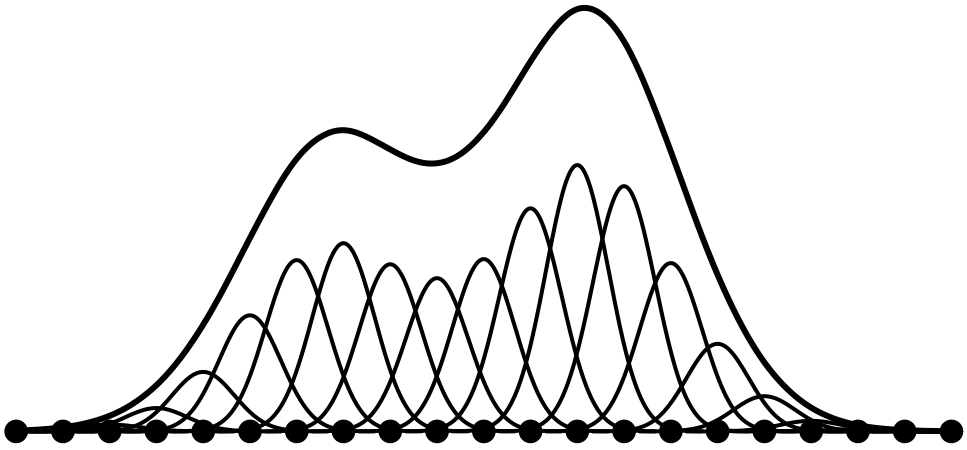
Katholieke Universiteit Leuven
Faculteit Wetenschappen

Arnošt Komárek

Accelerated Failure Time Models for Multivariate Interval-Censored Data with Flexible Distributional Assumptions



Promotoren:
Prof. Emmanuel Lesaffre
Prof. Jan Beirlant

Proefschrift ingediend tot
het behalen van de graad van
Doctor in de Wetenschappen

Mei 2006

ISBN 90-8649-014-X

Dankwoord

Ik wens mijn dank te betuigen aan iedereen die er op een of andere manier toe bijgedragen hebben aan goede afloop van mijn doctoraatstudies.

Bijzondere dank gaat uit naar mijn promotor, Professor Emmanuel Lesaffre, die mij altijd deskundig begeleidde en wist te steunen bij moeilijke momenten tijdens het voorbereiden van deze proefschrift. De feit dat deze proefschrift tot vijf artikels die in de internationale wetenschappelijke tijdschriften aanvaard werden, leidde, is vooral het resultaat van zijn bekwaamheid om mij op juiste moment in de juiste richting te duwen. Hij heeft mij ook vele mogelijkheden geboden tot kontakt met andere onderzoekers op zoals nationale als internationale niveau waarvoor ik hem ook bedank.

Verder wil ik bedanken mijn co-promotor, Professor Jan Beirlant, die altijd bereid was om mij te helpen als het nodig was.

Dank ook aan alle leden van de jury, Prof. Paul Janssen, Prof. An Carbonez, Prof. Irene Gijbels, Prof. Guadalupe Gómez en Prof. Dominique Declerck voor de kritische lezing van dit werk die tot een serieuze verbetering leidde.

Mijn (ex-)collega's van het Biostatistisch centrum, Dora, Kris, Samuel, Silvia, Steffen, Geert, Dimitris, Roula, Wendim, Luwis, Alejandro, María José, Ann, Roos, Annelies, Bart en Francis wil ik voor de aangename werkomgeving gedurende de laatste vijf jaren danken. Extra dank gaat uit naar Jeannine, voor haar perfecte administratieve steun waarop ik altijd kon leunen.

Voor een boeiend jaar dat mij in het gebied van de toegepaste statistiek en vooral biostatistiek geintroduceerd heb en dat tot het begin van mijn doctoraatstudies in Leuven leidde wil ik al mijn medestudenten en lesgevers van het Biostatistiek programma aan het Limburgs Universiteit Centrum in Diepenbeek in het academiejaar 2000 – 01 bedanken.

Deze proefschrift zou niet kunnen ontstaan zonder de financiële steun van de onderzoeksbeurzen van de Katholieke Universteit Leuven. De steun van de beurzen OT/00/35, OE/03/29, DB/04/031 en BDB-B/05/10 wordt diep geapprecieerd.

Als laatst maar niet in het minst wil ik Pascale, Filip, Sibe, Ine en Wout bedanken die voor mij een familie in België wisten te creëren.

Dank u wel!
Arnošt

Poděkování

Tento text by též nikdy nemohl vzniknout bez znalostí matematiky a statistiky, kterých jsem nabyl během pregraduálních studií na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze. Za základy svého statistického vědění chci potom poděkovat všem pracovníkům Katedry pravděpodobnosti a matematické statistiky.

Michalu Kulichovi potom děkuji, že mě přemluvil, abych v roce 2000 odjel na jeden rok do Belgie, čímž změnil na dalších nejméně šest let místo mého trvalého pobytu a přispěl nepřímo k naprosté změně tématu mé doktorské práce. Profesoru Jaromíru Antochovi, svému původnímu vedoucímu doktorské práce, děkuji, že na mě i přes moji dezerci, soudě dle našich následných ROBUSTních a jiných setkání, nezanevřel.

Závěrem děkuji Lence, že si mě ponechala i přes to, že mnohý čas, který bych mohl věnovat jí, jsem věnoval statistice. Děkuji též za dvou a půl kilový dárek, který mezitím trochu narostl a kterým mi zpestřila závěr jednoho COMPSTATu. Jindře děkuji za to, že prostě je. Bez tvých úsměvů a dalších projevů přízně i nepřízně by finální práce na tomto textu nebyly zdaleka tak úsměvné jak byly.

Děkuji!
Arnošt

Acknowledgement

There would be no need to develop the techniques presented in this thesis if there were no data posing interesting questions. I would like to thank to all who collected those interesting data sets and allowed me to use them in this thesis.

Data collection for the Signal Tandmobiel® project introduced in Section 1.1 was supported by Unilever, Belgium. The Signal Tandmobiel® project comprises the following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven), K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

The WIHS data introduced in Section 1.3 were collected by the Women's Interagency HIV Study Collaborative Study Group and its Oral Substudy with centers (Principal Investigators) at New York City/Bronx Consortium (K. Anastos, J. A. Phelan); Brooklyn, NY (H. Minkoff); Washington DC Metropolitan Consortium (M. Young); The Connie Wofsy Study Consortium of Northern California (R. Greenblatt, D. Greenspan, J. S. Greenspan); Los Angeles County/Southern California Consortium (A. Levine, R. Mulligan, M. Navazesh); Chicago Consortium (M. Cohen, M. Alves); Data Coordinating Center (A. Muñoz). The WIHS is funded by the National Institute of Allergy and Infectious Diseases, with supplemental funding from the National Cancer Institute, the National Institute of Child Health & Human Development, the National Institute on Drug Abuse, the National Institute of Dental and Craniofacial Research, the Agency for Health Care Policy and Research, the National Center for Research Resources, and the Centers for Disease Control and Prevention. U01-AI-35004, U01-AI-31834, U01-AI-34994, U01-AI-34989, U01-HD-32632 (NICHD), U01-AI-34993, U01-AI-42590, M01-RR00079, and M01-RR00083. The WIHS Oral Substudy is funded by the National Institute of Dental and Craniofacial Research.

The EBCP data introduced in Section 1.4 were kindly provided by Catherine Legrand and Richard Sylvester from the European Organisation for Research and Treatment of Cancer.

Thank You!
Arnošt

The majority of the material in this thesis is based on the original publications. Below, we give a list of the parts of the thesis based principally on these publications.

Sections 5.1, 7.7: Lesaffre, E., Komárek, A., and Declerck, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, **14**, 539–552.

Section 5.2: Komárek, A., Lesaffre, E., Härkänen, T., Declerck, D., and Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored data. *Biostatistics*, **6**, 145–155.

Chapter 7: Komárek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, **14**, 726–745.

Chapter 8: Komárek, A. and Lesaffre, E. (2006a). Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution. *To appear in Statistica Sinica.*

Chapter 9: Komárek, A. and Lesaffre, E. (2006b). Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Submitted.*

Chapter 10: Komárek, A. and Lesaffre, E. (2006c). Bayesian semiparametric accelerated failure time model for paired doubly-interval-censored data. *Statistical Modelling*, **6**, 3–22.

Contents

Notation

Here, we give a list of the most often used symbols within this thesis.

δ_i	⋆ censoring indicator, ⋆ 0 for right-censored, 1 for exactly observed, 2 for left-censored, 3 for interval-censored observations;
$\mathbf{1}$	⋆ vector of ones;
$\varphi(e)$	⋆ density of $\mathcal{N}(0,\, 1)$
$\varphi(e \,\vert\, \mu,\, \sigma^2)$	⋆ density of $\mathcal{N}(\mu,\, \sigma^2)$
$\varphi_q(\boldsymbol{e} \,\vert\, \boldsymbol{\mu},\, \Sigma)$	⋆ density of q-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ
$\Phi(e)$	⋆ cumulative distribution function of $\mathcal{N}(0,\, 1)$
$\Phi(e \,\vert\, \mu,\, \sigma^2)$	⋆ cumulative distribution function of $\mathcal{N}(\mu,\, \sigma^2)$
$\lfloor t^L,\, t^U \rfloor$	⋆ interval censored observation ⋆ according to the context, the interval might be closed, half closed or open
$\oint_{t_L}^{t^U} p(s)\, ds$	$= \int_{t^L}^{t^U} p(s)\, ds \quad$ if $t^L < t^U$ $= p(t^L) = p(t^U) \quad$ if $t^L = t^U$ ⋆ symbol used to write down the likelihood of the interval censored data

Preface

The accelerated failure time (AFT) model, the principal topic of this thesis is a regression model used to analyze survival data. The term *survival data* is usually used for data that measure the time to some event, not necessarily death. Precisely, *the event time* will be considered a positive real valued variable having a continuous distribution. In particular practical situations, data on event times are obtained by following subjects in the study over (calendar) time, recording the moments of the specified events of interest and computing the time spans between the event and some initial - *onset time* (e.g. enter to a study and disease progression, contagion by HIV virus and onset of AIDS, tooth emergence and the time it is attacked by caries for the first time).

A typical feature of survival data is the fact that the time to event is not always observed completely and observations are imposed to *censoring*. Most commonly, either the study is finished before all subjects involved encounter the specified event or the subject leaves for some reasons the study before encountering the event. In both situations, only the lower limit for the true event time is known and we talk about *right censoring* (see Sections 1.2 and 1.4 for examples).

In many areas of medical research, the occurrence of the event of interest can only be recorded at planned (or unplanned) visits. The exact event time is then only known to happen between two examination times (visits) and we encounter *interval censoring*. Typical examples are (a) time to caries development; (b) time emergence of a tooth (Section 1.1); (c) time to HIV seroconversion; (d) time to the onset of AIDS (Section 1.3). Indeed, in case of a cavity or of emergence the event is often observed after some delay, say at planned (or even unplanned) visits. Similarly, HIV seroconversion can only be determined by regular or irregular laboratory assessments. However, the

event may also happen before the first examination (e.g. a decayed tooth is detected already at the first dental examination) and we get a so called *left-censored* observation or it may happen after the last examination resulting in a right-censored observation. Hence interval censoring is a natural generalization of the commonly encountered right censoring.

Often not only the event time but also the time which specifies the origin of the time scale for the event (the onset time) can only be recorded in the same way as described in the previous paragraph. An example is the time to caries development on a tooth where the time of tooth emergence constitutes the onset time for caries (see Section 1.1). We then speak of *doubly interval censoring*. We further formalize the notion of censoring in Chapter 2

Furthermore, the independence between the event times cannot always be assumed thereby entering the area of *multivariate survival data*. The dependence can be caused by very different factors. Although many methods described in this thesis can be applied to any multivariate survival data the dependencies in our applications are all result of some type of *clustering*: emergence or caries times of several teeth of one child (Section 1.1), or progression free survival times of several patients within one hospital in a multicenter clinical trial (Section 1.4). Also *recurrent* infection times on one patient (Section 1.2) can be considered to result in clustered data.

The ultimate goal of the research presented in this thesis was to develop the AFT models which can be used to analyze multivariate survival data, possibly under the presence of doubly interval censoring. The scale of complexity considered in this thesis starts with interval censoring which can be handled by all methods introduced here. Possible dependencies between the observations (multivariate survival data) are viewed as the next step on the scale of complexity and finally, doubly interval censoring is regarded to be the final level of complexity treated by this thesis and only some methods shown here reached this final stage. With all the levels of complexity we strived for the model with distributional assumptions as flexible as possible. Two slightly different directions are followed in the thesis to address this issue. Both of them use a Gaussian mixture as a building block to model an unknown distribution. Whereas the first and more extensively explored approach uses the mixture with a higher number of fixed mixture components with mixture weights estimated using a kind of penalized methodology the second technique uses a classical mixture with both the number as well as the weights, locations and scales of the mixture components unknown.

Chapter 1 introduces several data sets that contain each survival data involving one or more issues discussed above and that will be used throughout the thesis to illustrate the developed methods. Terminology and notation

used in the thesis are formalized in Chapter 2 together with an explanation of some basic notions in the analysis of the survival data. The most popular regression models for the survival data are introduced in Chapter 3.

In Chapter 4 we give the likelihood for interval and doubly-interval-censored data and discuss briefly the difficulties encountered when using maximum-likelihood methods in the context of (doubly) interval-censored data. Subsequently, we show how the Bayesian inference together with the Markov chain Monte Carlo (MCMC) methodology can simplify the calculations.

Available methods for the analysis of interval-censored data will be reviewed in Chapter 5 and one of the methods, namely the Bayesian proportional hazards model with a piecewise constant baseline hazard function will be applied to the analysis of the dental clustered doubly-interval-censored data.

In Chapter 6 we explain in detail how the classical and the penalized normal mixtures can be used to specify unknown distributions in a flexible way.

The first AFT model presented in this thesis – the AFT model with an error distribution being a normal mixture with a high number of fixed components estimated using the penalized maximum-likelihood method – is shown in Chapter 7. However only univariate interval-censored data can be handled by this method. To move on to the area of multivariate or even doubly-interval-censored survival data we found it more advantageous to use a Bayesian methodology rather than the more classical maximum-likelihood based techniques. The Bayesian AFT model allowing for multivariate interval-censored data and using a classical normal mixture with both unknown number of mixture components as well as the mixture components themselves to specify the error distribution is presented in Chapter 8. Finally, Chapters 9 and 10 show the Bayesian AFT models suitable for multivariate doubly-interval-censored data that exploit a penalized normal mixture with higher number of fixed components. For all methods described in this thesis, software was written in the form of R (R Development Core Team, 2005) packages called smoothSurv and bayesSurv downloadable from *the Comprehensive R Archive Network* at `http://www.R-project.org`. The software is briefly described in Appendix C.

Part I

Introduction

Chapter 1

Motivating Data Sets

This chapter introduces the data sets which will be used throughout the thesis illustrating the developed techniques and showing their generality. Each data set involves one or more specific features of interest here, discussed briefly in the Preface. The Signal Tandmobiel® data set introduced in Section 1.1 involves clustered interval- and doubly-interval-censored dental observations. Section 1.2 describes a clinical trial with patients with a chronic granulomatous disease where times of possibly recurrent infections were of interest. At the same time, the time of the last infection is right-censored. The Women's Interagency HIV Study involved interval-censored data and is described in Section 1.3. In Section 1.4, a multicenter clinical trial is described which evaluated the effect of perioperative chemotherapy on disease progression in early breast cancer patients where the heterogeneity accross the centra plays an important role.

1.1 The Signal Tandmobiel® study

The Signal Tandmobiel® project is a longitudinal oral health study performed in Flanders from 1996 to 2001. It involved 4 468 schoolchildren (2 315 boys and 2 153 girls) born in 1989. Two stratification factors, i.e. geographical location (5 provinces) and educational system (3 school systems) establishing 15 strata, were taken into account. The sample represented about 7% of the corresponding Flemish population of school children. Detailed oral health data at tooth and tooth-surface level (caries experience, gingivitis, etc.) were annually collected by a team of 16 dentists whose examination method was calibrated every six months. In addition, data on dietary and oral hygiene

habits were collected using a questionnaire completed by the parents. Hence the data set consists of a series of at most 6 longitudinal dental observations and reported oral health habits. The details of the study design and research methods have been described in detail by Vanobbergen et al. (2000).

Here, we concentrate on the emergence and caries times of permanent premolars and molars (teeth $\kappa + 4$, $\kappa + 5$, $\kappa + 6$, $\kappa = 10, 20, 30, 40$ in European dental notation, see Figure 1.1). There is no doubt that an adequate knowledge of timing and patterns of tooth emergence and/or caries attacks are still essential for diagnosis and treatment planning in paediatric dentistry and orthodontics. Additionally, the effect of certain prespecified factors (like the caries status of the primary teeth – see Figure 1.2 for their notation, use of fluoride supplements, brushing habits etc.) on the emergence or caries processes are often of interest.

An interesting feature of this data set, though typical in dental applications, is the fact that both emergence and onset of caries are only observable when the child is examined (by a dentist). This leads to interval-censored emergence times and to doubly-interval-censored times for caries (see also Figure 2.1). Additionally, the teeth of a single mouth share common immeasurable or

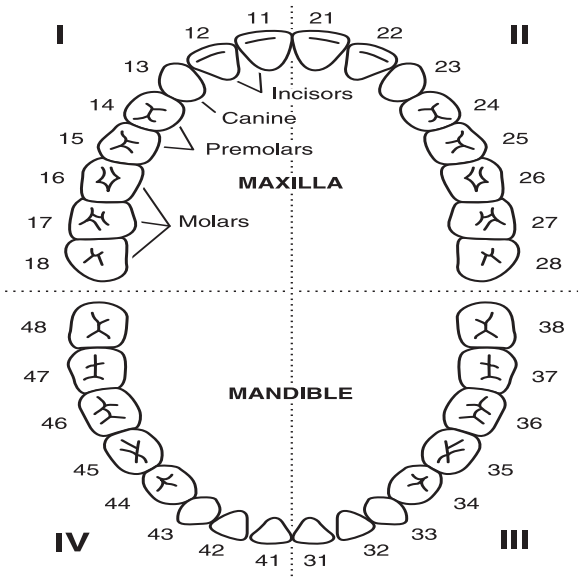


Figure 1.1: European notation for the position of permanent teeth. Maxilla = upper jaw, mandible = lower jaw. The first and the fourth quadrants are at the right-hand side of the subject, the second and the third quandrats are at the left-hand side of the subject.

only roughly measured factors like genetical dispositions or dietary habits. As a result, the emergence times or the times to caries of teeth in the same mouth are related. Hence, when studying the emergence time or the time to caries of several teeth, dependencies among the observations taken on a single child must be taken into account. Analysis of the emergence time or time to caries is reported in several sections of the thesis.

1.2 The Chronic Granulomatous Disease trial (CGD)

The Chronic Granulomatous Disease is a group of inherited rare disorders of the immune function characterized by *recurrent* pyogenic infections which may lead to death in childhood. There is evidence of a positive role of gamma interferon in restoring the immune functions of the patients. For that reason, a multicenter placebo-controlled randomized trial was conducted to study the ability of gamma interferon to reduce the rate of serious infections.

Between October 1988 and March 1989, 128 patients (63 taking gamma inter-

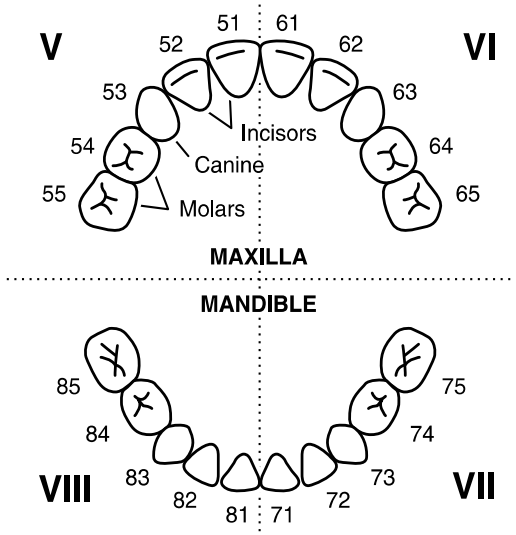


Figure 1.2: European notation for the position of deciduous (primary) teeth. The quadrants are numbered 5, 6, 7, 8. The fifth and the eight quadrants are at the right-hand side of the subject, the sixth and the seventh quadrants are at the left-hand side of the subject.

feron, 65 taking placebo) with CGD were accrued by 13 hospitals in Europe and the United States. The average follow-up time was 292 days, minimal and maximal follow-up times, respectively were 91 and 432 days, respectively. For each patient, times of initial and any recurrent serious infections were recorded. There is a minimum of one and a maximum of eight recurrent infection times per patient, with a total of 203 records.

Besides the gamma interferon there are other factors that may influence the times between the infections. In the course of the study the following additional information was recorded for each patient:

- Age at time of study entry (mean 14.6 years, range from 1 to 44 years, standard deviation 9.8 years);
- Gender: *male* ($n = 104$), *female* ($n = 24$);
- Pattern of inheritance: *autosomal recessive* ($n = 42$), *X-linked* ($n = 86$);
- Using corticosteroids at time of study entry: *yes* ($n = 3$), *no* ($n = 125$);
- Using prophylactic antibiotics at time of study entry: *yes* ($n = 111$), *no* ($n = 17$);
- Category of the hospital: *US – NIH* ($n = 26$), *US – other* ($n = 63$), *Europe – Amsterdam* ($n = 19$), *Europe – other* ($n = 20$).

The data can be found in Appendix D.2 of Fleming and Harrington (1991).

It is of interest here to set up a regression model with the time between the two consecutive infections as response and mentioned factors as covariates. It should be taken into account that the infection times of one patient cannot be assumed to be independent. We address this issue in Section 8.8.

1.3 The Woman's Interagency HIV Study (WIHS)

The Woman's Interagency HIV Study comprises the cohort of 2 058 seropositive women with a comparison cohort of 568 seronegative women being exposed to a higher risk of HIV infection than the common U.S. population. The study groups were enrolled between October 1994 and November 1995 through six clinical consortia at 23 sites throughout the United States. Barkan et al. (1998) provide full details on the setup of the study. In this thesis we concentrate only on the WIHS Oral Substudy involving 224 seropositive AIDS-free (at baseline) women.

The women participating in the Oral Substudy were regularly (on average every 7 months) examined for AIDS symptoms, the number of copies of

the HIV RNA virus (viral load) and CD4 T-lymphocyte counts per *ml* of blood. Additionally, the presence of one of the three oral lesion markers: oral candidiasis, hairy leukoplakia and angular cheilitis was checked. The average follow-up time was 41 months and the maximal follow-up time was 84 months. For each woman, the time of seroconversion (HIV infection) was externally estimated and assumed to be known. Clinical AIDS diagnoses were self-reported in 73.5% of cases, presumptive or definite in 17.5%, and indeterminate in 9%; the case definition did not depend on CD4 T-lymphocytes. For 66 women the onset of AIDS was interval-censored, while for 158 women it was right-censored.

For HIV positive people, it is of interest to describe the distribution of the time to the onset of an AIDS-related illness based on some measured quantities. We examine in Section 7.6 how the classical predictors like viral load and CD4 T-cells counts together with oral lesion markers can be used in describing this distribution.

1.4 Perioperative Chemotherapy in Early Breast Cancer Patients (EBCP)

To investigate whether a short intensive course of perioperative chemotherapy can change the course of early breast cancer compared to surgery alone, the European Organization for Research and Treatment of Cancer (EORTC) conducted a multicenter randomized clinical trial (EORTC Trial 10854). From 1986 to 1991, a total of 2 793 women with early breast cancer were randomized to receive either one perioperative course of an anthracycline-containing chemotherapeutic regimen within 24 h after surgery ($n = 1\,398$) or surgery alone ($n = 1\,395$). See Clahsen et al. (1996) for more details on the trial.

Patients were followed-up for several endpoints, however, we concentrate on the progression-free survival (PFS) time. The mean follow-up time was 8.15 years with a maximum of 14.13 years. Other factors that may influence the PFS time include:

- Category of the age of the patient: *<40 years* ($n = 321$), *40–50 years* ($n = 796$), *>50 years* ($n = 1\,676$);
- Type of surgery: *mastectomy* ($n = 1\,231$), *breast-conserving surgery* ($n = 1\,542$), missing data for $n = 20$ patients;
- Category of the tumor size: *<2 cm* ($n = 823$), *≥2 cm* ($n = 1\,915$), missing data for $n = 55$ patients;

- Pathological nodal status: *negative* ($n = 1\,467$), *positive* ($n = 1\,303$), missing data for $n = 23$ patients;
- Presence of other disease: *no* ($n = 2\,542$), *yes* ($n = 234$), missing for $n = 19$ patients.

The trial was conducted in 14 centra located in 5 geographical regions (the Netherlands, Poland, France, South of Europe and South Africa). Figure 1.3 shows Kaplan-Meier estimates of PFS time survival functions for the treatment and control group, separately for each center. Obviously, there is a huge heterogeneity among the centra. Not only the overall proportion of PFS patients at fixed time points differs from center to center but also the effect of treatment on PFS both quantitatively and qualitatively seems to vary accross centra. Models that measure the effect of covariates and that allow modelling heterogeneity between centra will be considered in Chapters 8 and 9.

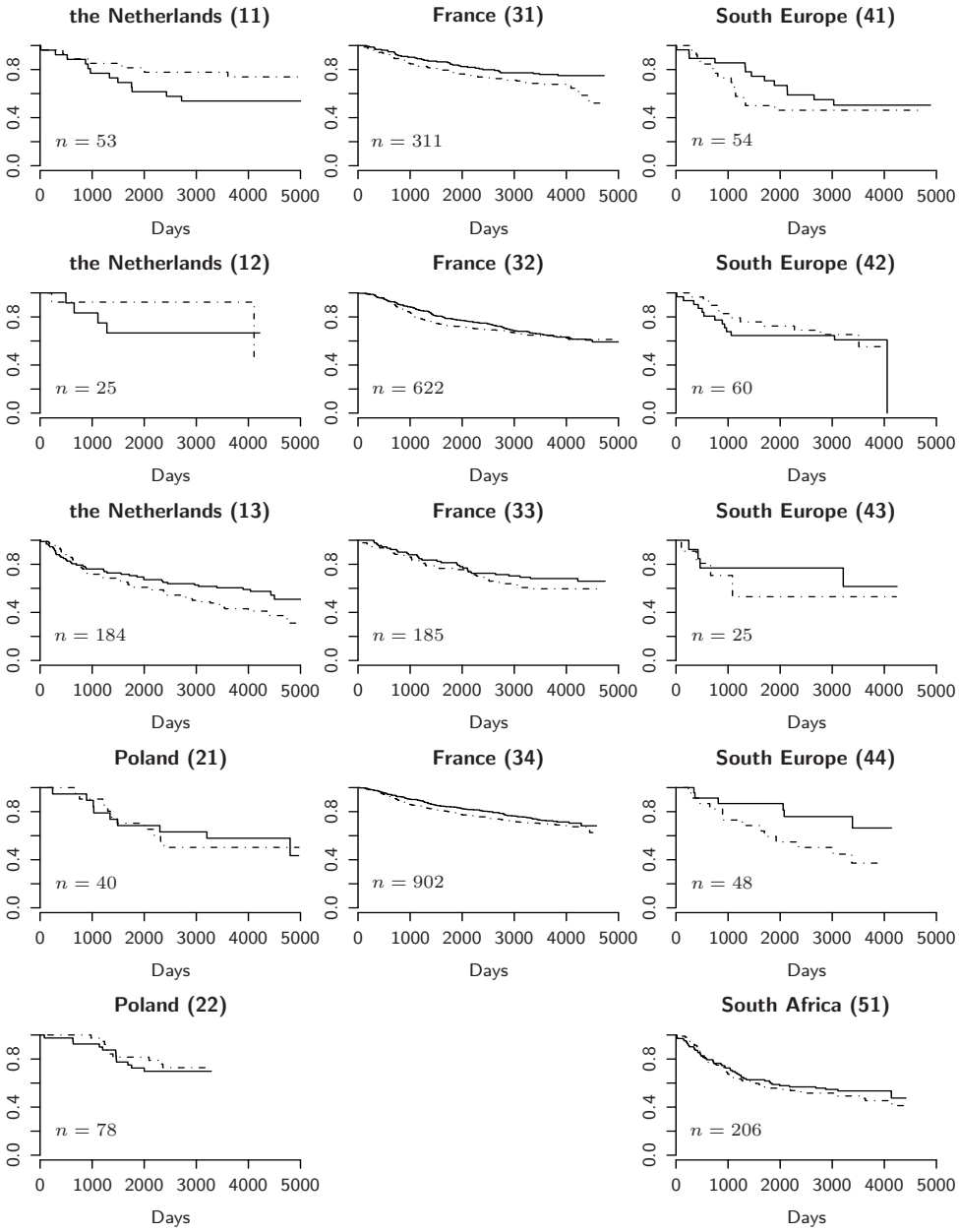


Figure 1.3: EBCP Data. Kaplan-Meier estimates of the PFS time distribution separately for each institution. Solid line: treatment arm, dotted-dashed line: control arm.

Chapter 2

Basic Notions

In this chapter we introduce some notation that will be used throughout the thesis and explain more in detail some basic notions like types and mechanisms of censoring considered.

2.1 Right, left and interval censoring

Let $T_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ be the exact *event time* for the lth observational unit of the ith cluster. It will be assumed throughout the thesis that $T_{i,l}$ is a nonnegative random variable with a continuous distribution with some density $p_{i,l}(t)$ which might depend on a vector of covariates, e.g., $\boldsymbol{x}_{i,l} = (x_{i,l,1}, \ldots, x_{i,l,m})'$. The time $T_{i,l}$ can either be known exactly or in a coarsened manner and is then called *censored*. Suppose first that knowing whether the event occurred or not requires a detailed examination (visit to a dentist, laboratory assessment) executed at pre-planned visits. Then it is only known that the event time occurred after, say $t^L_{i,l}$, and before, say $t^U_{i,l}$. According to the context, we either know $t^L_{i,l} < T_{i,l} \leq t^U_{i,l}$, $t^L_{i,l} \leq T_{i,l} < t^U_{i,l}$, $t^L_{i,l} \leq T_{i,l} \leq t^U_{i,l}$, or $t^L_{i,l} < T_{i,l} < t^U_{i,l}$. Thus, the true event time $T_{i,l}$ is known to lie in the interval whose lower and upper limits are equal to $t^L_{i,l}$ and $t^U_{i,l}$, respectively and the observation is called *interval-censored*. Note that all methods presented in Part II of the thesis lead to the same results irrespective of whether the interval is closed, open or half open. To cover all these situations we will write $T_{i,l} \in \lfloor t^L_{i,l}, t^U_{i,l} \rfloor$.

With the same notation *right-censored* observations are obtained, i.e. by setting $t^U_{i,l} = \infty$ and $t^L_{i,l}$ equal to the time the subject was last seen before leaving the study or before the study was terminated. Similarly, a *left-censored* ob-

servation is obtained with $t^L_{i,l} = 0$ and $t^U_{i,l}$ equal to the first time, the subject was seen after the event. Finally, an exactly observed time $t_{i,l}$ is recorded with $t^L_{i,l} = t^U_{i,l} = t_{i,l}$. Below, a censoring indicator $\delta_{i,l}$ is used, which will be equal to 0 for right-censored, 1 for exactly observed, 2 for left-censored and 3 for interval-censored observations, respectively.

2.2 Doubly interval censoring

Suppose that the event time $T_{i,l}$ is obtained as the difference of two random variables: $V_{i,l}$, here always called *the failure time* and $U_{i,l}$, here always called *the onset time*, i.e. $T_{i,l} = V_{i,l} - U_{i,l}$. The pair $U_{i,l}, V_{i,l}$ can be, for example, the emergence time of a tooth and the onset time of caries of that tooth. Doubly interval censoring is obtained in the situations when either $U_{i,l}$ and/or $V_{i,l}$ are interval-censored and it is only known $U_{i,l} \in \lfloor u^L_{i,l}, u^U_{i,l} \rfloor$ and $V_{i,l} \in \lfloor v^L_{i,l}, v^U_{i,l} \rfloor$. A scheme of a typical doubly-interval-censored observation is given in Figure 2.1 and an example is given by the Signal Tandmobiel® data of Section 1.1 with $U_{i,l}$ being the emergence time of the lth tooth of the ith child and $V_{i,l}$ being the time when the same tooth is attacked by caries for the first time.

In the following, we omit the subscript (i, l) from all expressions if it is not necessary to make an explicit distinction among different observations of one data set or use only a single subscript i if we do not deal with multivariate survival data.

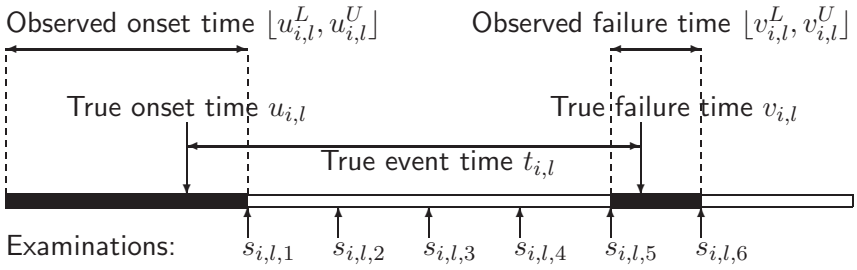


Figure 2.1: Doubly interval censoring. A scheme of a doubly-interval-censored observation obtained by performing examinations to check the event status at times $s_{i,l,1}, \ldots, s_{i,l,6}$. The onset time is left-censored at time $u^U_{i,l} = s_{i,l,1}$ (i.e. interval-censored in the interval $\lfloor u^L_{i,l}, u^U_{i,l} \rfloor = \lfloor 0, s_{i,l,1} \rfloor$), the failure time is interval-censored in the interval $\lfloor v^L_{i,l}, v^U_{i,l} \rfloor = \lfloor s_{i,l,5}, s_{i,l,6} \rfloor$.

2.3 Density, survival, hazard and cumulative hazard functions

A continuous distribution of an event time T is uniquely determined by its density $p(t)$. Equivalently, the distribution of T is determined by a non-increasing right-continuous *survival function* $S(t)$ defined as the probability that T exceeds a value t in its range, i.e.

$$S(t) = \Pr(T > t) = \int_t^\infty p(s)\, ds.$$

Another possibility is to specify the *hazard function* $\hbar(t)$ which gives the instantaneous rate at which an event occurs for an item that is still at risk for the event at time t, i.e.

$$\hbar(t) = \lim_{\Delta t \to 0_+} \frac{\Pr(t \le T < t + \Delta t \,|\, T \ge t)}{\Delta t} = \Pr\big(T \in N_t(dt) \,|\, T \ge t\big),$$

where

$$N_t(dt) = [t,\, t + dt).$$

The density and the survival function can be computed from the hazard function using the following relationships:

$$\begin{aligned} p(t) &= \hbar(t)\, \exp\big\{-H(t)\big\}, \\ S(t) &= \exp\big\{-H(t)\big\}, \end{aligned}$$

where $H(t) = \int_0^t \hbar(s)\, ds$ is the *cumulative hazard function*.

2.4 Independent noninformative censoring and simplified likelihood

Throughout the thesis we will assume independent noninformative censoring in the therminology of Kalbfleisch and Prentice (2002). In this section, we explain this concept first in the framework of right-censored data and then extend it to the area of interval-censored data. Finally, we introduce the term of simplified likelihood and remark that it can be used for the inference with censored data under the assumption of independent noninformative censoring.

2.4.1 Right-censored data

Kalbfleisch and Prentice (2002) introduce the concept of independent noninformative censoring in the context of right-censored data in the following way. Let C denote the random variable causing the censoring. That is, instead of observing the event time T we only observe $X = \min(T, C)$ and $\delta = I[T \leq C]$.

Independent censoring

They call the censoring mechanism *independent* when the hazard which applies to the censored population is at each time point the same as the hazard which applies would there have been no censoring. That is, the hazard functions have to satisfy

$$\Pr\big(T \in N_t(dt) \,\big|\, C \geq t,\, T \geq t\big) = \Pr\big(T \in N_t(dt) \,\big|\, T \geq t\big) \tag{2.1}$$

for any $t > 0$. Note that independence of random variables T and C implies that the condition (2.1) is satisfied. However, T and C are not necessarily independent when the condition (2.1) is fulfilled.

Further, Kalbfleisch and MacKay (1979) proved that the condition (2.1) is equivalent to so called *constant-sum* condition:

$$\Pr\big(\delta = 1 \,\big|\, T \in N_t(dt)\big) + \int_0^t \Pr\big(C \in N_x(dx),\, \delta = 0 \,\big|\, T \geq x\big) = 1 \tag{2.2}$$

for any $t > 0$, introduced by Williams and Lagakos (1977). The term $\Pr\big(\delta = 1 \,\big|\, T \in N_t(dt)\big)$ could be interpreted as the probability that a subject who would fail at time t is actually observed to fail and the term $\Pr\big(C \in N_x(dx),\, \delta = 0 \,\big|\, T \geq x\big)$ has the meaning that a subject who survives at least x time units is censored at time x. To relate the condition (2.2) to its interval-censored version which will be introduced in the following section, we rewrite it into the form:

$$\Pr\big(\delta = 1 \,\big|\, T \in N_t(dt)\big) + \int_0^t \frac{\Pr\big(C \in N_x(dx),\, T \in [x,\, \infty),\, \delta = 0\big)}{\Pr\big(T \in [x,\, \infty)\big)} = 1. \tag{2.3}$$

Noninformative censoring

Kalbfleisch and Prentice (2002) further call the censoring mechanism *noninformative* if the censoring random variable C does not depend on any parameters used to model the distribution of the event time T. In other words,

with the independent noninformative censoring, the censoring procedure or rules may depend arbitrarily during the course of the study on:

- previous event times of other subjects in the study;
- previous censoring times of other subjects in the study;
- random mechanisms external to the study;
- values of covariates possibly included in the model;

but must not contain any information on the parameters used to model the event time.

The independent noninformative censoring includes *type I censoring*. In this case, censoring can only happen at a pre-planned calendar time. This censoring scheme has been used for the CGD data introduced in Section 1.2 and for the EBCP data of Section 1.4.

2.4.2 Interval-censored data

Consider now the case of interval-censored data where the observed intervals are generated by a triplet $(T^L, T^U, T)'$. That is, we observe an interval $\lfloor t^L, t^U \rfloor$ if $T^L = t^L$, $T^U = t^U$ and $T \in \lfloor T^L, T^U \rfloor$. Note that since the observed interval $\lfloor T^L, T^U \rfloor$ must contain the event time T, the support of the random vector $(T^L, T^U, T)'$ is equal to

$$\big\{(t^L, t^U, t) : 0 \leq t^L \leq t \leq t^U \leq \infty\big\}.$$

Oller, Gómez, and Calle (2004) show that the interval-censored counterpart of the constant-sum condition (2.3) is given by

$$\iint\limits_{\left\{(t^L, t^U): t \in \lfloor t^L, t^U \rfloor\right\}} \frac{\Pr\big(T^L \in N_{t^L}(dt^L),\, T^U \in N_{t^U}(dt^U),\, T \in \lfloor t^L, t^U \rfloor\big)}{\Pr\big(T \in \lfloor t^L, t^U \rfloor\big)} = 1 \qquad (2.4)$$

for all $t > 0$. Further, they introduce the term *noninformative condition* and show that it is stronger than the constant-sum condition (2.4). It should be pointed out that Oller et al. use the term "noninformative" in a different context than Kalbfleisch and Prentice (2002) whose meaning of this word is adopted in this thesis.

In summary, we will call the interval censoring *independent* if it satisfies the constant-sum condition (2.4) and *noninformative* if the distribution of censoring random variables T^L and T^U does not depend on the parameters used to model the distribution of the event time T.

A typical example of an independent noninformative interval censoring can be found in the Signal Tandmobiel® data (Section 1.1) and in the WIHS data (Section 1.3). In both cases even a stronger condition of independence of T and $(T^L, T^U)'$ is satisfied. Indeed, either dental examinations or check ups of the AIDS status were performed at pre-planned time-points and thus external to the studied event time. Note that interval censoring would not be independent when the event induces an examination, namely when a child visits the dentist because of a decayed tooth.

2.4.3 Simplified likelihood for interval-censored data

We explain in Chapter 4 that *likelihood* is the corner stone for the inference on the event time T. Strictly speaking, with interval-censored data, the likelihood contribution is given by the density of observables, i.e. by the density of the vector $(T^L, T^U)'$ whose support is such that $T \in \lfloor T^L, T^U \rfloor$ with probability one. That is, the likelihood contribution of the observed $\lfloor t^L, t^U \rfloor$ is given by

$$L_{full} = \Pr\big(T^L \in N_{t^L}(dt^L),\, T^U \in N_{t^U}(dt^U),\, T \in \lfloor t^L, t^U \rfloor\big).$$

However, it is shown in Oller et al. (2004) that under the assumption of independent noninformative censoring, the likelihood contribution L_{full} is proportional to so called *simplified likelihood* contribution

$$L = \Pr\big(T \in \lfloor t^L, t^U \rfloor\big),$$

where a possible randomness of T^L and T^U is ignored. Consequently, the inference on the event time T can be based on this simplified likelihood. In the remainder of the thesis, we will use the simplified likelihood for the inference and omit the word 'simplified' for clarity.

Chapter 3

An Overview of Regression Models for Survival Data

Two regression models dominate the survival analysis to describe the dependence of the distribution of the event time T on covariates, say $\boldsymbol{x} = (x_1, \ldots, x_m)'$: (a) the proportional hazards (PH) model and (b) the accelerated failure time (AFT) model. In this chapter, we introduce these two models, compare them and show how they can be extended to handle multivariate survival data. We also review these models for the analysis of right-censored data however with an emphasis on the AFT model. For methods that allow interval- or doubly-interval-censored data we refer to Chapter 5.

3.1 Proportional hazards model

This model, introduced by Cox (1972), specifies that, for a given covariate vector $\boldsymbol{x}$, the hazard function is expressed as the product of an unspecified baseline hazard function $\hbar_0(t)$ and the exponential of a linear function of the covariates, i.e.

$$\hbar(t \,|\, \boldsymbol{x}) = \hbar_0(t) \, \exp(\boldsymbol{\beta}'\boldsymbol{x}). \tag{3.1}$$

The regression parameter vector $\boldsymbol{\beta}$ is estimated by maximizing a partial likelihood (Cox, 1975) which treats $\hbar_0$ as nuisance and does not estimate it. However, when the baseline hazard $\hbar_0$ is of interest as well, e.g. for prediction purposes, its non-parametric estimate can be obtained using the method of Breslow (1974). The survival function for an object with covariates $\boldsymbol{x}$,

$S(\cdot \,|\, \boldsymbol{x})$, is related to the baseline survival function S_0 by the relationship

$$S(t \,|\, \boldsymbol{x}) = \big\{S_0(t)\big\}^{\exp(\boldsymbol{\beta}'\boldsymbol{x})}.$$

An exhaustive treatment of the PH model and its extensions can be found, e.g., in Therneau and Grambsch (2000) or Kalbfleisch and Prentice (2002, Chapter 4). The software to fit the PH model using the method of maximal partial likelihood together with possibilities to compute residuals, draw diagnostic plots or assess goodness of fit is available in most modern statistical packages, e.g. function `coxph` in R/S-PLUS or procedure `PHREG` in SAS.

3.2 Accelerated failure time model

The accelerated failure time model is a useful, however less frequently used alternative to the PH model. For this model, the effect of a covariate implies on average an acceleration or deceleration of the event time. For a vector of covariates $\boldsymbol{x}$ the effect is expressed by the parameter vector $\boldsymbol{\beta}$ in the following way:

$$T = \exp(\boldsymbol{\beta}'\boldsymbol{x})\, T_0,$$

where T_0 is a baseline survival time. On the logarithmic scale, this model becomes a simple linear regression model

$$\log(T) = \boldsymbol{\beta}'\boldsymbol{x} + \varepsilon, \tag{3.2}$$

with $\varepsilon = \log(T_0)$. The hazard and survival functions of a subject with covariate vector $\boldsymbol{x}$ is related to the baseline hazard ($\hbar_0$) and survival function (S_0) by the relationships

$$\begin{aligned} \hbar(t \,|\boldsymbol{x}) &= \hbar_0\big\{\exp(-\boldsymbol{\beta}'\boldsymbol{x})\, t\big\}\, \exp(-\boldsymbol{\beta}'\boldsymbol{x}), \\ S(t \,|\, \boldsymbol{x}) &= S_0\big\{\exp(-\boldsymbol{\beta}'\boldsymbol{x})\, t\big\}. \end{aligned} \tag{3.3}$$

Usually one assumes that the error random variable ε has a density $g_\varepsilon(\varepsilon)$ from the location-scale family, i.e. $g_\varepsilon(\varepsilon) = \tau^{-1} g_\varepsilon^*\big\{\tau^{-1}(\varepsilon - \alpha)\big\}$, where $g_\varepsilon^*(\cdot)$ has location parameter $= 0$ and scale parameter $= 1$. The location parameter α and the scale parameter τ have to be estimated from the data as well as the regression parameter $\boldsymbol{\beta}$.

A parametric AFT model assumes that $g_\varepsilon^*(\cdot)$ is a density of a specific type (e.g. Gaussian, logistic or Gumbel). In that case, the parameters α, τ and $\boldsymbol{\beta}$ can easily be estimated using the method of maximum likelihood. However, the parametric assumptions affect evidently the shape and character

of the resultant survival or hazard curves which, in the case of an incorrect specification, is undesirable, especially when prediction is of interest.

On the other hand, semi-parametric procedures for the AFT model leave the density $g_\varepsilon(\varepsilon)$ unspecified and provide only the estimate of the regression parameter vector $\boldsymbol{\beta}$. In the past, primarily two semi-parametric methods for the AFT model with right-censored data have been examined. The first one is based on the generalization of the least squared method to censored data first proposed by Miller (1976) and in a different manner by Buckley and James (1979) giving their names to this approach. A slight modification of the Buckley-James estimator and its asymptotic properties was given by Lai and Ying (1991). However, a drawback of the Buckley-James method is that it may fail to converge or may oscillate between several solutions.

The second approach is based on linear-rank-tests for censored data and was developed by Prentice (1978), Gill (1980), and Louis (1981) in the case of one covariate. Tsiatis (1990) extended the method to the multiple regression context. The asymptotic equivalence of the Buckley-James method and the linear-rank-test-based estimators has been pointed out by Ritov (1990). The asymptotic properties of the linear-rank-test-based estimators were presented in greatest generality by Ying (1993). In contrast to the partial likelihood method for the PH model, the numerical aspect of the linear-rank-test-based estimation of the regression parameters of the AFT model could be computationally cumbersome. Only recently, Jin et al. (2003) suggested an algorithm to compute this estimate using a linear programming technique. They also provide an S-PLUS function. Further, there seems to exist no non-parametric method to estimate the baseline survival distribution like the method of Breslow (1974) for the PH model. Consequently, the semi-parametric procedures cannot be used when prediction is of interest.

Only parametric AFT models have been implemented in major statistical packages (functions `survreg` in R and `SurvReg` in S-PLUS and procedure `LIFEREG` in SAS).

3.3 Accelerated failure time model versus proportional hazards model

Both the PH as well as the AFT model make an explicit assumption about the effect of covariates on the hazard function. The effect of covariates on the hazard function in the PH model is given by (3.1), in the AFT model by (3.3). The assumed different effect of a covariate on the baseline hazard for the PH and AFT model is exemplified in Figure 3.1. It is seen that, like in the

PH model, in the AFT model the effect of covariates on the baseline hazard function is multiplicative, but additionally for the AFT model an acceleration or deceleration of the time scale is seen. Also, in the AFT model the hazard is increased for $\beta < 0$ whereas in the PH model for $\beta > 0$.

We point out (see Kalbfleisch and Prentice, 2002, Section 2.3.4) that the PH model and the AFT model are equivalent if and only if the distribution of the standardized error term $\varepsilon^* = \tau^{-1}(\varepsilon - \alpha)$ in the AFT model (3.2) is the Gumbel (extreme value distribution of a minimum), i.e. when

$$g^*_\varepsilon(\varepsilon^*) = \exp\{\varepsilon^* - \exp(\varepsilon^*)\}.$$

In that case, the distribution of the baseline survival time T_0 is Weibull and the baseline hazard function $\hbar_0(t)$ has the form

$$\hbar_0(t) = \gamma\,(\lambda\,t)^{\gamma-1},$$

where $\lambda = \exp(-\alpha)$ and $\gamma = \tau^{-1}$.

Further, it is generally true that it is not always possible (e.g. due to lack of knowledge) to include all relevant covariates in the model. One of the advantages of the AFT model is that the regression parameters of the included covariates do not change when other, important, covariates are omitted. Of course, the neglected covariates have an impact on the distribution of the error term ε in (3.2) hich is typically changed into one with larger variability. Such change, however, is of no major importance (except that it influences the precision with which the regression parameters of the included covariates are estimated) when semi-parametric methods or methods with a flexible distribution for ε are used. Unfortunately, a similar property does not hold for the PH model, see Hougaard (1999) for a more detailed discussion.

The fact that only parametric AFT models are implemented in major statistical packages, together with the computational difficulties associated with the semi-parametric AFT model may have caused that the PH model became far more popular in practice than the AFT model. See Nardi and Schemper (2003) for comparison of the PH model and parametric AFT models. Though, the property that the AFT model postulates a direct relationship between failure time and covariates led Sir David Cox (see Reid, 1994) to remark that "accelerated life models are in many ways more appealing" than the proportional hazards model "because of their quite direct physical interpretation." Indeed, in the AFT model, the regression indicates directly how is the time – a quantity being understandable also by non-statisticians – increased or decreased. Whereas, in the PH model, the direct effect of regression is on the hazard which might be more difficult to understand by practitioners.

3.4 Regression models for multivariate survival data

Both the PH model and the AFT model can be extended to handle multivariate survival data. In this section, we briefly discuss one extension of the PH model and concentrate mainly on the multivariate versions of the AFT model that will serve as a basis for developments presented in this thesis.

3.4.1 Frailty proportional hazards model

For multivariate survival data, a common extension of the PH model includes a cluster specific random effect Z_i, called the *shared frailty*, in the expression of the hazard function, i.e.

$$\hbar(t \,|\, \boldsymbol{x}_{i,l},\, Z_i) = \hbar_0(t)\; Z_i\; \exp(\boldsymbol{\beta}'\boldsymbol{x}_{i,l}). \tag{3.4}$$

The frailty component Z_i is most often assumed to have a parametric distribution such as a gamma or log-normal distribution. For more details, we refer to Aalen (1994), Hougaard (2000) and Therneau and Grambsch (2000) where also available software is described.

Nevertheless, the model (3.4) is rather simple, e.g., in the analysis of a multicenter clinical trial only the center effect and not the center by treatment

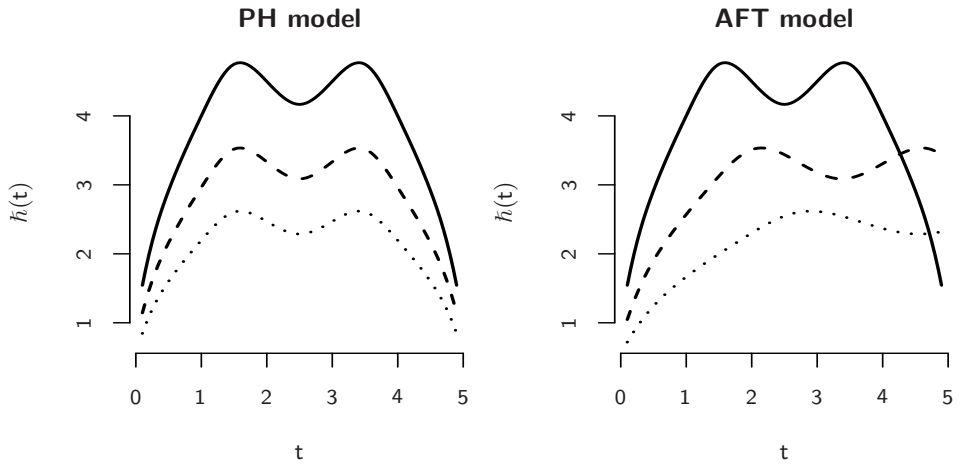


Figure 3.1: Effect of PH and AFT assumption on a hypothetical baseline hazard function (solid line) for a univariate covariate x taking a value of 0.6 (dashed line) and 1.2 (dotted line) with regression parameter $\beta = -0.5$ for the PH model and $\beta = 0.5$ for the AFT model.

interaction can be controlled for. This drawback led to further developments mimicking the classical linear mixed model of Laird and Ware (1982) by assuming

$$\hbar(t \,|\, \boldsymbol{x}_{i,l},\, \boldsymbol{b}_i) = \hbar_0(t)\ \exp(\boldsymbol{\beta}'\boldsymbol{x}_{i,l} + \boldsymbol{b}_i'\boldsymbol{z}_{i,l}), \tag{3.5}$$

where $\boldsymbol{z}_{i,l} = (z_{i,l,1}, \ldots, z_{i,l,q})'$ is an additional vector of covariates and $\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,q})'$ is a cluster specific random effect which is again usually assumed to follow a parametric distribution, most often multivariate normal. Such model is considered, e.g., by Vaida and Xu (2000). Note that the model (3.4) is a special case of (3.5) with $\boldsymbol{z}_{i,l} \equiv 1$ and $Z_i \equiv \exp(\boldsymbol{b}_i)$.

Besides the fact that the frailty PH model is not, similarly as the basic PH model, robust towards neglected covariates, it has another important drawback. Indeed, for most frailty distributions, the marginal hazard function obtained from (3.4) by integrating out Z_i is no more proportional with respect to the covariates $\boldsymbol{x}_{i,l}$. Moreover, the form in which the covariate vector $\boldsymbol{x}_{i,l}$ modifies the marginal baseline hazard function depends on the assumed frailty distribution. Consequently, the estimates of the regression parameters $\boldsymbol{\beta}$ can be highly sensitive towards the choice of the frailty distribution; see Hougaard (2000, Chapter 7) for more details.

3.4.2 Population averaged accelerated failure time model

A natural extension of the basic AFT model allowing for multivariate data, breaks down the assumption of i.i.d. error terms ε in the model expression (3.2) by assuming

$$\log(T_{i,l}) = \boldsymbol{\beta}'\boldsymbol{x}_{i,l} + \varepsilon_{i,l}, \quad i = 1, \ldots, N,\ l = 1, \ldots, n_i, \tag{3.6}$$

with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i})'$, $i = 1, \ldots, N$ being independent random vectors, each with *a multivariate* density $g_{\varepsilon,i}(\boldsymbol{\varepsilon}_i)$. Such model is often called *population-averaged* (PA) or *marginal*. When all clusters are of the same size, i.e. when $n_i = n$ for all i, it is usually assumed that the random error vectors $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, N$ are i.i.d. with a multivariate density $g_\varepsilon(\boldsymbol{\varepsilon})$. The main disadvantages of the PA model is that the model is designed only to account for within-cluster dependencies and consequently structured modelling of these dependencies is rather unnatural.

Early semi-parametric approaches to the population averaged AFT model (3.6) with right-censored data are given by Lin and Wei (1992); Lee, Wei, and Ying (1993) and are directed mainly towards the estimation of the regression parameter $\boldsymbol{\beta}$. They use the following estimation strategy. In the first step, they ignore the correlation and estimate the regression coefficient

$\boldsymbol{\beta}$ using one of the semi-parametric approaches for uncorrelated censored data outlined in Section 3.2 (the Buckley-James estimator or censored data linear-rank-test-based estimator). In the second step, they correct the standard errors of the estimate using a GEE approach (Liang and Zeger, 1986). However, we can point out that ignoring the dependence in the estimation step generally does not take full advantage of the information in the data and is likely not to be efficient. For that reason, Pan and Kooperberg (1999) suggest, in the case of bivariate survival data, i.e. $n_i = 2$ for all $i = 1, \ldots, N$, methods that account already in the estimation step for the within-cluster correlation. Briefly, their method iterates between (a) estimating the joint bivariate distribution of $(\varepsilon_{i,1}, \varepsilon_{i,2})'$ using the bivariate log-spline density estimate of Kooperberg (1998), (b) multiple imputation (Wei and Tanner, 1991) of censored observations, (c) estimating the regression parameter $\boldsymbol{\beta}$ using either ordinary or generalized least squares. Note that this procedure can be considered as a generalization of the basic Buckley-James estimator, for which in step (a) the Kaplan-Meier estimator of the survival distribution is used while ignoring the correlation and in step (b) a simple imputation using conditional expectations is employed.

Finally, Pan and Connett (2001) present an approach that, to some extent, combines the methods of Lee et al. (1993) and Pan and Kooperberg (1999). It iterates between (a) estimating the marginal distribution of $\varepsilon_{i,l}$ using the Kaplan-Meier estimate while ignoring the dependencies, (b) multiple imputation of censored observations, (c) GEE estimation of the regression parameter $\boldsymbol{\beta}$ using a general working correlation matrix.

3.4.3 Cluster specific accelerated failure time model

Another extension of the AFT model for multivariate data adds, similarly as the frailty PH model and analogously as the classical linear mixed model of Laird and Ware (1982), cluster specific random effect vector $\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,q})'$ combined with a vector of covariates $\boldsymbol{z}_{i,l} = (z_{i,l,1}, \ldots, z_{i,l,q})'$ into the model expression, i.e.

$$\log(T_{i,l}) = \boldsymbol{\beta}'\boldsymbol{x}_{i,l} + \boldsymbol{b}_i'\boldsymbol{z}_{i,l} + \varepsilon_{i,l}, \quad i = 1, \ldots, N,\ l = 1, \ldots, n_i. \tag{3.7}$$

The random effect vectors $\boldsymbol{b}_i$, $i = 1, \ldots, N$ are assumed to be i.i.d. with some (multivariate) density $g_b(\boldsymbol{b})$, the random error terms $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ are assumed to be i.i.d. with some density $g_\varepsilon(\varepsilon)$ and independent on the random effects. Besides the term *cluster-specific* (CS), the model (3.7) is sometimes called *conditional*, since the distribution of the event time $T_{i,l}$ is modelled conditionally, given the cluster specific characteristic $\boldsymbol{b}_i$.

In the literature, Pan and Louis (2000) and Pan and Connett (2001) consider model (3.7) with a univariate random effect b_i and $z_{i,l} \equiv 1$ for all i and l. The estimation procedure iterates between (a) estimating the distribution of independent error terms $\varepsilon_{i,l}$ using the Kaplan-Meier estimator, (b) multiple imputation of censored times, (c) a Monte Carlo EM algorithm of Wei and Tanner (1990) in Pan and Louis (2000) or restricted maximum likelihood in Pan and Connett (2001) to estimate the regression parameter $\boldsymbol{\beta}$.

Observe that in contrast to the frailty PH model, in the cluster-specific AFT model the meaning of the regression parameters $\boldsymbol{\beta}$ is the same conditionally given $\boldsymbol{b}_i$ as well as marginally. Indeed, when the random effects $\boldsymbol{b}_i$, $i = 1, \ldots, N$ are integrated out from model (3.7), we obtain the model (3.6) with the only change in the error distribution which is given as an appropriate convolution.

3.4.4 Population averaged model versus cluster specific model

When compared to the PA model, not only the CS model allows for structured modelling of within-cluster dependencies but is often preferred to it due to clear decomposition of the sources of variability and more natural interpretation of the regression parameters, see Lindsey and Lambert (1998) and Lee and Nelder (2004) for more details.

However, in some sense, the PA model is more general in the following sense. The CS model is specified hierarchically and always implies a particular PA model when the random effects are integrated out. On the other hand, the same PA model can correspond to several, very different CS models. Moreover, with the most common assumptions, i.e. when the error terms $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ in the CS model are assumed to be i.i.d. the random effects b_i, $i = 1, \ldots, N$ in the CS model i.i.d. and independent on the errors and the error term vectors $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, N$ in the PA model i.i.d., the PA model leads to a more general covariance structure than the CS model. To illustrate this, consider the CS model (3.7) with a random intercept only, i.e. $z_{i,l} \equiv 1$. Let $\mathrm{var}(\varepsilon_{i,l}) = \sigma_\varepsilon^2$ and $\mathrm{var}(b_i) = \sigma_b^2$, $i = 1, \ldots, N$. Such model implies a covariance matrix for the log-event times vector $\big(\log(T_{i,1}), \ldots, \log(T_{i,n_i})\big)'$ which is of the compound symmetry type, i.e.

$$\mathrm{var} \begin{pmatrix} \log(T_{i,1}) \\ \vdots \\ \log(T_{i,n_i}) \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_b^2 & \ldots & \sigma_b^2 \\ \vdots & \ddots & \vdots \\ \sigma_b^2 & \ldots & \sigma_\varepsilon^2 + \sigma_b^2 \end{pmatrix}.$$

That is, the variance is necessarily the same for all observations within a cluster and the correlation between the two observations is the same for all pairs within a cluster. On the other hand, with the PA model (3.6) both the variance and the correlation are allowed to vary across the cluster as usually unstructured covariance matrix for the error terms vector $\boldsymbol{\varepsilon}_i$ and subsequently also for the log-event times vector is assumed.

Chapter 4

Frequentist and Bayesian Inference

Both PH and AFT models determine a probabilistic mechanism that leads to survival data. The mechanism depends further on a vector of unknown parameters, denoted by $\boldsymbol{\theta}$, which represents the relevant information we wish to pick up from the observed data. For example, for the AFT model (3.2), the $\boldsymbol{\theta}$ vector is equal to $(\alpha, \boldsymbol{\beta}', \tau)'$ and the probabilistic mechanism is given by equation (3.2) together with the specification of the density of the error term ε. The assumed probabilistic mechanism together with the observed data determines the *likelihood* function, $L(\boldsymbol{\theta})$, which is the corner stone to draw the inference about the unknown parameter vector $\boldsymbol{\theta}$.

Two major paradigms exist in statistics of how to use the likelihood in order to draw the inference about $\boldsymbol{\theta}$, namely the *frequentist* and the *Bayesian* paradigms. In the classical frequentist point of view, the data are assumed to be a random sample generated by the random mechanism controlled by $\boldsymbol{\theta}$, which is unknown but *fixed*. Several methods exist to estimate the true value of the parameter $\boldsymbol{\theta}$, *maximum likelihood* (ML) being one of the most popular ones. The estimator, $\widehat{\boldsymbol{\theta}}$, maximizes the likelihood function over a set Θ of admissible $\boldsymbol{\theta}$ values – the parameter space. Hypotheses about the parameter vector $\boldsymbol{\theta}$ can be tested and accuracy of the estimates can be assessed by calculation of the confidence intervals. See, e.g., Cox and Hinkley (1974, Chapter 9) or Lehmann and Casella (1998, Chapter 6) for more details on ML estimation.

In Bayesian statistics, both the data and the parameter vector $\boldsymbol{\theta}$ are treated as random variables. Besides the probabilistic model to generate the data, *a prior distribution* $p(\boldsymbol{\theta})$ must be specified for the model parameters. Infer-

ence is then based on the *posterior distribution* $p(\boldsymbol{\theta} \,|\, \text{data})$ of the parameters given the data which is calculated using *Bayes*' rule:

$$p(\boldsymbol{\theta} \,|\, \text{data}) = \frac{L(\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_\Theta L(\boldsymbol{\theta}^*)\, p(\boldsymbol{\theta}^*)\, d\boldsymbol{\theta}^*} \propto L(\boldsymbol{\theta})\, p(\boldsymbol{\theta}). \tag{4.1}$$

As point estimate of $\boldsymbol{\theta}$, the posterior expectation, median or mode can be used. The uncertainty about the model parameters can be expressed using credible intervals constructed using the quantiles of the posterior distribution (see Section 4.6 for more details). For an extensive introduction into the area of Bayesian statistics, see, e.g., Carlin and Louis (2000); Gelman et al. (2004).

4.1 Likelihood for interval-censored data

We saw that the likelihood plays a principal role in drawing inference about unknown model parameters. In this section, we discuss the general form of the likelihood, first for univariate interval-censored and doubly-interval-censored data. The multivariate case will be discussed in the following section.

In this section, let T_i, $i = 1, \ldots, N$ be a set of independent event times each with a density $p_i(t;\, \boldsymbol{\theta})$. For instance, for AFT model (3.2) density $p_i(t;\, \boldsymbol{\theta})$ is given by

$$p_i(t;\, \boldsymbol{\theta}) = (\tau\, t)^{-1}\, g_\varepsilon^*\big\{\tau^{-1}(\log t - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i)\big\},$$

where $\boldsymbol{x}_i$ is a covariate vector for the ith observation.

4.1.1 Interval-censored data

Let $\lfloor t_i^L,\, t_i^U \rfloor$ be observed intervals and δ_i corresponding censoring indicators with the same convention as in Section 2.1. Let the corresponding survival functions be denoted by $S_i(t;\, \boldsymbol{\theta})$. The likelihood $L(\boldsymbol{\theta})$ is then the product of individual likelihood contributions $L_i(\boldsymbol{\theta})$, i.e. $L(\boldsymbol{\theta}) = \prod_{i=1}^N L_i(\boldsymbol{\theta})$, where

$$L_i(\boldsymbol{\theta}) = \begin{cases} \int_{t_i^L}^{\infty} p_i(s;\, \boldsymbol{\theta})\, ds = S_i(t_i^L;\, \boldsymbol{\theta}), & \delta_i = 0, \\ p_i(t_i;\, \boldsymbol{\theta}), & \delta_i = 1, \\ \int_0^{t_i^U} p_i(s;\, \boldsymbol{\theta})\, ds = \big\{1 - S_i(t^U;\, \boldsymbol{\theta})\big\}, & \delta_i = 2, \\ \int_{t_i^L}^{t_i^U} p_i(s;\, \boldsymbol{\theta})\, ds = \big\{S_i(t_i^L;\, \boldsymbol{\theta}) - S_i(t_i^U;\, \boldsymbol{\theta})\big\}, & \delta_i = 3. \end{cases}$$

This can be briefly written as

$$L_i(\boldsymbol{\theta}) = \oint_{t_i^L}^{t_i^U} p_i(s;\, \boldsymbol{\theta})\, ds \tag{4.2}$$

if we make use of the notation

$$\oint_{\tau^L}^{\tau^U} p(s)\,ds = \begin{cases} \displaystyle\int_{\tau^L}^{\tau^U} p(s)\,ds, & \text{if } \tau^L < \tau^U \\ p(\tau^L) = p(\tau^U), & \text{if } \tau^L = \tau^U, \end{cases} \tag{4.3}$$

i.e. the integral disappears whenever the event time is exactly observed. Note that already for simple interval-censored data, the likelihood involves integration of the density.

4.1.2 Doubly-interval-censored data

Let $\lfloor u_i^L, u_i^U \rfloor$, $i = 1, \ldots, N$ be observed intervals for the onset time U_i and $\lfloor v_i^L, v_i^U \rfloor$ observed intervals for the failure time V_i in the sense of Section 2.2. It is tempting to transform observations into single intervals of the form $\lfloor t_i^L, t_i^U \rfloor = \lfloor v_i^L - u_i^U, v_i^U - u_i^L \rfloor$ and then to use methods for simple interval-censored data with the likelihood (4.2). However, as pointed out by De Gruttola and Lagakos (1989), this approach would be only valid if the onset time U_i is uniformly distributed and independent of the event time T_i.

To write a likelihood contribution of each observation in the general case a bivariate density of an event and onset times must be considered. Let $q_i(t, u; \boldsymbol{\theta})$ be a density of the random vector $(T_i, U_i)'$, $i = 1, \ldots, N$. The likelihood contribution of the ith observation is then given by a double integral of the form

$$L_i(\boldsymbol{\theta}) = \oint_{u_i^L}^{u_i^U} \Big\{ \oint_{v_i^L - u}^{v_i^U - u} q_i(t, u; \boldsymbol{\theta})\,dt \Big\}\,du. \tag{4.4}$$

Note that whenever either the onset time U_i and/or the failure time V_i are exactly observed either both or one integrals disappear in the formula (4.4).

In most practical situations it can be assumed that, given the parameter vector $\boldsymbol{\theta}$, the onset and the event time are independent, i.e.

$$q_i(t, u; \boldsymbol{\theta}) = p_i(t; \boldsymbol{\theta})\, p_i^U(u; \boldsymbol{\theta}). \tag{4.5}$$

In the rest of this thesis we shall make use of assumption (4.5). The likelihood contribution of the ith subject can then be rewritten as

$$L_i(\boldsymbol{\theta}) = \oint_{u_i^L}^{u_i^U} \Big\{ \oint_{v_i^L - u}^{v_i^U - u} p_i(t; \boldsymbol{\theta})\,dt \Big\}\, p_i^U(u; \boldsymbol{\theta})\,du. \tag{4.6}$$

4.2 Likelihood for multivariate (doubly) interval-censored data

In the case of multivariate event times $T_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$, observed as intervals $\lfloor t_{i,l}^L, t_{i,l}^U \rfloor$, the likelihood contribution of the ith cluster equals

$$L_i(\boldsymbol{\theta}) = \oint_{t_{i,1}^L}^{t_{i,1}^U} \cdots \oint_{t_{i,n_i}^L}^{t_{i,n_i}^U} p_i(t_1, \ldots, t_{n_i}; \boldsymbol{\theta})\, dt_{n_i} \cdots dt_1, \tag{4.7}$$

where $p_i(t_1, \ldots, t_{n_i}; \boldsymbol{\theta})$ is the density of $(T_{i,1}, \ldots, T_{i,n_i})'$ implied by the assumed model.

When population averaged AFT model introduced in Section 3.4.2 is assumed, $p_i(t_1, \ldots, t_{n_i}; \boldsymbol{\theta})$ equals

$$p_i(t_1, \ldots, t_{n_i}; \boldsymbol{\theta}) = \frac{g_{\varepsilon,i}\big\{\log(t_1) - \boldsymbol{\beta}'\boldsymbol{x}_{i,1}, \ldots, \log(t_{n_i}) - \boldsymbol{\beta}'\boldsymbol{x}_{i,n_i}\big\}}{t_1 \cdots t_{n_i}}. \tag{4.8}$$

In the case of the cluster-specific AFT model described in Section 3.4.3, the density $p_i(t_1, \ldots, t_{n_i}; \boldsymbol{\theta})$ becomes

$$p_i(t_1, \ldots, t_{n_i}; \boldsymbol{\theta}) = \int_{\mathbb{R}^q} \left[\prod_{l=1}^{n_i} \frac{g_\varepsilon\big\{\log(t_l) - \boldsymbol{\beta}'\boldsymbol{x}_{i,l} - \boldsymbol{b}_i'\boldsymbol{z}_{i,l}\big\}}{t_l} \right] g_b(\boldsymbol{b}_i)\, d\boldsymbol{b}_i. \tag{4.9}$$

For doubly-interval-censored data, under assumption (4.5), the likelihood contribution of the ith cluster is obtained by an appropriate multivariate modification of the expression (4.6).

4.3 Bayesian data augmentation

The computation of the likelihood for interval- and doubly-interval-censored data is rather involved. The complexity even increases when multivariate survival data are introduced. Indeed, the maximum likelihood method involves multivariate integration combined with the optimization of the likelihood which becomes quickly intractable even for simple models.

On the other hand, in Bayesian statistics, where the unknown parameter vector $\boldsymbol{\theta}$ is assumed to be random and its posterior distribution $p(\boldsymbol{\theta} \,|\, \text{data})$ is used for inference, we are completely free to augment the vector of unknowns by arbitrary *auxiliary variables*, let say $\boldsymbol{\psi}$. Inference can then equally be based on the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\psi} \,|\, \text{data})$. Indeed, all (marginal)

posterior characteristics of $\boldsymbol{\theta}$ (mean, median, credible intervals) are the same regardless whether they are computed from $p(\boldsymbol{\theta} \,|\, \text{data})$ or $p(\boldsymbol{\theta},\, \boldsymbol{\psi} \,|\, \text{data})$ since

$$p(\boldsymbol{\theta} \,|\, \text{data}) = \int p(\boldsymbol{\theta},\, \boldsymbol{\psi} \,|\, \text{data})\, d\boldsymbol{\psi}.$$

In the case of censored data, matters simplify considerably if the unknown true event times t_i are explicitely considered to make a part of the vector of unknowns, i.e. $\boldsymbol{\psi} = (t_i \,:\, i = 1, \ldots, N,\, t_i \text{ is censored})'$. Assume now that all observations are censored. In this situation, it is obvious that $\boldsymbol{\psi}$ (uncensored *augmented data*) conveys more precise information about the model parameter $\boldsymbol{\theta}$ than the censored data which implies

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{\psi},\, \text{data}) = p(\boldsymbol{\theta} \,|\, \boldsymbol{\psi}).$$

The joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ then equals

$$p(\boldsymbol{\theta},\, \boldsymbol{\psi} \,|\, \text{data}) = p(\boldsymbol{\theta} \,|\, \boldsymbol{\psi},\, \text{data})\, p(\boldsymbol{\psi} \,|\, \text{data}) = p(\boldsymbol{\theta} \,|\, \boldsymbol{\psi})\, p(\boldsymbol{\psi} \,|\, \text{data}). \qquad (4.10)$$

The two terms on the right hand side of formula (4.10) are now easily computed. Indeed, $p(\boldsymbol{\theta} \,|\, \boldsymbol{\psi})$ is the posterior distribution of $\boldsymbol{\theta}$ if the uncensored data were available, i.e.

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{\psi}) \propto L^{augm}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

where the likelihood L^{augm} of the uncensored augmented data is simply

$$L^{augm}(\boldsymbol{\theta}) = \prod_{i=1}^{N} p_i(t_i;\, \boldsymbol{\theta}).$$

The second term of the right hand side of formula (4.10), $p(\boldsymbol{\psi} \,|\, \text{data})$, is under the assumption of independent noninformative censoring proportional to the product of indicator functions:

$$p(\boldsymbol{\psi} \,|\, \text{data}) \propto \prod_{i=1}^{N} I\big\{t_i \in \lfloor t_i^L,\, t_i^U \rfloor\big\}.$$

A similar procedure can be applied for doubly-censored data. In that case, both true onset times u_i and true event times t_i $i = 1, \ldots, N$ are augmented into the vector of unknowns. The situation where only the part of the data is censored is analogous, only with some change in notation. Finally, in the case of multivariate survival data and cluster specific models, the integrals of

the form (4.9) can easily be avoided by augmenting the vector of unknowns by the values of the random effects $\boldsymbol{b}_i$, $i = 1, \ldots, N$.

The idea of *data augmentation* was first introduced in the context of the EM algorithm (Dempster, Laird, and Rubin, 1977) and formalized in the context of Bayesian computation by Tanner and Wong (1987). For more complex models with censored data, this technique constitutes a highly appealing alternative to difficult maximum likelihood estimation. Moreover, it is quite natural to include the true event times or the values of latent random effects in the set of unknowns. For these reasons, most of the models developed in this thesis make use of the Bayesian estimation with augmented true event times.

4.4 Hierarchical specification of the model

In Bayesian statistics, the prior distribution $p(\boldsymbol{\theta})$ and the model assumed to generate the data, represented by the likelihood $L(\boldsymbol{\theta}) = p(\text{data} \,|\, \boldsymbol{\theta})$, are usually specified in a hierarchical manner. Firstly, remember that the parameter vector $\boldsymbol{\theta}$ contains not only the parameters in a classical sense but also all remaining latent factors like random effects or augmented times. Crudely, the vector $\boldsymbol{\theta}$ can usually be splitted into two parts $\boldsymbol{\theta} = (\boldsymbol{\psi}', \boldsymbol{\phi}')'$ where $\boldsymbol{\psi}$ refers to the latent factors and $\boldsymbol{\phi}$ to the parameters in a classical sense. The specification of the Bayesian model then proceeds in the following steps:

1. **Data Model** step specifies the likelihood function

   $$L(\boldsymbol{\theta}) = p(\text{data} \,|\, \boldsymbol{\theta}) = p(\text{data} \,|\, \boldsymbol{\psi}, \boldsymbol{\phi})$$

   and is actually equivalent to the frequentist specification of the model.

2. **Latent Process Model** step specifies

   $$p(\boldsymbol{\psi} \,|\, \boldsymbol{\phi}),$$

   i.e. the distribution of the latent factors, possibly given the classical parameters $\boldsymbol{\phi}$.

3. **Parameter Model (Prior)** specifies the prior distribution for the classical parameters $\boldsymbol{\phi}$, i.e. it specifies

   $$p(\boldsymbol{\phi}).$$

   Often, the components of $\boldsymbol{\phi}$ are assumed to be a priori independent and if no external information is available are assigned vague but proper prior distributions.

The overall prior distribution is then given by

$$p(\boldsymbol{\theta}) \propto p(\boldsymbol{\psi} \,|\, \boldsymbol{\phi}) \times p(\boldsymbol{\phi}),$$

and the posterior distribution is obtained using the relationship (4.1) as

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, \text{data}) &\propto L(\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \\ &\propto p(\text{data} \,|\, \boldsymbol{\psi}, \boldsymbol{\phi}) \times p(\boldsymbol{\psi} \,|\, \boldsymbol{\phi}) \times p(\boldsymbol{\phi}), \end{aligned} \tag{4.11}$$

i.e. it is proportional to the product of the distributions specified in the above three steps.

The hierarchical structure of more complex hierarchical models is usually best expressed using so called directed acyclic graphs (DAG) where each model quantity is represented by the node drawn as a circle for unknowns and drawn as a squared box for observed or fixed quantities (data, covariates). Solid arrows are used to represent stochastic dependencies and dashed arrows deterministic dependencies between the nodes. A simple DAG which only distinguishes among the data, latent quantities $\boldsymbol{\psi}$ and classical parameters $\boldsymbol{\phi}$ and which corresponds to the expression (4.11) is shown in Figure 4.1.

Further, it is assumed that given its parents, each node is conditionally independent on all its grandparents, i.e. schematically

$$p(\text{child} \,|\, \text{parents, grandparents}) = p(\text{child} \,|\, \text{parents}).$$

The posterior distribution of the hierarchical model is then proportional, analogously to the relationship (4.11), to the product of all conditional distributions of the type $p(\text{child} \,|\, \text{parents})$ times the product of the prior distributions for the nodes of the first generation (i.e. having no parents).

Illustration 4.1. *Linear mixed model.* As an illustration, consider a classical normal linear mixed model with data $= \{\boldsymbol{y}_i, \ldots, \boldsymbol{y}_N\}$ being a realization

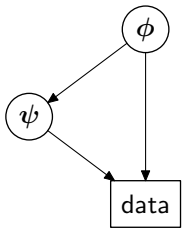


Figure 4.1: Directed acyclic graph – general scheme.

of independent random vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, N$, each of length n which in a frequentist sense can be specified as

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

$$\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\mathbf{0},\, \mathbb{D}), \qquad \boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_n(\mathbf{0},\, \Sigma),$$

where $\mathbb{X}_i$, $\mathbb{Z}_i$, $i = 1, \ldots, N$ are fixed covariate matrices. For the sake of the Bayesian modelling, the vector $\boldsymbol{\theta} = (\boldsymbol{\psi}', \boldsymbol{\phi}')'$ is given by

$$\boldsymbol{\psi} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_N')', \qquad \boldsymbol{\phi} = \Big(\boldsymbol{\beta}',\, \mathrm{vec}(\mathbb{D}),\, \mathrm{vec}(\Sigma)\Big)'.$$

The whole model can be represented by the DAG shown in Figure 4.2. The above mentioned three steps in the model building proceeds as follows. The Data Model is given by a normal likelihood

$$L(\boldsymbol{\theta}) = p(\text{data} \,|\, \boldsymbol{\theta}) = p(\text{data} \,|\, \boldsymbol{\psi},\, \boldsymbol{\phi}) = \prod_{i=1}^{N} \varphi_n(\boldsymbol{y}_i \,|\, \boldsymbol{\beta}'\boldsymbol{x}_i + \boldsymbol{b}_i'\boldsymbol{z}_i,\, \Sigma).$$

The Latent Process Model is determined by the normal distribution of the random effects, i.e.

$$p(\boldsymbol{\psi} \,|\, \boldsymbol{\phi}) = \prod_{i=1}^{N} \varphi_q(\boldsymbol{b}_i \,|\, \mathbf{0},\, \mathbb{D}).$$

Finally, some prior distributions $p(\boldsymbol{\beta})$, $p(\mathbb{D})$, $p(\Sigma)$ are assigned to the parameters of the main interest, i.e. to $\boldsymbol{\beta}$, $\mathbb{D}$, Σ and

$$p(\boldsymbol{\phi}) = p(\boldsymbol{\beta}) \,\times\, p(\mathbb{D}) \,\times\, p(\Sigma).$$

□

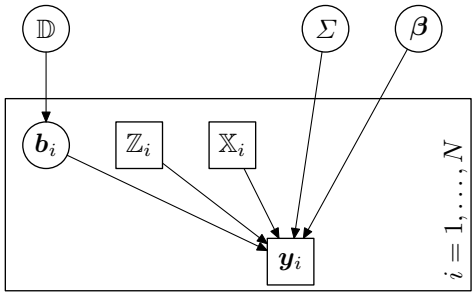


Figure 4.2: Directed acyclic graph for the linear mixed model.

4.5 Markov chain Monte Carlo

In previous sections, we stated that the inference in the Bayesian approach is based on the posterior distribution $p(\boldsymbol{\theta} \,|\, \text{data})$ which is obtained using the Bayes' formula (4.1) and is proportional to the product of the likelihood and the prior distribution. We also saw that difficult likelihood evaluations can be avoided by the introduction of a set of suitable auxiliary variables (augmented data). What needs to be discussed is how the posterior distribution can be computed and how to determine posterior summaries about $\boldsymbol{\theta}$. Most quantities related to the posterior summarization (posterior moments, quantiles, highest posterior density regions etc.) involve computation of the posterior expectation of some function $G(\boldsymbol{\theta})$, i.e. computation of

$$\mathrm{E}\big\{G(\boldsymbol{\theta}) \,\big|\, \text{data}\big\} = \int_{\Theta} G(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \text{data})\, d\boldsymbol{\theta} = \frac{\int_{\Theta} G(\boldsymbol{\theta})\, L(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{\Theta} L(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \quad (4.12)$$

The integration in the expression (4.12) is usually high-dimensional and only rarely analytically tractable in realistic practical situations.

Markov chain Monte Carlo (MCMC) methods avoid the explicit evaluations of integrals. Instead, we construct a Markov chain with state space Θ whose stationary distribution is equal to $p(\boldsymbol{\theta} \,|\, \text{data})$. After a sufficient number of burn-in iterations the current draws follow the stationary distribution, i.e. the posterior distribution of interest. We keep a sample of $\boldsymbol{\theta}$ values, let say $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ and approximate the posterior expectation (4.12) by

$$\overline{G}_M = \frac{1}{M} \sum_{m=1}^{M} G(\boldsymbol{\theta}^{(m)}). \quad (4.13)$$

The ergodic theorem implies that, under mild conditions, $\overline{G}_M$ converges almost surely to $\mathrm{E}\big\{G(\boldsymbol{\theta}) \,\big|\, \text{data}\big\}$ as $M \to \infty$ (see, e.g., Billingsley, 1995, Section 24).

Many methods are available to construct the Markov chains with desired properties. The most often used are the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs algorithm (Geman and Geman, 1984; Gelfand and Smith, 1990). Both of them, often properly dedicated will be used extensively throughout this thesis. A comprehensive introduction into the area of the MCMC can be found, e.g., in Geyer (1992); Tierney (1994); Besag et al. (1995). More details can be obtained from several books, e.g., Gilks, Richardson, and Spiegelhalter (1996); Gamerman (1997); Chen, Shao, and Ibrahim (2000); Robert and Casella (2004).

4.6 Credible regions and Bayesian p-values

With a frequentist approach, confidence intervals or regions and p-values are used to summarize the estimates and the inference for $\boldsymbol{\theta}$ – parameter of interest. In Bayesian statistics, the role of the confidence regions is played by the *credible regions* and p-values are replaced by the *Bayesian p-values*. In this section, we briefly discuss their construction.

4.6.1 Credible regions

For a given $\alpha \in (0, 1)$, the $100(1-\alpha)\%$ credible region Θ_α for a parameter of interest $\boldsymbol{\theta}$ is defined using the conditional distribution $\boldsymbol{\theta} \,|\, \text{data}$ (posterior distribution of $\boldsymbol{\theta}$) as

$$\Pr\big(\boldsymbol{\theta} \in \Theta_\alpha \,\big|\, \text{data}\big) = 1 - \alpha. \tag{4.14}$$

Equal-tail credible interval

Suppose first, the parameter of interest θ is univariate. The credible region Θ_α can then be obtained by setting $\Theta_\alpha = (\theta^L_\alpha, \theta^U_\alpha)$, such that

$$\Pr\big(\theta \le \theta^L_\alpha \,\big|\, \text{data}\big) = \Pr\big(\theta \ge \theta^U_\alpha \,\big|\, \text{data}\big) = \alpha/2.$$

Such an interval is easily constructed when a sample from the posterior distribution of θ (obtained, e.g., using the MCMC technique) is available. Indeed, θ^L_α and θ^U_α are $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$, respectively, quantiles of the posterior distribution $\theta \,|\, \text{data}$ and from the MCMC output they can be estimated using the sample quantiles.

Simultaneous credible bands

For the case the parameter of interest, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, is multivariate and we wish to calculate *simultaneous* probability statements, Besag et al. (1995, p. 30) suggest to compute *simultaneous credible bands*. In that case, Θ_α equals

$$\Theta_\alpha = (\theta^L_{1,\alpha_{uni}}, \theta^U_{1,\alpha_{uni}}) \times \cdots \times (\theta^L_{q,\alpha_{uni}}, \theta^U_{q,\alpha_{uni}}). \tag{4.15}$$

That is, Θ_α is given as a product of univariate equal-tail credible intervals of the same univariate level α_{uni} (typically $\alpha_{uni} \ge \alpha$). As shown by Besag et al. (1995), the simultaneous credible bands can easily be computed when the sample from the posterior distribution is available as only order statistics for each univariate sample are needed. From the computational point of view,

the most intensive part in computation of the simultaneous credible band is to sort the univariate samples. However, when simultaneous credible bands for different values of α are required this must be done only once. This property is advantageously used when computing the simultaneous Bayesian p-values (see Section 4.6.2).

As pointed by Held (2004), due to the fact the simultaneous credible band is by construction restricted to be hyperrectangular, it can cover a huge area actually not supported by the posterior distribution. Obviously, this problem becomes more severe when a high posterior correlation exists between the components of the vector $\boldsymbol{\theta}$.

Highest posterior density region

An alternative to the credible intervals and simultaneous credible bands is given by *the highest posterior density (HPD) region*. In that case, Θ_α is obtained by requiring (4.14) and additionally

$$p(\boldsymbol{\theta}_1 \,|\, \text{data}) > p(\boldsymbol{\theta}_2 \,|\, \text{data}) \quad \text{for all } \boldsymbol{\theta}_1 \in \Theta_\alpha,\ \boldsymbol{\theta}_2 \notin \Theta_\alpha.$$

Note that in the univariate case and for unimodal posterior densities $p(\theta \,|\, \text{data})$, the HPD region becomes an interval. However, it is clear that in contrast to the equal-tail credible interval or the simultaneous credible band the computation of the HPD region is much more complicated even when the sample from the posterior distribution is already available.

4.6.2 Bayesian p-values

The Bayesian counterpart of the p-value for the hypothesis $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ (typically $\boldsymbol{\theta}_0$ is a vector of zeros) – *the Bayesian p-value* – can be defined as 1 minus the content of the credible region which just covers $\boldsymbol{\theta}_0$, i.e.

$$p = 1 - \min\big\{\alpha :\ \boldsymbol{\theta}_0 \in \Theta_\alpha\big\}. \tag{4.16}$$

In the univariate case, a two-sided Bayesian p-value based on the *equal-tail credible interval* is computed quite easily once the sample from the posterior distribution is available since (4.16) can be expressed as

$$p = 2\min\Big\{\Pr(\theta \le \theta_0 \,|\, \text{data}),\ \Pr(\theta \ge \theta_0 \,|\, \text{data})\Big\}, \tag{4.17}$$

and $\Pr(\theta \le \theta_0 \,|\, \text{data})$, $\Pr(\theta \ge \theta_0 \,|\, \text{data})$ can be estimated as a proportion of the sample being higher or lower, respectively, than the point of interest θ_0.

In the multivariate case, a two-sided simultaneous Bayesian p-value based on the *simultaneous credible band* can be obtained by calculating the simultaneous credible bands Θ_α on various levels α and determining the smallest level, such that $\boldsymbol{\theta}_0 \in \Theta_\alpha$, i.e. by direct usage of the expression (4.16).

