hi}\right), \quad h = NR \end{aligned}$$

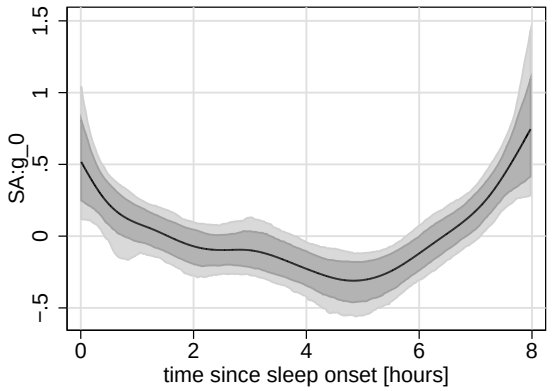
at which again cubic P–spline priors are assumed for the transition–specific log–baseline effects $g_{h0}(t)$, $h = AS, SA, RN, NR$, as well as for the time–varying effect of a high cortisol level on the transition from NREM to REM $g_{NR,1}(t)$. The term b_{hi} denotes transition– and patient–specific random effects with i.i.d. Gaussian priors $b_{hi} \sim N(0, \tau^2_{hb})$.

Estimated results for the time–varying baseline effects $g_{h0}(t)$ for the transitions $h =$

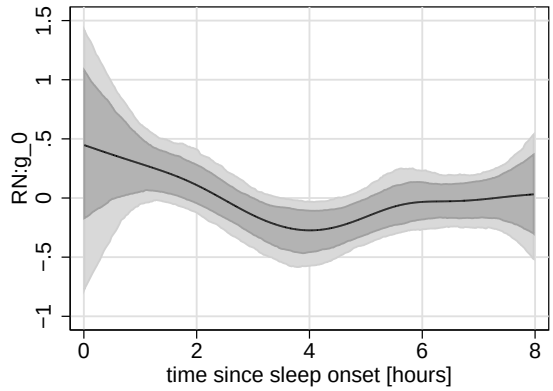
AS, SA, RN, NR and the time–varying effect of a high cortisol level on the transition from NREM to REM $g_{NR,1}(t)$ are displayed in Figure 4.5. As was to be expected the tendency to fall asleep again is particularly low for patients who awake in the beginning and at the end of the night, i.e. within the time spans $t < 1$ and $t > 7$, respectively. We further conclude from our results that the propensity to fall asleep is notably high around $t \in [2, 3.3]$ and seems to have a local minimum around five hours after sleep onset. By contrast, the tendency to wake up is roughly u–shaped and rather high in the beginning and especially high at the end of the night while it is lowermost around $t \in [4, 6]$. The intensity for the transition from REM to NREM sleep is highest directly after sleep onset and is then decreasing until $t \approx 4$, increasing again until $t \approx 6$ and staying rather constant from that time on. Concerning the inverse transition from NREM to REM sleep, the log–baseline effect $g_{NR,0}(t)$ marks the effect for a low level of cortisol, while $g_{NR,1}(t)$ describes deviations from this effect if the level of cortisol is high, i.e. exceeds 90 nmol/l. In case the cortisol level is low, the intensity for a transition from NREM to REM is initially very low, but steeply increasing within the first hour after initial sleep onset followed by some ups and downs with peaks at $t \approx 1.3$, $t \approx 3.0$, $t \approx 4.9$ and (possibly) $t \approx 7.8$, i.e. we observe a cyclic developing with two pronounced peaks in the first half of the night and two poor peaks in the second half of the night. Since high levels of cortisol appear very rarely within the first hours after sleep onset, the (pointwise) credible intervals for the time–varying effect $g_{NR,1}(t)$ of $\mathsf{c}_i(t)$ become quite broad for $t < 2$. Hence we can not draw any conclusions for this time span. Figure 4.5 f) however, which displays the time–varying effect $g_{NR,1}(t)$ for $t > 4$, exhibits an increased propensity to REM sleep for a time span around six hours after sleep onset. This is to say that our analysis only supports the hypothesis posted above (high cortisol concentrations have a positive effect on the propensity to REM sleep) for a time span around $t \in [5.5, 6.8]$. The estimated individual– and transition–specific random effects are displayed in Figure 4.6. There are several persons that show especially high or low tendencies for one or more transitions. Patient $i = 5$ for example has an exceptionally high tendency to awake, coming along with an exceptionally low tendency to fall asleep and a low propensity to REM sleep. The individual sleep process of patient $i = 5$ is displayed in Figure 4.7 and supports those results since it clearly differs from the prevailing sleep patterns as exemplarily displayed in Figure 4.3.



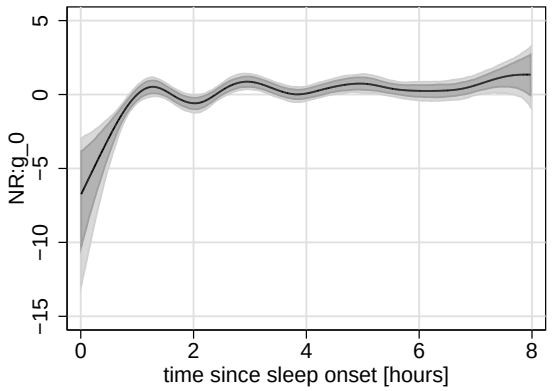
(a) AWAKE → SLEEP



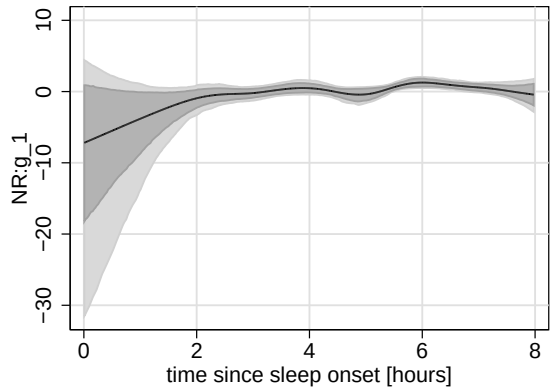
(b) SLEEP → AWAKE



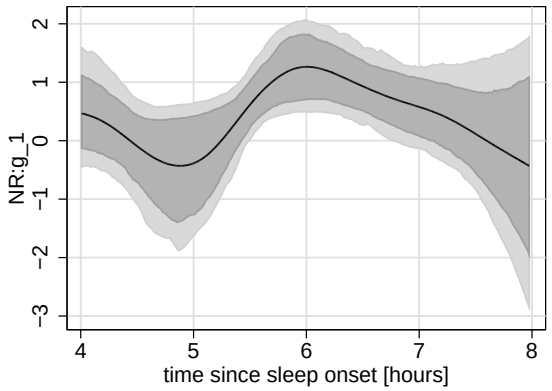
(c) REM → NREM



(d) NREM → REM

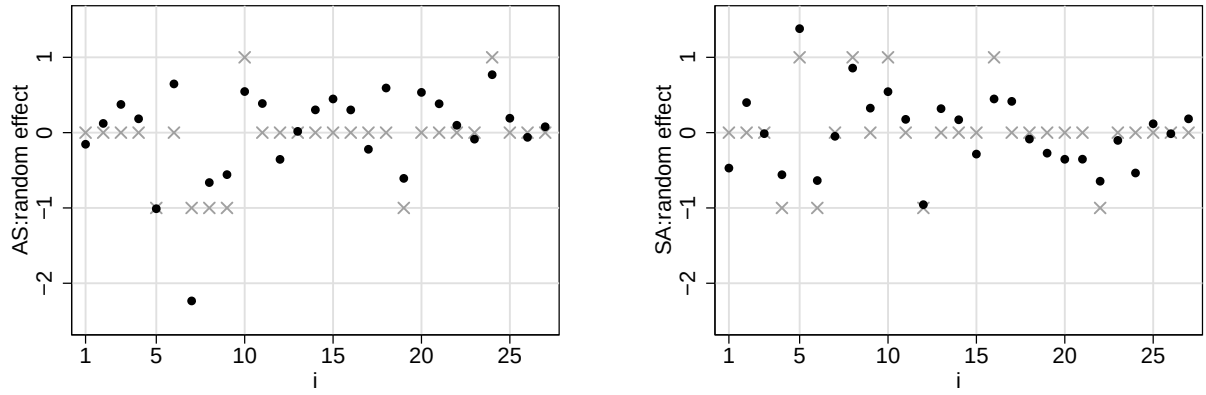


(e) NREM → REM



(f) NREM → REM

Figure 4.5: Human Sleep Processes: Posterior mean estimates for the time–dependent effects $g_{h0}(t), h = AS, SA, RN, NR$ and $g_{NR,1}(t)$ together with 80% and 95% credible intervals.



