August 1, 2005

Bayesian Accelerated Failure Time Model for Correlated Interval-Censored Data with a Normal Mixture as Error Distribution

Arnošt Komárek and Emmanuel Lesaffre

Biostatistical Centre, Katholieke Universiteit Leuven,

Kapucijnenvoer 35, B–3000, Leuven, Belgium

E-mail: emmanuel.lesaffre@med.kuleuven.be

Tel: +32-16-33 68 96; Fax: +32-16-33 69 00

Summary: A Bayesian approach for an accelerated failure time model with interval-censored data is proposed. The model allows for structured correlated data by inclusion of a random effect part that might depend on covariates as in a linear mixed model. The error distribution is modelled as a normal mixture with an unknown number of components. Also the means and variances of the components are not pre-specified to accommodate most continuous distributions which results, among other things, in a nearly correct estimation of the shape of hazard and survivor curves. A Markov chain Monte Carlo algorithm is described to sample from the posterior distribution. The approach is evaluated using a simulation study and is illustrated on modelling the emergence times of eight permanent teeth using the data from the Signal Tandmobiel® study.

Key words: Clustered Data; Multicenter Study; Regression; Reversible Jump Markov Chain Monte Carlo; Survival Data.

1 Introduction

Correlated survival data are encountered in many medical problems, for instance when the observations are clustered. Moreover, in some problems, the occurrence of the event can only be recorded at regular intervals leading to *interval-censored* data.

Correlated interval-censored survival times were encountered in the Signal Tandmobiel® study. This is a longitudinal prospective (1996–2001) oral health screening project performed in Flanders, Belgium. The 4 468 school-children, born in 1989 were examined on a yearly basis by one of 16 trained dental-examiners. The details of the study design can be found in Vanobbergen et al. (2000). For oral health researchers, two research questions are of interest. Firstly, what is the effect of decayed primary predecessors (described by a binarised $dmft$ score) on the emergence times of the permanent premolars (teeth 14, 15, 24, 25, 34, 35, 44, 45 in European dental notation)? The emergences were observed only in intervals of length equal to 1 year. Secondly, what is the correlation between the emergence times of different teeth? Leroy et al. (2003) have shown that there is *horizontal symmetry*, i.e. some information concerning the correlation structure is available as the same emergence distribution can be assumed at horizontally symmetric positions (e.g., for teeth 14 and 24). Further, it is known that factors, like gender, have an impact on the emergence time and thus should be controlled for.

In multicenter clinical trials, it is often necessary to control for center and to check the center by treatment interaction when evaluating the treatment effect. An important center effect or center by treatment interaction can be due to center differences in social-economics characteristics of the patients, different training of the medical staff, differences in the administration of treatment, etc. To obtain a valid statistical conclusion on treatment efficacy, it should be controlled for such effects. Furthermore, in situations when disease progression can

only be revealed by a laboratory assessment (AIDS, some types of cancer), the observed event times are interval-censored as well.

The paper is organized as follows. In Section 2, we review models for correlated censored data. In Section 3 we propose a new approach – the Bayesian mixture MEAFT model. In Section 4, a Markov chain Monte Carlo algorithm to sample from the posterior distribution is discussed. The approach is evaluated using a simulation study in Section 5 and used to analyze the Signal Tandmobiel® data in Section 6.

2 Models for correlated censored data

Several approaches to analyze correlated right-censored survival times have been proposed. One approach is to extend the Cox's proportional hazards (PH) model (Cox, 1972) by including a cluster-specific random effect, called *frailty* in the expression of the hazard function (see, e.g., Hougaard, 2000; Therneau and Grambsch, 2000). The frailty component is most often assumed to have a parametric distribution such as gamma or log-normal. However, the frailty PH model has some important drawbacks. Firstly, the implied correlation structure is too simple, e.g., in the analysis of the multicenter clinical trials only the center effect and not the center by treatment interaction can be controlled for. Secondly, the choice of the frailty distribution can have a crucial impact on the results for the regression parameters of interest (Hougaard, 2000, Chapter 7). Thirdly, the PH model is generally not robust towards neglected covariates (Hougaard, 1999).