To compute the Bayesian p-value based on the HPD region, the expression (4.16) takes the form

$$p = \Pr\big\{\boldsymbol{\theta} : \, p(\boldsymbol{\theta} \,|\, \text{data}) \leq p(\boldsymbol{\theta}_0 \,|\, \text{data}) \,\big|\, \text{data}\big\}. \tag{4.18}$$

An MCMC estimate of (4.18) can easily be obtained when $p(\boldsymbol{\theta} \,|\, \text{data})$ (any proportionality constants may be ignored) can efficiently be evaluated. Often, this is not the case however. Nevertheless, a technique how to overcome the problem of unknown or difficult to evaluate $p(\boldsymbol{\theta} \,|\, \text{data})$ using its estimate based on Rao-Blackwellization is given by Held (2004).

Mainly for computational reasons, we report in this thesis, if not stated otherwise, univariately equal-tail credible intervals and corresponding Bayesian p-values of the type (4.17) and multivariately simultaneous credible regions (4.15) and corresponding simultaneous Bayesian p-values computed using an iterative procedure to evaluate (4.16).

Chapter 5

An Overview of Methods for Interval-Censored Data

For right-censored data, a variety of methods (non-, semi- and fully parametric) have been developed. Further, commercial software is available to support these techniques. In contrast, for interval-censored data and multivariate (doubly-)interval-censored data commercial software is much more limited and only parametric approaches seem to be available for regression models besides of course the user-written programs. Further, until recently only few methods were available. That is why, in practice, modelling with interval-censored data is often mimicked by methods developed for right-censored data. For this, the interval needs to be replaced by an exact time or right-censored time. The most common assumption is that the event occurred at the midpoint of the interval. However, applying methods for right-censored data on these artificial fixed points can lead to biased and misleading results and the correctness of such approach depends strongly on the underlying distribution of the event times, see e.g., Rücker and Messerer (1988); Law and Brookmeyer (1992); Odell, Anderson, and D'Agostino (1992); Dorey, Little, and Schenker (1993).

In Section 5.1, we first review appropriate *frequentist* methods to deal with (doubly-)interval-censored data and link them to the corresponding (classical) method for right-censored data. We start with the estimation of the survival distribution, proceed to the two-sample tests for the survival distributions, continue with the proportional hazards and accelerated failure time models and end up with the remark on the problem of interval-censored covariates. Whenever feasible, we mention computational aspects of described methods applicable for R, SPLUS and SAS.

With suitable semi-parametric approaches, both PH and AFT models can be used not only for the estimation of the effect of covariates but also for both estimation of the baseline survival distribution or comparison of two or more samples. With the *Bayesian* approach, it is moreover relatively easy to set up and estimate the models for multivariate (doubly-)interval-censored data. We will illustrate this on the analysis of the Signal Tandmobiel® data using a semi-parametric Bayesian PH model in Section 5.2. As we are interested mainly in the AFT model, we give also an overview of available Bayesian developments for this model in Section 5.3. We end this chapter by highlighting our motivations for the further developments presented in this thesis.

5.1 Frequentist methods

5.1.1 Estimation of the survival function

In the case of simple i.i.d. survival data, often the aim is to estimate the survival function. When only categorical covariates are involved, the survival function can be estimated for each unique combination of covariate values and could be used to check the fitted regression model.

For right-censored data, the classical non-parametric maximum-likelihood estimate (NPMLE) of the survival function is given by Kaplan and Meier (1958). For interval-censored data Peto (1973) first proposed the NPMLE and used the constrained Newton-Raphson method to compute it. Nowadays, the NPMLE of the survival function based on the interval-censored data is known as the Turnbull's estimate (see Turnbull, 1976) who suggested a so called *iterative self-consistency* algorithm, which is, in fact, an EM-like (Dempster et al., 1977) algorithm. An improved version of the maximization algorithm which utilizes standard convex optimization technique was given by Gentleman and Geyer (1994) who also discussed the unicity of the estimate. For computation, a valuable alternative, the iterative convex minorant algorithm, was suggested by Groeneboom and Wellner (1992). Finally, strong consistency of the Turnbull's estimate has been proved under rather general assumptions by Yu, Li, and Wong (2000). The asymptotic distributional behaviour of the Turnbull's estimator for some special cases has been established by Yu et al. (1998) and Huang (1999). An extension of the NPMLE of the survival function for *bivariate* interval-censored data is discussed, e.g., by Bogaerts and Lesaffre (2004).

Several numerical algorithms to compute the non-parametric estimate of the survival function of the interval-censored data are implemented in Vandal's

and Gentleman's R package `Icens` downloadable from *the Comprehensive R Archive Network (CRAN)* or in the S-PLUS function `kaplanMeier`.

A valuable alternative to non-parametric procedures is obtained by smoothing the survival or equivalently the density function or the hazard function. In most practical situations, it can be assumed that the event-times are continuously distributed, and we even get more realistic, not step-wise, estimates. One such method, applicable directly also to interval-censored data is given by Kooperberg and Stone (1992) who smooth the density using splines. They also provide software in the form of the R package `logspline` downloadable from *CRAN* or the S-PLUS library `splinelib` downloadable from *StatLib*. Splines for the smoothing the hazard function are exploited by the approach of Rosenberg (1995).

Illustration 5.1. *Signal Tandmobiel® study.* As an illustration, we computed both the non-parametric estimate of Turnbull (1976) and the smooth estimate of Kooperberg and Stone (1992) of the cumulative distribution functions (cdf) for the emergence of the right mandibular permanent first premolar, separately for boys and girls based on the Signal Tandmobiel® data introduced in Section 1.1. The cdf function giving the proportion of children with the emerged tooth is called in this context the emergence curve and is preferred in this situation to the survival curve. The estimates are plotted in

Figure 5.1: Signal Tandmobiel® study: Cumulative distribution functions of emergence for right mandibular permanent first premolar, separately for girls and boys. Non-parametric estimate of Turnbull (solid line), smooth estimate of Kooperberg and Stone (dashed line).

Figure 5.1. Due to rather high sample size in each group (more than 2 000), the non-parametric estimate is almost the same as the smooth estimate, especially for boys. From the plots it is seen that the emergence for girls is somewhat fastened when compared to boys. □

Doubly-interval-censored data

Non-parametric estimation of the survival curve based on doubly-interval-censored data was first considered by De Gruttola and Lagakos (1989) who make use of discretization of data and generalization of the self-consistency algorithm of Turnbull (1976). The authors estimate simultaneously the onset and the event distributions by treating them as bivariate data. However, they point out that the large number of parameters resulting from discretization, especially if time is grouped too coarsely may cause identifiability problems. This gave rise to several two-step approaches. First, the distribution of the onset time is separately estimated and second, the estimated onset distribution is used as an input for estimation of the distribution of the event time. Bacchetti (1990); Bacchetti and Jewell (1991) assume piece-wise constant hazard and use penalized maximum-likelihood method to estimate the levels of the hazard on each interval. The roughness penalty in the likelihood prevents the method from identifiability problems reported by De Gruttola and Lagakos (1989). The original proposal of De Gruttola and Lagakos (1989) motivates the two-step approaches of Gómez and Lagakos (1994); Sun (1995). Finally, Gómez and Calle (1999) present an extension of the technique of Gómez and Lagakos (1994) which does not require discretization of the data.

5.1.2 Comparison of two survival distributions

If the data can be divided in two (or more) groups, e.g. boys and girls, one could compare the distributions of the event times in these two groups. For right-censored data, many non-parametric tests for comparing two survival curves are available, e.g. the log-rank test (Mantel, 1966), the Gehan generalization of the Wilcoxon test (Gehan, 1965), the Peto-Prentice generalization of the Wilcoxon test (Peto and Peto, 1972; Prentice, 1978) and the weighted Kaplan-Meier statistic of Pepe and Fleming (1989, 1991) which with unit weights is equal to the difference of means of the two survival distributions.

The Gehan-Wilcoxon test has been adopted to interval-censored data by Mantel (1967), while the interval-censored version of the Peto-Prentice-Wilcoxon test is presented by Self and Grossman (1986). The log-rank test for interval-censored data is given by Finkelstein (1986). Further, Petroni and

Wolfe (1994) discuss the weighted Kaplan-Meier statistic in the context of interval-censoring. The performance of above mentioned two-sample tests for interval-censored data is in detail studied and compared by Pan (1999a). Furthermore, Fay (1996, 1999) derived a general class of linear-rank tests for interval-censored data which covers, as special cases, the Wilcoxon-based tests. Finally, Fay and Shih (1998) present a class of tests called *distribution permutation tests* which besides the Wilcoxon-based tests covers also an improved version of the weighted Kaplan-Meier test. SPLUS programs to perform some distribution permutation tests are given by Gómez, Calle, and Oller (2004, Section 4.4) and can be downloaded from

`http://www-eio.upc.es/grass`.

Regrettably, the asymptotic properties of the above methods assume *the grouped continuous model*, which implies that the status of each subject is checked at the same timepoints (in the study time scale) whose number is fixed or that observed intervals are grouped in such a way. For example, for the Signal Tandmobiel® study this would mean that the emergence status of the teeth was checked at prespecified ages, the same for all children. Obviously, such setting is too restrictive in many practical situations. For instance, in the above example, each child was checked by a dentist-examiner on a prespecified day of the year, irrespective of his or her age.

The grouped continuous model assumption is necessary to be able to apply the standard maximum likelihood theory to interval-censored data measured on a continuous scale without making any parametric assumptions. Only recently, Fang, Sun, and Lee (2002) developed a test statistic, based on the weighted Kaplan-Meier statistic of Pepe and Fleming (1989) that does not require the grouped continuous model assumption. Finally, Pan (2000b) offers two-sample test procedures obtained by combining standard right-censored tests and multiple imputation that allows, in contrast to single (e.g. midpoint) imputation mentioned at the beginning of this chapter, to draw appropriately the statistical inference.

Illustration 5.2. *Signal Tandmobiel® study.* The emergence curves of the right mandibular permanent first premolar for boys and girls shown in Figure 5.1 were compared using the Wilcoxon-based, log-rank and Fay's and Shih's version of the difference in means tests. Not surprisingly, for all these tests, the p-value is practically equal to zero. The values of the test statistics, their mean and variance under the null hypothesis and the standardized value, which can asymptotically be compared to the quantile of the standard Gaussian distribution, are shown in Table 5.1. □

Table 5.1: Signal Tandmobiel® study: Two-sample tests comparing the emergence of the permanent right mandibular first premolar (tooth 44) for boys and girls.

Test	Test Statistic	Mean	Variance	Standardized Test Statistic
Gehan-Wilcoxon	554 812	0	2 865 333 000	10.365
Peto-Prentice-Wilcoxon	140.607	−37.634	284.255	10.572
Log-rank	212.316	−53.663	675.251	10.236
Difference in means	264.095	−76.486	1 102.340	10.258

5.1.3 Proportional hazards model

To extend the PH model to interval-censored data, basically four types of approaches can be found in the literature. Firstly, the baseline hazard $\hbar_0$ can be parametrically specified and standard maximum likelihood theory applied to estimate all the parameters. However, the parametric assumptions can cause bias in inference if incorrectly specified and especially with heavily censored data it is general difficult to assess them.

The second class of methods makes use of a combination of multiple imputation (see Rubin, 1987; Wei and Tanner, 1991) and methods for right-censored data represented by works of Satten (1996); Satten, Datta, and Williamson (1998); Goggins et al. (1998); Pan (2000a). A disadvantage of these methods is, however, that they are highly computationally demanding and the fact that the procedures they use to impute missing data have a relatively *ad hoc* nature.

The third approach, suggested by Finkelstein (1986), Pan (1999b), and Goetghebeur and Ryan (2000) resembles most the original method of Cox (1972) combined with that of Breslow (1974). Indeed, in all three papers the baseline hazard $\hbar_0$ is estimated non-parametrically on top of estimating the regression coefficients. Whereas the method of Finkelstein relies on the grouped data assumption, Goetghebeur and Ryan developed an EM-type procedure that relaxes that assumption. Moreover, the approach of Goetghebeur and Ryan seems to be the only one that reduces to a standard Cox model when interval-censoring reduces to right-censoring. Finally, the approach of Pan extends the iterative convex minorant method mentioned in Section 5.1.1 into the context of the PH model. His approach is also implemented as R package `intcox`.

Finally, methods that smoothly estimate $\hbar_0$ are a trade-off between para-

metric modelling that allows for a straightforward maximum likelihood estimation of the parameters and semi-parametric models with a completely unspecified baseline hazard $\hbar_0$. Kooperberg and Clarkson (1997) suggest to use regression splines to express the logarithm of $\hbar_0$, while Joly et al. (1998) employ monotone splines (Ramsay, 1988) directly for the baseline hazard $\hbar_0$. Betensky et al. (1999) use local likelihood smoothing to model the baseline hazard, firstly without covariates. Extension of their method into the regression setting is given by Betensky et al. (2002). Recently, Cai and Betensky (2003) propose to use penalized linear spline for the baseline hazard function.

A nice feature of these methods is that predictive survival and hazard curves are directly available and moreover, they are smooth rather than step-wise as in the case of semi-parametric estimation. The software for the approach of Kooperberg and Clarkson (1997) is included in the previously mentioned R package `logspline` or S-PLUS library `splinelib`.

Doubly-interval-censored data

One of the first approaches to the PH model with doubly-interval-censored data is given by Kim, De Gruttola, and Lagakos (1993) who, under the assumption of the grouped data, directly generalize the one-sample results of De Gruttola and Lagakos (1989). However, their method is highly computationally intensive. For the situation when only the onset time is interval-censored however the failure time is only right-censored or exactly observed, alternatives are offered by Goggins, Finkelstein, and Zaslavsky (1999); Sun, Liao, and Pagano (1999); Pan (2001).

5.1.4 Accelerated failure time model

A parametric AFT model estimated using the maximum likelihood method can be used with interval-censored data as well. It is also implemented in major statistical packages (functions `survreg` in R and `SurvReg` in SPLUS, procedure `LIFEREG` in SAS). On the other hand, semi-parametric methods which are not straightforward even with right-censored data are only with considerable difficulties extended to the interval-censored data, see Rabinowitz, Tsiatis, and Aragon (1995); Betensky, Rabinowitz, and Tsiatis (2001). Though, both approaches are practically applicable only with low-dimensional covariate vectors $\boldsymbol{x}$ and as well as for right-censored data there exists no non-parametric method to estimate the baseline survival distribution implying that the semi-parametric procedures cannot be used when prediction is of interest.

More promising alternatives are the methods that make use of multiple imputation and/or smoothing. Indeed, approaches of Pan and Kooperberg (1999); Pan and Louis (2000); Pan and Connett (2001) introduced in Sections 3.4.2 and 3.4.3 could relatively easily be extended to handle also (multivariate) interval-censored or even doubly-interval-censored data. However, it can be computationally demanding, especially with doubly-interval-censored data, to perform integration of the form (4.4) in the optimization of the likelihood.

5.1.5 Interval-censored covariates

Up to now, we concentrated on the problem of interval-censored response. In the regression context, it is however possible in practice, that we have to face the problem of interval-censored covariate. Such problem is considered, for example, by Gómez, Espinal, and Lagakos (2003) who studied, in the framework of an HIV/AIDS clinical trial, the association between waiting time between indinavir failure and enrolment (covariate) and subsequent viral load (response).

However, we will not consider problems of this type in this thesis. Recent developments in this field can be found, e.g., in Topp and Gómez (2004); Langohr, Gómez, and Muga (2004); Calle and Gómez (2005).

5.2 Bayesian proportional hazards model: An illustration

For an extensive overview of the Bayesian methods for the proportional hazards model we refer the reader to the book of Ibrahim, Chen, and Sinha (2001). Here, only the analysis based on the PH model, published as Komárek et al. (2005), will be presented and that of doubly-interval-censored data from the Signal Tandmobiel® study. Actually, the main purpose of this section is to illustrate typical features of a Bayesian analysis and show how it can be used to answer rather complex questions.

In Section 5.2.1, we formulate the research question and outline the problems related to this question. Section 5.2.2 presents a frequentist Cox's PH regression model using midpoints of the observed intervals as if they were exact observations, to compare our Bayesian approach to a more commonly used, however incorrect, approach. In Section 5.2.3, the Bayesian model suggested by Härkänen, Virtanen, and Arjas (2000) and modified for our purposes is explained and results are presented in Section 5.2.4. We finalize this part by a discussion.

5.2.1 Signal Tandmobiel® study: Research question and related data characteristics

In this section we will tackle the following research question: *Does fluoride-intake at a young age have a protective effect on caries in permanent teeth?* Our analyses will be limited to caries experience of the four permanent first molars (teeth number 16, 26, 36, 46 in Figure 1.1).

The data suggest that the use of fluoride reduces caries experience in primary teeth, see Vanobbergen et al. (2001) and that fluoride-intake delays the emergence of the permanent teeth, see Leroy et al. (2003a). The latter result raises the question whether the fluoride-intake only reduces the time at risk or whether it has also a direct protective effect on caries experience.

Unfortunately, fluoride-intake in children cannot be measured accurately. Indeed, fluoride-intake can come from: (1) fluoride supplements (systemic), (2) accidental ingestion of toothpaste or (3) tap water. Further, the intake from these sources can be recorded only crudely. Therefore it was decided to measure fluoride-intake by the degree of fluorosis on some reference teeth. Fluorosis is the most common side-effect of fluoride-intake and appears as white spots on the enamel of teeth. For this analysis, a child was considered fluoride-positive (covariate fluor $= 1$) if there were white spots on at least two permanent maxillary incisors during the fourth year of the study or during both the fifth and sixth year of the study.

The prevalence of fluorosis was relatively low (480 children, 10.8%). In our analysis, 480 fluorosis children and 960 randomly selected fluorosis-free children are included. Case-control subsampling was done to reduce computation time. To check that it did not destroy the stratification, we constructed a $5 \times 3 \times 2$ contingency table with factors province, school system and whether the child is in the subsample or not (subsample). A classical p-value of 0.13 was obtained for the significance of the interaction of the third factor with the other two using a likelihood-ratio test in a log-linear model, implying that the stratification is similar in the used and the discarded subsamples.

The prevalence of caries experience at the age of 12 was negligible (at most 1.4%) for all permanent teeth except for the first molars (teeth used in the analysis). For these teeth the prevalence was 25.8% in children with fluorosis compared to 29.4% in fluorosis-free children, with prevalence of 23.3% and 27.7% for boys, and 27.9% and 31.2% for girls, respectively. Thus, at first sight the impact of fluoride-intake seems to be minor. However, since the emergence of permanent teeth might be delayed by fluoride-intake, evaluating the impact of fluoride-intake should take into account the time at risk for caries. Hence, in our analysis the response will be the time between emergence

and the onset of caries development. Remember that both tooth emergence and onset of caries development are interval-censored, implying a doubly-interval-censored response. See Figure 2.1 for a graphical illustration of a possible evolution of a particular tooth.

At the onset of the study about 86% of the permanent first molars had already emerged. The severity of this censoring will affect the efficiency with which the effect of fluoride-intake can be estimated. We tried two strategies to improve the efficiency of our estimation procedure. Firstly, we included in our analysis the emergence times of teeth 14, 24, 34, 44, 12, 22, 33, 43 all of which had emerged in more than 60% of cases during the course of the study. By incorporating information on these teeth and using the association between teeth of the same subject (via the concept of "the birth time of dentition", see next section), it was attempted to better estimate the true emergence time of the permanent first molars. Secondly, emergence times from a Finnish longitudinal data set (Virtanen, 2001), involving 235 boys and 223 girls born in 1980–1981 with follow-up from 6 to 18 years, were added to our Flemish data. For these Finnish data almost all 28 permanent teeth emerged during the study period.

Our research question is not uncommon in dentistry, but cannot be addressed within any classical statistical package. For our analysis, we have used the software package BITE (Härkänen, 2003), based on a semi-parametric Bayesian survival model developed by Härkänen et al. (2000).

5.2.2 Proportional hazards modelling using midpoints

A standard frequentist Cox's PH model introduced in Section 3.1 could be applied, replacing interval-censored observations by the midpoints of the observed intervals and treating the resulting data as right-censored observations. In this way, we analyzed time to caries development for the four permanent first molars. For our analysis, the left-censored emergence times were first assumed to be interval-censored with a lower limit for emergence of 5 years, which is practically the youngest age for the emergence of these teeth (Nanda, 1960). Possible dependencies between the four teeth of the same child can be taken into account, for example by inclusion of a gamma–frailty component in the PH model as explained in Section 3.4.1.

Based on preliminary Bayesian modelling, we do not distinguish between opposite teeth in the same jaw and assume so called *horizontal symmetry*. However, we do make a distinction between maxillary (upper) and mandibular (lower) teeth and also between teeth in different positions (of a quadrant) in the mouth.

Table 5.2: Signal Tandmobiel® study. Naive PH models for the effect of fluorosis on caries on permanent first molars. Hazard ratios (95% confidence intervals (CI)) between a fluorosis and fluorosis-free group of children while controlling for gender and jaw.

	Model WITHOUT frailties		Model WITH frailties	
Group	Estimate	95% CI	Estimate	95% CI
Boys, maxilla	0.787	(0.541, 1.032)	0.704	(0.204, 1.204)
Boys, mandible	0.733	(0.532, 0.934)	0.613	(0.231, 0.995)
Girls, maxilla	0.871	(0.698, 1.044)	0.892	(0.610, 1.174)
Girls, mandible	0.812	(0.670, 0.953)	0.776	(0.559, 0.993)

For comparison purposes, we present the same PH model as the one shown in Section 5.2.3 but analyzed by Bayesian methods. Hence, the hazard for the time to caries of the lth tooth of the ith child depends on the tooth position, fluor and gender of the child (0 = boy, 1 = girl). More specifically:

$$\hbar(t|\mathsf{tooth}_l, \mathsf{gender}_i, \mathsf{fluor}_i) = \hbar_0(t) \cdot Z_i \cdot \exp(\boldsymbol{\beta}' \boldsymbol{x}_{i,l}), \tag{5.1}$$

$i = 1, \ldots, N,\ l = 16, 26, 36, 46$, where $\hbar_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)'$, and $\boldsymbol{x}_{i,l} = (\mathsf{fluor}_i,\ \mathsf{gender}_i,\ \mathsf{tooth}_l,\ \mathsf{fluor}_i \times \mathsf{gender}_i,\ \mathsf{fluor}_i \times \mathsf{tooth}_l)'$. The covariate "tooth" is a dummy variable that distinguishes teeth on different positions in the mouth (apart from horizontal symmetry). The term Z_i is either one, corresponding to a model without frailties, or a gamma distributed frailty term.

Estimates of hazard ratios between the fluorosis and fluorosis-free group controlling for gender and jaw are shown in Table 5.2. As seen, incorrectly ignoring dependencies between the responses of one child by using a model without frailties artificially decreases the size of the confidence interval. Although both models conclude that the effect of fluorosis on the development of caries on the permanent first molars is at the borderline of 5% significance (Table 5.2), the results are not reliable. As pointed on page 39, the correctness of the midpoint imputation depends strongly on the underlying distribution of the event times. For that reason, a more sophisticated analysis is needed.

5.2.3 The Bayesian survival model for doubly-interval-censored data

The non-parametric Bayesian intensity model of Härkänen et al. (2000) provides a flexible tool for analyzing multivariate survival data. Further, a software package written in C, called BITE and downloadable from

`http://www.rni.helsinki.fi/~tth`

together with scripts used to perform all analyses presented here, makes the analysis feasible in practice.

Model for emergence

Let $U_{i,l}$ be the (unknown) age at which tooth l of child i emerged. The hazard for emergence of tooth l of the ith child at time t is

$$\lambda_{i,l}^{(e)}(t) = \hbar^{(e)}(t - \eta_i | \mathsf{tooth}_l, \mathsf{gender}_i) \times I[\eta_i < t \leq U_{i,l}]. \tag{5.2}$$

The dependence between emergence times of one child is accounted for by using a subject-specific variable η_i called *birth time of dentition*. This is a latent variable which represents the common time marking the onset of the tooth eruption process and hereby "explains" the positive correlation between eruption times $U_{i,l}$ within a subject. Note that η_i is always less than the first emergence time of the permanent teeth. The intensity of emergence for a particular child is zero before that time, expressed by the indicator $I[\eta_i < t \leq U_{i,l}]$. The hazard function $\hbar^{(e)}(\cdot|\mathsf{tooth}_l, \mathsf{gender}_i)$ is defined as piece-wise constant for estimation purposes.

Model for caries experience

Let $V_{i,l}$ be the age at which the lth tooth of child i developed caries. The hazard for the caries process is given by

$$\lambda_{i,l}^{(c)}(t) = Z_i \times \hbar^{(c)}(t - U_{i,l} | \mathsf{tooth}_l, \mathsf{gender}_i, \mathsf{fluor}_i) \times I[U_{i,l} < t \leq V_{i,l}], \tag{5.3}$$

where the variable Z_i is an unknown subject-specific frailty coefficient modulating the hazard function. Again, we assume in (5.3) that h is piece-wise constant. We call the difference $V_{i,l} - U_{i,l}$ *the time-to-caries.*

The covariate "fluor" will be used in two ways. Firstly, for each combination of values of fluor, gender and tooth a piece-wise constant hazard function is specified and fitted. Secondly, the term $\hbar^{(c)}(\cdot|\mathsf{tooth}_l, \mathsf{gender}_i, \mathsf{fluor}_i)$ in (5.3)

is replaced by $\hbar_0^{(c)}(\cdot) \times \exp(\boldsymbol{\beta}' \boldsymbol{x}_{i,l})$, with $\boldsymbol{\beta}$ and $\boldsymbol{x}_{i,l}$ being the same as in (5.1), thus assuming a PH model for caries experience whilst retaining a piece-wise constant baseline hazard function $\hbar_0^{(c)}(\cdot)$.

Remarks

Our statistical model will involve the above two measurement models. Hence the possible dependencies among times of interest are taken into account by involving two types of subject-specific parameters, η_i and Z_i. The first subject-specific parameter η_i is included in the model for the emergence and will shift the hazard function in time, whereas the frailty Z_i recognizes that the teeth of one child can be more sensitive to caries than the corresponding teeth of another child, reflecting different dietary behavior, brushing habits, etc.

Priors for baseline hazard functions

In BITE the working assumption is that hazard functions are piece-wise constant. Further, for the emergence hazard functions $\hbar^{(e)}(\cdot|\mathsf{tooth}_l, \mathsf{gender}_i)$ the first level of the piece-wise constant and the increment levels are assigned gamma prior distributions. This will ensure a priori an increasing hazard function for emergence. In the case of caries experience, the first level of the piece-wise constant hazard function $\hbar^{(c)}(\cdot|\mathsf{tooth}_l, \mathsf{gender}_i, \mathsf{fluor}_i)$ in the non-parametric model and $\hbar_0^{(c)}(\cdot)$ in the PH model, say h_0, is assigned a gamma prior distribution. Further, the level h_m of the mth interval has, conditional on the previous levels $h_0, \ldots, h_{m-1}$, a Gamma(α, α/h_{m-1}) prior distribution. This gives a priori $\mathrm{E}[h_m|h_{m-1}, \ldots, h_0] = h_{m-1}$ and assures that there is no built-in prior assumption of trend in the hazard rate. Finally, the prior for the jump points of each piece-wise constant function is a homogeneous Poisson process, as suggested by Arjas and Gasbarra (1994). Because jump points are assumed to be random and not fixed, the posterior predictive hazard functions will be smooth, rather than piece-wise constant.

Priors for the random effect terms

The prior distribution for the birth time of dentition η_i illustrates how we have combined the Flemish data and the Finnish data, and how the timing of emergence of the Finnish data is included in our analysis. We assume that the shapes of the emergence hazard functions f for Finland and Flanders are the same, but we do allow for a shift in emergence times by assuming different

means for the birth time of dentition in the two countries. More precisely, the prior distribution of η_i is assumed normal $\mathcal{N}(\xi_0, \tau^{-2})$ for a Finnish child and normal $\mathcal{N}(\xi_1, \tau^{-2})$ for a Flemish child.

The Bayesian approach allows us to include the dentist's knowledge on the problem at hand by assigning to the parameters ξ_0 and ξ_1 independent normal prior distributions with mean 5.2 years and standard deviation 1 year. Both the normal distribution as well as the choice of the prior means and standard deviation of the hyperparameters ξ_0 and ξ_1 are motivated by the results found in the literature on the earliest emergence of permanent teeth, see Nanda (1960) or more recently Parner et al. (2001). This reflects the dentist's belief that permanent teeth on average emerge slightly after 5 years of age. The parameter τ^2 is assigned a Gamma(2, 2) prior distribution.

The individual frailties Z_i in the model for caries are a priori assumed to be conditional on the hyper-parameter ϕ, independent and identically gamma distributed with both shape and inverse scale equal to that hyper-parameter. The hyper-parameter itself is then given a Gamma(2, 2) prior distribution. Sensitivity of the results with respect to the choice of parameters for priors of hyperparameters ξ_0, ξ_1, τ and ϕ will be discussed in Section 5.2.4.

Treatment of censored data

Left- and interval-censoring are treated by Bayesian data augmentation introduced in Section 4.3. Additionally, the left-censored emergence times of all teeth are changed into interval-censored emergence times with a lower limit equal to 4 years, implying that less internal information is used here than previously with the frequentist PH model where the limit was 5 years. In the case that both emergence and caries development were observed within one observational interval we force sampled values of the MCMC to satisfy $V_{i,l} > U_{i,l}$.

Bayes inference on model components

The posterior distributions based on the model with prior assumptions described in the previous paragraphs are minor modifications of those derived in Härkänen et al. (2000). Our Bayesian model is complex and requires the use of Markov Chain Monte Carlo sampling techniques outlined in Section 4.5. The software package BITE (Härkänen, 2003), based on the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), was used to sample from the posterior distributions. Further, BITE employs the reversible jump approach of Green (1995) to sample piece-wise constant hazard

functions. We carried out two runs, each with 20 000 iterations of burn-in followed by 14 000 iterations with a 1:4 thinning to obtain a sample from the posterior distribution. We used the Gelman and Rubin (1992) test to check for convergence.

5.2.4 Results

A non-parametric model with Flemish and Finnish data

To evaluate the effect of fluoride-intake on the development of caries on the permanent first molars we have calculated the posterior expectations of hazard ratios

$$\frac{\hbar^{(c)}(t|\mathsf{tooth},\mathsf{gender},\text{fluorosis})}{\hbar^{(c)}(t|\mathsf{tooth},\mathsf{gender},\text{fluorosis-free})}.$$

These hazard ratios together with their 95% equal tail point-wise credible intervals can be found in Figure 5.2. The PH assumption with respect to covariate fluor seems to be satisfied since credible intervals in all cases cover

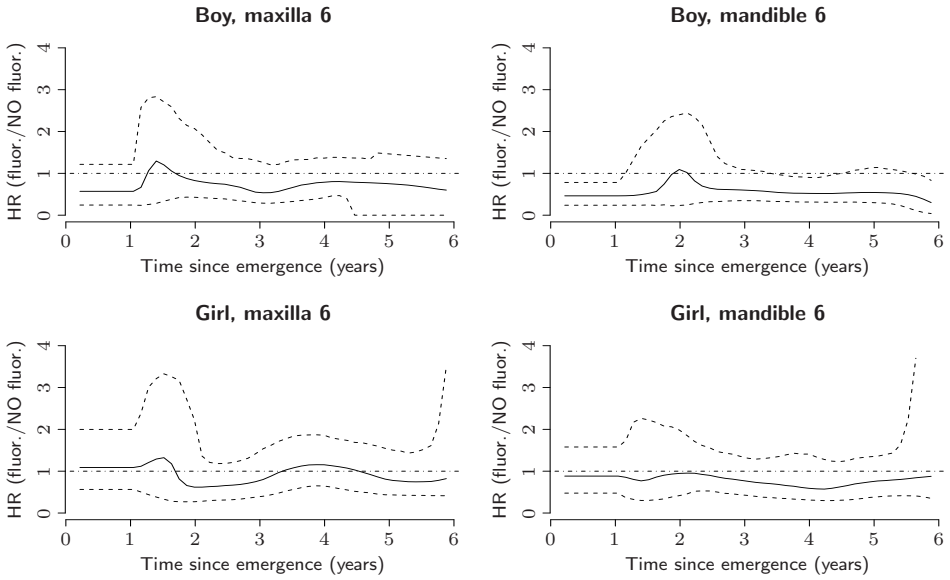


Figure 5.2: Signal Tandmobiel® study. Bayesian non-parametric model based on Flemish and Finnish Data. Posterior means of the hazard ratios between the fluorosis groups (solid line), 95% point-wise equal tail probability region (dashed line).

Table 5.3: Signal Tandmobiel® study. Bayesian PH models for the effect of fluorosis on caries on permanent first molars. Hazard ratios (95% equal-tail credible intervals (CI)) between fluorosis groups while controlling for gender and jaw for models fitted using both Flemish and Finnish data and Flemish data only.

	Flemish and Finnish data		Flemish data only	
	Posterior		Posterior	
Group	mean	95% CI	mean	95% CI
Boys, maxilla	0.674	(0.492, 1.010)	0.651	(0.463, 0.960)
Boys, mandible	0.572	(0.414, 0.850)	0.549	(0.386, 0.779)
Girls, maxilla	0.991	(0.721, 1.364)	1.002	(0.698, 1.333)
Girls, mandible	0.840	(0.608, 1.136)	0.844	(0.602, 1.135)

a horizontal line. In three cases, this horizontal line is close to the dotted-dashed line $y = 1$ implying no effect of fluoride-intake on caries development. A positive effect of fluoride intake seems to be present only for mandibular permanent first molars in boys. There are also no deviations from the PH assumption with respect to gender and tooth (plots are not shown). This allowed us to assume for the caries model a PH effect of the three covariates, possibly including some interaction terms. By this semi-parametric assumption it was hoped to see more clearly the effect of fluoride-intake on caries experience.

A proportional hazards model with Flemish and Finnish data

For reasons stated in the previous paragraph, we have fitted a model where the caries hazard function (5.3) was changed into

$$\lambda_{i,l}^{(c)}(t) = Z_i \times \hbar_0^{(c)}(t) \times \exp(\boldsymbol{\beta}' \boldsymbol{x}_{i,l}) \times I[U_{i,l} < t \leq V_{i,l}], \tag{5.4}$$

where $\boldsymbol{x}_{i,l}$ and $\boldsymbol{\beta}$ are the same as in (5.1). The additional β-parameters were given a $\mathcal{N}(0, 10^2)$ prior. However, the hazard function for emergence is still defined by (5.2). Posterior expectations of the hazard ratios between the fluorosis groups while controlling for the other covariates are given in the left part of Table 5.3.

The PH analysis for caries gives similar conclusions to the previous non-parametric analysis. A positive effect of fluoride-intake is now seen for the mandibular permanent first molars of boys and has a borderline positive

Table 5.4: Signal Tandmobiel® study. Bayesian models with Flemish and Finnish Data. Posterior means and 95% equal-tail credible intervals for the hyperparameters μ_0 – conditional expectation of η_i for Finland, μ_1 – conditional expectation of η_i for Flanders, τ^{-2} – conditional variance of η_i, ϕ^{-1} – conditional variance of frailties Z_i (top of the Table). Means of the posterior predictive distributions and 95% equal tail posterior predictive intervals for the birth time of dentition η_i in Finland and Flanders, respectively, and for the frailty term Z_i (bottom of the Table).

Posterior mean (95% credible interval)		
Hyperparameter	Non-parametric model	Cox regression model
μ_0	5.47 (5.40, 5.54)	5.45 (5.38, 5.52)
μ_1	5.69 (5.64, 5.73)	5.68 (5.64, 5.73)
τ^{-2}	0.48 (0.45, 0.52)	0.49 (0.45, 0.52)
ϕ^{-1}	3.85 (3.57, 4.17)	3.94 (3.58, 4.28)

Posterior predictive mean (95% posterior predictive interval)		
Parameter	Non-parametric model	Cox regression model
η_i (Finland)	5.48 (4.12, 6.79)	5.45 (4.05, 6.84)
η_i (Flanders)	5.69 (4.33, 7.09)	5.69 (4.34, 7.01)
Z_i	1.02 (10^{-6}, 6.90)	0.95 (10^{-6}, 6.45)

effect for the maxillary permanent first molars of boys. However, no effect of fluoride intake was seen for girls.

Remark concerning hyperparameters

The posterior expectations and 95% equal-tail credible intervals of the hyperparameters related to the birth times of dentition η_i and frailties Z_i are given in the upper part of Table 5.4. The non-parametric model and PH model for caries give similar results.

We now state our conclusions concerning the emergence process in Flanders and Finland. The emergence process starts slightly earlier in Finland (by approx. 0.2 years) than in Flanders, as is seen by the difference in the posterior expectations of the means of birth time of dentition. The MCMC output for the hyperparameters can also be used to estimate properties of the predictive distributions of birth time of dentition and frailties. Their means and 95% equal-tail posterior predictive intervals are shown in the bottom part of Table 5.4, which shows that the average of Finnish birth time of dentition is

close to 5.5 years of age, slightly higher than the prior expectation but close to the value obtained by Härkänen et al. (2000) on another Finnish data set. The 95% posterior predictive intervals show that the actual moment of birth time of dentition varies between about 4 and 7 years of age. Finally, the 95% posterior predictive interval of Z_i shows a clear heterogeneity in the frailty for caries experience.

Sensitivity analysis

Firstly, the model (5.4) was fitted using Flemish data only, to see how influential was inclusion of the Finnish data. As seen in Table 5.3, the hazard ratios changed only slightly. The same was true for the remaining parameters. Moreover, the Finnish data improved only slightly the precision with which the emergence of the first permanent molars was estimated. This is seen in Figure 5.3 which shows a comparison of 95% pointwise equal tail credible regions for the emergence hazard functions of the permanent first molars based on the analysis with both data sets and the Flemish data set only. Though,

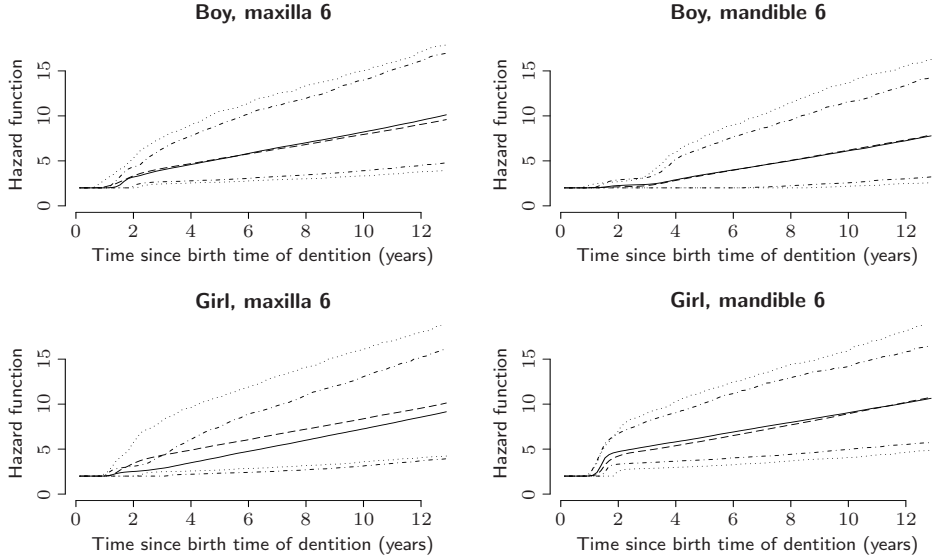


Figure 5.3: Signal Tandmobiel® study. Bayesian PH models. Posterior means of the emergence hazard functions $\hbar^{(e)}(\cdot|\mathsf{tooth},\mathsf{gender})$ for the permanent first molars together with their 95% pointwise equal tail probability regions. Comparison of the posterior mean with (solid line) and without additional Finnish data (dashed line), respectively together with 95% prob. regions (dotted-dashed line, dotted line respectively).

the credible regions are somewhat narrower when both databases are used.

To see how the behavior of the parameter estimates changes when informative priors for the hyperparameters are modified we have fitted the proportional hazards model with Flemish data only, using different choices of priors for the hyperparameters. Specifically, we used normal distributions $\mathcal{N}(3,\ 2)$, $\mathcal{N}(4,\ 1)$, $\mathcal{N}(5.2,\ 1)$, $\mathcal{N}(6,\ 1)$ as a priors for the expectation ξ_0 of birth time of dentition η_i. The standard deviation of the normal prior with mean 3 years was increased so as to cover realistic emergence times of permanent teeth. We used Gamma(0.1, 0.1), Gamma(2, 2), and Gamma(10, 10) distributions as priors for the precision τ of the variance of the birth time of dentition and for the precision ϕ of frailties Z_i. All other parameters were given flat priors and there is thus no reason to modify them.

Posterior means and 95% equal-tail credible intervals for hazard ratios between the fluorosis and fluorosis-free groups for different choices of the prior distributions are shown in Figure 5.4, which shows that the influence of the choice of the prior distribution is not strong.

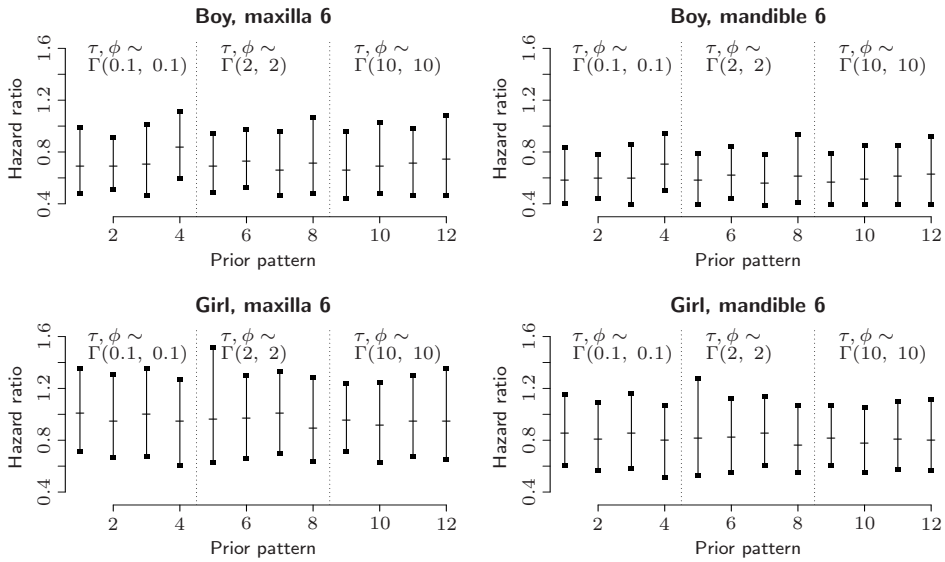


Figure 5.4: Sensitivity Analysis. Evolution of posterior mean and 95% credible intervals for the hazard ratios between the fluorosis and fluorosis-free groups with changing prior distributions for hyperparameters τ, ϕ and ξ_0. Prior patters number 1, 5 and 9 use $\mathcal{N}(3,\ 2)$ prior for ξ_0, patterns number 2, 6 and 10 use $\mathcal{N}(4,\ 1)$ prior for ξ_0, patterns number 3, 7 and 11 use $\mathcal{N}(5.2,\ 1)$ prior for ξ_0 and patterns number 4, 8 and 12 use $\mathcal{N}(6,\ 1)$ prior for ξ_0.

We argue that our other assumptions are not strong. Indeed, we assume that the distributions of the birth time of dentition differ between Finnish and Flemish populations only in their means. Moreover, as indicated above, the Finnish data had only a slight impact on the results for the Flemish data. Further, the baseline hazards were estimated non-parametrically. Finally, different choices for the priors of the hyperparameters led to similar results as discussed above.

5.2.5 Discussion

The model presented here allows for the analysis of survival data in dental research where (doubly-)interval-censored data and dependencies between observations (e.g. between teeth in the same mouth) are common. Our specific application is to a typical dental research question, i.e. whether fluoride-intake has a protective effect for caries. The results show that the protective effect of fluoride-ingestion is not convincing. We observed a positive effect only for mandibular teeth of boys. This agrees with current guidelines for the use of fluoride in caries prevention, where only the *topical* application (e.g. fluoride in tooth paste) is considered to be essential (Oulis, Raadal, and Martens, 2000).

We acknowledge that our analyses could have been more refined if the amount of left- and right censoring was less, for instance if the study had started approximately one year earlier and ended in high school. This would make our analyses less dependent on prior assumptions. Yet these prior assumptions are simply a reflection of basic dental knowledge and it would be a waste not to use them. Moreover, to our knowledge the Signal Tandmobiel® study is possibly the largest longitudinal study executed with such great detail on dental aspects.

This section has illustrated the usefulness of the Bayesian approach. Firstly, it was possible to incorporate prior information and to relax the parametric assumptions often made in survival analysis with interval-censored data. Secondly, even rather complex models could be specified for doubly-interval-censored data. However, we have to admit that this approach is computationally demanding. On a Pentium IV 2 GHz PC with 512 MB RAM one BITE run took about 5 days to converge. However, in an epidemiological analysis where there is correlation among the subjects, where the response and/or the covariates are (right-, left- or interval-) censored and when we wish to avoid parametric assumptions we doubt any classical approach will suffice.

5.3 Bayesian accelerated failure time model

Most contributions to the AFT model in the Bayesian literature work explicitely only with right-censored data. However, using the idea of Bayesian data augmentation (Section 4.3) they can all be quite easily extended to handle also interval-censored data. Additionally, actually all papers dealing with the Bayesian AFT model use a Bayesian non-parametric approach (see Walker et al., 1999 or the book Ghosh and Ramamoorthi, 2003) for the distributional parts of the AFT model. In this section, we give a brief overview.

Firstly, Christensen and Johnson (1988) and Johnson and Christensen (1989) consider the basic univariate AFT model (3.2) and use a Dirichlet process prior (Ferguson, 1973, 1974) for the underlying baseline survival distribution, i.e. the distribution of $\exp(\varepsilon)$. In the former paper, only a semi-Bayesian approach is used, whereas the latter paper presents a fully Bayesian analysis however, with uncensored data only. The authors state that "The analysis becomes totally intractable when there are censored observations." Additionally, as discussed in Johnson and Christensen (1989), difficulties might arise due to the discrete nature of a Dirichlet process (the baseline survival distribution is discrete with probability one if it is assigned the Dirichlet process prior). An improvement is presented by Kuo and Mallick (1997) who consider a Dirichlet processes *mixture* (Lo, 1984) for either ε or $\exp(\varepsilon)$.

Subsequently, Walker and Mallick (1999) suggest to use a diffuse, finite Pólya tree prior distribution described in Lavine (1992, 1994) and Mauldin, Sudderth, and Williams (1992) for the error term ε in the AFT model (3.2). The main advantages of the Pólya tree prior distribution are (1) it can assign probability one to the set of continuous distribution, (2) it is easy to constraint the resulting error term ε to have the median (or any other quantile) rather than the mean equal to zero (or any other fixed number) such that also the regression quantiles can be modelled, of which the median regression is the most important case. Additionally, Walker and Mallick (1999) break down the i.i.d. assumption of the error terms and assume also the population averaged AFT model (3.6).

Successive approaches to the Bayesian non-parametric AFT concentrate on the *median* regression. Namely, Kottas and Gelfand (2001) suggest to use the Dirichlet process mixture of either unimodal parametric densities or unimodal step functions for the distribution of the error term ε in the basic AFT model (3.2). Another median regression AFT model is given by Hanson and Johnson (2002) who use a mixture of Pólya trees centered about a standard, parametric family of probability distributions as a prior for the error term ε.

Finally, Hanson and Johnson (2004) consider a mixture of Dirichlet processes

introduced by Antoniak (1974) (which is distinct from the Dirichlet process mixture used by Kuo and Mallick, 1997 or Kottas and Gelfand, 2001) as the prior for the error term ε in the basic AFT model (3.2). They also consider explicitely the interval-censored data.

The area of multivariate survival data modelled by the mean of the Bayesian AFT model seems to be almost unexplored. Except the work Walker and Mallick (1999) we are not aware of any other contribution. Moreover, the structured modelling of dependencies by the mean of the cluster specific AFT model introduced in Section 3.4.3 seems to be absent at all in the literature.

5.4 Concluding remarks

In this chapter and in Chapter 3 we came across with two fundamental regression models for the survival data. We mentioned that the most frequently used PH model has several drawbacks so that in many practical situations it is worthy to consider alternatives of which the AFT model is an appealing one. We pointed out that the AFT model whose distributional parts are parametrically specified can relatively easily be estimated even using the method of maximum-likelihood. However, especially for prediction purposes, it is important to avoid incorrectly specified parametric models since due to the censoring any parametric assumption is very difficult to check with survival data. For that reason, one aims for methods that leave the distributional parts of the model either completely unspecified or specify them in a flexible way. For the PH model, the partial likelihood due Cox (1975) is available for this purpose. Unfortunately, no similar concept exists for the AFT model. Several frequentist semi-parametric methods were reviewed in Sections 3.2, 3.4.2, 3.4.3, and 5.1.4. Nevertheless, we saw that, especially with interval censoring, or let alone doubly interval censoring, most of them become computationally intractable in practical situations. Moreover, with multivariate data, the situation becomes even more complex.

On the other hand, the Bayesian approach together with data augmentation offer an appealing alternative allowing to formulate and also estimate realistically complex models even with multivariate and/or (doubly-)interval-censored data. We have illustrated this issue on the Bayesian semi-parametric PH model in Section 5.2. In Section 5.3, we have subsequently reviewed existing semi-parametric approaches to the AFT model. However, we mentioned that most of them were primarily developed to handle only univariate data. Nevertheless, many survival problems lead to the analysis of the multivariate data.

Concluding Remarks to Part I and Introduction to Part II

We have introduced two versions of the AFT model - the population-averaged and the cluster-specific model that can be used to analyze the multivariate survival data. We have also mentioned that, especially for the cluster-specific AFT model (3.7), with unspecified distributional parts of the model, there is almost no methodology developed in the literature.

In this thesis, we aim to present the methods to handle both the population-averaged AFT model (3.6) and the cluster-specific AFT model (3.7) under the presence of multivariate and/or (doubly-)interval-censored data. At the same time, we want to minimize the parametric assumptions concerning the distributional parts of the model as much as possible. One possibility to reach this target is to use smoothing methods for the unknown distributional parts. In the literature, more often the baseline hazard function is smoothed (Section 5.1.3: Kooperberg and Clarkson, 1997; Joly et al., 1998; Betensky et al., 1999; Section 5.2: Härkänen et al., 2000; Komárek et al., 2005).

However, with the AFT model, it is quite natural to use a flexible smooth expression for the density, either of the error term ε and/or the random effects $\boldsymbol{b}$. For example, for the bivariate population-averaged AFT model, Pan and Kooperberg (1999) use this idea in combination with the multiple imputation (see Section 3.4.2).

In principal, the methods presented in Part II of this thesis will be built on the same basis as that of Pan and Kooperberg (1999). Whereas they express the logarithm of the unknown density using the splines and use numerical integration to evaluate and optimize the likelihood we will model directly the density using a linear combination of suitable basis parametric functions and simplify thus the likelihood evaluation (see Section 6.2.4). In contrast to Pan

and Kooperberg (1999) we also exploit another strategy to determine the number of the basis functions. Whereas they choose the optimal number of basis functions using a criterion like AIC (Akaike, 1974) we will either take an overspecified number of the basis functions and prevent identifiability problems and overfitting the data using a penalty term (Chapters 7, 9, 10) or estimate the number of the basis functions simultaneously with the other model parameters (Chapter 8).

Further, we will show that for univariate survival data we are able, even under the interval-censoring to use maximum-likelihood based methods without the need for multiple imputation (Chapter 7). With the introduction of multivariate and doubly-interval-censored data we avoid multiple imputation by switching to the Bayesian approach (Chapters 8, 9, 10) which is more advantageous in such situation as was explained in Chapter 4.

Part II

Accelerated Failure Time Models with Flexible Distributional Assumptions

Chapter 6

Mixtures as Flexible Models for Unknown Distributions

We aim to develop the accelerated failure time models with flexibly specified distributional parts. We have already sketched that we wish to use flexible, yet smooth expressions for densities involved in the specification of these distributional parts. In this chapter, let $g(y)$ ($g(\boldsymbol{y})$) denote an unknown density of some generic univariate random variable Y (random vector $\boldsymbol{Y}$). We outline two similar, though conceptually different, methods to approximate $g(y)$ or $g(\boldsymbol{y})$ in a flexible and smooth way, namely

1. The classical mixture approach;
2. An approach based on penalized smoothing.

We introduce the classical mixture approach in Section 6.1. In Section 6.2, the penalized smoothing approach exploiting B-splines will be given. In Section 6.3, we replace the B-splines by normal densities and introduce the penalized normal mixture. Finally, we compare the classical and penalized normal mixture in Section 6.4.

6.1 Classical normal mixture

6.1.1 From general finite mixture to normal mixture

To model unknown distributional shapes *finite mixture* distributions have been advocated by, e.g., Titterington, Smith, and Makov (1985, Section 2.2) as appealing *semi-parametric* structures. Using a finite mixture the density

$g(\boldsymbol{y})$ is modelled in the following way:

$$g(\boldsymbol{y}) = g(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \sum_{j=1}^{K} w_j g_j(\boldsymbol{y}), \tag{6.1}$$

where g_j, $j = 1, \ldots, K$ are known densities and $\boldsymbol{\theta} = (K, w_1, \ldots, w_K)'$ is the vector of unknown parameters. Namely, K is the number of mixture components, and w_j, $j = 1, \ldots, K$ are unknown weights satisfying $w_j > 0$, $j = 1, \ldots, K$ and $\sum_j w_j = 1$. In general, the number of mixture components, K, is assumed unknown, however, due to difficulties outlined further in the text, estimation of K is often separated from estimation of the remaining parameters, especially when using maximum-likelihood based methods.

Further, it is often assumed that the mixture components, g_j, $j = 1, \ldots, K$ have a common parametric form $\tilde{g}$ and each mixture component depends on an unknown vector of parameters $\boldsymbol{\eta}_j$, $j = 1, \ldots, K$. Expression (6.1) changes then into

$$g(\boldsymbol{y}) = g(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \sum_{j=1}^{K} w_j \tilde{g}(\boldsymbol{y} \,|\, \boldsymbol{\eta}_j), \tag{6.2}$$

where $\boldsymbol{\theta} = (K, w_1, \ldots, w_K, \boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_K')'$. A frequently used particular form of (6.2) is a normal mixture where $\tilde{g}(\boldsymbol{y} \,|\, \boldsymbol{\eta}_j)$ equals $\varphi(\boldsymbol{y} \,|\, \boldsymbol{\mu}_j, \Sigma_j)$, a density of the (multivariate) normal distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix Σ_j. For instance, Verbeke and Lesaffre (1996) use a mixture of multivariate normal distributions with $\Sigma_j = \Sigma$ for all j to model a distribution of the random effects in the linear mixed model.

In this thesis, we use the classical normal mixture only in a univariate context, i.e. to express an unknown univariate density $g(y)$ as

$$g(y) = g(y \,|\, \boldsymbol{\theta}) = \sum_{j=1}^{K} w_j \varphi(y \,|\, \mu_j, \sigma_j^2). \tag{6.3}$$

In this case, the vector $\boldsymbol{\theta}$ equals

$$\boldsymbol{\theta} = (K, w_1, \ldots, w_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)'. \tag{6.4}$$

Figure 6.1 illustrates how two- or four-component, even homoscedastic, normal mixtures can be used to obtain densities of different shapes.

6.1.2 Estimation of mixture parameters

Let $\boldsymbol{\theta}$ be a vector given by the expression (6.4) and containing all unknown parameters of model (6.3). Suppose first that an i.i.d. sample $y_1, \ldots, y_n$

from a density $g(y \mid \boldsymbol{\theta})$ is available to estimate the unknown parameter vector $\boldsymbol{\theta}$. *Maximum-likelihood* based methods pose two main difficulties when estimating $\boldsymbol{\theta}$:

1. When K, the number of mixture components is unknown, one of the basic regularity conditions for the validity of the classical maximum-likelihood theory is violated. Namely, the parameter space does not have a fixed dimension. Indeed, the number of unknowns (number of unknown mixture weights, means and variances) is one of the unknowns. See, e.g., Titterington et al. (1985, Section 1.2.2) for a detailed discussion of this difficulty.

2. For a fixed $K \geq 2$, the likelihood becomes unbounded resulting in non-existence of the maximum-likelihood estimate when one of the mixture means, say μ_1, is equal to one of the observations y_i, $i = 1, \ldots, n$ and when the corresponding mixture variance, σ_1^2, converges to zero. See, e.g., McLachlan and Basford (1988, Section 2.1) for more details.

In classical frequentist approach, the first problem is tackled by consecutive fitting of several models with different numbers of mixture components and choosing the best one using some criterion, e.g., Akaike's information criterion (Akaike, 1974). To avoid the second problem, homoscedastic normal mixtures, i.e. with $\sigma_j^2 = \sigma^2$ for all j are used leading to a bounded likelihood.

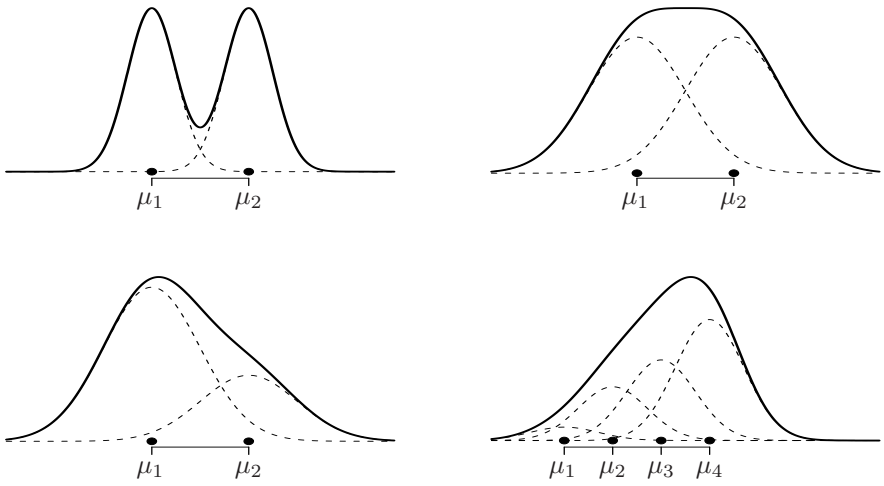


Figure 6.1: Several densities expressed as two- or four-component homoscedastic normal mixtures.

Bayesian methodology, on the other hand, offers a unified framework to estimate both the number of mixture components K and heteroscedastic normal mixtures in the same way as any other unknown parameters, i.e. using proper posterior summaries. A breakthrough in Bayesian analysis of models with a parameter space of varying dimension is the introduction of the *reversible jump* Markov chain Monte Carlo (RJMCMC) algorithm by Green (1995) which allows to explore a joint posterior distribution of the whole parameter vector $\boldsymbol{\theta}$ from model (6.3), including the number of mixture components K. Explicit application of the RJMCMC algorithm to normal mixtures is then described by Richardson and Green (1997).

The fact that the likelihood is unbounded for heteroscedastic normal mixtures leads to an improper posterior distribution in the Bayesian setting when a fully non-informative prior distribution is used for the variances of the mixture components (mixture variances), i.e. when $p(\sigma_1^2, \ldots, \sigma_K^2) \propto \prod_j \sigma_j^{-2}$. However, the problem is solved by using a slightly informative prior distribution for the mixture variances. For instance, replacing $\prod_j \sigma_j^{-2}$ by a product of inverse gamma distributions with parameters h_1 and h_2 where $h_1 = h_2 = 0.001$ or $h_1 = 1,\ h_2 = 0.005$, the classical vague priors, is already sufficient to prevent that the mixture variances will tend to zero causing an infinite likelihood.

We use a classical normal mixture model (6.3) for the density of the error distribution in the cluster-specific AFT model in Chapter 8. To avoid difficulties with the maximum-likelihood estimation outlined above and for other reasons (see Sections 4.1 and 4.2) only Bayesian methodology will be considered here. In Chapter 8 we also discuss the RJMCMC algorithm and the issue of the prior distribution for mixture variances in more detail.

6.2 Penalized B-splines

6.2.1 Introduction to B-splines

Different types of smoothing are routinely used in various places of modern statistics to express an unknown (smooth) function. Most often, either regression surfaces or densities are smoothed; see, e.g., Fahrmeir and Tutz (2001, Chapter 5) and Hastie, Tibshirani, and Friedman (2001) for an overview.

In this thesis, we concentrate on smoothing based on splines. For simplicity, we consider the univariate case first. The unknown function $g(y)$ (density in our case) is expressed as a linear combination (mixture) of suitable basis

spline functions $B_1(y), \ldots, B_K(y)$, i.e.

$$g(y) = g(y \,|\, \boldsymbol{\theta}) = \sum_{j=1}^{K} w_j B_j(y), \tag{6.5}$$

where $\boldsymbol{\theta} = \boldsymbol{w} = (w_1, \ldots, w_K)'$. Expression (6.5) is similar to (6.3) introduced in the previous section. Note however that in contrast to normal densities in (6.3), the basis spline functions $B_j(y)$, $j = 1, \ldots, K$ are always fully specified, including their location and scale, and the number of basis splines, K, is always fixed beforehand. The only quantities that have to be estimated are the spline coefficients (mixture weights) $\boldsymbol{w}$.

So called B-splines (de Boor, 1978; Dierckx, 1993) form, for their numerical stability and simplicity, a suitable system of basis spline functions. Their use in statistics was promoted especially by Eilers and Marx (1996). The B-spline is a piecewise polynomial function. To fully specify the B-spline basis,

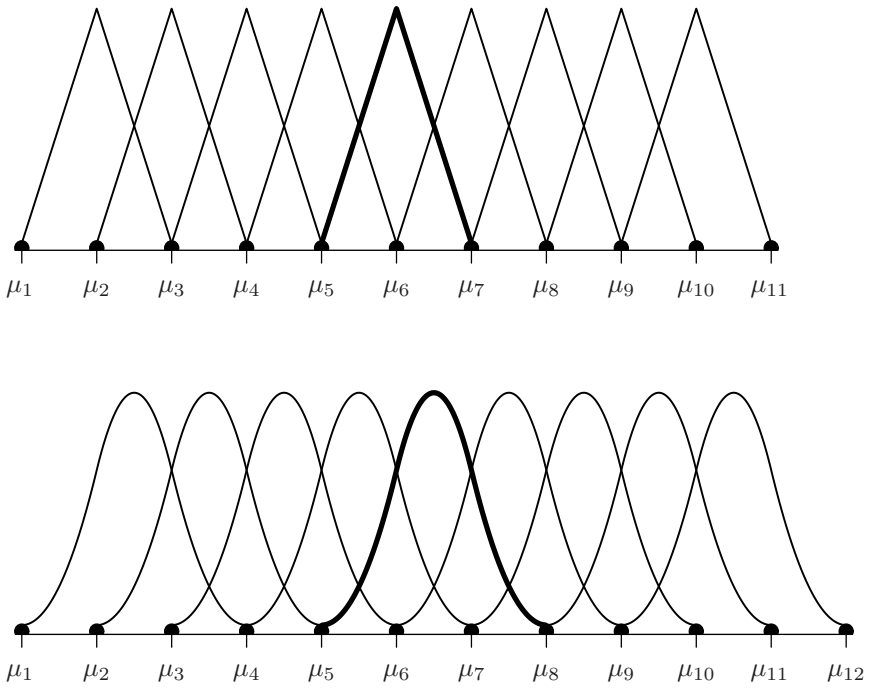


Figure 6.2: Basis B-splines of degree $d = 1$ (upper panel) and degree $d = 2$ (lower panel).

$B_1(y), \ldots, B_K(y)$, we have to determine

1. Degree d of the polynomial pieces;

2. A set of values (knots) $\mu_1 \leq \cdots \leq \mu_{d+1} < \cdots < \mu_{K+1} \leq \cdots \leq \mu_{K+d+1}$ such that the interval (μ_1, μ_{K+d+1}) covers the domain of the function $g(y)$ we wish to express using the B-splines.

Given that, the value of each basis B-spline can easily be computed at an arbitrary point $y \in \mathbb{R}$ (see de Boor, 1978). Figure 6.2 shows a basis of linear ($d = 1$) and quadratic ($d = 2$) B-splines with $K = 9$. It can be found that the jth basis B-spline of degree d

1. Consists of $d + 1$ polynomial pieces;

2. Is only positive on the interval (μ_j, μ_{j+d+1});

3. Has continuous derivatives up to order $d - 1$;

4. Except on boundaries it overlaps with $2d$ polynomial pieces of its neighbors.

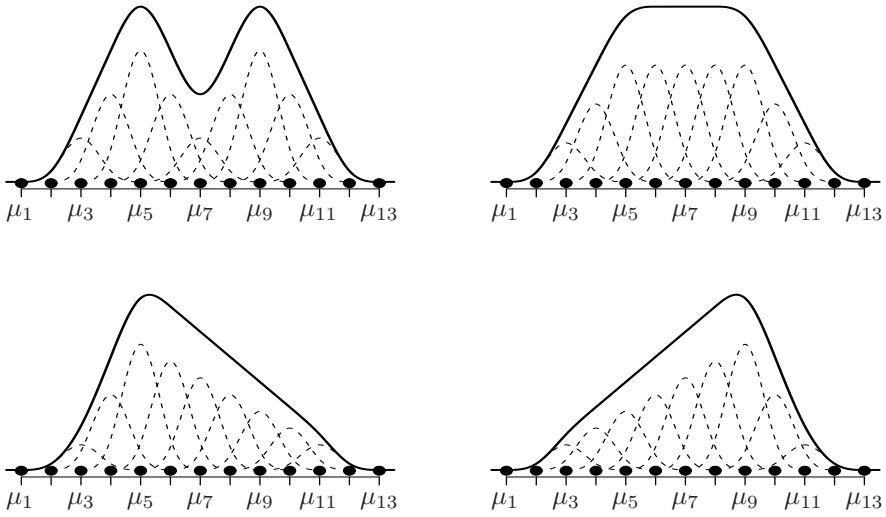


Figure 6.3: Several functions expressed as linear combinations of cubic B-splines with $K = 9$ and equidistant set of knots.

Furthermore, for all $y \in (\mu_1, \mu_{K+d+1})$ the basis B-splines sum up to one, i.e. $\sum_{j=1}^{K} B_j(y) = 1$. Finally, Dierckx (1993) gives simple recursive formulas to compute derivatives or integrals of the function $g(y)$ expressed by (6.5). Figure 6.3 illustrates that B-spline mixture can result in functions of various shapes.

6.2.2 Penalized smoothing

Choosing the optimal number and position of knots is generally a complex task in the area of spline smoothing. Too many knots leads to overfitting the data; too few knots leads to underfitting and inaccuracy. O'Sullivan (1986, 1988) proposed to take a relatively *large number* of knots and to restrict the flexibility of the fitted curve by putting a penalty on the second derivative.

In the context of B-splines, Eilers and Marx (1996) suggested

1. To use a large number of *equidistant* knots covering the domain of the function $g(y)$ one wishes to smooth;

2. To estimate the spline coefficients using the method of *penalized maximum-likelihood*. Further, they propose to base the penalty on squared finite higher-order differences between adjacent spline coefficients w_j.

They call their method as *penalized B-spline*, or *P-spline* smoothing. Eilers and Marx (1996) use the P-splines primarily to smooth regression surfaces although they propose also a methodology, based on the Poisson generalized linear model (GLM), for smooth estimation of the density with the i.i.d. data. We sketch this method in Section 6.2.4.

The strategy of several further developments (Chapters 7, 9, 10) in this thesis is based on the ideas of Eilers and Marx (1996), modified and adapted to regression modelling with censored data. Namely,

1. For reasons stated in Section 6.3 we replace the basis B-splines by normal densities with a common variance;

2. We base the penalty term on squared finite higher-order differences between appropriate *transformations* of the adjacent spline coefficients w_j, see Section 7.2.2 for a motivation;

3. More complex models in Chapters 9 and 10 will be estimated using the Bayesian methodology using the prior distributions inspired by the penalty term used in the penalized maximum-likelihood applications;

4. We will not use the Poisson GLM-based density estimation, see Section 6.2.4 for the reasons why.

In agreement with Eilers and Marx (1996) we use a set of equidistant knots in all penalized-based developments.

6.2.3 B-splines in the survival analysis

General splines have been suggested at several places in the survival literature to model flexibly either the (log-)density/(log-)hazard function or the effect of covariates replacing a linear predictor by a spline function. See the discussion section of Abrahamowicz, Ciampi, and Ramsay (1992), the introductory section of Kooperberg, Stone, and Truong (1995) or Chapter 5 of Therneau and Grambsch (2000) for an overview.

More specifically, B-splines have been used by Rosenberg (1995) who uses their cubic variant to express the baseline hazard function in the Cox's PH model. He choses the optimal number of knots according to Akaike's information criterion (Akaike, 1974) while placing the knots to the quantiles of uncensored observations. An approach based on the penalized maximum-likelihood is given by Joly, Commenges, and Letenneur (1998) who use monotone splines (close relatives of the B-splines introduced by Ramsay, 1988) to model the baseline hazard function, in the Cox's PH model as well. Tutz and Binder (2004) and Kauermann (2005b) use B-splines to extend the basic Cox's PH model by allowing for time-varying regression parameters.

Recently, Lambert and Eilers (2005) use a Bayesian version of penalized B-splines to model both the baseline hazard and the effect of covariates in the Cox's PH model in an actuarial way. To our best knowledge, we are not aware of any approach where the B-splines would be used to model the density of the survival times.

6.2.4 B-splines as models for densities

The function $g(y \,|\, \boldsymbol{\theta})$ expressed by (6.5) can serve as a model for the density of a continuous distribution with domain $(\mu_1,\, \mu_{K+d+1})$ provided

$$g(y \,|\, \boldsymbol{\theta}) \geq 0 \quad \text{for all } y \in (\mu_1,\, \mu_{K+d+1}), \tag{6.6}$$

$$\int_{\mu_1}^{\mu_{K+d+1}} g(y \,|\, \boldsymbol{\theta})\, dy = 1. \tag{6.7}$$

Condition (6.6) is satisfied if we require all the spline coefficients to be positive, i.e.

$$w_j > 0, \qquad j = 1, \ldots, K. \tag{6.8}$$

Constraint (6.7) can easily be avoided when we change the expression (6.5) for $g(y \,|\, \boldsymbol{\theta})$ into

$$g(y \,|\, \boldsymbol{\theta}) = Q^{-1} \sum_{j=1}^{K} w_j \, B_j(y), \tag{6.9}$$

$$Q = \int_{\mu_1}^{\mu_{K+d+1}} \Big\{ \sum_{j=1}^{K} w_j B_j(y) \Big\} \, dy$$

The constant Q can easily be computed using the formulas given by Dierckx (1993, Section 1.3). For example, in the case of coincident boundary knots (i.e. $\mu_1 = \cdots \mu_{d+1}$ and $\mu_{K+1} = \cdots = \mu_{K+d+1}$) the constant Q equals

$$Q = \frac{1}{d+1} \sum_{j=1}^{K} w_j \, (\mu_{j+d+1} - \mu_j).$$

We show in Section 6.3.2 how to avoid also the inequality constraints (6.8).

A somewhat different approach to estimate a density function using B-splines has been suggested in Eilers and Marx (1996, Section 8), namely by smoothing a histogram. They divide the range of the data into a large number K of bins, each of length h, and let the midpoints of the bins to define the knots $\mu_1, \ldots, \mu_K$. The raw continuous data, $y_1, \ldots, y_n$ are changed into counts $n_1, \ldots, n_K$ such that n_j, $j = 1, \ldots, K$ equals the number of raw observations y_i, $i = 1, \ldots, n$ with $\mu_j - h/2 \leq y_i < \mu_j + h/2$. The counts $n_1, \ldots, n_K$ constitute a histogram. They assume that each of these counts follows a Poisson distribution with expectation $\mathrm{E}(n_1), \ldots, \mathrm{E}(n_K)$, respectively. A smoothed histogram is obtained by expressing the Poisson log-expectations as the B-spline, namely

$$\log\big\{\mathrm{E}(n_j)\big\} = \sum_{k=1}^{K} w_k \, B_k(\mu_j), \qquad j = 1, \ldots, K.$$

The corresponding smooth density of the original continuous data is then given by

$$g(y \,|\, \boldsymbol{\theta}) = Q^{-1} \, \exp\Big\{ \sum_{k=1}^{K} w_k \, B_k(y) \Big\},$$

where Q is an appropriate proportionality constant. Eilers and Marx (1996) argue that the use of penalized maximum-likelihood estimation provides stable and useful results and does not lead to any pathological results resulting from discretization of the data.

For our developments in the context of the AFT model, we believe that the approach with the density directly expressed as a mixture of B-splines is more advantageous since it leads to a simpler likelihood evaluation. Remember that with censored observations the likelihood involves evaluation of integrals of the assumed density (see Section 4.1 and 4.2). With the density (6.9) these integrals are simply mixtures of integrated basis B-splines whose computation only involves integration of polynomials. Nevertheless, usage of the smoothed histogram approach in the censored data regression context is presented by Lambert and Eilers (2005).

6.2.5 B-splines for multivariate smoothing

The concept of B-splines can be extended to the multivariate setting, to smooth (estimate) a function $g(\boldsymbol{y})$ of several variables. For example the bivariate case is achieved by replacing the formula (6.5) by

$$g(\boldsymbol{y}) = g(y_1, y_2) = g(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} w_{j_1, j_2} \, B_{1, j_1}(y_1) \, B_{2, j_2}(y_2),$$

where B_{1, j_1}, $j_1 = 1, \ldots, K_1$ is a set of basis B-splines of degree d defined by knots $\mu_{1,1}, \ldots, \mu_{1, K_1+d+1}$, B_{2, j_2}, $j_2 = 1, \ldots, K_2$ a set of basis B-splines of degree d defined by a generally different set of knots $\mu_{2,1}, \ldots, \mu_{2, K_2+d+1}$, and $\boldsymbol{\theta} = (w_{1,1}, \ldots, w_{K_1,K_2})'$. Namely, $g(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ is expressed as a *Kronecker product* of univariate B-splines and this idea can be extended also to higher dimensions.

6.3 Penalized normal mixture

6.3.1 From B-spline to normal density

Using the B-spline expression (6.5) to model a survival density has one drawback. Namely, the support of the resulting density $g(y \,|\, \boldsymbol{\theta})$ is always bounded and equal to the interval (μ_1, μ_{K+d+1}). However, most continuous survival distributions are thought of as having a support of $(0, \infty)$ on the time scale and the real line on the log-scale. While in practice this might not constitute

any difficulty, in theory it might be more comfortable to approximate a density having an infinite support. Remember also that we aim to approximate densities of either the error distribution in the AFT model or the distribution of the random effects in the same model. This implies that it might be quite difficult in some settings to find a proper range of the density for the error terms and/or random effects as both distributions are seen from the data only indirectly. However, one can easily find that the basis B-spline of degree d is very close to the density of the standard normal distribution in the sense of the following proposition.

Proposition 6.1. *Let $B^d(y)$ be a basis B-spline of degree d defined on the grid of $d+2$ equidistant knots*

$$\mu_1^d = -\delta\,\frac{d+1}{2}, \quad \ldots, \quad \mu_{d+2}^d = \delta\,\frac{d+1}{2},$$

with $\delta = \mu_{j+1}^d - \mu_j^d$, $j = 1, \ldots, d+1$ equal to $\sqrt{12/(d+1)}$. Let

$$B_{st}^d(y) = \sqrt{\frac{d+1}{12}}\,B^d(y), \quad y \in \mathbb{R}$$

be a standardized basis B-spline of degree d. Then

$$\lim_{d\to\infty} B_{st}^d(y) = \varphi(y) \quad \textit{uniformly for all } y \in \mathbb{R},$$

where φ denotes a density of a standard normal distribution.

Proof. We give only main ideas of the proof. All technical details can be found in Unser, Aldroubi, and Eden (1992).

Firstly, an arbitrary basis B-spline of degree d is proportional to the density of a sum of $d+1$ independent uniformly distributed random variables. Properly standardized basis B-spline, $B_{st}^d(y)$, is then equal to a density of a zero mean, unit variance random variable given as a sum of $d+1$ independent uniformly distributed random variables. The proposition is then achieved using the central limit theorem (see, e.g., Billingsley, 1995, Section 27). □

The property outlined in Proposition 6.1 is illustrated in Figure 6.4. Moreover, the convergence is rather fast. Indeed, the standardized cubic basis B-spline is already quite close to the standard normal density.

This reasoning led us to replace the basis B-splines in the expression (6.5) by normal densities whose means are equal to the knots and whose variance is equal to a common value σ_0^2. In accordance with the idea of penalized B-splines (see Section 6.2.2), we use a larger number of equidistant knots

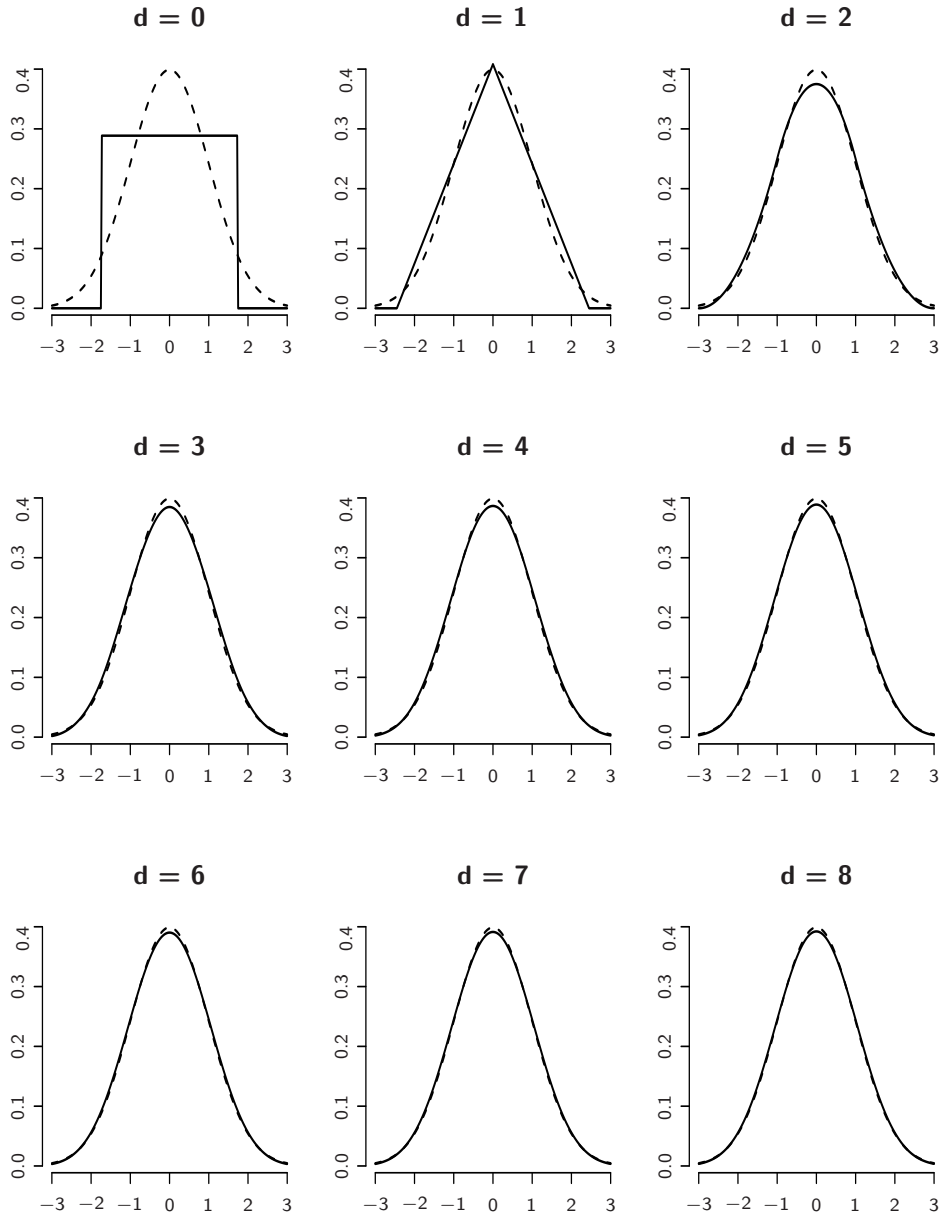


Figure 6.4: Standardized basis B-splines of degree 0 to 8 (solid line) compared to a standard normal density (dashed line).

chosen beforehand. Additionally, as explained in Section 6.3.3, we always use an odd number of knots symmetric around the middle knot. For this reason, the number of mixture components will be indicated by $2K+1$ and the knots – means denoted by $\mu_{-K}, \ldots, \mu_0, \ldots, \mu_K$. Namely, the unknown function $g(y)$ (density) is approximated by

$$g(y) = g(y \,|\, \boldsymbol{\theta}) = \sum_{j=-K}^{K} w_j \, \varphi(y \,|\, \mu_j, \, \sigma_0^2), \tag{6.10}$$

where $\boldsymbol{\theta} = (w_{-K}, \ldots, w_K)'$. The basis standard deviation, σ_0, is chosen beforehand as well as the knots. For its choice we adopted the value $2\delta/3$, where $\delta = \mu_{j+1} - \mu_j$, $j = -K, \ldots, K-1$ is the distance between the two consecutive knots – means. The motivation for this choice is provided by an attempt to keep a correspondence with the cubic B-splines. Remember, the basis cubic B-spline covers an interval of length 4δ. The same is nearly true for the normal density with the variance $(2\delta/3)^2$ if we admit that the $\mathcal{N}(\mu, \, \sigma^2)$ density is practically zero outside the interval $(\mu - 3\sigma, \, \mu + 3\sigma)$. In this context, we will call (6.10) *penalized normal mixture*.

6.3.2 Transformation of mixture weights

To ensure that the function $g(y \,|\, \boldsymbol{\theta})$ given by (6.10) is a density of some continuous distribution, we have to impose constraints analogous to (6.8) and (6.9) upon the mixture weights $\boldsymbol{w} = (w_{-K}, \ldots, w_K)'$. Namely, they have to satisfy

$$w_j > 0, \qquad j = -K, \ldots, K, \tag{6.11}$$

$$\sum_{j=-K}^{K} w_j = 1. \tag{6.12}$$

To avoid constrained estimation, one can use an alternative parametrization based on transformed mixture weights $\boldsymbol{a} = (a_{-K}, \ldots, a_K)'$

$$a_j(\boldsymbol{w}) = \log\Big(\frac{w_j}{w_0}\Big), \quad j = -K, \ldots, K, \tag{6.13}$$

Inversely, the original weights $\boldsymbol{w}$ are computed from the transformed weights $\boldsymbol{a}$ by

$$w_j(\boldsymbol{a}) = \frac{\exp(a_j)}{\sum_{k=-K}^{K} \exp(a_k)}, \quad j = -K, \ldots, K. \tag{6.14}$$

Instead of estimating the constrained weights $\boldsymbol{w}$, the vector $\boldsymbol{a}_{-0}$ of unconstrained transformed weights, except a_0 which is fixed to zero, is estimated.

Note that the weights $\boldsymbol{w}(\boldsymbol{a})$ expressed by (6.14) automatically satisfy both (6.11) and (6.12). Further, an arbitrary mixture component can be chosen to be the reference one having a corresponding a coefficient fixed to zero without any impact on the results. However, for notational convenience, without loss of generality, we will assume that $a_0 = 0$.

6.3.3 Penalized normal mixture for distributions with an arbitrary location and scale

Let Y be a random variable with a density $g(y)$ with

$$\mathrm{E}(Y) = \alpha, \qquad \mathrm{var}(Y) = \tau^2.$$

To be able to use the same grid of knots – means $\mu_{-K}, \ldots, \mu_K$ for distributions with an arbitrary location α and scale τ we incorporate these two parameters in the expression (6.10) for the unknown density $g(y)$, i.e., the density $g(y)$ will be approximated by

$$g(y) = g(y \,|\, \boldsymbol{\theta}) = \tau^{-1} \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \varphi\Big(\frac{y-\alpha}{\tau} \,\Big|\, \mu_j,\, \sigma_0^2\Big), \tag{6.15}$$

where $\boldsymbol{\theta} = (a_{-K}, \ldots, a_K,\, \alpha,\, \tau)'$. In other words, the density of the standardized random variable $Y^* = \tau^{-1}(Y - \alpha)$ is approximated by

$$g^*(y^* \,|\, \boldsymbol{\theta}^*) = \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \varphi\big(y^* \,\big|\, \mu_j,\, \sigma_0^2\big),$$

where $\boldsymbol{\theta}^* = (a_{-K}, \ldots, a_K)'$. The intercept α and the scale τ will be estimated simultaneously with the transformed mixture weights $\boldsymbol{a}$.

With expression (6.15), the knots $\mu_{-K}, \ldots, \mu_K$ have to cover a high probability region of the zero-mean, unit-variance distribution. In most practical situations, the choice with μ_{-K} equal to a value between -6 and -4.5 and μ_K equal to a value between 4.5 and 6 provides the range of the knots broad enough. Furthermore, a distance δ of 0.3 between two consecutive knots is small enough to approximate most smooth densities. As an illustration, we computed the L_2-distance between the standard normal density and its best approximation using a penalized mixture (6.15) with $\mu_{-K} = -6$, $\mu_K = 6$, different choices of $\delta = \mu_{j+1} - \mu_j$, and $\sigma_0 = 2\delta/3$. This distance is equal to 0.00570 for $\delta = 1$ ($K = 6$), and drops to 0.00104 for $\delta = 0.75$ ($K = 8$). When plotted, the penalized mixture (6.15) is indistinguishable from the normal density at $\delta = 0.75$. Further, for δ equal to 0.5 ($K = 12$), 0.4 ($K = 15$), 0.3

($K = 20$), 0.2 ($K = 30$), and 0.1 ($K = 60$) we obtain distances of 0.00031, 0.00022, 0.00017, 0.00014, and 0.00012, respectively. Clearly, for $\delta = 0.3$ the penalized mixture and the normal density are quite close.

6.3.4 Multivariate smoothing

In Section 6.2.5 we discussed how the Kronecker product of univariate B-splines can be used to model unknown multivariate functions. The same idea can be used also with the penalized normal mixture. In this thesis, we use the multivariate penalized normal mixture only in the bivariate setting which will be discussed now. Extensions to higher dimensions are obvious, only with more complex notation.

Firstly, we note that the bivariate basis formed of the Kronecker product of univariate normal densities is actually the basis formed of bivariate normal densities with diagonal covariance matrices. Indeed, for arbitrary $y_1 \in \mathbb{R}$ and $y_2 \in \mathbb{R}$

$$\varphi(y_1 \,|\, \mu_1,\, \sigma_1^2)\, \varphi(y_2 \,|\, \mu_2,\, \sigma_2^2) = \varphi_2(y_1,\, y_2 \,|\, \boldsymbol{\mu},\, \Sigma),$$

where $\varphi_2(\cdot \,|\, \boldsymbol{\mu},\, \Sigma)$ is a density of $\mathcal{N}_2(\boldsymbol{\mu},\, \Sigma)$ with $\boldsymbol{\mu} = (\mu_1,\, \mu_2)'$ and $\Sigma = \mathrm{diag}(\sigma_1^2,\, \sigma_2^2)$.

Analogously to the univariate formula (6.10), the unknown bivariate density $g(y_1,\, y_2) = g(\boldsymbol{y})$ is expressed by

$$g(\boldsymbol{y}) = g(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w_{j_1,j_2}\, \varphi(\boldsymbol{y} \,|\, \boldsymbol{\mu}_{(j_1,j_2)},\, \Sigma), \tag{6.16}$$

where $\boldsymbol{\mu}_{(j_1,j_2)} = (\mu_{1,j_1},\, \mu_{2,j_2})'$, $j_1 = -K_1, \ldots, K_1$, $j_2 = -K_2, \ldots, K_2$ is a fixed fine grid of knots, $\Sigma = \mathrm{diag}(\sigma_1^2,\, \sigma_2^2)$ is a fixed basis covariance matrix (the same for all mixture components) and $\mathbb{W} = (w_{j_1,j_2})$, $j_1 = -K_1, \ldots, K_1$, $j_2 = -K_2, \ldots, K_2$ a matrix of unknown mixture weights satisfying

$$w_{j_1,j_2} > 0, \qquad j_1 = -K_1, \ldots, K_1,\ j_2 = -K_2, \ldots, K_2 \tag{6.17}$$

$$\sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w_{j_1,j_2} = 1. \tag{6.18}$$

The vector $\boldsymbol{\theta}$ of unknown parameters contains the elements of the matrix $\mathbb{W}$. Similarly to Section 6.3.2, the constraints (6.17) and (6.18) are avoided by the reparametrization of the weight matrix $\mathbb{W}$ into the matrix $\mathbb{A} = (a_{j_1,j_2})$,

$j_1 = -K_1, \ldots, K_1$, $j_2 = -K_2, \ldots, K_2$ of transformed weights by

$$a_{j_1,j_2}(\mathbb{W}) = \log\Big(\frac{w_{j_1,j_2}}{w_{0,0}}\Big), \qquad j_1 = -K_1, \ldots, K_1,$$
$$w_{j_1,j_2}(\mathbb{A}) = \frac{\exp(a_{j_1,j_2})}{\sum\limits_{k_1=-K_1}^{K_1} \sum\limits_{k_2=-K_2}^{K_2} \exp(a_{k_1,k_2})}, \quad j_2 = -K_2, \ldots, K_2. \tag{6.19}$$

For notational convenience and without loss of generality, the mixture component $(0, 0)$ is chosen to be the baseline with $a_{0,0} = 0$.