(a) AWAKE → SLEEP

(b) SLEEP → AWAKE

(c) REM → NREM

(d) NREM → REM

Figure 4.6: Human Sleep Processes: Posterior mean estimates (black dots) of the transition– and individual–specific random effects b_{hi} for $h = AS, SA, RN, NR$ and $i = 1, \ldots, 27$. Grey crosses denote significance on a 95% level with -1: significant negative effect, 0: no significant effect and +1: significant positive effect.

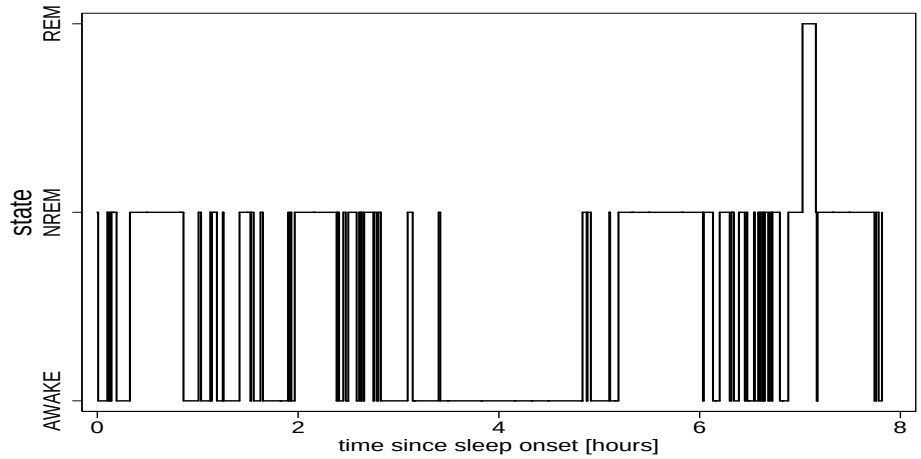


Figure 4.7: Individual sleep process for patient $i = 5$.

4.5 Conclusion

Within this chapter we have shown how the geoadditive survival models presented in Chapter 2 are generalized to multiplicative continuous–time multi–state models. Our approach allows the estimation of transition–specific nonlinear log–baseline effects as well as time–varying effects of covariates, nonlinear effects of continuous covariates and an appropriate consideration of unobserved unit– or cluster–specific and spatial heterogeneity.

So far we have only considered transition–specific effects and did not assume any correlation structure between transition rates. With some applications however, it might make sense to assume that unit– or cluster–specific random effects or structured spatial effects are correlated across (some or all) transition rates or are even transition–independent. Extensions towards this aspect are topics of future work.

Chapter 5

Bayesian survival and multi–state analysis with BayesX: a tutorial

All models presented in this thesis are implemented in the statistical software package *BayesX*. The focus of this chapter is to demonstrate how complex survival data and multi–state data may be analyzed within *BayesX* based on MCMC techniques. For this purpose the estimation of some of the survival models presented in Subsection 2.5.3 to analyze waiting times on CABG as well as the estimation of the relative survival model presented in Section 3.5 on the basis of simulated breast cancer data, and the estimation of the multi–state model presented in Subsection 4.4.2 to analyze human sleep processes are described in detail. For a description of the data sets we refer to the according subsections. Note that in addition to MCMC techniques *BayesX* also provides restricted maximum likelihood (REML) techniques as described in Kneib (2006) for the estimation of crude survival models and multi state models.

This chapter is organized as follows. After some comments on the overall capabilities of *BayesX* and the general structure of this software package given in Section 5.1 and in Section 5.2, respectively, we start with the analysis in Section 5.3, which is concerned with a description on how to create a *dataset object* to incorporate, handle and manipulate the data. Since we want to estimate a spatial effect of the ward with the CABG data, we need the boundaries of the districts to compute the neighborhood information of the map of London and Essex. This information will be stored in a *map object*. Section 5.4 describes how to create and handle these objects. Estimation of the regression models is carried

out in Section 5.5 using *bayesreg objects*. Section 5.6 describes post estimation commands which can be used to investigate the sampling paths and the autocorrelation functions of the estimated parameters.

5.1 BayesX

BayesX is a public domain software package for performing complex full and empirical Bayesian inference to estimate flexible regression models with structured additive predictors. Functions for handling and manipulating data sets and geographical maps, and for visualizing results are added for convenient use. *BayesX* is available at

`http://www.stat.uni-muenchen.de/~bayesx`

An overview over the capabilities of *BayesX* is given in Brezger, Kneib and Lang (2005). For more detailed information on all available features and the methodological background see the manuals that are provided in addition to the software *BayesX* and the references given therein.

5.2 Getting started

After having started *BayesX*, a main window with four sub–windows appears on the screen. These are a *command window* for entering and executing code, an *output window* for displaying results, a *review window* for easy access to past commands, and an *object browser* that displays all objects currently available.

BayesX is object oriented although the concept is limited, i.e. inheritance and other concepts of object oriented languages like C++ or S–plus are not supported. For every object type a number of object–specific methods may be applied to a particular object. The syntax for generating a new object in *BayesX* is

> *objecttype objectname*

where *objecttype* is the type of the object, e.g. `dataset`, and *objectname* is the name to be given to the new object.

5.3 Dataset objects

In a first step we read the available data set information into *BayesX*. This is done by creating three *dataset objects* named `cabg`, `cancer` and `sleep` for the CABG data, the (simulated) breast cancer data and the human sleep data, respectively, by typing:

```
> dataset cabg

> dataset cancer

> dataset sleep
```

in the *command window*. We store the data in `cabg`, `cancer` and `sleep` using the method `infile`. If the data is provided in the external ASCII files `c:\data\cabg.raw`, `c:\data\cancer.raw` and `c:\data\sleep.raw`, respectively, we may type

```
> cabg.infile using c:\data\cabg.raw

> cancer.infile using c:\data\cancer.raw

> sleep.infile using c:\data\sleep.raw
```

Note, that this command supposes that the variable names are given in the first row of the according external file. In case the variable names are not given in the file we would have to supply them right after the keyword `infile`. If a data set has more than 10000 observations it is recommended to set the option `maxobs` to the according number of rows. This option allows *BayesX* to allocate enough memory to store all the data right from the start, which speeds up the execution time of the `infile` command.

After having read in the data set information we can inspect the data visually. Executing the command

```
> cabg.describe
```

for example opens an *object–viewer* window containing the according CABG data in form of a spreadsheet. This can also be achieved by double–clicking on the according *dataset object* in the *object browser*.

Further methods allow to examine the variables in the *dataset object*. For a categorial variable the `tabulate` command may be used to produce a frequency table and for continuous variables the `descriptive` command prints several characteristics of the variable in the *output window*.

There are also methods to manipulate variables and generate new variables in a *dataset object*. Assume for example that `cagb` includes the categorical variable `numdv` that takes the values 1,2 and 3 and indicates the number of diseased vessels. Then the dummy variables `dv2` and `dv3` that are used for the estimation may be created and added to `cabg` using method `generate`. This might be done by executing the following commands

```
> cabg.generate dv2=0

> cabg.replace dv2=1 if numdv=2

> cabg.generate dv3=0

> cabg.replace dv3=1 if numdv=3
```

or in condensed form by executing the commands

```
> cabg.generate dv2=(numdv=2)

> cabg.generate dv3=(numdv=3)
```

Here `(numdv=2)` may be interpreted as the (row–wise) query "is `numdv` equal to 2 or not?" (written as `numdv==2` with some programming languages). Hence the first command causes *BayesX* to add a new covariate `dv2` to the *dataset object* `cabg`, that takes the value TRUE coded as 1 if the corresponding row of `numdv` equals 2 and the value FALSE coded as 0 otherwise.