A possible alternative to the PH model is the accelerated failure time (AFT) model which assumes that the covariates speed up or slow down the expected event time. We refer to Chapter 7 of Kalbfleisch and Prentice (2002) for an extensive review of classical approaches to the AFT model. In contrast to the PH model, in the AFT model, neglected covariates do not

cause bias in estimating the regression parameters for the included covariates (Hougaard, 1999). An extension of the AFT model – the mixed effects accelerated failure time (MEAFT) model – takes into account the within-cluster correlations explicitly by including random effects in the regression expression as in a classical linear mixed model of Laird and Ware (1982), namely

$$\log(T_{i,l}) \equiv Y_{i,l} = \boldsymbol{\beta}^T \boldsymbol{x}_{i,l} + \boldsymbol{b}_i^T \boldsymbol{z}_{i,l} + \varepsilon_{i,l}, \qquad i = 1, \ldots, N, \quad l = 1, \ldots, n_i, \tag{1}$$

where $T_{i,l}$ is the event time of the lth observation of the ith cluster, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the unknown regression coefficient vector, $\boldsymbol{x}_{i,l}$ the covariate vector for fixed effects, $\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,q})^T$, $i = 1, \ldots, N$ are i.i.d. random effects vectors with a density $g(\boldsymbol{b})$, $\boldsymbol{z}_{i,l}$ is the covariate vector for random effects and $\varepsilon_{i,l}$, $i = 1, \ldots, N$, $l = 1, \ldots, n_i$ are i.i.d. error random variables with a density $f(\varepsilon)$. As the roles of the regression parameters and dispersion parameters are clearly separated in the MEAFT, the regression parameters are robust against misspecification of the random effects distribution $g(\boldsymbol{b})$ (Keiding et al., 1997; Lambert et al., 2004). Model (1), restricted to the case of $z_{i,l} \equiv 1$, has been studied by Pan and Louis (2000); Pan and Connett (2001) for right-censored data. They make a working assumption concerning the normality of both $\varepsilon_{i,l}$ and b_i and use frequentist techniques for uncensored data combined with Monte Carlo EM algorithm and multiple imputation, respectively to overcome the problem of censoring.

Assume now that the (i,l)th true log-event time $y_{i,l}$ is only known to lie in the interval $(y_{i,l}^L, y_{i,l}^U]$, $-\infty \leq y_{i,l}^L \leq y_{i,l}^U \leq \infty$. For an uncensored observation: $y_{i,l}^L = y_{i,l}^U$, for a right-censored observation: $y_{i,l}^U = \infty$ and for a left-censored observation: $y_{i,l}^L = -\infty$. The likelihood contribution of the ith cluster is given by

$$L_i = \int_{\mathbb{R}^q} \Big\{ \prod_{l=1}^{n_i} \int_{y_{i,l}^L}^{y_{i,l}^U} f(y - \boldsymbol{\beta}^T \boldsymbol{x}_{i,l} - \boldsymbol{b}^T \boldsymbol{z}_{i,l})\, dy \Big\} g(\boldsymbol{b})\, d\boldsymbol{b}, \tag{2}$$

where the convention $\int_a^a f(s)\, ds \equiv f(a)$ applies to accommodate also uncensored observations. Due to multiple integration in the likelihood (2), it is rather cumbersome to use maximum-likelihood based methods for the MEAFT model with interval-censored observations even with

$f(\varepsilon)$ and $g(\boldsymbol{b})$ being parametrically specified. While stochastic versions of standard estimation techniques can be used, as was done by Pan and Louis (2000) or Pan and Connett (2001), we believe that a Bayesian approach is more natural and easier to use here.

Furthermore, for small samples or in situations when prediction and not only regression parameters themselves are of interest, it is desirable to avoid full parametric assumptions (like normality) concerning the error density $f(\varepsilon)$ since it determines the shape and character of resulting survival and hazard curves which are to be estimated from the data. For that reason, we sought after a method that offers enough flexibility in specifying the error density $f(\varepsilon)$ while being computationally tractable for both interval-censored data and a general covariate vector $\boldsymbol{z}_{i,l}$.

3 A Bayesian mixture MEAFT model

To our best knowledge, the MEAFT model with a general q-variate random effects covariate vector $\boldsymbol{z}_{i,l}$ being inevitable to solve problems outlined in the introduction and a flexible density $f(\varepsilon)$ has not been considered in the literature yet. 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. However, until the last decade the statistical analysis of mixtures has not been straightforward. The reversible jump MCMC algorithm (Green, 1995) to estimate unknown mixture parameters, suggested by Richardson and Green (1997) is a breakthrough in this area. We adopt their approach to model the error density $f(\varepsilon)$ in the MEAFT model and argue that this extension of the simple normal distribution for $f(\varepsilon)$ offers a rich family of distributions of various shapes suitable to model practically any survival data. See Section 5 for some examples of densities and corresponding hazard or survivor functions approximated by a normal mixture.