Moments of the bivariate penalized normal mixture

It is useful to stress that although all bivariate normal components in (6.16) are uncorrelated the covariance matrix of the random vector $(Y_1,\, Y_2)'$ with the density $g(\boldsymbol{y}) = g(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ defined by (6.16) is, except for a special combination of mixture weights, not diagonal. Namely,

$$\mathrm{E}(Y_1) = \sum_{j_1=-K_1}^{K_1} w_{j_1+}\mu_{1,j_1}, \qquad \mathrm{E}(Y_2) = \sum_{j_2=-K_2}^{K_2} w_{+j_2}\mu_{2,j_2},$$

$$\mathrm{var}(Y_1) = \sigma_1^2 \; + \sum_{j_1=-K_1}^{K_1} w_{j_1+}\Big\{\mu_{1,j_1} - \mathrm{E}(Y_1)\Big\}^2,$$
$$\mathrm{var}(Y_2) = \sigma_2^2 \; + \sum_{j_2=-K_2}^{K_2} w_{+j_2}\Big\{\mu_{2,j_2} - \mathrm{E}(Y_2)\Big\}^2,$$
$$\mathrm{cov}(Y_1,\, Y_2) = \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w_{j_1,j_2}\Big\{\mu_{1,j_1} - \mathrm{E}(Y_1)\Big\}\Big\{\mu_{2,j_2} - \mathrm{E}(Y_2)\Big\},$$

where subscript $+$ means summation over the range of the corresponding index.

Bivariate penalized normal mixture for distributions with an arbitrary location and scale

Analogously to Section 6.3.3 we introduce here an extra intercept parameter vector $\boldsymbol{\alpha} = (\alpha_1,\, \alpha_2)'$ and an extra scale parameter vector $\boldsymbol{\tau} = (\tau_1,\, \tau_2)'$ to

allow for modelling the bivariate densities of a random vector $\boldsymbol{Y} = (Y_1, Y_2)'$ with a general location and scales, i.e. with

$$\begin{aligned} \mathrm{E}(Y_1) &= \alpha_1, \qquad \mathrm{var}(Y_1) = \tau_1^2, \\ \mathrm{E}(Y_2) &= \alpha_2, \qquad \mathrm{var}(Y_2) = \tau_2^2. \end{aligned}$$

As before, the same values of the extreme knots $\mu_{1,-K_1}$, μ_{1,K_1}, $\mu_{2,-K_2}$, μ_{1,K_2} and the basis standard deviations σ_1, σ_2 can be used for distributions with different location and scale.

Namely the bivariate density $g(\boldsymbol{y})$ of a general distribution will be approximated by

$$\begin{aligned} &g(\boldsymbol{y}) = g(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \qquad (6.20) \\ &(\tau_1\tau_2)^{-1} \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w_{j_1,j_2}(\mathbb{A})\, \varphi_2\Big(\frac{y_1-\alpha_1}{\tau_1}, \frac{y_2-\alpha_2}{\tau_2} \,\Big|\, \boldsymbol{\mu}_{(j_1,j_2)}, \Sigma\Big), \end{aligned}$$

where $\boldsymbol{\theta} = (a_{-K_1,-K_2}, \ldots, a_{K_1,K_2}, \alpha_1, \alpha_2, \tau_1, \tau_2)'$. In other words, the density of the standardized random vector

$$\boldsymbol{Y}^* = \begin{pmatrix} Y_1^* \\ Y_2^* \end{pmatrix} = \begin{pmatrix} \tau_1^{-1} & 0 \\ 0 & \tau_2^{-1} \end{pmatrix} \begin{pmatrix} Y_1 - \alpha_1 \\ Y_2 - \alpha_2 \end{pmatrix}$$

is approximated by

$$g^*(\boldsymbol{y}^* \,|\, \boldsymbol{\theta}^*) = \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w_{j_1,j_2}(\mathbb{A})\, \varphi(\boldsymbol{y}^* \,|\, \boldsymbol{\mu}_{(j_1,j_2)}, \Sigma), \qquad (6.21)$$

where the vector $\boldsymbol{\theta}^*$ contains only the elements of the matrix $\mathbb{A}$ of transformed weights. The same guidelines as in the univariate case (Section 6.3.3) will be applied for the choice of the grid points and the basis standard deviations, i.e. both $\mu_{1,-K_1}, \ldots, \mu_{1,K_1}$ and $\mu_{2,-K_2}, \ldots, \mu_{2,K_2}$ being the univariate grids of equidistant knots with the distance between the two knots equal to $\delta \approx 0.3$, with the minimal knot lying between -6 and -4.5, the maximal knot lying between 4.5 and 6 and basis standard deviations equal $\sigma_1 = \sigma_2 = 2\delta/3$.

6.4 Classical versus penalized normal mixture

We finalize this chapter by an explicit comparison of the classical normal mixture and penalized normal mixture.

- With the penalized normal mixture, invariably a relatively large but fixed number of mixture components is needed and the smoothness of the resulting smoothed distribution is optimized via a penalty term. On the other hand, with the classical mixture, often a small number of mixture components is sufficient but, the number of components have to be estimated which might cause some difficulties as outlined in Section 6.1.2;

- The fine grid of fixed knots in the penalized mixture approach prevents inaccuracy in the estimate of the unknown density, while the penalization inhibits overfitting. In contrast, in the case of a classical mixture, the means and the standard deviations of the mixture components must be estimated;

- In order to use a standard grid of knots we have included explicitely the intercept and scale parameters in the model specification when using the penalized approach. This is not desirable with the classical mixture approach as both the overall intercept and the overall scale are implicitely defined by the mixture components means and standard deviations;

- Extension of the univariate smoothing into the multivariate smoothing is conceptually simple with the penalized approach as was shown in Section 6.3.4 using the Kronecker product of the basis functions. In higher dimensions, there are only some computational difficulties arising from the fact that the number of unknown parameters increases exponentially with the dimension.

  Extension of the classical mixture approach into higher dimensions is relatively easy with a fixed number of mixture components however is not straightforward when the number of mixture components have to be estimated simultaneously with the remaining parameters. Even with the Bayesian approach and the reversible jump MCMC algorithm mentioned in Section 6.1.2 the multivariate extensions are still an area of active research, see Dellaportas and Papageorgiou (2006) for recent developments.

Chapter 7

Maximum Likelihood Penalized AFT Model

In this chapter, we present the AFT model for the case of independent observations. The error distribution of the model will be based on the penalized normal mixture (Section 6.3) and penalized maximum-likelihood estimation. The basic version of this approach is given by Komárek, Lesaffre, and Hilton (2005) and an extension allowing also modelling the dependence of the scale parameter on the covariates can be found in Lesaffre, Komárek, and Declerck (2005).

In Section 7.1, we describe the model in detail. In Section 7.2, we show how the model parameters are estimated using the penalized maximum-likelihood method. Section 7.3 describes the inferential procedures. In Section 7.4, computation of predictive survival or hazard functions and predictive densities is discussed. Section 7.5 gives the results of a simulation study that evaluates the performance of the method. The proposed method is applied to the analysis of the WIHS data in Section 7.6 and to the analysis of the Signal Tandmobiel® data in Section 7.7. We finalize the chapter by a discussion in Section 7.8.

7.1 Model

Let T_i, $i = 1, \ldots, N$ be independent event times observed as intervals $\lfloor t_i^L, t_i^U \rfloor$ and δ_i be the corresponding censoring indicator with the same convention as in Section 2.1. Let $y_i^L = \log(t_i^L)$ and $y_i^U = \log(t_i^U)$. Further, let $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,m})'$ be a vector of covariates associated with the ith subject. The

effect of covariates on the event time T_i will be specified using the basic AFT model introduced in Section 3.2, i.e.

$$\log(T_i) = \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{7.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ is a vector of unknown regression parameters and $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. error random variables with the density $g_\varepsilon(\varepsilon)$.

7.1.1 Model for the error density

The density $g_\varepsilon(\varepsilon)$ of the error term will be expressed using the penalized normal mixture (6.15), i.e.

$$g_\varepsilon(\varepsilon) = \tau^{-1} \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \varphi\Big(\frac{\varepsilon - \alpha}{\tau} \,\Big|\, \mu_j,\, \sigma_0^2\Big), \tag{7.2}$$

where $\mu_{-K}, \ldots, \mu_K$ is a set of fixed equidistant knots, σ_0 fixed basis standard deviation, α unknown intercept and τ unknown scale parameter. Finally, $\boldsymbol{w} = (w_{-K}, \ldots, w_K)'$ are unknown mixture weights and $\boldsymbol{a} = (a_{-K}, \ldots, a_K)'$ their transformations obtained using the relationship (6.13).

Let $\varepsilon_1^*, \ldots, \varepsilon_N^*$ be standardized error terms, i.e. having the density

$$g_\varepsilon^*(\varepsilon^*) = \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \varphi(\varepsilon^* \,|\, \mu_j,\, \sigma_0^2). \tag{7.3}$$

Keeping the intercept α and the scale τ identifiable requires that the first two moments of the density (7.3) be fixed, i.e.,

$$\mathrm{E}(\varepsilon_i^*) = \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \mu_j = 0, \qquad \mathrm{var}(\varepsilon_i^*) = \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, (\mu_j^2 + \sigma_0^2) = 1. \tag{7.4}$$

Due to the fact that $\sum_{j=-K}^{K} w_j(\boldsymbol{a})\sigma_0^2 = \sigma_0^2$, the variance constraint can be rewritten into the form $\sum_{j=-K}^{K} w_j(\boldsymbol{a})\mu_j^2 = 1 - \sigma_0^2$. It is then easily seen that the basis standard deviation σ_0 must be smaller than 1 to be able to satisfy this constraint. Finally, the two equality constraints (7.4) can be avoided if two coefficients, say, a_{-1} and a_1, are expressed as functions of the remaining non-baseline coefficients, denoted together as a vector $\boldsymbol{d} = (a_{-K}, \ldots, a_{-2}, a_2, \ldots, a_K)'$:

$$a_k(\boldsymbol{d}) = \log\Big\{\omega_{0,k} + \sum_{j \notin \{-1,0,1\}} \omega_{j,k} \exp(a_j)\Big\}, \qquad k = -1,\ 1, \tag{7.5}$$

with

$$\begin{aligned}\omega_{j,-1} &= -\frac{\mu_j - \mu_1}{\mu_{-1} - \mu_1} \cdot \frac{1 - \sigma_0^2 + \mu_1\mu_j}{1 - \sigma_0^2 + \mu_1\mu_{-1}}, \\ \omega_{j,1} &= -\omega_{j,-1} \cdot \frac{\mu_{-1}}{\mu_1} - \frac{\mu_j}{\mu_1}, \qquad\qquad j = -K, \ldots, -2,\, 0,\, 2, \ldots, K.\end{aligned}$$

7.1.2 Scale regression

In most regression models, it is conventionally assumed that the covariates influence the mean, but it is presumed that it will *not* influence the scale parameter. With hindsight, this is simply one model choice and in many cases it may be untenable. Recently, there is interest in joint mean-covariance models in the context of longitudinal studies (Pourahmadi, 1999; Pan and MacKenzie, 2003). Our AFT model (7.1) with the error density (7.2) can be generalized in the same direction yielding *the mean-scale penalized AFT model*. With this generalization, we allow the scale parameter τ to vary across individuals. Moreover, for the ith individual, the scale parameter τ_i will depend on a vector of covariates, say $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,m_s})'$, as

$$\tau_i \equiv \tau(\boldsymbol{z}_i) = \exp(\boldsymbol{\gamma}'\boldsymbol{z}_i), \tag{7.6}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{m_s})'$ is a vector of unknown parameters. Note, that the covariate vector $\boldsymbol{z}_i$ usually contains the intercept term, i.e. $z_{i,1} = 1$ for all i. In that case, the original AFT model (7.1) with the error density (7.2) and the common scale parameter τ can be written as the mean-scale AFT model with $\boldsymbol{z}_i = 1$ for all i and $\tau = \exp(\gamma_1)$.

All parameters in the model (transformed mixture coefficients $\boldsymbol{d}$; regression parameters vector $\boldsymbol{\beta}$; intercept α; and log-scale $\log(\tau)$ or scale-regression parameters vector $\boldsymbol{\gamma}$) are estimated by means of a penalized maximum-likelihood method. In the next section, we construct the penalized log-likelihood function which consists of an ordinary log-likelihood and a difference penalty for the transformed spline coefficients. The penalized log-likelihood is subsequently maximized to obtain the estimates, see Appendix A for practical aspects of the optimization of the penalized log-likelihood.

7.2 Penalized maximum-likelihood

7.2.1 Penalized log-likelihood

Let $\boldsymbol{\theta}$ be the vector of all unknown parameters to be estimated, i.e., $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}', a_{-K}, \ldots, a_{-2}, a_2, \ldots, a_K)'$. Let $\ell_i(\boldsymbol{\theta}) = \log\{L_i(\boldsymbol{\theta})\}$, $i = 1, \ldots, N$

denote the ordinary log-likelihood contribution of the ith observation based on model (7.1) with error density (7.2), i.e., using the results of Section 4.1.1 and the convention (4.3),

$$\begin{aligned} L_i(\boldsymbol{\theta}) &= \tau_i^{-1} \oint_{t_i^L}^{t_i^U} t^{-1} g_\varepsilon^* \Big\{ \frac{\log(t) - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i}{\tau_i} \Big\}\, dt \\ &\propto \tau_i^{-1} \oint_{y_i^L}^{y_i^U} g_\varepsilon^* \Big(\frac{y - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i}{\tau_i} \Big)\, dy \\ &= \tau_i^{-1} \sum_{j=-K}^{K} w_j(\boldsymbol{a}) \oint_{y_i^L}^{y_i^U} \varphi\Big(\frac{y - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i}{\tau_i} \,\Big|\, \mu_j,\, \sigma_0^2 \Big)\, dy \end{aligned}$$

The proportionality constant is equal to $t_i^L = t_i^U$ for exactly observed event times ($\delta_i = 1$) and equal to 1 for all remaining observations ($\delta_i = 0,\, 2,\, 3$). For the purpose of maximum-likelihood based estimation, this constant can be ignored so for notational convenience we will assume that this constant equals one. Finally, let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \ell_i(\boldsymbol{\theta})$ be the ordinary log-likelihood of the whole data set.

As usual with censored data, the likelihood evaluation involves integration. With our model, however, this does not cause any considerable difficulties irrespective of the type of censoring (left-, right-, interval-). Indeed, all integrals involved in the computation of the likelihood are normal cumulative distribution functions, which can be easily and efficiently evaluated.

To construct the penalized log-likelihood function $\ell_P(\boldsymbol{\theta}; \lambda)$, we subtract a penalty term $q(\boldsymbol{a}; \lambda)$ based on the transformed mixture coefficients $\boldsymbol{a}$ from $\ell(\boldsymbol{\theta})$, i.e.,

$$\ell_P(\boldsymbol{\theta}; \lambda) = \ell(\boldsymbol{\theta}) - q(\boldsymbol{a}; \lambda), \tag{7.7}$$

where λ is a fixed tuning parameter that controls the smoothness of the fitted error distribution and inhibits identifiability problems due to overparametrization. For a given (reasonable) λ, Eilers and Marx (1996) proposed to base the penalty on squared higher-order finite differences of the coefficients of adjacent B-splines, and they used second-order difference in their examples. We base our penalty on squared finite differences of order s of the *transformed* coefficients of adjacent mixture components:

$$\begin{aligned} q(\boldsymbol{a}; \lambda) &= \frac{\lambda}{2} \sum_{j=-K+m}^{K} \big\{ \Delta^s a_j \big\}^2 \qquad (7.8) \\ &= \frac{\lambda}{2}\, \boldsymbol{a}'\, \mathbb{P}_s' \mathbb{P}_s\, \boldsymbol{a}, \end{aligned}$$

where $\Delta^1 a_j = a_j - a_{j-1}$, $\Delta^s a_j = \Delta^{s-1} a_j - \Delta^{s-1} a_{j-1}$, $s = 1, \ldots,$ and $\mathbb{P}_s$ is a $(2K + 1 - s) \times (2K + 1)$ difference operator matrix. According to our experience, $s = 2$ or $s = 3$ is sufficient to obtain a smooth estimate of the density. However, in our context the choice $s = 3$ has another interesting justification, as explained in Section 7.2.2 and will be used in all applications presented in this thesis.

7.2.2 Remarks on the penalty function

There are two reasons why we penalize the transformed mixture coefficients $\boldsymbol{a}$ instead of the original coefficients $\boldsymbol{w}$ and why we prefer the penalty of order $s = 3$.

First, the penalty based on $\boldsymbol{a}$ distinguishes between areas of the density where there are few datapoints (i.e., where the coefficients $\boldsymbol{w}$ are close to zero) and areas where there are many datapoints (i.e., where the coefficients $\boldsymbol{w}$ are well above zero); the penalty based on $\boldsymbol{w}$ cannot distinguish between these areas. For example,

$$\begin{aligned} \text{for} \quad & \breve{\boldsymbol{w}} = (0.001, 0.002, 0.001, 0.996)', \\ & \tilde{\boldsymbol{w}} = (0.201, 0.202, 0.201, 0.396)' \\ \text{we have} \quad & \breve{\boldsymbol{a}} = (-6.904, -6.211, -6.904, 0)', \\ & \tilde{\boldsymbol{a}} = (-0.678, -0.673, -0.678, 0)' \end{aligned}$$

$$\begin{aligned} \text{and} \quad & (\Delta^2 \breve{w}_3)^2 = \quad 0.000004 \quad = (\Delta^2 \tilde{w}_3)^2, \\ \text{while} \quad & (\Delta^2 \breve{a}_3)^2 = \quad 1.92 \gg 0.000099 \quad = (\Delta^2 \tilde{a}_3)^2. \end{aligned}$$

Indeed, in the areas with a sufficient amount of data, the estimated shape of the error distribution is mostly driven by the data themselves, whereas in the data-poor areas the shape of the fitted error distribution is inter- or extrapolated from the data-rich areas according to the flexibility allowed by the penalty term.

Second, the penalty of the third order ($s = 3$) based on transformed mixture coefficients $\boldsymbol{a}$ has an interesting property which can serve as a basis for an empirical test of normality (see Section 7.2.3). A basis for this property is given by the following proposition which is proved in Appendix A.

Proposition 7.1. *Let for* $K \in \mathbb{N}$

$$\begin{aligned} \boldsymbol{\mu}^K &= \{\mu_j^K = \frac{j}{K}, \quad j = -K^2, \ldots, K^2\} \\ &= \{-K, -K + \frac{1}{K}, \ldots, -\frac{1}{K}, 0, \frac{1}{K}, \ldots, K - \frac{1}{K}, K\} \end{aligned}$$

be a sequence of knots. Let for $\boldsymbol{a} \in \mathbb{R}^{2K^2+1}$ *a discrete distribution on* $\boldsymbol{\mu}^K$ *be given by*

$$Pr(\mu^K = \mu_j^K \,|\, \boldsymbol{a}) = \exp(a_j).$$

Let $\boldsymbol{a}^K$ *minimizes* $\sum_{j=-K^2+3}^{K^2} \{\Delta^3 a_j\}^2$ *under the constraints*

$$\begin{aligned} \sum_{j=-K^2}^{K^2} Pr(\mu^K = \mu_j^K | \boldsymbol{a}) &= 1, \\ E(\mu^K \,|\, \boldsymbol{a}) &= 0, \\ var(\mu^K \,|\, \boldsymbol{a}) &= 1 - \sigma_0^2 \quad \textit{for } \sigma_0 \in (0,1) \textit{ fixed.} \end{aligned} \tag{7.9}$$

Let

$$g_K(y) = \sum_{j=-K^2}^{K^2} Pr(\mu^K = \mu_j^K | \boldsymbol{a}^K)\, \varphi(y \,|\, \mu_j^K, \sigma_0^2),\ y \in \mathbb{R}.$$

Then for all $y \in \mathbb{R}$

$$\lim_{K\to\infty} g_K(y) = \varphi(y).$$

The empirical normality test is obtained using the following consideration. Suppose that for fixed K we have $2K^2+1$ knots $-K, -K+\frac{1}{K}, \ldots, -\frac{1}{K}, 0, \frac{1}{K}, \ldots, K-\frac{1}{K}, K$. Suppose further that we maximize the penalized log-likelihood (7.7) for $\lambda \to \infty$. This is equivalent (in the limit) to minimizing the penalty term (7.8) under the constraints (7.4). For fixed K, let $g^*_{\varepsilon,K}$ be the fitted standardized error density arising from the above-mentioned optimization problem. Using Proposition 7.1 with $w_j(\boldsymbol{a}) = \Pr(\mu^K = \mu_j^K \,|\, \boldsymbol{a})$, $j = -K^2, \ldots, K^2$ we get that $\lim_{K\to\infty} g^*_{\varepsilon,K}(\varepsilon^*) = \varphi(\varepsilon^*)$ for all $\varepsilon^* \in \mathbb{R}$. In practice, the set of knots and the basis standard deviation recommended in Sections 6.3.1 and 6.3.3 (e.g., knots from -6 to 6 by 0.3 and $\sigma_0 = 0.2$) give already rise to a fitted standardized error density $g^*_{\varepsilon,K}$ practically indistinguishable from the normal density, φ, when only the penalty term is minimized. This property does not hold for the order $s \neq 3$ of the penalty or when the penalty is based on the original mixture coefficients $\boldsymbol{w}$.

7.2.3 Selecting the smoothing parameter

In the area of density estimation, methods for selecting the smoothing parameter, λ, that rely on cross–validation are often used. The standard modified maximum-likelihood cross–validation score that we are attempting to minimize is

$$\mathrm{CV}(\lambda) = -\sum_{i=1}^{N} \ell_i(\hat{\boldsymbol{\theta}}^{(-i)}),$$

where $\hat{\boldsymbol{\theta}}$ is the penalized maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}^{(-i)}$ the penalized MLE based on the sample excluding the ith observation. However, computation and optimization of the cross–validation score is extremely computationally intensive in our case. In a similar context, O'Sullivan (1988) suggested a one-step Newton-Raphson approximation combined with a first-order Taylor series approximation. Applying his method in our setting results in an approximate cross-validation score given by

$$\overline{\mathrm{CV}}(\lambda) = -\Big\{\sum_{i=1}^{n} \ell_i(\hat{\boldsymbol{\theta}}) - \mathrm{trace}\big(\hat{\mathbb{H}}^{-1}\hat{\mathbb{I}}\big)\Big\}, \tag{7.10}$$

where

$$\hat{\mathbb{H}} = -\frac{\partial^2 \ell_P(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}, \qquad \hat{\mathbb{I}} = -\frac{\partial^2 \ell(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}.$$

We denote $\mathrm{trace}(\hat{\mathbb{H}}^{-1}\hat{\mathbb{I}})$ by $\mathrm{df}(\lambda)$ and call it the *effective degrees of freedom* or the *effective dimension* of the model since it necessarily plays the same role as the effective dimension of a linear smoother (Hastie and Tibshirani, 1990). Depending on a chosen order s of the differences in the penalty, the degrees of freedom decreases in λ from $\dim(\boldsymbol{\beta}) + 2 + (2K + 1 - 3)$ for $\lambda = 0$ (i.e., the ordinary log-likelihood) to $\dim(\boldsymbol{\beta}) + 2 + (s - 3)$ for $\lambda \to \infty$ and $s \geq 3$ (i.e., the penalized log-likelihood). For example, when $K = 20$, $\mu_{j+1} - \mu_j = 0.3$, $\sigma_0 = 0.2$ and $s = 3$, penalized likelihood estimation as $\lambda \to \infty$ depends effectively on $2K + 1 - s = 38$ fewer parameters than does ordinary likelihood estimation.

Further, minimizing the expression (7.10) is essentially the same as maximizing Akaike's information criterion $\mathrm{AIC}(\lambda) = \ell(\hat{\boldsymbol{\theta}}) - \mathrm{df}(\lambda)$ (Akaike, 1974). This can be a valuable way to compare different models and assess the importance of covariate contributions (see an example in Section 7.6).

In accompanying R programs (see Appendix C), a grid search using user-defined values $\lambda_1^*, \ldots, \lambda_L^*$ (in our applications we used values $\lambda_1^* = e^2, \lambda_2^* = e^1, \ldots, \lambda_L^* = e^{-9}$) is used to find the optimal AIC. Since the log-likelihood is of the order $O(N)$, using a factor of $N\lambda_l^*/2$ in the penalty term (7.8) instead of $\lambda/2$ allows one to use approximately the same grid for datasets of different sizes while also maintaining the proportional importance of the penalty term in the penalized log-likelihood at the same level.

The result immediately following Proposition 7.1 further implies that with a sufficiently dense set of knots, we can check the normality of the error term. When the optimal value of the tuning parameter λ becomes large the error density of the model can be considered to be normal.

Linear mixed model interpretation

Recently, Wand (2003) or Kauermann (2005a) pointed out the strong link between penalized maximum-likelihood estimation and linear mixed models which can be used for selection of the smoothing parameter. The idea, which underlies also the pseudo-variance estimate in Section 7.3.1 and the full Bayesian developments in Chapters 9 and 10, is the following. The coefficient vector $\boldsymbol{a}$ is considered to be a vector of random effects having the normal distribution

$$\boldsymbol{a} \sim \mathcal{N}\Big(\mathbf{0},\ \lambda^{-1}\left(\mathbb{P}_s'\mathbb{P}_s\right)^{-}\Big),$$

where $(\mathbb{P}_s'\mathbb{P}_s)^{-}$ is the generalized inverse of $\mathbb{P}_s'\mathbb{P}_s$. Smoothing parameter λ then determines (together with the fixed matrix $\mathbb{P}_s$) the variability of the "random effects" $\boldsymbol{a}$. Penalized likelihood (7.7) can then be interpreted as the likelihood of the mixed effects model with normal random effects $\boldsymbol{a}$. The optimal λ value is obtained as the maximum-likelihood or more frequently as the restricted maximum-likelihood estimate of the inverse variance component in such constructed mixed effects model. See, e.g., Cai and Betensky (2003) or Kauermann (2005b) for practical applications of this approach.

7.3 Inference based on the maximum likelihood penalized AFT model

With standard maximum-likelihood method the score vector (the first derivative of the log-likelihood) has a zero mean when its expectation is computed under the true parameter vector. Under a mild regularity conditions, it is then possible to prove that the MLE is an unbiased estimate. However, introduction of the penalty term with $\lambda > 0$ leads to the penalized score vector (the first derivative of the penalized log-likelihood) having a mean different from zero when its expectation is computed under the true parameter vector. Consequently, the penalized MLE $\hat{\boldsymbol{\theta}}$ is a biased estimator and its standard errors may not be very informative when that bias is high. However, there are two possibilities for drawing accurate inferences based on penalized MLE.

7.3.1 Pseudo-variance

Wahba (1983) described a pseudo-Bayesian technique for generating confidence bands around the cross-validated smoothing spline. O'Sullivan (1988) used this technique in the penalized ML framework and his approach can be adopted also here. Basically, the penalized log-likelihood ℓ_P is viewed

as a "posterior" log-density for the parameter $\boldsymbol{\theta}$ and the penalty term as a "prior" negative log-density of that parameter. Then, the second order Taylor series expansion of the "posterior" log-density around its mode $\hat{\boldsymbol{\theta}}$ leads to

$$\ell_P(\boldsymbol{\theta}) \approx \ell_P(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\mathbb{H}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Finally the Gaussian approximation gives "posterior" normal distribution for $\boldsymbol{\theta}$ with covariance matrix

$$\widehat{\text{var}}_P(\hat{\boldsymbol{\theta}}) = \hat{\mathbb{H}}^{-1}. \tag{7.11}$$

We call this estimate of the variance of the penalized MLE $\hat{\boldsymbol{\theta}}$ the "pseudo-variance estimate."

7.3.2 Asymptotic variance

More formal inference is possible under the following assumptions. Firstly, we assume independent noninformative censoring. Secondly, as the sample size N increases, the knots (both number and positions) and the basis standard deviation remain fixed. Let $\boldsymbol{\theta}_T$ be the true parameter value of $\boldsymbol{\theta}$, assuming it exists. To get asymptotically unbiased estimates we have to either keep the value of the smoothing parameter λ constant as $N \to \infty$ or let it increase at a rate lower than N (i.e., $\lambda = \lambda_N$ and $\lim_{N\to\infty} \lambda_N/N = 0$). Under these conditions, the penalty part of the penalized log-likelihood reduces its importance relative to the log-likelihood part as $N \to \infty$ (i.e., as the sample size N increases, the smoothness of the fitted error distribution is determined to greater extent by the data and to a lesser extent by the penalty). Then, in combination with standard maximum likelihood arguments, for arbitrary $\xi > 0$ the penalized MLE $\hat{\boldsymbol{\theta}}$ satisfies $\Pr_{\boldsymbol{\theta}_T}\big(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T| < \xi\big) \to 1$. Using the same arguments as in Gray (1992), one can further show that $\sqrt{N}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T\big)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\lim_{N\to\infty}(N\,\mathbb{W})$ where the matrix $\mathbb{W}$ can be consistently estimated by

$$\widehat{\text{var}}_A(\hat{\boldsymbol{\theta}}) = \hat{\mathbb{H}}^{-1} \hat{\mathbb{I}}\, \hat{\mathbb{H}}^{-1}, \tag{7.12}$$

which we call the "asymptotic variance estimate." As pointed out by Gray (1992), the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ remains the same if the smoothing parameters λ_N are replaced by estimates satisfying $\hat{\lambda}_N/\lambda_N \xrightarrow{\Pr} 1$.

7.3.3 The pseudo-variance versus the asymptotic variance

In various applications, the pseudo-variance estimate (7.11) has been shown to be useful. When smoothing a spline curve $g(t)$, Wahba (1983) showed

it yielded pointwise confidence intervals $\hat{g}(t) \pm z\sqrt{\widehat{\mathrm{var}}_P\{\hat{g}(t)\}}$, where z is a quantile of the normal distribution, that have good frequentist coverage properties. Verweij and Van Houwelingen (1994) used it in the context of penalized likelihood estimation in Cox regression; they called the square roots of its diagonal elements "pseudo-standard errors." Joly, Commenges, and Letenneur (1998) exploited this technique to get confidence bands on the hazard function smoothed using M-splines. In contrast, for the asymptotic variance estimate (7.12) there is no guarantee that for finite samples its middle matrix $\hat{\mathbb{I}}$ is positive semidefinite. Based on our experience, this problem is not rare. Finally, according to our simulations the pseudo-variance estimate (7.11) yields confidence intervals $\hat{\beta} \pm z\sqrt{\widehat{\mathrm{var}}_P(\hat{\beta})}$ for regression parameters with better coverage properties than the corresponding confidence intervals based on the asymptotic estimate (7.12).

7.3.4 Remarks

We have assumed in this section that the true parameter vector $\boldsymbol{\theta}_T$ exists. This does not have to be true. In particular, true $\boldsymbol{a}$ coefficients may fail to exist when the true error distribution is not a mixture of the normal densities determined by the choice of knots and the standard deviation σ_0. However, if the distance between two consecutive knots is small enough, we argue that the penalized mixture of the normal densities can approximate every continuous distribution sufficiently well, see Dalal and Hall (1983) or O'Hagan (1994, Sec. 6.47), that the assumption on the existence of the true parameter vector $\boldsymbol{\theta}_T$ is not restrictive at all. Loosely speaking, combining this with the asymptotic arguments given in Section 7.3.2 implies that by increasing the sample size, the estimated coefficients $\boldsymbol{a}$ will yield an estimated density which is close to the true error density.

7.4 Predictive survival and hazard curves and predictive densities

The penalized AFT model has actually a parametric nature given the weights $w_{-K}, \ldots, w_K$ in (7.2) are known. This makes it easy to compute predictive survival curves or predictive hazards or densities for a given combination of

covariates, say $\boldsymbol{x}_{new}$ and $\boldsymbol{z}_{new}$. The predictive survival function is given by

$$S(t \mid \boldsymbol{x}_{new},\, \boldsymbol{z}_{new}) = \tag{7.13}$$
$$1 - \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \Phi\Big\{ \frac{\log(t) - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_{new}}{\tau(\boldsymbol{z}_{new})} \;\Big|\; \mu_j,\, \sigma_0^2 \Big\}.$$

The predictive density is computed by

$$p(t \mid \boldsymbol{x}_{new},\, \boldsymbol{z}_{new}) = \tag{7.14}$$
$$\big\{ t\, \tau(\boldsymbol{z}_{new}) \big\}^{-1} \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \varphi\Big\{ \frac{\log(t) - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_{new}}{\tau(\boldsymbol{z}_{new})} \;\Big|\; \mu_j,\, \sigma_0^2 \Big\},$$

and finally the predictive hazard is obtained from the above quantities as

$$\hbar(t \mid \boldsymbol{x}_{new},\, \boldsymbol{z}_{new}) = \frac{p(t \mid \boldsymbol{x}_{new},\, \boldsymbol{z}_{new})}{S(t \mid \boldsymbol{x}_{new},\, \boldsymbol{z}_{new})}. \tag{7.15}$$

In practice, all unknown parameters are replaced by their penalized maximum-likelihood estimates.

7.5 Simulation study

To see how the proposed method performs, we carried out a simulation study. 'True' uncensored data were generated according to model (7.1) with error density (7.2). Two covariates, i.e. $\boldsymbol{x}_i = (x_{i,1},\, x_{i,2})'$ were included in the model and the values of the parameters were the following: $\alpha = 1.6$, $\tau = 1.4$ and $\boldsymbol{\beta} = (-0.8,\, 0.4)'$. The covariate $x_{i,1}$ was binary taking a value of 1 with probability 0.4 and covariate $x_{i,2}$ was generated according to the extreme value distribution of a minimum, with location 8.5 and scale 1. The model attempts to mimic an AFT model used for the dataset presented in Section 7.6 with $x_{i,1}$ playing the role of the covariate lesion and $x_{i,2}$ being distributed as $\log_2(1 + \text{CD4 count})$. Time to the event T is expressed in months. The standardized error term ε^* was generated from a standard normal distribution $\mathcal{N}(0, 1)$, from a standardized extreme value distribution, and from a mixture of two normal distributions $0.4\mathcal{N}(-1.4, 0.8^2) + 0.6\mathcal{N}(0.93, 0.8^2)$. Samples of sizes 50, 100, 300, and 600 were generated. Each simulation involved 100 replications.

For each uncensored dataset we created four censored datasets that were then used to compute the estimates: a dataset with (1) approximately 20% right-censored and 80% uncensored observations *(light RC)*; (2) approximately 20%

right and 80% interval-censored observations *(light R+IC)*; (3) approximately 60% right and 40% uncensored observations *(heavy RC)*; (4) approximately 60% right and 40% interval-censored observations *(heavy R+IC)*. The censoring was created by simulating consecutive 'visit times' for each subject in the dataset. Times of the first 'visits' were drawn from $\mathcal{N}(7, 1)$ distribution. Further, times between each consecutive 'visits' were simulated from $\mathcal{N}(6, 0.5^2)$. This approach reflects the idea that subjects in our Oral Substudy were seen for the first time about 7 months after the onset of the parent study and then approximately every 6 months for several years. At each visit, subjects were withdrawn (censored) according to a prespecified percentage (between 0.4% and 0.7% for light censoring and between 4.0% and 5.0% for heavy censoring) creating right-censored observations provided that the uncensored event time T_i was greater than the visit time at which the subject was withdrawn. To obtain interval-censored observations, we took the 'visit' interval that contained the uncensored event time T_i.

For comparison, estimates for each dataset were computed using our smoothed procedure and using two parametric models: an AFT model on the log scale with a correctly specified error distribution (normal, extreme value or mixture of normals, respectively) and a log-normal AFT model. For the smoothing procedure, the third order penalty, equidistant knots with a distance of 0.3 between consecutive knots, and the basis standard deviation of 0.2 were used.

Selected results of the simulation are given in Appendix B, Section B.1. Namely, Tables B.1 – B.6 show the results for the regression parameters. It is seen that, in most cases, our smoothed procedure performs better than the incorrectly specified log-normal AFT model and often only but slightly worse than the correctly specified parametric AFT model. Additionally, when our smoothing approach is used, the error distribution is reproduced rather satisfactory as can be seen in Figures B.1 – B.3. This property is quite important especially when the estimated model is to be used for prediction purposes. Further, it is seen that even for small samples the performance of our smoothing procedure is quite similar to the performance of a parametric AFT model with a correctly specified error distribution.

7.6 Example: WIHS data – interval censoring

In Section 1.2, we introduced the study comprising the cohort of seropositive women and the cohort of seronegative women with an increased risk of HIV infection. In this section, we concentrate on the data set collected in the framework of the Oral Substudy involving 224 seropositive AIDS-free (at baseline) women. We explore how the distribution of the time between the

Table 7.1: WIHS Data. Akaike's information criterion, degrees of freedom, the optimal $\log(\lambda/N)$ for the fitted models.

Model	AIC	df	$\log(\lambda/N)$
(1) lesion	−262.39	3.2	2
(2) lvload	−256.16	3.4	2
(3) lcd4	−256.94	3.4	2
(4) lesion + lvload	−255.63	4.4	2
(5) lesion + lcd4	−253.19	8.9	−7
(6) lvload + lcd4	−253.45	8.4	−6
(7) lesion + lvload + lcd4	−250.01	10.0	−7

baseline measurement and the onset of an AIDS-related illness can be explained using classical predictors which are the number of copies of the HIV RNA virus and the count of CD4 T-cells per ml of blood. Additionally, we examine whether presence of one of the three lesion markers, *oral candidiasis*, *hairy leukoplakia* and *angular cheilitis*, is useful, possibly together with one or both laboratory predictors, in describing the distribution of the residual time to onset of AIDS.

For the purpose of modelling, the three lesion markers were summarized in one binary covariate, lesion, equal to one if at least one of the above mentioned three lesion markers was present. Further, the laboratory predictors entered the models in an transformed way, classically used in the HIV research. Namely, the covariate lvload is equal to $\log_{10}(1 + \text{viral load})$ and the covariate lcd4 equals $\log_2(1 + \text{CD4 count})$. All three covariates are moderately to strongly associated with one another since, as AIDS progresses, viral load increases, CD4 count falls, and oral lesions occur more frequently. In our sample, for women with lesion $= 0$ and 1, respectively, the median lvload was 3.60 and 4.23 (Mann-Whitney p-value, 0.001), There was also a moderate negative correlation of -0.46 between lcd4 and lvload. These associations have to be taken into account when interpreting the results.

As a response, we used the time in months between the baseline visit, defined as the first visit at which the lesion markers were collected by dental professionals, and the onset of an AIDS-related illness. As mentioned in Section 1.3, the response time is *right-censored* for 158 women and *interval-censored* for 66 women with the average length of the observed interval equal to 7 months.

7.6.1 Fitted models

To obtain the results shown below, we used a sequence of 41 equidistant knots from -6 to 6 with a distance of 0.3 between each pair. The basis standard deviation was 0.2 and the third order difference was used in the penalty. Different models were compared using Akaike's information criterion and claims concerning the significance of the parameters were based on Wald's tests using the pseudo-variance estimate (7.11). Summary of the fitted models is shown in Tables 7.1 and 7.2.

If used alone (model (1)) the effect of lesion on the time to onset of AIDS is statistically significant ($p = 0.018$) and the estimated time is $\exp(-0.87) \approx 0.42$ times shorter for women with lesion $= 1$ than women with lesion $= 0$. According to the AIC values for models (2) and (3), the transformed CD4 count and viral load are equally good predictors of the time to onset of AIDS. Addition of the lesion marker (models (4) and (5)) improves the model with lcd4 considerably but improves the model with lvload only slightly. Finally, some additional improvement is gained by considering the model with all three predictors (model (7)).

7.6.2 Predictive survival and hazard curves, predictive densities

Figure 7.1 shows predictive survival and hazard curves and predictive densities for women with lesion $= 0$ and lesion $= 1$ based on the simplest model lesion and on the most complex model considered lesion + lvload + lcd4. The predictive survival curves based on the model lesion are further overlaid with the nonparametric estimate of Turnbull (1976) in each group. The two estimates are quite close to each other, illustrating the semiparametric nature of our approach. However, our procedure gives smooth estimates of the survival curves and moreover enables quantification of the difference in survival between the two groups. Notice further that due to the fact that the hazard is obtained as a ratio of the density and the survival function, which relatively slowly varies from one, only a slight difference is observed between the predictive density and the hazard.

Further, we point out that the predictive densities for models where lcd4 was not involved are very close to the log-normal density. This is not surprising since the optimal tuning parameter λ for these models was equal to $224 \cdot \exp(2)$, essentially a value of infinity in this practical situation and thus implying that the fitted error distributions are close to the normal distribution, as discussed in Section 7.2.3. On the other hand, models where lcd4 was

Table 7.2: WIHS Data. Estimates of the regression parameters (standard error; p-value) for the fitted models.

Model	lesion	logvload	logcd4
(1) lesion	-0.87		
	$(0.37;\ 0.018)$		
(2) lvload		-0.76	
		$(0.19;\ < 0.001)$	
(3) lcd4			0.44
			$(0.11;\ < 0.001)$
(4) lesion + lvload	-0.62	-0.70	
	$(0.36;\ 0.080)$	$(0.19;\ < 0.001)$	
(5) lesion + lcd4	-0.78		0.39
	$(0.26;\ 0.003)$		$(0.07;\ < 0.001)$
(6) lvload + lcd4		-0.39	0.38
		$(0.14;\ 0.004)$	$(0.06;\ < 0.001)$
(7) lesion + lvload + + lcd4	-0.60	-0.30	0.39
	$(0.23;\ 0.008)$	$(0.11;\ 0.005)$	$(0.05;\ < 0.001)$

used in combination with other covariates gave much lower optimal tuning parameters λ, implying also non-normal error densities. This is seen on the right-hand side of Figure 7.1. The phenomenon could indicate presence of a risk-group mixture in the data or absence of another important predictor. Indeed, a factor that could play an important role is antiretroviral therapy, which might have been used by some women in our sample. However, this factor requires modelling time–dependent covariates, which cannot be done with our model.

7.6.3 Conclusions

In conclusion, the time to AIDS onset in this study population is notably shorter in women with oral lesions. Further, this marker improves the prediction of that time based on any of the classical indicators (CD4 count and viral load). When interpreting these findings, one must bear in mind that only a limited number of WIHS women opted to participate in the Oral Substudy, the source of the dental data. Thus they may differ in unknown ways from the

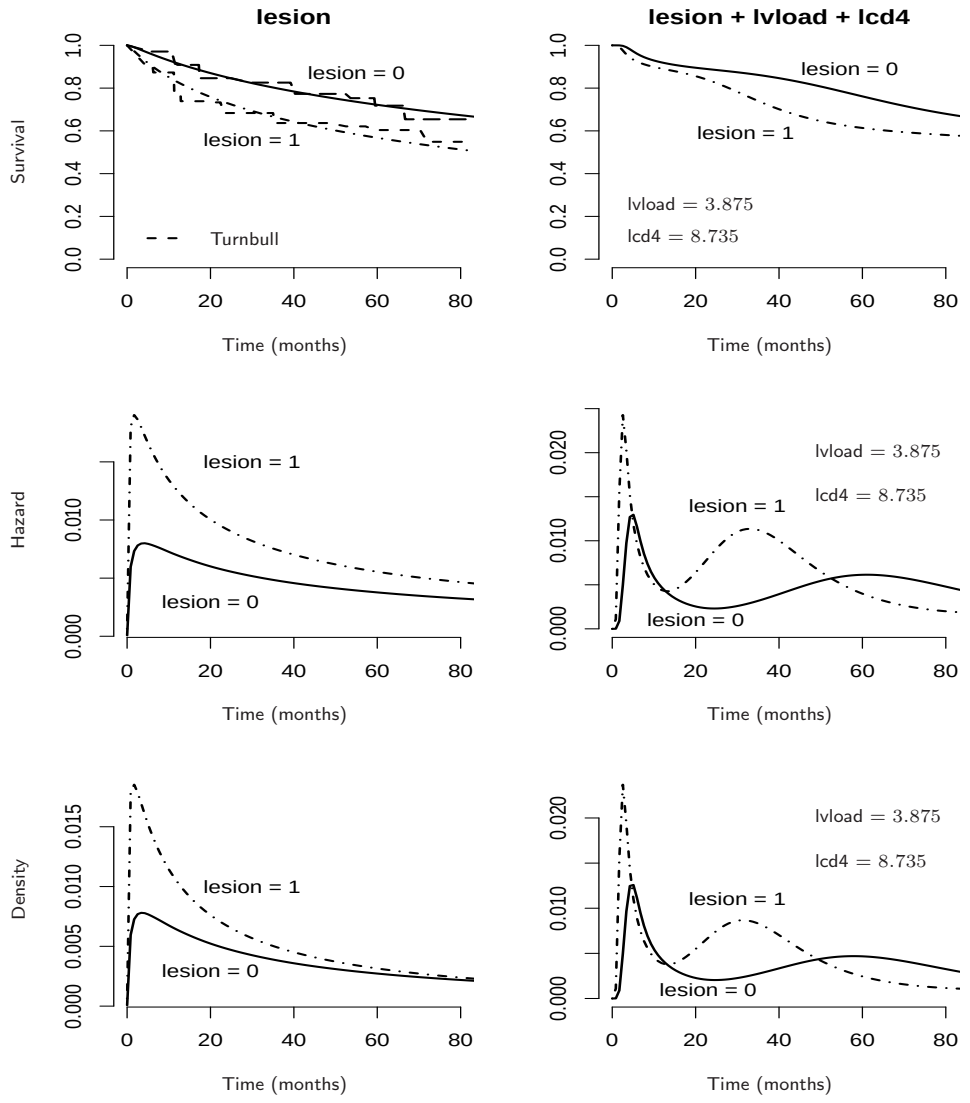


Figure 7.1: WIHS Data. Predicted survival curves, hazard curves and densities for women with lesion = 1 (dotted-dashed line) vs. women with lesion = 0 (solid line) based on models lesion (left part) and lesion + lvload + lcd4 (right part). Predictive curves for the latter model control for a median value of lvload = 3.875 and a median value of lcd4 = 8.735. Predictive survival curves for model lesion are further compared to the nonparametric estimate of Turnbull (1976) in each group.

Table 7.3: Signal Tandmobiel® study. Description of fitted models.

Models with constant scale	
gender	$\boldsymbol{x} = (\mathsf{gender})$
dmf	$\boldsymbol{x} = (\mathsf{dmf})$
gender + dmf	$\boldsymbol{x} = (\mathsf{gender}, \mathsf{dmf})'$
gender $*$ dmf	$\boldsymbol{x} = (\mathsf{gender}, \mathsf{dmf}, \mathsf{gender} \times \mathsf{dmf})'$
Mean-scale models	
gender $*$ dmf/scale(dmf)	$\boldsymbol{x} = (\mathsf{gender}, \mathsf{dmf}, \mathsf{gender} \times \mathsf{dmf})'$
	$\boldsymbol{z} = (\mathsf{dmf})$
gender $*$ dmf/scale(gender $*$ dmf)	$\boldsymbol{x} = (\mathsf{gender}, \mathsf{dmf}, \mathsf{gender} \times \mathsf{dmf})'$
	$\boldsymbol{z} = (\mathsf{gender}, \mathsf{dmf}, \mathsf{gender} \times \mathsf{dmf})'$

overall set. Nonetheless, our findings are consistent with those of others who have evaluated oral lesions as predictors of AIDS onset and they illustrate use of our method in the area of AIDS research. Our method restricts us to analysis of baseline covariates. Although this is a very widely applicable special case, extension of the method to accommodate time-dependent covariates would allow more complex relationships between outcomes and covariates.

7.7 Example: Signal Tandmobiel® study – interval-censored data

In paediatric dentistry and orthodontics, adequate knowledge of timing and patterns of tooth emergence is useful for diagnosis and treatment planning. This motivates an example in this section where we fit the distribution of emergence times of permanent maxillary right premolars (teeth 14 and 15 in Figure 1.1) based on the data from the Signal Tandmobiel® study introduced in Section 1.1.

It is anticipated, that the distribution of emergence times of a particular tooth is different for boys and girls. See Figure 5.1 and Table 5.1 where the emergence distributions for boys and girls are compared for tooth 44. However, a similar phenomenon is observed also for other teeth, 14 and 15 included. For that reason, we used the covariate gender (0 for boys and 1 for girls) in our models. Additionally, it was of dental interest to check whether the distribution of the emergence time of a permanent tooth changes when the primary predecessor of the permanent tooth experienced caries or not.

Table 7.4: Signal Tandmobiel® study. Akaike's information criteria for different models.

Model	Tooth 14	Tooth 15
gender	−5 532.59	−4 551.57
dmf	−5 538.03	−4 549.93
gender + dmf	−5 494.51	−4 526.85
gender ∗ dmf	−5 491.47	−4 522.76
gender ∗ dmf/scale(dmf)	−5 468.61	−4 506.66
gender ∗ dmf/scale(gender ∗ dmf)	−5 467.67	−4 507.59

For this, we included a binarised dmf score pertaining to the predecessor as a covariate, dmf = 1 if the primary predecessor of that permanent tooth was recorded as decayed, or missing due to caries, or filled and 0 otherwise.

As response, for a particular child, we consider the age of emergence of a particular permanent tooth (14 or 15), recorded in years. Due to the design of the study (annual planned examinations), the response variable is interval-censored with intervals of length equal to approximately 1 year. It should be stressed that in this section, the two teeth will be analyzed separately, i.e. ignoring their possible correlation. In Section 7.8, we indicate how the correlation between teeth can be incorporated in the analysis. For a better fit, we shifted the time origin of the AFT model to 5 years of age which is clinically minimal emergence time for the permanent teeth (see, e.g., Ekstrand, Christiansen, and Christiansen, 2003). Namely, we replaced T_i by $T_i - 5$ in the AFT model specification (7.1). Similarly as in Section 7.6, we used a sequence of 41 equidistant knots from −6 to 6 with a distance of 0.3 between each pair. The basis standard deviation was 0.2 and the third order difference was used in the penalty.

7.7.1 Fitted models

We fitted four penalized AFT models with constant scale parameter and two mean-scale penalized AFT models. The fitted models are described in Table 7.3 and AIC's for these models are given in Table 7.4. The model selection was based on the AIC.

Firstly, the model with the interaction term gender ∗ dmf seems to fit the data best and the interaction term cannot be omitted. Secondly, the models where the scale parameter τ depends on covariates give a considerably better fit. For tooth 15, only dmf included in the scale covariate vector leads to the

Table 7.5: Signal Tandmobiel® study. Estimates (standard errors) for the models gender $*$ dmf/scale(dmf).

Parameter	Tooth 14	Tooth 15
α	1.7734 (0.0073)	1.9143 (0.0091)
β(gender)	−0.0931 (0.0099)	−0.0803 (0.0110)
β(dmf)	−0.0990 (0.0116)	−0.0773 (0.0125)
β(gender $*$ dmf)	0.0401 (0.0166)	0.0473 (0.0172)
γ_1	−1.5613 (0.0219)	−1.6121 (0.0351)
γ(dmf)	0.2144 (0.0307)	0.2415 (0.0399)

best AIC. For tooth 14, the model with the scale depending only on dmf can be improved by inclusion of gender and its interaction with dmf however the improvement is minor. These findings lead us to conclude that the model that describes satisfactory well the data while being kept as simple as possible is the model gender $*$ dmf/scale(dmf). The estimates for this model are given in Table 7.5. It is seen that dmf $= 1$ accelerates the emergence for both genders and also increases the variability of the emergence distribution.

7.7.2 Predictive emergence and hazard curves

For our data, predictive emergence curves (cumulative distribution functions), which are prefered in this case to survival curves, based on the model gender $*$ dmf/scale(dmf) are shown in Figure 7.2 and predictive hazards in Figure 7.3. Further, Figure 7.2 shows also the non-parametric estimates of Turnbull (1976) computed separately for each combination of covariates. It is seen that model-based emergence curves agree with the non-parametric estimates indicating the goodness-of-fit of our model. Further, the figures show that the difference between children with dmf $= 0$ and dmf $= 1$ is higher for boys than for girls and that the emergence process for boys is indeed postponed compared to girls.

Non-decreasing predictive hazard curves reflect the nature of the problem at hand. Indeed, it can be expected that, provided the tooth of a child has not emerged yet, the probability that the tooth will emerge increases with age.

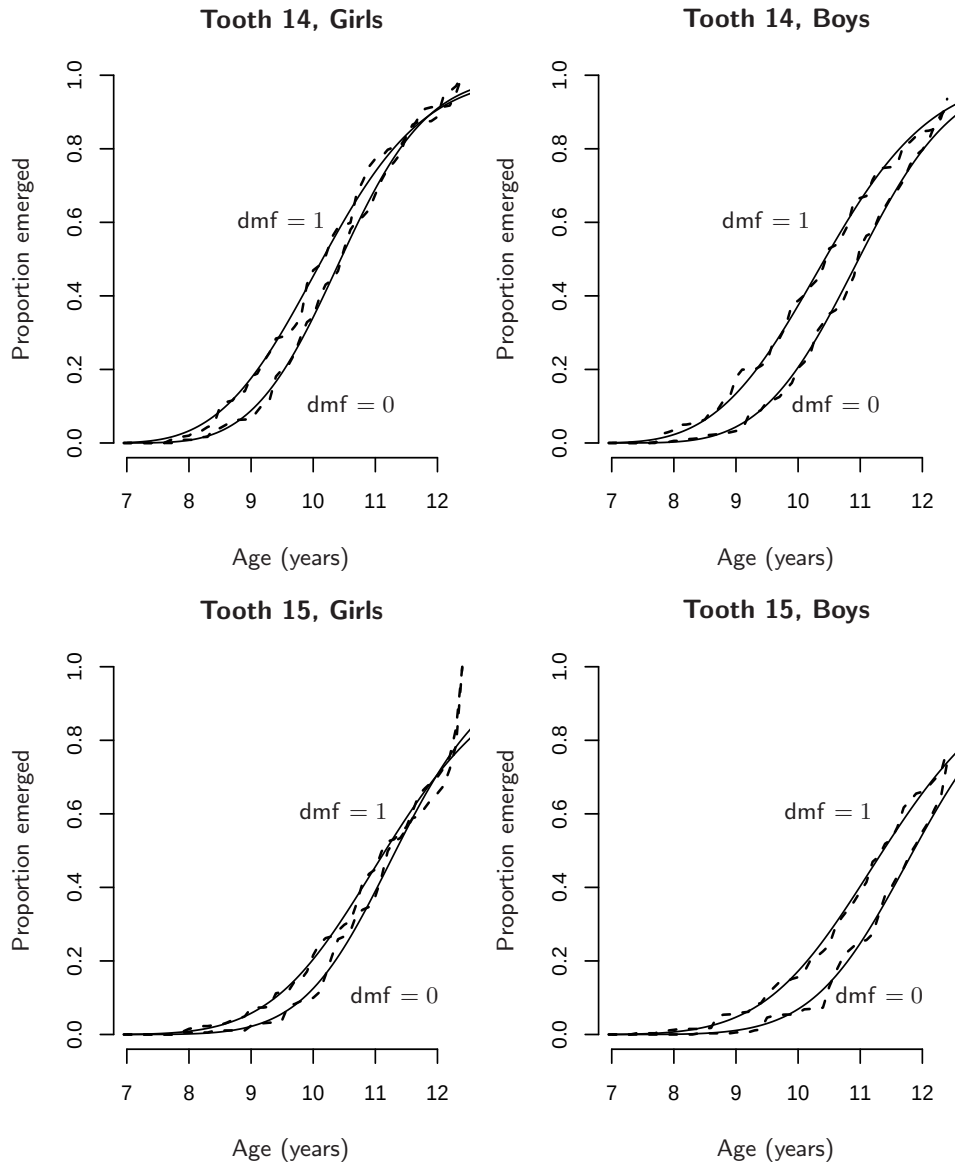


Figure 7.2: Signal Tandmobiel® study. Predictive emergence curves: solid lines for curves based on the model gender $*$ dmf/scale(dmf) (on each plot: left curve for dmf $= 1$, right curve for dmf $= 0$), dashed line for a non-parametric estimate of Turnbull.

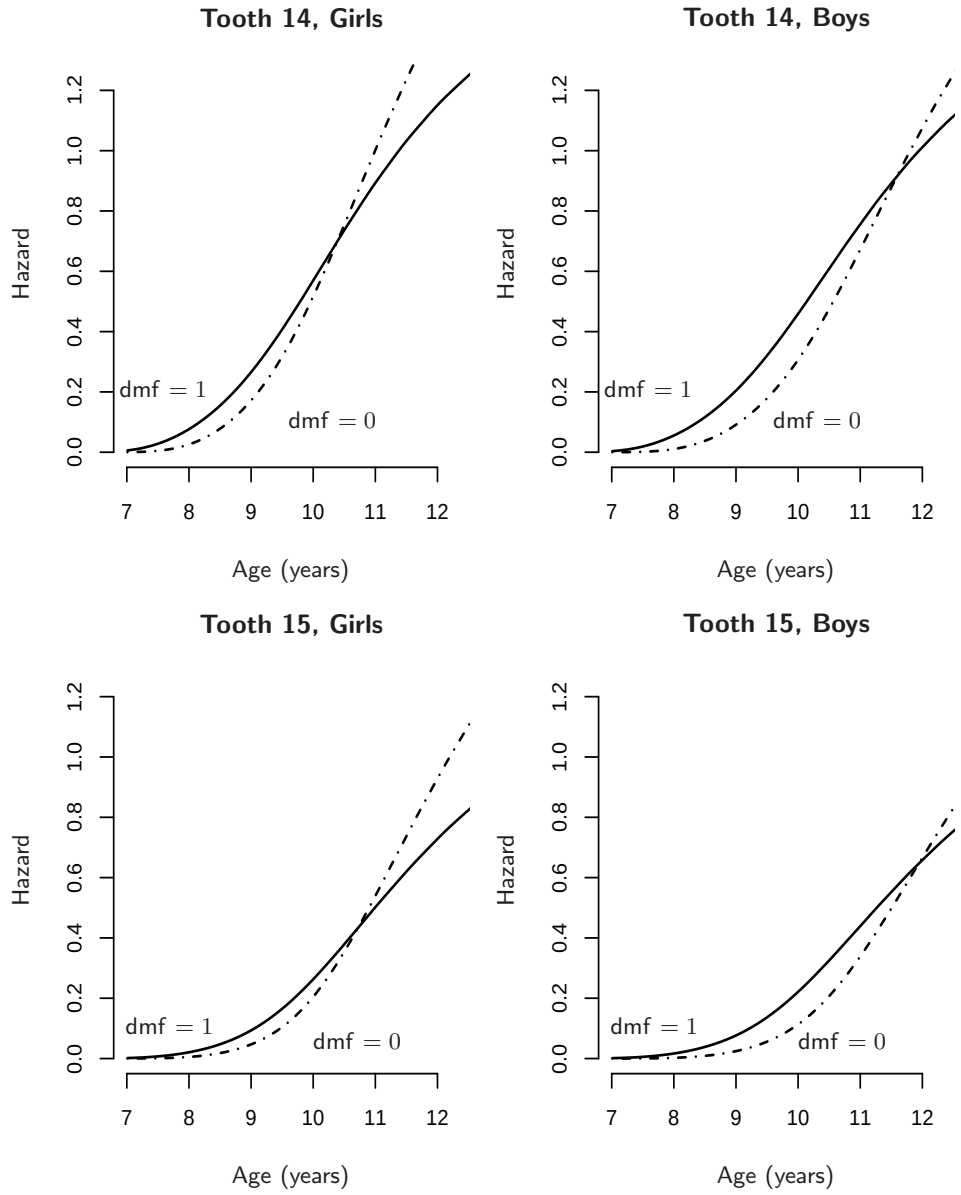


Figure 7.3: Signal Tandmobiel® study. Predictive hazard curves based on the model gender $*$ dmf/scale(dmf): solid line for dmf $= 1$, dotted-dashed line for dmf $= 0$.

Table 7.6: Signal Tandmobiel® study. Estimates (standard errors) for the models gender, dmf and gender + dmf.

Parameter	Model gender or dmf	Model gender + dmf
	Tooth 14	
β(gender)	−0.0740 (0.0080)	−0.0766 (0.0081)
β(dmf)	−0.0729 (0.0086)	−0.0741 (0.0085)
	Tooth 15	
β(gender)	−0.0564 (0.0085)	−0.0594 (0.0087)
β(dmf)	−0.0613 (0.0089)	−0.0628 (0.0090)

7.7.3 Comparison of emergence distributions between different groups

While the model gender $*$ dmf/scale(dmf) gives a parsimonious description of emergence distributions for different groups of children and serves as a solid basis for prediction as was shown in the previous section, it is not suitable to provide simple p-values for a comparison of emergence distributions between e.g. boys and girls. Due to the fact that an interaction term gender $*$ dmf appeared to be significantly important, we could only provide a p-value for a multiple comparison of the four groups (girls with dmf $= 1$ and 0 and boys with dmf $= 1$ and 0).

To simply compare two distributions, while averaging the effect of other covariates, the basic AFT model with a univariate covariate x (i.e. either the model gender or the model dmf) can be used together with a significance test for the group parameter. Additionally, it is possible to perform a test that compares two groups while controling for additional confounding variables (e.g. comparison of boys and girls while controling for dmf or vice versa). To do that, we perform significance tests of $\boldsymbol{\beta}$ parameters in the model gender + dmf.

The estimates of regression parameters $\boldsymbol{\beta}$ together with their standard errors, derived from the formula (7.11), in mentioned models are given in Table 7.4. The Wald tests of significance for each β parameter all yield p-values lower than 0.0001, which confirm the findings obtained previously that there is indeed a significant difference in emergence distributions of studied teeth between boys and girls and also between the group of children with dmf $= 0$ and dmf $= 1$. The difference remains both marginally (irrespective of value of dmf or irrespective of value of gender, respectively) and while controling for the other covariate.

The issue of the robustness of the AFT model against the omitted covariates, discussed in Section 3.3, is further illustrated in Table 7.4. The effect of gender remains almost unchanged in both models, gender and gender + dmf, and an analogous conclusion holds also for the effect of dmf.

7.7.4 Conclusions

It has been shown that the emergence process of teeth 14 and 15 is significantly different between boys and girls and that the caries experience status of a primary predecessor, expressed by the dmf score, has a significant effect on the timing of emergence of permanent successors.

Predictive emergence curves have been drawn that can be used for diagnosis and treatment planning in paediatric dentistry. Further, it was found that the acceleration effect of caries experience on a primary predecessor on the timing of emergence of its successor was stronger for boys than for girls.

7.8 Discussion

In this chapter, we have suggested a method useful for fitting the linear regression model for independent censored observations while avoiding overly restrictive parametric assumptions on the error distribution. Most classically, the logarithmic transformation of the response leads to the well known AFT model. However, other transformations of the response leading to its potential range covering the whole real line are also possible. The density of the error distribution is specified in a semi-parametric way as a mixture of the overspecified number of normal densities with fixed means – knots and given common standard deviation. Mixture coefficients are estimated using the penalized maximum-likelihood method. Such model specifications allow flexibility with respect to the resulting error distribution yet retain tractability such that data carrying censoring of several types, especially interval censoring, can be handled naturally.

The method of this chapter could generally be extended to handle also multivariate survival data. Namely, the population averaged AFT model (see Section 3.4.2) with a multivariate error distribution specified as a multivariate penalized mixture (see Section 6.3.4) could be used. Or alternatively, the cluster specific AFT model (see Section 3.4.3) with an error distribution given as a penalized mixture and random effects distribution specified either parametrically or as a (multivariate) penalized mixture could be considered. However, as outlined in Sections 4.2 and 4.3, the computation and let alone

optimization of the (penalized) likelihood is practically intractable. For this reason, we switch to fully Bayesian approaches using the MCMC methodology.

Chapter 8

Bayesian Normal Mixture Cluster-Specific AFT Model

In this chapter we present a cluster-specific AFT model (see Section 3.4.3) with a flexible error distribution. This model, introduced by Komárek and Lesaffre (2006a), allows us to analyze also data sets where not necessary all observations are independent. For example, we will be able to analyze jointly several teeth from the Signal Tandmobiel® study, analyze the CGD data where the times to recurrent infections are involved or to analyze the data from the multicenter studies like EBCP data. The approach presented here uses the classical normal mixture (see Section 6.1) to express the error density in the AFT model. For the random effects we use a parametric (multivariate) normal distribution. The full Bayesian approach with the Markov chain Monte Carlo methodology will be used for the inference.

In Section 8.1, we specify the cluster-specific AFT model and the distributional assumptions we use in this chapter. In Section 8.2, we specify the model from the Bayesian perspective and derive the corresponding posterior distribution. Details of the Markov chain Monte Carlo methodology to sample from the posterior distribution are given in Section 8.3. In Section 8.4, we show how the survival distributions for specific combinations of covariates can be estimated. Further, in Section 8.5, we give the estimates of the individual random effects that could be used, for example, for the discrimination. The performance of the method is evaluated using the simulation study in Section 8.6. The method is applied to the analysis of the interval-censored emergence times of 8 permanent teeth in Section 8.7, to the recurrent events analysis in Section 8.8 and to the analysis of the breast cancer multicenter study in Section 8.9. The chapter is finalized by the discussion in Section 8.10.

8.1 Model

Let $T_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ be the lth event time in the ith cluster or the lth recurrent event on the ith subject in the study. Let $T_{i,l}$ be observed as an interval $\lfloor t_{i,l}^L, t_{i,l}^U \rfloor$. Let logarithmic transformations of the event and observed event times be $Y_{i,l} = \log(T_{i,l})$, $y_{i,l}^L = \log(t_{i,l}^L)$, $y_{i,l}^U = \log(t_{i,l}^U)$. We will assume that the random vectors $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_N$, where $\boldsymbol{T}_i = (T_{i,1}, \ldots, T_{i,n_i})'$, $i = 1, \ldots, N$ are independent. However, the components of each $\boldsymbol{T}_i$ are not necessarily independent.

To model the effect of covariates on the event time we use the cluster-specific AFT model (3.7), i.e.

$$\log(T_{i,l}) = Y_{i,l} = \boldsymbol{\beta}'\boldsymbol{x}_{i,l} + \boldsymbol{b}_i'\boldsymbol{z}_{i,l} + \varepsilon_{i,l}, \quad i = 1, \ldots, N,\ l = 1, \ldots, n_i, \quad (8.1)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ is the unknown regression coefficient vector, $\boldsymbol{x}_{i,l}$ the covariate vector for fixed effects, $\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,q})'$, $i = 1, \ldots, N$ are the random effect vectors with the density $g_b(\boldsymbol{b})$ causing the possible correlation for the components of $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})'$. Further, $\boldsymbol{z}_{i,l}$ is the covariate vector for random effects and $\varepsilon_{i,l}$ are independent and identically distributed random variables with the density $g_\varepsilon(\varepsilon)$. Along the lines of Gelman et al. (2004, Chapter 15) we use the terms 'fixed' and 'random' effects throughout the thesis even in a Bayesian context where all unknown parameters are treated as random quantities.

For recurrent events, usually $\boldsymbol{z}_{i,l} = 1$ for all i and l and $\boldsymbol{b}_i = b_{i,1}$ expresses an individual-specific deviation from an overall mean log-event time which is not explained by fixed effects covariates (see the analysis of CGD data in Section 8.8). For clustered data, the vector $\boldsymbol{z}_{i,l}$ may define further sub-clusters (as in the analysis of the Signal Tandmobiel® data in Section 8.7) allowing for a higher dependence of observations within sub-clusters given by common values of appropriate components of the vector $\boldsymbol{b}_i$ while keeping the dependence also across the sub-clusters through the correlation between the components of $\boldsymbol{b}_i$. In multicenter clinical trials where the aim is to evaluate an effect of some treatment (e.g. the EBCP data analyzed in Section 8.9), the vector $\boldsymbol{z}_{i,l}$ might be equal to $(1, \mathsf{treatment}_{i,l})'$ allowing that both a baseline value of the expected event time and a treatment effect can vary across centra.

8.1.1 Distributional assumptions

The density $g_\varepsilon(\varepsilon)$ of the error term $\varepsilon_{i,l}$ in model (8.1) is specified in a flexible way as a classical normal mixture (6.3), i.e.

$$g_\varepsilon(\varepsilon) = \sum_{j=1}^{K} w_j \varphi(y \,|\, \mu_j,\, \sigma_j^2), \tag{8.2}$$

where K is the *unknown* number of mixture components and further, $\boldsymbol{w} = (w_1, \ldots, w_K)'$ are unknown mixture weights, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ unknown mixture means and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)'$ unknown mixture variances.

We have already mentioned in Section 6.1.2 that a heteroscedastic mixture (8.2) leads to the likelihood which is unbounded if the parameter space for variances is unconstrained. In a full Bayesian analysis, this difficulty is solved by using an appropriate prior distribution for the variances which plays the role of constraints. We discuss this issue in full detail in Section 8.2.1.

For the random effects $\boldsymbol{b}_i$, we take a suitable parametric distribution, namely the multivariate normal distribution, see Section 8.2.2 for details. The fact that we put more emphasis on a correct specification of the distribution of the error term $\varepsilon_{i,l}$ than on a specification of the distribution of random effects $\boldsymbol{b}_i$ is driven by the following reasoning.

For an AFT model, the regression parameters $\boldsymbol{\beta}$ express the effect of covariates ($\boldsymbol{x}_{i,l}$) both conditionally (given $\boldsymbol{b}_i$) and marginally (after integrating $\boldsymbol{b}_i$ out). Both interpretations do not change when different distributional assumptions are made on $\boldsymbol{b}_i$. Further, with a correctly specified distribution of $\varepsilon_{i,l}$ the conditional model is always correctly specified. However, when the distribution of $\varepsilon_{i,l}$ is incorrect neither conditional nor marginal models are specified correctly. Further, Keiding, Andersen, and Klein (1997) showed that for univariate (single-spell) Weibull AFT model the regression parameters are robust against the misspecification of the random effects distribution. This finding, also for non-Weibull models is further supported by the empirical results of Lambert et al. (2004). Finally, Verbeke and Lesaffre (1997) showed, in the context of normal linear mixed model with uncensored data, that the maximum-likelihood estimates of the regression parameters are unaffected by the misspecified random effects distribution.

Of course, in situations in which the variability of the random effects *considerably* exceeds the variability of the error term it becomes more important to specify correctly the distribution of the random effects rather than the distribution of the error term. However, in all applications presented in this chapter this is not the case.

8.1.2 Likelihood

The likelihood contribution of the ith cluster can be derived from expressions (4.7) and (4.9). Namely,

$$L_i = \int_{\mathbb{R}^q} \left\{ \prod_{l=1}^{n_i} \oint_{y_{i,l}^L}^{y_{i,l}^U} g_\varepsilon(y_l - \boldsymbol{\beta}' \boldsymbol{x}_{i,l} - \boldsymbol{b}_i' \boldsymbol{z}_{i,l})\, dy_l \right\} g_b(\boldsymbol{b}_i)\, d\boldsymbol{b}_i. \tag{8.3}$$

It might be useful to stress again that due to multivariate integration in the likelihood (8.3), it is rather cumbersome to use maximum-likelihood based methods for the cluster-specific AFT model with interval-censored observations even with $g_\varepsilon(\varepsilon)$ and $g_b(\boldsymbol{b})$ being parametrically specified. Mainly for this reason, the full Bayesian approach will be exploited.

8.2 Bayesian hierarchical model

The Bayesian specification of the model continues by specification of the prior distributions for all unknown parameters, denoted by $\boldsymbol{\theta}$. We assume a cluster-specific AFT model (8.1) with a hierarchical structure graphically represented by a directed acyclic graph (DAG) given in Figure 8.1. As explained in Section 4.4, the joint prior distribution of $\boldsymbol{\theta}$ is then given by the product of the conditional distributions of the nodes pertaining to unobserved quantities given their parents, namely

$$\begin{aligned} p(\boldsymbol{\theta}) \propto & \\ & \prod_{i=1}^{N} \left[\prod_{l=1}^{n_i} \left\{ p\big(t_{i,l} \,\big|\, \boldsymbol{\beta},\, \boldsymbol{b}_i,\, \varepsilon_{i,l}\big) \times p\big(\varepsilon_{i,l} \,\big|\, \boldsymbol{\mu},\, \boldsymbol{\sigma}^2,\, r_{i,l}\big) \times p\big(r_{i,l} \,\big|\, K,\, \boldsymbol{w}\big) \right\} \times \right. \\ & \left. p\big(\boldsymbol{b}_i \,\big|\, \boldsymbol{\gamma},\, \mathbb{D}\big) \right] \times \\ & p\big(\boldsymbol{\mu} \,\big|\, K\big) \times p\big(\boldsymbol{\sigma}^2 \,\big|\, K,\, \eta\big) \times p\big(\eta\big) \times p\big(\boldsymbol{w} \,\big|\, K\big) \times p\big(K\big) \times \\ & p\big(\boldsymbol{\beta}\big) \times p\big(\boldsymbol{\gamma}\big) \times p\big(\mathbb{D}\big). \end{aligned} \tag{8.4}$$

For clarity, we omitted all fixed hyperparameters and fixed covariates in the expression (8.4). As the DAG indicates, the unknown parameters can be split into two parts connected only through the node of the true event times. The conditional distribution for this node is simply a Dirac (degenerated) distribution driven by the AFT model (8.1), i.e.

$$\begin{gathered} p(t_{i,l} \mid \boldsymbol{\beta},\, \boldsymbol{b}_i,\, \varepsilon_{i,l}) = I[\log(t_{i,l}) = \boldsymbol{\beta}' \boldsymbol{x}_{i,l} + \boldsymbol{b}_i' \boldsymbol{z}_{i,l} + \varepsilon_{i,l}], \\ i = 1, \ldots, N,\ l = 1, \ldots, n_i. \end{gathered}$$

In the subsequent sections, we explain all the multiplicands of expression (8.4) and also the meaning of the newly introduced parameters $r_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$, $\boldsymbol{\gamma}$, $\mathbb{D}$, and η.

8.2.1 Prior specification of the error part

The prior conditional distributions pertaining to the error part of the model are inspired by the work of Richardson and Green (1997) (with some change in notation) who studied Bayesian estimation of the normal mixtures in the context of i.i.d. data. That is, they did not consider covariates or censoring.

To improve the computation of the posterior distribution, it is useful to assume that $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ come from a heterogeneous population consisting of groups $j = 1, 2, \ldots, K$ of sizes proportional to the mixture weights w_j and introduce latent allocation variables $r_{i,l}$ denoting the label of the group from which each random error variable $\varepsilon_{i,l}$ is drawn. By this we are introducing here the Bayesian implementation of the data augmentation algorithm (see Section 4.3). Together with distributional assumption (8.2) this

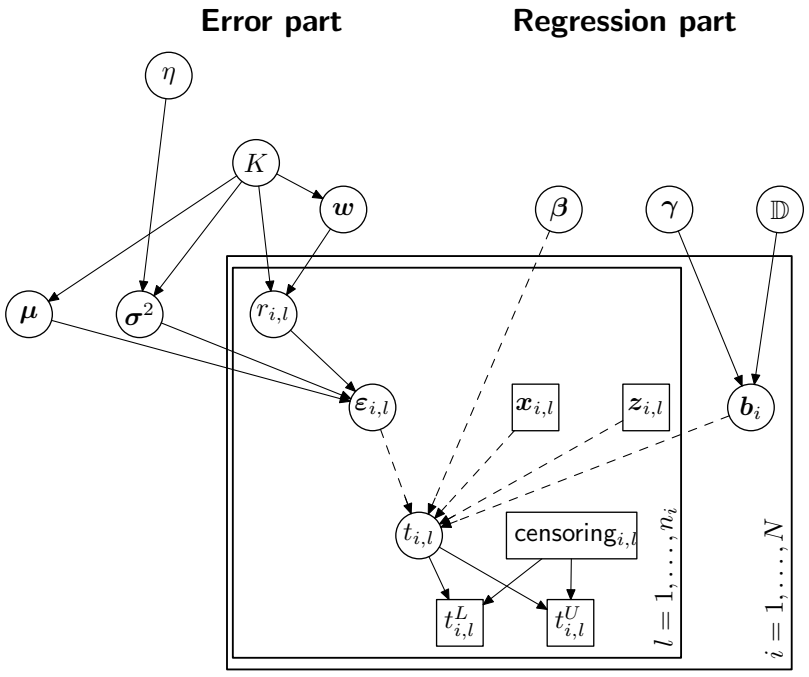


Figure 8.1: Directed acyclic graph for the Bayesian normal mixture cluster-specific AFT model.

leads to the following conditional distributions appearing in the prior (8.4):

$$\begin{aligned} &\Pr(r_{i,l} = j \mid K, \boldsymbol{w}) = w_j, && j \in \{1, \ldots, K\}, \\ &p(\varepsilon_{i,l} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2, r_{i,l}) = \varphi(\varepsilon_{i,l} \mid \mu_{r_{i,l}}, \sigma^2_{r_{i,l}}) && i = 1, \ldots, N,\ l = 1, \ldots, n_i. \end{aligned}$$

For the number of mixture components, K, we experimented with

1. a Poisson distribution with mean equal to a fixed hyper-parameter λ truncated at some prespecified (relatively large) value K_{max} and truncated zero, i.e.

$$\Pr(K = k) = \Big\{ \sum_{j=1}^{K_{max}} \frac{\lambda^j}{j!} \Big\}^{-1} \frac{\lambda^k}{k!}, \qquad k = 1, \ldots, K_{max};$$

2. a uniform distribution on $\{1, \ldots, K_{max}\}$, i.e.

$$\Pr(K = k) = \frac{1}{K_{max}}, \qquad k = 1, \ldots, K_{max}.$$

The prior for the mixture weights $\boldsymbol{w}$ is taken to be a symmetric K-dimensional Dirichlet with prior 'sample size' equal to $K\,\delta$, i.e.

$$p(\boldsymbol{w} \mid K) = \frac{\Gamma(K\,\delta)}{\big\{\Gamma(\delta)\big\}^K} \prod_{j=1}^{K} w_j^{\delta - 1},$$

where δ is a fixed hyperparameter.

Further, the mixture means μ_j and variances σ_j^2, $j = 1, \ldots, K$ are a priori all drawn independently with normal $\mathcal{N}(\xi, \kappa)$ and inverse-gamma $\mathrm{IG}(\zeta, \eta)$ priors respectively, i.e.

$$p(\boldsymbol{\mu} \mid K) = \prod_{j=1}^{K} \varphi(\mu_j \mid \xi, \kappa), \tag{8.5}$$

$$p(\boldsymbol{\sigma}^2 \mid K, \eta) = \prod_{j=1}^{K} \left\{ \frac{\eta^\zeta}{\Gamma(\zeta)} (\sigma_j^2)^{-(\zeta+1)} \exp\Big(-\frac{\eta}{\sigma_j^2}\Big) \right\}, \tag{8.6}$$

where ξ, κ and ζ are fixed hyperparameters. As in Richardson and Green (1997) we let the hyperparameter η follow a gamma distribution with fixed shape parameter h_1 and fixed rate parameter h_2, i.e.

$$p(\eta) = \frac{h_2^{h_1}}{\Gamma(h_1)} \eta^{h_1 - 1} \exp(-h_2 \eta).$$

A rationale for this construction is given in Section 8.2.3.

Since the error model is invariant to permutations of labels $j = 1, \ldots, K$, the joint prior distribution of a vector $\boldsymbol{\mu}$ is restricted to the set $\{\boldsymbol{\mu} : \mu_1 < \cdots < \mu_K\}$ for identifiability reasons, see Stephens (2000) or Jasra, Holmes, and Stephens (2005) for other approaches to establish identifiability. The joint prior distribution of the mixture means and variances is thus $K!$ times the products (8.5) and (8.6), restricted to above mentioned set of increasing means.

8.2.2 Prior specification of the regression part

The regression part of the model has the structure of a classical Bayesian linear mixed model (see, e.g., Gelman et al., 2004, Chapter 5). Let $\mathbb{X}$ be a $(\sum_{i=1}^{N} n_i) \times m$ matrix with vectors $\boldsymbol{x}'_{1,1}, \ldots, \boldsymbol{x}'_{N,n_N}$ as rows. Similarly, let $\mathbb{Z}$ be a $(\sum_{i=1}^{N} n_i) \times q$ matrix with vectors $\boldsymbol{z}'_{1,1}, \ldots, \boldsymbol{z}'_{N,n_N}$ as rows. Further, we will assume that the matrix $(\mathbb{X}, \mathbb{Z})$ is of full column rank $(m + q)$. In other words, covariates included in $\boldsymbol{x}_{i,l}$ are not included in $\boldsymbol{z}_{i,l}$ and vice versa. This gives rise to hierarchical centering which in general results in a better behavior of the MCMC algorithm (Gelfand, Sahu, and Carlin, 1995). Finally, since $g_\varepsilon(\varepsilon)$ does not have zero mean we do not allow a column of ones in the matrix $\mathbb{X}$ to avoid identifiability problems.

The prior distribution for each regression coefficient β_j, $j = 1, \ldots, m$ is assumed to be $\mathcal{N}(\nu_{\beta,j}, \psi_{\beta,j})$, and the β_j are assumed to be a priori independent, i.e.

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \varphi(\beta_j \,|\, \nu_{\beta,j},\, \psi_{\beta,j}).$$

The vectors $\boldsymbol{\nu}_\beta = (\nu_{\beta,1}, \ldots, \nu_{\beta,m})'$ and $\boldsymbol{\psi}_\beta = (\psi_{\beta,1}, \ldots, \psi_{\beta,m})'$ are fixed hyperparameters.

As already mentioned in Section 8.1.1, the (prior) distribution for the random effect vector $\boldsymbol{b}_i$, $i = 1, \ldots, N$ is assumed to be (multivariate) normal with a prior mean $\boldsymbol{\gamma}$ and a prior covariance matrix $\mathbb{D}$, i.e.

$$p(\boldsymbol{b}_i \,|\, \boldsymbol{\gamma},\, \mathbb{D}) = \varphi_q(\boldsymbol{b}_i \,|\, \boldsymbol{\gamma},\, \mathbb{D}), \tag{8.7}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$.

The prior distribution for each γ_j, $j = 1, \ldots, q$ is $\mathcal{N}(\nu_{\gamma,j}, \psi_{\gamma,j})$, independently for $j = 1, \ldots, q$, i.e.

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{q} \varphi(\gamma_j \,|\, \nu_{\gamma,j},\, \psi_{\gamma,j}).$$

The vectors $\boldsymbol{\nu}_\gamma = (\nu_{\gamma,1}, \ldots, \nu_{\gamma,q})'$ and $\boldsymbol{\psi}_\gamma = (\psi_{\gamma,1}, \ldots, \psi_{\gamma,q})'$ are fixed. Special care is needed when the random intercept is included in the model (i.e. when $\mathbb{Z}$ contains a column of ones, let say its first column). Hierarchical centering cannot be applied in this case since the overall intercept is given by the mean of the mixture (8.2). For that reason, γ_1 is fixed to zero (or equivalently, $\nu_{\gamma,1} = 0,\ \psi_{\gamma,1} = 0$).

The prior distribution for the covariance matrix $\mathbb{D}$ of random effects is assumed to be an inverse-Wishart with fixed degrees of freedom df and a fixed scale matrix $\mathbb{S}$, i.e.

$$p(\mathbb{D}) = \left\{ 2^{\frac{df\,q}{2}}\, \pi^{\frac{q(q-1)}{4}} \prod_{j=1}^{q} \Gamma\Big(\frac{df+1+j}{2}\Big) \right\}^{-1} \times \qquad (8.8)$$
$$|\mathbb{S}|^{\frac{df}{2}}\, |\mathbb{D}|^{-\frac{df+q+1}{2}}\, \exp\Big\{ -\frac{1}{2} \operatorname{trace}(\mathbb{S}\mathbb{D}^{-1}) \Big\}.$$

In the special case of a univariate random effect ($q = 1$), we use d instead of $\mathbb{D}$ and s instead of $\mathbb{S}$ in the notation. Note that in that case, the inverse-Wishart distribution is the same as the inverse-gamma distribution with the shape parameter equal to $df/2$ and the scale parameter equal to $s/2$.

Further, in the situation of $q = 1$, we considered alternatively (see Section 8.8) also the use of a uniform prior for standard deviation of the random effect which is often considered to be a better choice (see Gelman et al., 2004, pp. 136, 390 or Gelman, 2006), i.e. a priori

$$p(\sqrt{d}) = \frac{1}{\sqrt{s}}\, I[0 < d < s], \qquad (8.9)$$

for a large value of s. On the original variance scale the prior (8.9) transforms into

$$p(d) = \frac{1}{2\sqrt{s\,d}}\, I[0 < d < s],$$

which is formally a truncated inverse-gamma distribution with the shape parameter equal to $-1/2$ and the scale parameter equal to zero.

8.2.3 Weak prior information

In this problem, we have opted for specifying weak prior information on the parameters of interest. When a priori information is available, our prior assumptions could be appropriately modified.

For the regression part of the model, we use non-informative, however proper distributions, that is, the prior variances of regression parameters β ($\boldsymbol{\psi}_\beta$)

and $\boldsymbol{\gamma}$ ($\boldsymbol{\psi}_\gamma$) are chosen such that the posterior variance of the regression parameters is at least 100 times lower (which must be checked from the results). Prior hyperparameters for the covariance matrix $\mathbb{D}$ giving a weak prior information correspond to choices of $df = q-1+c$ and $\mathbb{S} = \text{diag}(c, \ldots, c)$ with c being a small positive number.

In the error part of the model, it is not possible to be fully non-informative, i.e. to use priors $p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mid K) \propto 1 \times \prod_{j=1}^{K} \sigma_j^{-2}$ and to obtain proper posterior distributions (Diebolt and Robert, 1994; Roeder and Wasserman, 1997). Richardson and Green (1997) offer, in the context of i.i.d. observations, for say $e_1, \ldots, e_N$, the following alternative: A rather flat prior $\mathcal{N}(\xi, \kappa)$ for each μ_j is achieved by letting ξ equal to $\bar{e} = N^{-1} \sum_{i=1}^{N} e_i$ and setting κ equal to a multiple of R^2, where $R = \max(e_i) - \min(e_i)$. They point out that it might be restrictive to suppose that knowledge of the range or variability of the data implies much about the size of each single σ_j^2 and therefore introduced an additional hierarchical level by allowing η to follow a gamma distribution with parameters h_1 and h_2. They recommend taking $\zeta > 1 > h_1$ to express the belief that the σ_j^2 are similar which is necessary to avoid a problem of unbounded likelihood, without being informative about their absolute size. Finally, they suggest setting the parameter h_2 to a small multiple of $1/R^2$. Here, the residuals $y_{i,l} - \boldsymbol{\beta}'\boldsymbol{x}_{i,l} - \boldsymbol{b}_i'\boldsymbol{z}_{i,l}$ play the role of the observations e_i. A rough estimate of their location and scale can be obtained through a maximum-likelihood fit of the AFT model, even without random effects (the scale of residuals can only increase), with an explicitly included intercept and scale parameters in the model. This can be done using standard software packages as R, SPLUS, SAS. The estimated intercept from this model can then be used instead of $\bar{e}$ and a multiple of the estimated scale parameter instead of R.

8.2.4 Posterior distribution

As we indicated in Section 4.4, the joint posterior distribution, $p(\boldsymbol{\theta} \,|\, \text{data})$, is proportional to the product of all DAG conditional distributions, i.e.

$$p\big(\boldsymbol{\theta} \,\big|\, \text{data}\big) \propto p\big(\boldsymbol{\theta}\big) \times \prod_{i=1}^{N} \prod_{l=1}^{n_i} p\big(t_{i,l}^L,\, t_{i,l}^U \,\big|\, t_{i,l},\, \mathsf{censoring}_{i,l}\big), \tag{8.10}$$

where $p(\boldsymbol{\theta})$ is given by (8.4) and $p(t_{i,l}^L,\, t_{i,l}^U \,|\, t_{i,l},\, \mathsf{censoring}_{i,l})$ is discussed below. A box called $\mathsf{censoring}_{i,l}$ in the DAG represents a realization of the random variable(s) causing the censoring of the (i,l)th event time. Note, that under the assumption of independent noninformative censoring (see Section 2.4) there is no need to specify a measurement model for the censoring mechanism

since it only acts as a multiplicative constant in the posterior. After omitting subscripts i, l for clarity, the expression of $p(t^L, t^U \mid t, \text{censoring})$ is rather obvious for most censoring mechanisms.