5.4 Map objects

In the following we want to estimate a spatially correlated effect of the ward a patient with coronary artery disease lives. Therefore we need the boundaries of the wards in London and Essex to compute the neighborhood information of the map of this part of Great Britain. We therefore create a *map object*

```
> map m
```

and read in the boundaries using the `infile` command of *map objects*:

```
> m.infile using c:\data\LondonEssex.bnd
```

Having read in the boundary information, *BayesX* automatically computes the neighborhood matrix of the map. The file following the keyword `using` is assumed to contain the boundaries in form of closed polygons. To give an example we print a small part of the boundary file of London and Essex. The map corresponding to the section of the boundary file can be found in Figure 5.1.

```
      ⋮
"8849",37
532351,181179
532407,181166
532404,181147
532399,181143
532399,181136
532409,181131
532412,181116
532418,181112
532424,181109
532446,181106
532446,181082
532463,181082
532511,181083
532532,181082
532528,181060
532530,181050
532558,181064
532579,181072
532572,181051
532571,181045
532563,181013
532561,180999
532608,180984
532589,180926
532502,180952
532491,180920
532445,180932
532448,180959
532450,180969
532383,180991
```

```
532373,180941
532345,180944
532310,180952
532320,181015
532324,181059
532350,181163
532351,181179
```

⋮

For each region of the map the boundary file must contain the identifying name of the region, the number of vertices the polygon consists of and the vertices of the polygons that form the boundary of the region. The first line always contains the region code surrounded by quotation marks and the number of vertices the polygon of the region consists of. Note that the first vertex (532351,181179 with our example) has to be repeated at the end to obtain a closed polygon and hence the number of vertices of a pentagon would for example be 6. The region code and the number of vertices must be separated by a comma. The subsequent lines contain the vertices that are to be connected by straight lines and thus form the boundary of the region. The vertices are represented by the according coordinates, which must be separated by a comma. Compare Chapter 5 of the complete *BayesX* manual for a detailed description of some special cases, e.g. regions divided into subregions.

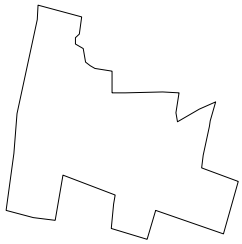


Figure 5.1: Corresponding graph of the section of the boundary file

Map objects may be visualized using method `describe`:

```
> m.describe
```

resulting in the graph shown in Figure 5.2. Additionally, `describe` prints further information about the *map object* in the *output window* including the name of the object, the number of regions, the minimum and maximum number of neighbors and the bandwidth of the corresponding adjacency or neighborhood matrix:

```
MAP m
Number of regions: 488
Minimum number of neighbors: 1
Maximum number of neighbors: 14
Bandsize of corresponding adjacency matrix: 39
```

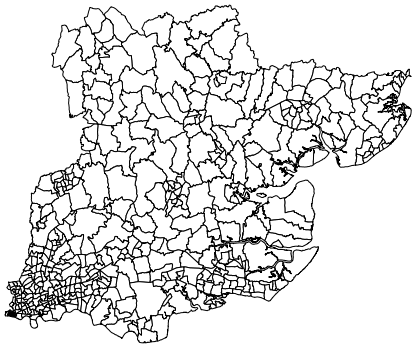


Figure 5.2: The wards of London and Essex.

The numerical complexity associated with the estimation of structured spatial effects using MCMC techniques depends essentially on the structure of the neighborhood matrix. Often the geographical information stored in a boundary file does not represent the "ideal" ordering (as regards to the estimation problem) of the districts or regions. Therefore it may be useful to reorder the map using method `reorder`:

```
> m.reorder
```

Usually reordering results in a smaller bandwidth although the bandwidth is not the criterion that is minimized by `reorder`. Instead the *envelope* of the neighborhood matrix is minimized (compare George and Liu, 1981).

In order to avoid reordering the *map object* every time you start *BayesX* it is useful to store the reordered version in a separate file. This can be achieved using the `outfile` command of *map objects*:

```
> m.outfile, replace using c:\data\LondonEssexSort.bnd
```

The reordered map is now stored in the given file. Note, that specifying the option `replace` allows *BayesX* to overwrite an existing file with the same name. Without this option an error message would be raised if the given file is already existing.
Reading the boundary information from an external file and computing the neighborhood matrix may be a computationally intensive task if the map contains a large number of regions or if the polygons are given in great detail. To avoid doing these computation in every *BayesX* session, we store the neighborhood information in a so–called *graph file* using method `outfile` together with the `graph` option:

```
> m.outfile, replace graph using c:\data\LondonEssexSort.gra
```

For more information on *graph files* we refer to Chapter 5 of the complete *BayesX* manual.

5.5 Bayesreg objects

We start with a detailed description of the estimation of survival models presented in Subsection 2.5.3 to analyze waiting times on CABG. The description of the estimation of relative survival models presented in Section 3.5, and the description of the multi–state models presented in Subsection 4.4.2 to analyze human sleep processes follow thereafter.

5.5.1 Survival models

To estimate a survival model using MCMC techniques we first create a *bayesreg object* which we name `surv_m8`:

```
> bayesreg surv_m8
```

By default estimation results are written to the subdirectory `output` of the installation directory. In this case the default filenames are composed of the name of the *bayesreg object* and the type of the specific file. Usually it is more convenient to store the results in a user–specified directory. To define this directory we use the `outfile` command of *bayesreg objects*:

```
> surv_m8.outfile = c:\data\m8
```

Note, that `outfile` does not only specify a directory but also a base filename (the characters 'm8' in our example). Therefore executing the command above leads to storage of the results in the directory '`c:\data`' and all generated filenames start with the characters 'm8'.
In addition to parameter estimates *BayesX* also gives acceptance rates for the different effects and some further information on the estimation process. In contrast to parameter estimates this information is not stored automatically but is printed in the *output window*. Therefore it is useful to store the contents of the *output window*. This can be achieved automatically by opening a *log file* using the `logopen` command

```
> logopen, replace using c:\data\cabg_log.txt
```

After opening a *log file*, every information written to the output window is also stored in this file. Option `replace` allows *BayesX* to overwrite an existing file with the same name as the specified *log file*. Without `replace` results are appended to an existing file.
Our *dataset object* `cabg` contains the imported variables `ward` (electorial ward a patient resides in), `time` (time since diagnosis), `delta` (indicator of non–censoring), `sex` (1=male, 0=female), `numdv` (number of diseased vessels) and `age` (age of patient at time of diagnosis) as well as the newly generated dummy variables `dv2` and `dv3`. Models 7 and 8 presented in Subsection 2.5.3 correspond to a continuous–time survival model with hazard rate:

$$\lambda(\texttt{t}) = \exp(g_0(\texttt{t}) + f_{age}(\texttt{age}) + f_{spat}(\texttt{ward}) + \gamma_1\texttt{sex} + \gamma_2\texttt{dv2} + \gamma_3\texttt{dv3}),$$

The log–baseline effect g_0 and the continuous covariate `age` are assumed to have a possibly nonlinear effect on the hazard and are therefore modelled nonparametrically via P–splines. The effect of the spatial covariate `ward` is assumed to be spatially correlated, at which model 7 assumes a GRF prior and model 8 assumes a MRF prior. Note that the neighborhood matrix and possible weights associated with the neighbors are obtained from the *map object* `m` (compare Section 5.4).
To estimate model 8 (MRF prior for the spatial effect) we use method `regress` of *bayesreg objects*:

```
> surv_m8.regress delta = time(baseline) + age(psplinerw2)
  + ward(spatial,map=m,proposal=iwlsmode) + sex + dv2 + dv3,
  family=cox iterations=30000 burnin=10000 step=20 predict using cabg
```

Note that with `family=cox` *BayesX* expects the indicator of non–censoring (named `delta` in our example) to be entered on the left side of the equals sign. This indicator has to be a 0–1 coded variable taking the value 0 if an observation is censored and the value 1 otherwise. Furthermore a `baseline` term has to be entered on the right side of the equals sign, which is modelled by a P–spline with second order random walk prior. Note that the variable `time` which indicates the observed survival time has to be greater than zero. In case the global option `begin` is not specified after the comma, it is assumed that each row in the data set represents an observation from $t = 0$ to $t =$ `time`, i.e. no left truncation and time–varying covariates are present. The effect of `age` is also modelled by a P–spline with second order random walk prior, which is specified by `psplinerw2`. By default, the degree of a spline is 3 and the number of inner knots is 20. Full details about all possible options for P–splines are given in Section 7.1 of the *BayesX* reference manual. Concerning the spatial covariate `ward`, the term `spatial` defines a MRF prior where the neighborhood matrix is specified via the option `map`. The additional option `proposal` may be used to specify the type of proposal density, with `proposal=iwlsmode` indicating an iteratively weighted least squares (IWLS) proposal based on posterior mode estimation (see Brezger and Lang (2006) for details). With this example `iwlsmode` turned out to yield higher acceptance rates than the IWLS proposal based on the posterior mean which would be used by default.

Options `iterations`, `burnin` and `step` define properties of the MCMC–algorithm. The total number of MCMC iterations is given by `iterations` while the number of burn in iterations is given by `burnin`. Therefore we obtain a sample of 20000 random numbers with the above specifications. Since, in general, these random numbers are correlated, we do not use all of them but thin out the Markov chain by the thinning parameter `step`. Specifying `step=20` as above forces *BayesX* to store only every 20th sampled parameter which leads to a random sample of length 1000 for every parameter in our example. With `iterations=30000` the simulation run time of model 8 is about 40 minutes (Pentium 4 CPU 2.8 GHz).