At the same time, the MCMC methodology easily overcomes a problem of the difficult likelihood, expression (2). Indeed, there is no need to maximize this likelihood since the sample from the posterior distribution obtained using the MCMC method is used to draw the inference. Furthermore, MCMC replaces both integrals in (2) by sampling exact event times and values of latent random effects from appropriate, ease-to-sample distributions as will be shown in Sections 4.1 and 4.4.

We assume a Bayesian mixture MEAFT model (1) with a hierarchical structure graphically represented by a directed acyclic graph (DAG) given in Figure 1 where the usual convention of graphical models is used, i.e. square boxes represent fixed or observed quantities and circles the unknown parameters, solid lines represent stochastic while dashed lines express deterministic dependencies, respectively. The joint prior distribution is given by the product of the conditional distributions of each node given its parents which we discuss in this section. As the DAG indicates, the unknown parameters can be split into two parts connected only through the node of true log-event times.

[Figure 1 about here.]

3.1 Prior specification of the error part

The density $f(\varepsilon)$ of the error term $\varepsilon_{i,l}$ in model (1) is specified as

$$f(\varepsilon) = \sum_{j=1}^{k} w_j \, \varphi(\varepsilon \mid \mu_j, \, \sigma_j^2), \tag{3}$$

with $\varphi(\cdot \mid \mu_j, \, \sigma_j^2) \equiv$ density of $\mathcal{N}_1(\mu_j, \sigma_j^2)$. Note that the number of mixture components, k, is unknown as well as the mixture weights $\boldsymbol{w} = (w_1, \ldots, w_k)^T$, means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$, and variances $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)^T$. It is well known (McLachlan and Basford, 1988, Chapter 2) that a heteroscedastic mixture (3) leads to the likelihood which is unbounded if the parameter

space for variances is unconstrained. In a Bayesian analysis, this difficulty is solved by using an appropriate prior distribution for the variances which plays the role of constraints.

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 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. The corresponding DAG conditional distributions are then given by

$$\varepsilon_{i,l} \,|\, \boldsymbol{\mu},\, \boldsymbol{\sigma}^2,\, r_{i,l} \sim \mathcal{N}_1(\mu_{r_{i,l}}, \sigma^2_{r_{i,l}}), \qquad i = 1, \ldots, N,\ l = 1, \ldots, n_i, \tag{4}$$

$$\Pr(r_{i,l} = j \,|\, k,\, \boldsymbol{w}) = w_j, \qquad j = 1, \ldots, k. \tag{5}$$

DAG conditional distributions of the remaining parameters of the error part of the model are inspired by the work of Richardson and Green (1997) (with some change in notation). We give a brief summary. For the number of mixture components, k, we experimented with (1) a Poisson distribution with mean equal to a hyper-parameter λ truncated at some prespecified (relatively large) value k_{max} and (2) a uniform distribution on $\{1, \ldots, k_{max}\}$ (the node λ in the DAG in Figure 1 becomes redundant then). 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. $\boldsymbol{w}\,|k, \delta \sim \mathrm{D}(\delta, \delta, \ldots, \delta)$, where δ is a fixed hyper-parameter. Further, the mixture means μ_j and variances σ_j^2, respectively are a priori all drawn independently with normal and inverse-gamma priors $\mu_j \mid k, \xi, \kappa \sim \mathrm{N}(\xi, \kappa)$ and $\sigma_j^2 \mid k, \zeta, \eta \sim \mathrm{IG}(\zeta, \eta)$, respectively. As in Richardson and Green (1997) we let η have a gamma distribution $\mathrm{G}(g, h)$ with fixed hyper-parameters g and h, see Section 3.4 for more details.

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) for other approaches to establish identifiability. The joint prior distribution

of the mixture means and variances is thus $k!$ times the product of the individual normal and inverse-gamma densities, restricted to above mentioned set of increasing means.