For example with interval censoring resulting from checking the survival status at (random) times $\boldsymbol{C} = \{c_0, \ldots, c_{S+1}\}$, where $c_0 = 0$, $c_{S+1} = \infty$ we obtain a Dirac density

$$p(t^L = c_s, t^U = c_{s+1} \mid t, \boldsymbol{C}) = I\big\{t \in \lfloor c_s, c_{s+1} \rfloor\big\}, \quad s = 0, \ldots, S.$$

With standard right-censoring driven by the (random) censoring time $C = c$, the following Dirac densities are obtained

$$p(t^L = t^U = t \mid t, c) = I[t \leq c],$$
$$p(t^L = t, t^U = \infty \mid t, c) = I[t > c].$$

8.3 Markov chain Monte Carlo

Inference is based on a sample from the posterior distribution obtained using the MCMC methodology (see Section 4.5). The parameters of the error part of the model are updated using the combination of the reversible jump MCMC algorithm of Green (1995) and a conventional Gibbs algorithm (Geman and Geman, 1984). For the remaining parameters of the model, each iteration of the MCMC is conducted using the Gibbs sampler. Both the reversible jump MCMC algorithm and the full conditional distributions needed to implement the Gibbs sampler are discussed below.

8.3.1 Update of the error part of the model

Details on how to implement the update of the parameters of the error part of the model are given in Richardson and Green (1997). Their guidelines, now based on residuals $\varepsilon_{i,l} = y_{i,l} - \boldsymbol{\beta}'\boldsymbol{x}_{i,l} - \boldsymbol{b}_i'\boldsymbol{z}_{i,l}$, can be immediately applied. We give only a brief summary and for details we refer therein.

Six move types are suggested by Richardson and Green (1997), namely

(i) Updating the mixture weights $\boldsymbol{w}$ while keeping K fixed;

(ii) Updating the mixture means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2$ while keeping K fixed;

(iii) Updating the allocation parameters $r_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$;

(iv) Updating the variance-hyperparameter η;

(v) *Split-combine* move, i.e. splitting one mixture component into two, or combining two into one;

(vi) *Birth-death* move, i.e. the birth or death of an empty mixture component.

In our context, due to the regression and the presence of censored data, we add one more move type, i.e.
(vii) Updating the residuals $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$.
Note that only move types (v) and (vi) change the dimension of the parameter vector by changing K to $K-1$ or $K+1$ and are performed using the reversible jump MCMC algorithm. The moves (i)–(iv) and the move (vii) are performed by sampling from the full conditional distributions given below.

Full conditional for mixture weights $\boldsymbol{w}$

The full conditional distribution for the mixture weights is Dirichlet with parameters $\delta + N_j$, $j = 1, \ldots, K$, i.e.

$$p(\boldsymbol{w} \mid \cdots) = \frac{\Gamma(K\delta + n)}{\prod_{j=1}^{K} \Gamma(\delta + N_j)} \prod_{j=1}^{K} w_j^{\delta + N_j - 1},$$

where $n = \sum_{i=1}^{N} n_i$ is the total sample size and N_j, $j = 1, \ldots, K$ is the number of observations currently allocated in the jth mixture component, i.e.

$$N_j = \sum_{i=1}^{N} \sum_{l=1}^{n_i} I[r_{i,l} = j], \quad j = 1, \ldots, K.$$

Full conditional for mixture means

The full conditional for each mixture mean is normal with the mean and variance

$$\mathrm{E}(\mu_j \mid \cdots) = \frac{\sigma_j^{-2} \sum\limits_{(i,l):\, r_{i,l} = j} \varepsilon_{i,l} + \kappa^{-1}\xi}{\sigma_j^{-2} N_j + \kappa^{-1}}, \qquad j = 1, \ldots, K,$$

$$\mathrm{var}(\mu_j \mid \cdots) = \frac{1}{\sigma_j^{-2} N_j + \kappa^{-1}}, \qquad j = 1, \ldots, K.$$

Note that due to the ordering constraint $\mu_1 < \cdots < \mu_K$, the full conditional only generates a proposal which is accepted provided it does not break this ordering.

Full conditional for mixture variances

The full conditional for each mixture variance is an inverse gamma distribution

$$\sigma_j^2 \mid \cdots \sim \text{I-Gamma}\Big\{\zeta + \frac{N_j}{2},\, \eta + \frac{1}{2}\sum_{(i,l):\, r_{i,l}=j} (\varepsilon_{i,l} - \mu_j)^2\Big\}.$$

Full conditional for the allocation variables

The full conditional for each allocation variable $r_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ is discrete with

$$\Pr(r_{i,l} = j \mid \cdots) \propto \frac{w_j}{\sigma_j} \exp\Big\{-\frac{(\varepsilon_{i,l} - \mu_j)^2}{2\sigma_j^2}\Big\}, \qquad j \in \{1, \ldots, K\}.$$

Full conditional for the variance-hyperparameter

The full conditional for the variance hyperparameter η is a gamma distribution

$$\eta \mid \cdots \sim \text{Gamma}(h_1 + K\,\zeta,\, h_2 + \sum_{j=1}^{K} \sigma_j^{-2}).$$

Split-combine move

To perform the split-combine move, firstly a *random* choice is made whether to try to perform the split or combine move, namely, given K, the probability of attempting the split move is π_K^{split} and the probability of attempting the combine move is $\pi_K^{combine} = 1 - \pi_K^{split}$. Obviously, $\pi_1^{split} = 1$ and $\pi_{K_{max}}^{split} = 0$. Otherwise we use $\pi_K^{split} = \pi_K^{combine} = 0.5$, $K = 2, \ldots, K_{max} - 1$.
When the ***combine*** move is attempted the new mixture with $K - 1$ components is proposed as follows:

1. Choose *at random* a pair of mixture components $(j_1,\, j_2)$ such that for the current values of the mixture means holds

$$\mu_{j_1} < \mu_{j_2} \text{ and there is no other } \mu_j \text{ in the interval } [\mu_{j_1},\, \mu_{j_2}]; \quad (8.11)$$

2. Propose a new mixture component by merging the j_1th and the j_2th component. Label this new component by j^*. Set the weight, mean and variance of the new component such that its 0th, 1st and 2nd moments are the same as those of the combination of the merged components, i.e.

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2}, \\ \mu_{j^*} &= \frac{w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2}}{w_{j^*}}, \\ \sigma^2_{j^*} &= \frac{w_{j_1}(\mu^2_{j_1} + \sigma^2_{j_1}) + w_{j_2}(\mu^2_{j_2} + \sigma^2_{j_2})}{w_{j^*}} - \mu^2_{j^*}. \end{aligned} \tag{8.12}$$

3. Propose new values for the allocation variables $r_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ that were equal to j_1 or to j_2, i.e. set such allocation variables equal to j^*.

4. Accept the proposed mixture with $K-1$ components with the probability

$$\Pr^{combine}_{accept} = \min\{1,\, A^{-1}_{sc}(K-1)\},$$

where the acceptance ratio $A_{sc}(K-1)$ is discussed below. If not accepted keep the current K-component mixture.

The ***split*** move must be reversible in the sense described in Green (1995) to the combine move. Namely it consists of the following steps:

1. Choose *at random* a component j^* which is proposed to be splitted;

2. Propose two new mixture components, labeled j_1 and j_2. To keep reversibility, set their weights, means and variances such that the equation (8.12). This can be done by sampling a three-dimensional auxiliary random vector $\boldsymbol{u} = (u_1, u_2, u_3)'$ from some distribution with a density $p_u(\boldsymbol{u})$ and setting

$$\begin{aligned} w_{j_1} &= w_{j^*}u_1, & w_{j_2} &= w_{j^*}(1-u_1), \\ \mu_{j_1} &= \mu_{j^*} - u_2\sigma_{j^*}\sqrt{\frac{w_{j_2}}{w_{j_1}}}, & \mu_{j_2} &= \mu_{j^*} + u_2\sigma_{j^*}\sqrt{\frac{w_{j_1}}{w_{j_2}}}, \\ \sigma^2_{j_1} &= u_3(1-u_2^2)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j_1}}, & \sigma^2_{j_2} &= (1-u_3)(1-u_2^2)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j_2}}. \end{aligned} \tag{8.13}$$

Check whether the condition (8.11) holds. If not, *reject* directly the split-proposal otherwise continue;

3. Propose new values (either j_1 or j_2) for these allocation variables $r_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ that were equal to j^*. This is done randomly with

$$\Pr_{alloc}(r_{i,l} = j_1) \propto \frac{w_{j_1}}{\sigma_{j_1}} \exp\Big\{-\frac{(\varepsilon_{i,l} - \mu_{j_1})^2}{2\sigma_{j_1}^2}\Big\},$$
$$\Pr_{alloc}(r_{i,l} = j_2) \propto \frac{w_{j_2}}{\sigma_{j_2}} \exp\Big\{-\frac{(\varepsilon_{i,l} - \mu_{j_2})^2}{2\sigma_{j_2}^2}\Big\}.$$

4. Accept the proposed mixture with $K + 1$ components with the probability

$$\Pr_{accept}^{split} = \min\big\{1,\, A_{sc}(K)\big\},$$

see below for the expression of the acceptance ratio $A_{sc}(K)$. If not accepted keep the current K-component mixture.

The acceptance ratio $A_{sc}(K)$ has the following general structure:

$$A_{sc}(K) = [\text{posterior ratio}] \times [\text{proposal ratio}](K) \times [\text{Jacobian}].$$

The individual components of the above product have the following meaning.

$$[\text{posterior ratio}] = \frac{p(\boldsymbol{\theta}^{j_1,j_2} \mid \text{data})}{p(\boldsymbol{\theta}^{j^*} \mid \text{data})},$$

where the posterior density $p(\cdot\,|\text{data})$ is given by (8.10). Further, $\boldsymbol{\theta}^{j_1,j_2}$ refers to the parameter vector pertaining to the proposal in the case of the split move and to the current values of parameters in the case of the combine move. Similarly, $\boldsymbol{\theta}^{j^*}$ refers to the current parameter vector in the case of the split move and to the proposal in the case of the combine move. The proposal ratio is given by

$$[\text{proposal ratio}](K) = \frac{\pi_{K+1}^{combine}}{\pi_K^{split} p_u(\boldsymbol{u}) \prod\limits_{(i,l):\, r_{i,l}=j^*} \Pr_{alloc}(r_{i,l})}.$$

Finally, the Jacobian refers to the transformation (8.13) from $(w_{j^*},\, \mu_{j^*},\, \sigma_{j^*}^2,\, u_1,\, u_2,\, u_3)'$ to $(w_{j_1},\, w_{j_2},\, \mu_{j_1},\, \mu_{j_2}, \sigma_{j_1}^2,\, \sigma_{j_2}^2)'$, i.e.

$$[\text{Jacobian}] = \left| \frac{w_{j^*}\, \sigma_{j_1}^2\, \sigma_{j_2}^2\, (\mu_{j_2} - \mu_{j_1})}{\sigma_{j^*}^2\, u_2\, (1 - u_2^2)\, u_3\, (1 - u_3)} \right|$$

What leaves to be discussed is the choice of the density $p_u(\boldsymbol{u})$ of the auxiliary random vector $\boldsymbol{u}$ used to generate the proposal in the split move. Richardson

and Green (1997) suggest to generate u_1, u_2 and u_3 independently from the following beta distributions:

$$u_1 \sim \text{Beta}(2,\, 2), \qquad u_2 \sim \text{Beta}(2,\, 2), \qquad u_3 \sim \text{Beta}(1,\, 1).$$

Note that at each iteration of the MCMC a new auxiliary vector $\boldsymbol{u}$ is generated also independently on the previous iteration. Brooks, Giudici, and Roberts (2003) showed that some improvement of the MCMC sampling can be achieved by allowing (a) a correlation between the components of $\boldsymbol{u}$; (b) a serial correlation between the auxiliary vectors $\boldsymbol{u}$ generated at successive iterations of the MCMC. In our practical applications (Sections 8.7, 8.8 and 8.9) we exploited their methodology as well.

Birth-death move

Similarly as in the split-combine move, it is *randomly* chosen whether the birth or the death move will be attempted. If the current number of mixture components is K, the birth move is attempted with the probability π_K^{birth} and the death move with the probability $\pi_K^{death} = 1 - \pi_K^{birth}$. Analogously to the probabilities of the split and combine moves we use $\pi_1^{birth} = 1$, $\pi_{K_{max}}^{birth} = 0$ and $\pi_K^{birth} = \pi_K^{birth} = 0.5$, $K = 2, \ldots, K_{max} - 1$.
When the ***birth*** move is attempted the new mixture with $K+1$ components is proposed in the following steps:

1. Sample the weight, mean and the variance for the new component from the following distributions:

$$\begin{aligned} w_{j^*} &\sim \text{Beta}(1,\, K), \\ \mu_{j^*} &\sim \mathcal{N}(\xi,\, \kappa), \\ \sigma^2_{j^*} &\sim \text{I-Gamma}(\zeta,\, \eta). \end{aligned} \tag{8.14}$$

   Note that the expectation of the new weight is equal to $1/(K+1)$, i.e. a reciprocal of the number of components in the proposed mixture;

2. In the proposed mixture, rescale the weights such that they, together with the new weight w_{j^*}, sum to one, i.e. the weights of the proposed mixture are $w'_1, \ldots, w'_K, w_{j^*}$ with

$$w'_j = w_j(1 - w_{j^*}), \qquad j = 1, \ldots, K. \tag{8.15}$$

3. Accept the proposed mixture with $K+1$ components with the probability

$$\text{Pr}_{accept}^{birth} = \min\bigl\{1, A_{bd}(K)\bigr\},$$

see below for the form of the acceptance ratio $A_{bd}(K)$. If not accepted, keep the current K-component mixture.

When it is chosen to propose the ***death*** move, the new mixture with $K-1$ components is proposed in the follwoing way

1. Check whether there are any *empty* mixture components, i.e. the components for which $N_j = \sum_{i,l} I[r_{i,l} = j]$ is equal to zero. If not the death move is directly rejected;

2. Choose *randomly* an empty mixture component. Let j^* be the label of this component;

3. In the proposed $(K-1)$-component mixture, delete the j^*th component and rescale the remaining weights such that they sum to one, i.e. the proposed mixture has the weights w'_j, $j = 1, \ldots, K$, $j \neq j^*$.

$$w'_j = \frac{w_j}{1 - w_{j^*}}, \qquad j = 1, \ldots, K, \ j \neq j^*;$$

4. Accept the proposed mixture with $K-1$ components with the probability

$$\Pr_{accept}^{death} = \min\{1, A_{bd}^{-1}(K-1)\},$$

where the acceptance ratio $A_{bd}^{-1}(K-1)$ is given below. If not accepted keep the current K-component mixture.

Analogous to the split-combine move, the acceptance ratio $A_{bd}(K)$ has the general structure

$$A_{bd}(K) = [\text{posterior ratio}] \times [\text{proposal ratio}](K) \times [\text{Jacobian}],$$

where

$$[\text{posterior ratio}] = \frac{p(\boldsymbol{\theta}^+ \mid \text{data})}{p(\boldsymbol{\theta}^- \mid \text{data})}.$$

The vector $\boldsymbol{\theta}^+$ refers to the set of the parameters containing the proposed mixture in the case of the birth move and the set of the current parameter values in the case of the death move. Similarly, the vector $\boldsymbol{\theta}^-$ refers to the set of the current parameter values in the case of the birth move and to the set of parameters contaning the proposed mixture in the case of the death move. Further, the proposal ratio is given by

$$[\text{proposal ratio}](K) = \frac{\pi_{K+1}^{death}}{\pi_K^{birth}\, p_{prop}(w_{j^*},\, \mu_{j^*},\, \sigma_{j^*}^2)},$$

where $p_{prop}(w_{j^*},\, \mu_{j^*},\, \sigma^2_{j^*})$ is the density of the proposal step given by (8.14), i.e.

$$p_{prop}(w_{j^*},\, \mu_{j^*},\, \sigma^2_{j^*}) = \\ \left\{K\,(1-w_{j^*})^{K-1}\right\} \,\times\, \varphi(\mu_{j^*}\,|\,\xi,\,\kappa) \,\times\, \left\{\frac{\eta^\zeta}{\Gamma(\zeta)}\,(\sigma^2_{j^*})^{-(\zeta+1)}\,\exp\Big(-\frac{\eta}{\sigma^2_{j^*}}\Big)\right\}.$$

Finally, the Jacobian refers to the transformation (8.15), i.e.

$$[\text{Jacobian}] = (1-w_{j^*})^K.$$

Updating the residuals

The update of the residuals $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ is fully deterministic provided the (i,l)th residual correspond to an uncensored observation $t_{i,l} = t^L_{i,l} = t^U_{i,l}$. In such case, the update of $\varepsilon_{i,l}$ consists of using the AFT expression (8.1) with the current values of the parameters, i.e. the updated $\varepsilon_{i,l}$ is equal to $\log(t_{i,l}) - \boldsymbol{\beta}'\boldsymbol{x}_{i,l} - \boldsymbol{b}_i'\boldsymbol{z}_{i,l}$.

When the residual $\varepsilon_{i,l}$ corresponds to the censored observation with an observed interval $\lfloor t^L_{i,l},\, t^U_{i,l}\rfloor$ its update consists of sampling from the full conditional distribution of $\varepsilon_{i,l}$ which appears to be a truncated normal distribution, namely $\mathcal{N}(\mu_{r_{i,l}},\, \sigma^2_{r_{i,l}})$ truncated on $\Big\lfloor \log(t^L_{i,l}) - \boldsymbol{\beta}'\boldsymbol{x}_{i,l} - \boldsymbol{b}_i'\boldsymbol{z}_{i,l},\ \log(t^U_{i,l}) - \boldsymbol{\beta}'\boldsymbol{x}_{i,l} - \boldsymbol{b}_i'\boldsymbol{z}_{i,l}\Big\rfloor$.

8.3.2 Update of the regression part of the model

The regression part of the model is updated by sampling from the full conditional distribution of each parameter or a set of parameters.

Full conditional for the fixed effects $\boldsymbol{\beta}$

Let $\boldsymbol{\beta}_{(S)}$ be an arbitrary sub-vector of vector $\boldsymbol{\beta}$, and $\boldsymbol{x}_{i,l(S)}$ the corresponding sub-vectors of covariate vectors $\boldsymbol{x}_{i,l}$, and further let $\boldsymbol{x}_{i,l(-S)}$ be their complementary sub-vectors. Similarly, let further $\boldsymbol{\nu}_{\beta(S)}$ and $\boldsymbol{\psi}_{\beta(S)}$ be appropriate sub-vectors of hyperparameters $\boldsymbol{\nu}_\beta$ and $\boldsymbol{\psi}_\beta$, respectively. Finally, let $\Psi_{\beta(S)} = \mathrm{diag}(\boldsymbol{\psi}_{\beta(S)})$. Then

$$\boldsymbol{\beta}_{(S)} \mid \cdots \sim \mathcal{N}\Big(\mathrm{E}(\boldsymbol{\beta}_{(S)} \mid \cdots),\ \mathrm{var}(\boldsymbol{\beta}_{(S)} \mid \cdots)\Big),$$

with

$$\mathrm{E}(\boldsymbol{\beta}_{(S)} \mid \cdots) = \mathrm{var}(\boldsymbol{\beta}_{(S)} \mid \cdots) \times \Big\{ \Psi^{-1}_{\beta(S)} \boldsymbol{\nu}_{\beta(S)} + \sum_{i=1}^{N} \sum_{l=1}^{n_i} \sigma^{-2}_{r_{i,l}} \boldsymbol{x}_{i,l(S)} e^{(F)}_{i,l(S)} \Big\},$$

$$\mathrm{var}(\boldsymbol{\beta}_{(S)} \mid \cdots) = \Big(\Psi^{-1}_{\beta(S)} + \sum_{i=1}^{N} \sum_{l=1}^{n_i} \sigma^{-2}_{r_{i,l}} \boldsymbol{x}_{i,l(S)} \boldsymbol{x}'_{i,l(S)} \Big)^{-1},$$

where $e^{(F)}_{i,l(S)} = \log(t_{i,l}) - \mu_{r_{i,l}} - \boldsymbol{\beta}'_{(-S)} \boldsymbol{x}_{i,l(-S)} - \boldsymbol{b}'_i \boldsymbol{z}_{i,l}$.

Full conditional for the means of random effects $\boldsymbol{\gamma}$

There is no loss of generality to assume that $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_{(S)}, \boldsymbol{\gamma}'_{(-S)})'$. Further, let $\boldsymbol{b}_{i(S)}$, $\boldsymbol{b}_{i(-S)}$, $\boldsymbol{\nu}_{\gamma(S)}$, $\boldsymbol{\psi}_{\gamma(S)}$ the corresponding sub-vectors or complementary sub-vectors of indicated quantities and $\Psi_{\gamma(S)} = \mathrm{diag}(\boldsymbol{\psi}_{\gamma(S)})$. Furthermore, let the inversion of the matrix $\mathbb{D}$ be decomposed in the following way

$$\mathbb{D}^{-1} = \begin{pmatrix} \mathbb{V}_{(S)} & \mathbb{V}_{(S,-S)} \\ \mathbb{V}'_{(S,-S)} & \mathbb{V}_{(-S)} \end{pmatrix},$$

then

$$\boldsymbol{\gamma}_{(S)} \mid \cdots \sim \mathrm{N}\Big(\mathrm{E}(\boldsymbol{\gamma}_{(S)} \mid \cdots),\ \mathrm{var}(\boldsymbol{\gamma}_{(S)} \mid \cdots) \Big),$$

with

$$\mathrm{E}(\boldsymbol{\gamma}_{(S)} \mid \cdots) = \mathrm{var}(\boldsymbol{\gamma}_{(S)} \mid \cdots) \times \Big\{ \Psi^{-1}_{\gamma(S)} \boldsymbol{\nu}_{\gamma(S)} + \mathbb{V}_{(S)} \sum_{i=1}^{N} \boldsymbol{b}_{i(S)} + \mathbb{V}_{(S,-S)} \sum_{i=1}^{N} \big(\boldsymbol{b}_{i(-S)} - \boldsymbol{\gamma}_{(-S)} \big) \Big\},$$

$$\mathrm{var}(\boldsymbol{\gamma}_{(S)} \mid \cdots) = \Big(\Psi^{-1}_{\gamma(S)} + N\, \mathbb{V}_{(S)} \Big)^{-1},$$

Full conditional for the random effects $\boldsymbol{b}_i$

For the random effects vectors $\boldsymbol{b}_i$:

$$\boldsymbol{b}_i \mid \cdots \sim \mathcal{N}\Big(\mathrm{E}(\boldsymbol{b}_i \mid \cdots), \mathrm{var}(\boldsymbol{b}_i \mid \cdots) \Big), \qquad i = 1, \ldots, N,$$

with

$$\begin{aligned}\mathrm{E}(\boldsymbol{b}_i \mid \cdots) &= \mathrm{var}(\boldsymbol{b}_i \mid \cdots) \times \\ &\qquad \Big[\mathbb{D}^{-1}\boldsymbol{\gamma} + \sum_{l=1}^{n_i} \sigma_{r_{i,l}}^{-2} \boldsymbol{z}_{i,l}\big\{\log(t_{i,l}) - \mu_{r_{i,l}} - \boldsymbol{\beta}'\boldsymbol{x}_{i,l}\big\}\Big], \\ \mathrm{var}(\boldsymbol{b}_i \mid \cdots) &= \Big(\mathbb{D}^{-1} + \sum_{l=1}^{n_i} \sigma_{r_{i,l}}^{-2} \boldsymbol{z}_{i,l}\boldsymbol{z}_{i,l}'\Big)^{-1}.\end{aligned}$$

Full conditional for the covariance matrix of random effects $\mathbb{D}$

Finally, $\mathbb{D} \mid \cdots$ is an inverse-Wishart distribution with degrees of freedom equal to $df + N$ and a scale matrix equal to

$$\mathbb{S} + \sum_{i=1}^{N} (\boldsymbol{b}_i - \boldsymbol{\gamma})(\boldsymbol{b}_i - \boldsymbol{\gamma})'.$$

8.4 Bayesian estimates of the survival distribution

Simple posterior median or mean are suitable overall estimates for the components of the parameter vector $\boldsymbol{\theta}$. To characterize a survival distribution underlying the data we also need an estimate for the survival and hazard function or for the survival density or the density of the error term in the AFT model. All these quantities are functions with an expression that depends on the parameter vector $\boldsymbol{\theta}$. In the Bayesian statistics they are estimated by the mean of (posterior) predictive quantities to be discussed in this section.

8.4.1 Predictive survival and hazard curves and predictive survival densities

For a specific value of covariates, say $\boldsymbol{x}_{new}$ and $\boldsymbol{z}_{new}$, the predictive survival function is given by

$$S(t \mid \text{data}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) = \int S(t \mid \boldsymbol{\theta}, \text{data}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new})\, p(\boldsymbol{\theta} \mid \text{data})\, d\boldsymbol{\theta}$$

for any $t > 0$. Further, once the parameter vector $\boldsymbol{\theta}$ is known the data do not bring any additional information and hence

$$S(t \mid \boldsymbol{\theta}, \text{data}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) = S(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}).$$

Additionally, analogously to Section 7.4, the quantity $S(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new})$ is expressed using the model parameters as

$$S(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) = 1 - \sum_{j=1}^{K} w_j \Phi\big\{\log(t) - \boldsymbol{\beta}'\boldsymbol{x}_{new} - \boldsymbol{b}'\boldsymbol{z}_{new} \,\big|\, \mu_j, \sigma_j^2\big\}. \quad (8.16)$$

The MCMC estimate of the predictive survival function is then given, using the expression (4.13):

$$\hat{S}(t \mid \text{data}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) = \frac{1}{M} \sum_{m=1}^{M} S(t \mid \boldsymbol{\theta}^{(m)}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}), \quad (8.17)$$

where $\boldsymbol{\theta}^{(m)}$, $m = 1, \ldots, M$ is the MCMC sample from the posterior (predictive) distribution. All components of $\boldsymbol{\theta}^{(m)}$ are directly available except $\boldsymbol{b}^{(m)}$. These must be additionally sampled from $\mathcal{N}_q(\boldsymbol{\gamma}^{(m)}, \mathbb{D}^{(m)})$.

Analogously, predictive hazard curves and predictive survival densities are obtained using the relationship

$$p(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) = t^{-1} \sum_{j=1}^{K} w_j \varphi\big\{\log(t) - \boldsymbol{\beta}'\boldsymbol{x}_{new} - \boldsymbol{b}'\boldsymbol{z}_{new} \,\big|\, \mu_j, \sigma_j^2\big\} \quad (8.18)$$

for the survival density and the relationship

$$\hbar(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new}) = \frac{p(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new})}{S(t \mid \boldsymbol{\theta}, \boldsymbol{x}_{new}, \boldsymbol{z}_{new})} \quad (8.19)$$

for the hazard.

8.4.2 Predictive error densities

Averaging the error density (8.2) across the MCMC run, conditionally on fixed values of K, gives a Bayesian predictive error density estimate of the mixture with K components, i.e. an estimate of

$$\mathrm{E}\big\{g_\varepsilon(e) \,\big|\, K, \text{data}\big\} = \int_{\Theta_K} g_\varepsilon(e)\, p(\boldsymbol{\theta} \mid K, \text{data})\, d\boldsymbol{\theta}, \quad e \in \mathbb{R}, \quad (8.20)$$

where the domain of integration, Θ_K, is the subset of the overall parameter space pertaining to mixtures with a fixed number K of the mixture components.

Averaging further across values of K gives an estimate of

$$\mathrm{E}\big\{g_\varepsilon(e) \,\big|\, \text{data}\big\} = \int g_\varepsilon(e)\, p(\boldsymbol{\theta} \mid \text{data})\, d\boldsymbol{\theta}, \quad e \in \mathbb{R}, \quad (8.21)$$

the overall Bayesian predictive density estimate of the error distribution.

8.5 Bayesian estimates of the individual random effects

In some situations, for example when discrimination between clusters is of interest, an estimate of the individual random effects must be provided. In the Bayesian statistics, their estimates are given by some characteristic of the posterior distribution, for instance by the posterior mean $\mathrm{E}(\boldsymbol{b}_i \,|\, \text{data})$. The precision of the estimate can be evaluated using the credible interval.

When using the MCMC to draw the sample from the posterior distribution, we estimate each individual random effect vector $\boldsymbol{b}_i$ by the average of the sampled values, i.e.

$$\hat{\boldsymbol{b}}_i = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{b}_i^{(m)},$$

where M is the number of MCMC iterations and $\boldsymbol{b}_i^{(m)}$ the value of $\boldsymbol{b}_i$ sampled at the mth iteration. The credible interval is obtained by taking sample quantiles from the MCMC sample.

8.6 Simulation study

A simulation study was carried out to explore the performance of the proposed method. The setting mimics a study with clustered data where a continuous covariate as well as a dichotomous covariate might influence the distribution of the event time. At the same time there might be an overall heterogeneity between clusters present as well as a possible interaction between the cluster effect and the effect of the dichotomous covariate. The factual setting used to generate the 'true' data was motivated by the results of the WIHS analysis presented in Section 7.6.

Namely, 'true' uncensored data were generated according to the model

$$\log(T_{i,l}) = 1.5 + \beta\, x_{i,l} + b_{i,1} + b_{i,2}\, z_{i,l} + \varepsilon_{i,l}, \quad i = 1, \ldots, N,\ l = 1, \ldots, n_i,$$

where

$$\begin{aligned} &\beta = 0.4, && (b_{i,1},\, b_{i,2})' \sim \mathcal{N}_2\big((0, \gamma)', \mathbb{D}\big), && \\ &\gamma = -0.8, && \mathrm{var}(b_{i,1}) = 0.5^2, && \mathrm{var}(b_{i,2}) = 0.1^2, \\ & && \mathrm{corr}(b_{i,1}, b_{i,2}) = 0.4. && \end{aligned}$$

The covariate $x_{i,l}$ was generated according to the extreme-value distribution of a minimum, with location equal to 8.5 and scale equal to 1 inspired more

or less by the $\log_2(1 + \text{CD4 count})$ covariate in the WIHS data set. The covariate $z_{i,l}$ was binary taking a value of 1 with probability equal to 0.4. The error term $\varepsilon_{i,l}$ was generated from a standard normal distribution, from a Cauchy distribution, from a Student t_2 distribution, from a standardized extreme value distribution, and from a normal mixture $0.4\mathcal{N}_1(-2.000,\ 0.25) + 0.6\mathcal{N}_1(1.333,\ 0.36)$, respectively. Two sample sizes were considered: (1) $N = 50$, $n_i = 5$ for all i (small sample size) and (2) $N = 100$, $n_i = 10$ for all i (large sample size). Each simulation involved 100 replications.

All event times were interval-censored by simulating 120 consecutive 'assessment times' for each 'patient' in the dataset (the first assessment time was drawn from $\mathcal{N}(7,\ 1)$, times between each consecutive assessments from $\mathcal{N}(6, 0.25)$). At each assessment, between 0.2% and 0.6% randomly selected patients were withdrawn from the study resulting in approximately 15% of right-censored observations. For each dataset, the estimates were computed using the Bayesian normal mixture cluster-specific AFT model, using the Bayesian cluster-specific model with a normal error and using the maximum-likelihood AFT model with a normal error and ignoring the random effects structure.

Appendix B, Section B.2 gives selected results of the simulation. Average estimates of the regression parameters, their standard and mean squared errors are given in Tables B.7 and B.8. The results related to the covariance matrix $\mathbb{D}$ of the random effects are given in Tables B.9 – B.11. It is seen that, in most cases, the Bayesian mixture approach performs better than the incorrectly specified models. A large difference in favour of the Bayesian mixture model is seen in the case of a normal mixture or Cauchy for the error distribution.

Additionally, when the Bayesian mixture approach is used, also the error distribution and consequently also the hazard or survival functions are reproduced closely which is not always the case when the Bayesian normal model is used. See Figures B.4 – B.9.

8.7 Example: Signal Tandmobiel® study – clustered interval-censored data

In Section 7.7 we analyzed separately the emergence times of teeth 14 and 15. In this section, we extend this analysis by inclusion of all permanent premolars, i.e. teeth 14, 15, 24, 25, 34, 35, 44, 45 in Figure 1.1 and additionally, all eight teeth will be analyzed jointly. This allows not only to answer the question what the impact of different covariates on the emergence time is but

also the question concerning the relationship between the emergence times of different teeth. A random sample of 500 boys and 500 girls will be used for the inference.

The response variable $T_{i,l}$, $i = 1, \ldots, 1\,000$, $l = 1, \ldots, 8$, refers to the age of emergence of the lth permanent premolar of the ith child. As indicated in Sections 1.1 and 7.7 the response variable is interval-censored with intervals of length equal to approximately 1 year. For reasons stated in Section 7.7 we shifted the time origin of the AFT model to 5 years of age, i.e. by replacing $T_{i,l}$ by $T_{i,l} - 5$ in the model (8.1).

Further, Leroy et al. (2003b) have shown that there is *horizontal symmetry* with respect to emergence, i.e. the same emergence distribution can be assumed at horizontally symmetric positions (e.g., for teeth 14 and 24). In model (8.1), this leads to the random effect vector

$$\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,4})' \quad \text{with } \boldsymbol{z}_{i,l} = (1, \mathsf{man4}_{i,l}, \mathsf{max5}_{i,l}, \mathsf{man5}_{i,l})',$$

where $\mathsf{man4}_{i,l}, \mathsf{max5}_{i,l}, \mathsf{man5}_{i,l}$, respectively are dummies for the mandibular first premolars (teeth 34, 44), maxillary second premolars (teeth 15, 25) and mandibular second premolars (teeth 35, 45), respectively. With this model specification, apart of the random variation given by the error term $\varepsilon_{i,l}$, the terms

$$\begin{aligned} b^*_{i,max4} &= b_{i,1}, & b^*_{i,man4} &= b_{i,1} + b_{i,2}, \\ b^*_{i,max5} &= b_{i,1} + b_{i,3}, & b^*_{i,man5} &= b_{i,1} + b_{i,4} \end{aligned}$$

determine how the log-emergence time of a pair of horizontally symmetric teeth of a single child differ from the population average. As fixed effects we used $\mathsf{gender} \equiv$ girl, dmf, interaction between gender and dmf, and all two-way interaction terms between gender, dmf and dummies for the pairs of horizontal symmetric teeth, i.e.

$$\begin{aligned} \boldsymbol{x}_{i,l} = (&\mathsf{gender}_i, \mathsf{dmf}_{i,l}, \mathsf{gender}_i * \mathsf{dmf}_{i,l}, \\ &\mathsf{gender}_i * \mathsf{man4}_{i,l}, \mathsf{gender}_i * \mathsf{max5}_{i,l}, \mathsf{gender}_i * \mathsf{man5}_{i,l}, \\ &\mathsf{dmf}_{i,l} * \mathsf{man4}_{i,l}, \mathsf{dmf}_{i,l} * \mathsf{max5}_{i,l}, \mathsf{dmf}_{i,l} * \mathsf{man5}_{i,l})'. \end{aligned}$$

See Section 7.7 for the definition of the covariate dmf.

For the inference we sampled two chains, each of length 20 000 with 1:3 thinning which took about 27 hours on a Pentium IV 2 GHz PC with 512 MB RAM. The first 1 500 iterations of each chain were discarded. Convergence was evaluated by the method of Gelman and Rubin (1992).

Table 8.1: Signal Tandmobiel® study. Posterior medians, 95% equal-tail credible intervals for the effect of different covariates and error variance.

Effect	Posterior median	95% CI	Posterior median	95% CI
	Maxilla 4		Maxilla 5	
intercept	1.7566	(1.7338, 1.7822)	1.9001	(1.8729, 1.9283)
gender	−0.0680	(−0.1003, −0.0368)	−0.0504	(−0.0844, −0.0163)
dmf	−0.0457	(−0.0631, −0.0284)	−0.0317	(−0.0500, −0.0135)
	Mandible 4		Mandible 5	
intercept	1.7242	(1.7019, 1.7484)	1.9060	(1.8805, 1.9323)
gender	−0.0668	(−0.0972, −0.0375)	−0.0654	(−0.0965, −0.0323)
dmf	−0.0201	(−0.0378, −0.0032)	−0.0090	(−0.0283, 0.0098)

Effect	Posterior median	95% CI
	All teeth	
gender ∗ dmf	0.0105	(−0.0073, 0.0279)
log(scale) $\log(\sigma)$	−2.2580	(−2.3111, −2.1721)
error scale σ	0.1046	(0.0992, 0.1139)

8.7.1 Prior distribution

The initial maximum-likelihood AFT model, for each tooth separately, with a normal error distribution and without random effects estimated the intercept as 1.8 and scale as 0.25. According to the suggestions of Section 8.2.3 we used the following values of hyperparameters: $\xi = 1.8$, $\kappa = (3 \cdot 0.25)^2$, $\zeta = 2$, $h_1 = 0.2$, $h_2 = 0.1$, $\delta = 1$. For the number of mixture components, K, a truncated Poisson prior with $\lambda = 5$ reflecting our prior belief that the error distribution is skewed and $K_{max} = 30$ was used. All β and γ parameters were assigned a spread $\mathcal{N}(0, 100)$ prior. For the covariance matrix $\mathbb{D}$ of random effects we used an inverse Wishart prior with $df = 4$. Though, due to the fact that 1 000 clusters are involved in the data set, even a higher value could be used with a negligible impact on results. Prior scale matrix $\mathbb{S}$ was equal to diag(0.002) (corresponding to inverse-gamma$(df, 0.001)$ in the univariate case).

Table 8.2: Signal Tandmobiel® study. Posterior medians, 95% equal-tail credible intervals and Bayesian two-sided p-values for the effect of dmf > 0 for the two genders and different teeth.

Tooth	Gender	Post. median	95% CI	p-value
Maxilla 4	Girl	-0.0352	$(-0.0522, -0.0185)$	< 0.001
	Boy	-0.0457	$(-0.0631, -0.0284)$	< 0.001
Maxilla 5	Girl	-0.0212	$(-0.0390, -0.0035)$	0.019
	Boy	-0.0317	$(-0.0500, -0.0135)$	< 0.001
Mandible 4	Girl	-0.0098	$(-0.0267, 0.0070)$	0.255
	Boy	-0.0201	$(-0.0378, -0.0032)$	0.021
Mandible 5	Girl	0.0015	$(-0.0162, 0.0193)$	0.870
	Boy	-0.0090	$(-0.0283, 0.0098)$	0.353

8.7.2 Results for the regression and error parameters

The effect of different covariates on the emergence, separately for each tooth is given in Table 8.1. The results in Table 8.1 were obtained as MCMC summary for proper combinations of model parameters. For example, the intercept effect for the maxillary teeth 4 equals the error mean $\alpha = \sum_{j=1}^{K} w_j \mu_j$. For the maxillary teeth 5, the intercept effect equals $\alpha+\gamma(\mathsf{max5})$ where $\gamma(\mathsf{max5})$ is the mean of the random effect $b_{i,3}$. The intercept effects for the remainig teeth are defined in an analogous manner. The effect of gender in Table 8.1 is defined as $\beta(\mathsf{gender})$ for the maxillary teeth 4, $\beta(\mathsf{gender}) + \beta(\mathsf{gender} * \mathsf{max5})$ for the maxillary teeth 5 and analogously for the remaining teeth. Finally, the effect of dmf is given by $\beta(\mathsf{dmf})$ for the maxillary teeth 4, by $\beta(\mathsf{dmf})+\beta(\mathsf{dmf}*\mathsf{max5})$ for the maxillary teeth 5 and analogously for remaining teeth. The error scale refers to the summary for the standard deviation σ of the error distribution, i.e. $\sigma = \sqrt{\sum_{j=1}^{K} w_j(\mu_j^2 + \sigma_j^2) - \alpha^2}$. The row labeled as log(scale) refers to the summary for $\log(\sigma)$.

Most of the quantities in Table 8.1 are comparable to the results of the earlier analysis (see Section 7.7) given in Table 7.5. Remember however that in Section 7.7 we analyzed separately only one maxillary tooth 4 (14) and one maxillary tooth 5 (15). Furthermore, in contrast to the recent analysis we allowed the dependence of the error variance on the covariates in Section 7.7.

In this analysis, the main interest lies in the effect of dmf on emergence. This can be evaluated from Table 8.2 that shows posterior summary statistics for the effect of dmf (appropriate linear combinations of β parameters) for boys

Table 8.3: Signal Tandmobiel® study. Posterior medians, 95% equal-tail credible intervals for variances and correlations between tooth-specific linear combinations of random effects.

Parameter	Posterior median	95% CI
$\mathrm{sd}(b^*_{i,max4})$	0.204	(0.192, 0.218)
$\mathrm{sd}(b^*_{i,man4})$	0.198	(0.186, 0.211)
$\mathrm{sd}(b^*_{i,max5})$	0.205	(0.190, 0.221)
$\mathrm{sd}(b^*_{i,man5})$	0.202	(0.187, 0.218)
$\mathrm{corr}(b^*_{max4}, b^*_{man4})$	0.887	(0.856, 0.914)
$\mathrm{corr}(b^*_{max4}, b^*_{max5})$	0.914	(0.887, 0.938)
$\mathrm{corr}(b^*_{max4}, b^*_{man5})$	0.842	(0.804, 0.874)
$\mathrm{corr}(b^*_{man4}, b^*_{max5})$	0.793	(0.749, 0.832)
$\mathrm{corr}(b^*_{man4}, b^*_{man5})$	0.895	(0.864, 0.923)
$\mathrm{corr}(b^*_{max5}, b^*_{man5})$	0.847	(0.810, 0.880)

and girls and the four pairs of horizontally symmetric teeth. It is seen that caries on the primary predecessor accelerates significantly the emergence of the permanent successor in the case of maxillary teeth. For the mandibular teeth, a slight effect is observed only for the first premolar on boys. Additionally, besides the effect of dmf the emergence process for girls is ahead of boys.

8.7.3 Inter-teeth relationship

Further, Table 8.3 shows posterior summary statistics for standard deviations and correlations of above defined tooth-specific linear combinations $b^*_{i,max4}$, $b^*_{i,man4}$, $b^*_{i,max5}$, $b^*_{i,man5}$ of random effects $b_{i,1}, \ldots, b_{i,4}$. It shows how the child effect is important and how the different teeth in one mouth are strongly correlated. The posterior medians of all standard deviations in Table 8.3 are all about 0.2 which is approximately two times higher than the posterior median of the standard deviation of the error distribution which was equal to 0.1. Posterior medians of all correlation parameters lie between 0.79 and 0.91.

8.7.4 Predictive emergence and hazard curves

Predictive emergence curves (predictive cumulative distribution functions) computed using an approach described in Section 8.4.1 are shown in Fig-

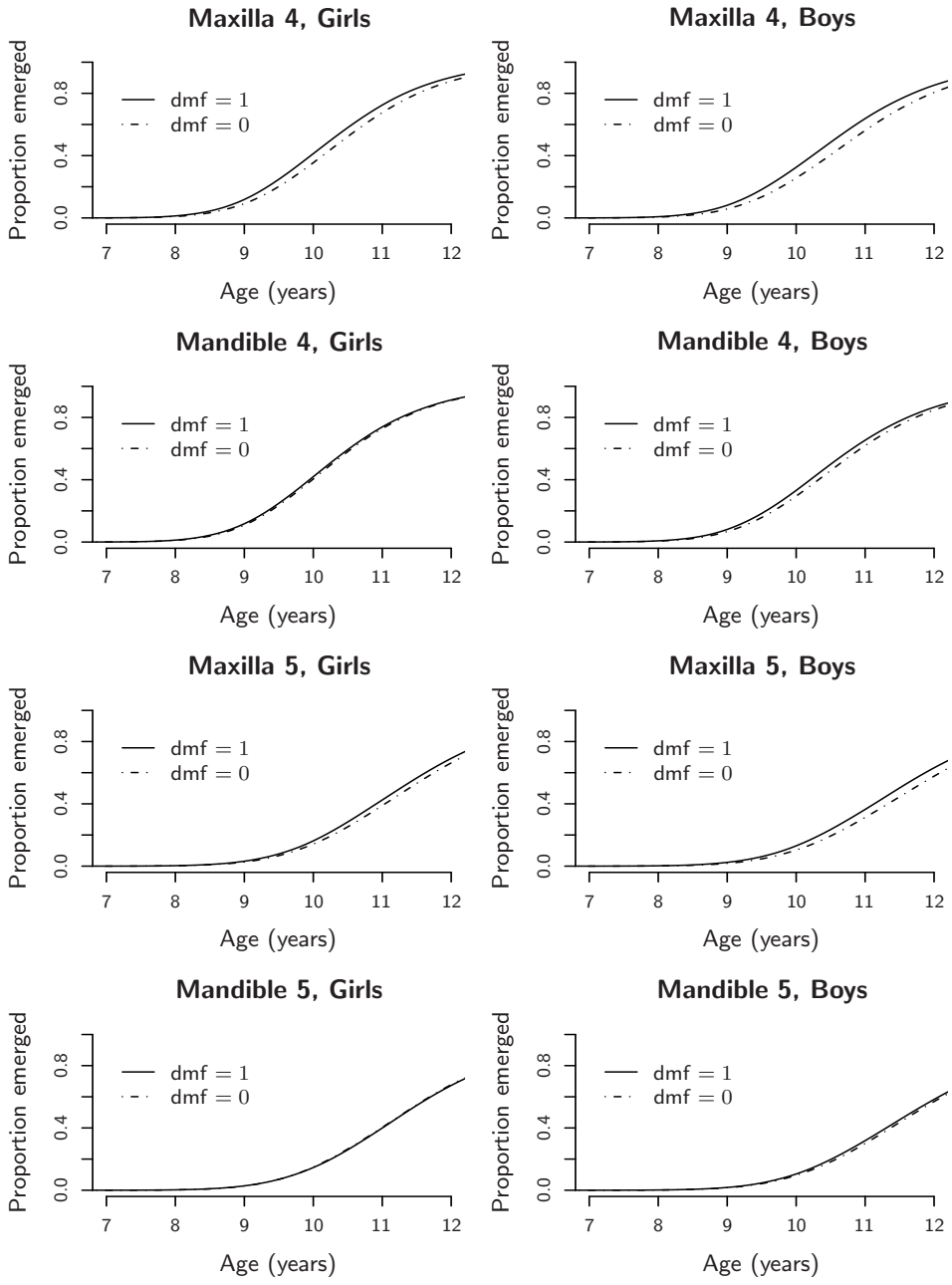


Figure 8.2: Signal Tandmobiel® study. Posterior predictive emergence curves. Solid line: dmf = 1, dotted-dashed line: dmf = 0.

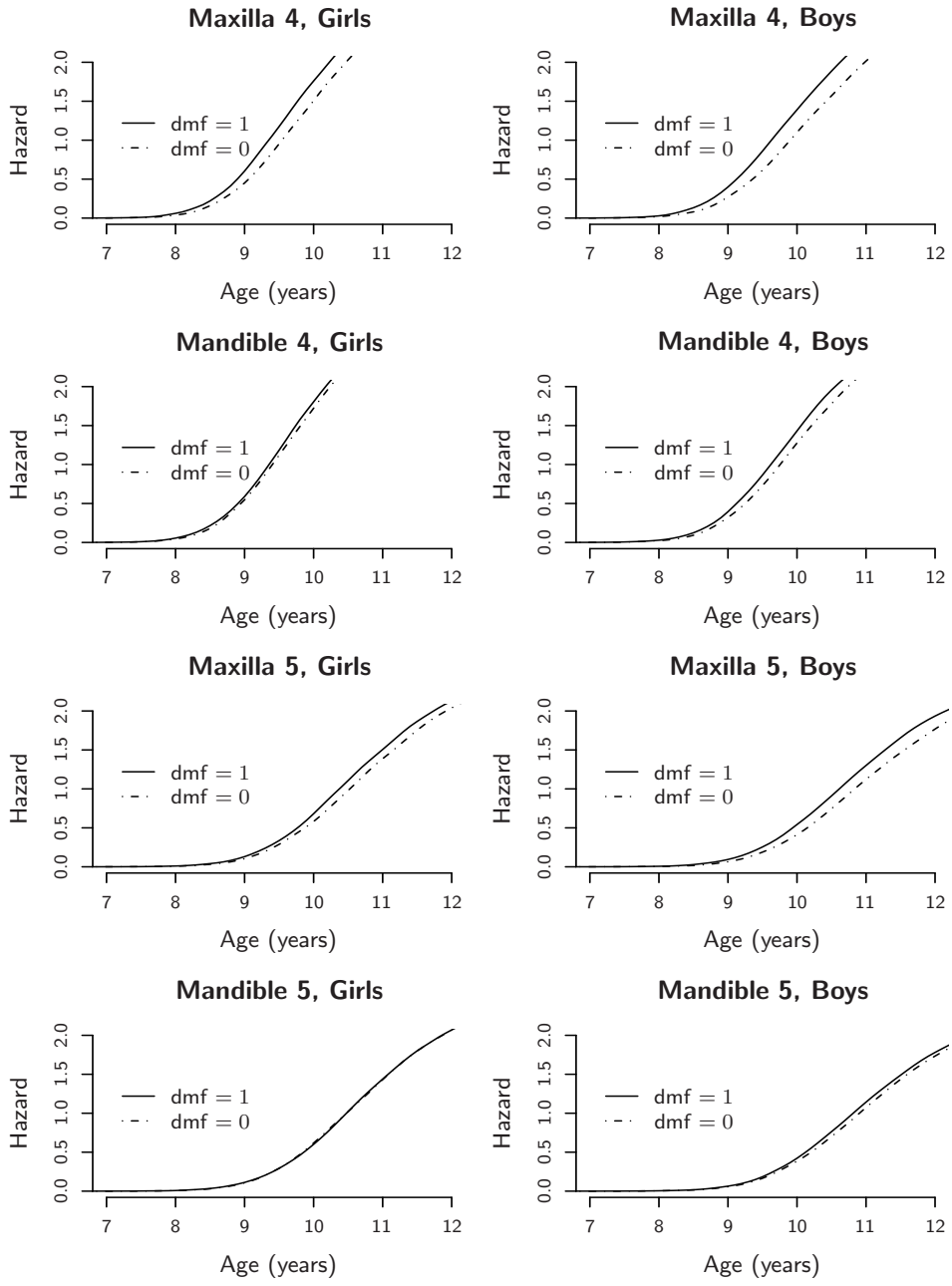


Figure 8.3: Signal Tandmobiel® study. Posterior predictive hazard curves. Solid line: dmf $= 1$, dotted-dashed line: dmf $= 0$.

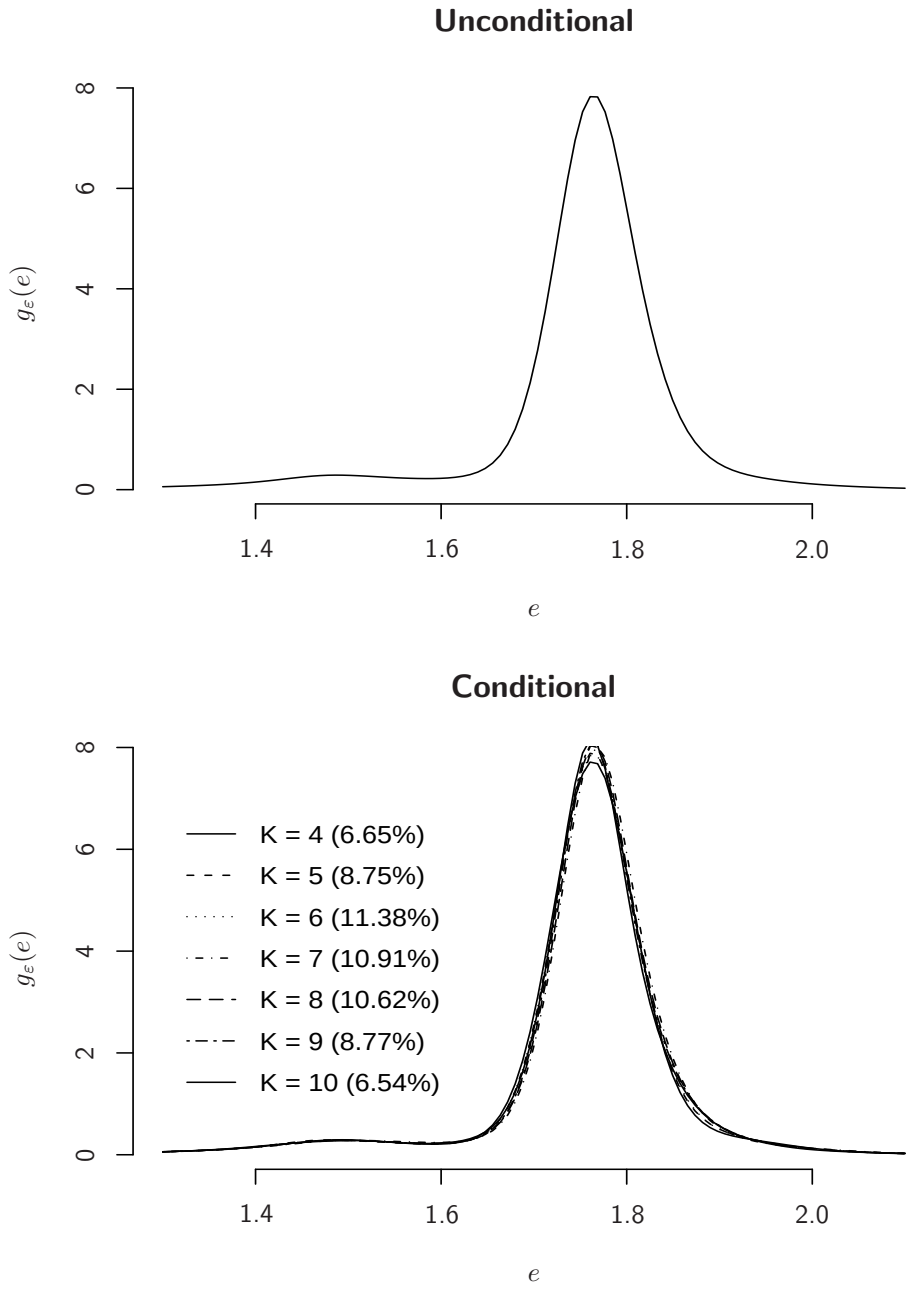


Figure 8.4: Signal Tandmobiel® study. Posterior predictive error density.

ure 8.2. In agreement with the results discussed in Section 8.7.2 almost negligible difference is observed between the predictive emergence curves for $\mathsf{dmf} > 0$ and $\mathsf{dmf} = 0$ for mandibular teeth. The same is true for the predictive hazard functions of emergence shown in Figure 8.3. As expected (see Section 7.7.2 for the reasons why) the predictive hazard functions are all increasing.

8.7.5 Predictive error density

In our sample, the number of mixture components K ranged from 2 to 24 while mixtures with $K \in \{6, 7, 8\}$ occupied each more than 10% of the sample, with the highest frequency for $K = 7$ (11.2%). Mixtures with $K \geq 17$ took each less than 1.5% of the sample. Apparently, the model did not suffer from the technical restriction given by $K_{max} = 30$.

Figure 8.4 further shows both the overall estimate of the predictive error density (8.21) and the conditional (given K) estimate of the predictive error density (8.20). It is seen that the mixtures with the most frequent numbers of components are all almost the same.

8.7.6 Conclusions

This section showed an analysis of clustered data where moreover closer dependence between some observations within the cluster could be assumed. Since in Section 7.7 we have shown on the similar analysis of the same data set that the error variance might depend on covariates the model presented in this section might be improved if we allow to depend the variances of the mixture components determining the error distribution on covariates as well. However, in the current mixture setting such extension is not trivial and requires further research.

8.8 Example: CGD data – recurrent events analysis

The chronic granulomatous disease (CGD) trial has been introduced in Section 1.2. The response variable $T_{i,l}$ is a time to the lth (recurrent) infection on the ith patient, $i = 1, \ldots, 128$, $l = 1, \ldots, n_i$, $1 \leq n_i \leq 8$. So that a patient represents a cluster and the infection times the individual observations.

The problem of recurrent events in this data set was discussed by several authors in the literature. Among others, Therneau and Hamilton (1997) used

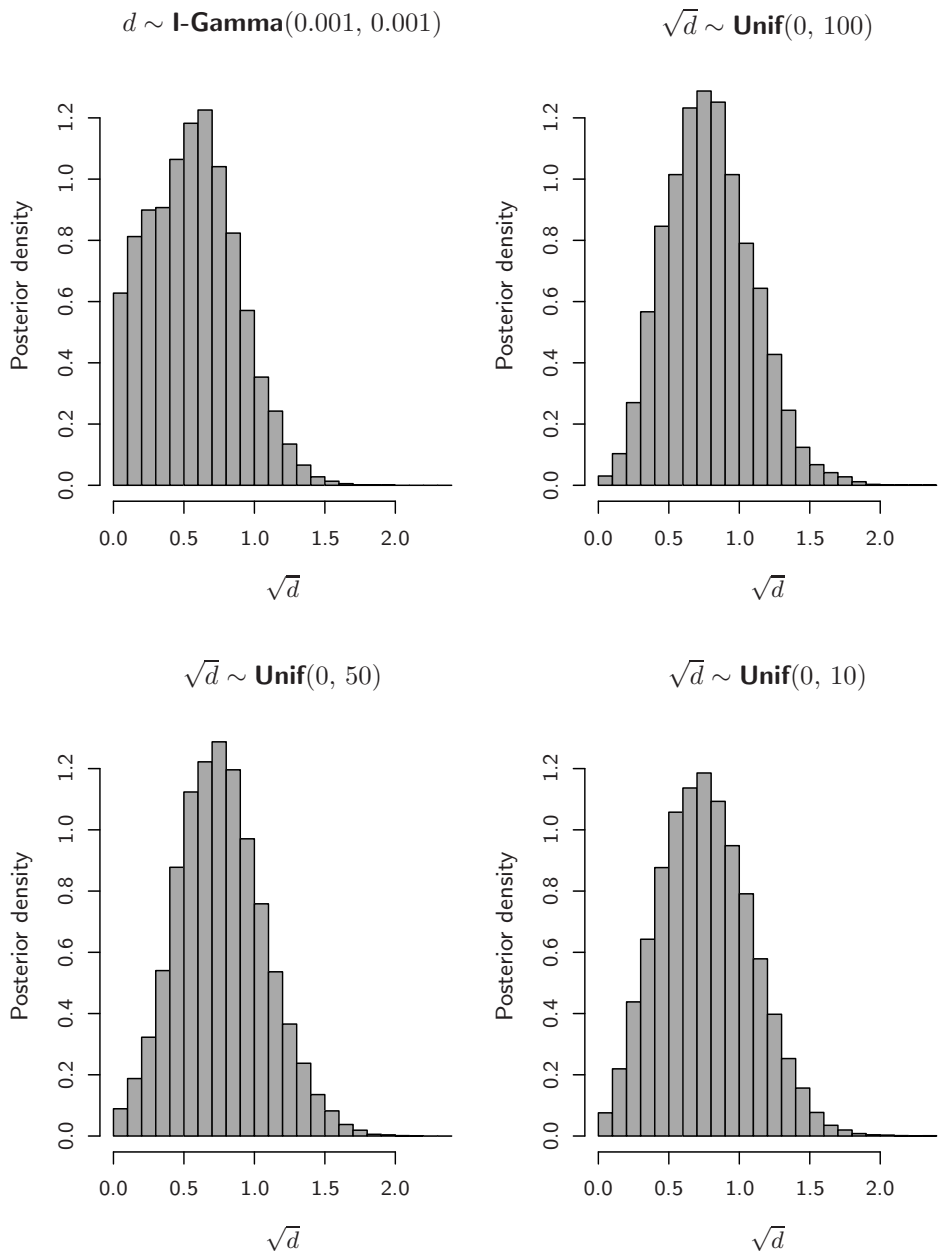


Figure 8.5: CGD data. Scaled histograms of sampled standard deviations of the random effect b_i for different prior distributions.

the CGD data to illustrate several approaches for recurrent event analysis based on the Cox's PH model. Vaida and Xu (2000) used this dataset to illustrate the PH model with random effects. They specify the hazard function for the $(i,l)th$ event as

$$\hbar(t \,|\, \boldsymbol{x}_{i,l},\, \boldsymbol{z}_{i,l},\, \boldsymbol{b}_i) = \hbar_0(t)\ \exp(\boldsymbol{\beta}'\boldsymbol{x}_{i,l} + \boldsymbol{b}_i'\boldsymbol{z}_{i,l}),$$

where $\hbar_0$ is a baseline hazard function, $\boldsymbol{\beta}$ regression parameters vector for 'fixed' effects, $\boldsymbol{x}$ a covariate vector of 'fixed' effects, $\boldsymbol{b}_i$ a random effect vector and $\boldsymbol{z}_{i,l}$ corresponding covariates, see also Section 3.4.1. They use a normal distribution for $\boldsymbol{b}_i$.

In this section we present an analysis of the CGD data using the Bayesian CS normal mixture AFT model that could be considered as an AFT counterpart of the random effects PH model of Vaida and Xu (2000). In the model formula (8.1) a univariate random effect b_i is used with $z_{i,l} \equiv 1$. As fixed effects covariates we used the same covariates as Vaida and Xu (2000), namely the $\boldsymbol{x}_{i,l}$ vector equals

$$\begin{aligned}\boldsymbol{x}_{i,l} = (&\mathsf{trtmt}_i,\, \mathsf{inher}_i,\, \mathsf{age}_i,\, \mathsf{cortic}_i,\, \mathsf{prophy}_i,\, \mathsf{gender}_i,\\ &\mathsf{hcatUSother}_i,\, \mathsf{hcatEUAmster}_i,\, \mathsf{hcatEUother}_i)',\end{aligned}$$

where trtmt equals 1 for the gamma inferon group and equals 0 for the placebo group, inher equals 1 for patients with the autosomal recessive and equals 0 for patients with X-linked pattern of inheritance, age is the age of the patient in years, cortic equals 1 if the corticosteroids are used and equals 0 otherwise, prophy equals 1 if the prophylactic antibiotics are used and equals 0 otherwise, gender equals to 1 for females and equals 0 for males and finally hcatUSother, hcatEUAmster, and hcatEUother are dummies for the hospital categories *US–other*, *EU-Amsterdam*, and *EU-other*, respectively.

For the inference we sampled two chains, each of length 60 000 with 1:6 thinning which took about 5 minutes on a Pentium IV 2 GHz PC with 512 MB RAM. The first 30 000 iterations of each chain were discarded. The convergence was evaluated by a critical examination of the trace and autocorrelation plots and using the method of Gelman and Rubin (1992).

8.8.1 Prior distribution

The initial maximum-likelihood AFT model with a normal error distribution and without random effects gave an estimate of the intercept equal to 3.66 and a scale equal to 1.69. Along the suggestions made in Section 8.2.3 we used the following values of hyperparameters: $\xi = 3.66$, $\kappa = 25 \approx (3 \cdot 1.69)^2$,

$\zeta = 2$, $h_1 = 0.2$, $h_2 = 0.1$, $\delta = 1$. For the number of mixture components, K, a truncated Poisson prior with $\lambda = 5$ reflecting our prior belief that the error distribution is skewed and $K_{max} = 30$ was used. Prior means of all regression parameters were equal to 0 and their prior variances to 1 000.

For the variance d of the random effect we tried either an inverse-gamma I-Gamma(0.001, 0.001) prior ($df = 0.002$, $s = 0.002$ in the terms of the inverse-Wishart distribution) or a uniform Unif$(0, \sqrt{s})$ prior on $\sqrt{d}$ with $\sqrt{s} = 100$, 50, 10. As discussed in Gelman (2006, Sections 2.2 and 4.3), with the I-Gamma(ϵ, ϵ) prior the inference might become very sensitive to the choices of ϵ. This is not the case of the uniform distribution on $\sqrt{d}$ where the choice of the range of the uniform distribution has practically no impact on the results (provided the upper limit of the uniform distribution is not chosen too small). In Figure 8.5, we show scaled histograms of sampled values of $\sqrt{d}$ for above mentioned prior distributions. It is seen that the inverse-gamma prior leads to a high posterior probability mass close to zero. The phenomenon driven by the prior distribution which has a peak close to zero. On the other hand, with the uniform prior on $\sqrt{d}$, the posterior distribution is clearly separated from zero with the region of the support obviously driven by the data. Moreover, in agreement with the findings of Gelman (2006), the posterior distribution is practically the same irrespective the choice of the range of the uniform prior. The results presented below will be based on Unif(0, 100) prior on $\sqrt{d}$ (practically the same results were obtained also with the remaining uniform priors on $\sqrt{d}$).

8.8.2 Effect of covariates on the time to infection

Table 8.4 shows posterior summary statistics for the effect of the included covariates on the distribution of the time to infection. Reported Bayesian p-value is simultaneous in the case of the covariate *hospital category*. It is seen that the effect of *gamma interferon* is highly significant increasing the time to the infection by the factor of $\exp(1.273) = 3.57$. The effect of the *pattern of inheritance* is slightly not-significant on a conventional 5% level. On the other hand, the increase of age by 1 year increases significantly the infection free time by the factor of $\exp(0.047) = 1.05$. Further, the use of *corticosteroids* should be avoided as it decreases significantly the infection free time by the factor of $\exp(-2.767) = 0.06$ whereas the use of *prophylactic antibiotics* increases significantly the infection free time by the factor of $\exp(1.191) = 3.29$. The infection free time is further significantly higher for *females*, being $\exp(1.476) = 4.38$ times higher than in the case of *males*. Finally, the effect of the *hospital category* is slightly not significant however

Table 8.4: CGD data. Posterior medians, 95% equal-tail credible intervals and Bayesian two-sided (simultaneous) p-values for the effect of covariates.

Parameter	Posterior median	95% CI
Treatment group		$p = 0.001$
gamma interferon	1.273	(0.437, 2.195)
Pattern of inheritance		$p = 0.067$
autosomal recessive	−0.914	(−1.829, 0.071)
Age		$p = 0.022$
	0.047	(0.007, 0.092)
Use of corticosteroids		$p = 0.038$
yes	−2.767	(−5.727, −0.161)
Use of prophylactic antibiotics		$p = 0.023$
yes	1.191	(0.150, 2.330)
Gender		$p = 0.042$
female	1.476	(0.050, 3.111)
Hospital category		$p = 0.065$
US – other	0.461	(−0.481, 1.451)
Europe – Amsterdam	1.729	(0.183, 3.377)
Europe – other	1.268	(0.017, 2.637)

Table 8.5: CGD data. Posterior medians and 95% equal-tail credible intervals for the moments of the error distribution and standard deviation of the random effects.

Parameter	Posterior median	95% CI
Moments of the error distribution		
Intercept α	4.088	(2.532, 5.527)
Error scale σ	2.495	(1.399, 4.083)
Standard deviation of the random effects		
$\mathrm{sd}(b_i)$	0.748	(0.183, 1.395)

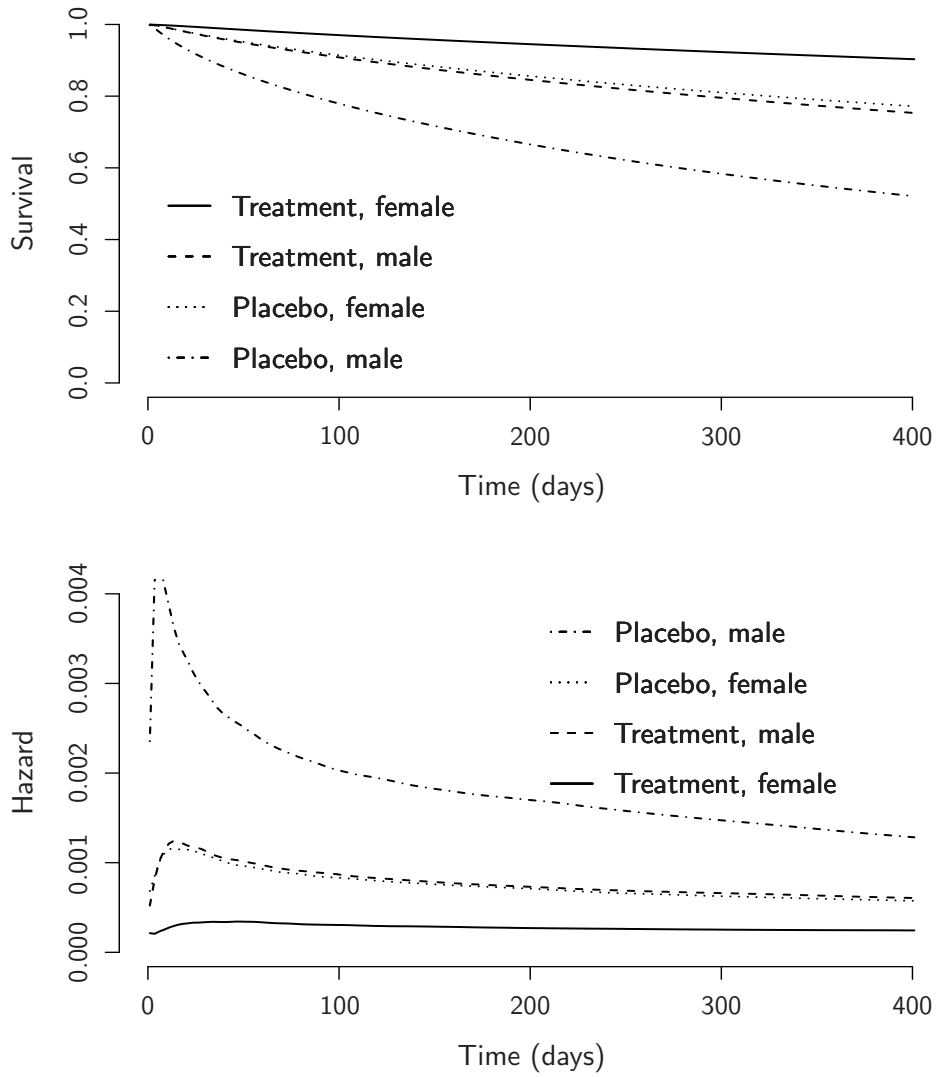


Figure 8.6: CGD data. Predictive survival (upper panel) and hazard (lower panel) curves for males and females taking either treatment or placebo. Remaining covariates were fixed either to the mean value (age = 14.6) or to the most common value (*X-linked* pattern of inheritance, *no use* of corticosteroids, *use* of prophylactic corticosteroids, and a hospital category *US-other*).

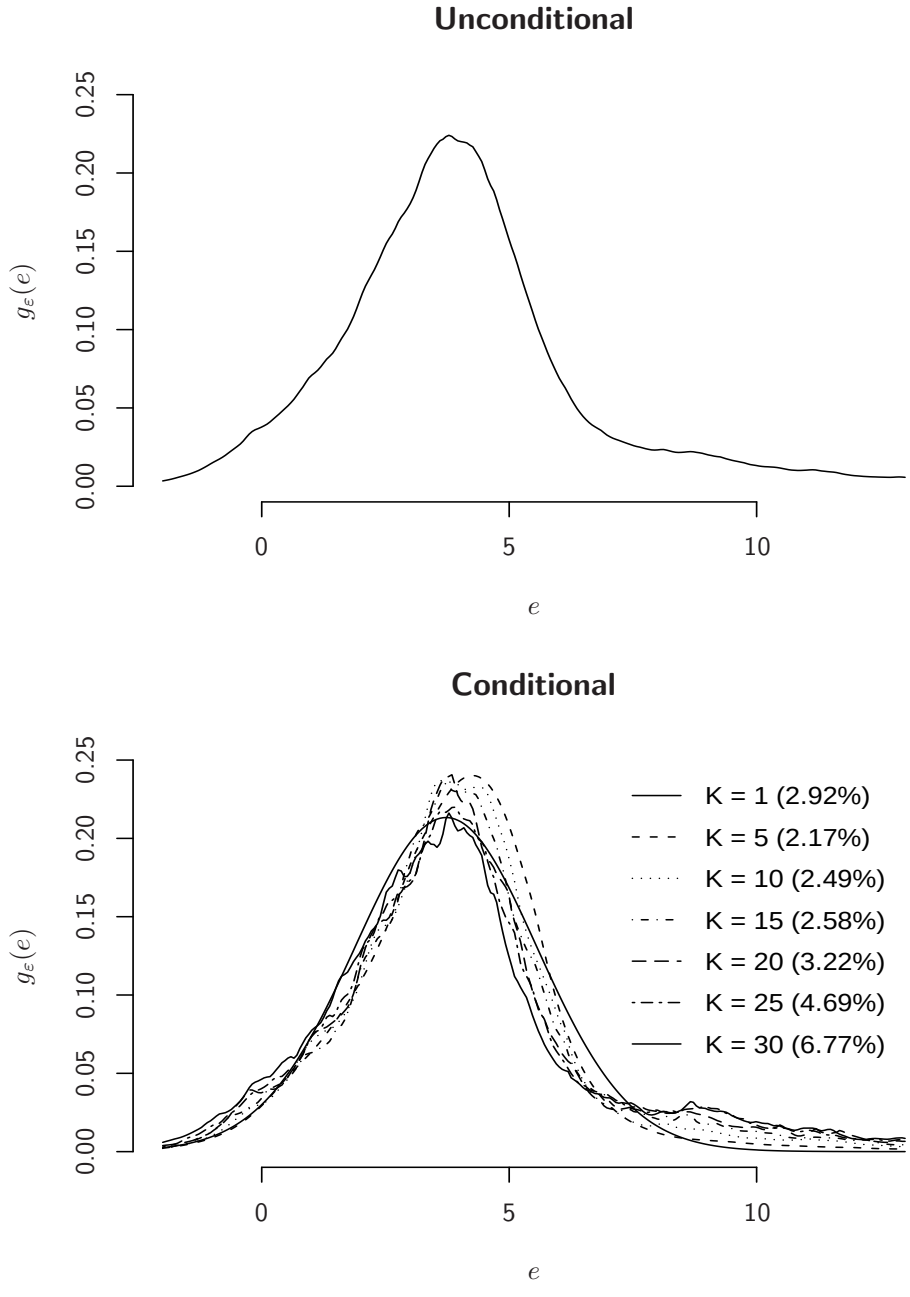


Figure 8.7: CGD data. Posterior predictive error density.

the posterior median suggests the best results are obtained in the hospitals of category *Europe – Amsterdam* whereas the worst results in the hospital category *US – NIH*.

Although the parameters of the AFT model are not directly comparable to the parameters of the PH model, we can compare at least the direction of the relationship obtained here and by Vaida and Xu (2000) who used the PH model. Care must be taken as Vaida and Xu (2000) use different 0-1 coding of dichotomous variables than we do. However, we conclude that the directions of the relationships between the covariates and the time to infection found by the AFT model is the same compared to the findings obtained using the PH model.

The effect of the treatment (*gamma interferon*) is seen also in Figure 8.6 where we plot predictive survival and hazard curves for *males* and *females* taking either *gamma interferon* or *placebo*. Remaining covariates were fixed either to their mean or the most common value.

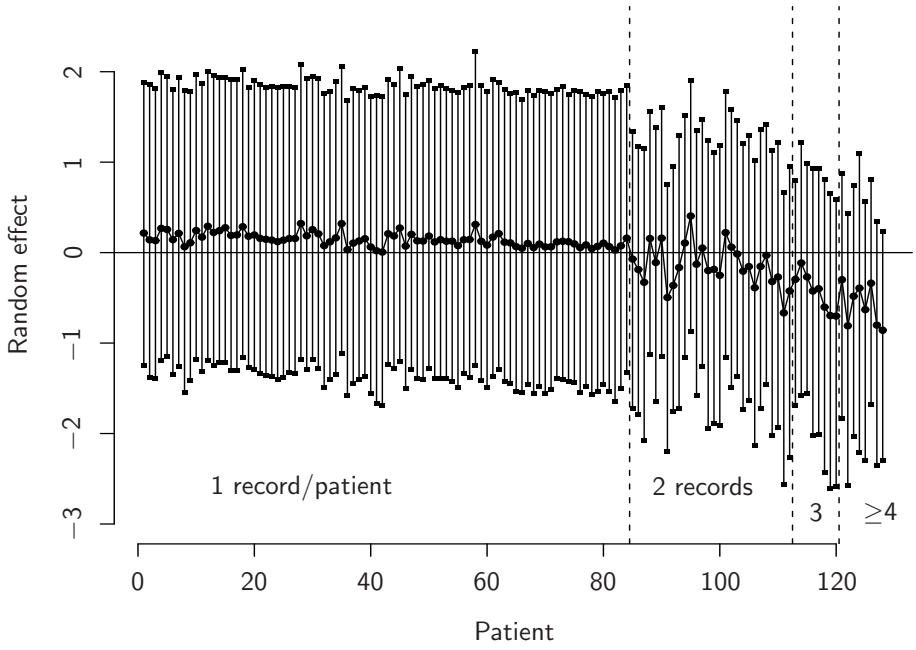


Figure 8.8: CGD data. Posterior means and 95% equal-tail credible intervals for individual random effects. Patients are sorted according to the number of records in the data set.

8.8.3 Predictive error density and variability of random effects

Posterior summary statistics for the moments of the error distribution, computed in the same way as indicated in Section 8.7.2, and for the standard deviation of the random effects are given in Table 8.5.

The estimate of the error density is given in Figure 8.7. In this case, also mixtures with a high number of components were quite highly represented in the sample. For a clarity, the conditional estimates of the error density (given K) are plotted only for chosen values of K. Higher number of components is needed firstly because of clear skewness of the error density and secondly because of somewhat higher probability mass in the right tail of the density.

8.8.4 Estimates of individual random effects

Figure 8.8 shows posterior means and 95% equal-tail posterior credible intervals for the values of individual random effects b_i, $i = 1, \ldots, 128$. For the purpose of plotting, the patients were sorted according to the number of records they have in the data set. Since there are no big differences in the follow-up times for different patients, less records in the data set generally implies longer infection-free periods. Indeed, for the patients with only one recorded infection time practically all estimated individual random effects lie above zero, the mean for b_i. Furthermore, there can be observed a decreasing trend in the estimated individual random effects as the number of recorded infection times increases.

8.8.5 Conclusions

In this section we have shown how the Bayesian normal mixture CS AFT model can be used to analyse recurrent events data. It might be useful to include the covariate *number of infections* in the model. However, such covariate would be time-dependent and it is not possible to include covariates of this type in any model where the (baseline) survival distribution is modelled via density and not hazard function.

8.9 Example: EBCP data – multicenter study

In Section 1.4 we have introduced a multicenter randomized clinical trial aiming to evaluate the effect of perioperative chemotherapy given besides

the surgery on the progression-free survival (PFS) time compared to surgery alone for early breast cancer patients while controlling for several baseline covariates.

In Figure 1.3 we have indicated there possibly exists heterogeneity between centra with respect to the PFS distribution. Additionally, there is some evidence for the heterogeneity with respect to the treatment effect. In this section, we perform an analysis using the Bayesian normal mixture cluster-specific AFT model that addresses all these issues.

The cluster is represented by the center, i.e. $i = 1, \ldots, 14$, within the ith center n_i patients were involved in the trial with $25 \leq n_i \leq 902$. As response $T_{i,l}$, $i = 1, \ldots, 14$, $l = 1, \ldots, n_i$ we use the PFS time in days of the lth patient treated by the ith center.

To allow for the baseline heterogeneity across centra and also for the heterogeneity with respect to the treatment effect we include a bivariate random effect $\boldsymbol{b}_i = (b_{i,1}, b_{i,2})'$ in the CS AFT model (8.1). The covariate vector $\boldsymbol{z}_{i,l}$ for the random effects has the form

$$\boldsymbol{z}_{i,l} = (1, \mathsf{trtmtGroup}_{i,l})',$$

where $\mathsf{trtmtGroup}_{i,l}$ equals one if the (i,l)th patient underwent surgery alone and equals zero if she additionally got the course of perioperative chemotherapy.

Additionally, as fixed effects we include all baseline factors mentioned in Section 1.4 in the model. Namely, the covariate vector $\boldsymbol{x}_{i,l}$ in the model (8.1) equals

$$\begin{aligned}\boldsymbol{x}_{i,l} = (&\mathsf{ageMid}_{i,l}, \mathsf{ageOld}_{i,l}, \mathsf{tySu}_{i,l}, \mathsf{tumSiz}_{i,l}, \mathsf{nodSt}_{i,l}, \mathsf{otDis}_{i,l},\\ &\mathsf{regionNL}_i, \mathsf{regionPL}_i, \mathsf{regionSE}_i, \mathsf{regionSA}_i)',\end{aligned}$$

where ageMid and ageOld are dummies for the age groups 40–50 years and older than 50 years, respectively with the group younger than 40 years as the baseline, tySu being equal to 1 for the breast-conserving surgery and equal to 0 for mastectomy, tumSiz being equal to 1 for the tumors of size $\geq$ 2cm and equal to 0 for tumors of size $<$ 2cm, nodSt being equal to 1 for the positive and equal to 0 for the negative pathological nodal status, otDis being equal to 1 if there was another disease present and equal to 0 otherwise. Finally, covariates regionNL, regionPL, regionSE, regionSA are dummies for the geographical location of the center with France as the baseline.

Since the covariate region is categorical and center-specific it should be possible to reveal, at least partially, the regional structure of the centra from the estimates of their individual random effects $b_{i,1}$, $i = 1, \ldots, 14$ when we

omit the covariate region from the model. To show this, we fitted additionally a model where all dummies for the region were omitted from the covariate vector $\boldsymbol{x}$ (model **without region**).

For the inference we sampled two chains, each of length 200 000 with 1:5 thinning which took about 32 hours on a Pentium IV 2 GHz PC with 512 MB RAM. The first 150 000 iterations of each chain were discarded. The convergence was evaluated by a critical examination of the trace and autocorrelation plots and using the method of Gelman and Rubin (1992).

8.9.1 Prior distribution

The initial maximum-likelihood AFT model, without random effects gave the estimate of the intercept equal to 9.43 and the estimate of the error scale equal to 1.73. As the prior mean for the mixture components, ξ, we have taken zero to show that the posterior for the mixture means manages to shift from slightly misspecified location. To set up the remaining hyperparameters we followed closely the guidelines given in Section 8.2.3, namely $\kappa = 40$ which is slightly higher than $(3 \cdot 1.73)^2$, $\zeta = 2$, $h_1 = 0.2$, $h_2 = 0.1$, $\delta = 1$. For the number of mixture components, K, we used a truncated Poisson distribution prior with $\lambda = 5$ and $K_{max} = 30$. Both γ_2 (mean of the random effects $b_{i,2}$) as well as all β regression parameters were assigned a spread $\mathcal{N}(0,\, 100)$ prior. The covariance matrix $\mathbb{D}$ of the random effects got an inverse Wishart prior with $df = 2$ and $\mathbb{S} = \mathrm{diag}(0.002)$.

8.9.2 Effect of covariates on PFS time

The effect of considered covariates, in both models with included or excluded covariate region, on the progression-free survival time can be evaluated from Table 8.6 where we report posterior medians, 95% equal-tail credible intervals and Bayesian p-values (simultaneous for categorical covariates with more than 2 levels) for the β and γ parameters.

It is seen that the results for the model with region included are almost the same as these in the model with region excluded. This is in agreement with the general property of the AFT model mentioned in Section 3.3 that the regression parameters for included covariates do not change when an important factor is omitted from the model. If we base our conclusions on the model with region included then we see that, after adjustment for the remaining covariates, *surgery alone* decreases the time to the cancer progression by the factor of $\exp(-0.173) = 0.84$ compared to the surgery given together with the *perioperative chemotherapy.* However the difference is not significant at conventional 5% level.

Table 8.6: Early breast cancer patients data. Posterior medians, 95% equal-tail credible intervals and Bayesian two-sided (simultaneous) p-values for the effect of covariates.

	Model with region		Model without region	
Parameter	Poster. median	95% CI	Poster. median	95% CI
Treatment group		$p = 0.070$		$p = 0.086$
surgery alone	-0.173	$(-0.350, 0.016)$	-0.166	$(-0.342, 0.026)$
Age		$p = 0.005$		$p = 0.003$
40–50 years	0.417	(0.140, 0.695)	0.429	(0.154, 0.715)
> 50 years	0.260	(0.002, 0.520)	0.295	(0.036, 0.558)
Type of surgery		$p = 0.056$		$p = 0.029$
breast conserving	0.174	$(-0.005, 0.357)$	0.197	(0.021, 0.379)
Tumor size		$p < 0.001$		$p < 0.001$
≥ 2cm	-0.494	$(-0.686, -0.306)$	-0.507	$(-0.697, -0.314)$
Nodal status		$p < 0.001$		$p < 0.001$
positive	-0.653	$(-0.819, -0.488)$	-0.657	$(-0.822, -0.490)$
Other disease		$p = 0.008$		$p = 0.008$
present	-0.385	$(-0.666, -0.099)$	-0.394	$(-0.683, -0.102)$
Region		$p = 0.033$		
the Netherlands	-0.512	$(-0.878, -0.068)$		
Poland	0.119	$(-0.394, 0.663)$		
South Europe	-0.450	$(-0.857, -0.038)$		
South Africa	-0.864	$(-1.343, -0.371)$		

Further, the prognosis for the cancer progression is the most optimistic in the middle age group *40 – 50 years* where the time to the progression of the disease is increased by the factor of $\exp(0.417) = 1.52$ compared to the youngest group *<40 years*. In the oldest group *>50 years* the time to the disease progression is still increased, by the factor of $\exp(0.260) = 1.30$, compared to the youngest group. The estimates for the effect of age further suggests a non-linear relationship between the age and log-progression-free survival time.

The effect of the type of surgery on the disease progression is slightly not significant at 5% level when basing the inference on the model with region. However the posterior median of the β parameter for this covariate suggest that *breast conserving surgery* increases the time to the cancer progression by

Table 8.7: Early breast cancer patients data. Posterior medians and 95% equal-tail credible intervals for the moments of the error distribution and variance components of the random effects.

	Model with region		Model without region	
Parameter	Poster. median	95% CI	Poster. median	95% CI
Moments of the error distribution				
Intercept α	9.453	(8.983, 9.853)	9.229	(8.822, 9.796)
Error scale σ	1.741	(1.600, 1.859)	1.749	(1.597, 2.376)
Variance components of the random effects				
$\mathrm{sd}(b_{i,1})$	0.126	(0.026, 0.392)	0.348	(0.192, 0.616)
$\mathrm{sd}(b_{i,2})$	0.060	(0.020, 0.228)	0.085	(0.023, 0.275)
$\mathrm{corr}(b_{i,1}, b_{i,2})$	-0.071	$(-0.988, 0.973)$	-0.842	$(-0.995, 0.978)$

the factor of $\exp(0.174) = 1.20$ when compared to *mastectomy*. The effect of remaining patient-specific covariates is highly significant and in the direction expected from the clinical point of view. Namely, the tumor of size $\geq$*2 cm* decreases the time to the cancer progression by the factor of $\exp(-0.494) = 0.61$ compared to the smaller tumors of size $<$*2 cm*. A *positive* pathological nodal status decreases drastically the time to the cancer progression by the factor of $\exp(-0.653) = 0.52$ compared to the *negative* result. The *presence* of other related disease decreases the PFS time by the factor of $\exp(-0.385) = 0.68$.

Finally, a significant effect of the geographical region on the PFS time is seen. The best performing region is found to be *Poland*, followed by *France*, *South Europe* and *the Netherlands*. The region which performs the worst is then *South Africa*.

Relatively small effect of the perioperative therapy compared to surgery alone is also seen from the posterior predictive survival curves shown in Figure 8.9 and drawn for region = France and two typical combinations of covariates.