If option `predict` is specified, samples of the unstandardized deviance, the effective number of parameters p_D, and the deviance information criterion DIC of the model are computed, see Spiegelhalter et al. (2002). In addition, estimates for the linear predictor and the expectation of every observation are obtained.

For the estimation of model 7 (GRF prior with 100 knots for the spatial effect) we enter the commands

```
> bayesreg surv_m7
> surv_m7.outfile = c:\data\m7
> surv_m7.regress delta = time(baseline) + age(psplinerw2)
  + ward(geokriging,map=m,nrknots=100) + sex + dv2 + dv3,
  family=cox iterations=30000 burnin=10000 step=20 predict using cabg
```

For clarity we created a new *bayesreg object* `surv_m7` and specified the base filename `m7` by the `outfile` command. Note that using the *bayesreg object* `surv_m8` without changing the base filename would also be possible, but would lead to overwriting result files. With `iterations=30000` the simulation run time of model 7 is about 700 minutes (Pentium 4 CPU 2.8 GHz).
Recall the hazard rate of Model 10

$$\lambda(\texttt{t}) = \exp\left(g_0(\texttt{t}) + f_{age}(\texttt{age}) + f_{spat}(\texttt{ward}) + \gamma_1\texttt{sex} + g_1(\texttt{t})\texttt{dv2} + g_2(\texttt{t})\texttt{dv3}\right),$$

where the effect of the number of diseased vessels is modelled as a time–varying effect. This model is estimated as follows

```
> bayesreg surv_m10
> surv_m10.outfile = c:\data\m10
> surv_m10.regress delta = time(baseline) + age(psplinerw2)
  + ward(spatial,map=m,proposal=iwlsmode) + sex
  + dv2*time(baseline) + dv3*time(baseline),
  family=cox iterations=30000 burnin=10000 step=20 predict using cabg
```

The third command specifies cubic P–spline priors for the time–varying effects of the dummy variables `dv2` and `dv3`. With `iterations=30000` the simulation run time of model 10 is about 70 minutes (Pentium 4 CPU 2.8 GHz).
To shed some light on the influence of different choices for hyperpriors we presented some additional results of model 8 that were obtained with other choices of $IG(a;b)$ priors. The following command may for example be used to specify uniform priors on the standard deviations (i.e. set $a = -0.5$ and $b = 0$)

```
> bayesreg surv_m8u
> surv_m8u.outfile = c:\data\m8_uniform
> surv_m8u.regress delta=time(baseline,a=-0.5,b=0)+age(psplinerw2,a=-0.5,b=0)
  +ward(spatial,map=m,proposal=iwlsmode,a=-0.5,b=0)+sex+dv2+dv3,
  family=cox iterations=12000 burnin=2000 step=10 predict using cabg
```

In case the options `a` and `b` are not specified the parameters a and b are set to the default values $a = b = 0.001$.
In addition to the information being printed to the *output window* results for each effect are written to external ASCII files. The names of these files are given in the *output window*. By default the files contain the posterior mean and median, the posterior 2.5%, 10%, 90% and 97.5% quantiles, and the corresponding 95% and 80% posterior probabilities of the estimated effects. The posterior quantiles and posterior probabilities may be changed by the user using the global options `level1` and `level2`.
The *output window* also contains information on how to visualize the estimation results. For more details on visualizing estimation results we refer to Chapter 9 of the *BayesX* reference manual.
Having finished the estimation we may close the *log file* by typing `logclose`. Note, that the *log file* is closed automatically when you exit *BayesX*.

5.5.2 Relative survival analysis

To estimate the relative survival model presented in Section 3.5 we again start by creating a *bayesreg object* by typing

```
> bayesreg rs
```

Note that we could also use the existing *bayesreg object* `surv`, but we prefer to create a new one named `rs` for reasons of clarity. To store the results in the directory `c:\data` and to specify `rs` as a base filename we enter the command

```
> rs.outfile = c:\data\rs
```

A *log file* where the contents of the *output window* are stored is then opened by

```
> logopen, replace using c:\data\breastcancer_log.txt
```

The first lines of the *dataset object* `cancer` are given by

time	delta	age	meta1	meta2	age_plus_time	lambda_e
5.718018	0	46.6694	0	0	52.38742	.0037526
7.738525	0	50.91581	0	0	58.65434	.0070227
3.518052	1	63.14305	1	0	66.6611	.0156398
.5538804	1	78.95688	1	0	79.51076	.0565308
7.741333	0	80.16427	0	0	87.9056	.1308785
5.816528	0	82.85284	0	0	88.66936	.1412662
4.694092	0	83.3922	0	0	88.08629	.1332648
.9661008	1	83.95345	0	1	84.91956	.0970926
4.458618	0	80.96372	1	0	85.42234	.102099
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

where `time` is time `t` since diagnosis (in years), `delta` is the indicator of non–censoring, which takes the value one if the patient died and the value zero if the observation is right–censored. The covariate `age` denotes the age of the patient at time of diagnosis and the dummy variables `meta1` and `meta2` indicate whether the number of metastases is one or more than one, respectively. The variable `age_plus_time` is an auxiliary variable that was used to generate the expected hazard rate. It is given by the sum of `age` and `time` and denotes the age of the patient at the end of the observation, i.e. the age at death or at the time, when the observation was right–censored. Finally, `lambda_e` denotes the expected hazard rate, which with our example is given by `lambda_e`(`age_plus_time`) = $\exp\left((\texttt{age\_plus\_time} - 30)/10\right)/2500$. Note that a *dataset object* used for the estimation of a relative survival model has to contain the expected hazard rate λ^e, that usually depends on the age at death, the sex of a patient and possibly the date of death or further covariates. Typically this variable will have to be generated in advance with the help of mortality tables. Here it is important to consider that the observed survival time t and the hazard rate λ^e refer to the same time unit. Mortality tables usually contain annual data. In that case the survival times would have to be given in years as well.

The model presented in Section 3.5 corresponds to a relative survival model with hazard rate:

$$\begin{aligned}\lambda &= \lambda^e(\texttt{age},\texttt{t}) + \lambda^c(\texttt{t},\texttt{age},\texttt{meta1},\texttt{meta2}) \\ &= \lambda^e(\texttt{age}+\texttt{t}) + \exp(g_0(\texttt{t}) + f_{age}(\texttt{age}) + \gamma_1\texttt{meta1} + \gamma_2\texttt{meta2}),\end{aligned}$$

where the hazard rate is additively composed of the known expected hazard rate λ^e and the unknown disease–specific hazard rate λ^c. The log–baseline effect g_0 and the continuous covariate `age` are assumed to have a possibly nonlinear effect on the (disease–specific) hazard and are therefore modelled nonparametrically via P–splines.
To estimate this model we again use method `regress` of *bayesreg objects*:

```
> rs.regress delta = time(baseline) + age(psplinerw2)
  + meta1 + meta2 + lambda_e(offest),
  family=cox iterations=30000 burnin=10000 step=20 using rs
```

Note that the only difference to the estimation of crude survival models as presented in the previous subsection is the additional term `lambda_e(offset)`, that is used to specify the variable `lambda_e` as the expected hazard rate. With `iterations=30000` the simulation run time of this model is about 30 minutes (Pentium 4 CPU 2.8 GHz).
Again, additionally to the information being printed to the *output window* results for each effect are written to external ASCII files, with the names of these files being given in the *output window*. Having finished the estimation we may close the *log file* by typing `logclose`.