3.2 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 an $\sum_{i=1}^{N} n_i \times p$ matrix with vectors $\boldsymbol{x}_{1,1}^T, \ldots, \boldsymbol{x}_{N,n_N}^T$ as rows. Similarly, let $\mathbb{Z}$ be an $\sum_{i=1}^{N} n_i \times q$ matrix with vectors $\boldsymbol{z}_{1,1}^T, \ldots, \boldsymbol{z}_{N,n_N}^T$ as rows. Further, we will assume that the matrix $(\mathbb{X}, \mathbb{Z})$ is of full column rank $(p+q)$. 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 $f(\varepsilon)$ 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 is assumed to be $\mathrm{N}(\nu_{\beta,j}, \psi_{\beta,j})$, $j = 1, \ldots, p$ and the β_j are assumed to be a priori independent. The vectors $\boldsymbol{\nu}_\beta = (\nu_{\beta,1}, \ldots, \nu_{\beta,p})^T$ and $\boldsymbol{\psi}_\beta = (\psi_{\beta,1}, \ldots, \psi_{\beta,p})^T$ are fixed hyper-parameters.

The prior distribution for the random effect vector $\boldsymbol{b}_i$ is assumed to be (multivariate) normal, i.e.

$$\boldsymbol{b}_i \,|\, \boldsymbol{\gamma}, \mathbb{D} \;\sim\; \mathrm{N}_q(\boldsymbol{\gamma}, \mathbb{D}), \quad \text{independently for } i = 1, \ldots, N, \tag{6}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$. The prior distribution for each γ_j, is $\mathrm{N}(\nu_{\gamma,j}, \psi_{\gamma,j})$, independently for $j = 1, \ldots, q$. The vectors $\boldsymbol{\nu}_\gamma = (\nu_{\gamma,1}, \ldots, \nu_{\gamma,q})^T$ and $\boldsymbol{\psi}_\gamma = (\psi_{\gamma,1}, \ldots, \psi_{\gamma,q})^T$ 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 (3). 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 IW(τ, $\mathbb{S}$) (parametrized such that the mean is $(\tau - q - 1)^{-1}\mathbb{S}$), where τ denotes 'degrees of freedom' ($\tau > q - 1$) and $\mathbb{S}$ is a scale matrix.

Finally, the DAG conditional distributions of the (unknown) log-event times, i.e. nodes that connect the regression and error parts of the DAG, are all deterministic given by the MEAFT model (1).

3.3 Censoring

To complete the specification of the DAG we need to specify $p(y^L_{i,l}, y^U_{i,l} \,|\, y_{i,l}, \text{censoring})$. Firstly, the censoring mechanism in this paper is assumed to be non-informative about the failure distribution. A box called 'censoring' in the DAG represents a realization of the random variable(s) causing the censoring. Note that there is no need to specify a measurement model for the censoring mechanism since the inference relies on the posterior distribution of parameters given the data and the data consist of the realized censoring variables as well.

After omitting subscripts i, l for clarity, the form of $p(y^L, y^U \,|\, y, \text{censoring})$ is rather obvious for most censoring mechanisms. In the case of right censoring driven by a censoring random variable C, $p(y^L, y^U \,|\, y, c)$ is a discrete density with $P\big[(y^L, y^U) = (y, y) \,\big|\, y, c\big] = I[y \leq c]$, $P\big[(y^L, y^U) = (c, \infty) \,\big|\, y, c\big] = I[y > c]$. For interval censoring resulting from a realization of random variables $C_1, \ldots, C_m$ representing the times when a failure status was checked the density $p(y^L, y^U \,|\, y, c_1, \ldots, c_m)$ is again a discrete density with $P\big[(y^L, y^U) = (c_j, c_{j+1}) \,\big|\, y, c_1, \ldots, c_m\big] = I[c_j < y \leq c_{j+1}]$, $j = 0, \ldots, m$ with $c_0 = -\infty,\ c_{m+1} = \infty$.

3.4 Weak Prior Information

In this paper, 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 $\boldsymbol{\beta}$ ($\boldsymbol{\psi}_\beta$) 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 hyper-parameters for the covariance matrix $\mathbb{D}$ giving a weak prior information correspond to choices of $\tau = q - 1 + d$ and $\mathbb{S} = \text{diag}(d, \ldots, d)$ with d 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 $\text{N}(\xi, \kappa)$ for μ_j is achieved by letting ξ equal to $\bar{e} = n^{-1} \sum_{j=1}^{n} e_j$ and setting κ equal to a multiple of R^2, where $R = \max(e_i) - \min(e_i)$. They further 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 g and h. They further recommend taking $\zeta > 1 > g$ 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 to a small multiple of $1/R^2$. Here, the residuals $y_{i,l} - \boldsymbol{\beta}^T \boldsymbol{x}_{i,l} - \boldsymbol{b}_i^T \boldsymbol{z}_{i,l}$ 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.