8.9.3 Predictive error density and variance components of random effects

Posterior summary statistics for the moments of the error distribution and the variance components of the random effects are given in Table 8.7. The

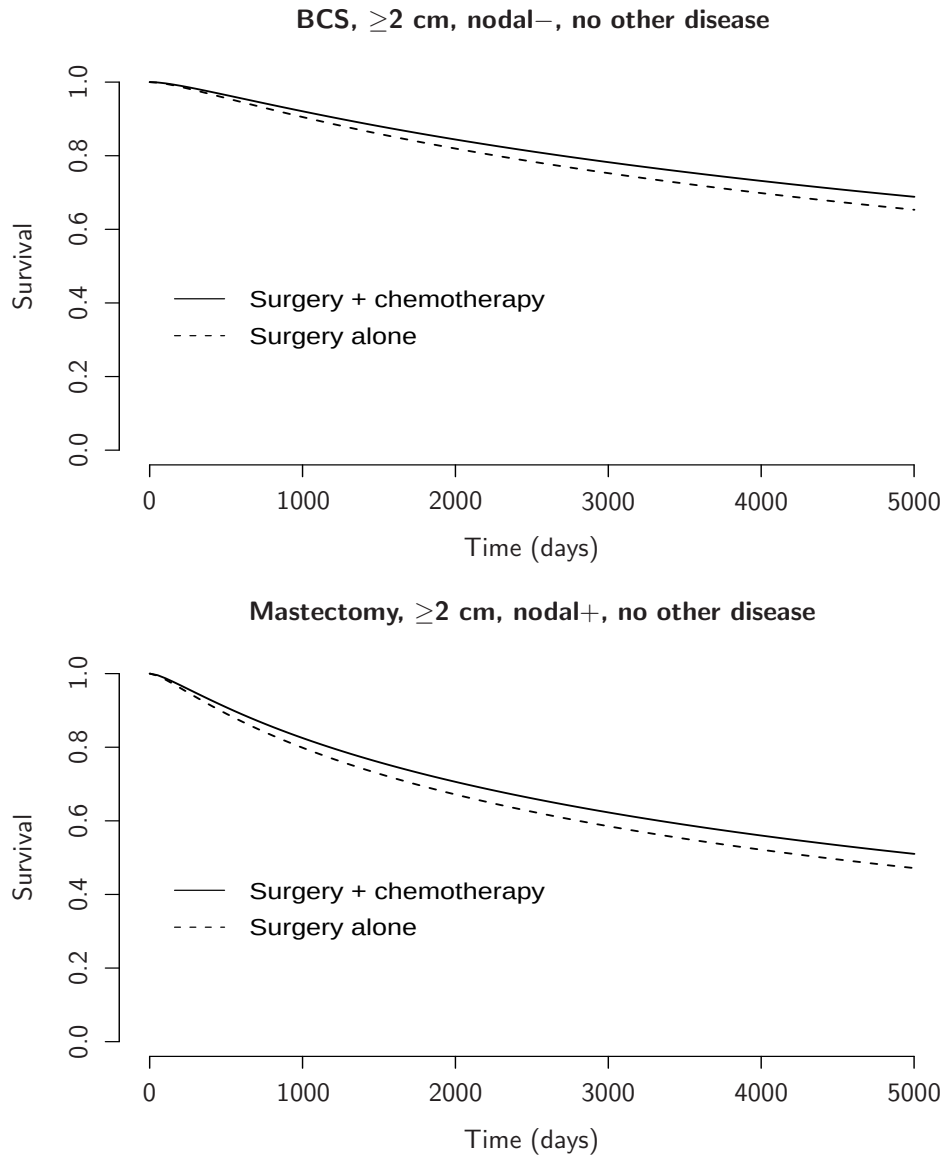


Figure 8.9: Early breast cancer patients data. Predictive survival curves based on the model with region for region = *France*, and two typical combinations of covariates: (1) *breast conserving surgery*, tumor size $\geq$*2 cm*, *negative* nodal status and *no* other associated disease (9.79% of the sample), (2) *mastectomy*, tumor size $\geq$*2 cm*, *positive* nodal status and *no* other associated disease (13.88% of the sample).

moments of the error distribution are computed in the same way as indicated in Section 8.7.2. It is seen that although there is heterogeneity between centra, the within-center variability given by the variance of the error distribution is much higher than the between-centra variability given by the variance of the random effects. Furthermore, as expected, the variability of the random intercept term $b_{i,1}$ increased considerably when we omitted the covariate region.

According to the posterior median there exists very low negative correlation between the overall center level and the treatment $\times$ center interaction in the model with region and relatively high negative correlation in the model with region excluded. However, in both cases the 95% equal-tail credible interval covers almost the whole range $(-1, 1)$ of possible values for ϱ forcing us to conclude that almost nothing can be said about the random effects correlation ϱ, probably due to the fact that effectively only a sample of size 14 is used to estimate this correlation. The reason for quite huge difference

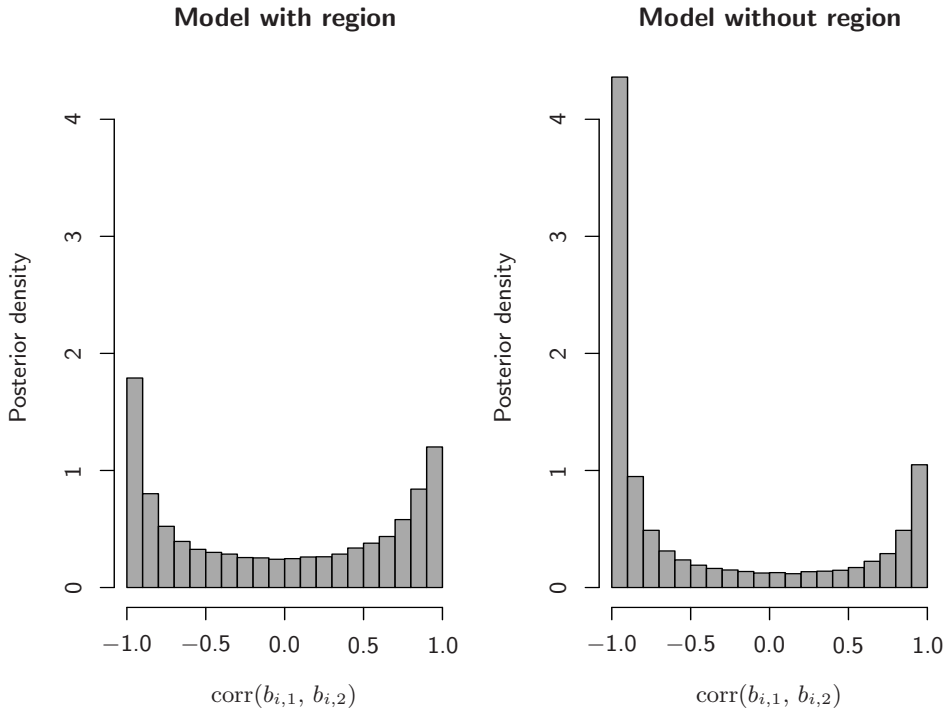


Figure 8.10: Early breast cancer patients data. Scaled histograms for sampled corr($b_{i,1}$, $b_{i,2}$).

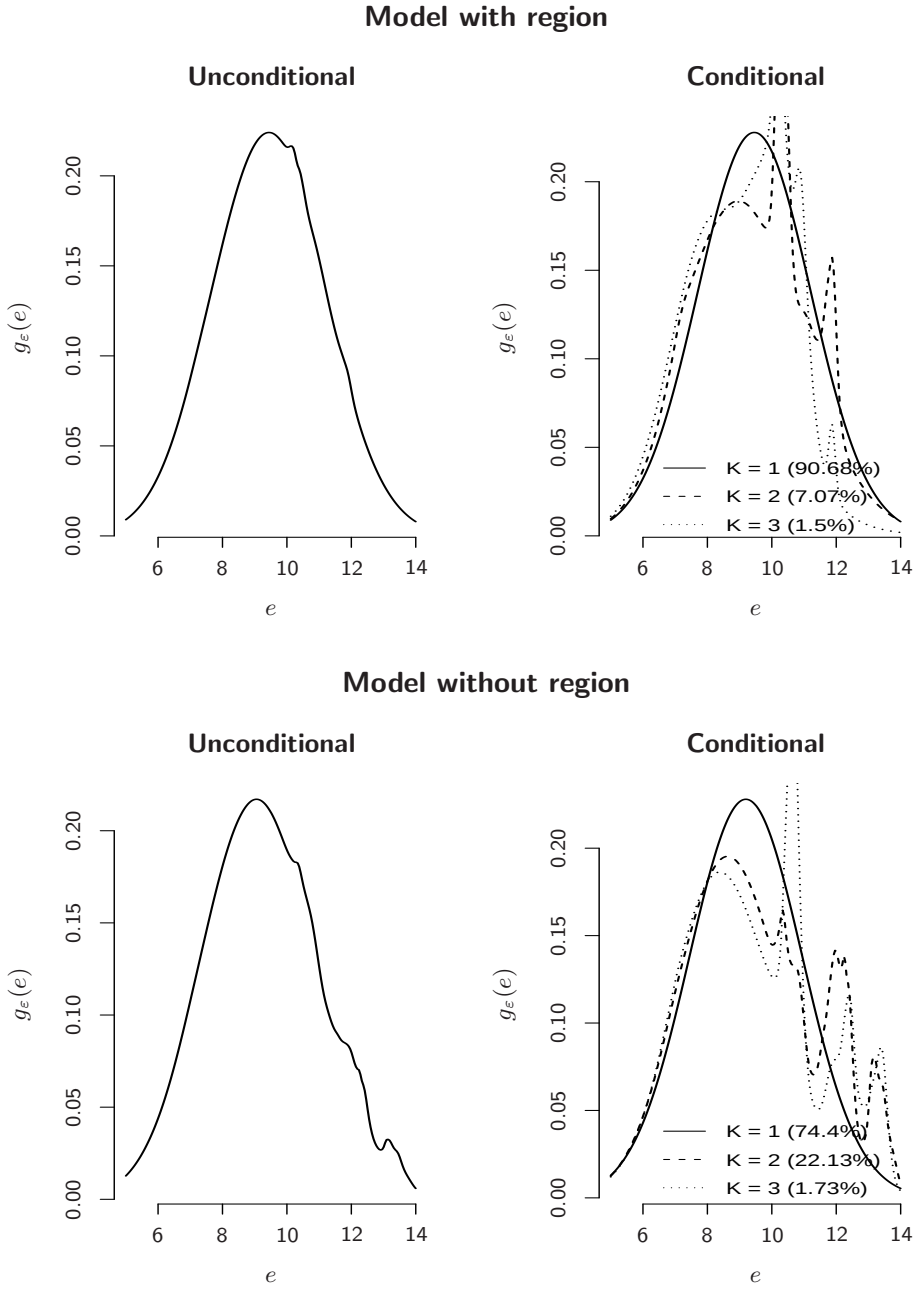


Figure 8.11: Early breast cancer patients data. Posterior predictive error densities.

in the posterior median for ϱ in the two models can be found in Figure 8.10 where we show scaled histograms of sampled values of ϱ, i.e. estimates of the posterior density of ϱ. It is seen that the posterior density has, in both cases, a 'U' shape, while putting somewhat more mass on negative values in the case of the model without region.

In the sample, mostly error densities with a low number of mixture components were presented. Namely, in the model with region, 90.68% of the sample was formed by a one-component density, 7.07% of the sample was formed by a two-component mixture, 1.50% of the sample contained a three-component mixtures and mixtures with more than 3 components were all represented in only 0.75% of the sample. In the model with omitted covariate region the proportion of densities with at least two components quite logically increased, namely one-component density is now represented only in 74.40% of the sample, two-component mixtures in 22.13% of the sample and three-component mixtures in 1.73% of the sample. Mixtures with more than 3 components are still quite rare, being all together represented only in 1.74% of the sample. The estimates of the error density (both uncondionally and conditionally given the number of mixture components) are given in Figure 8.11.

8.9.4 Estimates of individual random effects

Estimates of individual random effects that could serve to discriminate the centra are given in Figure 8.12. To be able to compare directly the models with and without covariate region the plots related to the random intercept $b_{i,1}$ take into account also the overall intercept α (mean of the error distribution) and in the case of the model with region also the appropriate main effect of region (β(regionNL), β(regionPL), β(regionSE), and β(regionSA) respectively). It is seen that the estimates of individual random intercepts in the model without region managed quite nicely to capture also the region effect, of course for the price of decreased precision of the estimates.

8.9.5 Conclusions

In this section, we have shown an analysis of a typical multicenter clinical trial with heterogeneity with respect to the overall center effect as well as center $\times$ treatment interaction. Among others we have further shown how the center-specific random effects may capture the effect of an omitted center-specific covariate.

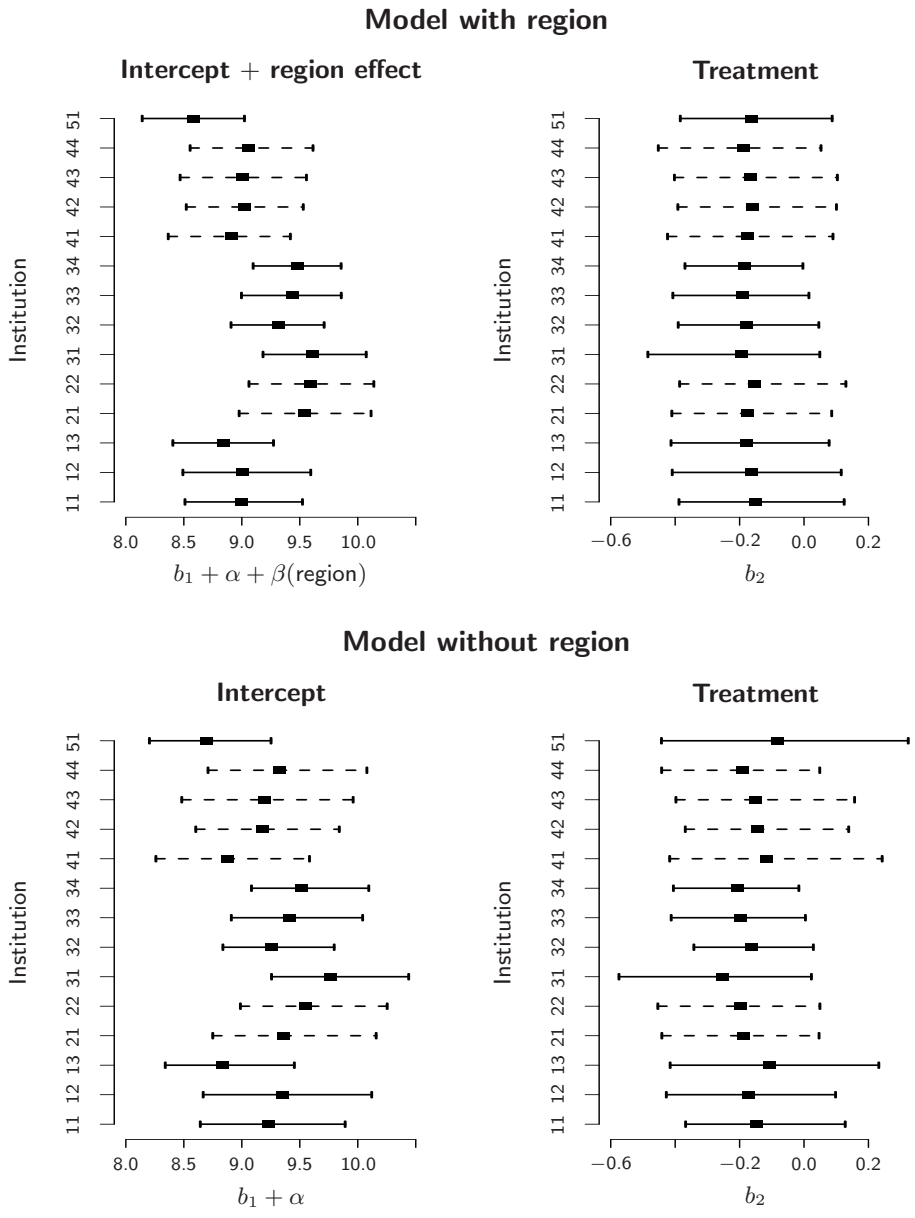


Figure 8.12: Early breast cancer patients data. Posterior means and 95% equal-tail credible intervals for individual random effects. Random intercepts are further shifted by an overall intercept α and in the model with region also by a corresponding region main effect β(region).

8.10 Discussion

In this chapter, we have proposed a Bayesian cluster-specific accelerated failure time model whose error distribution is modelled in a flexible way as a finite normal mixture. An advantage of the full Bayesian approach is the fact that a general random effect vector can be easily included in the model. Subsequently, the effect of covariates can be evaluated jointly with the association among clustered responses. Further, interval-, right-, or left-censored data are easy to handle and finally, the MCMC sampling-based implementation of the model offers a straightforward way to obtain credible intervals of model parameters as well as predictive survival or hazard curves.

Observe that the Bayesian approach is used here mainly for technical convenience. Indeed, in practice likelihood (8.3) is hardly tractable using the maximum-likelihood method. On the other hand, the Bayesian estimation using the MCMC does not pose any real difficulties. Further, since all our prior distributions are non-informative (or close to, cfr. variance parameters) and we use (on a proper scale) more or less posterior modes as point estimates the classical maximum-likelihood estimation would lead to almost the same results.

The proposed methodology aims to contribute to the area of *semi-parametric* modelling of *correlated* and at the same time *interval-censored* data. Furthermore, our approach allows to bring in a structure into the dependencies between observations in one cluster. For instance, in multicenter studies, the vector $\boldsymbol{z}_{i,l} = (1, \mathsf{treatment}_{i,l})'$ in the model formula (8.1) allows to consider not only the random center effect but also a random center-by-treatment interaction which can sometimes be substantial.

Unfortunately, our approach cannot handle time-dependent covariates. However, the same is true for any model where the distribution of the response is specified by the density and not by the hazard function. To include also the time-dependent covariates, usually the Cox's proportional hazards model is used. For example, Kooperberg and Clarkson (1997); Betensky et al. (1999); Goetghebeur and Ryan (2000) consider independent interval-censored data. Vaida and Xu (2000) offer an approach based on the proportional hazards linear mixed model with right-censored data.

Finally, our approach can be quite easily extended along the lines presented in Chapters 9 and 10 to handle also doubly-interval-censored data, i.e. the data where the response is given as the difference of two interval-censored observations.

Chapter 9

Bayesian Penalized Mixture Cluster-Specific AFT Model

This chapter continues with the developments in the framework of the cluster-specific AFT model. However, to model unknown distributional shapes a penalized normal mixture introduced in Section 6.3 will be exploited instead of the classical normal mixture that was used in Chapter 8. Furthermore, we directly describe a model for doubly-interval-censored data although it can also be used with interval- or right-censored data. This approach, introduced by Komárek and Lesaffre (2006b), will allow us to analyze the caries times in the Signal Tandmobiel® study.

The cluster-specific AFT model for doubly-interval-censored data is specified in Section 9.1. In Section 9.2, we specify the prior distributions of all model parameters and derive their posterior distribution. Markov chain Monte Carlo methodology for the model of this chapter is described in Section 9.3. Estimation of the survival distribution and of the individual random effects is described in Sections 9.4 and 9.5, respectively. Results of the simulation study aiming to evaluate the performance of the proposed method are shown in Section 9.6. Section 9.7 presents the analysis of doubly-interval-censored caries times of the four permanent first molars. The analysis of the breast cancer multicenter study is given in Section 9.8. Discussion finalizes the chapter in Section 9.9.

9.1 Model

Let $\sum_{i=1}^{N} n_i$ observational units be divided into N clusters, the ith one of size n_i. Let $U_{i,l}$ and $V_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ denote the true chronological onset and failure time, respectively and $T_{i,l} = V_{i,l} - U_{i,l}$ the true event time. With doubly interval censoring, it is only known that $U_{i,l}$ occurred within an interval of time $\lfloor u_{i,l}^L,\ u_{i,l}^U \rfloor$, where $u_{i,l}^L \leq u_{i,l}^U$. Similarly, the failure time $V_{i,l}$ is only known to lie in an interval $\lfloor v_{i,l}^L,\ v_{i,l}^U \rfloor$, with $v_{i,l}^L \leq v_{i,l}^U$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$. As in the whole thesis, it is assumed that observed intervals result from an independent noninformative censoring process (see Section 2.4). Further, as indicated in Section 4.1.2, we will assume that, given the model parameters, the true event time $T_{i,l}$ is independent of the true onset time $U_{i,l}$ for all i and l. Below, we discuss this issue further.

To account for possible dependencies of different individuals within a cluster, the cluster-specific random effects $\boldsymbol{d}_i = (d_{i,1}, \ldots, d_{i,q_d})'$ and $\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,q_b})'$ are introduced and incorporated in the cluster-specific AFT model for doubly-interval-censored data:

$$\log(U_{i,l}) = \boldsymbol{\delta}' \boldsymbol{x}_{i,l}^u + \boldsymbol{d}_i' \boldsymbol{z}_{i,l}^u + \zeta_{i,l}, \tag{9.1}$$

$$\log(V_{i,l} - U_{i,l}) = \log(T_{i,l}) = \boldsymbol{\beta}' \boldsymbol{x}_{i,l}^t + \boldsymbol{b}_i' \boldsymbol{z}_{i,l}^t + \varepsilon_{i,l}, \tag{9.2}$$

$$i = 1, \ldots, N, \quad l = 1, \ldots, n_i,$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{m_u})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_t})'$ are unknown regression parameter vectors, $\boldsymbol{z}_{i,l}^u$ is the covariate vector for random effects influencing the distribution of the onset time, $\boldsymbol{z}_{i,l}^t$ the covariate vector for random effects influencing the distribution of the event time and similary, $\boldsymbol{x}_{i,l}^u$ is the covariate vector for fixed effects having possibly an impact on the onset time and $\boldsymbol{x}_{i,l}^t$ the covariate vector for fixed effects having possibly an impact on the event time. The error terms $\zeta_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ are i.i.d. random variables with some density $g_\zeta(\zeta)$. Analogously, the error terms $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ are i.i.d. random variables with density $g_\varepsilon(\varepsilon)$. The random effects $\boldsymbol{d}_i$, $i = 1, \ldots, N$ and $\boldsymbol{b}_i$, $i = 1, \ldots, N$, respectively are assumed to be i.i.d. with a density $g_d(\boldsymbol{d})$ and $g_b(\boldsymbol{b})$, respectively. Furthermore we assume that ε_{i_1,l_1}, ζ_{i_2,l_2}, $\boldsymbol{b}_{i_3}$ and $\boldsymbol{d}_{i_4}$ are independent for all i_1, i_2, i_3, i_4 and l_1, l_2. This assumption implies that, given the model parameters and the random effects $\boldsymbol{b}_i$ and $\boldsymbol{d}_i$, $U_{i,l}$ and $T_{i,l}$ are independent for each i and l and the vectors $\boldsymbol{U}_i = (U_{i,1}, \ldots, U_{i,n_i})'$ and $\boldsymbol{T}_i = (T_{i,1}, \ldots, T_{i,n_i})'$ are independent for each i. Furthermore, for example in the context of the Signal Tandmobiel® application (see Section 9.7) where $U_{i,l}$ and $T_{i,l}$ are the emergence time and the time to caries, respectively, for the lth tooth of the ith child, it also

implies the following decomposition

(a) Whether a child is an early or late emerger is independent of whether a child is more or less sensitive against caries (independence of $\boldsymbol{d}_i$ and $\boldsymbol{b}_i$);

(b) Whether a specific tooth emerges early or late is independent of whether that tooth is more or less sensitive against caries (independence of $\zeta_{i,l}$ and $\varepsilon_{i,l}$).

9.1.1 Distributional assumptions

To finalize the specification of the measurement model we have to specify the densities g_ζ, g_ε of the random errors and the densities g_d, g_b of the random effects. According to the dimensionality of the problem, we distinguish two situations.

Model U

In the case of *univariate* densities, i.e. for the densities g_ζ and g_ε and for the densities g_d and g_b if the corresponding random effects are univariate (in which case we will use the notation $\boldsymbol{d}_i = (d_{i,1}) \equiv d_i$ and/or $\boldsymbol{b}_i = (b_{i,1}) \equiv b_i$), a penalized normal mixture as introduced in Section 6.3 will be used.

That is, a generic density $g(y)$ of a random variable Y (substitute $\zeta_{i,l}$, $\varepsilon_{i,l}$, d_i or b_i) is modelled as a location-and-scale transformed weighted sum of normal densities over a *fixed fine* grid of knots $\boldsymbol{\mu} = (\mu_{-K}, \ldots, \mu_K)'$ centered around $\mu_0 = 0$. The means of the normal components are equal to the knots and their variances are all equal and fixed to σ^2, i.e.

$$g(y) = \tau^{-1} \sum_{j=-K}^{K} w_j(\boldsymbol{a})\, \varphi\Big(\frac{y-\alpha}{\tau} \,\Big|\, \mu_j,\, \sigma^2\Big), \tag{9.3}$$

where the unknown intercept term α and the unknown scale parameter τ have to be estimated as well as the vector $\boldsymbol{a} = (a_{-K}, \ldots, a_K)'$ of the transformed weights. See (6.14) for the relationship between $\boldsymbol{a}$ and $\boldsymbol{w} = (w_{-K}, \ldots, w_K)'$.

Model M

In the case when a random effect vector $\boldsymbol{d}_i$ or $\boldsymbol{b}_i$ is *multivariate* it is assumed, analogously to Chapter 8 that it follows a multivariate normal distribution. This choice is driven mainly by computational convenience. Note however,

that the densities g_ζ and g_ε are still modelled using the penalized normal mixture (9.3). Finally, the same reasoning as in Section 8.1.1 can be used to explain why we put more emphasis on a correct specification of the error distribution.

For notational convenience and clarity of the exposition we will assume that in *Model U*, both random effects are univariate ($q_d = q_b = 1$) whereas in *Model M*, both random effects are multivariate ($q_d > 1$ and $q_b > 1$). However, in practical situations both cases can be mixed. For example the distribution of the univariate d_i can be specified as a penalized normal mixture (9.3) whereas for the multivariate $\boldsymbol{b}_i$ a multivariate normal distribution can be used.

9.1.2 Likelihood

Denoting p a generic density, the likelihood contribution of the ith cluster is given by

$$
\begin{aligned}
L_i &= \int_{\mathbb{R}^{q_d}} \int_{\mathbb{R}^{q_b}} \left\{ \prod_{l=1}^{n_i} \oint_{u_{i,l}^L}^{u_{i,l}^U} \oint_{v_{i,l}^L - u_{i,l}}^{v_{i,l}^U - u_{i,l}} p(t_{i,l},\, \boldsymbol{b}_i,\, u_{i,l},\, \boldsymbol{d}_i)\, dt_{i,l}\, du_{i,l} \right\} d\boldsymbol{b}_i\, d\boldsymbol{d}_i \\
&= \int_{\mathbb{R}^{q_d}} \int_{\mathbb{R}^{q_b}} \left\{ \prod_{l=1}^{n_i} \oint_{u_{i,l}^L}^{u_{i,l}^U} \oint_{v_{i,l}^L - u_{i,l}}^{v_{i,l}^U - u_{i,l}} p(t_{i,l} \,|\, \boldsymbol{b}_i,\, u_{i,l},\, \boldsymbol{d}_i)\, p(\boldsymbol{b}_i \,|\, u_{i,l},\, \boldsymbol{d}_i) \right. \\
&\qquad \left. p(u_{i,l} \,|\, \boldsymbol{d}_i)\, p(\boldsymbol{d}_i)\, dt_{i,l}\, du_{i,l} \right\} d\boldsymbol{b}_i\, d\boldsymbol{d}_i \qquad (9.4) \\
&= \int_{\mathbb{R}^{q_d}} \int_{\mathbb{R}^{q_b}} \left[\prod_{l=1}^{n_i} \oint_{u_{i,l}^L}^{u_{i,l}^U} \left\{ \oint_{v_{i,l}^L - u_{i,l}}^{v_{i,l}^U - u_{i,l}} p(t_{i,l} \,|\, \boldsymbol{b}_i)\, dt_{i,l} \right\} p(u_{i,l} \,|\, \boldsymbol{d}_i)\, du_{i,l} \right] \\
&\qquad p(\boldsymbol{b}_i)\, p(\boldsymbol{d}_i)\, d\boldsymbol{b}_i\, d\boldsymbol{d}_i,
\end{aligned}
$$

where

$$
\begin{aligned}
p(t_{i,l} \,|\, \boldsymbol{b}_i) &= t_{i,l}^{-1}\, g_\varepsilon\big\{ \log(t_{i,l}) - \boldsymbol{b}_i' \boldsymbol{z}_{i,l}^t - \boldsymbol{\beta}' \boldsymbol{x}_{i,l}^t \big\} \\
p(u_{i,l} \,|\, \boldsymbol{d}_i) &= u_{i,l}^{-1}\, g_\zeta\big\{ \log(u_{i,l}) - \boldsymbol{d}_i' \boldsymbol{z}_{i,l}^u - \boldsymbol{\delta}' \boldsymbol{x}_{i,l}^u \big\}
\end{aligned}
$$

are modelled using the expression (9.3) for g_ε and g_ζ.

Further, in the *Model U*, $p(\boldsymbol{b}_i) = g_b(b_i)$ and $p(\boldsymbol{d}_i) = g_d(d_i)$ are penalized normal mixtures (9.3). Since it is not possible to distinguish between the intercept terms of the error and the random effect the intercepts $\alpha = \alpha^d$ for g_d and $\alpha = \alpha^b$ for g_b are fixed to zero for identifiability reasons. In the case

of the *Model M*, the densities $p(\boldsymbol{b}_i) = g_b(\boldsymbol{b}_i)$ and $p(\boldsymbol{d}_i) = g_d(\boldsymbol{d}_i)$ are densities of an appropriate multivariate normal distribution (see also Section 9.2.3).

The method of penalized maximum-likelihood, suggested in Chapter 7, is computationally quite demanding for likelihood (9.4). Instead, a Bayesian approach together with MCMC methodology will be used here to avoid explicit integration and optimization.

9.2 Bayesian hierarchical model

To specify the model from a Bayesian point of view, prior distributions for all unknown parameters have to be given. For our model we assume a hierarchical structure described by a directed acyclic graph (DAG). The DAG for *Model U* where the distributions of the univariate random effects and the error terms are estimated using the penalized mixture is given in Figure 9.1.

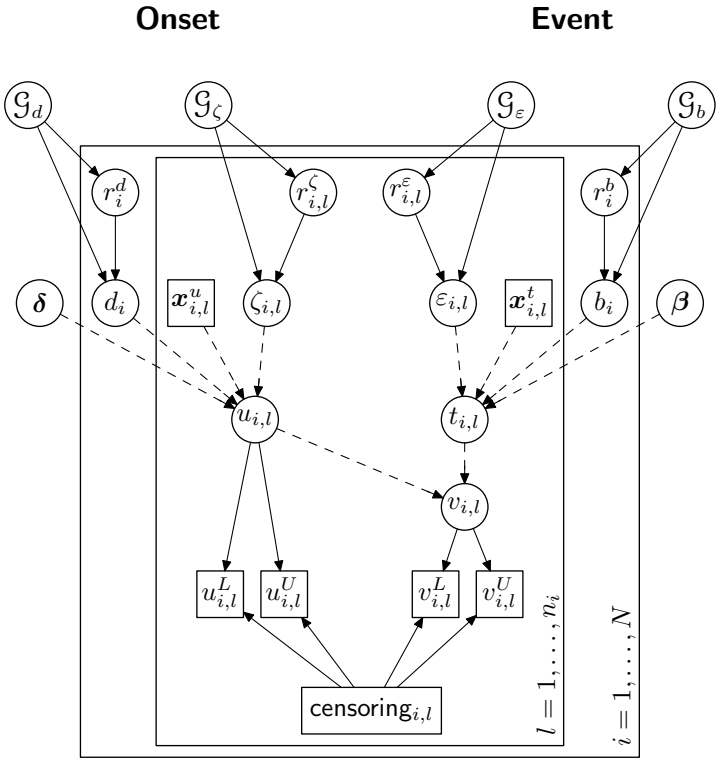


Figure 9.1: Directed acyclic graph for the Bayesian penalized mixture cluster-specific AFT model with univariate random effects (*Model U*).

The DAG for *Model M* with multivariate normal random effects and error terms expressed using the penalized mixture is given in Figure 9.2.

For *Model U*, the joint prior distribution of the total parameter vector $\boldsymbol{\theta}$ is given by

$$
\begin{aligned}
p(\boldsymbol{\theta}) \propto & \\
& \prod_{i=1}^{N} \Bigg[\prod_{l=1}^{n_i} \Big\{ p\big(v_{i,l} \,\big|\, u_{i,l},\, t_{i,l}\big) \times p\big(t_{i,l} \,\big|\, \boldsymbol{\beta},\, b_i,\, \varepsilon_{i,l}\big) \times p\big(u_{i,l} \,\big|\, \boldsymbol{\delta},\, d_i,\, \zeta_{i,l}\big) \times \\
& p\big(\varepsilon_{i,l} \,\big|\, \mathcal{G}_\varepsilon,\, r^{\varepsilon}_{i,l}\big) \times p\big(\zeta_{i,l} \,\big|\, \mathcal{G}_\zeta,\, r^{\zeta}_{i,l}\big) \times p\big(r^{\varepsilon}_{i,l} \,\big|\, \mathcal{G}_\varepsilon\big) \times p\big(r^{\zeta}_{i,l} \,\big|\, \mathcal{G}_\zeta\big) \Big\} \times \\
& p\big(b_i \,\big|\, \mathcal{G}_b,\, r^{b}_{i}\big) \times p\big(d_i \,\big|\, \mathcal{G}_d,\, r^{d}_{i}\big) \times p\big(r^{b}_{i} \,\big|\, \mathcal{G}_b\big) \times p\big(r^{d}_{i} \,\big|\, \mathcal{G}_d\big) \Bigg] \times \\
& p\big(\mathcal{G}_\varepsilon\big) \times p\big(\mathcal{G}_\zeta\big) \times p\big(\mathcal{G}_b\big) \times p\big(\mathcal{G}_d\big) \times p\big(\boldsymbol{\delta}\big) \times p\big(\boldsymbol{\beta}\big).
\end{aligned}
\tag{9.5}
$$

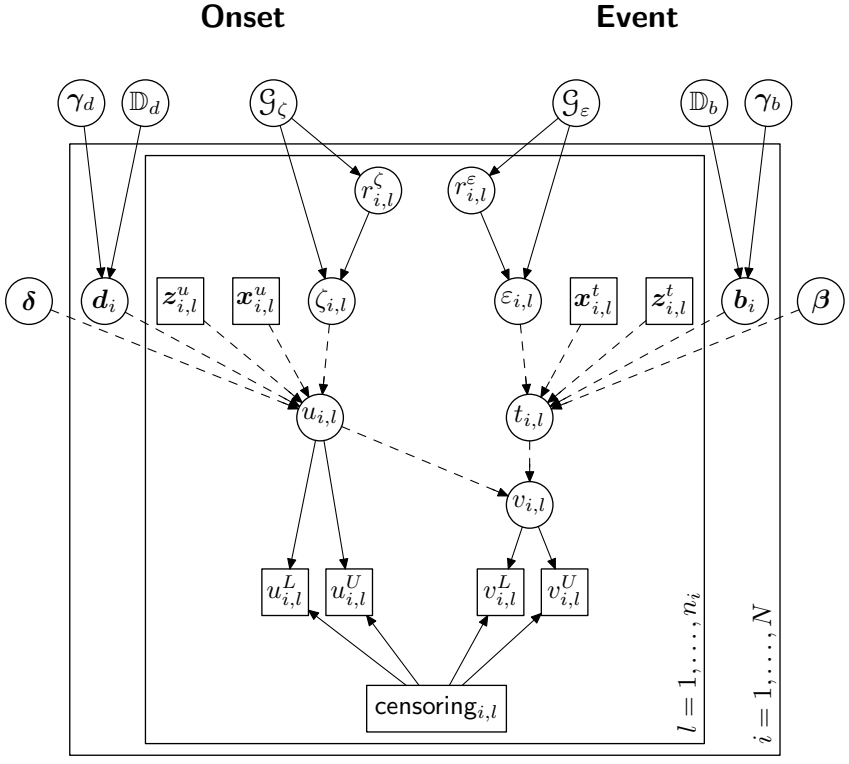


Figure 9.2: Directed acyclic graph for the Bayesian penalized mixture cluster-specific AFT model with multivariate normal random effects (*Model M*).

The node $\mathcal{G}_\varepsilon$ refers to the set $\{\sigma^\varepsilon,\ \boldsymbol{\mu}^\varepsilon,\ \alpha^\varepsilon,\ \tau^\varepsilon,\ \boldsymbol{w}^\varepsilon,\ \boldsymbol{a}^\varepsilon,\ \lambda^\varepsilon\}$ which contains the parameters of formulas (9.3) and (6.14) and a smoothing parameter λ^ε which will be further discussed in Section 9.2.1. The sets $\mathcal{G}_\zeta$, $\mathcal{G}_b$, $\mathcal{G}_d$ are defined in an analogous manner. Further, let $\mathcal{G}$ be a generic symbol for its subscriped counterpart (i.e. for $\mathcal{G}_\varepsilon$, $\mathcal{G}_\zeta$, $\mathcal{G}_b$, $\mathcal{G}_d$) and let y be a generic symbol for $\varepsilon_{i,l}, \zeta_{i,l}, b_i$, or d_i, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$, respectively. The sub-DAG for the generic Y random variable is shown in Figure 9.3 and the corresponding DAG conditional distributions are discussed in Sections 9.2.1 and 9.2.2.

In the case of *Model M*, the joint prior distribution is given by

$$
\begin{aligned}
p(\boldsymbol{\theta}) \propto\ & \\
& \prod_{i=1}^{N} \Bigg[\prod_{l=1}^{n_i} \Big\{ p\big(v_{i,l} \,\big|\, u_{i,l},\, t_{i,l}\big) \times p\big(t_{i,l} \,\big|\, \boldsymbol{\beta},\, \boldsymbol{b}_i,\, \varepsilon_{i,l}\big) \times p\big(u_{i,l} \,\big|\, \boldsymbol{\delta},\, \boldsymbol{d}_i,\, \zeta_{i,l}\big) \times \\
& p\big(\varepsilon_{i,l} \,\big|\, \mathcal{G}_\varepsilon,\, r^{\varepsilon}_{i,l}\big) \times p\big(\zeta_{i,l} \,\big|\, \mathcal{G}_\zeta,\, r^{\zeta}_{i,l}\big) \times p\big(r^{\varepsilon}_{i,l} \,\big|\, \mathcal{G}_\varepsilon\big) \times p\big(r^{\zeta}_{i,l} \,\big|\, \mathcal{G}_\zeta\big) \Big\} \times \\
& p\big(\boldsymbol{b}_i \,\big|\, \boldsymbol{\gamma}_b,\, \mathbb{D}_b\big) \times p\big(\boldsymbol{d}_i \,\big|\, \boldsymbol{\gamma}_d,\, \mathbb{D}_d\big) \Bigg] \times \\
& p\big(\mathcal{G}_\varepsilon\big) \times p\big(\mathcal{G}_\zeta\big) \times p\big(\boldsymbol{\gamma}_b\big) \times p\big(\mathbb{D}_b\big) \times p\big(\boldsymbol{\gamma}_d\big) \times p\big(\mathbb{D}_d\big) \times p\big(\boldsymbol{\delta}\big) \times p\big(\boldsymbol{\beta}\big),
\end{aligned}
\tag{9.6}
$$

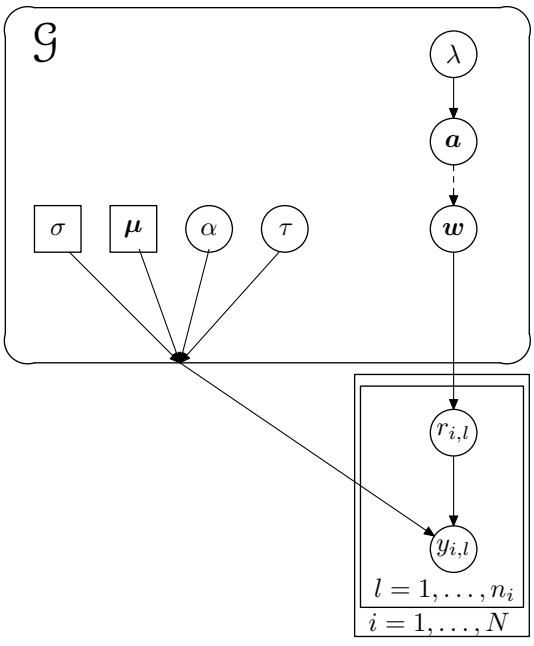


Figure 9.3: Directed acyclic graph for the penalized mixture.

where $\boldsymbol{\gamma}_d$ and $\mathbb{D}_d$ are the mean and the covariance matrix for the random effect vectors $\boldsymbol{d}_i$ and $\boldsymbol{\gamma}_b$ and $\mathbb{D}_b$ are the mean and the covariance matrix for the random effect vectors $\boldsymbol{b}_i$. These parameters will be discussed in detail in Section 9.2.3.

All the multiplicands of expressions (9.5) and (9.6) will be discussed in detail in the following sections.

9.2.1 Prior distribution for $\mathcal{G}$

The prior distribution of a generic node $\mathcal{G}$ whose structure is given in Figure 9.3 equals

$$p(\mathcal{G}) \propto p(\boldsymbol{a} \,|\, \lambda)\, p(\lambda)\, p(\alpha)\, p(\tau).$$

Prior for transformed mixture weights

Although often the grid length $(2\,K + 1)$ is of moderate size it results in a rather large number of unknown $\boldsymbol{a}$ parameters. To avoid overfitting of the data and identifiability problems, a restriction on the $\boldsymbol{a}$ parameters is needed. In Chapter 7 we added a penalty term for the transformed weights to the log-likelihood for this purpose. This penalty term can be interpreted as an informative log-prior distribution (e.g., Silverman, 1985, Section 6). Therefore the prior distribution $p(\boldsymbol{a} \,|\, \lambda)$ is defined as the exponential of the penalty term used in Chapter 7, i.e.

$$\begin{aligned} p(\boldsymbol{a} \mid \lambda) &\propto \exp\Big\{-\frac{\lambda}{2} \sum_{j=-K+s}^{K} \big(\Delta^s\, a_j\big)^2\Big\} \\ &= \exp\Big\{-\frac{\lambda}{2}\, \boldsymbol{a}'\, \mathbb{P}'\mathbb{P}\, \boldsymbol{a}\Big\}, \end{aligned} \tag{9.7}$$

where Δ^s denotes a difference operator of order s and $\mathbb{P}$ the corresponding difference operator matrix. The hyperparameter λ controls the smoothness of the resulting density $g(y)$.

Expression (9.7) is that of a multivariate normal density with zero mean and covariance matrix $\lambda^{-1}\big(\mathbb{P}'\mathbb{P}\big)^{-}$, where $\big(\mathbb{P}'\mathbb{P}\big)^{-}$ denotes a generalized inverse of the matrix $\mathbb{P}'\mathbb{P}$. This distribution is known as a Gaussian Markov random field (GMRF) and is extensively used in spatial statistics. Although the distribution (9.7) is improper (the matrix $\mathbb{P}'\mathbb{P}$ has a deficiency of s in its rank) the resulting posterior distribution is proper as soon as there is some informative data available, see Besag et al. (1995).

As a consequence of the findings discussed in Section 7.2, prior distribution (9.7) favours smooth estimates of the estimated densities ($g_\varepsilon, g_\zeta, g_b$ or g_d). Due to the correspondence of the prior (9.7) with the penalty term in the penalized maximum-likelihood approach we will call the mixture model (9.3) with this prior *a penalized mixture*.

Prior for the smoothing parameter

The smoothing hyperparameter λ can be interpreted as a component of the prior precision of the transformed weights $\boldsymbol{a}$. See Section 7.2.3 for the approaches to determine the optimal value of λ in the context of penalized maximum-likelihood estimation. For our full Bayesian inference, the unknown smoothing parameter λ is considered stochastic and is estimated simultaneously with all the remaining parameters of the model. Therefore, here a hyperprior has been assigned to λ, i.e. a highly dispersed Gamma($h_{\lambda,1}$, $h_{\lambda,2}$) prior, i.e.

$$p(\lambda) = \frac{h_{\lambda,2}^{h_{\lambda,1}}}{\Gamma(h_{\lambda,1})}\, \lambda^{h_{\lambda,1}-1}\, \exp\bigl(-h_{\lambda,2}\,\lambda\bigr),$$

where $h_{\lambda,1}$ is the fixed shape parameter and $h_{\lambda,2}$ the fixed rate parameter. A dispersed gamma distribution is obtained for instance with $h_{\lambda,1} = h_{\lambda,2} = 0.001$ or $h_{\lambda,1} = 1$, $h_{\lambda,2} = 0.005$.

Prior for the mixture intercept

Finally, in the case when the intercept term α is not fixed to zero (intercept of error distributions), a highly dispersed normal distribution has been taken for $p(\alpha)$, i.e.

$$p(\alpha) = \varphi(\alpha \,|\, \nu_\alpha,\, \psi_\alpha),$$

where ν_α is the fixed prior mean and ψ_α is the fixed large prior variance.

Prior for the mixture scale

For the precision τ^{-2} we have taken a highly dispersed Gamma($h_{\tau,1}$, $h_{\tau,2}$) distribution, see above the paragraph on the prior for the smoothing parameter. Alternatively a uniform distribution on τ (formally a truncated gamma distribution for τ^{-2} with $h_{\tau,1} = -1/2$ and $h_{\tau,2} = 0$) which is sometimes preferred for hierarchical models (Gelman et al., 2004, pp. 136, 390) could be taken.

9.2.2 Prior distribution for the generic node Y

To specify the prior distribution of generic Y ($\varepsilon_{i,l}$, $\zeta_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ in *Models U* and *M* and b_i, d_i, $i = 1, \ldots, N$ in *Model U*) we introduce, analogously to Section 8.2.1, a latent *allocation variable* r taking values in $\{-K, \ldots, K\}$. Actually, data augmentation (Tanner and Wong, 1987) is introduced which simplifies the MCMC procedure. The DAG conditional distribution $p(y \,|\, \mathcal{G}, r)$ is simply a normal distribution:

$$p(y \,|\, \mathcal{G}, r) = p(y \,|\, \sigma, \boldsymbol{\mu}, \alpha, \tau, r) = \varphi\left(y \,|\, \alpha + \tau\, \mu_r, (\tau\, \sigma)^2\right).$$

Further, $p(r \,|\, \mathcal{G}) = p(r \,|\, \boldsymbol{w})$ is given by

$$\Pr\big(r = j \;\big|\; \boldsymbol{w}\big) = w_j, \quad j \in \{-K, \ldots, K\}.$$

Had the latent allocation variable r not been introduced we would have had to work with the conditional distribution $p(y \,|\, \mathcal{G}) = p(y \,|\, \sigma, \boldsymbol{\mu}, \alpha, \tau, \boldsymbol{w})$ which is a normal *mixture* given by the formula (9.3).

9.2.3 Prior distribution for multivariate random effects in Model M

As was mentioned in Section 9.1.1, the multivariate random effects $\boldsymbol{b}_i$ and $\boldsymbol{d}_i$, $i = 1, \ldots, N$ in *Model M* are assumed to be a priori normally distributed. That is, the densities $p(\boldsymbol{b}_i \,|\, \boldsymbol{\gamma}_b, \mathbb{D}_b)$ and $p(\boldsymbol{d}_i \,|\, \boldsymbol{\gamma}_d, \mathbb{D}_d)$ in the expression (9.6) are

$$p(\boldsymbol{b}_i \,|\, \boldsymbol{\gamma}_b, \mathbb{D}_b) = \varphi_{q_b}(\boldsymbol{b}_i \,|\boldsymbol{\gamma}_b, \mathbb{D}_b), \qquad p(\boldsymbol{d}_i \,|\, \boldsymbol{\gamma}_d, \mathbb{D}_d) = \varphi_{q_d}(\boldsymbol{d}_i \,|\boldsymbol{\gamma}_d, \mathbb{D}_d),$$

where $\boldsymbol{\gamma}_b = (\gamma_{b,1}, \ldots, \gamma_{b,q_b})'$ is the prior mean of the random effects $\boldsymbol{b}_i$, $\boldsymbol{\gamma}_d = (\gamma_{d,1}, \ldots, \gamma_{d,q_d})'$ the prior mean of the random effects $\boldsymbol{d}_i$, $\mathbb{D}_b$ is the prior covariance matrix of the random effects $\boldsymbol{b}_i$ and $\mathbb{D}_d$ is the prior covariance matrix of the random effects $\boldsymbol{d}_i$.

Both prior random effect means $\boldsymbol{\gamma}_b$ and $\boldsymbol{\gamma}_d$ as well as random effect covariance matrices $\mathbb{D}_b$ and $\mathbb{D}_d$ are further assigned hyperpriors. These hyperpriors are chosen analogously to Section 8.2.2. That is, the prior distribution for each $\gamma_{b,j}$, $j = 1, \ldots, q_b$ and γ_{d,j^*}, $j^* = 1, \ldots, q_d$, respectively is $\mathcal{N}(\nu_{\gamma_b,j}, \psi_{\gamma_b,j})$ and $\mathcal{N}(\nu_{\gamma_d,j^*}, \psi_{\gamma_d,j^*})$, respectively, independently for $j = 1, \ldots, q_b$ and $j^* = 1, \ldots, q_d$, i.e.

$$p(\boldsymbol{\gamma}_b)\, p(\boldsymbol{\gamma}_d) = \Big\{\prod_{j=1}^{q_b} \varphi(\gamma_{b,j} \mid \nu_{\gamma_b,j}, \psi_{\gamma_b,j})\Big\} \times \Big\{\prod_{j^*=1}^{q_d} \varphi(\gamma_{d,j^*} \mid \nu_{\gamma_d,j^*}, \psi_{\gamma_d,j^*})\Big\}.$$

The vectors $\boldsymbol{\nu}_{\gamma_b} = (\nu_{\gamma_b,1}, \ldots, \nu_{\gamma_b,q_b})'$, $\boldsymbol{\nu}_{\gamma_d} = (\nu_{\gamma_d,1}, \ldots, \nu_{\gamma_d,q_d})'$, $\boldsymbol{\psi}_{\gamma_b} = (\psi_{\gamma_b,1}, \ldots, \psi_{\gamma_b,q_b})'$, and $\boldsymbol{\psi}_{\gamma_d} = (\psi_{\gamma_d,1}, \ldots, \psi_{\gamma_d,q_d})'$ are fixed hyperparameters. Special care is needed when the random intercept is included in the model. If for example $\boldsymbol{z}^t_{i,l,1} \equiv 1$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$, then for identifiability reasons $\gamma_{b,1}$ must be fixed to zero (or equivalently, $\nu_{\gamma_b,1} = 0$, $\psi_{\gamma_b,1} = 0$) as the overall intercept is given by the intercept α^ε of the error terms $\varepsilon_{i,l}$.

The prior distributions for the covariance matrices $\mathbb{D}_b$ and $\mathbb{D}_d$ are inverse-Wishart with fixed degrees of freedom df_b and df_d, respectively and fixed scale matrices $\mathbb{S}_b$ and $\mathbb{S}_d$, respectively. See formula (8.8) for the expression of the corresponding density.

9.2.4 Prior distribution for the regression parameters

The prior specification for the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is analogous to Section 8.2.2. Firstly, also here, we use the hierarchical centering. That is, the covariates included in $\boldsymbol{x}^t_{i,l}$ or $\boldsymbol{x}^u_{i,l}$, respectively are not included in $\boldsymbol{z}^t_{i,l}$ or $\boldsymbol{z}^u_{i,l}$, respectively and vice versa. Further, the covariate vectors $\boldsymbol{x}^t_{i,l}$ and $\boldsymbol{x}^u_{i,l}$, respectively never contain an intercept term since the overall intercept are already included in the model in the form of the parameters α^ε and α^ζ, respectively.

The prior distribution for each regression coefficient β_j, $j = 1, \ldots, m_t$ and δ_{j^*}, $j^* = 1, \ldots, m_u$ is $\mathcal{N}(\nu_{\beta,j}, \psi_{\beta,j})$ and $\mathcal{N}(\nu_{\delta,j^*}, \psi_{\beta,j^*})$, respectively, independently for $j = 1, \ldots, m_t$ and $j^* = 1, \ldots, m_u$, i.e.

$$p(\boldsymbol{\beta})\, p(\boldsymbol{\delta}) = \Big\{ \prod_{j=1}^{m_t} \varphi(\beta_j \mid \nu_{\beta,j},\, \psi_{\beta,j}) \Big\} \times \Big\{ \prod_{j^*=1}^{m_u} \varphi(\delta_{j^*} \mid \nu_{\delta,j^*},\, \psi_{\delta,j^*}) \Big\}.$$

The vectors $\boldsymbol{\nu}_\beta = (\nu_{\beta,1}, \ldots, \nu_{\beta,m_t})'$, $\boldsymbol{\nu}_\delta = (\nu_{\delta,1}, \ldots, \nu_{\delta,m_t})'$, $\boldsymbol{\psi}_\beta = (\psi_{\beta,1}, \ldots, \psi_{\beta,m_t})'$, and $\boldsymbol{\psi}_\delta = (\psi_{\delta,1}, \ldots, \psi_{\delta,m_t})'$ are fixed hyperparameters.

9.2.5 Prior distribution for the time variables

The terms $p(v_{i,l} \mid u_{i,l},\, t_{i,l})$, $p(t_{i,l} \mid\, \boldsymbol{\beta},\, \boldsymbol{b}_i,\, \varepsilon_{i,l})$ and $p(u_{i,l} \mid \boldsymbol{\delta},\, \boldsymbol{d}_i,\, \zeta_{i,l})$ appearing in the expressions (9.5) and (9.6) are all Dirac (degenerated) densities driven by the AFT models (9.1) and (9.2). Namely:

$$\begin{aligned}
p(v_{i,l} \mid u_{i,l},\, t_{i,l}) &= I[v_{i,l} = u_{i,l} + t_{i,l}], \\
p(u_{i,l} \mid \boldsymbol{\delta},\, \boldsymbol{d}_i,\, \zeta_{i,l}) &= I[\log(u_{i,l}) = \boldsymbol{\delta}' \boldsymbol{x}^u_{i,l} + \boldsymbol{d}_i' \boldsymbol{z}^u_{i,l} + \zeta_{i,l}], \\
p(t_{i,l} \mid \boldsymbol{\beta},\, \boldsymbol{b}_i,\, \varepsilon_{i,l}) &= I[\log(t_{i,l}) = \boldsymbol{\beta}' \boldsymbol{x}^t_{i,l} + \boldsymbol{b}_i' \boldsymbol{z}^t_{i,l} + \varepsilon_{i,l}], \\
& \qquad i = 1, \ldots, N, \quad l = 1, \ldots, n_i.
\end{aligned}$$

9.2.6 Posterior distribution

The product of all DAG conditional distributions determines the joint posterior distribution $p(\boldsymbol{\theta} \,|\, \text{data})$, i.e.

$$p(\boldsymbol{\theta} \,|\, \text{data}) \propto p(\boldsymbol{\theta}) \times \prod_{i=1}^{N} \prod_{l=1}^{n_i} \Big\{ p\big(u_{i,l}^L,\, u_{i,l}^U \,|\, u_{i,l},\, \mathsf{censoring}_{i,l}\big) \times p\big(v_{i,l}^L,\, v_{i,l}^U \,|\, v_{i,l},\, \mathsf{censoring}_{i,l}\big) \Big\},$$

where $p(\boldsymbol{\theta})$ is given by (9.5) for *Model U* and by (9.6) for *Model M*, respectively. Further, the terms $p\big(u_{i,l}^L,\, u_{i,l}^U \,|\, u_{i,l},\, \mathsf{censoring}_{i,l}\big)$ and $p\big(v_{i,l}^L,\, v_{i,l}^U \,|\, v_{i,l},\, \mathsf{censoring}_{i,l}\big)$, where $\mathsf{censoring}_{i,l}$ represents a realization of the random variable(s) causing the censoring of the (i,l)th onset and failure time, are the same as in Section 8.2.4, with an obvious change in notation.

9.3 Markov chain Monte Carlo

As indicated in Section 4.5 we base the inference on a sample from the posterior distribution obtained using MCMC methods. Here, Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) was chosen necessitating to sample from all full conditional distributions of blocks of model parameters. Below, the full conditional distributions are discussed.

9.3.1 Updating the parameters related to the penalized mixture $\mathcal{G}$

Let y_{i^*}, $i^* = 1, \ldots, n$ be the current values of the appropriate generic nodes y and r_{i^*}, $i^* = 1, \ldots, n$ corresponding latent allocation variables. That is,

- For $\mathcal{G}^{\varepsilon}$ we have $\{y_{i^*} : i^* = 1, \ldots, n\} = \{\varepsilon_{i,l} : i = 1, \ldots, N,\, l = 1, \ldots, n_i\}$, $\{r_{i^*} : i^* = 1, \ldots, n\} = \{r_{i,l}^{\varepsilon} : i = 1, \ldots, N,\, l = 1, \ldots, n_i\}$, and $n = \sum_{i=1}^{N} n_i$;
- For $\mathcal{G}^{\zeta}$ we have $\{y_{i^*} : i^* = 1, \ldots, n\} = \{\zeta_{i,l} : i = 1, \ldots, N,\, l = 1, \ldots, n_i\}$, $\{r_{i^*} : i^* = 1, \ldots, n\} = \{r_{i,l}^{\zeta} : i = 1, \ldots, N,\, l = 1, \ldots, n_i\}$, and $n = \sum_{i=1}^{N} n_i$;
- For $\mathcal{G}^{b}$ we have $\{y_{i^*} : i^* = 1, \ldots, n\} = \{b_i : i = 1, \ldots, N\}$, $\{r_{i^*} : i^* = 1, \ldots, n\} = \{r_i^b : i = 1, \ldots, N\}$, and $n = N$;

- For $\mathcal{G}^d$ we have $\{y_{i^*} : i^* = 1, \ldots, n\} = \{d_i : i = 1, \ldots, N\}$, $\{r_{i^*} : i^* = 1, \ldots, n\} = \{r_i^d : i = 1, \ldots, N\}$, and $n = N$.

Full conditional for transformed mixture weights

The full conditional of each element of $\boldsymbol{a}$ is given by

$$p(a_j \mid \cdots) \propto \frac{\exp(N_j\, a_j)}{\Big\{\sum_{k=-K}^{K} \exp(a_k)\Big\}^n} \times \exp\left[-\frac{\Big\{a_j - \mathrm{E}\big(a_j \mid \boldsymbol{a}_{-(j)},\, \lambda\big)\Big\}^2}{2\ \mathrm{var}\big(a_j \mid \boldsymbol{a}_{-(j)},\, \lambda\big)}\right], \quad j = -K, \ldots, K, \tag{9.8}$$

where N_j is the number of y_{i^*} for which the latent allocation variable r_{i^*} is equal to j, i.e.

$$N_j = \sum_{i^*=1}^{n} I[r_{i^*} = j].$$

Further, $\mathrm{E}\big(a_j \mid \boldsymbol{a}_{-(j)},\, \lambda\big)$ and $\mathrm{var}\big(a_j \mid \boldsymbol{a}_{-(j)},\, \lambda\big)$ are the mean and the variance resulting from the GMRF prior (9.7). For example, for the third order differences ($s = 3$), which have been used in all applications in this thesis (Sections 9.7 and 9.8), we have

$$\begin{aligned}
\mathrm{E}\big(a_j \,\big|\, \boldsymbol{a}_{-(j)}\big) &= \frac{a_{j-3} - 6\,a_{j-2} + 15\,a_{j-1} + 15\,a_{j+1} - 6\,a_{j+2} + a_{j+3}}{20}, \\
j &= -K+3, \ldots, K-3, \\
\mathrm{E}\big(a_{-K+2} \,\big|\, \boldsymbol{a}_{-(-K+2)}\big) &= \frac{-3\,a_{-K} + 12\,a_{-K+1} + 15\,a_{-K+3} - 6\,a_{-K+4} + a_{-K+5}}{19}, \\
\mathrm{E}\big(a_{K-2} \,\big|\, \boldsymbol{a}_{-(K-2)}\big) &= \frac{-3\,a_{K} + 12\,a_{K-1} + 15\,a_{K-3} - 6\,a_{K-4} + a_{K-5}}{19}, \\
\mathrm{E}\big(a_{-K+1} \,\big|\, \boldsymbol{a}_{-(-K+1)}\big) &= \frac{3\,a_{-K} + 12\,a_{-K+2} - 6\,a_{-K+3} + a_{-K+4}}{10}, \\
\mathrm{E}\big(a_{K-1} \,\big|\, \boldsymbol{a}_{-(K-1)}\big) &= \frac{3\,a_{K} + 12\,a_{K-2} - 6\,a_{K-3} + a_{K-4}}{10}, \\
\mathrm{E}\big(a_{-K} \,\big|\, \boldsymbol{a}_{-(-K)}\big) &= 3\,a_{-K+1} - 3\,a_{-K+2} + a_{-K+3}, \\
\mathrm{E}\big(a_{K} \,\big|\, \boldsymbol{a}_{-(K)}\big) &= 3\,a_{K-1} - 3\,a_{K-2} + a_{K-3},
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{var}\big(a_j \,\big|\, \boldsymbol{a}_{-(j)}\big) &= (20\,\lambda)^{-1}, \qquad j = -K+3, \ldots, K-3,\\
\operatorname{var}\big(a_{-K+2} \,\big|\, \boldsymbol{a}_{-(-K+2)}\big) &= \operatorname{var}\big(a_{K-2} \,\big|\, \boldsymbol{a}_{-(K-2)}\big) = (19\,\lambda)^{-1},\\
\operatorname{var}\big(a_{-K+1} \,\big|\, \boldsymbol{a}_{-(-K+1)}\big) &= \operatorname{var}\big(a_{K-1} \,\big|\, \boldsymbol{a}_{-(K-1)}\big) = (10\,\lambda)^{-1},\\
\operatorname{var}\big(a_{-K} \,\big|\, \boldsymbol{a}_{-(-K)}\big) &= \operatorname{var}\big(a_{K} \,\big|\, \boldsymbol{a}_{-(K)}\big) = \lambda^{-1}.
\end{aligned}$$

Distribution (9.8) is log-concave so we experimented both with the slice sampler of Neal (2003) as well as with the adaptive rejection sampling (ARS) method of Gilks and Wild (1992) to update the elements of $\boldsymbol{a}$. However, in our applications no method was found to be superior with respect to the performance of the MCMC. The results presented in Sections 9.7 and 9.8 were obtained using slice sampling.

Furthermore, it is seen that the full conditional distribution for each transformed mixture weight depends only on the weights of the neighboring mixture components. For a better performance of the MCMC, especially to decrease the autocorrelation of the sampled chain, it is thus advantageous to update in one iteration of the MCMC the transformed mixture weights in such an order that the full conditional of a we are updating does not depend on a which has just been updated. This is obtained, for example, using the following update order:

$$\cdots \to a_0 \to a_{s+1} \to a_{2(s+1)} \to \cdots \to a_1 \to a_{1+s+1} \to a_{1+2(s+1)} \to \cdots .$$

Full conditional for the smoothing parameter

For the smoothing parameter λ, the full conditional distribution is Gamma $(h^*_{\lambda,1},\, h^*_{\lambda,2})$ where

$$h^*_{\lambda,1} = h_{\lambda,1} + \frac{2K+1-s+1}{2}, \qquad h^*_{\lambda,2} = h_{\lambda,2} + \frac{1}{2}\,\boldsymbol{a}'\mathbb{P}'\mathbb{P}\boldsymbol{a}.$$

Full conditional for the mixture intercept

The full conditional for the mixture intercept α is a normal distribution with the mean and variance

$$\begin{aligned}
\mathrm{E}(\alpha \,|\, \cdots) &= \operatorname{var}(\alpha \,|\, \cdots) \,\times\, \Big\{(\sigma\tau)^{-2} \sum_{i^*=1}^{n} (y_{i^*} - \tau\mu_{r_{i^*}}) \,+\, \psi_\alpha^{-1}\nu_\alpha\Big\},\\
\operatorname{var}(\alpha \,|\, \cdots) &= \big\{(\sigma\tau)^{-2} n \,+\, \psi_\alpha^{-1}\big\}^{-1},
\end{aligned}$$

respectively and is thus easily sampled from.

Full conditional for the mixture scale

The full conditional distribution of τ^{-2} has the form

$$p(\tau^{-2} \mid \cdots) \propto (\tau^{-2})^{\xi_1 - 1} \exp\Big(\xi_3 \sqrt{\tau^{-2}} - \xi_2 \tau^{-2}\Big), \tag{9.9}$$

with

$$\begin{aligned}
\xi_1 &= h_{\tau,1} + 0.5\, n, \\
\xi_2 &= h_{\tau,2} + 0.5\, \sigma^{-2} \sum_{i^*=1}^{n} (y_{i^*} - \alpha)^2, \\
\xi_3 &= \sigma^{-2} \sum_{i^*=1}^{n} \mu_{r_{i^*}} (y_{i^*} - \alpha).
\end{aligned}$$

Distribution (9.9) is generally not log-concave so that the adaptive rejection sampling (ARS) method of Gilks and Wild (1992), successfully used in many situations when the full conditional distribution does not have a standard form, cannot be used here. However, it can easily be shown that the density (9.9) is always unimodal and the slice sampler of Neal (2003) can be used to update the parameter τ^{-2} in an MCMC run.

Full conditional for the allocation variables

The full conditional for each allocation variable r_{i^*}, $i^* = 1, \ldots, n$ is discrete with

$$\Pr(r_{i^*} = j \mid \cdots) \propto w_j \exp\Big\{-\frac{(y_{i^*} - \alpha - \tau\mu_j)^2}{2(\sigma\tau)^2}\Big\}, \qquad j \in \{-K, \ldots, K\}.$$

9.3.2 Updating the generic node Y

The update of the generic node Y is of two types: (1) update of the residuals $\varepsilon_{i,l}$, $\zeta_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ (2) update of the univariate random effects b_i, d_i, $i = 1, \ldots, N$ in *Model U*.

Updating the residuals

The update of the 'onset' residuals $\zeta_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ is fully deterministic provided the (i,l)th onset time $u_{i,l} = u_{i,l}^L = u_{i,l}^U$ is uncensored. The update of $\zeta_{i,l}$ consists then of using the AFT expression (9.1) with the

current values of the parameters, i.e. the updated $\zeta_{i,l}$ is equal to $\log(u_{i,l}) - \boldsymbol{\delta}'\boldsymbol{x}^u_{i,l} - \boldsymbol{d}'_i\boldsymbol{z}^u_{i,l}$.

When the (i,l)th onset time is interval-censored with an observed interval $\lfloor u^L_{i,l},\, u^U_{i,l}\rfloor$, its update consists of the sampling from a truncated normal distribution, namely

$$\mathcal{N}\Big(\alpha^\zeta + \tau^\zeta \mu_{r^\zeta_{i,l}},\, (\sigma^\zeta\tau^\zeta)^2\Big) \text{ truncated on}$$
$$\Big]\log(u^L_{i,l}) - \boldsymbol{\delta}'\boldsymbol{x}^u_{i,l} - \boldsymbol{d}'_i\boldsymbol{z}^u_{i,l},\, \log(u^U_{i,l}) - \boldsymbol{\delta}'\boldsymbol{x}^u_{i,l} - \boldsymbol{d}'_i\boldsymbol{z}^u_{i,l}\Big].$$

A similar procedure is used when updating the 'event' residuals $\varepsilon_{i,l}$, $i = 1,\ldots,N$, $l = 1,\ldots,n_i$. It is useful to stress that for the update of $\varepsilon_{i,l}$ also the 'onset' residual $\zeta_{i,l}$ and subsequently also the true onset time $u_{i,l} = \exp(\boldsymbol{\delta}'\boldsymbol{x}^u_{i,l} + \boldsymbol{d}'_i\boldsymbol{z}^u_{i,l} + \zeta_{i,l})$ make a part of the condition when exploiting the full conditional distribution. This implies that the update of $\varepsilon_{i,l}$ is fully deterministic provided the (i,l)th failure time $v_{i,l} = v^L_{i,l} = v^U_{i,l}$ is uncensored, irrespective whether the onset time is censored or not. The update of $\varepsilon_{i,l}$ consists then of using the AFT expression (9.2) with the current values of the parameters, i.e. the updated $\varepsilon_{i,l}$ is equal to $\log(v_{i,l} - u_{i,l}) - \boldsymbol{\beta}'\boldsymbol{x}^t_{i,l} - \boldsymbol{b}'_i\boldsymbol{z}^t_{i,l}$.

When the residual $\varepsilon_{i,l}$ corresponds to the censored failure time with an observed interval $\lfloor v^L_{i,l},\, v^U_{i,l}\rfloor$ its update consists of the sampling from the full conditional distribution of $\varepsilon_{i,l}$ which is here a truncated normal distribution, namely

$$\mathcal{N}\Big(\alpha^\varepsilon + \tau^\varepsilon \mu_{r^\varepsilon_{i,l}},\, (\sigma^\varepsilon\tau^\varepsilon)^2\Big) \text{ truncated on}$$
$$\Big]\log(v^L_{i,l} - u_{i,l}) - \boldsymbol{\beta}'\boldsymbol{x}^t_{i,l} - \boldsymbol{b}'_i\boldsymbol{z}^t_{i,l},\, \log(v^U_{i,l} - u_{i,l}) - \boldsymbol{\beta}'\boldsymbol{x}^t_{i,l} - \boldsymbol{b}'_i\boldsymbol{z}^t_{i,l}\Big].$$

Updating the univariate random effects in Model U

In *Model U*, the full conditional distributions for the univariate random effects b_i and/or d_i, $i = 1,\ldots,N$ are normal distributions, namely

$$b_i \mid \cdots \sim \mathcal{N}\Big(\mathrm{E}(b_i \mid \cdots), \mathrm{var}(b_i \mid \cdots)\Big), \qquad i = 1,\ldots,N,$$

with

$$\mathrm{E}(b_i \mid \cdots) = \mathrm{var}(b_i \mid \cdots) \times$$
$$\Big[(\sigma^b\tau^b)^{-2}\,\tau^b\mu^b_{r^b_i} \;+\; (\sigma^\varepsilon\tau^\varepsilon)^{-2}\sum_{l=1}^{n_i}\big\{\log(t_{i,l}) - \alpha^\varepsilon - \boldsymbol{\beta}'\boldsymbol{x}^t_{i,l} - \tau^\varepsilon\mu^\varepsilon_{r^\varepsilon_{i,l}}\big\}\Big],$$
$$\mathrm{var}(b_i \mid \cdots) = \Big\{(\sigma^b\tau^b)^{-2} \;+\; (\sigma^\varepsilon\tau^\varepsilon)^{-2}\,n_i\Big\}^{-1},$$

Analogous formulas, with an obvious change in notation, hold for d_i, $i = 1, \ldots, N$.

9.3.3 Updating the parameters related to the multivariate random effects in Model M

In the case of the multivariate random effects $\boldsymbol{b}_i$ and/or $\boldsymbol{d}_i$ having a multivariate normal prior distribution the following full conditionals are used to update the related parameters.

Full conditionals for the multivariate random effects $\boldsymbol{b}_i$ and $\boldsymbol{d}_i$

The full conditional of the multivariate random effects vector $\boldsymbol{b}_i$, $i = 1, \ldots, N$ is multivariate normal distribution, i.e.

$$\boldsymbol{b}_i \mid \cdots \sim \mathcal{N}\Big(\mathrm{E}(\boldsymbol{b}_i \mid \cdots), \mathrm{var}(\boldsymbol{b}_i \mid \cdots)\Big), \qquad i = 1, \ldots, N,$$

with

$$\begin{aligned}
\mathrm{E}(\boldsymbol{b}_i \mid \cdots) &= \mathrm{var}(\boldsymbol{b}_i \mid \cdots) \times \\
&\quad \Big[\mathbb{D}_b^{-1}\boldsymbol{\gamma}_b + (\sigma^{\varepsilon}\tau^{\varepsilon})^{-2} \sum_{l=1}^{n_i} \boldsymbol{z}_{i,l}^t \big\{\log(t_{i,l}) - \alpha^{\varepsilon} - \boldsymbol{\beta}'\boldsymbol{x}_{i,l}^t - \tau^{\varepsilon}\mu_{r_{i,l}^{\varepsilon}}^{\varepsilon}\big\}\Big], \\
\mathrm{var}(\boldsymbol{b}_i \mid \cdots) &= \Big\{\mathbb{D}_b^{-1} \; + \; (\sigma^{\varepsilon}\tau^{\varepsilon})^{-2} \sum_{l=1}^{n_i} \boldsymbol{z}_{i,l}^t (\boldsymbol{z}_{i,l}^t)'\Big\}^{-1}.
\end{aligned}$$

The full conditional distribution of the multivariate random effects $\boldsymbol{d}_i$, $i = 1, \ldots, N$ is analogous with an obvious change in notation.

Full conditionals for the means $\boldsymbol{\gamma}_b$, $\boldsymbol{\gamma}_d$ and the covariance matrices $\mathbb{D}_b$, $\mathbb{D}_d$ of the multivariate random effects

For the means $\boldsymbol{\gamma}_b$, $\boldsymbol{\gamma}_d$ and the covariance matrices $\mathbb{D}_b$, $\mathbb{D}_d$ of the multivariate random effects, the full conditional distributions are exactly the same as these derived for the Bayesian normal mixture CS AFT model in Section 8.3.2. Only appropriate subscripts have to be added to expressions appearing in formulas given in Section 8.3.2.

9.3.4 Updating the regression parameters

Full conditionals for the fixed effects $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$

Let $\boldsymbol{\beta}_{(S)}$ be an arbitrary sub-vector of vector $\boldsymbol{\beta}$, and $\boldsymbol{x}_{i,l(S)}$ the corresponding sub-vectors of covariate vectors $\boldsymbol{x}^t_{i,l}$, and further let $\boldsymbol{x}_{i,l(-S)}$ be their complementary sub-vectors. Similarly, let further $\boldsymbol{\nu}_{\beta(S)}$ and $\boldsymbol{\psi}_{\beta(S)}$ be appropriate sub-vectors of hyperparameters $\boldsymbol{\nu}_\beta$ and $\boldsymbol{\psi}_\beta$, respectively. Finally, let $\Psi_{\beta(S)} = \mathrm{diag}(\boldsymbol{\psi}_{\beta(S)})$. Then

$$\boldsymbol{\beta}_{(S)} \mid \cdots \sim \mathcal{N}\Big(\mathrm{E}(\boldsymbol{\beta}_{(S)} \mid \cdots),\ \mathrm{var}(\boldsymbol{\beta}_{(S)} \mid \cdots)\Big),$$

with

$$\begin{aligned}\mathrm{E}(\boldsymbol{\beta}_{(S)} \mid \cdots) &= \mathrm{var}(\boldsymbol{\beta}_{(S)} \mid \cdots)\times \\ &\qquad \Big\{\Psi^{-1}_{\beta(S)}\boldsymbol{\nu}_{\beta(S)} + (\sigma^\varepsilon\tau^\varepsilon)^{-2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\boldsymbol{x}_{i,l(S)}e^{(F)}_{i,l(S)}\Big\},\\ \mathrm{var}(\boldsymbol{\beta}_{(S)} \mid \cdots) &= \Big\{\Psi^{-1}_{\beta(S)} + (\sigma^\varepsilon\tau^\varepsilon)^{-2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\boldsymbol{x}_{i,l(S)}\boldsymbol{x}'_{i,l(S)}\Big\}^{-1},\end{aligned}$$

where $e^{(F)}_{i,l(S)} = \log(t_{i,l}) - \alpha^\varepsilon - \boldsymbol{\beta}'_{(-S)}\boldsymbol{x}_{i,l(-S)} - \boldsymbol{b}'_i\boldsymbol{z}^t_{i,l} - \tau^\varepsilon\mu^\varepsilon_{r^\varepsilon_{i,l}}$.
The full conditional distribution for an arbitrary subvector of the vector $\boldsymbol{\delta}$ is analogous with an obvious change in notation.

9.4 Bayesian estimates of the survival distribution

9.4.1 Predictive survival and hazard curves and predictive survival densities

Analogously to Section 8.4, the survival and hazard functions or the survival densities for a specific combination of covariates are estimated by the mean of (posterior) predictive quantities.

Almost all expressions given in Section 8.4.1 apply also here with the following changes. To get the Bayesian estimate of the survival function of the event time T, given the covariates $\boldsymbol{x}^t_{new}$ and $\boldsymbol{z}^t_{new}$, the expression (8.16) changes

into

$$S(t \mid \boldsymbol{\theta}, \boldsymbol{x}^t_{new}, \boldsymbol{z}^t_{new}) = \tag{9.10}$$
$$1 - \sum_{j=-K}^{K} w_j^{\varepsilon} \Phi\big\{\log(t) - \boldsymbol{\beta}'\boldsymbol{x}^t_{new} - \boldsymbol{b}'\boldsymbol{z}^t_{new} \,\big|\, \alpha^{\varepsilon} + \tau^{\varepsilon}\mu_j^{\varepsilon},\, (\sigma^{\varepsilon}\tau^{\varepsilon})^2\big\}.$$

Similarly, to get the estimate of the survival density, we use

$$p(t \mid \boldsymbol{\theta}, \boldsymbol{x}^t_{new}, \boldsymbol{z}^t_{new}) = \tag{9.11}$$
$$t^{-1} \sum_{j=-K}^{K} w_j^{\varepsilon} \varphi\big\{\log(t) - \boldsymbol{\beta}'\boldsymbol{x}^t_{new} - \boldsymbol{b}'\boldsymbol{z}^t_{new} \,\big|\, \alpha^{\varepsilon} + \tau^{\varepsilon}\mu_j^{\varepsilon},\, (\sigma^{\varepsilon}\tau^{\varepsilon})^2\big\}$$

instead of the expression (8.18).

To be able to use a relationship analogous to (8.17) we need a sample $\{\boldsymbol{b}^{(m)} : m = 1, \ldots, M\}$ of the posterior predictive values of the random effects. In the case of a univariate random effect b in *Model U*, $b^{(m)}$ is sampled from the normal mixture $\sum_{j=-K}^{K} w_j^{(m)} \mathcal{N}\Big(\tau^{b,(m)}\mu_j^b,\, (\sigma^b\tau^{b,(m)})^2\Big)$. In the case of a multivariate random effect $\boldsymbol{b}$ in *Model M*, $\boldsymbol{b}^{(m)}$ is sampled from $\mathcal{N}_{q_b}(\boldsymbol{\gamma}_b^{(m)}, \mathbb{D}_b^{(m)})$. The predictive quantities for the onset time U are obtained in an analogous manner.

9.4.2 Predictive error and random effect densities

The estimate of the smoothed densities g_ε, g_ζ, g_b, g_d is obtained by the mean of the (posterior) predictive density which is given, for example in the case of g_ε, by

$$\mathrm{E}\big\{g_\varepsilon(e) \,\big|\, \text{data}\big\} = \int g_\varepsilon(e) p(\boldsymbol{\theta} \mid \text{data})\, d\boldsymbol{\theta}, \qquad e \in \mathbb{R}. \tag{9.12}$$

The MCMC estimate of (9.12) is obtained by averaging the error density (9.3) over the MCMC run, i.e.

$$\hat{g}_\varepsilon(e) = \frac{1}{M} \sum_{m=1}^{M} \Big\{ \big(\tau^{\varepsilon,(m)}\big)^{-1} \sum_{j=-K}^{K} w_j^{\varepsilon,(m)}\, \varphi\Big(\frac{e - \alpha^{\varepsilon,(m)}}{\tau^{\varepsilon,(m)}} \,\Big|\, \mu_j^{\varepsilon},\, (\sigma^{\varepsilon})^2\Big)\Big\}. \tag{9.13}$$

9.5 Bayesian estimates of the individual random effects

As explained in Section 8.5 in the context of the Bayesian normal mixture CS AFT model, in some situation estimates of the individual random effects

must be provided. In the context of this chapter these can be computed in the same way as shown in Section 8.5.

9.6 Simulation study

To validate our approach we conducted a simulation study which mimics to a certain extent the Signal Tandmobiel® data. From each of 150 clusters we simulated 4 observations. The onset time $U_{i,l}$ and the event time $T_{i,l}$, $i = 1, \ldots, 150$, $l = 1, \ldots, 4$ were generated according to the AFT models (9.1) and (9.2) with $\boldsymbol{x}^u_{i,l} = (x^u_{i,l,1}, x^u_{i,l,2})'$, $\boldsymbol{z}^u_{i,l} \equiv 1$, $\boldsymbol{\delta} = (0.20, -0.10)'$ and $\boldsymbol{x}^t_{i,l} = (x^t_{i,l,1}, x^t_{i,l,2})'$, $\boldsymbol{z}^t_{i,l} \equiv 1$, $\boldsymbol{\beta} = (0.30, -0.15)'$. The covariates $x^u_{i,l,1}$ and $x^t_{i,l,1}$ are continuous and generated independently from a uniform distribution on (0, 1), the covariates $x^u_{i,l,2}$ and $x^t_{i,l,2}$ are binary with the equal probabilities for zeros and ones.

The error terms $\zeta_{i,l}$ and $\varepsilon_{i,l}$ are obtained from $\zeta_{i,l} = \alpha^\zeta + \tau^\zeta \zeta^*_{i,l}$ ($\alpha^\zeta = 1.75$, $\zeta^*_{i,l} \sim g^*_\zeta$) and $\varepsilon_{i,l} = \alpha^\varepsilon + \tau^\varepsilon \varepsilon^*_{i,l}$ ($\alpha^\varepsilon = 2.00$, $\varepsilon^*_{i,l} \sim g^*_\varepsilon$), respectively. Further, the random effects d_i and b_i are obtained from $d_i = \tau^d d^*_i$ ($d^*_i \sim g^*_d$) and $b_i = \tau^b b^*_i$ ($b^*_i \sim g^*_b$), respectively. The scale parameters were chosen such that $(\tau^d)^2 + (\tau^\zeta)^2 = \tau^2_{onset} = 0.1$ and $(\tau^b)^2 + (\tau^\varepsilon)^2 = \tau^2_{event} = 1.0$, see below for the individual values. The choice of τ^2_{onset} and τ^2_{event} was motivated by the results of the analysis in Section 9.7.

Two scenarios for the distributional parts of the model were considered. In scenario I, both densities g^*_ζ and g^*_ε (of the standardized error terms) are a mixture of normals, i.e. equal to $0.4\mathcal{N}(-2.000, 0.25) + 0.6\mathcal{N}(1.333, 0.36)$ standardized to have unit variance. For the densities g^*_d and g^*_b (of the standardized random effects) the density of a standardized extreme value of minimum distribution was taken. In scenario II, we reversed the setting, i.e. we have taken an extreme value distribution for the error terms and a normal mixture for the random effects. Additionally, within each scenario, the variances τ^2_{onset} and τ^2_{event} were decomposed such that the ratios $\tau^d/\tau^\zeta = \tau^b/\tau^\varepsilon$ were equal to 5, 3, 2, 1, 1/2, 1/3, and 1/5, respectively.

The true onset and event times were interval-censored by simulating the 'visit' times for each subject in the data set. The first visit was drawn from $\mathcal{N}(1, 0.2^2)$. Each of the distances between the consecutive visits was drawn from $\mathcal{N}(0.5, 0.05^2)$.

The results for the simulation study are shown in Appendix B.3. Tables B.12 and B.13 give the results for the regression parameters and show that they are estimated practically unbiasedly and with a reasonable precision. It is further seen that the precision of the estimation decreases when the

within-cluster variability (variance of the error terms) increases compared to the between-cluster variability (variance of the random effects). In practice however, the between-cluster variability is often much higher than the within-cluster variability. Further, Tables B.14 and B.15 show results for the standard deviations of the error terms and random effects. Here, the precision is sometimes somewhat worse. However, also the standard deviations are, in most cases, estimated with minimal bias. Furthermore, the shape of the survival functions or survival densities is correctly estimated as is illustrated in Figures B.10–B.17 which show results for the fitted survival functions and survival densities for selected combinations of covariates.

9.7 Example: Signal Tandmobiel® study – clustered doubly-interval-censored data

This analysis of the Signal Tandmobiel® data, introduced in Section 1.1, involves

(a) doubly-interval-censored data, i.e. the time from tooth emergence to onset of caries;

(b) clustering. Indeed, we will examine several teeth jointly and the teeth from the same mouth are related.

The primary interest of the present analysis is to address the influence of *sound* versus affected (*decayed/filled/missing due to caries*) deciduous second molars (in Figure 1.2, teeth 55, 65, 75, 85, respectively) on the caries susceptibility of the adjacent permanent first molars (in Figure 1.1, teeth 16, 26, 36, 46, respectively). Note that for about five years the deciduous second molars are in the mouth together with the permanent first molars.

It is possible that the caries processes on the primary and the permanent molar occur simultaneously. In this case it is difficult to know whether caries on the deciduous molar caused caries on the permanent molar or vice versa. For this reason, the permanent first molar was excluded from the analysis if caries was present when emergence was recorded. This implies that the data are not balanced with respect to the size of the clusters. In total, 3 520 children were included in the analysis of which 187 contributed 1 tooth, 317 2 teeth, 400 3 teeth and 2 616 all 4 teeth.

Additionally, we considered the impact of gender (*boy/girl*), presence of sealants in pits and fissures of the permanent first molar (*none/present*), occlusal plaque accumulation on the permanent first molar (*none/in pits and*

fissures/on total surface), and reported oral brushing habits (*not daily/daily*). Note that pits and fissures sealing is a preventive action which is expected to protect the tooth against caries development. The presence of plaque on the occlusal surfaces of the permanent first molars was assessed using a simplified version of the index described by Carvalho, Ekstrand, and Thylstrup (1989). All explanatory variables were obtained at the examination where the presence of the permanent first molar was first recorded.

The choice of explanatory variables is motivated by the results of Leroy et al. (2005) where a GEE multivariate log-logistic AFT model was used to analyze the time to caries. Multiple imputation was used to deal with the interval-censored emergence times. Further, on top of that, the caries status of the deciduous first molars (in Figure 1.2, teeth 54, 64, 74, 84, respectively) was included in the covariate part of the model. We will not use this factor as an explanatory variable due to its high dependence with the status of the deciduous second molar (in all quadrants of the mouth, the χ^2 test statistics with 9 degrees of freedom exceeded 1 100).

The onset time $U_{i,l}$, $l = 1, \ldots, 4$ is the age (in years) of the ith child (ith cluster) at which the lth permanent first molar emerged. The failure time, $V_{i,l}$, indicates the onset of caries of the lth permanent first molar. The time from tooth emergence to the onset of caries, $T_{i,l}$, is doubly-interval-censored. Here, both the time of tooth emergence and the onset of caries experience are only known to lie in an interval of about 1 year.

Further, in our example about 85% of the permanent first molars had emerged at the first examination giving rise to a huge amount of left-censored onset times. However, at each examination the permanent teeth were scored according to their clinical eruption stage using a grading that starts at P0 (tooth not visible in the mouth) and ends with P4 (fully erupted tooth with full occlusion). Based on the clinical eruption stage at the moment of the first examination, all left-censored emergence times were transformed into interval-censored ones with the lower limit of the observed interval equal to the age at examination minus 0.25 year, 0.5 year and 1 year, respectively for the teeth with the eruption stage P1, P2 and P3, respectively and with the lower limit equal to 5 years for the teeth with the eruption stage P4. We refer to Leroy et al. (2005) for details and motivation.

9.7.1 Basic Model

The analysis starts with the **Basic Model** where we allowed for a different effect of the covariates on both emergence and caries experience for the four permanent first molars. Namely, the Basic Model was based on the AFT

models (9.1) and (9.2) with the covariate vector $\boldsymbol{x}_{i,l}^u$ for emergence:

$$\begin{aligned}\boldsymbol{x}_{i,l}^u = (&\mathsf{gender}_i,\ \mathsf{tooth26}_{i,l},\ \mathsf{tooth36}_{i,l},\ \mathsf{tooth46}_{i,l},\\ &\mathsf{tooth26}_{i,l} * \mathsf{gender}_i,\ \mathsf{tooth36}_{i,l} * \mathsf{gender}_i,\ \mathsf{tooth46}_{i,l} * \mathsf{gender}_i)',\end{aligned}$$

and the covariate vector $\boldsymbol{x}_{i,l}^t$ for caries:

$$\begin{aligned}\boldsymbol{x}_{i,l}^t = (&\tilde{\boldsymbol{x}}_{i,l}^t,\ \mathsf{tooth26}_{i,l},\ \mathsf{tooth36}_{i,l},\ \mathsf{tooth46}_{i,l},\\ &\mathsf{tooth26}_{i,l} * \tilde{\boldsymbol{x}}_{i,l}^t,\ \mathsf{tooth36}_{i,l} * \tilde{\boldsymbol{x}}_{i,l}^t,\ \mathsf{tooth46}_{i,l} * \tilde{\boldsymbol{x}}_{i,l}^t)',\end{aligned}$$

where

$$\begin{aligned}\tilde{\boldsymbol{x}}_{i,l}^t = (&\mathsf{gender}_i,\ \mathsf{statusD}_{i,l},\ \mathsf{statusF}_{i,l},\ \mathsf{statusM}_{i,l},\\ &\mathsf{brushing}_i,\ \mathsf{sealants}_{i,l},\ \mathsf{plaquePF}_{i,l},\ \mathsf{plaqueT}_{i,l}).\end{aligned}$$

The covariates tooth26, tooth36, tooth46 are dummies for the position of the permanent first molar with the molar 16 as the baseline, the covariate gender equals 1 for *boys* and equals 0 for girls. The covariates statusD, statusF, statusM are dummies for the status of the adjacent deciduous molar: *decayed, filled, missing due to caries* with *sound* being the baseline. The covariate brushing is dichotomous (1 = *daily*, 0 = *not daily*) as well as the covariate sealants (1 = *present*, 0 = *not present*). Finally, the covariates plaquePF and plaqueT are dummies for the plaque accumulation: *in pits and fissures, on total surface* with *no plaque* as the baseline.