5.5.3 Multi–state models

To estimate the multi–state models presented in Subsection 4.4.2 we again start by creating a *bayesreg object* by typing

```
> bayesreg ms
```

To store the results in the directory `c:\data` and to specify `ms` as a base filename we enter the command

`id`	identification number of subject
`beg`	time of transition to the current state (admission time)
`end`	time of transition to the next state (emission time)
`tas`	1 a transition AWAKE→SLEEP is observed at $t =$ `end`
	0 else
`tsa`	1 a transition SLEEP→AWAKE is observed at $t =$ `end`
	0 else
`trn`	1 a transition REM→NREM is observed at $t =$ `end`
	0 else
`tnr`	1 a transition NREM→REM is observed at $t =$ `end`
	0 else
`st`	1 subject is currently in state AWAKE
	2 subject is currently in state NREM
	3 subject is currently in state REM
`cort`	cortisol level in nmol/l

Table 5.1: Original variables of the dataset object sleep.

```
> ms.outfile = c:\data\ms
```

A *log file* where the contents of the *output window* are stored is then opened by

```
> logopen, replace using c:\data\humansleep_log.txt
```

The original variables of the *dataset object* `sleep` are summarized and explained in Table 5.1. The additional dummy coded covariate `corthigh`, which indicates whether or not the cortisol secretion is higher than 90 nmol/l is generated by typing

```
> sleep.generate corthigh = (cort>90)
```

Now the first lines of the *dataset object* `sleep` are given by

```
id  st  beg   end  tas tsa trn tnr cort corthigh
1   2   0     1    0   1   0   0   52.6  0
1   1   1     5    1   0   0   0   52.6  0
1   2   5     8    0   1   0   0   52.6  0
1   1   8     10   1   0   0   0   52.6  0
1   2   10    36   0   0   0   0   52.6  0
1   2   36    76   0   0   0   0   46.9  0
1   2   76    108  0   0   0   1   47.5  0
1   3   108   109  0   0   1   0   47.5  0
1   2   109   110  0   0   0   1   47.5  0
1   3   110   111  0   0   1   0   47.5  0
1   2   111   115  0   0   0   1   47.5  0
1   3   115   116  0   0   0   0   47.5  0
1   3   116   126  0   0   1   0   37.4  0
.   .    .     .   .   .   .   .     .   .
.   .    .     .   .   .   .   .     .   .
.   .    .     .   .   .   .   .     .   .
```

Note that the states have to be numbered consecutively from 1 to H, at which numbers are exchangeable. Since we are considering continuous time scales, an observation should start at $t = 0$ (unless the observation is left truncated) and the variables `beg` and `end` should be generated so that within each observation process `beg` equals the value of `end` in the previous row (unless observations are fragmentary only).

The transition rates of the multi–state model analyzed in Subsection 4.4.2 are given by

$$\begin{aligned} \lambda_h &= \exp\left(g_{h0}(t) + b_h\right), \quad h = AS, SA, RN \\ \lambda_h &= \exp\left(g_{h0}(t) + \texttt{c}(t) \cdot g_{h1}(t) + b_h\right), \quad h = NR \end{aligned}$$

This model is estimated with *BayesX* by entering the following command

```
> ms.mregress tas = end(baseline) + id(random):
              tsa = end(baseline) + id(random):
              trn = end(baseline) + id(random):
              tnr = end(baseline) + corthigh*end(baseline) + id(random),
  family=multistate begin=beg state=st iterations=30000 burnin=10000
  step=20 using sleep
```

Note that the command `regress` used with the estimation of Cox models (and other models with univariate response) is now replaced by the command `mregress` which is used to analyze models with multivariate responses. With `family=multistate` *BayesX* expects the specification of at least two transitions separated by a colon, at which the corresponding 0–1 coded transition indicators are to be entered on the left side of the equals sign. With the command above *BayesX* assumes cubic P–spline priors with 20 knots and second order random walk priors for the log–baseline effects as well as the time–varying effect of `corthigh`, diffuse priors for the fixed effects of `sex` and i.i.d. Gaussian priors with mean zero for each individual and transition specific random effect. With `iterations=30000` the simulation run time is about 160 minutes (Pentium 4 CPU 2.8 GHz). Concerning the state structure, *BayesX* assumes that an observation with current state `st` is at risk of experiencing a transition of type h if the data set contains at least one type h transition with the accordant state `st`. For checking purposes the following matrix, that indicates the number of type h transitions observed with every single state, is printed in the output window.

```
Matrix of possible transitions:

    Transition  1   2   3   4
State
1           460 0   0   0
2           0   399 0   306
3           0   77  234 0
```

Again, additionally to the information being printed to the *output window* results for each effect are written to external ASCII files, with the names of these files being given in the *output window*. Having finished the estimation we may close the *log file* by typing `logclose`.

5.6 Post estimation commands

Bayesreg objects provide some post estimation commands to get sampled parameters or to plot autocorrelation functions of sampled parameters. For example

```
> surv_m8.plotautocor, maxlag=250
```

computes the autocorrelation functions for all parameters estimated with the `regress` command (lastly) entered with the *bayesreg object* `surv_m8`. Here verb+maxlag+ specifies the maximum lag number.

If the number of parameters is large this may be computationally expensive, so *BayesX* provides a second possibility to compute autocorrelation functions. Adding the option `mean` to the `plotautocor` command as in

```
> surv_m8.plotautocor, mean
```

leads to the computation of only the minimum, mean and maximum autocorrelation functions.
Note, that executing the `plotautocor` command also stores the computed autocorrelation functions in a file named `autocor.raw` in the output directory of the *bayesreg object*.

To save memory, the sampling paths of the estimated parameters are only stored temporarily by default and will be destroyed, when the corresponding *bayesreg object* is deleted. If we want to store the sampling paths permanently, we have to execute the `getsample` command

```
> surv_m8.getsample
```

which stores the sampled parameters in ASCII files in the output directory. To avoid too large files, the samples are typically partitioned into several files.

Appendix A

Calculation of IWLS weights

A.1 Geoadditive survival analysis

In Subsection 2.3.1, which is concerned with Bayesian inference for geoadditive survival models, we describe how to update parameter vectors corresponding to time–independent effects by an MH–algorithm based on IWLS proposals. The IWLS weights w_i and working observations $\tilde{y}_i$ used with this algorithm are derived as follows.
As specified in equation (2.4) the geoadditive predictor of our survival models is given by

$$\eta_i(t) = g_0(t) + \sum_{j=1}^{p} g_j(t) z_{ij} + \sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i'\boldsymbol{\gamma} + b_{g_i}.$$

Suppose for example we want to update the parameter vector $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j,d_j})'$ corresponding to a time–independent function $f_j(x_{ij})$, which is modelled by a P–spline. With the generic notation of Subsection 2.2.1 this function may be written as

$$f_j(x_{ij}) = \sum_{m=1}^{d_j} \beta_{jm} B_m(x_{ij}) = \mathbf{Z}_{ji}\boldsymbol{\beta}_j$$

with B_m denoting B–spline basis functions and $\mathbf{Z}_{ji} = \big(B_1(x_{ij}), \ldots, B_{d_j}(x_{ij})\big)$ denoting the i–th row of the design matrix $\mathbf{Z}_j$ introduced in Subsection 2.2.1. The predictor $\eta_i(t)$ and the vector of predictors $\boldsymbol{\eta} = (\eta_1(t_1), \ldots, \eta_n(t_n))'$, respectively, may now be rewritten as

$$\eta_i(t) = \mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(t), \quad \boldsymbol{\eta} = \mathbf{Z}_j\boldsymbol{\beta}_j + \tilde{\boldsymbol{\eta}}$$

The following proportionality holds for the full conditional of $\boldsymbol{\beta}_j$

$$p(\boldsymbol{\beta}_j|\cdot) \propto L(\boldsymbol{\beta}_j) \cdot p(\boldsymbol{\beta}_j|\tau_j^2) \tag{A.1}$$

with the first factor denoting the likelihood that depends among others upon $\boldsymbol{\beta}_j$ and the second factor denoting the prior of $\boldsymbol{\beta}_j$. The dependency of the likelihood on $\boldsymbol{\beta}_j$ may be expressed as follows

$$\begin{aligned}
L(\boldsymbol{\beta}_j) &= \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} \cdot \exp\left[-\int_0^{t_i} \lambda_i(u)du\right] \\
&= \exp\left[\sum_{i=1}^{n}\left(\delta_i\eta_i(t) - \int_0^{t_i}\exp(\eta_i(u))du\right)\right] \\
&= \exp\left[\sum_{i=1}^{n}\left(\delta_i(\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(t)) - \int_0^{t_i}\exp(\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(u))du\right)\right] \\
&= \exp\left[\sum_{i=1}^{n} l_i(\boldsymbol{\beta}_j)\right] = \exp\left[l(\boldsymbol{\beta}_j)\right]
\end{aligned}$$

with l_i denoting the individual log–likelihood. As specified in equation (2.7), the general form of a prior for $\boldsymbol{\beta}_j$ is