4 Markov chain Monte Carlo algorithm

Details of the implementation of the MCMC algorithm for the parameters of the error part of the model are given in Richardson and Green (1997). Their guidelines, now based on residuals $\varepsilon_{i,l} = y_{i,l} - \boldsymbol{\beta}^T \boldsymbol{x}_{i,l} - \boldsymbol{b}_i^T \boldsymbol{z}_{i,l}$ can be immediately applied with some obvious changes in notation. For the actual implementation of the reversible jump MCMC algorithm we additionally employed the auxiliary variable (AV) method of Brooks, Giudici, and Roberts (2003, Section 9) for the dimension changing steps (split-combine and birth-death moves).

For the regression part of the model, each iteration of the MCMC is conducted using the Gibbs sampler (Geman and Geman, 1984). The full conditional distributions needed to implement the Gibbs sampler are given below. The notation $| \ldots$ indicates that conditioning is done on all remaining parameters.

4.1 True log-event times $y_{i,l}$

The full conditional distribution of each $y_{i,l}$ is a truncated normal, i.e.

$$y_{i,l} \mid \cdots \sim \mathrm{N}(\mu_{r_{i,l}} + \boldsymbol{\beta}^T \boldsymbol{x}_{i,l} + \boldsymbol{b}_i^T \boldsymbol{z}_{i,l},\ \sigma^2_{r_{i,l}}) \text{ truncated on } (y^L_{i,l}, y^U_{i,l}]. \tag{7}$$

4.2 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 hyper-parameters $\boldsymbol{\nu}_{\beta}$ and $\boldsymbol{\psi}_{\beta}$, respectively. Finally, let $\Psi_{\beta(S)} = \text{diag}(\boldsymbol{\psi}_{\beta(S)})$. Then

$$\boldsymbol{\beta}_{(S)} \mid \cdots \sim \text{N}\Big(E[\boldsymbol{\beta}_{(S)} \mid \cdots], \text{var}[\boldsymbol{\beta}_{(S)} \mid \cdots]\Big), \tag{8}$$

$$\text{with} \quad \text{var}[\boldsymbol{\beta}_{(S)} \mid \cdots] = \Big(\Psi^{-1}_{\beta(S)} + \sum_{i=1}^{N}\sum_{l=1}^{n_i} \sigma^{-2}_{r_{i,l}} \boldsymbol{x}_{i,l(S)} \boldsymbol{x}^T_{i,l(S)}\Big)^{-1},$$

$$E[\boldsymbol{\beta}_{(S)} \mid \cdots] = \text{var}[\boldsymbol{\beta}_{(S)} \mid \cdots] \times \Big\{\Psi^{-1}_{\beta(S)} \boldsymbol{\nu}_{\beta(S)} + \sum_{i=1}^{N}\sum_{l=1}^{n_i} \sigma^{-2}_{r_{i,l}} \boldsymbol{x}_{i,l(S)} e^{(F)}_{i,l(S)}\Big\},$$

where $e^{(F)}_{i,l(S)} = y_{i,l} - \mu_{r_{i,l}} - \boldsymbol{\beta}^T_{(-S)} \boldsymbol{x}_{i,l(-S)} - \boldsymbol{b}^T_i \boldsymbol{z}_{i,l}$.

4.3 Means of random effects $\boldsymbol{\gamma}$

There is no loss of generality to assume that $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^T_{(S)}, \boldsymbol{\gamma}^T_{(-S)})^T$. 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 $\Psi_{\gamma(S)} = \text{diag}(\boldsymbol{\psi}_{\gamma(S)})$. Furthermore, let the inversion of the matrix $\mathbb{D}$ be decomposed in the following way

$$\mathbb{D}^{-1} = \begin{pmatrix} \mathbb{V}_{(S)} & \mathbb{V}_{(S,-S)} \\ \mathbb{V}^T_{(S,-S)} & \mathbb{V}_{(-S)} \end{pmatrix}, \tag{9}$$

then

$$\boldsymbol{\gamma}_{(S)} \mid \cdots \sim \text{N}\Big(E[\boldsymbol{\gamma}_{(S)} \mid \cdots], \text{var}[\boldsymbol{\gamma}_{(S)} \mid \cdots]\Big), \tag{10}$$