To account for clustering, univariate child-specific random effects d_i and b_i are included in the model expressions (9.1) and (9.2), respectively with $z_{i,l}^u = z_{i,l}^t \equiv 1$. Finally, analogously to Sections 7.7 and 8.7, we subtracted 5 years from all observed times, i.e. $\log(U_{i,l} - 5)$ was used in the left-hand side of the model formula (9.1).

As discussed already in Section 9.1, our model assumes that, given the covariates and child-specific random effects, the emergence time $U_{i,l}$ and the time to caries $T_{i,l}$ are independent for each i and l. Specifically, we assume that the caries process on a specific tooth only depends on the time when that tooth is at risk for caries and not on the chronological time. This assumption seems reasonable for the Signal Tandmobiel® data taking into account the results of Leroy et al. (2005) who evaluated also the effect of the emergence time on the time to caries and found it non-significant ($p = 0.78$).

9.7.2 Final Model

Based on the results for the Basic Model (see below) we fitted the **Final Model** where we omitted all two-way interactions with the covariates

tooth26, tooth36, tooth46. Additionally, we binarized the covariates statusD, statusF, statusM into a new covariate status which was equal to 1 for *decayed, filled* or *missing due to caries* deciduous molars and was equal to 0 for *sound* deciduous molars. Also the covariates plaquePF and plaqueT were binarized into the covariate plaque equal to 1 for the teeth with plaque present either *in pits and fissures* or *on total surface* and equal to 0 otherwise. That is, the onset and event covariate vectors are equal to

$$\begin{aligned}\boldsymbol{x}^u_{i,l} &= (\mathsf{gender}_i, \mathsf{tooth26}_{i,l}, \mathsf{tooth36}_{i,l}, \mathsf{tooth46}_{i,l})',\\ \boldsymbol{x}^t_{i,l} &= (\mathsf{gender}_i, \mathsf{status}_{i,l}, \mathsf{brushing}_i, \mathsf{sealants}_{i,l}, \mathsf{plaque}_{i,l},\\ &\qquad \mathsf{tooth26}_{i,l}, \mathsf{tooth36}_{i,l}, \mathsf{tooth46}_{i,l})'.\end{aligned}$$

9.7.3 Prior distribution

Firstly, for all penalized mixtures we used the same grid of equidistant knots of length 31 ($K = 15$) defined on $[-4.5, 4.5]$ with the basis standard deviation $\sigma = 2(\mu_j - \mu_{j-1})/3 = 0.2$. Secondly, the third order difference ($s = 3$) was used in the prior (9.7). Further, the prior distributions of the nodes in DAGs (Figures 9.1 and 9.3) without parents were taken highly dispersed. That is all λ and τ^{-2} parameters were a priori Gamma(1, 0.005) distributed, all α, β and δ parameters were given a $\mathcal{N}(0, 100)$ prior.

9.7.4 Results

For each considered model we ran 500 000 iterations with 1:3 thinning which took about 44 hours on a 3 GHz Pentium IV PC with 1 GB RAM. We kept the last 100 000 iterations for inference.

Results for the Basic Model

The analysis of the **Basic Model** revealed that all interaction terms with tooth covariates are redundant implying that the effect of all these covariates is the same for all four permanent first molars. To evaluate this we used simultaneous Bayesian p-values computed using the method described in Section 4.6.2. For the emergence part, the simultaneous p-value for the tooth:gender interactions is higher than 0.5. For the caries part of the model, the p-values are higher than 0.5 for the interactions of tooth with gender and plaque and higher than 0.1 for the interactions with brushing, sealants and status. Also the covariate tooth is not significant however we kept it in the

Table 9.1: Signal Tandmobiel® study, Final Model. Posterior medians, 95% equal-tail credible regions (CR) and Bayesian two-sided p-values for the model parameters. For the parameter Tooth the CR and the p-value are simultaneous.

	Emergence		Caries	
Parameter	Posterior median	95% CR	Posterior median	95% CR
Tooth		$p > 0.5$		$p > 0.5$
tooth 26	−0.003	(−0.013, 0.007)	−0.006	(−0.045, 0.031)
tooth 36	0.001	(−0.008, 0.011)	−0.009	(−0.051, 0.034)
tooth 46	0.002	(−0.008, 0.012)	−0.016	(−0.059, 0.026)
Gender		$p = 0.008$		$p = 0.085$
girl	−0.023	(−0.039, −0.007)	−0.071	(−0.155, 0.009)
Status				$p < 0.001$
dmf			−0.140	(−0.193, −0.091)
Brushing				$p < 0.001$
daily			0.337	(0.233, 0.436)
Sealants				$p < 0.001$
present			0.119	(0.060, 0.178)
Plaque				$p < 0.001$
present			−0.114	(−0.171, −0.067)
E(error)	0.442	(0.427, 0.456)	1.920	(1.810, 2.059)
sd(error)	0.029	(0.025, 0.034)	0.767	(0.712, 0.834)
sd(random)	0.199	(0.191, 0.210)	0.672	(0.614, 0.734)

model to address the question whether the emergence and caries timing are the same for the four permanent first molars.

Further, for none of the four permanent first molars a significant difference was found between the status groups *decayed, filled* or *missing*, and between the plaque groups *present in pits and fissures* or *present on total surface*. This finding, together with the fact that the group with *extracted* deciduous molar and the group with the plaque *present on total surface* had very low prevalence (1.45% and 3.13%, respectively), led to the simplification of these two covariates in the Final Model.

Results for the Final Model

Table 9.1 shows posterior medians, 95% equal-tail credible intervals and Bayesian two-sided p-values for the parameters in the **Final Model**. It is seen that neither for the emergence and nor for the caries process there is a significant difference between the four permanent first molars. However, the molars of girls emerge significantly earlier than those of boys. With respect to caries experience, the difference between boys and girls is not significant at 5%. However all remaining covariates have a significant impact on the caries process. Namely, daily brushing increases the time to caries with a factor of $\exp(0.337) = 1.40$ compared to less frequent brushing. Presence of sealants increases the time to caries with a factor of $\exp(0.119) = 1.13$. On the other hand, the presence of the plaque decreases the time to caries with a factor of $\exp(-0.114) = 0.89$ and the fact that the neighboring deciduous second molar is either decayed, filled or extracted due to caries decreases the time time to caries with a factor of $\exp(-0.140) = 0.87$.

Figure 9.4 shows the posterior predictive survival and hazard functions for the time to caries on the upper right permanent first molar of boys, for 'the best', 'the worst' and two intermediate combinations of covariates (the curves for the remaining teeth and girls are similar). It is seen that when the teeth are daily brushed, plaque-free and sealed the hazard for caries starts to increase approximately 1 year after emergence however then remains almost constant. Whereas, when the teeth are not brushed daily and are exposed to other risk factors the hazard starts to increase already approximately 6 months after emergence. After a period of constant risk then the hazard starts to increase again.

The peak in the hazard for caries approximately 1 year after emergence was observed also by Leroy et al. (2005) and can be explained by the fact that teeth are most vulnerable for caries soon after the emergence when the enamel is not yet fully developed. This peak is also present, although with a different size and with a slight shift, for all covariate combinations. On the other hand, for covariate combinations reflecting good oral health and hygiene habits, the hazard remains almost constant after the initial period of highly increasing risk whereas for combinations of covariates reflecting bad oral conditions the hazard starts to increase again approximately 3 years after emergence. This shows clearly the relationship between caries experience and oral health and hygiene habits.

Finally, Figure 9.5 shows Bayesian predictive error and random effect density estimates. The estimate of the emergence random effect density g_d suggests the children could be divided, even after conditioning on gender, into two

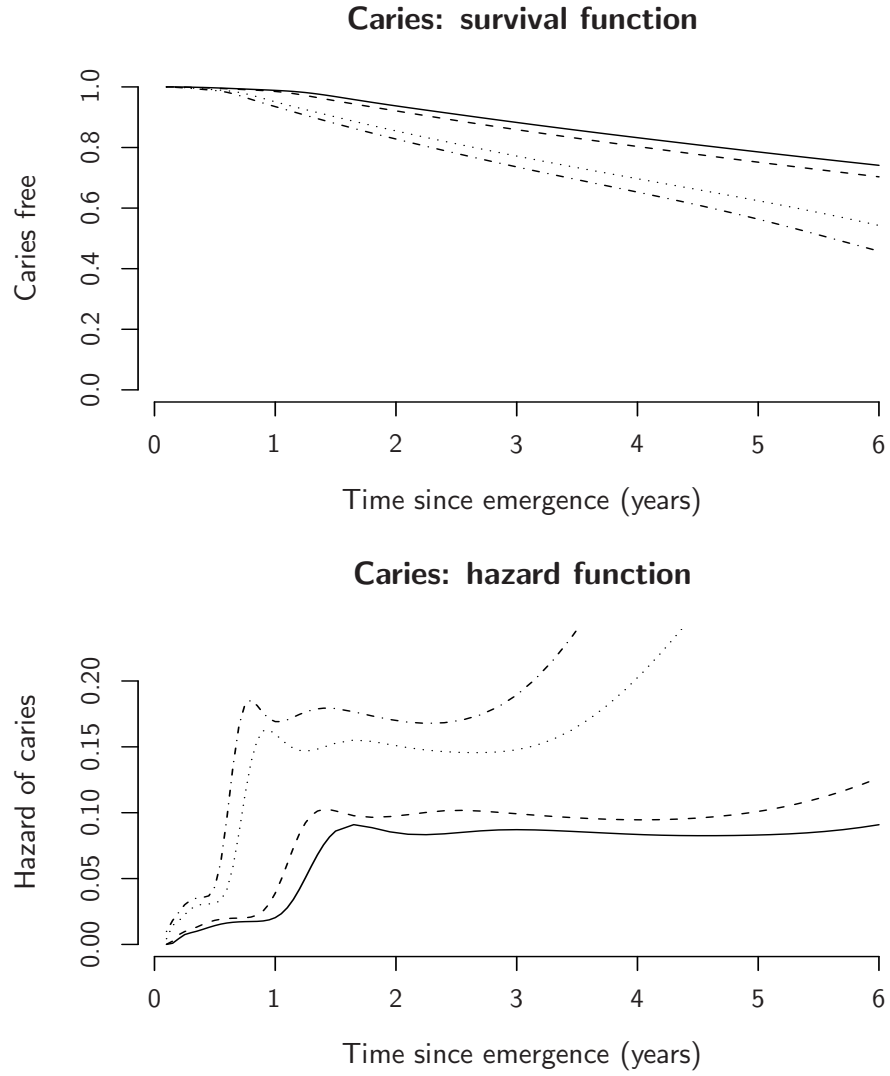


Figure 9.4: Signal Tandmobiel® study, Final Model. Posterior predictive caries free (survival) and caries hazard curves for tooth 16 of boys and the following combinations of covariates: solid and dashed lines for no plaque, present sealing, daily brushing and sound primary second molar (solid line) or dmf primary second molar (dashed line), dotted and dotted-dashed lines for present plaque, no sealing, not daily brushing and sound primary second molar (dotted line) or dmf primary second molar (dotted-dashed line).

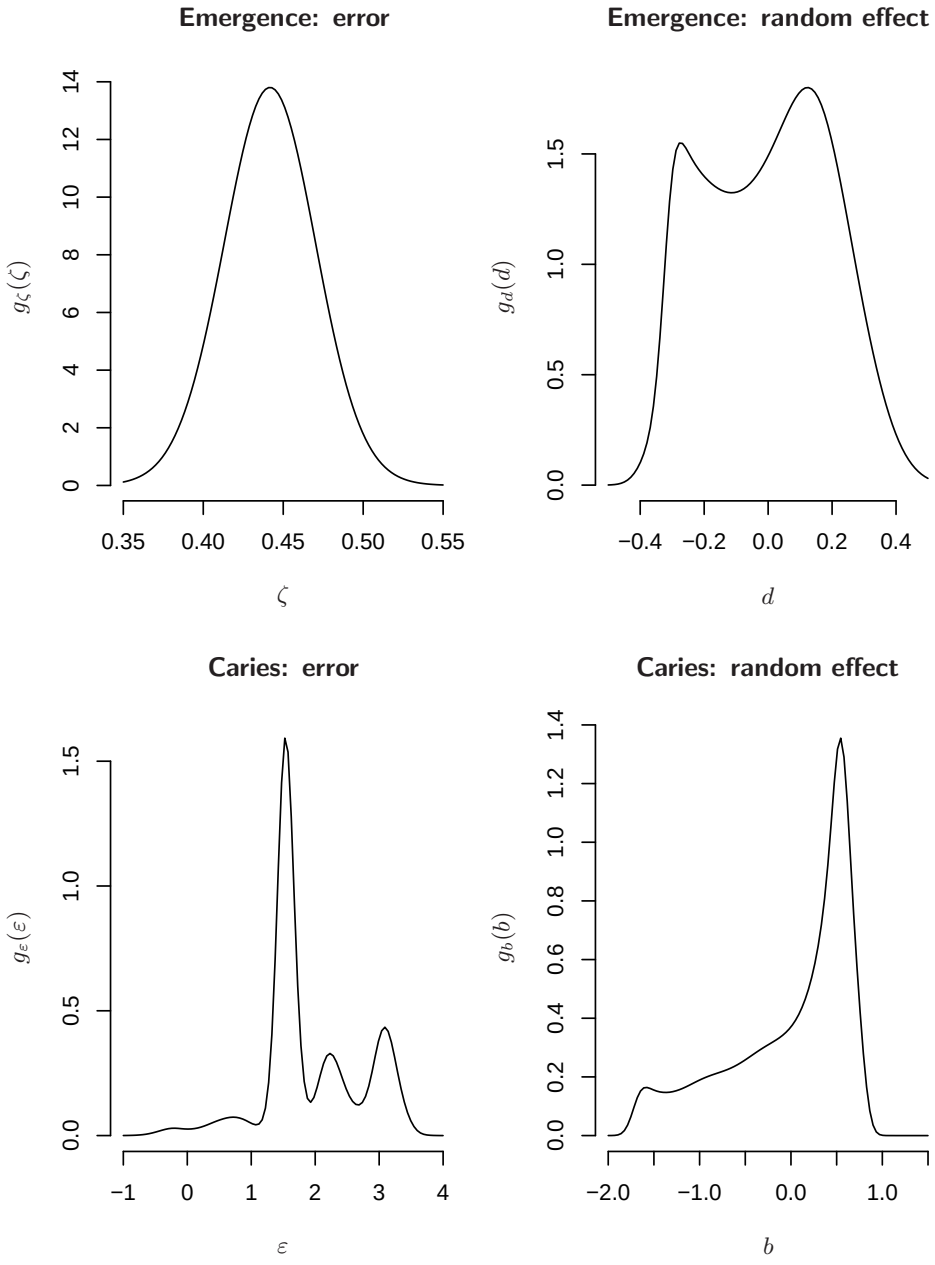


Figure 9.5: Signal Tandmobiel® study, Final Model. Estimates of the densities of the error terms and random effects.

groups: early and late emergers. Also, the children can be divided into two groups with respect to caries sensitivity (see random effect density g_b). Finally, as the estimate of the caries error density g_ε shows three modes it seems that there are other important factors influencing the caries process besides the included covariates.

9.7.5 Conclusions

This section showed how the Bayesian penalized mixture CS AFT model can be used to analyze clustered doubly-interval-censored data. Owing to flexible distributional assumptions it was not here necessary to perform the classical checks for correct distributional specification. Clearly, this step cannot be avoided when using fully parametric methods. However, for censored, or let alone doubly-interval-censored data, this is far from trivial. As was illustrated in this section new important findings concerning the distribution of the event time, derived e.g. from the shape of the hazard function, can be discovered when avoiding strong parametric assumptions.

Further, we point out that the Basic Model corresponded, for comparison purposes, as closely as possible to the model used by Leroy et al. (2005). The differences were in detail outlined above. The most important one is that we used here the flexible and cluster-specific (conditional) model fitted in the Bayesian way, whereas in Leroy et al. (2005) a parametric and population-averaged (marginal) model fitted using a frequentist method. The results for the regression parameters of the caries part of the model correspond quite closely to the earlier findings of Leroy et al. (2005) where, however, no attempts were done to simplify the model. Nevertheless, our results largely confirmed their findings. Namely, they found the overall effect (on all four teeth) of all factors except gender to be significant with p-value < 0.001. For the effect of gender they observed a p-value of 0.060 compared to 0.085 found by us. Due to the fact that Leroy et al. (2005) used a parametric log-logistic AFT model, they could not reveal the second period of increased hazard found here.

Finally, we have to admit that some covariates used in our dental application should actually be treated as time-dependent. Unfortunately, with our and any other method where the distribution of the event time is specified using a density and not using an instantaneous quantity like the hazard function, inclusion of time-dependent covariates is difficult.

9.8 Example: EBCP data – multicenter study

In this section, we re-analyze the Early Breast Cancer Patients data introduced in Section 1.4 using the penalized mixture cluster-specific AFT model and compare the results to the results of the earlier analysis conducted using the classical normal mixture cluster-specific AFT model (see Section 8.9).

Except for the model for the error distribution of the AFT model, we fitted exactly the same models as in Section 8.9. Here is their brief overview. The response event time $T_{i,l}$, $i = 1, \ldots, 14$, $l = 1, \ldots, n_i$, $25 \leq n_i \leq 902$ is the progression-free survival (PFS) time of the lth patient treated by the ith center. In the CS AFT model (9.2), a bivariate random effect $\boldsymbol{b}_i = (b_{i,1}, b_{i,2})'$ with the covariate vector $\boldsymbol{z}_{i,l}^t = (1, \mathsf{trtmtGroup}_{i,l})'$ is included to allow for the baseline heterogeneity as well as the heterogeneity with respect to the treatment effect across centra. The covariate vector for the fixed effects is given by

$$\begin{aligned}\boldsymbol{x}_{i,l}^t = (&\mathsf{ageMid}_{i,l}, \mathsf{ageOld}_{i,l}, \mathsf{tySu}_{i,l}, \mathsf{tumSiz}_{i,l}, \mathsf{nodSt}_{i,l}, \mathsf{otDis}_{i,l},\\ &\mathsf{regionNL}_i, \mathsf{regionPL}_i, \mathsf{regionSE}_i, \mathsf{regionSA}_i)'.\end{aligned}$$

See Section 8.9 for explanation of the meaning of the single covariates.

Analogously to Section 8.9, besides the model **with region** described above we fitted also the model **without region** for which the covariates regionNL, regionPL, regionSE, and regionSA were omitted from the covariate vector $\boldsymbol{x}_{i,l}^t$. The motivation for this step was an attempt to see whether the the regional structure can be revealed from the estimates of the individual random effects $b_{i,1}$, $i = 1, \ldots, 14$.

For the inference we sampled two chains, each of length 125 000 with 1:5 thinning which took about 2.5 hour on a Pentium IV 2 GHz PC with 512 MB RAM. For the inference we kept the last 25 000 iterations of each chain.

9.8.1 Prior distribution

To specify the penalized mixture defining the distribution of the error terms $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ we used the grid of equidistant knots of length 31 ($K = 15$) defined on the interval $[-4.5, 4.5]$ with the basis standard deviation $\sigma = 2(\mu_j - \mu_{j-1})/3 = 0.2$. In the prior (9.7), we used the third order differences ($s = 3$). Further, the smoothing parameter λ^ε as well as the error precision parameter $(\tau^\varepsilon)^{-2}$ were given a dispersed Gamma(1, 0.005) prior. The intercept parameter α^ε as well as all fixed effect regression parameters β and the parameter $\gamma_{b,2}$ – the mean of the treatment random effects $b_{i,2}$

Table 9.2: Early breast cancer patients data. Posterior medians, 95% equal-tail credible intervals and Bayesian two-sided (simultaneous) p-values for the effect of covariates.

	Model with region		Model without region	
Parameter	Poster. median	95% CI	Poster. median	95% CI
Treatment group		$p = 0.084$		$p = 0.074$
surgery alone	-0.153	$(-0.325, 0.026)$	-0.150	$(-0.310, 0.015)$
Age		$p = 0.026$		$p = 0.014$
40–50 years	0.325	(0.059, 0.585)	0.344	(0.088, 0.619)
> 50 years	0.285	(0.041, 0.520)	0.313	(0.073, 0.565)
Type of surgery		$p = 0.008$		$p = 0.007$
breast conserving	0.229	(0.053, 0.404)	0.248	(0.078, 0.420)
Tumor size		$p < 0.001$		$p < 0.001$
≥ 2cm	-0.462	$(-0.643, -0.283)$	-0.470	$(-0.656, -0.288)$
Nodal status		$p < 0.001$		$p < 0.001$
positive	-0.599	$(-0.758, -0.442)$	-0.605	$(-0.771, -0.440)$
Other disease		$p = 0.016$		$p = 0.015$
present	-0.323	$(-0.605, -0.059)$	-0.335	$(-0.609, -0.067)$
Region		$p = 0.007$		
the Netherlands	-0.403	$(-0.737, -0.017)$		
Poland	0.349	$(-0.113, 0.802)$		
South Europe	-0.339	$(-0.729, 0.033)$		
South Africa	-0.737	$(-1.161, -0.320)$		

– were given a dispersed $\mathcal{N}(0, 100)$ prior. Finally, the covariance matrix $\mathbb{D}_b$ of the random effects got an inverse Wishart prior with $df_b = 2$ and $\mathbb{S}_b = \text{diag}(0.002)$.

9.8.2 Effect of covariates on PFS time

Table 9.2 shows the posterior summary for the effect of considered covariates in both models with included or excluded covariate region. In the model with region included, *surgery alone* decreases the time to the cancer progression by the factor of $\exp(-0.153) = 0.86$ compared to the surgery given together with the *perioperative chemotherapy*. However, as well as in the previous analysis in Section 8.9, the difference is not significant at conventional 5% level.

Table 9.3: Early breast cancer patients data. Posterior medians and 95% equal-tail credible intervals for the moments of the error distribution and variance components of the random effects.

	Model with region		Model without region	
Parameter	Poster. median	95% CI	Poster. median	95% CI
Moments of the error distribution				
$\mathrm{E}(\varepsilon)$	9.155	(8.771, 9.525)	8.967	(8.570, 9.353)
$\mathrm{sd}(\varepsilon)$	1.481	(1.356, 1.663)	1.470	(1.352, 1.639)
Variance components of the random effects				
$\mathrm{sd}(b_{i,1})$	0.111	(0.024, 0.336)	0.302	(0.157, 0.541)
$\mathrm{sd}(b_{i,2})$	0.057	(0.020, 0.217)	0.074	(0.022, 0.245)
$\mathrm{corr}(b_{i,1}, b_{i,2})$	-0.219	$(-0.987, 0.963)$	-0.675	$(-0.993, 0.980)$

Also the results for the effect of remaining covariates is very similar to the results of the earlier analysis given in Table 8.6. Firstly, again, the estimates in both models – with and without region – are almost the same. Further, according to the model with region included, in the middle age group *40 – 50 years*, the time to the progression of cancer is increased by the factor of $\exp(0.325) = 1.38$ compared to the youngest group *<40 years*. For the patients from the oldest group >50 years, this time is increased by the factor of $\exp(0.285) = 1.33$ compared to the youngest group. The variable *breast conserving surgery* increases the PFS time by the factor $\exp(0.229) = 1.26$ compared to *mastectomy*. Further, the tumor of size *≥2 cm* decreases the PFS time by the factor of $\exp(-0.462) = 0.63$ compared to the smaller tumors of size *<2 cm*. A *positive* pathological nodal status decreases the PFS time by the factor of $\exp(-0.599) = 0.55$ compared to the *negative* result. The *presence* of other related disease decreases the PFS time by the factor of $\exp(-0.323) = 0.72$. Analogously to Section 8.9, the effect of the geographical reason on the PFS time is highly significant with the same ordering of regions, namely *Poland* performs the best, followed by *France*, *South Europe*, *the Netherlands* and *South Africa*.

Finally, Figure 9.6 illustrates rather small effect of the perioperative therapy compared to surgery alone on the posterior predictive survival curves drawn for region = France and two typical combinations of covariates. More or less the same picture has been seen also in Figure 8.9 referring to the results of the earlier analysis.

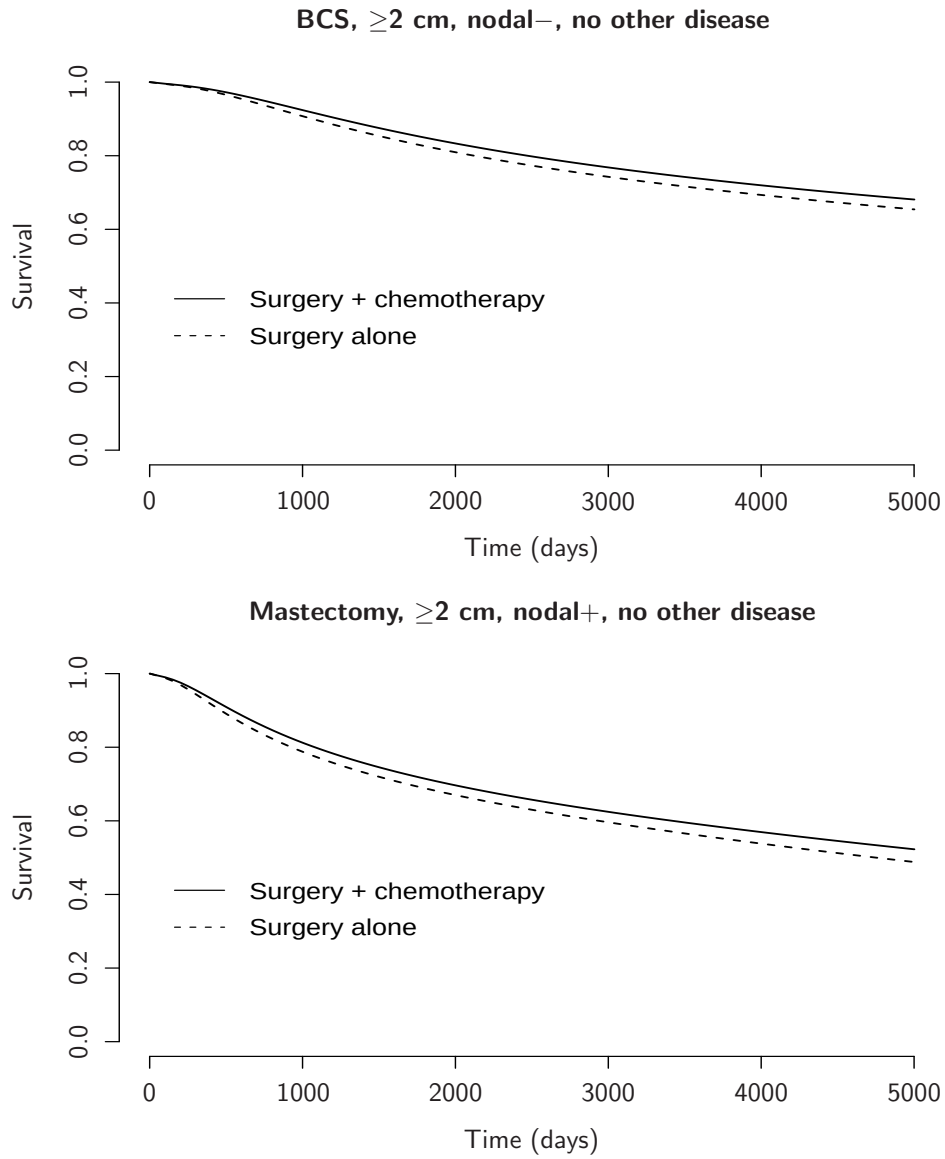


Figure 9.6: Early breast cancer patients data. Predictive survival curves based on the model with region for region = *France*, and two typical combinations of covariates: (1) *breast conserving surgery*, tumor size $\geq$*2 cm*, *negative* nodal status and *no* other associated disease (9.79% of the sample), (2) *mastectomy*, tumor size $\geq$*2 cm*, *positive* nodal status and *no* other associated disease (13.88% of the sample).

9.8.3 Predictive error density and variance components of random effects

Table 9.3 gives posterior summary statistics for the moments of the error distribution and the variance components of the random effects. Also in this case, the results are very similar to these related to the earlier analysis and shown in Table 8.7. Furthermore, the 95% equal-tail credible interval for the correlation between the overall center level and the treatment × center interaction covers again almost the whole range $(-1, 1)$ of possible value. This is also seen on the scaled histograms of sampled values of ϱ in Figure 9.7.

The estimates of the error densities in both models with and without the covariate region are shown in Figure 9.8. It is seen that exclusion of the covariate region had hardly an effect on the estimated error distribution. Indeed, since this covariate only groups different centra (clusters), its omission approved itself mainly in the variability of the random intercept $b_{i,1}$ (see Table 9.3).

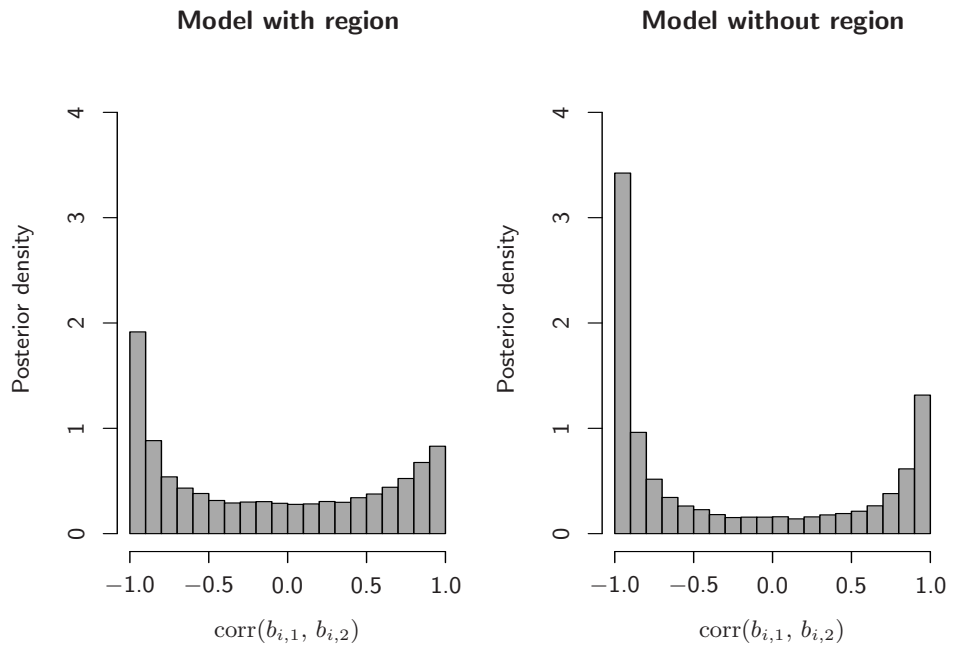


Figure 9.7: Early breast cancer patients data. Scaled histograms for sampled corr($b_{i,1}$, $b_{i,2}$).

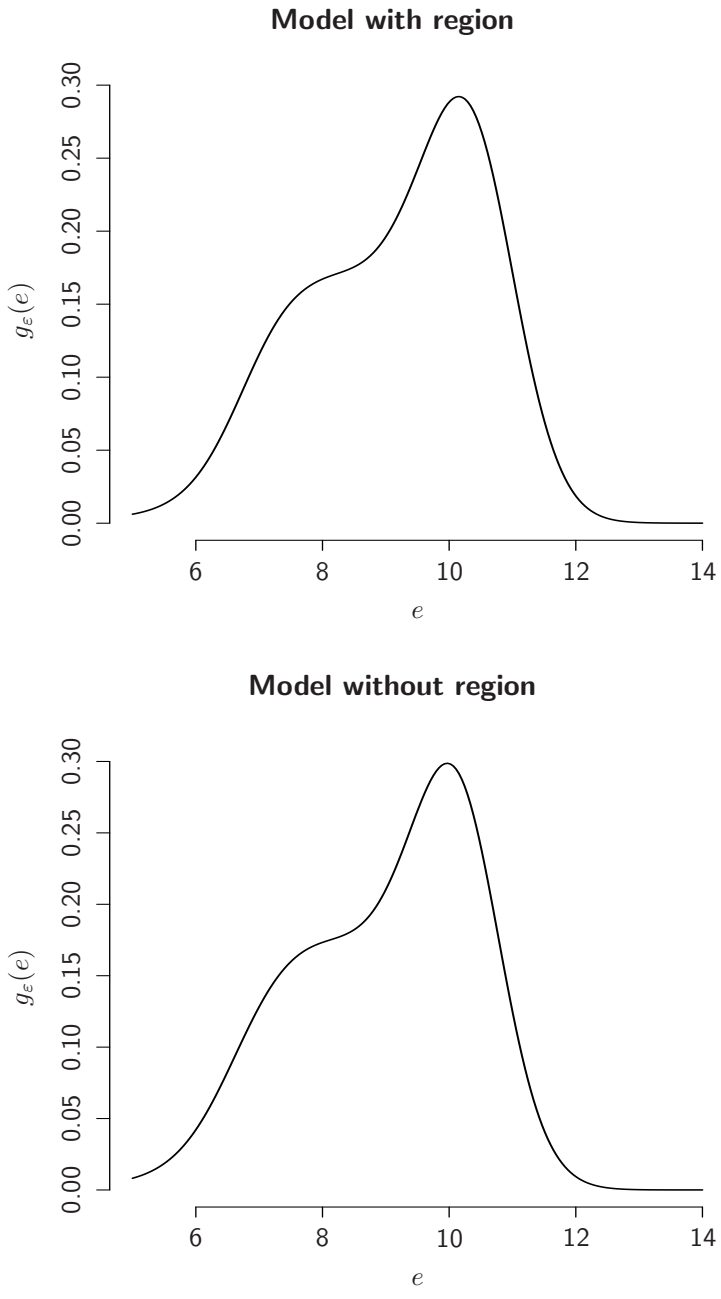


Figure 9.8: Early breast cancer patients data. Posterior predictive error densities.

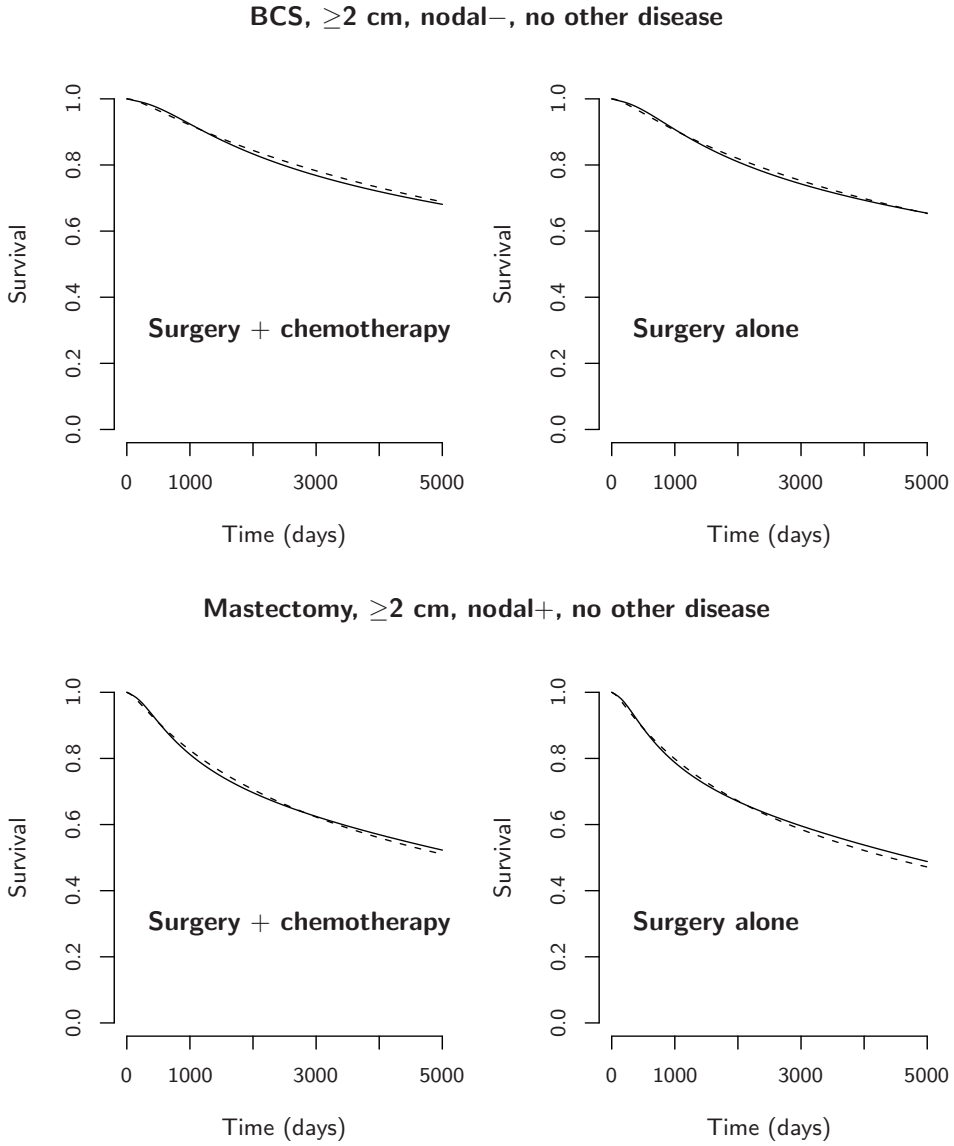


Figure 9.9: Early breast cancer patients data, comparison of the penalized mixture CS AFT model (solid lines) and the classical mixture CS AFT model (dashed lines). Predictive survival curves based on the model with region for region = *France*, and two typical combinations of covariates: (1) *breast conserving surgery*, tumor size $\geq$*2 cm*, *negative* nodal status and *no* other associated disease (9.79% of the sample), (2) *mastectomy*, tumor size $\geq$*2 cm*, *positive* nodal status and *no* other associated disease (13.88% of the sample).

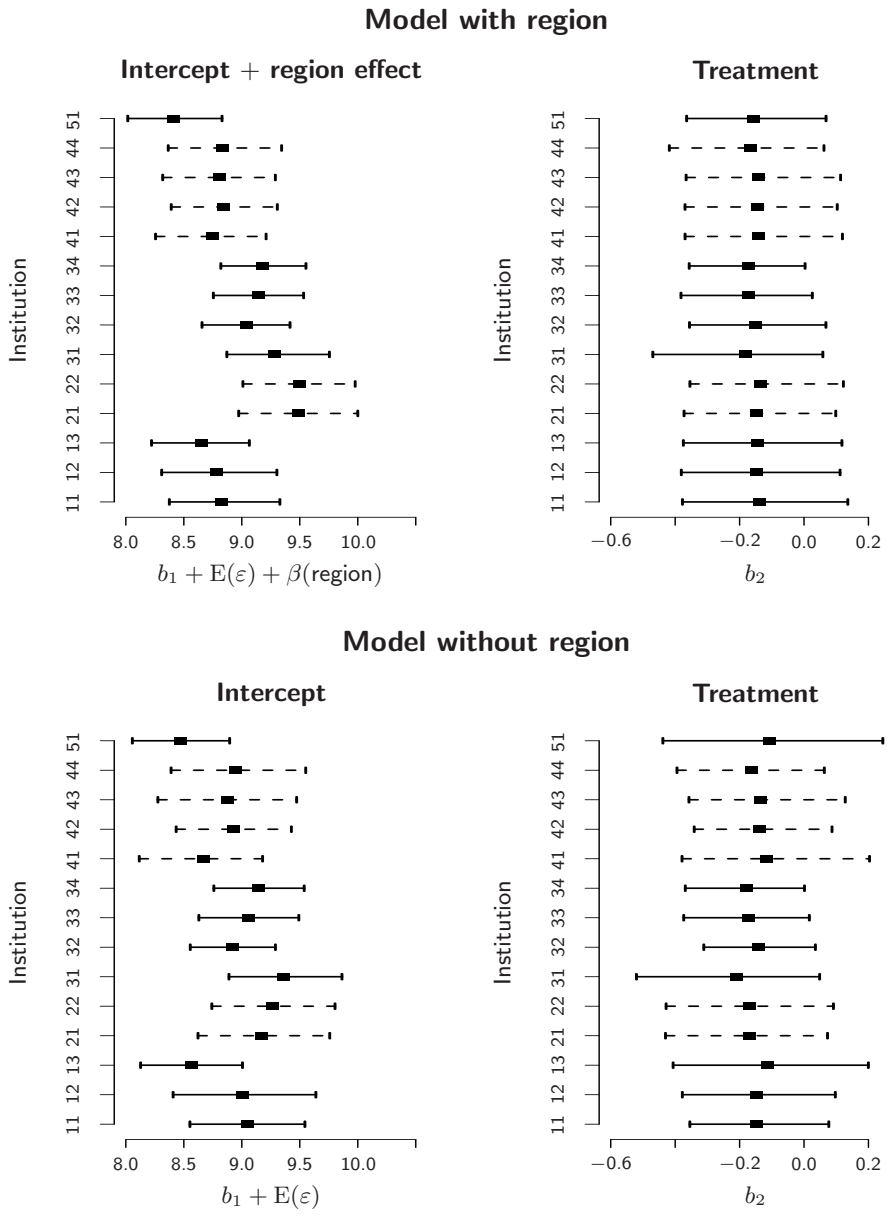


Figure 9.10: Early breast cancer patients data. Posterior means and 95% equal-tail credible intervals for individual random effects. Random intercepts are further shifted by the error mean $\mathrm{E}(\varepsilon)$ and in the model with region also by a corresponding region main effect $\beta(\mathsf{region})$.

The shape of the estimated error density seems to be somewhat different from what has been found using the classical mixture model in Section 8.9 (Figure 8.11). In Figure 8.11, the shape similar to what is seen now (Figure 9.8) can only be found when looking at conditional predictive densities, given $K > 1$. However, both estimated error distributions lead to almost the same estimates of the survival curves as it is seen in Figure 9.9.

9.8.4 Estimates of individual random effects

Finally, Figure 9.10 shows estimates of individual random effects. Analogously to Figure 8.12, the plots related to the random intercept takes into account also the mean of the error term and in the case of the model with region also the appropriate main effect of region. It can be seen that, analogous to the remaining model characteristics, Figure 9.10 resembles quite closely Figure 8.12. Among other things, also here the estimates of individual random intercepts in the model without region managed quite nicely to capture also the region effect.

9.8.5 Conclusions

The main purpose of this section was to explore how the chosen method for the estimation of the error distribution influences the results of a particular analysis. We have seen that, except for the estimate of the error distribution itself, the differences were almost negligible. Moreover, although the estimated shapes of the error distribution were somewhat different they both led to almost identical survival curves.

9.9 Discussion

A semiparametric method to perform a regression analysis with clustered doubly-interval-censored data was suggested in this chapter. We opted for a fully Bayesian approach and MCMC methodology. Note however, that similarly as in Chapter 8, the Bayesian approach is used only for technical convenience to avoid difficult optimization unavoidable with more classical maximum-likelihood based estimation. Remember that we use a penalty-like prior distribution for the transformed mixture weights $\boldsymbol{a}$ and vague priors for all remaining parameters. We did not make any attempt to use any prior information although it could have been utilized. Taking into account the above reasoning, we conclude that similar results would have been obtained if the penalized maximum-likelihood estimation had been used.

Chapter 10

Bayesian Penalized Mixture Population-Averaged AFT Model

In Section 9.7, we evaluated the impact of several covariates on the *time to caries* of the permanent first molars which are the teeth most often attacked by caries during childhood. It was also of interest to know whether the covariates have the same effect on all teeth. Hence all four teeth had to be modelled jointly. In the same section, univariate cluster-specific random effects have been included in the model expression to account for within-cluster dependencies. Given these random effects, the observations within each cluster were assumed to be independent. Distributional parts of the model were specified as penalized *univariate* normal mixtures.

However, it is also of interest to evaluate the association between the times-to-caries of the studied teeth. Nevertheless, the approach of Chapter 9 treats the within-cluster association as nuisance and, except for the estimated variance of the random effects, it does not give a direct measure of the within-cluster association. For this reason, we modify the method of Chapter 9 and assume *a multivariate* error distribution as a penalized multivariate normal mixture with a high number of components with equidistant means and constant covariance matrices.

For the explanatory and also computational reasons we describe only *a bivariate* version of the model as given by Komárek and Lesaffre (2006c) and apply it to the analysis of right permanent first molars in Section 10.6. The approach of this chapter allows to visualize the estimated bivariate error distribution and evaluate the association of paired responses.

In Section 10.1, we specify the penalized mixture population-averaged AFT

model. Further, the prior distributions are given and posterior distribution is derived in Section 10.2. Section 10.3 provides the details of the Markov chain Monte Carlo method in the context of the model of this chapter. In Section 10.4, we show how the association between the paired responses can be evaluated. Estimation of the survival distribution is discussed in Section 10.5. The analysis of the doubly-interval-censored caries times of the right permanent first molars is given in Section 10.6. Finally, we provide discussion in Section 10.7.

10.1 Model

A similar notation as in Chapter 9 will be used here. That is, let $U_{i,l}$ and $V_{i,l}$, $i = 1, \ldots, N$, $l = 1, 2$ be the *onset* time and the *failure* time, respectively for the lth unit of the ith cluster in the study. Let $T_{i,l} = V_{i,l} - U_{i,l}$ denote the corresponding *event* time The onset time $U_{i,l}$ is only observed in an interval $\lfloor u^L_{i,l},\, u^U_{i,l} \rfloor$. Similarly, we only know that the event time $V_{i,l}$ lies in an interval $\lfloor v^L_{i,l},\, v^U_{i,l} \rfloor$.

Further, let $\boldsymbol{x}^u_{i,l}$ be the vector of covariates which might have an effect on the onset time $U_{i,l}$ and $\boldsymbol{x}^t_{i,l}$ be the vector of covariates which can possibly influence the event time $T_{i,l}$. Additionally, we assume that the onset times vector $(U_{i,1},\, U_{i,2})'$ and the time-to-event vector $(T_{i,1},\, T_{i,2})'$ are, given the covariates, for each i independent (see Chapter 9 for a detailed discussion of this assumption) and that the interval censoring is independent and noninformative (e.g. pre-scheduled visits, see Section 2.4).

The distribution of $(U_{i,1},\, U_{i,2},\, T_{i,1},\, T_{i,2})'$, $i = 1, \ldots, N$, given the covariates, is given by the following accelerated failure time model:

$$\log(U_{i,l}) = \boldsymbol{\delta}' \boldsymbol{x}^u_{i,l} + \zeta_{i,l}, \tag{10.1}$$

$$\log(V_{i,l} - U_{i,l}) = \log(T_{i,l}) = \boldsymbol{\beta}' \boldsymbol{x}^t_{i,l} + \varepsilon_{i,l}, \tag{10.2}$$

$$i = 1, \ldots, N, \quad l = 1, 2,$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{m_u})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_t})'$ are unknown regression parameter vectors, $\boldsymbol{\zeta}_i = (\zeta_{i,1}, \zeta_{i,2})'$, $i = 1, \ldots, N$ are i.i.d. random vectors with a bivariate density $g_\zeta(\zeta_1, \zeta_2)$ and similarly, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \varepsilon_{i,2})'$, $i = 1, \ldots, N$ i.i.d. random vectors with a bivariate density $g_\varepsilon(\varepsilon_1, \varepsilon_2)$.

10.1.1 Distributional assumptions

Our model for the unknown bivariate densities $g_\varepsilon(\varepsilon_1,\, \varepsilon_2)$ and $g_\zeta(\zeta_1,\, \zeta_2)$ is motivated by a penalized smoothing as introduced in Section 6.3.4 and directly

generalizes the method used in Chapter 9 into higher dimensions.

Let $\boldsymbol{Y} = (Y_1, Y_2)'$ be a generic symbol for either $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)'$ or $\boldsymbol{\zeta} = (\zeta_1, \zeta_2)'$ and $g(\boldsymbol{y}) = g(y_1, y_2)$ be a generic symbol for its density. We express the unknown density $g(\boldsymbol{y})$ as a location-and-scale transformed finite mixture of bivariate normal densities with zero correlation over a *fixed fine* grid with knots $\boldsymbol{\mu}_{(j_1,j_2)} = (\mu_{1,j_1}, \mu_{2,j_2})'$, $j_1 = -K_1, \ldots, K_1$, $j_2 = -K_2, \ldots, K_2$ that are centered around zero, i.e. $\boldsymbol{\mu}_{(0,0)} = (0, 0)'$. The means of the bivariate normal components are equal to the knots and their covariance matrices are all equal but fixed to $\Sigma = \operatorname{diag}\left(\sigma_1^2, \sigma_2^2\right)$. Thus,

$$g(\boldsymbol{y}) = \tag{10.3}$$
$$(\tau_1\tau_2)^{-1} \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w_{j_1,j_2}(\mathbb{A})\, \varphi_2\Big(\frac{y_1-\alpha_1}{\tau_1}, \frac{y_2-\alpha_2}{\tau_2} \,\Big|\, \boldsymbol{\mu}_{(j_1,j_2)}, \Sigma\Big).$$

In expression (10.3), the intercept term $\boldsymbol{\alpha} = (\alpha_1, \alpha_1)'$ and the scale parameters vector $\boldsymbol{\tau} = (\tau_1, \tau_2)'$ have to be estimated as well as the matrix $\mathbb{A} = (a_{j_1,j_2})$, $j_1 = -K_1, \ldots, K_1$, $j_2 = -K_2, \ldots, K_2$ of the transformed weights. See (6.19) for the relationship between $\mathbb{A}$ and $\mathbb{W} = (w_{j_1,j_2})$, $j_1 = -K_1, \ldots, K_1$, $j_2 = -K_2, \ldots, K_2$. The density of the zero-mean, unit-variance random vector $\boldsymbol{Y}^* = \Big(\tau_1^{-1}(Y_1 - \alpha_1), \tau_2^{-1}(Y_2 - \alpha_2)\Big)'$ is a density of the bivariate normal mixture with uncorrelated components given by (6.21).

In the following, let $\mathcal{G}_\varepsilon$ refers to the set $\{\Sigma^\varepsilon, \boldsymbol{\mu}^\varepsilon, \boldsymbol{\alpha}^\varepsilon, \boldsymbol{\tau}^\varepsilon, \mathbb{W}^\varepsilon, \mathbb{A}^\varepsilon, \boldsymbol{\lambda}^\varepsilon\}$ which contains the parameters defining the distribution of $\boldsymbol{\varepsilon}$ and a smoothing parameter vector $\boldsymbol{\lambda}^\varepsilon$ which we will discuss in Section 10.2.1. Similarly, let $\mathcal{G}_\zeta$ refers to the set $\boldsymbol{\mu}^\zeta, \boldsymbol{\alpha}^\zeta, \boldsymbol{\tau}^\zeta, \mathbb{W}^\zeta, \mathbb{A}^\zeta, \boldsymbol{\lambda}^\zeta\}$ which contains the parameters defining the distribution of $\boldsymbol{\zeta}$ and a corresponding smoothing parameter vector $\boldsymbol{\lambda}^\zeta$. Finally, let $\mathcal{G}$ be a generic symbol for $\mathcal{G}_\varepsilon$ or $\mathcal{G}_\zeta$.

10.1.2 Likelihood

Let p denote a generic density. The likelihood contribution of the ith paired response equals

$$L_i = \oint_{u_{i,1}^L}^{u_{i,1}^U} \oint_{u_{i,2}^L}^{u_{i,2}^U} \oint_{v_{i,1}^L - u_{i,1}}^{v_{i,1}^U - u_{i,1}} \oint_{v_{i,2}^L - u_{i,2}}^{v_{i,2}^U - u_{i,2}} p(u_{i,1}, u_{i,2}, t_{i,1}, t_{i,2})$$
$$dt_{i,2}\, dt_{i,1}\, du_{i,2}\, du_{i,1}$$

$$= \oint_{u_{i,1}^L}^{u_{i,1}^U} \oint_{u_{i,2}^L}^{u_{i,2}^U} p(u_{i,1},\, u_{i,2}) \qquad (10.4)$$
$$\left\{ \oint_{v_{i,1}^L - u_{i,1}}^{v_{i,1}^U - u_{i,1}} \oint_{v_{i,2}^L - u_{i,2}}^{v_{i,2}^U - u_{i,2}} p(t_{i,1},\, t_{i,2})\, dt_{i,2}\, dt_{i,1} \right\} du_{i,2}\, du_{i,1},$$

where

$$p(t_{i,1},\, t_{i,2}) = (t_{i,1}\, t_{i,2})^{-1}\, g_\varepsilon\big\{\log(t_{i,1}) - \boldsymbol{\beta}'\boldsymbol{x}_{i,1},\, \log(t_{i,2}) - \boldsymbol{\beta}'\boldsymbol{x}_{i,2}\big\},$$
$$p(u_{i,1},\, u_{i,2}) = (u_{i,1}\, u_{i,2})^{-1}\, g_\zeta\big\{\log(u_{i,1}) - \boldsymbol{\delta}'\boldsymbol{z}_{i,1},\, \log(u_{i,2}) - \boldsymbol{\delta}'\boldsymbol{z}_{i,2}\big\},$$

are obtained using the expression (10.3) for g_ε and g_ζ.

In another context, Ghidey, Lesaffre, and Eilers (2004) used an expression similar to (10.3) to model a density of the random intercept and slope in the linear mixed model with uncensored data. Further, Bogaerts and Lesaffre (2006) used this approach to model a density of bivariate simply-interval-censored data without covariates. In both papers, a penalized maximum likelihood method has been used to estimate unknown parameters. In our context, however, a maximum likelihood procedure is difficult and computationally almost intractable. Like in Chapter 9 we suggest to use the Bayesian approach together with MCMC methodology.

10.2 Bayesian hierarchical model

Let $\boldsymbol{\theta}$ be a vector of all unknown parameters in our model. We assume the hierarchical structure represented by the directed acyclic graph (DAG) shown in Figure 10.1. The DAG implies the following prior distribution:

$$\begin{aligned} p(\boldsymbol{\theta}) \propto \prod_{i=1}^{N} \Big\{ & p\big(v_{i,1},\, v_{i,2} \,\big|\, u_{i,1},\, u_{i,2},\, t_{i,1},\, t_{i,2}\big) \times \\ & p\big(t_{i,1},\, t_{i,2} \,\big|\, \boldsymbol{\beta},\, \varepsilon_{i,1},\, \varepsilon_{i,2}\big) \times p\big(u_{i,1},\, u_{i,2} \,\big|\, \boldsymbol{\delta},\, \zeta_{i,1},\, \zeta_{i,2}\big) \times \\ & p\big(\varepsilon_{i,1},\, \varepsilon_{i,2} \,\big|\, \mathcal{G}_\varepsilon,\, r^\varepsilon_{i,1},\, r^\varepsilon_{i,2}\big) \times p\big(\zeta_{i,1},\, \zeta_{i,2} \,\big|\, \mathcal{G}_\zeta,\, r^\zeta_{i,1},\, r^\zeta_{i,2}\big) \times \\ & p\big(r^\varepsilon_{i,1},\, r^\varepsilon_{i,2} \,\big|\, \mathcal{G}_\varepsilon\big) \times p\big(r^\zeta_{i,1},\, r^\zeta_{i,2} \,\big|\, \mathcal{G}_\zeta\big) \Big\} \times \\ & p\big(\mathcal{G}_\varepsilon\big) \times p\big(\mathcal{G}_\zeta\big) \times p\big(\boldsymbol{\delta}\big) \times p\big(\boldsymbol{\beta}\big). \end{aligned} \qquad (10.5)$$

The DAG child-parent conditional distributions and priors for the parameters residing on the top of the hierarchy are similar to these used in Chapter 9. We give a brief overview and highlight the differences with the bivariate model considered here.

10.2.1 Prior distribution for $\mathcal{G}$

The structure of the prior distribution of the generic node $\mathcal{G}$ is the same as in Section 9.2.1, i.e.

$$p(\mathcal{G}) \propto p(\mathbb{A} \mid \boldsymbol{\lambda})\, p(\boldsymbol{\lambda})\, p(\boldsymbol{\alpha})\, p(\boldsymbol{\tau}).$$

With the bivariate setting, the number of unknown elements of the matrix $\mathbb{A}$ is naturally much higher than with the univariate setting used in Chapter 9, namely it is equal to $(2K_1+1)\times(2K_2+1)$ (e.g. equal to 961 in the analysis of

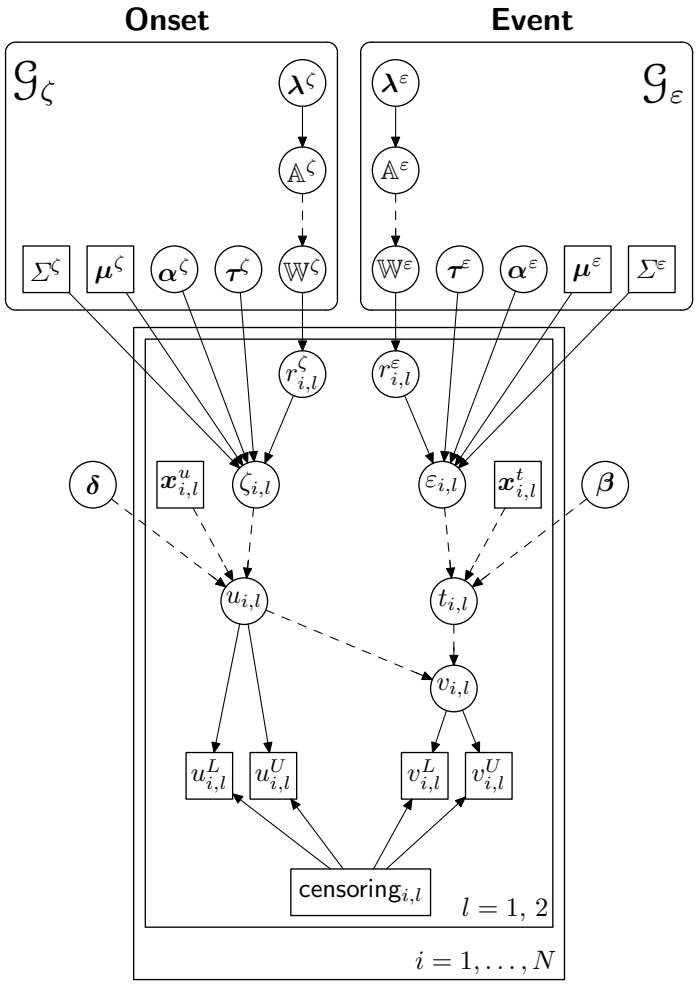


Figure 10.1: Directed acyclic graph for the Bayesian penalized mixture population-averaged AFT model.

the Signal Tandmobiel® data in Section10.6). With an uninformative prior for $\mathbb{A}$, this could cause overfitting of the data or identifiability problems.

Spatial prior for $\mathbb{A}$

Since the (transformed) mixture weights correspond to spatially located normal components, a Gaussian Markov random field (GMRF) prior (see, e.g., Besag et al., 1995, Section 3), common in spatial statistics, can be exploited here. Such a prior distribution can be defined by specifying the conditional distribution of each a_{j_1,j_2} given remaining a_{k_1,k_2}, $(k_1,k_2) \neq (j_1,j_2)$, here denoted as $\mathbb{A}_{-(j_1,j_2)}$, and the hyperparameter λ that controls the smoothness. Usually, only a few neighboring coefficients are effectively used in the specification of $p(a_{j_1,j_2} \,|\, \mathbb{A}_{-(j_1,j_2)}, \lambda)$. A commonly used conditional distribution is a normal distribution with expectation and variance equal to

$$\begin{aligned} \mathrm{E}\big(a_{j_1,j_2} \,|\, \mathbb{A}_{-(j_1,j_2)}, \lambda\big) &= \frac{a_{j_1-1,j_2} + a_{j_1+1,j_2} + a_{j_1,j_2-1} + a_{j_1,j_2+1}}{2} - \\ &\qquad \frac{a_{j_1-1,j_2-1} + a_{j_1-1,j_2+1} + a_{j_1+1,j_2-1} + a_{j_1+1,j_2+1}}{4}, \\ \mathrm{var}\big(a_{j_1,j_2} \,|\, \mathbb{A}_{-(j_1,j_2)}, \lambda\big) &= (4\lambda)^{-1}, \end{aligned} \tag{10.6}$$

respectively, based on the eight nearest neighbors and local quadratic smoothing. Note that the expectation and variance formulas have to be changed appropriately on edges where only five neighbors are available and in corners where we have only three neighbors out of the original eight. Namely, for the edge given by $j_1 = K_1$:

$$\begin{aligned} \mathrm{E}\big(a_{K_1,j_2} \,|\, \mathbb{A}_{-(K_1,j_2)}, \lambda\big) &= a_{K_1-1,j_2} + \frac{a_{K_1,j_2-1} + a_{K_1,j_2+1}}{2} - \\ &\qquad \frac{a_{K_1-1,j_2-1} + a_{K_1-1,j_2+1}}{2} \\ \mathrm{var}\big(a_{K_1,j_2} \,|\, \mathbb{A}_{-(K_1,j_2)}, \lambda\big) &= (2\lambda)^{-1}, \qquad j_2 = -K_2+1, \ldots, K_2-1, \end{aligned}$$

and similarly for the remaining edges. In the corner given by $(j_1,j_2) = (K_1,K_2)$:

$$\begin{aligned} \mathrm{E}\big(a_{K_1,K_2} \,|\, \mathbb{A}_{-(K_1,K_2)}, \lambda\big) &= a_{K_1-1,K_2} + a_{K_1,K_2-1} - a_{K_1-1,K_2-1}, \\ \mathrm{var}\big(a_{K_1,K_2} \,|\, \mathbb{A}_{-(K_1,K_2)}, \lambda\big) &= \lambda^{-1}, \end{aligned}$$

and similarly for the remaining corners.

Let $\boldsymbol{a}$ denote the matrix $\mathbb{A}$ stacked into a column vector. Using a bivariate difference operator

$$\Delta\, a_{j_1,j_2} = a_{j_1,j_2} - a_{j_1+1,j_2} - a_{j_1,j_2+1} + a_{j_1+1,j_2+1},$$

and denoting $\mathbb{D}$ the associated difference operator matrix, the joint prior of all transformed weights $\mathbb{A}$ given the smoothing hyperparameter λ can be written as

$$p(\mathbb{A} \mid \lambda) \propto \exp\Big\{-\frac{\lambda}{2} \sum_{j_1=-K_1}^{K_1-1} \sum_{j_2=-K_2}^{K_2-1} \big(\Delta\, a_{j_1,j_2}\big)^2\Big\} = \exp\Big(-\frac{\lambda}{2}\, \boldsymbol{a}'\mathbb{D}'\mathbb{D}\boldsymbol{a}\Big) \quad (10.7)$$

which shows that the DAG conditional distribution $p(\mathbb{A} \mid \lambda)$ specified as a GMRF is multivariate normal with covariance matrix $\lambda^{-1}\big(\mathbb{D}'\mathbb{D}\big)^-$, where $\big(\mathbb{D}'\mathbb{D}\big)^-$ denotes a generalized inverse of $\mathbb{D}'\mathbb{D}$. Although this distribution is improper (the matrix $\mathbb{D}'\mathbb{D}$ has a deficiency of $2(K_1 + K_2) + 1$ in its rank) the resulting posterior distribution is proper as soon as there is some informative data available, see Besag et al. (1995).

Conditionally univariate difference prior

An alternative prior, still belonging to the class of GMRF, corresponding closely to the prior for $\mathbb{A}$ used in Chapter 9 is obtained by considering a univariate difference operator for each row and each column of the matrix $\mathbb{A}$ with possibly two different smoothing hyperparameters stacked in a vector $\boldsymbol{\lambda} = (\lambda_1,\, \lambda_2)'$ acting on rows and columns separately. Then

$$\begin{aligned} p(\mathbb{A} \mid \boldsymbol{\lambda}) \propto \exp\Big\{ &-\frac{\lambda_1}{2} \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2+s}^{K_2} \big(\Delta_1^s\, a_{j_1,j_2}\big)^2 \\ &-\frac{\lambda_2}{2} \sum_{j_2=-K_2}^{K_2} \sum_{j_1=-K_1+s}^{K_1} \big(\Delta_2^s\, a_{j_1,j_2}\big)^2 \Big\} \\ = \exp\Big\{ &-\frac{1}{2}\, \boldsymbol{a}' \left(\lambda_1\, \mathbb{D}_1'\mathbb{D}_1 \;+\; \lambda_2\, \mathbb{D}_2'\mathbb{D}_2\right) \boldsymbol{a}\Big\} \end{aligned} \quad (10.8)$$

where Δ_l^s, $l = 1, 2$ denotes a difference operator of order s for the lth dimension, e.g. $\Delta_1^3\, a_{j_1,j_2} = a_{j_1,j_2} - 3\, a_{j_1,j_2-1} + 3\, a_{j_1,j_2-2} - a_{j_1,j_2-3}$ and $\mathbb{D}_1$ and $\mathbb{D}_2$ are the corresponding difference operator matrices for each dimension. This prior distribution corresponds to a local polynomial smoothing of degree $s-1$ in each row and each column of the matrix $\mathbb{A}$. For example, the conditional mean and variance are given (for $s = 3$ and except on the corners and on edges) by

$$\begin{aligned} \mathrm{E}\big(a_{j_1,j_2} \,|\, \mathbb{A}_{-(j_1,j_2)},\, \boldsymbol{\lambda}\big) &= \frac{\lambda_1\, A_{j_2\,|\,j_1} \;+\; \lambda_2\, A_{j_1\,|\,j_2}}{\lambda_1 \;+\; \lambda_2} \\ \mathrm{var}\big(a_{j_1,j_2} \,|\, \mathbb{A}_{-(j_1,j_2)},\, \boldsymbol{\lambda}\big) &= \frac{1}{20(\lambda_1 + \lambda_2)}, \end{aligned} \quad (10.9)$$

where

$$A_{k\,|\,j} = \frac{a_{j,k-3} - 6\,a_{j,k-2} + 15\,a_{j,k-1} + 15\,a_{j,k+1} - 6\,a_{j,k+2} + a_{j,k+3}}{20}.$$

Both the spatial prior for $\mathbb{A}$ and the conditionally univariate difference prior for $\mathbb{A}$ put higher probability mass in areas where spatially close coefficients of the matrix $\mathbb{A}$ do not substantially differ. In other words, a priori, we believe that the estimated densities $g_\zeta(\zeta_1,\, \zeta_2)$ and $g_\varepsilon(\varepsilon_1,\, \varepsilon_2)$ are smooth. In general, prior (10.8) leads to better a fit in our context and hence is preferred.

Prior for the smoothing parameter

The λ parameter in the prior (10.7) or the components λ_1, λ_2 of the $\boldsymbol{\lambda}$ parameter in the prior (10.8) determine, together with the fixed difference operator matrix $\mathbb{D}$, the precision of the transformed weights $\mathbb{A}$. We assign these parameters standardly used highly dispersed (but proper) Gamma priors.

Prior for the mixture intercepts and scales

The intercept parameters α_1^ε, α_2^ε, α_1^ζ, α_2^ζ can obtain a vague normal prior unless there is some external information available. For the scale parameters τ_1^ε, τ_2^ε, τ_1^ζ, τ_2^ζ we suggest to use either the uniform prior or a highly dispersed inverse-Gamma prior for the squared scale parameters.

10.2.2 Prior distribution for the generic node $\boldsymbol{Y}$

To specify the prior distribution of the generic node $\boldsymbol{Y}$, i.e. of the nodes $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\zeta}_i$, $i = 1, \ldots, N$, we introduce, analogously to Chapter 9 and using the idea of Bayesian data augmentation (see Section 4.3), latent *allocation vector* $\boldsymbol{r} = (r_1,\, r_2)'$ that can take discrete values from $\{-K_1, \ldots, K_1\} \times \{-K_2, \ldots, K_2\}$. Its DAG conditional distribution is given by

$$\begin{aligned}\Pr(\boldsymbol{r} = (j_1,\, j_2)' \,|\, \mathcal{G}) = \Pr(\boldsymbol{r} = (j_1,\, j_2)' \,|\, \mathbb{W}) = w_{j_1,\, j_2},\\ j_1 \in \{-K_1, \ldots, K_1\},\ j_2 \in \{-K_2, \ldots, K_2\}.\end{aligned}$$

The DAG conditional distribution of the generic node $\boldsymbol{Y}$ is then simply bivariate normal with independent margins:

$$p(y_1,\, y_2 \,|\, \mathcal{G},\, r_1,\, r_2) = \varphi_2\big(\boldsymbol{y} \,\big|\, \boldsymbol{\alpha} + \mathrm{diag}(\boldsymbol{\tau})\boldsymbol{\mu}_{(r_1,\, r_2)},\, \mathrm{diag}(\boldsymbol{\tau})\,\Sigma\,\mathrm{diag}(\boldsymbol{\tau})\big).$$

Without introducing the latent allocation vectors we would have to work with $p(\boldsymbol{y} \,|\, \mathcal{G}) = p(\boldsymbol{y} \,|\, \boldsymbol{\mu},\, \Sigma,\, \boldsymbol{\alpha},\, \boldsymbol{\tau}, \mathbb{W})$ which is a bivariate normal *mixture* given by (10.3).

10.2.3 Prior distribution for the regression parameters and time variables

The prior distribution of the regression parameters and the time variables is exactly the same as in Chapter 9. That is, the regression parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are given a vague normal prior unless there is some external information available. Finally, the nodes $u^L_{i,l}$, $u^U_{i,l}$, $v^L_{i,l}$, $v^U_{i,l}$, $t^L_{i,l}$ and $t^U_{i,l}$ have, conditionally on their parents, the Dirac distribution driven by the censoring mechanism and the true onset, failure or event time, respectively. See Section 9.2.5 with an obvious change in notation. Finally, remember that we do not have to specify an exact form of the censoring mechanism as soon as it is noninformative and independent.

10.2.4 Posterior distribution

The posterior distribution is given as a product of all DAG conditional distributions. See Section 9.2.6 for details.

10.3 Markov chain Monte Carlo

In practice we obtain a sample from the posterior distribution using the Markov chain Monte Carlo method and base our inference on this sample. Analogously to Chapter 9, the basis for the MCMC algorithm is Gibbs sampling (Geman and Geman, 1984) using the full conditional distributions. In the situations when the full conditional distribution was not of standard form we used either slice sampling (Neal, 2003) or adaptive rejection sampling (Gilks and Wild, 1992). For most parameters the full conditionals are identical (with only a slight change in notation) to those given in Section 9.3 and we refer the reader thereinto.

Here we mention only the full conditional distribution for the transformed mixture weights which, due to the bivariate nature considered here, differs

from that in Chapter 9 and is equal to

$$p(a_{j_1,j_2} \mid \cdots) \propto \frac{\exp(N_{j_1,j_2} a_{j_1,j_2})}{\Big\{ \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} \exp(a_{k_1,k_2}) \Big\}^N} \times \exp\left[- \frac{\Big\{ a_{j_1,j_2} - \mathrm{E}\big(a_{j_1,j_2} \mid \mathbb{A}_{-(j_1,j_2)},\, \boldsymbol{\lambda}\big) \Big\}^2}{2\ \mathrm{var}\big(a_{j_1,j_2} \mid \mathbb{A}_{-(j_1,j_2)},\, \boldsymbol{\lambda}\big)} \right],$$
$$j_1 = -K_1, \ldots, K_1,\ j_2 = -K_2, \ldots, K_2,$$

where N_{j_1,j_2} denotes the number of latent allocation vectors $\boldsymbol{r}_i$ that are equal to $(j_1,\, j_2)'$ and $\mathrm{E}\big(a_{j_1,j_2} \mid \mathbb{A}_{-(j_1,j_2)},\, \boldsymbol{\lambda}\big)$ and $\mathrm{var}\big(a_{j_1,j_2} \mid \mathbb{A}_{-(j_1,j_2)},\, \boldsymbol{\lambda}\big)$ follow from (10.6) or (10.9).

10.4 Evaluation of association

The association between the paired responses, after adjustment for the effect of covariates, can be evaluated for example using the Pearson correlation coefficient of the error terms $\zeta_{i,1}$ and $\zeta_{i,2}$, or $\varepsilon_{i,1}$ and $\varepsilon_{i,2}$, respectively. For example, the Pearson correlation coefficient of the error terms $\varepsilon_{i,1}$ and $\varepsilon_{i,2}$ equals

$$\varrho^{\varepsilon} = \frac{\sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} w^{\varepsilon}_{j_1,j_2} \big(\mu^{\varepsilon}_{1,j_1} - M^{\varepsilon}_1\big)\big(\mu^{\varepsilon}_{2,j_2} - M^{\varepsilon}_2\big)}{\Big\{ (\sigma^{\varepsilon}_1)^2 + \sum_{j_1=-K_1}^{K_1} w^{\varepsilon}_{j_1+} \big(\mu^{\varepsilon}_{1,j_1} - M^{\varepsilon}_1\big) \Big\}^{\frac{1}{2}} \Big\{ (\sigma^{\varepsilon}_2)^2 + \sum_{j_2=-K_2}^{K_2} w^{\varepsilon}_{+j_2} \big(\mu^{\varepsilon}_{2,j_2} - M^{\varepsilon}_2\big) \Big\}^{\frac{1}{2}}},$$

where

$$w^{\varepsilon}_{j_1+} = \sum_{j_2=-K_2}^{K_2} w^{\varepsilon}_{j_1,j_2}, \quad j_1 = -K_1, \ldots, K_1, \qquad M^{\varepsilon}_1 = \sum_{j_1=-K_1}^{K_1} w^{\varepsilon}_{j_1+} \mu^{\varepsilon}_{1,j_1},$$
$$w^{\varepsilon}_{+j_2} = \sum_{j_1=-K_1}^{K_1} w^{\varepsilon}_{j_1,j_2}, \quad j_2 = -K_2, \ldots, K_2, \qquad M^{\varepsilon}_2 = \sum_{j_2=-K_2}^{K_2} w^{\varepsilon}_{+j_2} \mu^{\varepsilon}_{2,j_2}.$$

Another popular measure of association for censored data is the Kendall's tau, denoted by τ_{Kend}, of which one advantage is that it is invariant towards monotone transformations. This implies in our context that after adjustment for the effect of covariates, the same value of the Kendall's tau is obtained for both the original event times and for their logarithmic transformation

represented by the error terms. For example for the time-to-event part of the model, given the model parameters, the Kendall's tau $\tau^{\varepsilon}_{Kend}$ is equal to

$$\tau^{\varepsilon}_{Kend} = 4 \cdot \sum_{j_1=-K_1}^{K_1} \sum_{j_2=-K_2}^{K_2} \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} w^{\varepsilon}_{j_1,j_2}\, w^{\varepsilon}_{k_1,k_2} \\ \Phi\Big(\frac{\mu^{\varepsilon}_{1,j_1} - \mu^{\varepsilon}_{1,k_1}}{\sqrt{2}\sigma^{\varepsilon}_1}\Big)\, \Phi\Big(\frac{\mu^{\varepsilon}_{2,j_2} - \mu^{\varepsilon}_{2,k_2}}{\sqrt{2}\sigma^{\varepsilon}_2}\Big) \;-\; 1,$$

see Bogaerts and Lesaffre (2006) for details.

10.5 Bayesian estimates of the survival distribution

10.5.1 Predictive survival nad hazard curves and predictive survival densities

The estimates of the survival and hazard functions or the survival densites for a specific combination of covariates are estimated by the mean of (posterior) predictive quantities. In practice, this is done analogously to Sections 8.4 and 9.4. However, due to the bivariate approach in this chapter, we have to distinguish between the quantities for the first margin: the onset time U_1 and the event time T_1 and for the second margin: the onset time U_2 and the event time T_2.

For example, to get the Bayesian estimate of the predictive survival function of the event time T_1, given the covariates $\boldsymbol{x}^t_{new}$ and $\boldsymbol{z}^t_{new}$, we can use the relationship (8.17) while replacing the expression (8.16) by

$$S_1(t_1 \mid \boldsymbol{\theta}, \boldsymbol{x}^t_{new}, \boldsymbol{z}^t_{new}) = \tag{10.10}$$
$$1 - \sum_{j_1=-K_1}^{K_1} w^{\varepsilon}_{j_1,+}\, \Phi\big\{\log(t_1) - \boldsymbol{\beta}'\boldsymbol{x}^t_{new} - \boldsymbol{b}'\boldsymbol{z}^t_{new} \,\big|\, \alpha^{\varepsilon}_1 + \tau^{\varepsilon}_1\mu^{\varepsilon}_{1,j_1},\, (\sigma^{\varepsilon}_1\tau^{\varepsilon}_1)^2\big\}.$$

To get the Bayesian estimate of the predictive survival density of the event time T_1, we replace the expression (8.18) by

$$p_1(t_1 \mid \boldsymbol{\theta}, \boldsymbol{x}^t_{new}, \boldsymbol{z}^t_{new}) = \tag{10.11}$$
$$t_1^{-1} \sum_{j_1=-K_1}^{K_1} w^{\varepsilon}_{j_1,+}\, \varphi\big\{\log(t_1) - \boldsymbol{\beta}'\boldsymbol{x}^t_{new} - \boldsymbol{b}'\boldsymbol{z}^t_{new} \,\big|\, \alpha^{\varepsilon}_1 + \tau^{\varepsilon}_1\mu^{\varepsilon}_{1,j_1},\, (\sigma^{\varepsilon}_1\tau^{\varepsilon}_1)^2\big\}.$$

Analogously, the quantities for the event time T_2 in the second margin and for the onset times U_1 and U_2 are obtained.

10.5.2 Predictive error densities

The MCMC estimates of the predictive error densities are obtained in the same way as explained in Section 9.4.2. We only have to use a bivariate counterpart of the expression (9.13), i.e. for the event error density we use

$$\hat{g}_{\varepsilon}(e_1,\, e_2) = \frac{1}{M}\sum_{m=1}^{M}\Bigg\{ \big(\tau_1^{\varepsilon,(m)}\tau_2^{\varepsilon,(m)}\big)^{-1} \sum_{j_1=-K_1}^{K_1}\sum_{j_2=-K_2}^{K_2} w_{j_1,j_2}^{\varepsilon,(m)}\, \varphi_2\Big(\frac{e_1-\alpha_1^{\varepsilon,(m)}}{\tau_1^{\varepsilon,(m)}},\, \frac{e_2-\alpha_2^{\varepsilon,(m)}}{\tau_2^{\varepsilon,(m)}} \,\Big|\, \boldsymbol{\mu}_{(j_1,j_2)}^{\varepsilon},\, \Sigma^{\varepsilon}\Big)\Bigg\}. \tag{10.12}$$

10.6 Example: Signal Tandmobiel® study – paired doubly-interval-censored data

In Section 9.7, we have analyzed the time to caries of the permanent first molars based on the data from the Signal Tandmobiel® study using the cluster-specific AFT model. The results were compared to the earlier analysis of Leroy et al. (2005). In this section, we perform a similar analysis. However, for practical reasons (see the introduction to this chapter) it is only possible to analyze a pair of teeth. In our analysis, we concentrated on differences between the maxillary (upper) and mandibular (lower) teeth and analyzed separately the pair of right teeth (teeth 16 and 46) and the pair of left teeth (teeth 26 and 36). The results for both pairs were very similar so we report only the results for the right teeth in this thesis. Due to the fact that a (parametric) population-averaged AFT model was used by Leroy et al. (2005), the results presented in this section can even more closely be compared to their findings.

The analysis proceeded in a similar way as in Section 9.7 with only changes related to the fact we analyze only two teeth now. Specifically, the onset time $U_{i,l}$, $i = 1,\ldots,N$, $l = 1,2$ refers to the age (in years) of the ith child at which the lth tooth ($l = 1 \equiv$ tooth 16, $l = 2 \equiv$ tooth 46) emerged. The failure time $V_{i,l}$, $i = 1,\ldots,N$, $l = 1,\ldots,2$ refers to the onset of caries and the event time $T_{i,l}$ to the time between the emergence and the onset of caries. As explained in Section 9.7, left-censored emergence times were transformed into interval-censored ones based on the clinical eruption stage. Finally, as in Section 9.7, we subtracted 5 years from all observed times, i.e. $\log(U_{i,l}-5)$ was used in the left-hand side of the model formula (10.1). Analogously to Section 9.7, we started the analysis with the **Basic Model** and based on

the results for the Basic Model we subsequently fitted its simplified version, referred as the **Final Model**.

10.6.1 Basic Model

In the **Basic Model** we allowed for a different effect of the covariates on both emergence and caries experience for the maxillary and mandibular tooth, respectively. That is, in the AFT models (10.1) and (10.2) we used the following covariate vectors $\boldsymbol{x}^u_{i,l}$ and $\boldsymbol{x}^t_{i,l}$ for the emergence and caries parts of the model, respectively.

$$\begin{aligned}\boldsymbol{x}^u_{i,l} &= (\mathsf{gender}_i,\ \mathsf{jaw}_{i,l} * \mathsf{gender}_i)',\\ \boldsymbol{x}^t_{i,l} &= (\tilde{\boldsymbol{x}}^t_{i,l},\ \mathsf{jaw}_{i,l} * \tilde{\boldsymbol{x}}^t_{i,l}),\end{aligned}$$

where

$$\begin{aligned}\tilde{\boldsymbol{x}}^t_{i,l} = (&\mathsf{gender}_i,\ \mathsf{statusD}_{i,l},\ \mathsf{statusF}_{i,l},\ \mathsf{statusM}_{i,l},\\ &\mathsf{brushing}_i,\ \mathsf{sealants}_{i,l},\ \mathsf{plaquePF}_{i,l},\ \mathsf{plaqueT}_{i,l}).\end{aligned}$$

The covariate jaw is dichotomous ($1 =$ *maxilla*, $0 =$ *mandible*) and distinguishes between the maxillary and mandibular tooth. It replaces the covariates tooth26, tooth36, tooth46 used in Section 9.7.1. Note that as well in the caries part as in the emergence part of the model the main effect of jaw is expressed by the intercept terms $\boldsymbol{\alpha}^{\varepsilon}$ and $\boldsymbol{\alpha}^{\zeta}$, respectively. See Section 9.7.1 for the explanation of the remaining covariates.

10.6.2 Final Model

In the **Final Model**, we excluded all interaction terms with the covariate jaw, i.e. we assumed that the studied factors have the same effect on the emergence and caries for both the maxillary and mandibular tooth. Additionally, as in Section 9.7, we binarized the covariates plaquePF, plaqueT and statusD, statusF, statusM into new covariates plaque and status, respectively. Bayesian two-sided p-values and for factors with more than two levels simultaneous two-sided Bayesian p-values (see Section 4.6.2) were used to arrive at the Final Model.

10.6.3 Prior distribution

To model the bivariate densities g_{ζ} and g_{ε} we used in both cases a grid of 31×31 ($K_1 = K_2 = 15$) knots with the distance d between the two knots in each

margin equal to 0.3 and the basis standard deviations $\sigma_1^\varepsilon = \sigma_2^\varepsilon = \sigma_1^\zeta = \sigma_2^\zeta = 0.2$. The grid of knots is defined on a square $[-4.5, 4.5] \times [-4.5, 4.5]$ which covers the support of most standardized unimodal distributions (unimodality was checked after the analysis).

For the transformed mixture weights $\mathbb{A}^\varepsilon$ and $\mathbb{A}^\zeta$ we used the prior (10.8) with the differences of the third order ($s = 3$). The smoothing parameters λ_1^ε, λ_2^ε, λ_1^ζ, λ_2^ζ were all assigned dispersed Gamma(1, 0.005) priors. The same priors were used also for the scale parameters τ_1^ε, τ_2^ε, τ_1^ζ, τ_2^ζ. The intercept terms α_1^ε, α_2^ε, α_1^ζ, α_2^ζ as well as regression parameters contained in vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ were all assigned dispersed $\mathcal{N}(0, 100)$ priors.

10.6.4 Results

For each model we ran 250 000 iterations with 1:3 thinning and kept last 25 000 iterations for the inference. Sampling for each model took about 68 hours on a 3 GHz Pentium IV PC with 1 GB RAM.

Results for the Basic Model

Table 10.1 shows the posterior medians, (simultaneous) 95% equal-tail credible intervals and (simultaneous) Bayesian two-sided p-values for the effect of each considered factor on emergence and caries experience, separately for the maxillary and the mandibular tooth.

It is seen that the results for the mandibular and the maxillary tooth are very similar. Indeed, the interaction terms between jaw and the remaining factor variables were all non-significant at 5%, namely, the p-values were > 0.5, > 0.5, > 0.5, 0.262, > 0.5, 0.145, respectively for the interaction with gender in the emergence and the caries part of the model, and for the interaction with brushing, sealants, plaque, and status, respectively.

Additionally, we computed the (simultaneous) Bayesian two-sided p-values for the two contrasts justifying the simplification of the covariates plaque and status for the Final Model, again separately for the mandibular and the maxillary tooth. For the variable status contrast *decayed* vs. *filled* vs. *missing due to caries*, the p-values were equal to 0.342 and 0.308, respectively for the maxillary and the mandibular tooth, respectively. For the variable plaque contrast *in pits and fissures* vs. *on total surface*, the p-values were equal to 0.262 and 0.301, respectively for the maxillary and the mandibular tooth, respectively.

Table 10.1: Signal Tandmobiel® study, Basic Model. Posterior medians, 95% equal-tail credible regions (CR) and Bayesian two-sided p-values for each factor variable, separately for the maxillary tooth 16 and the mandibular tooth 46.

	Maxillary tooth 16		Mandibular tooth 46	
Effect	Posterior median	95% CR	Posterior median	95% CR
		Emergence		
Gender	$p = 0.094$		$p = 0.142$	
girl	-0.018	$(-0.039, 0.003)$	-0.016	$(-0.036, 0.005)$
		Caries		
Gender	$p = 0.534$		$p = 0.403$	
girl	-0.035	$(-0.139, 0.073)$	-0.049	$(-0.162, 0.063)$
Status	$p < 0.001$		$p < 0.001$	
decayed	-0.449	$(-0.704, -0.224)$	-0.379	$(-0.641, -0.151)$
filled	-0.627	$(-0.844, -0.414)$	-0.375	$(-0.588, -0.175)$
missing	-0.470	$(-1.377, 0.138)$	-0.726	$(-1.398, -0.208)$
Brushing	$p = 0.003$		$p < 0.001$	
daily	0.226	(0.086, 0.386)	0.265	(0.097, 0.426)
Sealants	$p = 0.019$		$p = 0.401$	
present	0.158	(0.028, 0.283)	0.055	$(-0.077, 0.180)$
Plaque	$p = 0.014$		$p = 0.002$	
in pits and fissures	-0.183	$(-0.333, -0.031)$	-0.252	$(-0.404, -0.107)$
on total surface	-0.389	$(-0.819, -0.015)$	-0.468	$(-0.997, -0.038)$

Results for the Final Model

Results for the Final Model are given in Table 10.2. This table contains also the main effect of jaw which is given by $\mathrm{E}(\zeta_2) - \mathrm{E}(\zeta_1)$ and by $\mathrm{E}(\varepsilon_2) - \mathrm{E}(\varepsilon_1)$ in the case of emergence and caries, respectively.

It is seen that the lower tooth 46 emerges slightly later than the upper tooth 16. On the other hand, emergence occurs slightly earlier for girls than for boys. However, neither the position of the tooth nor gender have a significant

Table 10.2: Signal Tandmobiel® study, Final Model. Posterior medians, 95% equal-tail credible regions (CR) and Bayesian two-sided p-values for each factor variable.

	Emergence		Caries	
Effect	Posterior median	95% CR	Posterior median	95% CR
Jaw	$p = 0.021$		$p = 0.816$	
lower	0.017	(0.003, 0.032)	0.024	(−0.158, 0.218)
Gender	$p = 0.018$		$p = 0.267$	
girl	−0.017	(−0.033, −0.003)	−0.044	(−0.120, 0.033)
Status			$p < 0.001$	
dmf			−0.482	(−0.576, −0.388)
Brushing			$p < 0.001$	
daily			0.249	(0.139, 0.369)
Sealants			$p = 0.022$	
present			0.110	(0.019, 0.195)
Plaque			$p < 0.001$	
present			−0.228	(−0.313, −0.141)

Table 10.3: Signal Tandmobiel® study, Final Model. Posterior medians and 95% equal-tail credible regions (CR) for the mean, standard deviation, Pearson correlation and Kendall's tau of the error terms.