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right),$$

Suppose the current value of the chain is $\boldsymbol{\beta}_j^c$. Then a new value $\boldsymbol{\beta}_j^p$ is proposed by drawing a random vector from a multivariate Gaussian proposal distribution, which is obtain from a quadratic approximation of the log–likelihood by a second order Taylor expansion with respect to $\boldsymbol{\beta}_j^c$ given by

$$l(\boldsymbol{\beta}_j^p) \approx l(\boldsymbol{\beta}_j^c) + (\boldsymbol{\beta}_j^p - \boldsymbol{\beta}_j^c)'s(\boldsymbol{\beta}_j^c) + \frac{1}{2}(\boldsymbol{\beta}_j^p - \boldsymbol{\beta}_j^c)'\mathbf{H}(\boldsymbol{\beta}_j^c)(\boldsymbol{\beta}_j^p - \boldsymbol{\beta}_j^c) \tag{A.2}$$

at which $\mathbf{s}$ and $\mathbf{H}$ are the score function and the Hessian matrix (with respect to $\boldsymbol{\beta}_j$), respectively. Inserting (A.2) in (A.1) yields

$$
\begin{aligned}
p(\boldsymbol{\beta}_j^p|\cdot) &\propto \exp\left(l(\boldsymbol{\beta}_j^p) - \frac{1}{2}\boldsymbol{\beta}_j'\frac{\boldsymbol{K}_j}{\tau_j^2}\boldsymbol{\beta}_j\right) \\
&\approx \exp\left((\boldsymbol{\beta}_j^p)'\mathbf{s}(\boldsymbol{\beta}_j^c) + \frac{1}{2}(\boldsymbol{\beta}_j^p)'\mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^p - (\boldsymbol{\beta}_j^p)'\mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^c - \frac{1}{2}(\boldsymbol{\beta}_j^p)'\frac{\boldsymbol{K}_j}{\tau_j^2}\boldsymbol{\beta}_j^p\right) \\
&= \exp\left((\boldsymbol{\beta}_j^p)'\mathbf{s}(\boldsymbol{\beta}_j^c) + \frac{1}{2}(\boldsymbol{\beta}_j^p)'\mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^p - (\boldsymbol{\beta}_j^p)'\mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^c - \frac{1}{2}(\boldsymbol{\beta}_j^p)'\frac{\boldsymbol{K}_j}{\tau_j^2}\boldsymbol{\beta}_j^p\right) \\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_j^p)'\left(-\mathbf{H}(\boldsymbol{\beta}_j^c) + \frac{\boldsymbol{K}_j}{\tau_j^2}\right)\boldsymbol{\beta}_j^p + (\boldsymbol{\beta}_j^p)'\left(\mathbf{s}(\boldsymbol{\beta}_j^c) - \mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^c\right)\right)
\end{aligned}
$$

which is proportional to a multivariate Gaussian distribution with precision matrix and mean

$$\mathbf{P}_j = -\mathbf{H}(\boldsymbol{\beta}_j^c) + \frac{\boldsymbol{K}_j}{\tau_j^2}, \quad \mathbf{m}_j = (\mathbf{P}_j)^{-1}\left(\mathbf{s}(\boldsymbol{\beta}_j^c) - \mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^c\right). \tag{A.3}$$

For the calculation of $\mathbf{P}_j$ and $\mathbf{m}_j$ we need to compute the score function $\mathbf{s}$ and the Hessian matrix $\mathbf{H}$. The score function $\mathbf{s}$ is given by

$$\mathbf{s}(\boldsymbol{\beta}_j) = \left(\sum_i^n \frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial \beta_{j1}}, \ldots, \sum_i^n \frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial \beta_{j,d_j}}\right)'$$

with

$$
\begin{aligned}
l_i(\boldsymbol{\beta}_j) &= \delta_i(\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(t)) - \int_0^{t_i}\exp(\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(u))du \\
&= \delta_i\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \delta_i\tilde{\eta}_i(t) - \exp(\mathbf{Z}_{ji}\boldsymbol{\beta}_j)\int_0^{t_i}\exp(\tilde{\eta}_i(u))du
\end{aligned}
$$

and thus

$$
\begin{aligned}
\frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial \beta_{jm}} &= \delta_i\mathbf{Z}_{jim} - \mathbf{Z}_{jim}\exp(\mathbf{Z}_{ji}\boldsymbol{\beta}_j)\int_0^{t_i}\exp(\tilde{\eta}_i(u))du \\
&= \delta_i\mathbf{Z}_{jim} - \mathbf{Z}_{jim}\int_0^{t_i}\exp(\eta_i(u))du
\end{aligned}
$$

at which $\mathbf{Z}_{jim}$ is the element in the i–th row and m–th column of the design matrix $\mathbf{Z}_j$. Hence the score vector $\mathbf{s}(\boldsymbol{\beta}_j)$ is given by

$$\mathbf{s}(\boldsymbol{\beta}_j) = \mathbf{Z}_j'\boldsymbol{\Delta} - \mathbf{Z}_j'\tilde{\mathbf{W}}(\boldsymbol{\beta}_j) \tag{A.4}$$

at which $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_n)'$ and $\tilde{\mathbf{W}} = (w_1, \ldots, w_n)'$ with

$$w_i = \int_0^{t_i} \exp(\eta_i(u))du = \Lambda_i(t_i). \tag{A.5}$$

The Hessian matrix $\mathbf{H}$ is defined as follows

$$\mathbf{H}(\boldsymbol{\beta}_j) = \left(\sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial \beta_{jm} \partial \beta_{jk}} \right)_{m,k=1,\ldots,d_j}$$

Computing the partial derivatives delivers

$$\frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial \beta_{jm} \partial \beta_{jk}} = -\mathbf{Z}_{jim}\mathbf{Z}_{jik} \int_0^{t_i} \exp(\eta_i(u))du$$

leading to

$$\mathbf{H}(\boldsymbol{\beta}_j) = -\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j)\mathbf{Z}_j \tag{A.6}$$

with the weight matrix $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$ with w_i as defined in (A.5).
Inserting (A.4) and (A.6) in (A.3) yields

$$\begin{aligned}
\mathbf{P}_j &= \mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\mathbf{Z}_j + \frac{\boldsymbol{K}_j}{\tau_j^2} \\
\mathbf{m}_j &= (\mathbf{P}_j)^{-1} \left(\mathbf{Z}_j'\boldsymbol{\Delta} - \mathbf{Z}_j'\tilde{\mathbf{W}}(\boldsymbol{\beta}_j^c) + \mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\mathbf{Z}_j\boldsymbol{\beta}_j^c \right) \\
&= (\mathbf{P}_j)^{-1}\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c) \left(\mathbf{W}^{-1}(\boldsymbol{\beta}_j^c)\Delta - \mathbb{1} + \mathbf{Z}_j\boldsymbol{\beta}_j^c \right) \\
&= (\mathbf{P}_j)^{-1}\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c) \left(\mathbf{W}^{-1}(\boldsymbol{\beta}_j^c)\Delta - \mathbb{1} + \boldsymbol{\eta} - \tilde{\boldsymbol{\eta}} \right) \\
&= (\mathbf{P}_j)^{-1}\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c) \left(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\eta}} \right)
\end{aligned}$$

with the n–dimensional vector of working observations

$$\tilde{\mathbf{y}} = \mathbf{W}^{-1}(\boldsymbol{\beta}_j^c)\Delta - \mathbb{1} + \boldsymbol{\eta} = \left(\eta_1(t_1) + \frac{\delta_1}{w_1} - 1, \ldots, \eta_n(t_n) + \frac{\delta_n}{w_n} - 1 \right)'$$

A.2 Relative survival analysis

As mentioned in Section 3.3 updating of parameter vectors $\boldsymbol{\beta}_j$ corresponding to time–independent effects within a relative survival model is performed according to the same principle as described above for crude survival models. However, as a consequence of the slightly more complex likelihood the weights and working observations that are used within the IWLS–MH algorithm are slightly more complex as well. Using the results of (A.3) those quantities are derived as follows.