with

$$\text{var}[\boldsymbol{\gamma}_{(S)} \mid \cdots] = \Big(\Psi^{-1}_{\gamma(S)} + N\, \mathbb{V}_{(S)}\Big)^{-1},$$

$$E[\boldsymbol{\gamma}_{(S)} \mid \cdots] = \text{var}[\boldsymbol{\gamma}_{(S)} \mid \cdots] \times \Big\{\Psi^{-1}_{\gamma(S)} \boldsymbol{\nu}_{\gamma(S)} + \mathbb{V}_{(S)} \sum_{i=1}^{N} \boldsymbol{b}_{i(S)} + \mathbb{V}_{(S,-S)} \sum_{i=1}^{N} \big(\boldsymbol{b}_{i(-S)} - \boldsymbol{\gamma}_{(-S)}\big)\Big\}.$$

4.4 Random effects $\boldsymbol{b}_i$

For the random effects vectors $\boldsymbol{b}_i$:

$$\boldsymbol{b}_i \mid \cdots \sim \text{N}\Big(E[\boldsymbol{b}_i \mid \cdots], \text{var}[\boldsymbol{b}_i \mid \cdots]\Big), \qquad i = 1, \ldots, N, \tag{11}$$

$$\text{with} \quad \text{var}[\boldsymbol{b}_i \mid \cdots] = \Big(\mathbb{D}^{-1} + \sum_{i=1}^{N}\sum_{l=1}^{n_i} \sigma^{-2}_{r_{i,l}} \boldsymbol{z}_{i,l}\boldsymbol{z}^T_{i,l}\Big)^{-1},$$

$$E[\boldsymbol{b}_i \mid \cdots] = \text{var}[\boldsymbol{b}_i \mid \cdots] \times \Big\{\mathbb{D}^{-1}\boldsymbol{\gamma} + \sum_{i=1}^{N}\sum_{l=1}^{n_i} \sigma^{-2}_{r_{i,l}} \boldsymbol{z}_{i,l}(y_{i,l} - \mu_{r_{i,l}} - \boldsymbol{\beta}^T\boldsymbol{x}_{i,l})\Big\}.$$

4.5 Covariance matrix of random effects $\mathbb{D}$

Finally, $\mathbb{D} \mid \cdots$ is an inverse-Wishart distribution with degrees of freedom equal to $\tau + N$ and a scale matrix equal to $\mathbb{S} + \sum_{i=1}^{N}(\boldsymbol{b}_i - \boldsymbol{\gamma})(\boldsymbol{b}_i - \boldsymbol{\gamma})^T$.

4.6 Software

Programs in C++ have been written with an interface to the R language (R Development Core Team, 2005) as a contributed package `bayesSurv` and can be downloaded together with a comprehensive description of how to perform analyzes presented in this paper from the Comprehensive R Archive Network (CRAN) on `http://www.R-project.org`.

5 Simulation study

A simulation study was carried out to explore the performance of the proposed method. The setting mimics a typical multicenter study with a possible center by treatment interaction. 'True' uncensored data were generated according to the MEAFT model

$$\log(T_{i,l}) = 1.5 + \beta\, x_{i,l} + b_{i,1} + b_{i,2}\, z_{i,l} + \varepsilon_{i,l}, \quad i = 1, \ldots, N,\ l = 1, \ldots, n_i, \tag{12}$$

where $\beta = 0.4$, $(b_{i,1},\, b_{i,2})' \sim \mathcal{N}_1\big((0, \gamma)', \mathbb{D}\big)$, $\gamma = -0.8$, $\text{var}(b_{i,1}) = 0.5^2$, $\text{var}(b_{i,2}) = 0.1^2$, $\text{corr}(b_{i,1}, b_{i,2}) = 0.4$. 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 AIDS dataset analyzed by Komárek, Lesaffre, and

Hilton (2005). The covariate $z_{i,l}$ (*treatment vs. placebo*) – 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 $\mathcal{N}_1(7,\, 1)$, times between each consecutive assessments from $\mathcal{N}_1(6, 0.25)$). 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 mixture MEAFT model, using the Bayesian MEAFT model with a normal error and using the maximum-likelihood AFT model with a normal error and ignoring the random effects structure.