	Posterior			Posterior	
Param.	median	95% CR	Param.	median	95% CR
	Emergence			Caries	
$\mathrm{E}(\zeta_1)$	0.392	(0.379, 0.404)	$\mathrm{E}(\varepsilon_1)$	2.846	(2.645, 3.043)
$\mathrm{E}(\zeta_2)$	0.409	(0.397, 0.421)	$\mathrm{E}(\varepsilon_2)$	2.870	(2.706, 3.040)
$\mathrm{sd}(\zeta_1)$	0.170	(0.163, 0.178)	$\mathrm{sd}(\varepsilon_1)$	1.737	(1.631, 1.855)
$\mathrm{sd}(\zeta_2)$	0.170	(0.164, 0.177)	$\mathrm{sd}(\varepsilon_2)$	1.812	(1.722, 1.918)
ϱ^{ζ}	0.037	(0.030, 0.050)	ϱ^{ε}	0.023	(0.018, 0.028)
τ^{ζ}_{Kend}	0.022	(0.016, 0.030)	$\tau^{\varepsilon}_{Kend}$	0.011	(0.008, 0.013)

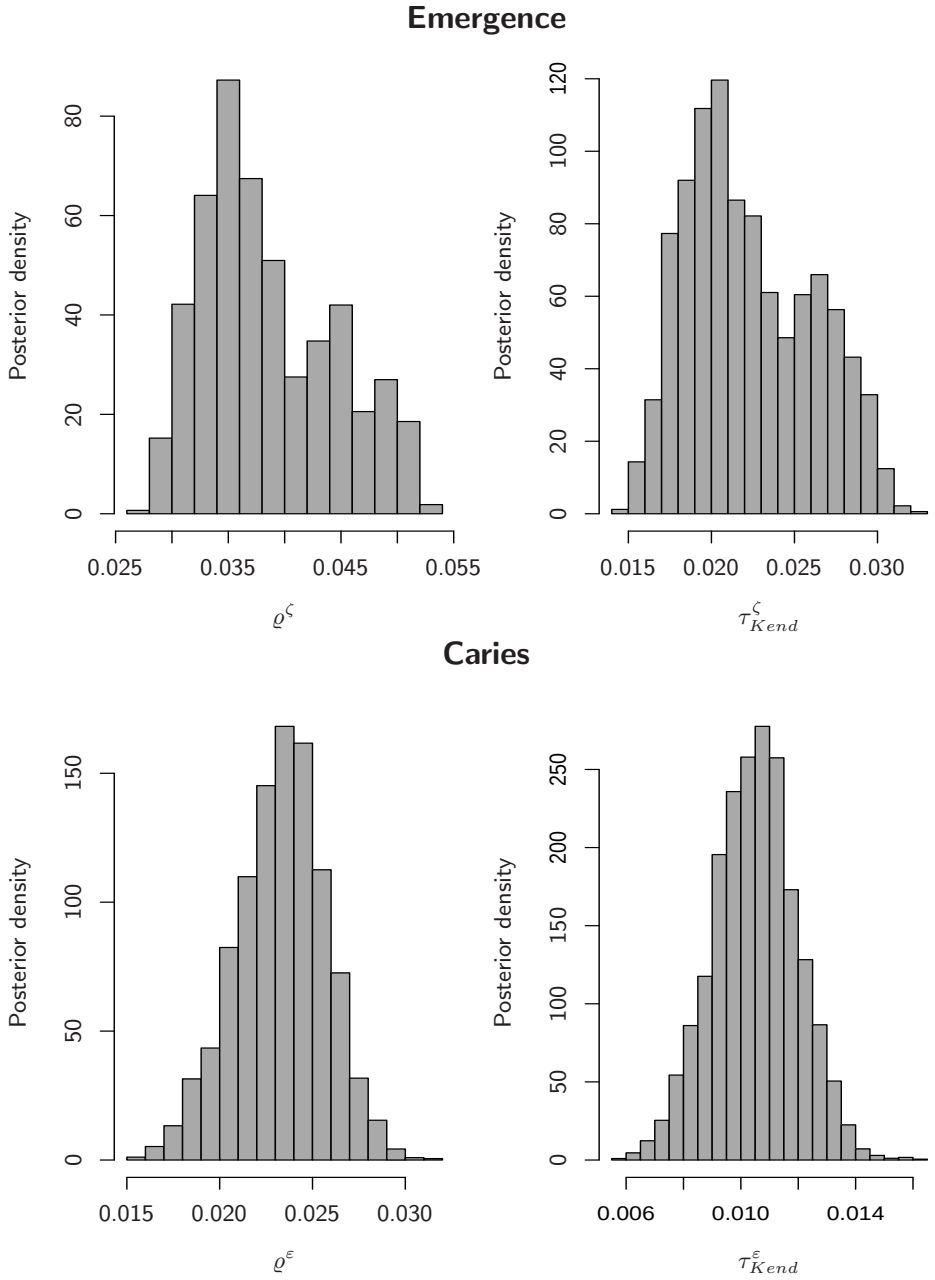


Figure 10.2: Signal Tandmobiel® study, Final Model. Scaled histograms for sampled Pearson correlation and Kendall's tau between the error terms.

effect on the time to caries. The remaining factors do influence significantly the time to caries, namely, daily brushing increases this time with a factor of $\exp(0.249) = 1.283$, presence of sealants with a factor of $\exp(0.110) = 1.116$. The factor for presence of plaque is $\exp(-0.228) = 0.796$ and when the adjacent deciduous second molar was not sound the factor is $\exp(-0.482) = 0.618$.

It is seen that the results given in Table 10.2 are slightly different from the summary given in Table 9.1 which relates to the earlier joint analysis of all four permanent first molars using the cluster-specific (conditional) AFT model. Especially, the effect of the covariate status appears to be more profound when evaluated using the population-averaged (marginal) model. However, the conclusions concerning a beneficial effect of sealing and daily brushing and an indisposed effect of not sound primary predecessors or plaque on the caries process on the permanent first molars are the same irrespective of the used model.

Further, Table 10.3 shows the mean and standard deviation of the error terms and also the residual association (after adjustment for the effect of covariates) between the maxillary and the mandibular tooth. For both the emergence and the caries processes, a very low posterior median for the Pearson correlation coefficient was found on the log-scale and the same is true also for the Kendall's tau. Moreover, as seen in Figure 10.2, the whole posterior distribution for the correlation coefficients and the Kendall's taus is concentrated in the neighborhood of zero.

Figures 10.3 and 10.4 show the estimates of the error densities $g_\zeta(\boldsymbol{\zeta})$ and $g_\varepsilon(\boldsymbol{\varepsilon})$ and their margins and illustrate the smoothing nature of our approach. These figures also reveal the low association between error terms for the upper and lower tooth. For the interpretation of the figure, we must take into account that about 75% of the caries times were right-censored and practically all around 12 years of age, which is 5 to 6 years after emergence. This implies that in fact each margin is identifiable from the data only up to approximately the first quartile. The right tail of the density is extrapolated from the left tail using the weights distributed according to the GMRF prior. It also implies that the association might be underestimated, see, e.g. Bogaerts and Lesaffre (2006).

Figure 10.5 shows the predictive survival and hazard functions for caries on the upper tooth 16 of boys and 'the best', 'the worst' and two intermediate combinations of covariates. Corresponding curves for the lower tooth 46 or for girls are almost the same due to the non-significant effect of the covariates gender and jaw on the caries. For teeth that are not brushed daily and are exposed to other risk factors, a high peak in the hazard function is

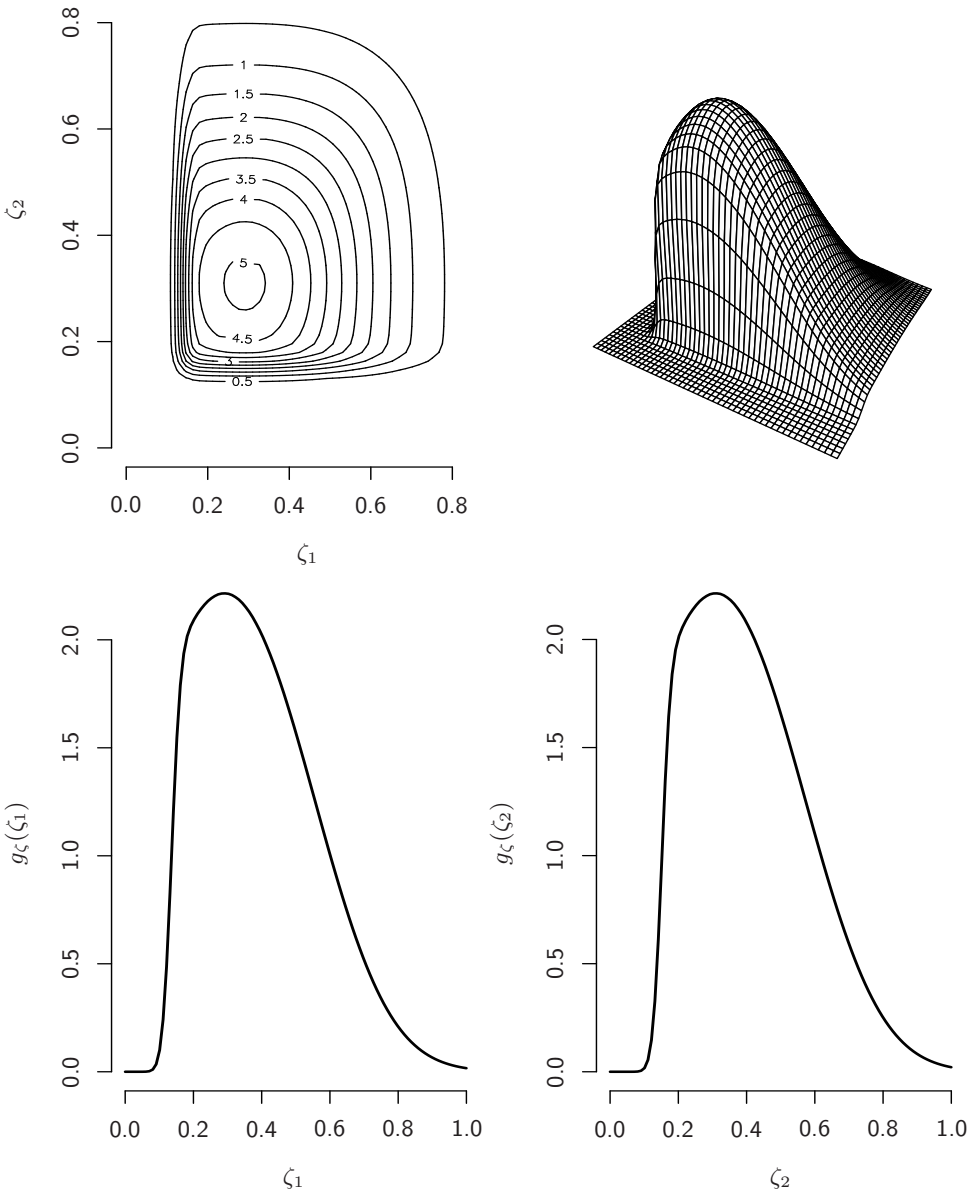


Figure 10.3: Signal Tandmobiel® study, Final Model. Estimate of the density $g_\zeta(\zeta_1, \zeta_2)$ and the corresponding marginal densities $g_\zeta(\zeta_1)$ and $g_\zeta(\zeta_2)$ of the error terms in the emergence part of the model.

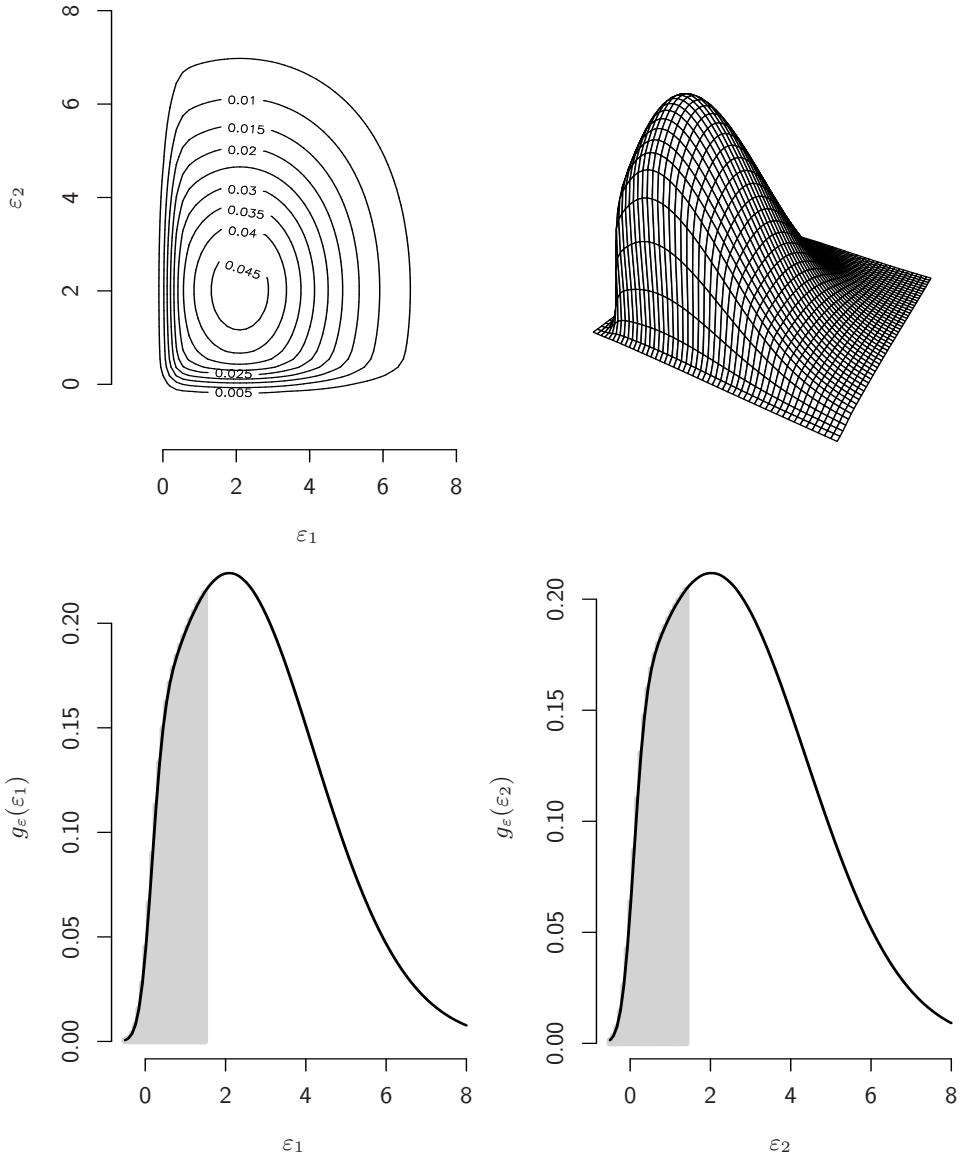


Figure 10.4: Signal Tandmobiel® study, Final Model. Estimate of the density $g_\varepsilon(\varepsilon_1, \varepsilon_2)$ and the corresponding marginal densities $g_\varepsilon(\varepsilon_1)$ and $g_\varepsilon(\varepsilon_2)$ of the error terms in the caries part of the model. The shaded part in the marginal densities extends to the first quartile.

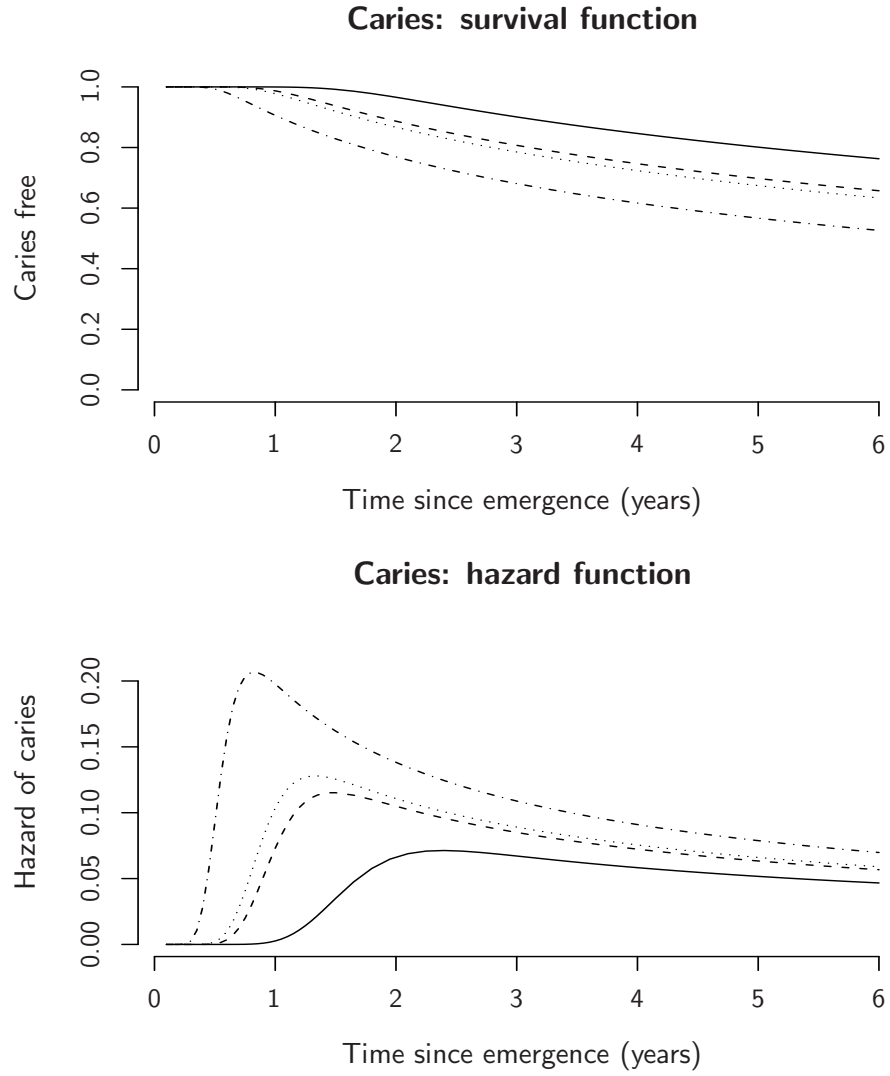


Figure 10.5: Signal Tandmobiel® study, Final Model. Posterior predictive caries free (survival) and caries hazard curves for tooth 16 of boys and the following combinations of covariates: solid and dashed lines for no plaque, present sealing, daily brushing and sound primary second molar (solid line) or dmf primary second molar (dashed line), dotted and dotted-dashed lines for present plaque, no sealing, not daily brushing and sound primary second molar (dotted line) or dmf primary second molar (dotted-dashed line).

observed already less than 1 year after emergence. A similar peak, however shifted to right and of much lower magnitude is seen also for other covariate combinations. The same peak, also approximately of the same magnitude, has already been found when analyzing all four permanent first molars using the cluster-specific AFT model in Chapter 9 (see Figure 9.4) and can be explained by the fact that permanent first molars are most vulnerable by caries soon after they emerge, possibly because of not yet fully developed enamel on their surfaces. However, when using the population-averaged model we do not see the second period of increased hazard for the 'worse' combinations of covariates as we have seen in Figure 9.4. This alleged difference between the results of the population-averaged and cluster-specific model could be caused by a failure to compare like with like, see Lee and Nelder (2004) for a deeper discussion to this point.

10.7 Discussion

In this chapter, we have suggested a semiparametric method to analyze bivariate doubly-interval-censored data in the presence of covariates. The method was applied to the analysis of a dental data set where all covariates were categorical. However, continuous covariates would not cause any difficulties and could have been used as well. Although the method was presented to deal with doubly-interval-censored data it can be used to analyze also simple interval- or right-censored data.

Further, using the ideas outlined in Section 6.3.4, the method of this chapter could theoretically be extended to handle not only bivariate data but also data of an arbitrary dimension (i.e. $n_i > 2$ for all i). However, the number of unknown parameters increases exponentially and the estimation becomes quite fast computationally intractable.

A disadvantage of the current method is that it requires balanced data, i.e. exactly two observations must be supplied for each cluster and if only one observation of the cluster is missing the whole cluster must be removed from the analysis. Missingness in one event time out of the pair could have been solved using the Bayesian data augmentation in the same way as it solves the problem of censoring. However, if the missingness is caused by a missing covariate value, the Bayesian data augmentation would not help unless a measurement model is set up also for the covariates. With unbalanced data, the cluster specific approach of Chapter 9 can be used, however.

Chapter 11

Overview and Further Research

In this thesis, we have developed several modifications of the accelerated failure time model for the analysis of the multivariate (doubly-)interval-censored data while making only weak distributional assumptions. We will now state an overview and give topics for future research.

11.1 Overview

Chapter 1 brings several data sets that motivate the developments presented in the thesis. The data sets are then used to illustrate the usage of presented methods in practical situations. Chapter 2 explains briefly several notions used in the area of survival data and introduces the notation used in the thesis.

An overview of the regression models for the analysis of the survival data is given in Chapter 3. We described the Cox's proportional hazards (PH) model and the accelerated failure time (AFT) model as the most popular models in the given area. For reasons stated in Section 3.3 we chose the accelerated failure time model as the basis for all developments in this thesis.

In Chapter 4, we discuss the likelihood form in the case of (multivariate) (doubly-)interval-censored data and show several advantages of the Bayesian inference compared to the maximum-likelihood estimation in such situations. Further, we suggest to use the Markov chain Monte Carlo methodology as the mean of Bayesian estimation.

The final chapter of the introductory part of the thesis, Chapter 5, gives an overview of existing methods for the analysis of the interval-censored data and shows in detail a Bayesian analysis of the dental multivariate doubly-

interval-censored data using a PH model with piecewise constant baseline hazard functions.

The main part of the thesis starts with Chapter 6 where we describe two slightly different classes of models for a flexible modelling of continuous densities. Firstly, a classical normal mixture is introduced and secondly, we propose a penalized normal mixture motivated by penalized B-splines as a useful tool to model unknown densities. Both approaches are subsequently used in the AFT models to express either the error density or the density of the random effects.

Chapter 7 gives the AFT model for univariate interval-censored data where the error distribution is specified as the penalized normal mixture. The inference is based on the maximum-likelihood paradigm. The model is further extended to allow not only the mean response but also the scale of the response to depend on covariates.

The AFT models presented in subsequent chapters can already handle also the multivariate (doubly-)interval-censored data. However, due to reasons discussed in Chapter 4 and in Section 7.8, we switch to the Bayesian inference. Firstly, Chapter 8 gives the AFT model with normal random effects (cluster-specific model) and the distribution of the error term specified as the classical normal mixture.

Secondly, Chapter 9 shows the cluster-specific AFT model where the error distribution and in the case of univariate random effects also the distribution of the random effects is specified as the penalized normal mixture. In this chapter, we also explicitely show and illustrate the usage of the proposed methods in the context of doubly-interval-censored data.

Finally, Chapter 10 gives the population-averaged AFT model for paired (doubly-)interval-censored data where the error distribution is given by a bivariate penalized normal mixture.

11.2 Generalizations and improvements

In this section, we list several topics to generalize or improve the models presented in this thesis.

Time-dependent covariates and joint modelling of survival data and longitudinal profiles

In many applications of the survival analysis, it is of interest to evaluate an effect of factors that can evolve over time. The values of such factors (e.g., blood pressure, dose of medication, etc.) are typically determined at (prespecified) occasions and it is assumed that they remain constant (deterministic) until the next occasion. In the last decade, several models were developed for joint modelling of the evolution of factors evolving over time (longitudinal data analysis) and the time-to-event, see Tsiatis and Davidian (2004) for an overview. That is, a stochastic component is included in the evolution of the time-dependent factors possibly influencing the survival time.

To include the time-dependent covariates, both deterministic and stochastic, in the survival model, it is necessary to specify the dependence of the survival time on the covariates using a local characteristic like the hazard function. However, in all models presented in Part II of this thesis, the covariates modified a global characteristic of the survival time, i.e. the mean log-time. The possibility on how to extend the models of this thesis to handle also the time-dependent covariates would be to use the hazard specification (3.3) of the AFT model and use a mixture model for the baseline hazard function $\hbar_0$.

Dependence of the scale parameters on covariates

In Section 7.1.2 we suggested to extend the basic AFT model by allowing the dependence of the scale parameter on the covariates. The same extension could quite easily be applied to both Bayesian penalized approaches in Chapters 9 and 10. However, in the case of the classical mixture (Chapter 8), a similar extension would be much more complicated due to the fact that the scale of the response is derived from the unknown number of the estimated variances of the mixture components.

Dependent censoring

The models in this thesis assumed all that the censoring mechanism is independent on the time-to-event (see Section 2.4). Generally, this does not

always have to be true. All Bayesian models (Chapters 8–10) could relatively easily be extended to handle also dependent censoring. However, a reasonable measurement model has to be specified for the censoring mechanism.

Goodness-of-fit

An important topic, not discussed in this thesis is the evaluation of goodness-of-fit. Indeed, in all models in this thesis, the distribution of the response is specified in a flexible manner and there is less need to evaluate the distributional assumptions. Nevertheless, one should also check an appropriatness of the AFT assumption with respect to the form in which the covariates modify the distribution of the response. On few places in this thesis, and in the case of categorical covariates, this was only graphically checked by comparing the fitted survival curves with their nonparametric estimates.

Classical goodness-of-fit methods are based on residuals whose form is straightforward in the case of a linear regression with uncensored data. In the case of right censored data, various forms of residuals are derived from the counting process specification of the survival models, see, e.g., Therneau and Grambsch (2000, Chapter 4). However, the definition of residuals for interval-censored data is not straightforward and only recently (Topp and Gómez, 2004) a work in this direction appeared in the literature.

Model selection

A general model selection is another important topic somewhat neglectful in this thesis. In Chapter 7, we based the model selection on the Akaike's information criterion whereas in Chapters 8–10 on the (simultaneous) Bayesian p-values for model contrasts. In general, also in the Bayesian framework some form of the information criterion could be used for the model selection. Recently, the most popular one seems to be the deviance information criterion (Spiegelhalter et al., 2002).

The use of specifically developed optimizers

Due to the complexity of the likelihood, we have considered the estimation through the method of penalized maximum-likelihood only in Chapter 7. However, there are currently several convenient gateways to optimization software and services available on the Internet. For example, the Kestrel interface to the NEOS server (Czyzyk, Mesnier, and Moré, 1998; Ferris, Mesnier, and Moré, 2000) together with the modeling language for mathematical

programming AMPL (Fourer, Gay, and Kernighan, 2003) enables to optimize complicated functions subject to different types of constraints. These possibilities could be explored as promising alternatives to the full Bayesian approaches presented in Chapters 8–10.

11.3 The use of penalized mixtures in other application areas

Finally, we show how we intend to use the ideas used in this thesis in the future work.

11.3.1 Generalized linear mixed models with random effects having a flexible distribution

Firstly, we aim to develop a generalized linear mixed model (GLMM) with random effects distribution specified as the penalized mixture. The proposed work has the following objectives.

Let $Y_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ be *discrete* random variables for which the components of the vector $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})'$ are possibly dependent. Typically, $\boldsymbol{Y}_i$ represents the outcomes of the ith subject at n_i different time points $t_{i,1}, \ldots, t_{i,n_i}$ in a *longitudinal study* or outcomes of n_i subjects forming the ith cluster in the case of *clustered* data. Further, let $\mu_{i,l} = \mathrm{E}(Y_{i,l})$. Using the GLMM, the expected outcome $\mu_{i,l}$ is expressed as

$$\mu_{i,l} = h^{-1}(\boldsymbol{x}'_{i,l}\boldsymbol{\beta} + \boldsymbol{z}'_{i,l}\boldsymbol{b}_i), \quad i = 1, \ldots, N,\ l = 1, \ldots, n_i,$$

where h is a known link function (e.g. log, logit, probit), $\boldsymbol{\beta}$ is the vector of unknown regression parameters (fixed effects), $\boldsymbol{x}_{i,l}$ the vector of covariates for fixed effects, $\boldsymbol{b}_i$ the vector of random effects and $\boldsymbol{z}_{i,l}$ the vector of covariates for random effects, see, e.g., Molenberghs and Verbeke (2005) for more details. We aim to concentrate mainly on longitudinal studies where usually $\boldsymbol{z}_{i,l} = (1,\ t_{i,l})'$, and $\boldsymbol{b}_i = (b_{i,1},\ b_{i,2})'$.

Classically, it is assumed that the random effects $\boldsymbol{b}_i$, $i = 1, \ldots, N$ are i.i.d. following a (multivariate) normal distribution. However, it has been shown (see Molenberghs and Verbeke, 2005, Chapter 23) that the incorrect assumption of normality of the random effects may lead to biased estimates of the regression parameters $\boldsymbol{\beta}$. But, due to the fact that the random effects $\boldsymbol{b}_i$ are latent, it is very difficult to check the normality assumption. That is why one strives for more flexible methods with respect to the distribution of the ran-

dom effects. One possibility, we wish to explore is to specify the distribution of the random effects as a penalized bivariate mixture (10.3).

11.3.2 Spatial models with the intensity specified by the penalized mixture

Secondly, we would like to explore the possibilities of the penalized mixtures in the context of spatial models. The motivation is the following. In epidemiology, it is of interest to model the prevalence or incidence of a disease in a spatial manner in order to represent the true risk in a honest manner. Let A denote the study area, R a region within A, and $\boldsymbol{y} = (y_1, y_2)$ coordinates of a location in A. Generation of the disease cases can be formalized by considering an underlying *point process* described by a *counting measure* N on A, i.e. $N(R)$ denotes the number of disease cases in R. Finally, let

$$\lambda(\boldsymbol{y}) = \lim_{\|\Delta\,\boldsymbol{y}\|\to 0} \frac{\mathrm{E}\{N(\Delta\boldsymbol{y})\}}{\|\Delta\,\boldsymbol{y}\|},$$

where $\Delta\,\boldsymbol{y}$ is an infinitesimal region around $\boldsymbol{y}$ and $\|\Delta\,\boldsymbol{y}\|$ its area, be the *intensity* of the point process. Different approaches have been suggested in the literature to express $\lambda(\boldsymbol{y})$ of which one uses an expression

$$\lambda(\boldsymbol{y}) = \varrho\, g(\boldsymbol{y})\, f(\boldsymbol{y};\, \boldsymbol{\theta}), \tag{11.1}$$

where ϱ denotes an overall region-wide rate, $g(\boldsymbol{y})$ a known background function representing the reference population and $f(\boldsymbol{y};\, \boldsymbol{\theta})$ represents a function of spatial location and possibly other parameters and associated covariates as well (see Lawson et al., 1999). However, there exists no gold standard for the expression of $f(\boldsymbol{y};\, \boldsymbol{\theta})$. The main requirement for $f(\boldsymbol{y};\, \boldsymbol{\theta})$ is, however, that it varies smoothly across A.

To model smoothly the variation of the intensity $\lambda(\boldsymbol{y})$ across the region of interest A, a penalized mixture could be used to express $f(\boldsymbol{y};\, \boldsymbol{\theta})$ as part of expression (11.1) as

$$f(\boldsymbol{y};\, \boldsymbol{\theta}) = 1 + \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} w_{k_1,k_2} \varphi_{k_1}(y_1)\, \varphi_{k_2}(y_2), \tag{11.2}$$

where the weights are, in contrast to the approaches used in this thesis, not constrained.

Further, it is here of interest to develop efficient procedures (a) to test a null hypothesis of $w_{-K_1,-K_2} = \cdots = w_{K_1,K_2} = 0$, corresponding to a constant ratio $\lambda(\boldsymbol{y})/g(\boldsymbol{y})$ which is known as *a standardized mortality rate*, and (b) to develop a general procedure for model selection.

Further, to allow for the dependence of the intensity $\lambda(\boldsymbol{y})$ on other (region-specific) covariates $\boldsymbol{x}(\boldsymbol{y})$, we would like to explore a generalization of the model (11.2) of the form

$$f(\boldsymbol{y};\, \boldsymbol{\theta},\, \boldsymbol{\beta}) = h\big\{\boldsymbol{x}(\boldsymbol{y}),\, \boldsymbol{\beta}\big\} + \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} w_{k_1,k_2} \varphi_{k_1}(y_1)\, \varphi_{k_2}(y_2),$$

where h is an unknown (nonlinear) function and $\boldsymbol{\beta}$ a vector of unknown regression parameters.

Appendix A

Technical details for the Maximum Likelihood Penalized AFT Model

This appendix provides the technical details for the practical computation of the penalized maximum-likelihood estimate for the AFT model of Chapter 7. Namely, we give more details concerning the optimization algorithm, provide the formulas for computation of the first and second derivatives of the penalized log-likelihood needed to implement this algorithm and give the proof of Proposition 7.1.

Notation introduced in Chapter 7 will be used throughout this appendix. Additionally, the following notation is employed.

$$
\begin{aligned}
e_i^L &= \tau_i^{-1}(y_i^L - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i), & e_i^U &= \tau_i^{-1}(y_i^U - \alpha - \boldsymbol{\beta}'\boldsymbol{x}_i),\\
\tilde{e}_{i,j}^L &= \sigma^{-1}(e_i^L - \mu_j), & \tilde{e}_{i,j}^U &= \sigma^{-1}(e_i^U - \mu_j),\\
\varphi_{i,j}^L &= \varphi(\tilde{e}_{i,j}^L), & \varphi_{i,j}^U &= \varphi(\tilde{e}_{i,j}^U),\\
\bar{\varphi}_{i,j}^L &= \tilde{e}_{i,j}^L\varphi(\tilde{e}_{i,j}^L), & \bar{\varphi}_{i,j}^U &= \tilde{e}_{i,j}^U\varphi(\tilde{e}_{i,j}^U),\\
\breve{\varphi}_{i,j}^L &= \left\{\left(\tilde{e}_{i,j}^L\right)^2 - 1\right\}\varphi(\tilde{e}_{i,j}^L), & \breve{\varphi}_{i,j}^U &= \left\{\left(\tilde{e}_{i,j}^U\right)^2 - 1\right\}\varphi(\tilde{e}_{i,j}^U),\\
\Phi_{i,j}^L &= \Phi(\tilde{e}_{i,j}^L) & \Phi_{i,j}^U &= \Phi(\tilde{e}_{i,j}^U),\\
i &= 1,\ldots,N, & j &= -K,\ldots,K.
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\varphi}_i^L &= (\varphi_{i,-K}^L, \ldots, \varphi_{i,K}^L)', & \boldsymbol{\varphi}_i^U &= (\varphi_{i,-K}^U, \ldots, \varphi_{i,K}^U)', \\
\bar{\boldsymbol{\varphi}}_i^L &= (\bar{\varphi}_{i,-K}^L, \ldots, \bar{\varphi}_{i,K}^L)', & \bar{\boldsymbol{\varphi}}_i^U &= (\bar{\varphi}_{i,-K}^U, \ldots, \bar{\varphi}_{i,K}^U)', \\
\breve{\boldsymbol{\varphi}}_i^L &= (\breve{\varphi}_{i,-K}^L, \ldots, \breve{\varphi}_{i,K}^L)', & \breve{\boldsymbol{\varphi}}_i^U &= (\breve{\varphi}_{i,-K}^U, \ldots, \breve{\varphi}_{i,K}^U)', \\
\boldsymbol{\Phi}_i^L &= (\bar{\Phi}_{i,-K}^L, \ldots, \bar{\Phi}_{i,K}^L)', & \boldsymbol{\Phi}_i^U &= (\bar{\Phi}_{i,-K}^U, \ldots, \bar{\Phi}_{i,K}^U)', \\
& i = 1, \ldots, N.
\end{aligned}
$$

We omit the superscripts 'L' and 'U' in the case of exactly observed event times ($\delta_i = 1$) resulting in $y_i^L = y_i^U = y_i$. Finally, in all formulas, we omit the Jacobian term (t_i^{-1} for exactly observed event times with $t_i^L = t_i^U = t_i$) resulting from the logarithmic transformation of the event times in the log-likelihood.

A.1 Optimization algorithm

To compute the penalized maximum-likelihood estimate we firstly maximize the penalized log-likelihood (7.7) with respect to $\tilde{\boldsymbol{\theta}} = (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{a}_{-0}')'$ under the constraints (7.4) and upon the convergence we compute the second derivative matrix of ℓ_P with respect to $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{d}')'$ to get the variance estimates.

Constrained optimization is conducted using the sequential quadratic programming (SQP) algorithm, see Han (1977); Fletcher (1987, Section 12.4). The idea of this algorithm is to iteratively maximize slightly modified quadratic approximation of the objective function subject to the linear approximation of the constraints.

Let

$$
c_1(\tilde{\boldsymbol{\theta}}) = \sum_{j=-K}^{K} w_j \mu_j, \qquad c_2(\tilde{\boldsymbol{\theta}}) = 1 - \sigma_0^2 - \sum_{j=-K}^{K} w_j \mu_j^2 \tag{A.1}
$$

be the constraint equations resulting from (7.4), and let

$$
\mathcal{L}(\tilde{\boldsymbol{\theta}}, \xi_1, \xi_2) = \ell_P(\tilde{\boldsymbol{\theta}}) \; + \; \xi_1 c_1(\tilde{\boldsymbol{\theta}}) \; + \; \xi_2 c_2(\tilde{\boldsymbol{\theta}})
$$

be the Lagrange function with the Lagrange multipliers ξ_1 and ξ_2 corresponding to the maximization problem $\max_{\tilde{\boldsymbol{\theta}}} \ell_P(\tilde{\boldsymbol{\theta}})$ subject to $c_1(\tilde{\boldsymbol{\theta}}) = 0$ and $c_2(\tilde{\boldsymbol{\theta}}) = 0$.

Let $\mathrm{QP}(\tilde{\boldsymbol{\theta}}, \mathbb{H})$ be a quadratic programming problem

$$
\max_{\boldsymbol{\delta}} \Big\{ \boldsymbol{\delta}' \frac{\partial \ell_P}{\partial \tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}}) \; + \; 0.5\, \boldsymbol{\delta}'\, \mathbb{H}\, \boldsymbol{\delta} \Big\} \tag{A.2}
$$

subject to

$$c_1(\tilde{\boldsymbol{\theta}}) \;+\; \boldsymbol{\delta}'\frac{\partial c_1}{\partial \tilde{\boldsymbol{\theta}}}\big(\tilde{\boldsymbol{\theta}}\big) = 0, \qquad c_2(\tilde{\boldsymbol{\theta}}) \;+\; \boldsymbol{\delta}'\frac{\partial c_2}{\partial \tilde{\boldsymbol{\theta}}}\big(\tilde{\boldsymbol{\theta}}\big) = 0, \tag{A.3}$$

where

$$\mathbb{H} = \mathbb{H}(\tilde{\boldsymbol{\theta}},\, \xi_1,\, \xi_2) = \frac{\partial^2 \mathcal{L}}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}'}\big(\tilde{\boldsymbol{\theta}},\, \xi_1,\, \xi_2\big). \tag{A.4}$$

Note that the objective function in (A.2) is the second order Taylor approximation of $\ell_P(\tilde{\boldsymbol{\theta}})$ around some fixed point $\tilde{\boldsymbol{\theta}}_0$ with $\boldsymbol{\delta} = \tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_0$, omitted constant term and the matrix of second derivatives $\partial^2 \ell_P / \partial \tilde{\boldsymbol{\theta}} \partial \boldsymbol{\theta}'$ replaced by the $\tilde{\boldsymbol{\theta}}$-$\tilde{\boldsymbol{\theta}}$ block of the second derivative matrix of the Lagrange function $\mathcal{L}$.

The SQP algorithm proceeds in the following steps

Step 0. Give the initial estimate $\tilde{\boldsymbol{\theta}}^{(0)}$ and the initial guesses $\xi_1^{(0)}$, $\xi_2^{(0)}$ for the Lagrange multipliers. Set $\mathbb{H}^{(0)} = \mathbb{H}\big(\tilde{\boldsymbol{\theta}}^{(0)},\, \xi_1^{(0)},\, \xi_2^{(0)}\big)$.

In the sth iteration:

Step 1. Find the point $\boldsymbol{\delta}^{(s)}$ which solves the quadratic program $\mathrm{QP}(\tilde{\boldsymbol{\theta}}^{(s)},\, \mathbb{H}^{(s)})$;

Step 2. Set

$$\tilde{\boldsymbol{\theta}}^{(s+1)} = \tilde{\boldsymbol{\theta}}^{(s)} \;+\; \boldsymbol{\delta}^{(s)}.$$

If $\tilde{\boldsymbol{\theta}}^{(s+1)}$ does not lead to increase of ℓ_P use step-halving procedure;

Step 3. Set $\xi_1^{(s+1)}$ and $\xi_2^{(s+1)}$ to the optimal Lagrangian multipliers of the quadratic program $\mathrm{QP}(\tilde{\boldsymbol{\theta}}^{(s)},\, \mathbb{H}^{(s)})$;

Step 4. Check the convergence, if it is not reached go to Step 1.

A.2 Individual log-likelihood contributions

$$\ell_i(\tilde{\boldsymbol{\theta}}) = \begin{cases} \log(1 - \boldsymbol{w}'\boldsymbol{\Phi}_i^L), & \delta_i = 0, \\ -\log(\tau_i) + \log(\boldsymbol{w}'\boldsymbol{\varphi}_i), & \delta_i = 1, \\ \log(\boldsymbol{w}'\boldsymbol{\Phi}_i^U), & \delta_i = 2, \\ \log\big\{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)\big\}, & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N.$$

A.3 First derivatives of the log-likelihood

A.3.1 With respect to the regression parameters and the intercept

$$\frac{\partial \ell}{\partial \alpha} = (\tau_i \sigma_0)^{-1}\, \boldsymbol{w}' \Big(\sum_{i=1}^{N} d\boldsymbol{b}_i \Big),$$
$$\frac{\partial \ell}{\partial \beta_l} = (\tau_i \sigma_0)^{-1}\, \boldsymbol{w}' \Big(\sum_{i=1}^{N} x_{i,l}\, d\boldsymbol{b}_i \Big), \qquad l = 1, \ldots, m,$$

where $d\boldsymbol{b}_i$ is a vector of length $2K+1$ of the form

$$d\boldsymbol{b}_i = \begin{cases} (1 - \boldsymbol{w}' \boldsymbol{\Phi}_i^L)^{-1}\, \boldsymbol{\varphi}_i^L, & \delta_i = 0, \\ (\boldsymbol{w}' \boldsymbol{\varphi}_i)^{-1}\, \bar{\boldsymbol{\varphi}}_i, & \delta_i = 1, \\ -(\boldsymbol{w}' \boldsymbol{\Phi}_i^U)^{-1}\, \boldsymbol{\varphi}_i^U, & \delta_i = 2, \\ \big\{ \boldsymbol{w}' (\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L) \big\}^{-1} (\boldsymbol{\varphi}_i^L - \boldsymbol{\varphi}_i^U), & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N.$$

A.3.2 With respect to the log-scale and the scale-regression parameters

Firstly, we consider the case when the scale parameter τ does not depend on covariates, i.e. $\log(\tau) = \gamma_1$.

$$\frac{\partial \ell}{\partial \gamma_1} = -\sum_{i=1}^{N} I[\delta_i = 1] + \sigma_0^{-1}\, \boldsymbol{w}' \Big(\sum_{i=1}^{N} d\boldsymbol{l}_i \Big).$$

Secondly, we consider the case when $\log(\tau_i) = \boldsymbol{\gamma}' \boldsymbol{z}_i$, where $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,m_s})'$. Then

$$\frac{\partial \ell}{\partial \gamma_l} = -\sum_{i=1}^{N} I[\delta_i = 1]\, z_{i,l} + \sigma_0^{-1}\, \boldsymbol{w}' \Big(\sum_{i=1}^{N} z_{i,l}\, d\boldsymbol{l}_i \Big), \qquad l = 1, \ldots, m_s.$$

In both formulas, $d\boldsymbol{l}_i$ is a vector of length $2K+1$ of the form

$$d\boldsymbol{l}_i = \begin{cases} \Big\{ \boldsymbol{w}' (\mathbf{1} - \boldsymbol{\Phi}_i^L) \Big\}^{-1} e_i^L\, \boldsymbol{\varphi}_i^L, & \delta_i = 0, \\ (\boldsymbol{w}' \boldsymbol{\varphi}_i)^{-1}\, e_i\, \bar{\boldsymbol{\varphi}}_i, & \delta_i = 1, \\ -(\boldsymbol{w}' \boldsymbol{\Phi}_i^U)^{-1}\, e_i^U\, \boldsymbol{\varphi}_i^U, & \delta_i = 2, \\ \big\{ \boldsymbol{w}' (\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L) \big\}^{-1} (e_i^L\, \boldsymbol{\varphi}_i^L - e_i^U\, \boldsymbol{\varphi}_i^U), & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N.$$

A.3.3 With respect to the transformed mixture weights

Let $\boldsymbol{a}_{-0}$ be the vector of transformed mixture weights except the baseline coefficient which is fixed to zero (without loss of generality $a_0 = 0$). Then

$$\frac{\partial \ell}{\partial \boldsymbol{a}_{-0}} = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}} \sum_{i=1}^{N} d\boldsymbol{a}_i,$$

where $d\boldsymbol{a}_i$ is a vector of length $2K+1$ of the form

$$d\boldsymbol{a}_i = \begin{cases} \left\{\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)\right\}^{-1} (\mathbf{1} - \boldsymbol{\Phi}_i^L), & \delta_i = 0, \\ (\boldsymbol{w}'\boldsymbol{\varphi}_i)^{-1}\, \boldsymbol{\varphi}_i, & \delta_i = 1, \\ (\boldsymbol{w}'\boldsymbol{\Phi}_i^U)^{-1}\, \boldsymbol{\Phi}_i^U, & \delta_i = 2, \\ \left\{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)\right\}^{-1} (\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L), & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N,$$

and $\partial \boldsymbol{w}/\partial \boldsymbol{a}_{-0}$ is a $2K \times (2K+1)$ matrix whose (j,k)th element equals $\partial w_k / \partial a_j$, $j = -K, \ldots, -1, 1, \ldots, K$, $k = -K, \ldots, K$. Namely

$$\begin{aligned} \frac{\partial w_j}{\partial a_j} &= w_j\,(1 - w_j), && j = -K, \ldots, -1, 1, \ldots, K, \\ \frac{\partial w_k}{\partial a_j} &= -w_j\, w_k, && j = -K, \ldots, -1, 1, \ldots, K,\ k = -K, \ldots, K,\ j \neq k. \end{aligned}$$

A.4 Second derivatives of the log-likelihood

Let $\tilde{\boldsymbol{\beta}}$ be the vector of regression parameters extended by the intercept, i.e. $\tilde{\boldsymbol{\beta}} = (\alpha,\, \boldsymbol{\beta}')'$ and $\tilde{\boldsymbol{x}}_i$, $i = 1, \ldots, N$ be the covariate vectors extended by the intercept term, i.e. $\tilde{\boldsymbol{x}}_i = (1,\, \boldsymbol{x}_i')'$.

A.4.1 With respect to the extended regression parameters

$$\frac{\partial^2 \ell}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} = \sum_{i=1}^{N} (ddbb_{i,1} - ddbb_{i,2}^2)\, \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i',$$

where $ddbb_{i,1}$ and $ddbb_{i,2}$ are scalars of the following form

$$ddbb_{i,1} = \begin{cases} (\tau_i\sigma_0)^{-2}\,\dfrac{\boldsymbol{w}'\bar{\boldsymbol{\varphi}}_i^L}{\boldsymbol{w}'(\mathbf{1}-\boldsymbol{\Phi}_i^L)}, & \delta_i = 0, \\ (\tau_i\sigma_0)^{-2}\,\dfrac{\boldsymbol{w}'\breve{\boldsymbol{\varphi}}_i}{\boldsymbol{w}'\boldsymbol{\varphi}_i}, & \delta_i = 1, \\ -(\tau_i\sigma_0)^{-2}\,\dfrac{\boldsymbol{w}'\bar{\boldsymbol{\varphi}}_i^U}{\boldsymbol{w}'\boldsymbol{\Phi}_i^U}, & \delta_i = 2, \\ (\tau_i\sigma_0)^{-2}\,\dfrac{\boldsymbol{w}'(\bar{\boldsymbol{\varphi}}_i^L-\bar{\boldsymbol{\varphi}}_i^U)}{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U-\boldsymbol{\Phi}_i^L)}, & \delta_i = 3, \end{cases} \qquad i = 1,\ldots,N,$$

$$ddbb_{i,2} = \begin{cases} (\tau_i\sigma_0)^{-1}\dfrac{\boldsymbol{w}'\boldsymbol{\varphi}_i^L}{\boldsymbol{w}'(\mathbf{1}-\boldsymbol{\Phi}_i^L)}, & \delta_i = 0, \\ (\tau_i\sigma_0)^{-1}\dfrac{\boldsymbol{w}'\bar{\boldsymbol{\varphi}}_i}{\boldsymbol{w}'\boldsymbol{\varphi}_i}, & \delta_i = 1, \\ -(\tau_i\sigma_0)^{-1}\dfrac{\boldsymbol{w}'\boldsymbol{\varphi}_i^U}{\boldsymbol{w}'\boldsymbol{\Phi}_i^U}, & \delta_i = 2, \\ (\tau_i\sigma_0)^{-1}\dfrac{\boldsymbol{w}'(\boldsymbol{\varphi}_i^L-\boldsymbol{\varphi}_i^U)}{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U-\boldsymbol{\Phi}_i^L)}, & \delta_i = 3, \end{cases} \qquad i = 1,\ldots,N.$$

A.4.2 Mixed with respect to the extended regression parameters and the log-scale or the scale-regression parameters

In the case when the scale parameter does not depend on covariates we have

$$\frac{\partial^2\ell}{\partial\tilde{\boldsymbol{\beta}}\partial\gamma_1} = \sum_{i=1}^{N}\big\{ddbl_{i,1} - ddbb_{i,2}(1+ddbl_{i,2})\big\}\,\tilde{\boldsymbol{x}}_i.$$

In the case of $\log(\tau_i) = \boldsymbol{\gamma}'\boldsymbol{z}_i$ we have

$$\frac{\partial^2\ell}{\partial\tilde{\boldsymbol{\beta}}\partial\boldsymbol{\gamma}'} = \sum_{i=1}^{N}\big\{ddbl_{i,1} - ddbb_{i,2}(1+ddbl_{i,2})\big\}\,\tilde{\boldsymbol{x}}_i\boldsymbol{z}_i'.$$

In both formulas, $ddbb_{i,2}$ is given in Section A.4.1, $ddbl_{i,1}$ and $ddbl_{i,2}$ are scalars of the form

$$ddbl_{i,1} = \begin{cases} \dfrac{e_i^L}{\tau_i\sigma_0^2} \cdot \dfrac{\boldsymbol{w}'\bar{\boldsymbol{\varphi}}_i^L}{\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)}, & \delta_i = 0, \\ \dfrac{e_i}{\tau_i\sigma_0^2} \cdot \dfrac{\boldsymbol{w}'\breve{\boldsymbol{\varphi}}_i}{\boldsymbol{w}'\boldsymbol{\varphi}_i}, & \delta_i = 1, \\ -\dfrac{e_i^U}{\tau_i\sigma_0^2} \cdot \dfrac{\boldsymbol{w}'\bar{\boldsymbol{\varphi}}_i^U}{\boldsymbol{w}'\boldsymbol{\Phi}_i^U}, & \delta_i = 2, \\ \dfrac{1}{\tau_i\sigma_0^2} \cdot \dfrac{\boldsymbol{w}'(e_i^L\,\bar{\boldsymbol{\varphi}}_i^L - e_i^U\,\bar{\boldsymbol{\varphi}}_i^U)}{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)}, & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N,$$

$$ddbl_{i,2} = \begin{cases} \dfrac{e_i^L}{\sigma_0} \cdot \dfrac{\boldsymbol{w}'\boldsymbol{\varphi}_i^L}{\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)} & \delta_i = 0, \\ \dfrac{e_i}{\sigma_0} \cdot \dfrac{\boldsymbol{w}'\bar{\boldsymbol{\varphi}}_i}{\boldsymbol{w}'\boldsymbol{\varphi}_i} & \delta_i = 1, \\ -\dfrac{e_i^U}{\sigma_0} \cdot \dfrac{\boldsymbol{w}'\boldsymbol{\varphi}_i^U}{\boldsymbol{w}'\boldsymbol{\Phi}_i^U} & \delta_i = 2, \\ \dfrac{1}{\sigma_0} \cdot \dfrac{\boldsymbol{w}'(e_i^L\,\boldsymbol{\varphi}_i^L - e_i^U\,\boldsymbol{\varphi}_i^U)}{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)} & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N.$$

A.4.3 Mixed with respect to the extended regression parameters and the transformed mixture weights

$$\frac{\partial^2 \ell}{\partial\tilde{\boldsymbol{\beta}}\partial\boldsymbol{a}_{-0}'} = \left[\sum_{i=1}^{N}\Big\{dd\boldsymbol{ba}_i - (\tau_i\sigma_0)^{-1}\,(\boldsymbol{w}'d\boldsymbol{b}_i)\,\tilde{\boldsymbol{x}}_i\,d\boldsymbol{a}_i'\Big\}\right]\Big(\frac{\partial\boldsymbol{w}}{\partial\boldsymbol{a}_{-0}}\Big)',$$

where $dd\boldsymbol{ba}_i$ is a $(m+1)\times(2K+1)$ matrix of the form

$$dd\boldsymbol{ba}_i = \begin{cases} \big\{\tau_i\sigma_0\,\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)\big\}^{-1}\,\tilde{\boldsymbol{x}}_i\boldsymbol{\varphi}_i^{L\prime}, & \delta_i = 0, \\ (\tau_i\sigma_0\,\boldsymbol{w}'\boldsymbol{\varphi}_i)^{-1}\,\tilde{\boldsymbol{x}}_i\bar{\boldsymbol{\varphi}}_i', & \delta_i = 1, \\ -(\tau_i\sigma_0\,\boldsymbol{w}'\boldsymbol{\Phi}_i^U)^{-1}\,\tilde{\boldsymbol{x}}_i\boldsymbol{\varphi}_i^{U\prime}, & \delta_i = 2, \\ \big\{\tau_i\sigma_0\,\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)\big\}^{-1}\,\tilde{\boldsymbol{x}}_i(\boldsymbol{\varphi}_i^L - \boldsymbol{\varphi}_i^U)', & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N.$$

Further, $d\boldsymbol{b}_i$ is a vector of length $2K+1$ given in Section A.3.1. Finally, $d\boldsymbol{a}_i$ and $\partial\boldsymbol{w}/\partial\boldsymbol{a}_{-0}$ are a vector of length $2K+1$ and a $2K\times(2K+1)$ matrix, respectively, given in Section A.3.3.

A.4.4 With respect to the log-scale or the scale-regression parameters

In the case when the scale parameter does not depend on covariates we have

$$\frac{\partial^2 \ell}{\partial \gamma_1^2} = \sum_{i=1}^{N} \Big\{ ddll_i - ddbl_{i,2}(1 + ddbl_{i,2}) \Big\}.$$

In the case of $\log(\tau_i) = \boldsymbol{\gamma}' \boldsymbol{z}_i$ we have

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \sum_{i=1}^{N} \Big\{ ddll_i - ddbl_{i,2}(1 + ddbl_{i,2}) \Big\} \boldsymbol{z}_i \boldsymbol{z}_i'.$$

In both formulas, $ddbl_{i,2}$ is a scalar given in Section A.4.2 and $ddll_i$ is a scalar given by the formula

$$ddll_i = \begin{cases} \left(\frac{e_i^L}{\sigma_0}\right)^2 \cdot \dfrac{\boldsymbol{w}' \bar{\boldsymbol{\varphi}}_i^L}{\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)}, & \delta_i = 0, \\ \left(\frac{e_i}{\sigma_0}\right)^2 \cdot \dfrac{\boldsymbol{w}' \breve{\boldsymbol{\varphi}}_i}{\boldsymbol{w}' \boldsymbol{\varphi}_i}, & \delta_i = 1, \\ -\left(\frac{e_i^U}{\sigma_0}\right)^2 \cdot \dfrac{\boldsymbol{w}' \bar{\boldsymbol{\varphi}}_i^U}{\boldsymbol{w}' \boldsymbol{\Phi}_i^U}, & \delta_i = 2, \\ \sigma_0^{-2} \dfrac{\boldsymbol{w}' \Big\{ (e_i^L)^2 \bar{\boldsymbol{\varphi}}_i^L - (e_i^U)^2 \bar{\boldsymbol{\varphi}}_i^U \Big\}}{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)}, & \delta_i = 3, \end{cases} \qquad i = 1, \ldots, N.$$

A.4.5 Mixed with respect to the log-scale or the scale-regression parameters and the transformed mixture weights

In the case when the scale parameter does not depend on covariates we have

$$\frac{\partial^2 \ell}{\partial \gamma_1 \partial \boldsymbol{a}_{-0}'} = \left[\sum_{i=1}^{N} \Big\{ ddl\boldsymbol{a}_i - \sigma_0^{-1} \left(\boldsymbol{w}' d\boldsymbol{l}_i \right) d\boldsymbol{a}_i' \Big\} \right] \left(\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}} \right)'.$$

In the case of $\log(\tau_i) = \boldsymbol{\gamma}' \boldsymbol{z}_i$ we have

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{a}_{-0}'} = \boldsymbol{z}_i \left[\sum_{i=1}^{N} \Big\{ ddl\boldsymbol{a}_i - \sigma_0^{-1} \left(\boldsymbol{w}' d\boldsymbol{l}_i \right) d\boldsymbol{a}_i' \Big\} \right] \left(\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}} \right)'.$$

In both formulas, $d\boldsymbol{a}_i$ and $\partial \boldsymbol{w}/\partial \boldsymbol{a}_{-0}$ are a vector of length $2K+1$ and a $2K \times (2K+1)$ matrix, respectively, given in Section A.3.3, and $ddl\boldsymbol{a}_i$ is a row vector of length $2K+1$ of the form

$$ddl\boldsymbol{a}_i = \begin{cases} \sigma_0^{-1} \dfrac{e_i^L}{\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)} \boldsymbol{\varphi}_i^{L\prime}, & \delta_i = 0, \\ \sigma_0^{-1} \dfrac{e_i}{\boldsymbol{w}'\boldsymbol{\varphi}_i} \bar{\boldsymbol{\varphi}}_i', & \delta_i = 1, \\ -\sigma_0^{-1} \dfrac{e_i^U}{\boldsymbol{w}'\boldsymbol{\Phi}_i^U} \boldsymbol{\varphi}_i^{U\prime}, & \delta_i = 2, \\ \left\{\sigma_0 \boldsymbol{w}'\big(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L\big)\right\}^{-1} \big(e_i^L \boldsymbol{\varphi}_i^L - e_i^U \boldsymbol{\varphi}_i^U\big)', & \delta_i = 3, \end{cases}$$
$$i = 1, \ldots, N.$$

A.4.6 With respect to the transformed mixture weights

$$\frac{\partial^2 \ell}{\partial \boldsymbol{a}_{-0} \partial \boldsymbol{a}_{-0}'} = \sum_{i=1}^N dd\boldsymbol{a}\boldsymbol{a}_i \; - \; \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}} \left(\sum_{i=1}^N d\boldsymbol{a}_i d\boldsymbol{a}_i' \right) \left(\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}} \right)',$$

where $d\boldsymbol{a}_i$ and $\partial \boldsymbol{w}/\partial \boldsymbol{a}_{-0}$ are a vector of length $2K+1$ and a $2K \times (2K+1)$ matrix, respectively, given in Section A.3.3. Further, $dd\boldsymbol{a}\boldsymbol{a}_i$ is a $2K \times 2K$ matrix given by

$$dd\boldsymbol{a}\boldsymbol{a}_i = \begin{cases} \big\{\boldsymbol{w}'(\mathbf{1} - \boldsymbol{\Phi}_i^L)\big\}^{-1} \sum_{j=-K}^K (1 - \Phi_{i,j}^L) \dfrac{\partial^2 w_j}{\partial \boldsymbol{a}_{-0} \partial \boldsymbol{a}_{-0}'}, & \delta_i = 0, \\ (\boldsymbol{w}'\boldsymbol{\varphi}_i)^{-1} \sum_{j=-K}^K \varphi_{i,j} \dfrac{\partial^2 w_j}{\partial \boldsymbol{a}_{-0} \partial \boldsymbol{a}_{-0}'}, & \delta_i = 1, \\ (\boldsymbol{w}'\boldsymbol{\Phi}_i^U)^{-1} \sum_{j=-K}^K \Phi_{i,j}^U \dfrac{\partial^2 w_j}{\partial \boldsymbol{a}_{-0} \partial \boldsymbol{a}_{-0}'}, & \delta_i = 2, \\ \big\{\boldsymbol{w}'(\boldsymbol{\Phi}_i^U - \boldsymbol{\Phi}_i^L)\big\}^{-1} \sum_{j=-K}^K (\Phi_{i,j}^U - \Phi_{i,j}^L) \dfrac{\partial^2 w_j}{\partial \boldsymbol{a}_{-0} \partial \boldsymbol{a}_{-0}'}, & \delta_i = 3, \end{cases}$$
$$i = 1, \ldots, N,$$

where $\partial^2 w_j/\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}$, $j = -K, \ldots, K$ is a $2K \times 2K$ matrix with the elements $ddwaa^j_{k,l}$, $k, l = -K, \ldots, -1, 1, \ldots, K$ given by

$$\begin{aligned}
ddwaa^j_{j,j} &= w_j\,(1-w_j)\,(1-2w_j), && j \neq 0,\\
ddwaa^j_{k,k} &= -w_j\,w_k\,(1-2w_k), && k \neq j,\\
ddwaa^j_{j,k} &= -w_j\,w_k\,(1-2w_j), && j \neq 0, \quad k \neq j,\\
ddwaa^j_{k,j} &= -w_j\,w_k\,(1-2w_j), && j \neq 0, \quad k \neq j,\\
ddwaa^j_{k,l} &= 2\,w_j\,w_k\,w_l, && k \neq j, \quad l \neq j, \quad k \neq l.
\end{aligned}$$

A.5 Derivatives of the penalty term

The penalty term depends only on the $\boldsymbol{a}_{-0}$ part of $\tilde{\boldsymbol{\theta}}$ so we have to provide only the derivatives with respect to this parameter sub-vector.

$$\begin{aligned}
\frac{\partial q}{\partial \boldsymbol{a}_{-0}} &= \lambda\,\mathbb{D}'\mathbb{D}\boldsymbol{a} \quad \text{with removed 0th element,}\\
\frac{\partial^2 q}{\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}} &= \lambda\,\mathbb{D}'\mathbb{D} \quad \text{with removed 0th row and 0th column.}
\end{aligned}$$

A.6 Derivatives of the constraints

To be able to compute the $\mathbb{H}$ matrix (A.4) derivatives of the constraint functions (A.1) are needed. Since the constraints (A.1) depend only on the $\boldsymbol{a}_{-0}$ part of $\tilde{\boldsymbol{\theta}}$ we have to provide only the derivatives with respect to this parameter sub-vector. The first derivatives are computed by

$$\frac{\partial c_1}{\partial \boldsymbol{a}_{-0}} = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}}\,\boldsymbol{\mu}, \qquad \frac{\partial c_2}{\partial \boldsymbol{a}_{-0}} = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{a}_{-0}}\,\boldsymbol{\mu}^2,$$

where $\boldsymbol{\mu} = (\mu_{-K}, \ldots, \mu_K)'$, $\boldsymbol{\mu}^2 = (\mu^2_{-K}, \ldots, \mu^2_K)'$, and $\partial \boldsymbol{w}/\partial \boldsymbol{a}_{-0}$ is a $2K \times (2K+1)$ matrix given in Section A.3.3.

The second derivatives are given by

$$\frac{\partial^2 c_1}{\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}} = \sum_{j=-K}^{K} \mu_j \frac{\partial^2 w_j}{\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}}, \qquad \frac{\partial^2 c_2}{\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}} = \sum_{j=-K}^{K} \mu_j^2 \frac{\partial^2 w_j}{\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}},$$

where $\partial^2 w_j/\partial \boldsymbol{a}_{-0}\partial \boldsymbol{a}'_{-0}$, $j = -K, \ldots, K$ is a $2K \times 2K$ matrix introduced in Section A.4.6.

A.7 Proof of Proposition 7.1

It is easily seen that the unconstrained minimizer of $\sum_{j=-K^2+3}^{K^2}\{\Delta^3 a_j\}^2$ is not unique and is given by an arbitrary quadratic function of knots, i.e.

$$a_j^K = b_0^K - b_2^K(\mu_j^K - b_1^K)^2, \qquad j = -K^2, \ldots, K^2.$$

Under the constraints (7.9), the minimizer becomes unique with $b_1^K = 0$, $b_0^K = -\log\Big[\sum_{j=-K^2}^{K^2}\exp\{-b_2^K(\mu_j^K)^2\}\Big]$ and b_2^K being a solution to $C_K(b) = 0$, where

$$C_K(b) = \frac{\sum_{j=-K^2}^{K^2}(\mu_j^K)^2\exp\{-b\,(\mu_j^K)^2\}}{\sum_{j=-K^2}^{K^2}\exp\{-b\,(\mu_j^K)^2\}} - (1-\sigma_0^2).$$

The function $C_K(b)$ has the following properties:

- It is continuous on $[0,\,\infty)$;
- For all $b \in [0,\,\infty)$

  $$\frac{d}{db}C_K(b) = \Big[\mathrm{E}\big\{(\mu^K)^2\big|b_2^K = b\big\}\Big]^2 - \mathrm{E}\big\{(\mu^K)^4\big|b_2^K = b\big\},$$

  and from the Hölder's inequality (see, e.g., Billingsley, 1995, p. 80) $\frac{d}{db}C_K(b) < 0$. I.e. $C_K(b)$ is decreasing on $[0,\,\infty)$;
- $C_K(0) = (K^2+1)/3 - (1-\sigma_0^2) > 0$ for all $K \geq 2$;
- $\lim_{b\to\infty} C_K(b) = -(1-\sigma_0^2) < 0$.

So that for all $K \geq 2$ there exists exactly obe root $b_2^K \in (0\,\infty)$ of the equation $C_K(b) = 0$.

Let function $C(b)$ be defined as

$$C(b) = \frac{\int_{-\infty}^{\infty} s^2\,\exp\big(-b\,s^2\big)\,ds}{\int_{-\infty}^{\infty}\exp\big(-b\,s^2\big)\,ds} - (1-\sigma_0^2) = (2b)^{-1} - (1-\sigma_0^2).$$

The equation $C(b) = 0$ has a unique solution $b_2 = \big\{2(1-\sigma_0^2)\big\}^{-1} \in (0.5,\,\infty)$. It follows from the property of the integral that for all $b \in (0,\,\infty)$

$$\lim_{K\to\infty} C_K(b) = C(b)$$

and consequently, using the properties of $C_K(b)$ also

$$\lim_{K\to\infty} b_2^K = b_2.$$

Let $F_K(\mu)$ be a cumulative distribution function of μ^K under b_2^K, i.e.

$$F_K(\mu) = \frac{\sum_{j=-K^2}^{\min(K\mu,\, K^2)} \exp\{-b_2^K\,(\mu_j^K)^2\}}{\sum_{j=-K^2}^{K^2} \exp\{-b_2^K\,(\mu_j^K)^2\}}$$

and $\Phi(\mu\,|\,0,\,1-\sigma_0^2)$ be a cumulative distribution function of the normal distribution $\mathcal{N}(0,\,1-\sigma_0^2)$, i.e.

$$\Phi(\mu\,|\,0,\,1-\sigma_0^2) = \frac{\int_{-\infty}^{\mu} \exp\bigl(-b_2\,s^2\bigr)\,ds}{\int_{-\infty}^{\infty} \exp\bigl(-b_2\,s^2\bigr)\,ds}.$$

It can be now shown that for all $\mu \in \mathbb{R}$

$$\lim_{K\to\infty} F_K(\mu) = \Phi(\mu\,|\,0,\,1-\sigma_0^2),$$

i.e. the random variable μ^K under b_2^K converges in distribution to a $\mathcal{N}(0, 1-\sigma_0^2)$ random variable.
Finally, for all $y \in \mathbb{R}$

$$g_K(y) = \int_{-\infty}^{\infty} \varphi(y\,|\,\mu,\,\sigma_0^2)\,dF_K(\mu)$$

and

$$\varphi(y) = \int_{-\infty}^{\infty} \varphi(y\,|\,\mu,\,\sigma_0^2)\,d\Phi(\mu\,|\,0,\,1-\sigma_0^2).$$

The assertion of the proposition now follows from the fact that function $\varphi(y\,|\,\mu,\,\sigma_0^2)$ is for all $y \in \mathbb{R}$ bounded and continuous function of μ. □

Appendix B

Simulation results

B.1 Simulation for the maximum likelihood penalized AFT model

Here we present selected results of the simulation study introduced in Section 7.5. Tables B.1 – B.6 show the results for the regression parameters. In the first third of the tables, results based on the penalized AFT model are shown. The second third of the tables shows the results based on the parametric AFT model estimated using the maximum-likelihood method assuming a correct (true) error distribution. Finally, the last third of the tables shows the results obtained by the parametric AFT model estimated using the maximum-likelihood method while assuming (in most case incorrectly) normal error distribution.

Figures B.1 – B.3 show the fitted error distributions. For comparison purposes, we plot also the true error distribution.

Table B.1: Results for the regression parameter $\beta_1 = -0.800$ related to the binary covariate. True error distribution: ***normal***. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Assumed Error Distribution					
	Smoothed		True		Normal	
N	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
			light RC			
600	−0.792 (0.118)	138.93	−0.792 (0.114)	130.57	−0.792 (0.114)	130.57
300	−0.812 (0.175)	307.92	−0.812 (0.168)	282.47	−0.812 (0.168)	282.47
100	−0.787 (0.337)	1140.71	−0.778 (0.316)	1005.28	−0.778 (0.316)	1005.28
50	−0.772 (0.478)	2290.06	−0.762 (0.401)	1623.70	−0.762 (0.401)	1623.70
			light R+IC			
600	−0.794 (0.119)	142.59	−0.794 (0.117)	136.20	−0.794 (0.117)	136.20
300	−0.817 (0.176)	311.80	−0.812 (0.172)	295.97	−0.812 (0.172)	295.97
100	−0.775 (0.351)	1235.08	−0.778 (0.323)	1045.28	−0.778 (0.323)	1045.28
50	−0.792 (0.513)	2635.81	−0.769 (0.424)	1806.80	−0.769 (0.424)	1806.80
			heavy RC			
600	−0.780 (0.140)	200.25	−0.782 (0.135)	186.94	−0.782 (0.135)	186.94
300	−0.798 (0.198)	391.34	−0.799 (0.198)	391.21	−0.799 (0.198)	391.21
100	−0.789 (0.491)	2412.45	−0.793 (0.413)	1708.48	−0.793 (0.413)	1708.48
50	−0.629 (0.622)	4156.99	−0.652 (0.490)	2616.76	−0.652 (0.490)	2616.76
			heavy R+IC			
600	−0.787 (0.150)	2280.00	−0.786 (0.141)	201.63	−0.786 (0.141)	201.63
300	−0.811 (0.212)	449.06	−0.800 (0.206)	425.72	−0.800 (0.206)	425.72
100	−0.837 (0.487)	2387.49	−0.799 (0.425)	1809.93	−0.799 (0.425)	1809.93
50	−0.680 (0.717)	5278.77	−0.664 (0.514)	2826.40	−0.664 (0.514)	2826.40

Table B.2: Results for the regression parameter $\beta_1 = -0.800$ related to the binary covariate. True error distribution: ***extreme value***. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Assumed Error Distribution					
	Smoothed		True		Normal	
N	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
			light RC			
600	−0.791 (0.112)	126.39	−0.786 (0.104)	110.56	−0.819 (0.136)	187.24
300	−0.827 (0.151)	235.96	−0.824 (0.151)	233.70	−0.864 (0.188)	393.02
100	−0.796 (0.300)	901.70	−0.782 (0.267)	714.10	−0.842 (0.349)	1234.08
50	−0.888 (0.467)	2257.10	−0.883 (0.323)	1112.5	−0.912 (0.465)	2290.56
			light R+IC			
600	−0.795 (0.109)	118.64	−0.786 (0.104)	109.58	−0.793 (0.123)	152.84
300	−0.826 (0.156)	250.12	−0.824 (0.151)	232.65	−0.833 (0.173)	309.89
100	−0.796 (0.299)	896.48	−0.782 (0.266)	712.12	−0.808 (0.320)	1024.61
50	−0.869 (0.428)	1883.02	−0.884 (0.324)	1117.59	−0.885 (0.430)	1919.72
			heavy RC			
600	−0.788 (0.149)	222.96	−0.785 (0.140)	198.32	−0.869 (0.173)	348.10
300	−0.851 (0.218)	499.81	−0.853 (0.200)	427.42	−0.935 (0.249)	802.15
100	−0.813 (0.459)	2104.51	−0.777 (0.360)	1301.67	−0.877 (0.460)	2176.66
50	−0.891 (0.732)	5437.10	−0.921 (0.546)	3132.58	−0.973 (0.648)	4493.43
			heavy R+IC			
600	−0.800 (0.156)	242.35	−0.786 (0.138)	191.75	−0.819 (0.152)	233.59
300	−0.855 (0.229)	552.22	−0.856 (0.203)	442.33	−0.880 (0.229)	589.85
100	−0.853 (0.469)	2225.78	−0.786 (0.368)	1357.85	−0.833 (0.420)	1778.50
50	−0.872 (0.684)	4725.93	−0.936 (0.563)	3360.86	−0.944 (0.620)	4048.72

Table B.3: Results for the regression parameter $\beta_1 = -0.800$ related to the binary covariate. True error distribution: ***normal mixture***. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Assumed Error Distribution					
	Smoothed		True		Normal	
N	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
			light RC			
600	−0.817 (0.154)	239.76	−0.813 (0.142)	203.08	−0.845 (0.173)	319.10
300	−0.817 (0.201)	408.04	−0.814 (0.187)	350.64	−0.850 (0.262)	713.18
100	−0.829 (0.386)	1498.05	−0.809 (0.319)	1019.39	−0.814 (0.438)	1917.94
50	−0.845 (0.624)	3912.84	−0.819 (0.502)	2526.53	−0.836 (0.628)	3963.00
			light R+IC			
600	−0.824 (0.159)	258.53	−0.819 (0.150)	229.01	−0.877 (0.184)	399.17
300	−0.834 (0.226)	523.20	−0.819 (0.201)	408.76	−0.880 (0.283)	865.88
100	−0.803 (0.411)	1686.57	−0.803 (0.323)	1043.16	−0.833 (0.466)	2184.66
50	−0.871 (0.688)	4781.04	−0.807 (0.567)	3209.78	−0.867 (0.692)	4839.10
			heavy RC			
600	−0.80 (0.213)	451.77	−0.797 (0.187)	349.75	−0.743 (0.194)	407.02
300	−0.752 (0.318)	1036.14	−0.763 (0.285)	827.02	−0.715 (0.310)	1033.05
100	−0.781 (0.558)	3114.28	−0.780 (0.485)	2357.78	−0.716 (0.520)	2771.73
50	−0.723 (0.915)	8426.26	−0.810 (0.746)	5568.92	−0.728 (0.788)	6257.40
			heavy R+IC			
600	−0.826 (0.263)	700.18	−0.808 (0.223)	497.55	−0.821 (0.230)	531.61
300	−0.789 (0.376)	1412.76	−0.759 (0.342)	1189.15	−0.782 (0.366)	1345.87
100	−0.752 (0.640)	4118.05	−0.776 (0.548)	3012.35	−0.779 (0.609)	3711.43
50	−0.846 (1.183)	14012.96	−0.868 (0.969)	9440.90	−0.851 (0.981)	9655.56

Table B.4: Results for the regression parameter $\beta_2 = 0.400$ related to the continuous covariate. True error distribution: ***normal***. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Assumed Error Distribution					
	Smoothed		True		Normal	
N	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
			light RC			
600	0.406 (0.046)	21.58	0.406 (0.046)	21.20	0.406 (0.046)	21.20
300	0.399 (0.064)	41.34	0.397 (0.059)	34.48	0.397 (0.059)	34.48
100	0.380 (0.134)	182.30	0.388 (0.121)	147.19	0.388 (0.121)	147.19
50	0.398 (0.202)	407.62	0.391 (0.176)	311.05	0.391 (0.176)	311.05
			light R+IC			
600	0.407 (0.049)	24.60	0.406 (0.048)	23.10	0.406 (0.048)	23.10
300	0.397 (0.063)	39.78	0.397 (0.062)	38.69	0.397 (0.062)	38.69
100	0.389 (0.133)	178.22	0.391 (0.121)	147.62	0.391 (0.121)	147.62
50	0.402 (0.215)	461.64	0.398 (0.184)	338.59	0.398 (0.184)	338.59
			heavy RC			
600	0.404 (0.051)	26.57	0.405 (0.050)	25.05	0.405 (0.050)	25.05
300	0.398 (0.070)	48.90	0.402 (0.068)	46.37	0.402 (0.068)	46.37
100	0.385 (0.173)	299.81	0.392 (0.140)	197.65	0.392 (0.140)	197.65
50	0.400 (0.264)	697.82	0.407 (0.214)	460.12	0.407 (0.214)	460.12
			heavy R+IC			
600	0.408 (0.056)	31.72	0.406 (0.054)	29.01	0.406 (0.054)	29.01
300	0.403 (0.087)	75.08	0.403 (0.074)	54.82	0.403 (0.074)	54.82
100	0.404 (0.166)	275.28	0.399 (0.142)	200.59	0.399 (0.142)	200.59
50	0.438 (0.314)	997.94	0.424 (0.244)	600.63	0.424 (0.244)	600.63

Table B.5: Results for the regression parameter $\beta_2 = 0.400$ related to the continuous covariate. True error distribution: ***extreme value***. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Assumed Error Distribution					
	Smoothed		True		Normal	
N	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
			light RC			
600	0.402 (0.040)	15.96	0.400 (0.039)	15.33	0.420 (0.048)	27.40
300	0.415 (0.061)	39.21	0.413 (0.057)	33.84	0.432 (0.076)	68.03
100	0.414 (0.101)	104.29	0.408 (0.093)	86.87	0.415 (0.113)	129.37
50	0.428 (0.188)	361.29	0.436 (0.158)	260.77	0.438 (0.186)	359.47
			light R+IC			
600	0.403 (0.041)	17.17	0.400 (0.039)	15.53	0.404 (0.045)	20.06
300	0.416 (0.059)	37.93	0.412 (0.056)	32.98	0.417 (0.067)	48.18
100	0.416 (0.101)	103.67	0.409 (0.093)	87.23	0.410 (0.105)	111.89
50	0.433 (0.182)	343.47	0.436 (0.160)	268.39	0.429 (0.174)	311.89
			heavy RC			
600	0.407 (0.061)	38.05	0.403 (0.058)	34.09	0.453 (0.073)	80.63
300	0.427 (0.086)	82.19	0.420 (0.077)	63.28	0.463 (0.098)	135.13
100	0.389 (0.155)	241.04	0.398 (0.138)	190.95	0.431 (0.155)	248.61
50	0.454 (0.294)	895.33	0.441 (0.229)	540.61	0.464 (0.256)	698.09
			heavy R+IC			
600	0.413 (0.061)	38.47	0.403 (0.059)	34.94	0.426 (0.066)	50.42
300	0.432 (0.084)	80.49	0.420 (0.077)	63.71	0.440 (0.087)	92.29
100	0.419 (0.164)	271.55	0.405 (0.143)	203.32	0.425 (0.151)	234.82
50	0.445 (0.268)	736.90	0.452 (0.241)	607.49	0.461 (0.250)	662.22

Table B.6: Results for the regression parameter $\beta_2 = 0.400$ related to the continuous covariate. True error distribution: ***normal mixture***. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Assumed Error Distribution					
	Smoothed		True		Normal	
N	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
			light RC			
600	0.405 (0.051)	26.07	0.403 (0.050)	24.79	0.412 (0.068)	48.18
300	0.401 (0.075)	56.31	0.400 (0.072)	51.28	0.418 (0.090)	84.56
100	0.386 (0.154)	239.62	0.386 (0.125)	158.23	0.397 (0.176)	311.23
50	0.361 (0.274)	763.56	0.358 (0.250)	640.84	0.369 (0.282)	806.32
			light R+IC			
600	0.408 (0.059)	35.94	0.407 (0.056)	31.74	0.424 (0.076)	62.83
300	0.408 (0.079)	62.88	0.401 (0.071)	50.22	0.432 (0.098)	105.56
100	0.403 (0.183)	336.76	0.397 (0.152)	230.94	0.417 (0.196)	386.41
50	0.376 (0.313)	983.42	0.391 (0.306)	935.42	0.390 (0.316)	997.41
			heavy RC			
600	0.400 (0.078)	60.87	0.396 (0.069)	48.17	0.368 (0.081)	74.92
300	0.392 (0.110)	121.55	0.404 (0.092)	85.67	0.373 (0.100)	106.46
100	0.367 (0.201)	414.59	0.380 (0.172)	301.03	0.363 (0.206)	437.47
50	0.315 (0.409)	1747.62	0.332 (0.347)	1253.25	0.327 (0.331)	1148.82
			heavy R+IC			
600	0.410 (0.091)	84.33	0.402 (0.084)	69.92	0.401 (0.096)	92.04
300	0.418 (0.107)	117.03	0.408 (0.095)	90.05	0.405 (0.113)	128.28
100	0.434 (0.302)	924.86	0.427 (0.249)	628.79	0.418 (0.267)	713.71
50	0.385 (0.479)	2296.79	0.392 (0.441)	1941.29	0.381 (0.429)	1843.27

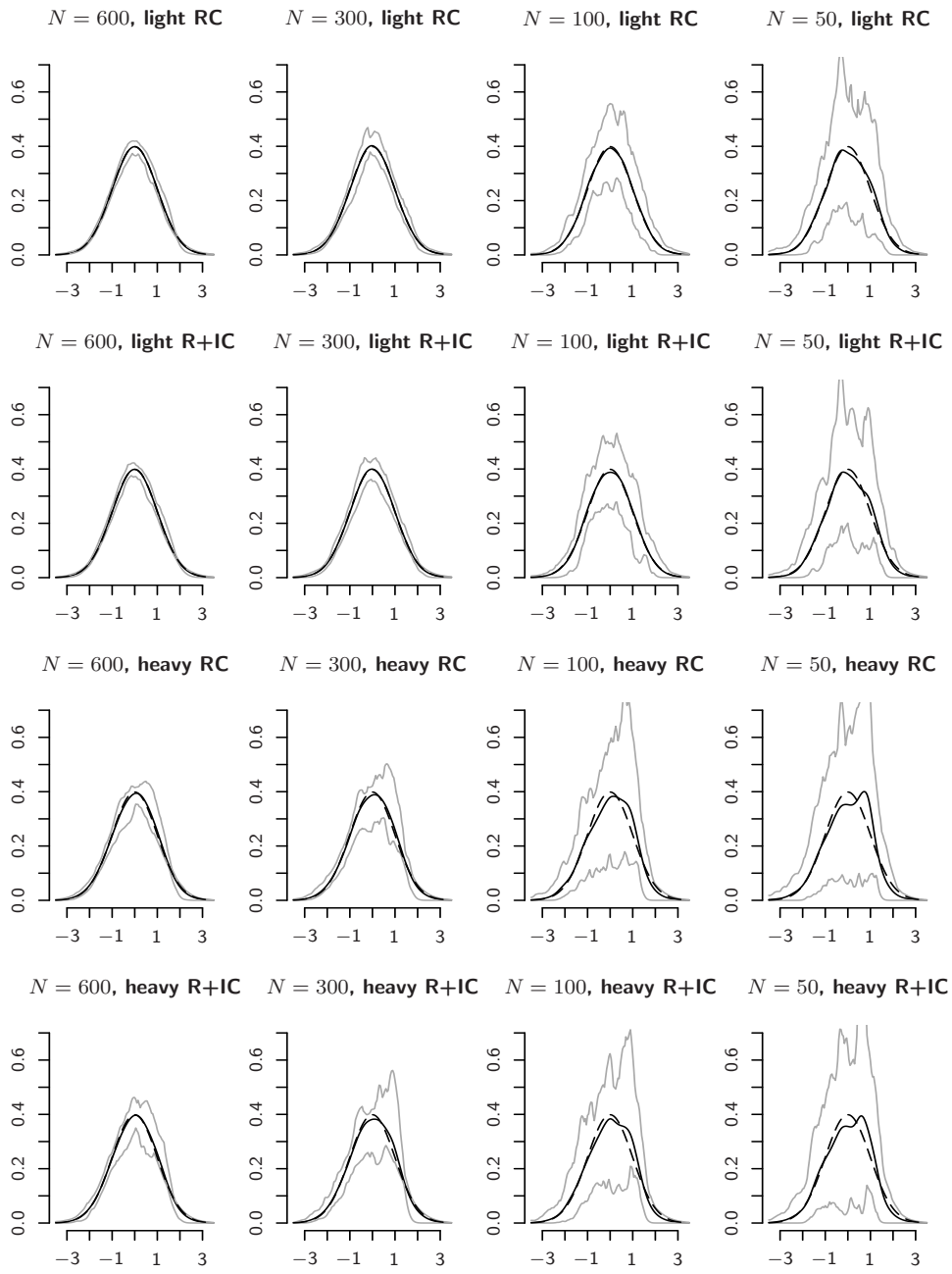


Figure B.1: Results for the standardized error distribution. True error distribution: ***normal***. Solid line: average fitted density, grey lines: 95% pointwise confidence band, dashed line: true error density.

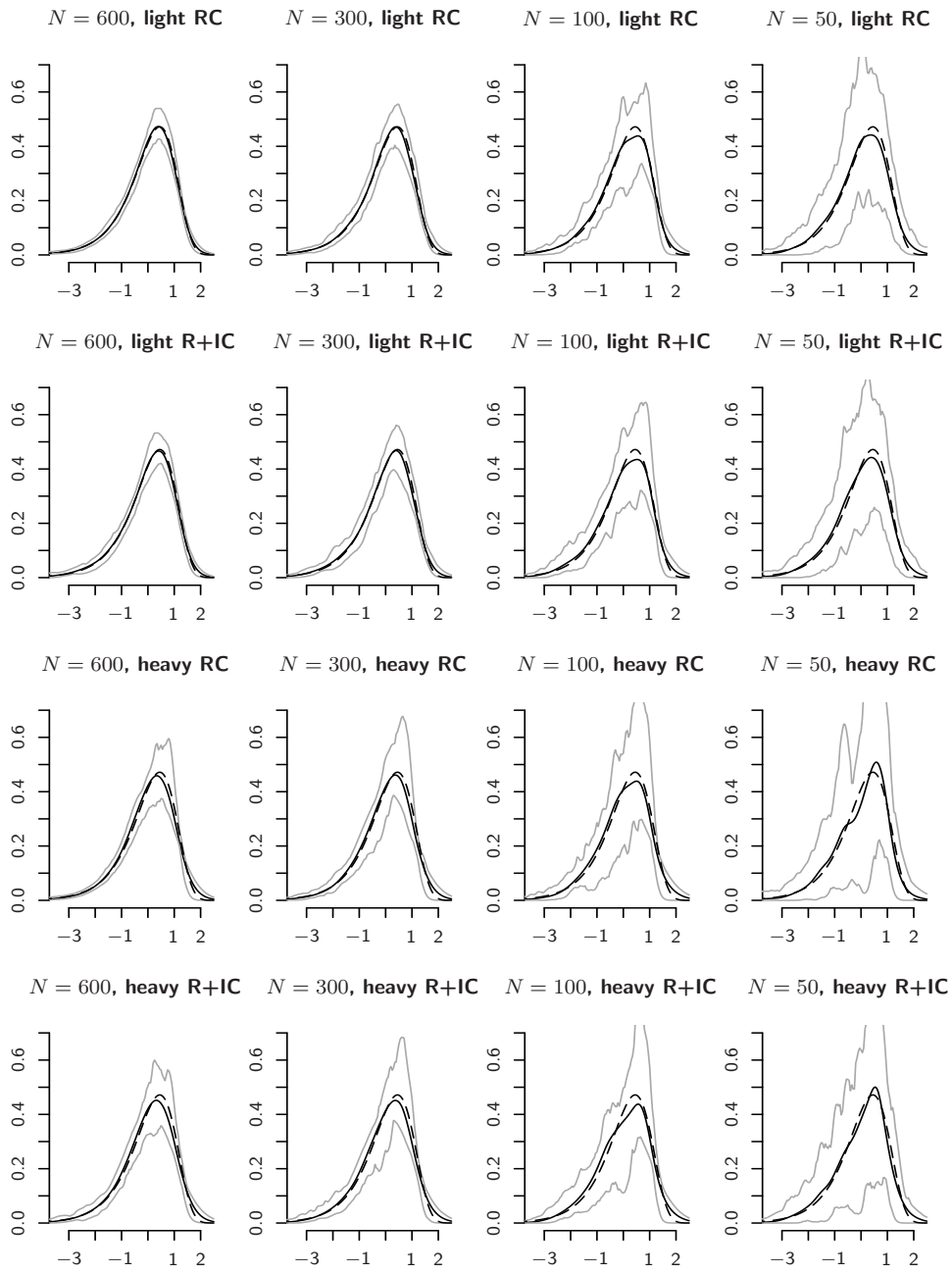


Figure B.2: Results for the standardized error distribution. True error distribution: ***extreme value***. Solid line: average fitted density, grey lines: 95% pointwise confidence band, dashed line: true error density.