For updating the parameter vector $\boldsymbol{\beta}_j$ again consider the following decomposition of $\eta_i(t) = \log\left(\lambda_i^c(t)\right)$ and $\boldsymbol{\eta} = (\eta_1(t_i), \ldots, \eta_n(t_n))'$, respectively

$$\eta_i(t) = \mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(t), \quad \boldsymbol{\eta} = \mathbf{Z}_j\boldsymbol{\beta}_j + \tilde{\boldsymbol{\eta}}$$

With $\lambda_i = \lambda_i^e + \lambda_i^c$, where $\lambda_i^e := \lambda_i^e(a_i + t_i)$ denotes the expected hazard and $\lambda_i^c = \lambda_i^c(t_i)$ denotes the disease related hazard, it can be seen easily from (3.3) that the individual log–likelihood $l_i(\boldsymbol{\beta}_j)$ is given by

$$\begin{aligned}
l_i(\boldsymbol{\beta}_j) &= \delta_i \log\left(\lambda_i^e + \lambda_i^c\right) - \int_0^{t_i} \lambda_i^e(a_i + u)du - \int_0^{t_i} \lambda_i^c(u)du \\
&= \delta_i \log\left(\lambda_i^e + \exp\left(\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(t_i)\right)\right) \\
&\quad - \int_0^{t_i} \lambda_i^e(a_i + u)du - \int_0^{t_i} \exp\left(\mathbf{Z}_{ji}\boldsymbol{\beta}_j + \tilde{\eta}_i(u)\right) du \\
&= \delta_i \log\left(\lambda_i^e + \exp\left(\mathbf{Z}_{ji}\boldsymbol{\beta}_j\right)\exp\left(\tilde{\eta}_i(t_i)\right)\right) \\
&\quad - \int_0^{t_i} \lambda_i^e(a_i + u)du - \exp\left(\mathbf{Z}_{ji}\boldsymbol{\beta}_j\right)\int_0^{t_i} \exp\left(\tilde{\eta}_i(u)\right) du
\end{aligned}$$

Hence the partial derivative is given by

$$\begin{aligned}
\frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial \beta_{jm}} &= \frac{\delta_i}{\lambda_i^e + \lambda_i^c}\mathbf{Z}_{jim}\exp\left(\mathbf{Z}_{ji}\boldsymbol{\beta}_j\right)\exp\left(\tilde{\eta}_i(t_i)\right) \\
&\quad -0 - \mathbf{Z}_{jim}\exp\left(\mathbf{Z}_{ji}\boldsymbol{\beta}_j\right)\int_0^{t_i}\exp\left(\tilde{\eta}_i(u)\right) du \\
&= \frac{\delta_i\lambda_i^c}{\lambda_i^e + \lambda_i^c}\mathbf{Z}_{jim} - \mathbf{Z}_{jim}\int_0^{t_i}\lambda_i^c(u)du
\end{aligned}$$

and thus we get

$$\mathbf{s}(\boldsymbol{\beta}_j) = \mathbf{Z}_j'\Delta - \mathbf{Z}_j'\tilde{\mathbf{W}} \tag{A.7}$$

with $\Delta = \left(\frac{\delta_1\lambda_1^c}{\lambda_1^e+\lambda_1^c}, \ldots, \frac{\delta_n\lambda_n^c}{\lambda_n^e+\lambda_n^c}\right)'$ and $\tilde{\mathbf{W}} = (\tilde{w}_1, \ldots, \tilde{w}_n)'$ with

$$\tilde{w}_i = \int_0^{t_i} \lambda_i^c(u)du = \Lambda_i^c(t_i) \tag{A.8}$$

By means of the quotient rule the elements of the Hessian matrix $\mathbf{H}$ are computed as follows

$$\begin{aligned}
\frac{\partial l_i(\boldsymbol{\beta}_j)}{\partial\beta_{jm}\partial\beta_{jk}} &= \frac{\left(\lambda_i^e+\lambda_i^c\right)\frac{\partial\delta_i\lambda_i^c\mathbf{Z}_{jim}}{\partial\beta_{jk}} - \delta_i\lambda_i^c\mathbf{Z}_{jim}\frac{\partial\lambda_i^e+\lambda_i^c}{\partial\beta_{jk}}}{\left(\lambda_i^e+\lambda_i^c\right)^2} - \mathbf{Z}_{jim}\mathbf{Z}_{jik}\int_0^{t_i}\lambda_i^c(u)du \\
&= \frac{\left(\lambda_i^e+\lambda_i^c\right)\delta_i\mathbf{Z}_{jim}\mathbf{Z}_{jik}\lambda_i^c - \delta_i\lambda_i^c\mathbf{Z}_{jim}\mathbf{Z}_{jik}\lambda_i^c}{\left(\lambda_i^e+\lambda_i^c\right)^2} - \mathbf{Z}_{jim}\mathbf{Z}_{jik}\int_0^{t_i}\lambda_i^c(u)du \\
&= \frac{\lambda_i^e\delta_i\mathbf{Z}_{jim}\mathbf{Z}_{jik}\lambda_i^c}{\left(\lambda_i^e+\lambda_i^c\right)^2} - \mathbf{Z}_{jim}\mathbf{Z}_{jik}\int_0^{t_i}\lambda_i^c(u)du \\
&= \mathbf{Z}_{jim}\frac{\lambda_i^e\lambda_i^c\delta_i}{\left(\lambda_i^e+\lambda_i^c\right)^2}\mathbf{Z}_{jik} - \mathbf{Z}_{jim}\int_0^{t_i}\lambda_i^c(u)du\mathbf{Z}_{jik} \\
&= -\mathbf{Z}_{jim}\left(\int_0^{t_i}\lambda_i^c(u)du - \frac{\lambda_i^e\lambda_i^c\delta_i}{\left(\lambda_i^e+\lambda_i^c\right)^2}\right)\mathbf{Z}_{jik}
\end{aligned}$$

The Hessian matrix $\mathbf{H}$ may now be written as

$$\mathbf{H}(\boldsymbol{\beta}_j) = -\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j \tag{A.9}$$

with $\mathbf{W} = diag(w_1, \ldots, w_n)$ at which

$$w_i = \tilde{w}_i - \frac{\lambda_i^e\lambda_i^c\delta_i}{\left(\lambda_i^e+\lambda_i^c\right)^2} = \Lambda_i^c(t_i) - \frac{\lambda_i^e\lambda_i^c\delta_i}{\lambda_i^2}$$

Inserting (A.7) and (A.9) in (A.3) yields the precision matrix and the mean of the Gaussian proposal density for $\boldsymbol{\beta}_j^p$ as well as the working observations $\tilde{\mathbf{y}}$, which are derived as follows

$$
\begin{aligned}
\mathbf{m}_j &= (\mathbf{P}_j)^{-1}\left(\mathbf{s}(\boldsymbol{\beta}_j^c) - \mathbf{H}(\boldsymbol{\beta}_j^c)\boldsymbol{\beta}_j^c\right) \\
&= (\mathbf{P}_j)^{-1}\left(\mathbf{Z}_j'\Delta - \mathbf{Z}_j'\tilde{\mathbf{W}}(\boldsymbol{\beta}_j^c) + \mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\mathbf{Z}_j\boldsymbol{\beta}_j^c\right) \\
&= (\mathbf{P}_j)^{-1}\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\left(\mathbf{W}^{-1}(\boldsymbol{\beta}_j^c)\Delta - \mathbf{W}^{-1}(\boldsymbol{\beta}_j^c)\tilde{\mathbf{W}}(\boldsymbol{\beta}_j^c) + \mathbf{Z}_j\boldsymbol{\beta}_j^c\right) \\
&= (\mathbf{P}_j)^{-1}\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\left(\mathbf{W}^{-1}(\boldsymbol{\beta}_j^c)\left(\Delta - \tilde{\mathbf{W}}(\boldsymbol{\beta}_j^c)\right) + \boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}\right) \\
&= (\mathbf{P}_j)^{-1}\mathbf{Z}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\left(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\eta}}\right)
\end{aligned}
$$

with $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)$, at which

$$\tilde{y}_i = \eta_i(t_i) + \frac{\delta_i\lambda_i^c/\lambda_i - \tilde{w}_i}{w_i}.$$

Bibliography

Andersen, P.K., Borgan, ∅., Gill, R.D., and Keiding, N. (1993), *Statistical models based on counting processes*, New York: Springer.

Andersen, P.K., and Keiding, N. (2002), "Multi–state models for event history analysis," *Statistical Methods in Medical Research*, 11, 91–15.

Banerjee, S., and Carlin, B.P. (2003), "Semiparametric Spatiotemporal Frailty Modelling," *Environmetrics*, 14, 523–535.

Banerjee, S., and Carlin, B.P. (2004), "Parametric Spatial Cure Rate Models for Interval–Cencored Time–to–Relapse Data," *Biometrics*, 60, 268–275.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC, Boca Raton.

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, 123–142.

Bender, R., Augustin, T., and Blettner, M. (2005), "Generating survival times to simulate Cox proportional hazards models," *Statistics in Medicine*, 24, 1713–1723.

Besag, J. and Kooperberg, C. (1995), "On Conditional and Intrinsic Autoregressions," *Biometrika*, 82, 733–746.

Blossfeld, H.-P., Hamerle, A., and Mayer, K.U. (1989). *Event History Analysis.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bolard, P., Quantin, C., Esteve, J., Faivre, J., and Abrahamowicz, M. (2001), "Modelling time–dependent hazard ratios in relative survival: Application to colon cancer," *Journal of Clinical Epidemiology*, 54, 986–996.

Brezger, A., Kneib, T., and Lang, S. (2005), "BayesX: Analysing Bayesian Semiparametric Regression Models," *Journal of statistical software*, Vol. 14, Issue 11. Open domain software available from `http://www.stat.uni-muenchen.de/~bayesx/`.

Brezger, A., and Lang, S. (2006), "Generalized structured additive regression based on Bayesian P–splines," *Computational Statistics and Data Analysis*, 50, 967–991.

Cai, T., and Betensky, R. A. (2003), "Hazard Regression for Interval Censored Data with Penalized Spline," *Biometrics*, 59, 570–9.

Cai, T., Hyndman, R., and Wand, M. (2002), "Mixed model-based hazard estimation," *Journal of Computational and Graphical Statistics*, 11, 784–798.

Carlin, B. P., and Banerjee, S. (2002), "Hierarchical Multivariate CAR Models for Spatio–Temporally Correlated Data," In: *Bayesian Statistics 7*, eds. J.M. Bernardo et al., Oxford: Oxford University Press.

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

Crook, A., Knorr-Held, L., and Hemingway, H. (2003), "Measuring spatial effects in time to event data: a case study using months from angiography to coronary artery bypass graft (CABG)," *Statistics in Medicine*, 22, 2943–2961.

Czado, C., and Rudolph, F. (2002), "Application of survival analysis methods to long–term care insurance," *Insurance: Mathematics and Economics*, 31 (3), 395–413.

De Boor, C. (2001), *A practical guide to Splines*, New York.

Devroye L. (1986), *Non–uniform random variate generation*, New York: Springer.

Eilers, P.H.C., and Marx, B.D. (1996), "Flexible smoothing using B-splines and penalized likelihood" (with comments and rejoinder), *Statistical Science*, 11 (2), 89–121.

Esteve, J., Benhamou, E., Croasdale, M., and Raymond, L. (1990), "Relative Survival and the estimation of net survival: elements for further discussion," *Statistics in Medicine*, 9, 529–538.

Fahrmeir, L., and Klinger, A. (1998), "A nonparametric multiplicative hazard model for event history data," *Biometrika*, 85(3), 581–592.

Fahrmeir, L., and Lang, S. (2001a), "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Journal of the Royal Statistical Society*, Ser. C, 50, 201–220.

Fahrmeir, L., and Lang, S. (2001b), "Bayesian semiparametric regression analysis of multicategorical time–space data," *Annals of the Institute of Statistical Mathematics*, 53, 11–30.

Fahrmeir, L., Lang, S.,Wolff, J., and Bender, S. (2003), "Semiparametric Bayesian Time-Space Analysis of Unemployment Duration," *Journal of the German Statistical Society (Allgemeines Statistisches Archiv)*, 87, 281–307.

Fahrmeir, L., and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer–Verlag, New York.

Gamerman, D. (1997), "Efficient Sampling from the Posterior Distribution in Generalized Linear Models," *Statistics and Computing*, 7, 57–68.

Gelfand, A.E., and Gosh, S.K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

Gelman, A. (2004), "Prior distributions for variance parameters in hierarchical models," provided by *Economics Working Paper Archive at WUSTL* in its series *Econometrics* with number 0404001.

George, A. and Liu, J.W. (1981), *Computer Solution of Large Sparse Positive Definite Systems*, Prentice–Hall.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Esteve, J., Gouvernet, J. and Faivre, J. (2003), "A relative survival regression model using B–spline functions to model non–proportional hazards," *Statistics in Medicine*, 22, 2767–2784.

Gould, A., and Lawless, J.F. (1988), "Estimation Efficiency in Lifetime Regression Models when Responses are Censored or Grouped," *Comm. Statist. Simul.,* 17, 689–712.

Henderson, R., Shimakura, S., and Gorst, D. (2002), "Modeling Spatial Variation in Leukemia Survival Data," *Journal of the American Statistical Assosiation*, 97, 965–972.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2005), "Geoadditive Survival Models: A Supplement," *SFB 386 Discussion Paper 454*, University of Munich. Available from `http://www.stat.uni-muenchen.de/sfb386/`.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2005), "Geoadditive Survival Models," *Journal of the American Statistical Association, Theory and Methods*, to appear.

Ibrahim, J.G., Chen, M.H., and Sinha, D. (2001), *Bayesian Survial Analysis*. Springer Series in Statistics, New York.

Kaempchen, S., Guenther, T., Toschke, M., Grunkemeier, G.L., Wottke, m., and Lange, R. (2003), "Assessing the benefit of biological valve prostheses: cumulative incidence (actual) vs. Kaplan–Meier (actuarial) analysis," *European Journal of Cardio–thoracic Surgery*, 23, 710–714.

Kalbfleich, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2nd edition.* Hoboken: John Wiley & Sons.

Kammann, E.E., and Wand, M.P. (2003), "Geoadditive models," *Journal of the Royal Statistical Society*, Ser. C, 52, 1–18.

Klein, J.P. and Moeschberger, M.L. (2003). *Survival analysis.* Springer, New York.

Kneib, T. (2006). *Mixed model based inference in structured additive regression.* PhD thesis, Dr. Hut Verlag.

Kneib, T. and Fahrmeir, L. (2004), "A mixed model approach for structured hazard regression," *SFB 386 Discussion Paper 400*, University of Munich. Available from `http://www.stat.uni-muenchen.de/sfb386/`, accepted for publication in the *Scandinavian Journal of Statistics*.

Kneib, T. and Fahrmeir, L. (2005), "Structured additive regression for multicategorical space-time data: A mixed model approach," *Biometrics*, to appear.

Knorr–Held, L. (1999), "Conditional Prior Proposals in Dynamic Models," *Scandinavian Journal of Statistics*, 26, 129–144.

Komárek, A., Lesaffre, E., and Hilton, J.F. (2005), "Accelerated failure time model for arbitrarly censored data with smoothed error distribution," *Journal of Computational and Graphical Statistics*, 45, 726–745.

Lang, S., and Brezger, A. (2004), "Bayesian P–splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

Lesaffre, E., Komárek, A., and Declerck, D. (2004), "An overview of methods for interval–censored data with an emphasis on applications in dentistry," *Technical report 0453*, IAP network. Available from `http://www.stat.ucl.ac.be/IAP`.

Lewis, P. A. W. and Shedler, G. S. (1979), "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics Quarterly*, 26, 403–414.

Li, Y., and Ryan, L. (2002), "Modeling Spatial Survival Data Using Semiparametric Frailty Models," *Biometrics*, 58, 287–297.

Marx, B.D., and Eilers, P. (1998), "Direct Generalized Additive Modeling with Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

Percy, C.L., Stanek, E., and Gloeckler, L. (1981), "Accuracy of cancer death certificates and its effect on cancer mortality statistics," *American Journal of Public Health*, 71, 242–250.

Rue, H. (2001), "Fast sampling of Gaussian Markov random fields," *Journal of the Royal Statistical Society*, Ser. B, 63, 325–338.

Sauleau, E.-A., Hennerfeind, A., Buemi, A., and Held, L. (2006), "Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geoadditive models," *Statistics in Medicine*, to appear.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit" (with discussion and rejoinder), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Stein, M.L. (1999), *Interpolation of spatial data. Some theory for kriging*, Springer, New York.

Sun, D., Tsutakawa, R.K. and Speckman, P.L. (1999), "Posterior distribution of hierarchical models using CAR(1) distributions," *Biometrika*, 86, 341–350.

Sun, D., Tsutakawa, R.K. and He, Z. (2001), "Propriety of posteriors with improper priors in hierarchical linear mixed models," *Statistica Sinica*, 11, 77–95.

Thompson, W.A.,Jr. (1977), "On the Treatment of Grouped Observations in Life Studies," *Biometrics*, 33, 463–470.

LEBENSLAUF

Andrea Hennerfeind

geboren am 11. August 1976 in Freising

Schulbildung:

1982–1986	Grundschule St. Lantbert in Freising
1986–1995	Josef–Hofmiller–Gymnasium in Freising (mathematisch–naturwissenschaftlich)

Studium:

1995–2001	Studium der Statistik an der Ludwig–Maximilians–Universität München mit dem Anwendungsgebiet Versicherungswissenschaften und der speziellen Ausrichtung mathematische Statistik
Sep. 1997	Diplom–Vorprüfung
Mai 2001	Diplom–Hauptprüfung

Beruf:

seit Juli 2001	wissenschaftlicher Mitarbeiter im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" bei Prof. Dr. L. Fahrmeir am Institut für Statistik der Ludwig–Maximilians–Universität München