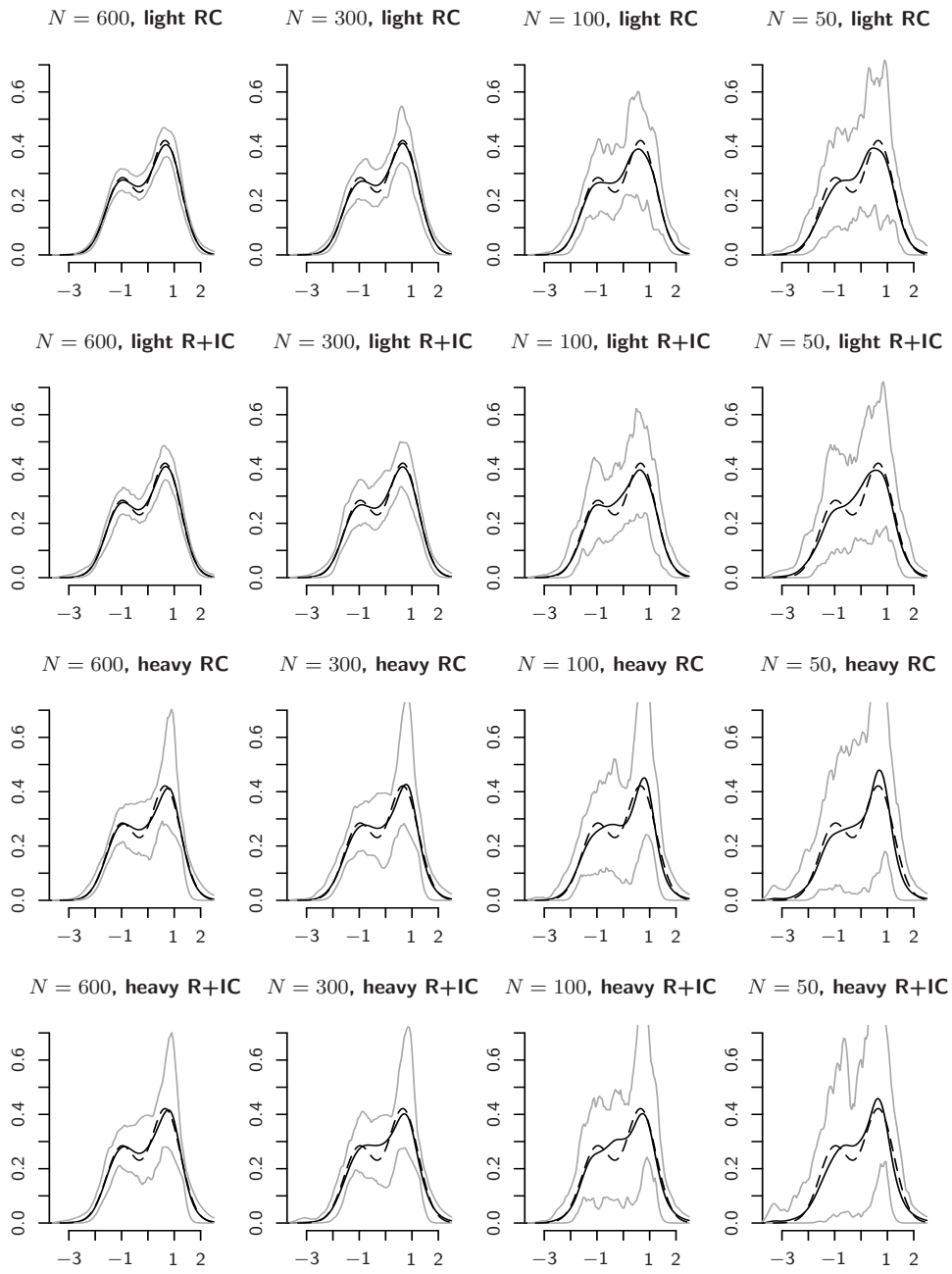


Figure B.3: Results for the standardized error distribution. True error distribution: ***normal mixture***. Solid line: average fitted density, grey lines: 95% pointwise confidence band, dashed line: true error density.

B.2 Simulation for the Bayesian normal mixture cluster-specific AFT model

In this section we give the results of the simulation study introduced in Section 8.6. Tables B.7 and B.8 show the results for the regression parameters. Further, Tables B.9 – B.11 give the results related to the covariance matrix $\mathbb{D}$ of the random effects In the first third (or half) of the tables, results based on the Bayesian normal mixture AFT model are shown. The second third (half) of the tables shows the results based on Bayesian AFT model with assumed normal error distribution and finally the last third of Tables B.7 and B.8 show the results obtained using the parametric AFT model with assumed normal distribution, no random effects included and estimated using the maximum likelihood.

Figures B.4 and B.5 give the fitted standardized (in the case of Cauchy and Student t_2 distribution only centered) error distribution compared to the true density. Figures B.6 and B.7 show the fitted hazard function for a combination of covariates $z_{i,l} = 0$ and $x_{i,l} = 8.13$ (median value). Always a comparison between the Bayesian normal mixture and the Bayesian model with (incorrectly) specified normal error distribution is given. The same comparison, however with respect to the fitted survivor functions is given in Figures B.8 and B.9.

Table B.7: Results for the mean of the covariate random effect $\gamma = -0.800$ related to the binary covariate. Mean, standard deviation and MSE ($\times 10^{-4}$) are clculated over the simulations.

	Estimation method					
	Bayesian mixture		Bayesian normal		ML, no random effects	
N, n_i	$\hat{\gamma}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\gamma}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\gamma}$ (SD)	MSE ($\times 10^{-4}$)
True error = ***normal***						
100, 10	−0.798 (0.069)	47.01	−0.798 (0.069)	48.17	−0.798 (0.078)	60.45
50, 5	−0.813 (0.155)	240.43	−0.811 (0.149)	222.78	−0.812 (0.153)	235.67
True error = ***Cauchy*** t_1						
100, 10	−0.8100 (0.103)	107.16	−0.736 (0.139)	234.52	−0.738 (0.142)	238.91
50, 5	−0.766 (0.224)	512.72	−0.719 (0.255)	716.39	−0.721 (0.253)	703.97
True error = ***Student*** t_2						
100, 10	−0.793 (0.100)	99.91	−0.761 (0.104)	123.72	−0.7600 (0.108)	132.14
50, 5	−0.778 (0.218)	479.02	−0.759 (0.196)	401.28	−0.761 (0.200)	415.46
True error = ***extreme value***						
100, 10	−0.797 (0.069)	47.39	−0.80 (0.075)	56.62	−0.802 (0.082)	66.90
50, 5	−0.815 (0.137)	191.09	−0.811 (0.138)	192.4	−0.809 (0.142)	202.26
True error = ***normal mixture***						
100, 10	−0.804 (0.051)	26.60	−0.926 (0.144)	366.82	−0.923 (0.148)	369.41
50, 5	−0.787 (0.097)	95.70	−0.869 (0.291)	894.99	−0.863 (0.283)	840.46

Table B.8: Results for the regression parameter $\beta = 0.400$ related to the continuous covariate. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Estimation method					
	Bayesian mixture		Bayesian normal		ML, no random effects	
N, n_i	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}$ (SD)	MSE ($\times 10^{-4}$)
True error = ***normal***						
100, 10	0.402 (0.027)	7.28	0.402 (0.027)	7.28	0.402 (0.030)	9.01
50, 5	0.397 (0.051)	26.31	0.397 (0.051)	25.6	0.399 (0.059)	34.67
True error = ***Cauchy*** t_1						
100, 10	0.392 (0.036)	13.23	0.361 (0.051)	41.84	0.357 (0.057)	50.73
50, 5	0.412 (0.071)	52.54	0.383 (0.081)	68.17	0.378 (0.109)	124.59
True error = ***Student*** t_2						
100, 10	0.394 (0.033)	11.51	0.379 (0.038)	19.02	0.378 (0.041)	21.62
50, 5	0.393 (0.076)	58.15	0.386 (0.069)	49.62	0.382 (0.084)	72.93
True error = ***extreme value***						
100, 10	0.404 (0.021)	4.34	0.403 (0.023)	5.41	0.402 (0.026)	6.76
50, 5	0.393 (0.042)	17.92	0.395 (0.045)	20.89	0.395 (0.051)	25.87
True error = ***normal mixture***						
100, 10	0.400 (0.019)	3.46	0.448 (0.048)	45.62	0.450 (0.052)	52.93
50, 5	0.394 (0.042)	17.65	0.432 (0.076)	68.1	0.444 (0.104)	127.26

Table B.9: Results for the standard deviation of the random intercept $\mathrm{sd}(b_{i,1}) = 0.500$. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Estimation method			
	Bayesian mixture		Bayesian normal	
N, n_i	$\widehat{\mathrm{sd}}(b_{i,1})$ (SD)	MSE	$\widehat{\mathrm{sd}}(b_{i,1})$ (SD)	MSE
True error = ***normal***				
100, 10	0.476 (0.069)	52.98	0.476 (0.068)	52.52
50, 5	0.321 (0.154)	559.95	0.324 (0.156)	551.48
True error = ***Cauchy*** t_1				
100, 10	0.381 (0.120)	284.36	0.188 (0.121)	1117.45
50, 5	0.118 (0.060)	1492.37	0.086 (0.013)	1718.19
True error = ***Student*** t_2				
100, 10	0.452 (0.094)	111.52	0.418 (0.106)	179.32
50, 5	0.160 (0.128)	1321.32	0.125 (0.086)	1480.14
True error = ***extreme value***				
100, 10	0.489 (0.061)	38.41	0.495 (0.069)	48.37
50, 5	0.343 (0.144)	453.41	0.305 (0.156)	625.15
True error = ***normal mixture***				
100, 10	0.493 (0.047)	22.33	0.428 (0.176)	360.99
50, 5	0.446 (0.093)	115.32	0.105 (0.048)	1583.26

Table B.10: Results for the standard deviation of the covariate random effect $\mathrm{sd}(b_{i,2}) = 0.100$. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Estimation method			
	Bayesian mixture		Bayesian normal	
N, n_i	$\widehat{\mathrm{sd}}(b_{i,2})$ (SD)	MSE	$\widehat{\mathrm{sd}}(b_{i,2})$ (SD)	MSE
	True error = ***normal***			
100, 10	0.125 (0.040)	22.13	0.125 (0.040)	22.30
50, 5	0.152 (0.059)	61.67	0.153 (0.059)	62.53
	True error = ***Cauchy*** t_1			
100, 10	0.156 (0.058)	64.95	0.124 (0.054)	34.43
50, 5	0.093 (0.017)	3.36	0.083 (0.008)	3.60
	True error = ***Student*** t_2			
100, 10	0.135 (0.031)	22.03	0.142 (0.033)	28.49
50, 5	0.105 (0.039)	15.61	0.097 (0.029)	8.30
	True error = ***extreme value***			
100, 10	0.109 (0.027)	8.22	0.112 (0.028)	9.16
50, 5	0.151 (0.059)	60.72	0.142 (0.052)	44.71
	True error = ***normal mixture***			
100, 10	0.094 (0.029)	8.51	0.174 (0.062)	93.20
50, 5	0.139 (0.043)	33.42	0.090 (0.020)	5.24

Table B.11: Results for the random effects correlation $\text{corr}(b_{i,1}, , b_{i,2}) = 0.400$. Mean, standard deviation and MSE ($\times 10^{-4}$) are calculated over the simulations.

	Estimation method			
	Bayesian mixture		Bayesian normal	
N, n_i	$\widehat{\text{corr}}(b_{i,1}, b_{i,2})$ (SD)	MSE	$\widehat{\text{corr}}(b_{i,1}, b_{i,2})$ (SD)	MSE
	True error = ***normal***			
100, 10	0.391 (0.457)	2086.96	0.395 (0.459)	2108.60
50, 5	0.293 (0.343)	1292.59	0.290 (0.343)	1299.25
	True error = ***Cauchy*** t_1			
100, 10	0.380 (0.372)	1385.34	0.210 (0.226)	873.41
50, 5	0.061 (0.090)	1232.64	0.014 (0.032)	1502.97
	True error = ***Student*** t_2			
100, 10	0.266 (0.434)	2066.67	0.240 (0.423)	2045.53
50, 5	0.100 (0.197)	1284.85	0.070 (0.118)	1230.40
	True error = ***extreme value***			
100, 10	0.388 (0.428)	1831.69	0.244 (0.469)	2442.23
50, 5	0.327 (0.409)	1722.62	0.249 (0.352)	1466.79
	True error = ***normal mixture***			
100, 10	0.376 (0.401)	1617.48	0.228 (0.415)	2019.53
50, 5	0.307 (0.433)	1958.50	0.027 (0.051)	1413.30

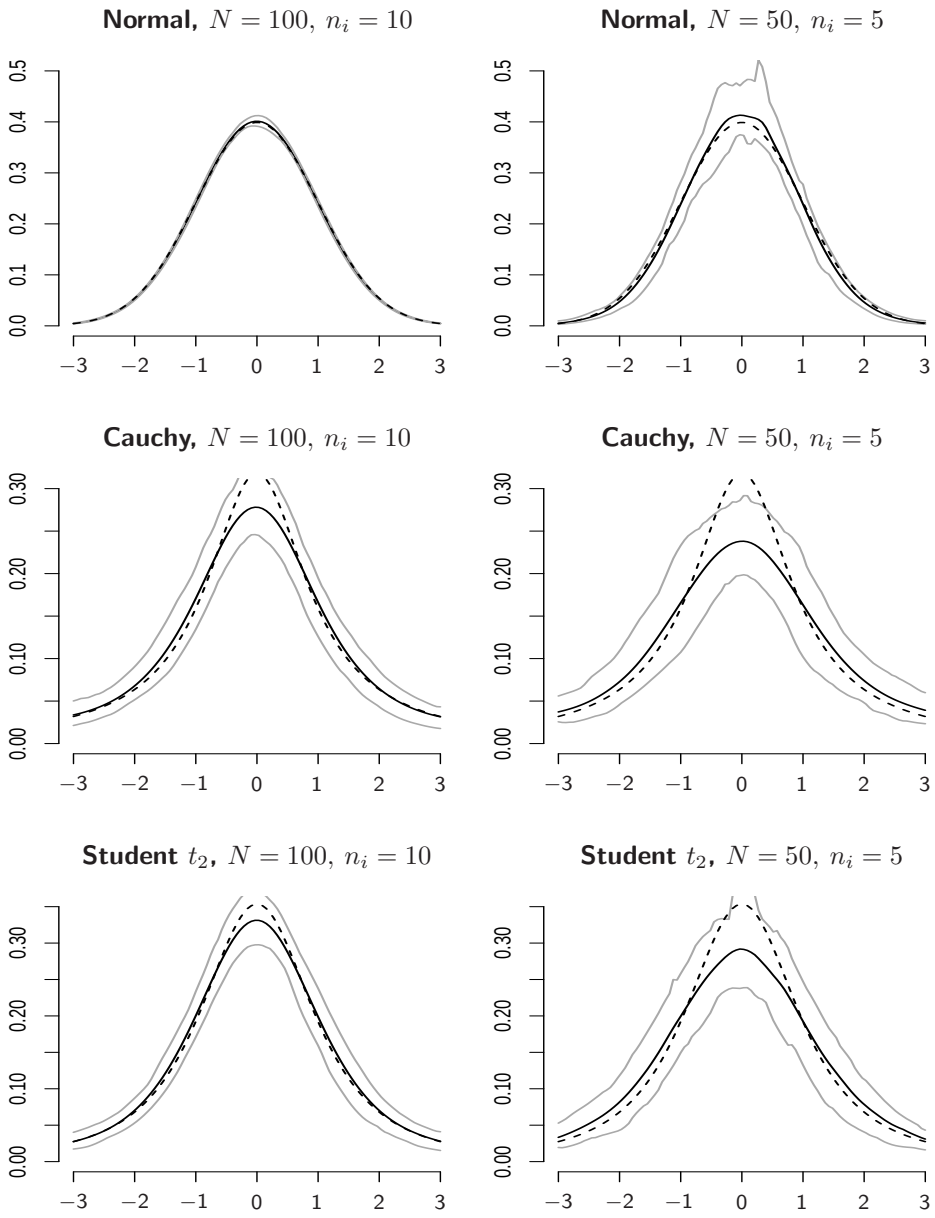


Figure B.4: Results for the standardized error density, estimated using the Bayesian mixture model. Solid line: average fitted standardized density, grey lines: 95% pointwise confidence band, dashed line: true standardized error density.

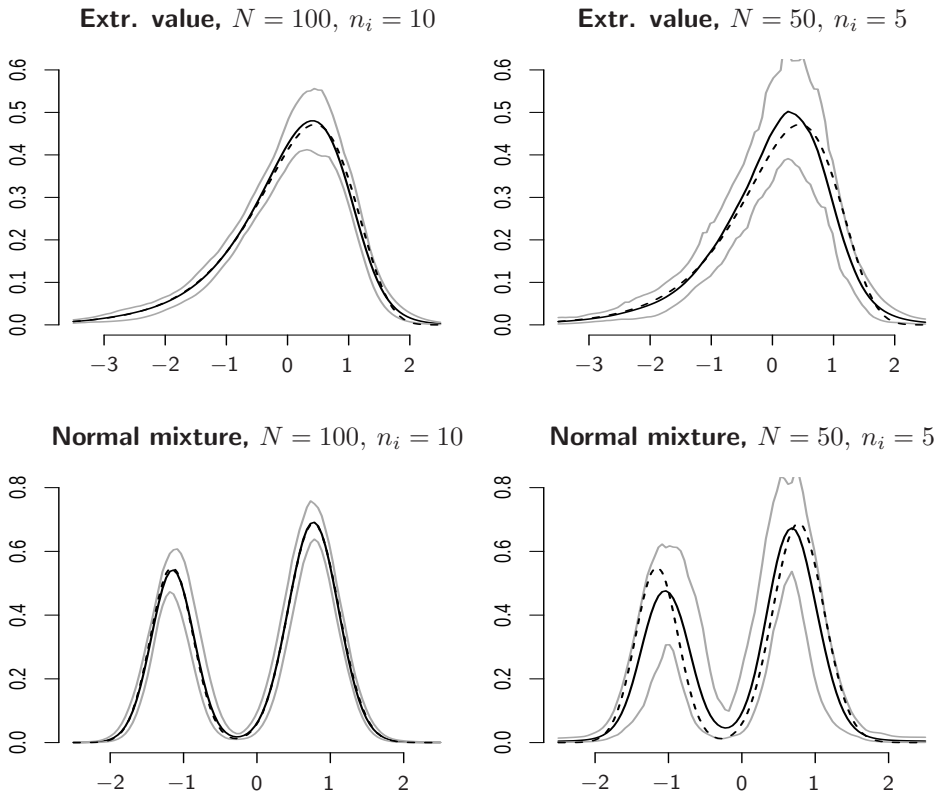


Figure B.5: Results for the standardized error density, estimated using the Bayesian mixture model. Solid line: average fitted standardized density, grey lines: 95% pointwise confidence band, dashed line: true standardized error density.

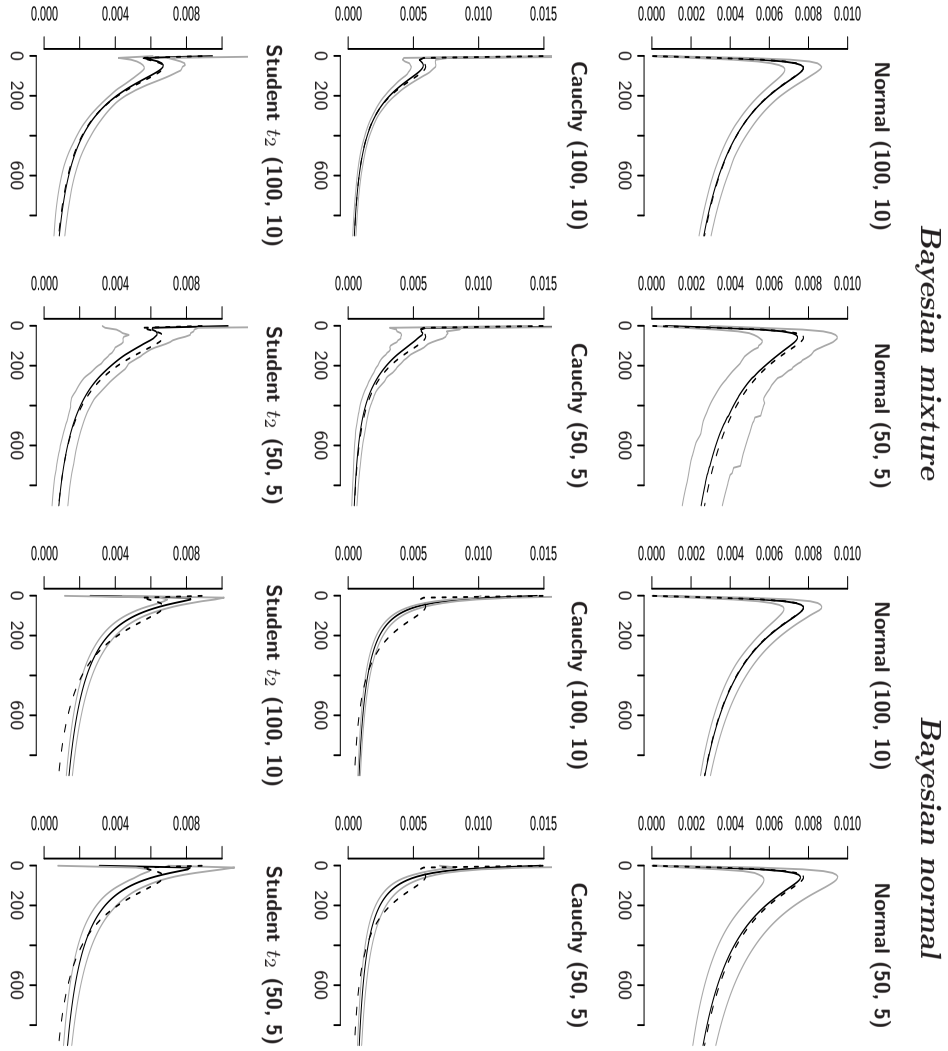


Figure B.6: Results for the hazard function, estimated using the Bayesian mixture model (left part) and the Bayesian normal model (right part). Each row shows the results for different true error densities. Solid line: average fitted hazard, grey lines: 95% pointwise confidence band, dashed line: true hazard function.

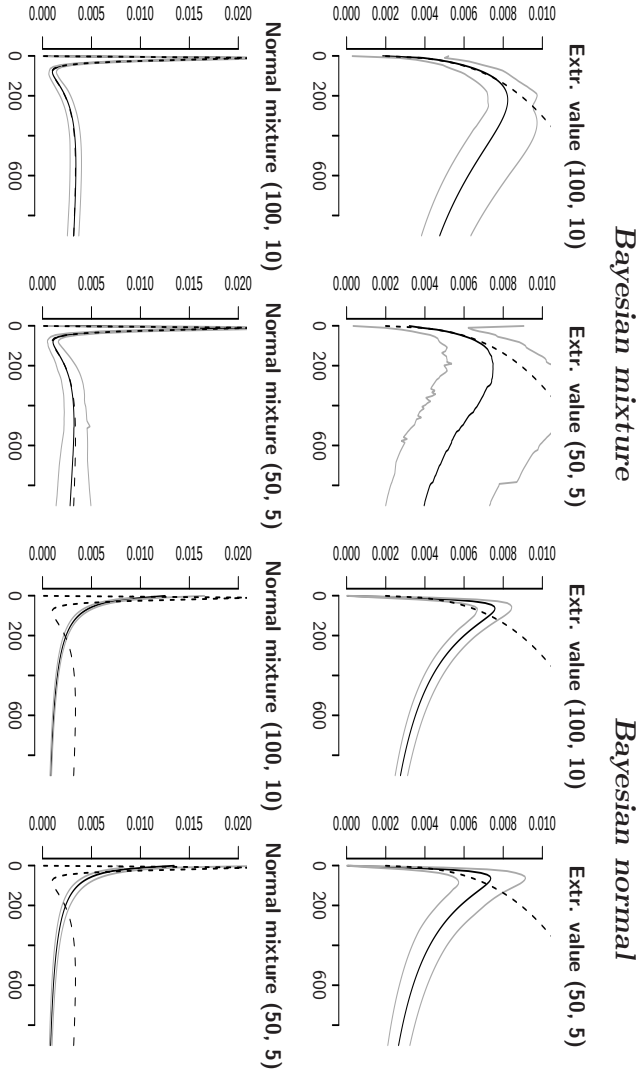


Figure B.7: Results for the hazard function, estimated using the Bayesian mixture model (left part) and the Bayesian normal model (right part). Each row shows the results for different true error densities. Solid line: average fitted hazard, grey lines: 95% pointwise confidence band, dashed line: true hazard function.

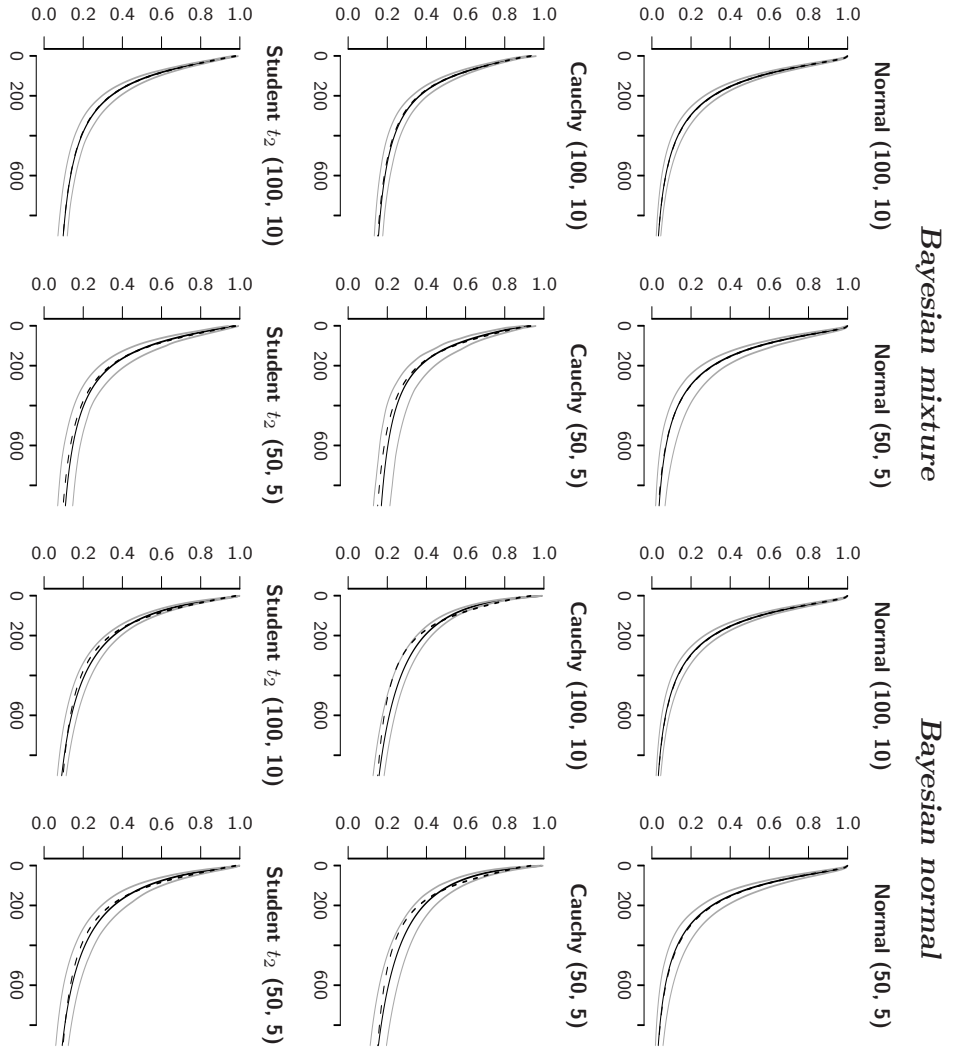


Figure B.8: Results for the survivor function, estimated using the Bayesian mixture model (left part) and the Bayesian normal model (right part). Each row shows the results for different true error densities. Solid line: average fitted survivor function, grey lines: 95% pointwise confidence band, dashed line: true survivor function.

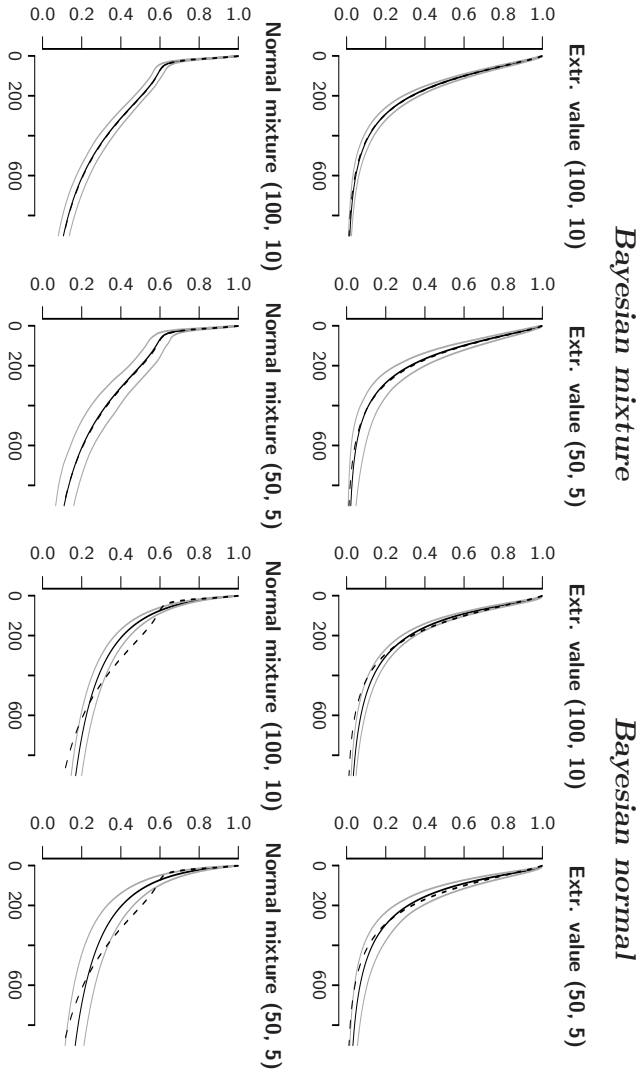


Figure B.9: Results for the survivor function, estimated using the Bayesian mixture model (left part) and the Bayesian normal model (right part). Each row shows the results for different true error densities. Solid line: average fitted survivor function, grey lines: 95% pointwise confidence band, dashed line: true survivor function.

B.3 Simulation for the Bayesian penalized mixture cluster-specific AFT model

This section presents selected results of the simulation study introduced in Section 9.6. Tables B.12 and B.13 show the results for the regression parameters. Tables B.14 and B.15 give the results for the variance components of the model.

Figures B.10 and B.11 show the fitted survivor densities for the onset part of the model for a combination of covariates $x^u_{i,l,1} = 0.5$ (median value) and $x^u_{i,l,2} = 1$. Figures B.12 and B.13 give the fitted survivor densities for the event part of the model for a combination of covariates $x^t_{i,l,1} = 0.5$ (median value) and $x^t_{i,l,2} = 1$. Corresponding fitted survivor functions are given in Figures B.14 – B.17.

Table B.12: Results for the regression parameters from the onset part of the model. Mean, standard deviation and MSE ($\times 10^{-4}$) over the simulation.

	$\delta_1 = 0.200$		$\delta_2 = -0.100$	
$\tau^d/\tau^\zeta =$		MSE		MSE
τ^b/τ^ε	$\hat{\delta}_1$ (SD)	($\times 10^{-4}$)	$\hat{\delta}_2$ (SD)	($\times 10^{-4}$)
Scenario I				
(error $\sim$ normal mixture, random effect $\sim$ extreme value)				
5	0.199 (0.007)	0.56	−0.101 (0.004)	0.17
3	0.201 (0.008)	0.68	−0.100 (0.005)	0.20
2	0.198 (0.011)	1.30	−0.100 (0.006)	0.37
1	0.199 (0.014)	1.84	−0.100 (0.009)	0.76
1/2	0.200 (0.018)	3.14	−0.100 (0.010)	0.92
1/3	0.201 (0.019)	3.74	−0.101 (0.010)	1.02
1/5	0.198 (0.019)	3.51	−0.100 (0.010)	0.95
Scenario II				
(error $\sim$ extreme value, random effect $\sim$ normal mixture)				
5	0.200 (0.010)	0.93	−0.101 (0.005)	0.30
3	0.202 (0.015)	2.38	−0.101 (0.008)	0.72
2	0.200 (0.019)	3.44	−0.099 (0.011)	1.27
1	0.196 (0.029)	8.73	−0.099 (0.019)	3.45
1/2	0.194 (0.038)	14.46	−0.097 (0.025)	6.30
1/3	0.201 (0.041)	16.73	−0.099 (0.024)	5.90
1/5	0.203 (0.043)	18.12	−0.100 (0.020)	4.12

Table B.13: Results for the regression parameters from the event part of the model. Mean, standard deviation and MSE ($\times 10^{-4}$) over the simulation.

	$\beta_1 = 0.300$		$\beta_2 = -0.150$	
$\tau^d/\tau^\zeta =$ τ^b/τ^ε	$\hat{\beta}_1$ (SD)	MSE ($\times 10^{-4}$)	$\hat{\beta}_2$ (SD)	MSE ($\times 10^{-4}$)
Scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value)				
5	0.302 (0.014)	2.12	−0.149 (0.008)	0.64
3	0.301 (0.032)	9.99	−0.149 (0.021)	4.47
2	0.298 (0.056)	30.55	−0.150 (0.034)	11.75
1	0.304 (0.054)	29.04	−0.148 (0.028)	7.55
1/2	0.301 (0.043)	18.07	−0.147 (0.031)	9.66
1/3	0.311 (0.058)	34.67	−0.150 (0.035)	11.88
1/5	0.299 (0.050)	25.11	−0.151 (0.031)	9.68
Scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture)				
5	0.298 (0.031)	9.40	−0.148 (0.016)	2.74
3	0.291 (0.040)	16.44	−0.152 (0.022)	4.99
2	0.306 (0.065)	42.02	−0.146 (0.036)	13.01
1	0.299 (0.103)	105.54	−0.149 (0.057)	32.60
1/2	0.304 (0.121)	144.59	−0.151 (0.070)	48.40
1/3	0.296 (0.126)	157.36	−0.146 (0.071)	50.06
1/5	0.308 (0.112)	125.51	−0.142 (0.065)	42.10

Table B.14: Results for the scale parameters from the onset part of the model. Mean, standard deviation and MSE ($\times 10^{-4}$) over the simulation.

	τ^d			τ^ζ		
$\tau^d/\tau^\zeta =$ τ^b/τ^ε	True τ^d	$\hat{\tau}^d$ (SD)	MSE ($\times 10^{-4}$)	True τ^ζ	$\hat{\tau}^\zeta$ (SD)	MSE ($\times 10^{-4}$)
Scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value)						
5	0.310	0.341 (0.035)	21.20	0.062	0.062 (0.002)	0.04
3	0.300	0.324 (0.037)	19.13	0.100	0.100 (0.002)	0.04
2	0.283	0.283 (0.031)	9.31	0.141	0.141 (0.003)	0.08
1	0.224	0.219 (0.024)	5.85	0.224	0.223 (0.006)	0.39
1/2	0.141	0.143 (0.018)	3.24	0.283	0.283 (0.006)	0.31
1/3	0.100	0.103 (0.035)	12.23	0.300	0.301 (0.012)	1.35
1/5	0.062	0.110 (0.097)	116.93	0.310	0.325 (0.034)	13.74
Scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture)						
5	0.310	0.311 (0.009)	0.86	0.062	0.061 (0.003)	0.11
3	0.300	0.318 (0.112)	128.44	0.100	0.116 (0.099)	99.25
2	0.283	0.299 (0.141)	198.36	0.141	0.159 (0.126)	159.21
1	0.224	0.218 (0.011)	1.54	0.224	0.224 (0.012)	1.35
1/2	0.141	0.132 (0.021)	5.23	0.283	0.285 (0.013)	1.70
1/3	0.100	0.065 (0.037)	25.76	0.300	0.304 (0.013)	1.83
1/5	0.062	0.040 (0.030)	13.87	0.310	0.314 (0.015)	2.34

Table B.15: Results for the scale parameters from the event part of the model. Mean, standard deviation and MSE ($\times 10^{-4}$) over the simulation.

	τ^b			τ^ε		
$\tau^d/\tau^\zeta =$ τ^b/τ^ε	True τ^b	$\hat{\tau}^b$ (SD)	MSE ($\times 10^{-4}$)	True τ^ε	$\hat{\tau}^\varepsilon$ (SD)	MSE ($\times 10^{-4}$)
Scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value)						
5	0.981	0.980 (0.393)	1532.34	0.196	0.202 (0.005)	0.60
3	0.949	0.987 (0.517)	2660.34	0.316	0.417 (0.160)	356.10
2	0.894	0.827 (0.065)	87.10	0.447	0.663 (0.217)	932.57
1	0.707	0.647 (0.046)	57.34	0.707	0.741 (0.090)	91.97
1/2	0.447	0.428 (0.039)	18.27	0.894	0.901 (0.017)	3.47
1/3	0.316	0.307 (0.037)	14.42	0.949	0.954 (0.018)	3.57
1/5	0.196	0.180 (0.037)	15.93	0.981	0.984 (0.018)	3.49
Scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture)						
5	0.981	0.971 (0.030)	9.67	0.196	0.202 (0.012)	1.73
3	0.949	0.941 (0.040)	16.72	0.316	0.325 (0.054)	29.79
2	0.894	0.884 (0.049)	24.54	0.447	0.532 (0.237)	626.75
1	0.707	0.671 (0.040)	28.78	0.707	0.886 (0.273)	1056.64
1/2	0.447	0.394 (0.092)	111.54	0.894	1.160 (0.230)	1228.67
1/3	0.316	0.079 (0.115)	695.18	0.949	1.286 (0.214)	1589.69
1/5	0.196	0.024 (0.026)	302.36	0.981	1.345 (0.235)	1873.27

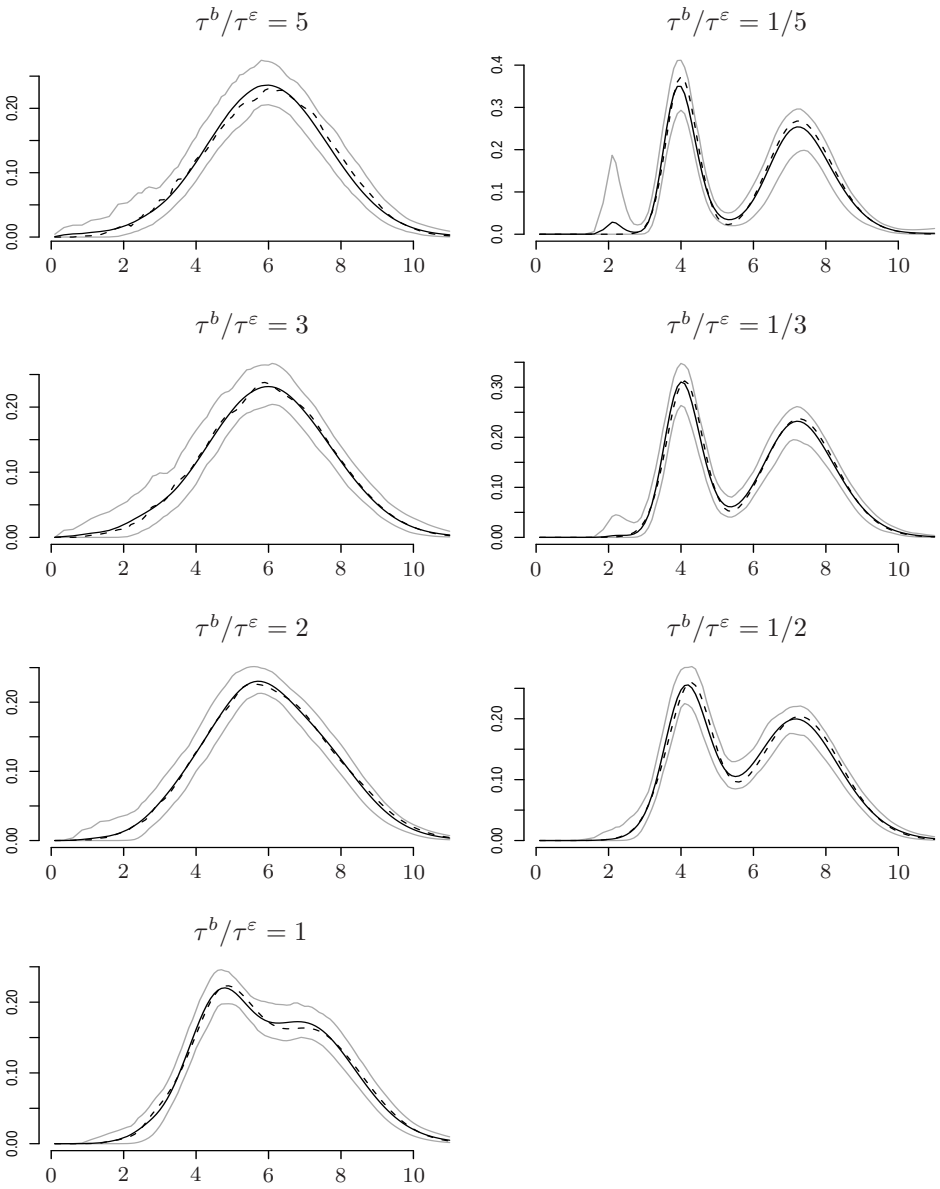


Figure B.10: Results for the survivor density of the onset time, for the combination of covariates $\boldsymbol{x}^u_{i,l} = (0.5, 1)'$, scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value). Solid line: average fitted survivor density, grey lines: 95% pointwise confidence band, dashed line: true survivor density..

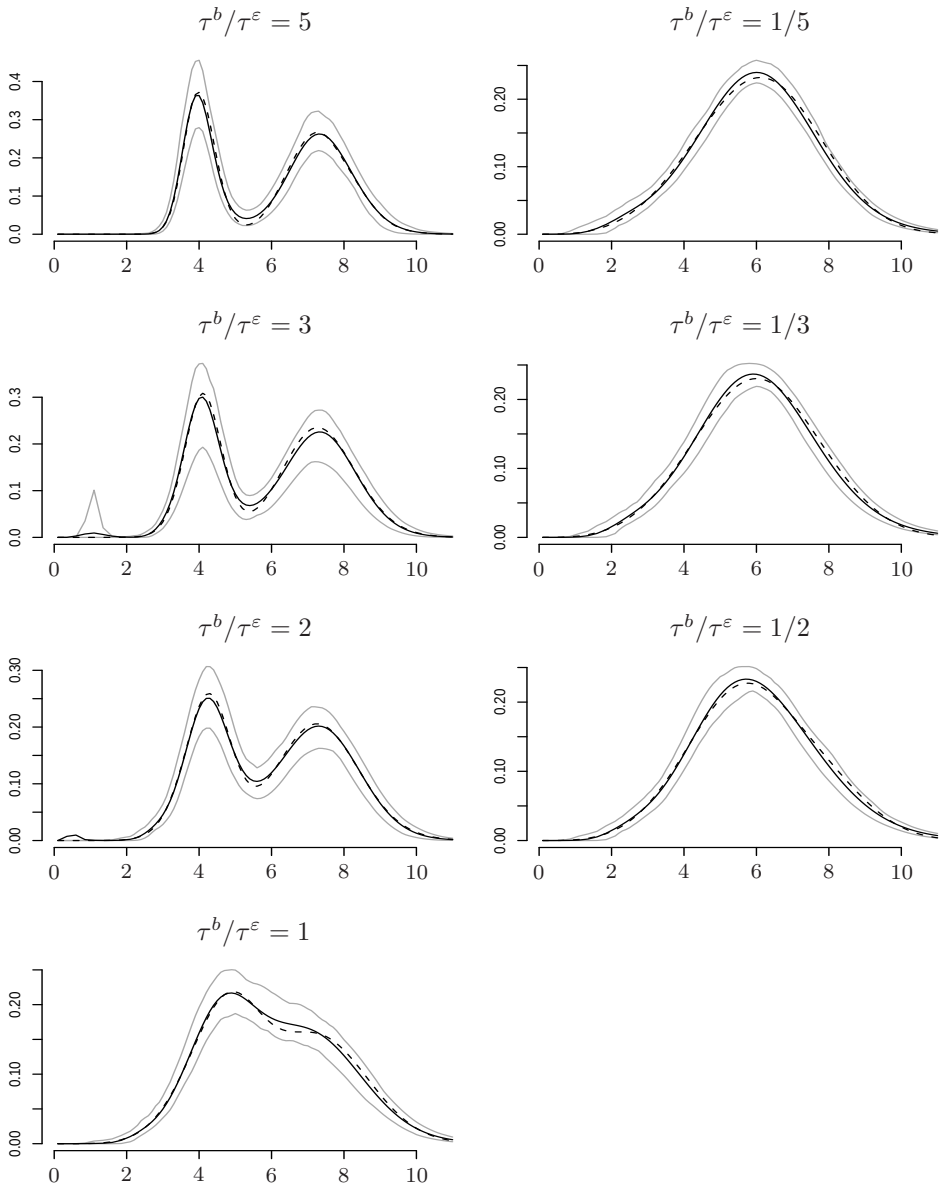


Figure B.11: Results for the survivor density of the onset time, for the combination of covariates $\boldsymbol{x}^u_{i,l} = (0.5, 1)'$, scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture). Solid line: average fitted survivor density, grey lines: 95% pointwise confidence band, dashed line: true survivor density..

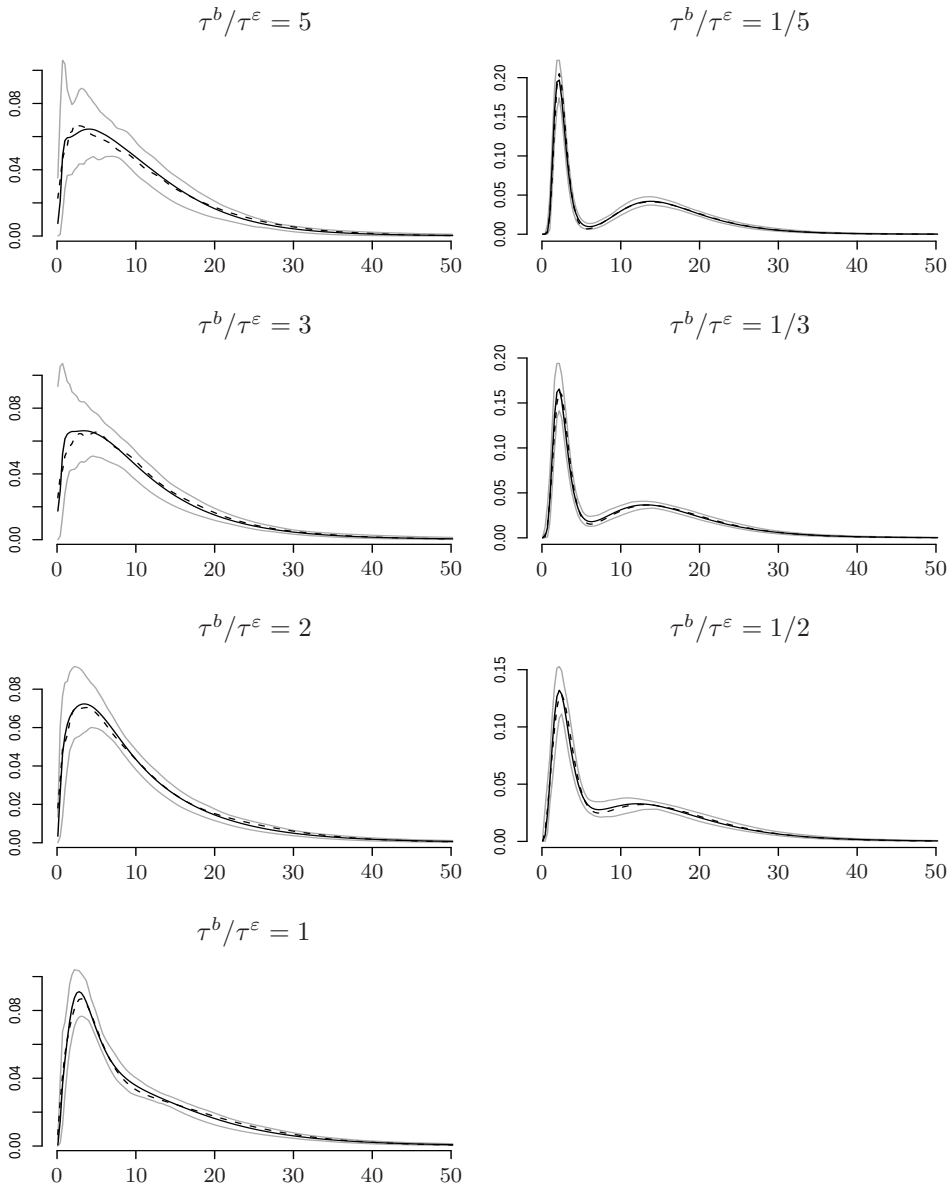


Figure B.12: Results for the survivor density of the event time, for the combination of covariates $\boldsymbol{x}_{i,l}^t = (0.5, 1)'$, scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value). Solid line: average fitted survivor density, grey lines: 95% pointwise confidence band, dashed line: true survivor density..

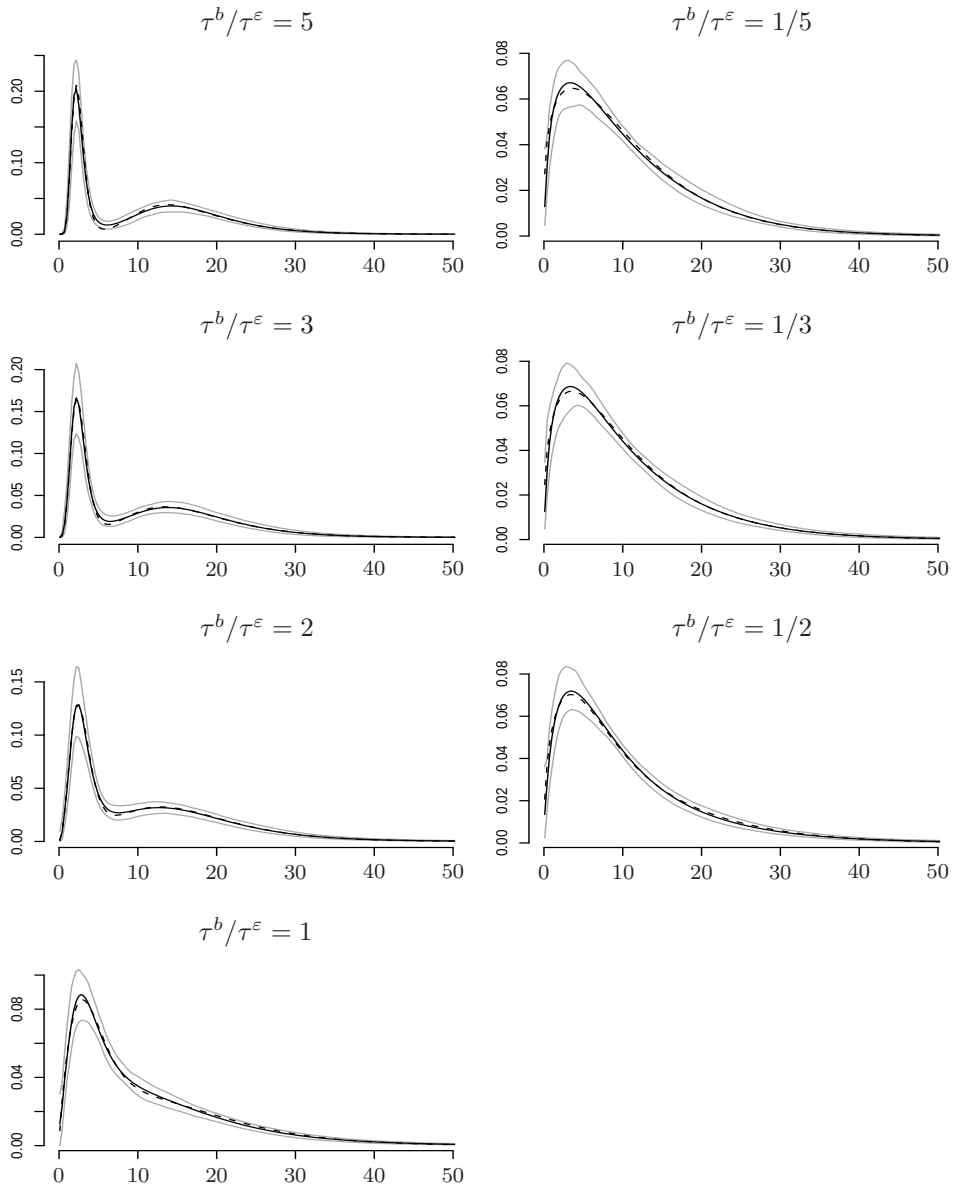


Figure B.13: Results for the survivor density of the event time, for the combination of covariates $\boldsymbol{x}_{i,l}^t = (0.5, 1)'$, scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture). Solid line: average fitted survivor density, grey lines: 95% pointwise confidence band, dashed line: true survivor density..

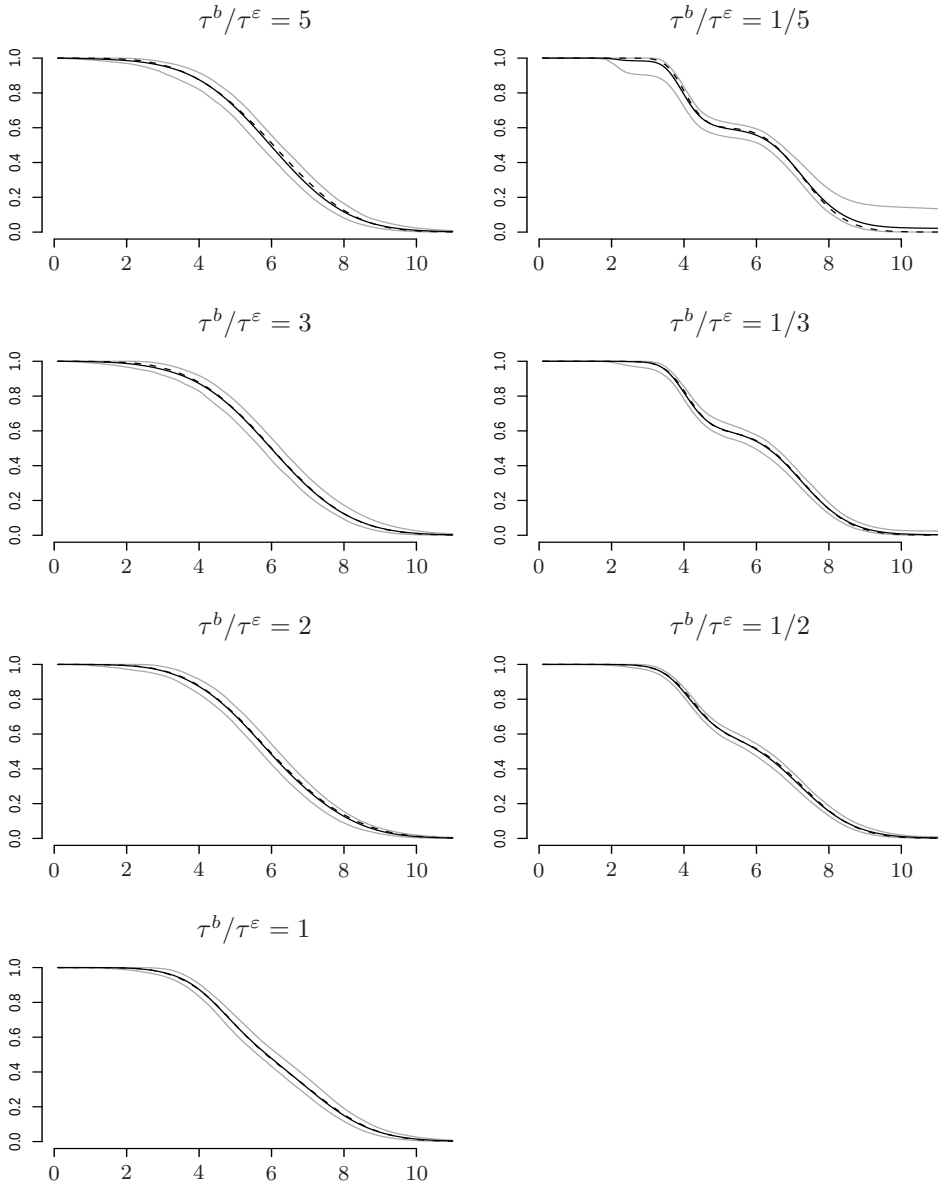


Figure B.14: Results for the survivor function of the onset time, for the combination of covariates $\boldsymbol{x}^u_{i,l} = (0.5, 1)'$, scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value). Solid line: average fitted survivor function, grey lines: 95% pointwise confidence band, dashed line: true survivor function..

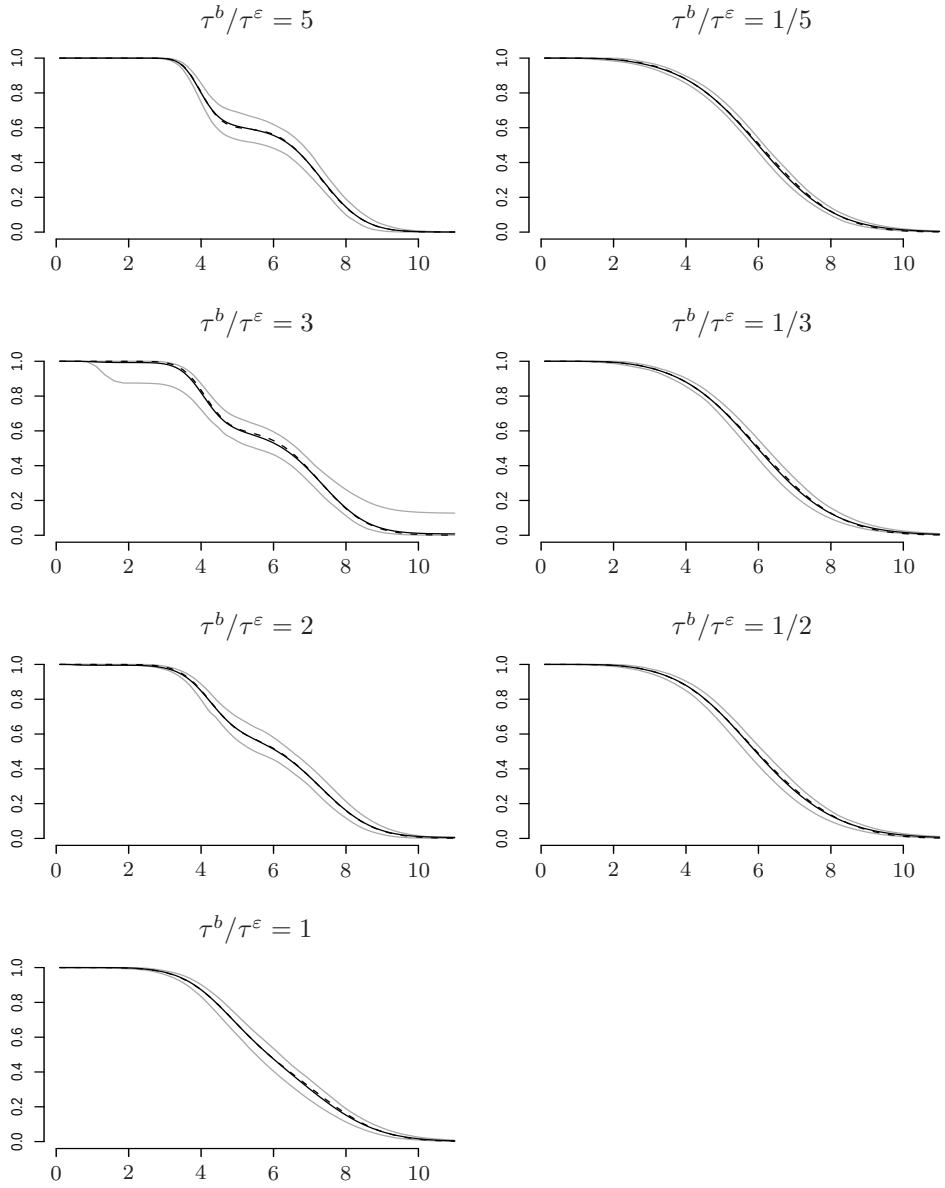


Figure B.15: Results for the survivor function of the onset time, for the combination of covariates $\boldsymbol{x}^u_{i,l} = (0.5, 1)'$, scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture). Solid line: average fitted survivor function, grey lines: 95% pointwise confidence band, dashed line: true survivor function..

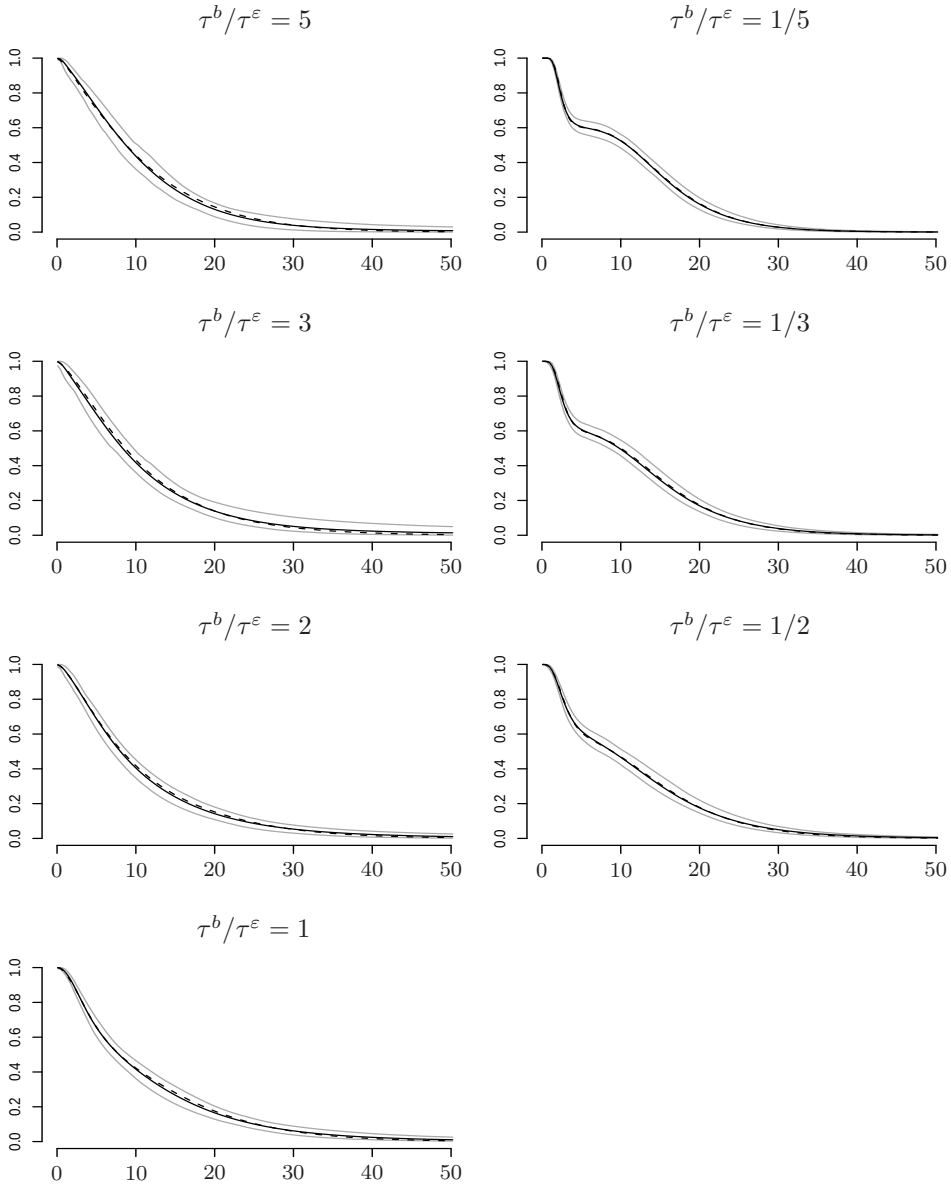


Figure B.16: Results for the survivor function of the event time, for the combination of covariates $\boldsymbol{x}_{i,l}^t = (0.5, 1)'$, scenario I (error $\sim$ normal mixture, random effect $\sim$ extreme value). Solid line: average fitted survivor function, grey lines: 95% pointwise confidence band, dashed line: true survivor function..

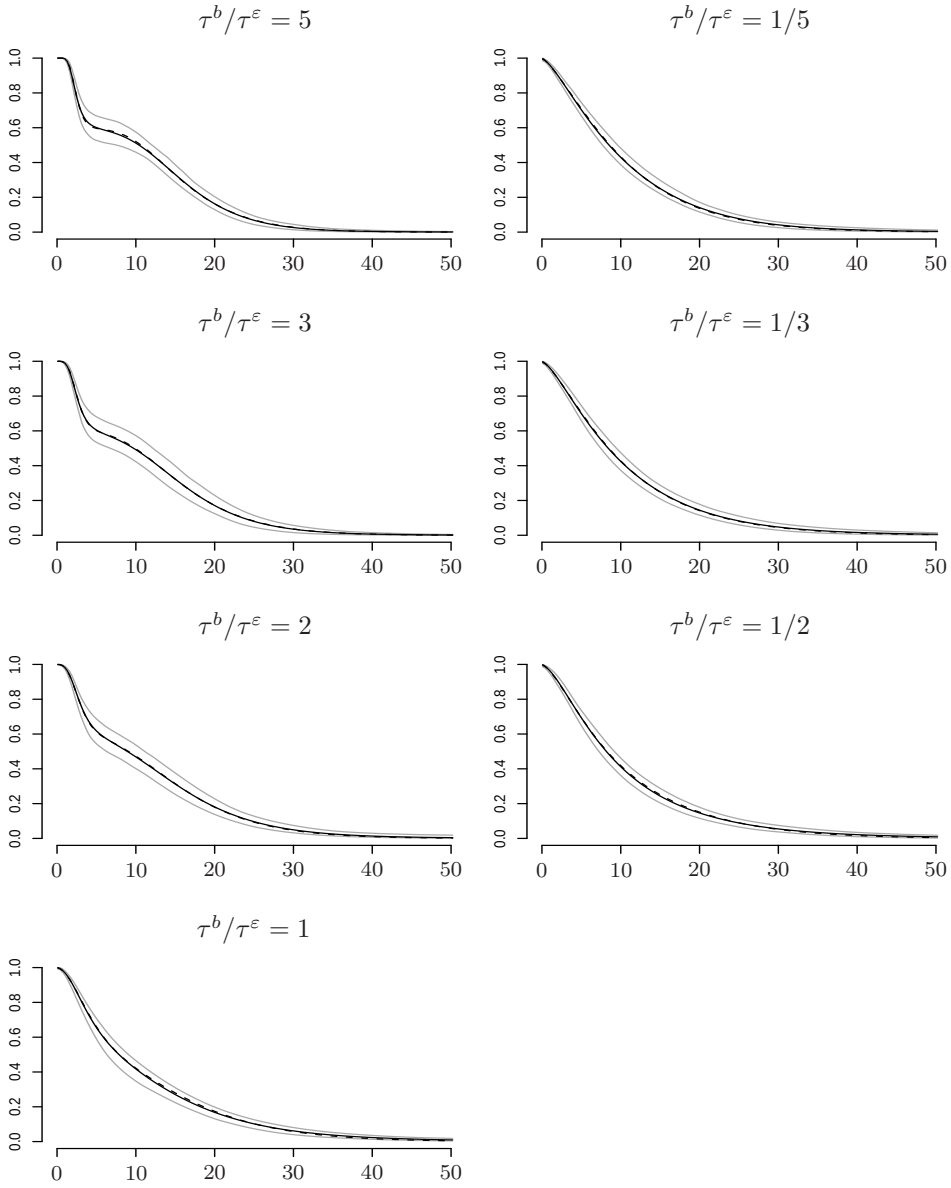


Figure B.17: Results for the survivor function of the event time, for the combination of covariates $\boldsymbol{x}_{i,l}^t = (0.5,\ 1)'$, scenario II (error $\sim$ extreme value, random effect $\sim$ normal mixture). Solid line: average fitted survivor function, grey lines: 95% pointwise confidence band, dashed line: true survivor function..

Appendix C

Software

For all methodologies described in Part II of the thesis a software in the form of R (R Development Core Team, 2005) packages smoothSurv and bayesSurv has been written and can be downloaded, together with extensive manuals and description on how to perform analyses shown in this thesis from the *Comprehensive R Archive Network* at `http://www.R-project.org`. To optimize the computational time, all time consuming computation is performed using the C++ compiled code. In this appendix, we only briefly list the most important functions from both packages.

C.1 Package smoothSurv

This package implements the methods for the penalized maximum-likelihood AFT model as described in Chapter 7 and involves, among others, the following functions:

smoothSurvReg fits the AFT model (7.1) with the error density (7.2) using the method of penalized maximum-likelihood. It also allows for the scale regression (7.6);

plot.smoothSurvReg computes and plots the fitted error density (7.2);

survfit.smoothSurvReg computes and plots the fitted survival function (7.13) for a specified combination of covariates;

fdensity computes and plots the fitted survival density (7.14) for a specified combination of covariates;

hazard computes and plots the fitted hazard function (7.15) for a specified combination of covariates;

estimTdiff estimates expected survival time for a specified combination of covariates or estimates expected value of the difference between the survival times for two specified combinations of covariates based on the AFT model fitted using the function smoothSurvReg.

C.2 Package bayesSurv

This package implements the Bayesian methods described in Chapters 8 – 10.

For the Bayesian normal mixture cluster-specific AFT model of Chapter 8, the core functions include:

bayessurvreg1 runs the MCMC simulation for the AFT model (8.1) with the error density (8.2) and normally distributed (multivariate) random effects;

bayesDensity computes the estimate of the predictive error densities (8.20) and (8.21);

predictive computes the MCMC estimate of the predictive survival, density or hazard function for a specified combination of covariates based on the formulas (8.16), (8.18) and (8.19).

For the Bayesian penalized mixture cluster-specific and population-averaged AFT models of Chapters 9 and 10, the core functions include:

bayessurvreg2 runs the MCMC simulation for the cluster-specific AFT model (9.1), (9.2) with the error densities specified by (9.3) and normally distributed (multivariate) random effects (*Model M*);

bayessurvreg3 runs the MCMC simulation for the cluster-specific AFT model (9.1), (9.2) with the error densities specified by (9.3) and univariate random effects whose distribution is specified by (9.3) (*Model U*);

bayesBisurvreg runs the MCMC simulation for the population-averaged AFT model (10.1), (10.2) with the error densities specified by (10.3);

bayesGspline computes the estimate of the predictive density of the factors whose distribution was specified as the penalized normal mixture (9.3) or (10.3). The function is based on formulas (9.13) and (10.12);

marginal.bayesGspline computes the estimates of the predictive marginal densities of the factors whose distribution was specified as the bivariate penalized normal mixture (10.3).

predictive2 computes the MCMC estimate of the predictive survival, density or hazard function for a specified combination of covariates based on the formulas (9.10), (9.11) or (10.10), (10.11).

Bibliography

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, **3**, 227–243.

Abrahamowicz, M., Ciampi, A., and Ramsay, J. O. (1992). Nonparametric density estimation for censored survival data: regression-spline approach. *The Canadian Journal of Statistics*, **20**, 171–185.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–1174.

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

Bacchetti, P. (1990). Estimation the incubation period of AIDS comparing population infection and diagnosis patterns. *Journal of the American Statistical Association*, **85**, 1002–1008.

Bacchetti, P. and Jewell, N. P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics*, **47**, 947–960.

Barkan, S. E., Melnick, S. L., Preston-Martin, S., Weber, K., Kalish, L. A., Miotti, P., Young, M., Greenblatt, R., Sacks, H., and Feldman, J. (1998). The Women's Interagency HIV Study. *Epidemiology*, **9**, 117–125.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with Discussion). *Statistical Science*, **10**, 3–66.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (1999). Local EM estimation of the hazard function for interval-censored data. *Biometrics*, **55**, 238–245.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, **21**, 263–275.

Betensky, R. A., Rabinowitz, D., and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika*, **88**, 703–711.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York, Third edition. ISBN 0-471-00710-2.

Bogaerts, K. and Lesaffre, E. (2004). A new, fast algorithm to find the regions of possible support for bivariate interval-censored data. *Journal of Computational and Graphical Statistics*, **13**, 330–340.

Bogaerts, K. and Lesaffre, E. (2006). Estimating Kendall's tau for bivariate interval censored data with a smooth estimate of the density. *Submitted*.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with Discussion). *Journal of the Royal Statistical Society, Series B*, **65**, 3–55.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–436.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, **59**, 570–579.

Calle, M. L. and Gómez, G. (2005). A semiparametric hierarchical method for a regression model with an interval-censored covariate. *Australian and New Zealand Journal of Statistics*, **47**, 351–364.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, Boca Raton, Second edition. ISBN 1-58488-170-4.

Carvalho, J. C., Ekstrand, K. R., and Thylstrup, A. (1989). Dental plaque and caries on occlusal surfaces of first permanent molars in relation to stage of eruption. *Journal of Dental Research*, **68**, 773–779.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York. ISBN 0-387-98935-8.

Christensen, R. and Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process. *Biometrika*, **75**, 693–704.

Clahsen, P. C., van de Velde, C. J., Julien, J. P., Floiras, J. L., Delozier, T., Mignolet, F. Y., and Sahmoud, T. M. (1996). Improved local control and disease-free survival after perioperative chemotherapy for early-stage breast cancer. A European Organization for Research and Treatment of Cancer Breast Cancer Cooperative Group Study. *Journal of Clinical Oncology*, **14**, 745–753.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. ISBN 0-412-16160-5.

Czyzyk, J., Mesnier, M. P., and Moré, J. J. (1998). The NEOS server. *IEEE Journal on Computational Science and Engineering*, **5**, 68–75.

Dalal, S. R. and Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society, Series B*, **45**, 278–286.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York. ISBN 0-387-90356-9.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1–11.

Dellaportas, P. and Papageorgiou, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing*, **16**, 57–68.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon, Oxford. ISBN 0-19-853440-X.

Dorey, F. J., Little, R. J., and Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, **12**, 1589–1603.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statistical Science*, **11**, 89–121.

Ekstrand, K. R., Christiansen, J., and Christiansen, M. E. (2003). Time and duration of eruption of first and second permanent molars: a longitudinal investigation. *Community Dentistry and Oral Epidemiology*, **31**, 344–350.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, Second edition.

Fang, H.-B., Sun, J., and Lee, M.-L. T. (2002). Nonparametric survival comparisons for interval-censored continuous data. *Statistica Sinica*, **12**, 1073–1083.

Fay, M. P. (1996). Rank invariant tests for interval censored data under grouped continuous model. *Biometrics*, **52**, 811–822.

Fay, M. P. (1999). Comparing several score tests for interval censored data. *Statistics in Medicine*, **18**, 273–285.

Fay, M. P. and Shih, J. H. (1998). Permutation tests using estimated distribution functions. *Journal of the American Statistical Association*, **93**, 387–396.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615–629.

Ferris, M. C., Mesnier, M. P., and Moré, J. (2000). NEOS and Condor: Solving nonlinear optimization problems over the Internet. *ACM Transactions on Mathematical Software*, **26**, 1–18.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845–854.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York. ISBN 0-471-52218-X.

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley & Sons, Chichester, Second edition. ISBN 0-471-49463-1.

Fourer, R., Gay, D. M., and Kernighan, B. W. (2003). *AMPL: A Modeling Language for Mathematical Programming*. Duxbury Press, Second edition. ISBN 0-534-388094.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, London. ISBN 0-412-81820-5.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203–223.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–499.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *To appear in Bayesian Analysis*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Second edition. ISBN 1-58488-388-X.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences (with Discussion). *Statistical Science*, **7**, 457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayes restoration of image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometrika*, **81**, 618–623.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with Discussion). *Statistical Science*, **7**, 473–511.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945–953.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag, New York. ISBN 0-387-95537-2.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London. ISBN 0-412-05551-1.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Number 124 in Mathematical Centre Tracts. Mathematisch Centrum, Amsterdam. ISBN 90-6196-197-1.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics*, **56**, 1139–1144.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics*, **54**, 1498–1507.

Goggins, W. B., Finkelstein, D. M., and Zaslavsky, A. M. (1999). Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Statistics in Medicine*, **18**, 2737–2747.

Gómez, G. and Calle, M. L. (1999). Non-parametric estimation with doubly censored data. *Journal of Applied Statistics*, **26**, 45–58.

Gómez, G., Calle, M. L., and Oller, R. (2004). Frequentist and Bayesian approaches for interval-censored data. *Statistical Papers*, **45**, 139–173.

Gómez, G., Espinal, A., and Lagakos, S. W. (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, **22**, 409–425.

Gómez, G. and Lagakos, S. W. (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics*, **50**, 204–212.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, **87**, 942–951.

Green, P. J. (1995). Reversible jump Markov chain computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser-Verlag, Boston. ISBN 0-8176-2794-4.

Han, S. P. (1977). A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications*, **22**, 297–309.

Hanson, T. and Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**, 1020–1033.

Hanson, T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics*, **13**, 341–361.

Härkänen, T. (2003). BITE: A Bayesian intensity estimator. *Computational Statistics*, **18**, 565–583.

Härkänen, T., Virtanen, J. I., and Arjas, E. (2000). Caries on permanent teeth: a nonparametric Bayesian analysis. *Scandinavian Journal of Statistics*, **27**, 577–588.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005–1016.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York. ISBN 0-387-95284-5.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, **13**, 20–35.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, **55**, 13–22.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York. ISBN 0-387-98873-4.

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, **9**, 501–519.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York. ISBN 0-387-95277-2.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20**, 50–67.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, **90**, 341–353.

Johnson, W. and Christensen, R. (1989). Nonparametric Bayesian analysis of the accelerated failure time model. *Statistics and Probability Letters*, **8**, 179–184.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age–specific incidence of dementia. *Biometrics*, **54**, 185–194.

Kalbfleisch, J. D. and MacKay, R. J. (1979). On constant-sum models for censored survival data. *Biometrika*, **66**, 87–90.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Chichester, Second edition. ISBN 0-471-36357-X.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kauermann, G. (2005a). A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planning and Inference*, **127**, 53–69.

Kauermann, G. (2005b). Penalised spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis*, **49**, 169–186.

Keiding, N., Andersen, P. K., and Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, **16**, 215–225.

Kim, M. Y., De Gruttola, V. G., and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, **49**, 13–22.

Komárek, A. and Lesaffre, E. (2006a). Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution. *To appear in Statistica Sinica*.

Komárek, A. and Lesaffre, E. (2006b). Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Submitted*.

Komárek, A. and Lesaffre, E. (2006c). Bayesian semiparametric accelerated failure time model for paired doubly-interval-censored data. *Statistical Modelling*, **6**, 3–22.

Komárek, A., Lesaffre, E., Härkänen, T., Declerck, D., and Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored data. *Biostatistics*, **6**, 145–155.

Komárek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, **14**, 726–745.

Kooperberg, C. (1998). Bivariate density estimation with an application to survival analysis. *Journal of Computational and Graphical Statistics*, **7**, 322–341.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics*, **53**, 1485–1494.

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, **1**, 301–328.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, **90**, 78–94.

Kottas, A. and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **96**, 1458–1468.

Kuo, L. and Mallick, B. (1997). Bayesian semiparametric inference for the accelerated failure time model. *The Canadian Journal of Statistics*, **25**, 457–472.

Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics*, **19**, 1370–1402.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lambert, P., Collett, D., Kimber, A., and Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*, **23**, 3177–3192.

Lambert, P. and Eilers, P. H. C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: A penalized Poisson regression approach. *Statistics in Medicine*, **24**, 3977–3989.

Langohr, K., Gómez, G., and Muga, R. (2004). A parametric survival model with an interval-censored covariate. *Statistics in Medicine*, **23**, 3159–3175.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, **20**, 1222–1235.

Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, **22**, 1161–1176.

Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, **11**, 1569–1578.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., and Bertollini, R., editors (1999). *Disease Mapping and Risk Assessment for Public Health*. John Wiley & Sons, Chichester. ISBN 0-471-98634-8.

Lee, E. W., Wei, L. J., and Ying, Z. (1993). Linear regression analysis for highly stratified failure time data. *Journal of the American Statistical Association*, **88**, 557–565.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view (with Discussion). *Statistical Science*, **19**, 219–238.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York, Second edition. ISBN 0-387-98502-6.

Leroy, R., Bogaerts, K., Lesaffre, E., and Declerck, D. (2003a). The effect of fluorides and caries in primary teeth on permanent tooth emergence. *Community Dentistry and Oral Epidemiology*, **31**, 463–470.

Leroy, R., Bogaerts, K., Lesaffre, E., and Declerck, D. (2003b). The emergence of permanent teeth in Flemish children (Belgium). *Community Dentistry and Oral Epidemiology*, **31**, 30–39.

Leroy, R., Bogaerts, K., Lesaffre, E., and Declerck, D. (2005). Effect of caries experience in primary molars on cavity formation in the adjacent permanent first molar. *Caries Research*, **39**, 342–349.

Lesaffre, E., Komárek, A., and Declerck, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, **14**, 539–552.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lin, J. S. and Wei, L. J. (1992). Linear regression analysis for multivariate failure time observations. *Journal of the American Statistical Association*, **87**, 1091–1097.

Lindsey, J. K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine*, **17**, 447–469.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, **12**, 351–357.

Louis, T. A. (1981). Nonparametric analysis of an accelerated failure time model. *Biometrika*, **68**, 381–390.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**, 163–170.

Mantel, N. (1967). Ranking procedures for arbitrarily restricted observations. *Biometrics*, **23**, 65–78.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Pólya trees and random distributions. *The Annals of Statistics*, **20**, 1203–1221.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York. ISBN 0-8247-7691-7.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63**, 449–464.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Science+Business Media, New York. ISBN 0-387-25144-8.

Nanda, R. S. (1960). Eruption of human teeth. *American Journal of Orthodontics*, **46**, 363–378.

Nardi, A. and Schemper, M. (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, **22**, 3597–3610.

Neal, R. M. (2003). Slice sampling (with Discussion). *The Annals of Statistics*, **31**, 705–767.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992). Maximum likelihood estination for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, **48**, 951–959.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, London, Sixth edition. ISBN 0-340-52922-9.

Oller, R., Gómez, G., and Calle, M. L. (2004). Interval censoring: model characterization for the validity of the simplified likelihood. *The Canadian Journal of Statistics*, **32**, 315–326.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problem (with Discussion). *Statistical Science*, **1**, 502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log–density and log–hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, **9**, 363–379.

Oulis, C. J., Raadal, M., and Martens, L. (2000). Guidelines on the use of fluoride in children: an EAPD policy document. *European Journal of Paediatric Dentistry*, **1**, 7–12.

Pan, J. and MacKenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244.

Pan, W. (1999a). A comparison of some two-sample tests with interval censored data. *Nonparametric Statistics*, **12**, 133–146.

Pan, W. (1999b). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. *Journal of Computational and Graphical Statistics*, **8**, 109–120.

Pan, W. (2000a). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, **56**, 199–203.

Pan, W. (2000b). A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine*, **19**, 1–11.

Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics*, **57**, 1245–1250.

Pan, W. and Connett, J. E. (2001). A multiple imputation approach to linear regression with clustered censored data. *Lifetime Data Analysis*, **7**, 111–123.

Pan, W. and Kooperberg, C. (1999). Linear regression for bivariate censored data via multiple imputation. *Statistics in Medicine*, **18**, 3111–3121.

Pan, W. and Louis, T. A. (2000). A linear mixed-effects model for multivariate censored data. *Biometrics*, **56**, 160–166.

Parner, E. T., Heidmann, J. M., Væth, M., and Poulsen, S. (2001). A longitudinal study of time trendsin the eruption of permanent teeth in Danish children. *Archives of Oral Biology*, **46**, 425–431.

Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, **45**, 497–507.

Pepe, M. S. and Fleming, T. R. (1991). Weighted Kaplan-Meier statistics: large sample and optimality considerations. *Journal of the Royal Statistical Society, Series B*, **53**, 341–352.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, **22**, 86–91.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank-invariant test procedures (with Discussion). *Journal of the Royal Statistical Society, Series A*, **135**, 185–206.

Petroni, G. R. and Wolfe, R. A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics*, **50**, 77–87.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parametrisation. *Biometrics*, **86**, 677–690.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167–179.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, **82**, 501–513.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, **3**, 425–461.

Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, **9**, 439–455.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components (with Discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, **18**, 303–328.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, Second edition. ISBN 0-387-21239-6.

Roeder, K. and Wasserman (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, **51**, 874–887.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York. ISBN 0-471-08705-X.

Rücker, G. and Messerer, D. (1988). Remission duration: an example of interval-censored observations. *Statistics in Medicine*, **7**, 1139–1145.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, **83**, 355–370.

Satten, G. A., Datta, S., and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association*, **93**, 318–327.

Self, S. G. and Grossman, E. A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics*, **42**, 521–530.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non–parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, **47**, 1–52.

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

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics*, **51**, 1096–1104.

Sun, J., Liao, Q., and Pagano, M. (1999). Regression analysis of doubly censored failure time data with application to AIDS studies. *Biometrics*, **55**, 909–914.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York. ISBN 0-387-98784-3.

Therneau, T. M. and Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, **16**, 2029–2047.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *The Annals of Statistics*, **22**, 1701–1762.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester. ISBN 0-471-90763-4.

Topp, R. and Gómez, G. (2004). Residual analysis in linear regression models with an interval-censored covariate. *Statistics in Medicine*, **23**, 3377–3391.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, **18**, 354–372.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14**, 809–834.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **37**, 290–295.

Tutz, G. and Binder, H. (2004). Flexible modelling of discrete failure time including time-varying smooth effects. *Statistics in Medicine*, **23**, 2445–2461.

Unser, M., Aldroubi, A., and Eden, M. (1992). On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory*, **38**, 864–872.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309–3324.

Vanobbergen, J., Martens, L., Lesaffre, E., Bogaerts, K., and Declerck, D. (2001). Assessing risk indicators for dental caries in the primary dentition. *Community Dentistry and Oral Epidemiology*, **29**, 424–434.

Vanobbergen, J., Martens, L., Lesaffre, E., and Declerck, D. (2000). The Signal-Tandmobiel® project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, **2**, 87–96.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, **23**, 541–556.

Verweij, P. J. M. and Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, **13**, 2427–2436.

Virtanen, J. I. (2001). Changes and trends in attack distributions and progression of dental caries in three age cohorts in Finland. *Journal of Epidemiology and Biostatistics*, **6**, 325–329.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross–validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, **45**, 133–150.

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B*, **61**, 485–527.

Walker, S. G. and Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics*, **55**, 477–483.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223–249.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.

Wei, G. C. G. and Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, **47**, 1297–1309.

Williams, J. S. and Lagakos, S. W. (1977). Models for censored survivaal analysis: Constant-sum and variable-sum models. *Biometrika*, **64**, 215–224.

Ying, Z. (1993). A large sample study of rank estimation for censored refression data. *The Annals of Statistics*, **21**, 76–99.

Yu, Q., Li, L., and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, **27**, 35–44.

Yu, Q., Schick, A., Li, L., and Wong, G. Y. C. (1998). Asymptotic properties of the GLME in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics*, **26**, 619–627.

Curriculum Vitae

Arnošt Komárek was born on March 28, 1977 in Hradec Králové in the Czech Republic. After secondary school at Božena Němcová Secondary Grammar School (Gymnázium Boženy Němcové) in Hradec Králové, he started undergraduate studies in Mathematics in September 1995 at the Faculty of Mathematics and Physics of the Charles University (Univerzita Karlova) in Prague, the Czech Republic where he chose direction of Mathematical Statistics and graduated as Master of Science in Mathematical Statistics in May 2000. From October 2000 till September 2001, he was enrolled as an Erasmus Exchange Student at the University of Limburg (Limburgs Universitair Centrum, nowadays Universiteit Hasselt) in Diepenbeek, Belgium and obtained a degree of Master of Science in Biostatistics. From October 2001 he started as a predoctoral student his career of a researcher at the Biostatistical Centre of the Catholic University of Leuven (Katholieke Universiteit Leuven) in Leuven, Belgium. At the same place, he started the doctoral programme in October 2002 of which this thesis is the most important outcome.