Table 1 shows the simulation results, i.e. average estimates of the regression parameters and their mean squared errors. 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 survivor functions are reproduced closely which is not always the case when the Bayesian normal model is used. Figure 2 shows this behaviour of our model on hazard functions of two heavy-tailed distributions estimated with the small sample size using either the Bayesian mixture or the Bayesian normal model. A similar comparison for the survivor functions of the extreme value or mixture error is shown in Figure 3. Figure 4 shows the behaviour of the Bayesian mixture method when the sample size increases for the extreme

value distribution and mixture distribution.

[Table 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

6 Analysis of Signal Tandmobiel® data

The first research question outlined in the introduction was already considered by Lesaffre, Komárek, and Declerck (2005) who analyzed each tooth separately using the penalized AFT model of Komárek et al. (2005). With the Bayesian MEAFT model of this paper we will analyze all teeth jointly and will be able to answer also the second research question. A random sample of 500 boys and 500 girls will be used for the inference.

For a better fit, we shifted the time origin of the MEAFT model to 5 years of age, i.e. by replacing $T_{i,l}$ by $T_{i,l} - 5$ in the model (1). The random effect vector $\boldsymbol{b}_i = (b_{i,1}, \ldots, b_{i,4})'$ with $\boldsymbol{z}_{i,l} = (1, man4_{i,l}, max5_{i,l}, man5_{i,l})'$ where $man4_{i,l}, max5_{i,l}, man5_{i,l}$, 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 is assumed in the model (1). With such model specification, apart of the random variation given by the error term $\varepsilon_{i,l}$, the terms $d_{i,max4} = b_{i,1}$, $d_{i,man4} = b_{i,1} + b_{i,2}$, $d_{i,max5} = b_{i,1} + b_{i,3}$, $d_{i,man5} = b_{i,1} + b_{i,4}$ determine how the log-emergence time of a pair of horizontally symmetric teeth of a single child differ from the population average. As the fixed effects we used $gender \equiv girl$, $dmft$ and all two-way interaction terms between $girl$, $dmft$ and dummies for the pairs of horizontal symmetric teeth.

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 3.4 we used the following values of hyper-parameters: $\xi = 1.8$, $\kappa = (3 \cdot 0.25)^2$, $\zeta = 2$, $g = 0.2$, $h = 0.1$, $\delta = 1$. 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 $\mathrm{N}(0, 100)$ prior. For the covariance matrix $\mathbb{D}$ of random effects we used an inverse Wishart prior with $\tau = 4$ which is a minimal possible value for prior degrees of freedom. 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$(\tau, 0.001)$ in the univariate case).

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. The first 1 500 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).

6.1 Regression parameters

In this analysis, the main interest lies in the effect of $dmft$ on emergence. This can be evaluated from Table 2 that shows posterior summary statistics for the effect of $dmft > 0$ (appropriate linear combinations of β parameters) for the two genders and the four pairs of horizontally symmetric teeth. As point estimate we report the posterior median which can easily be obtained from an MCMC sample and still corresponds roughly to the maximum-likelihood estimate, the most frequently used estimate in the classical statistics. Indeed, if the (log-)posterior distribution is unimodal and symmetric (which happened for practically all regression parameters)

then the posterior median is the same as the posterior mode. For skewed posterior distributions (the variance components), the log-posterior median is practically the same as the log-posterior mode.

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 $dmft$ the emergence process for girls is ahead of boys.

[Table 2 about here.]

6.2 Predictive emergence curves

Predictive cumulative distribution functions (cdf) are preferred in dentistry over the survivor functions in the case of emergence and they are known as *emergence curves*. Let $\boldsymbol{\theta}$ denote all unknown quantities of the model. For a specific value of covariates, say $\boldsymbol{x}_{new}$ and $\boldsymbol{z}_{new}$, the predictive cdf is given by

$$F(t \mid \text{data}) = \int F(t \mid \boldsymbol{\theta}, \text{data})\, p(\boldsymbol{\theta} \mid \text{data})\, d\boldsymbol{\theta}$$

for any $t > 0$. Further

$$F(t \mid \boldsymbol{\theta}, \text{data}) = F(t \mid \boldsymbol{\theta}) = \sum_{j=1}^{k} w_j \Phi\big\{\log(t) - \boldsymbol{\beta}^T \boldsymbol{x}_{new} - \boldsymbol{b}^T \boldsymbol{z}_{new} \;\big|\; \mu_j, \sigma_j^2\big\},$$

where $\Phi(\cdot \mid \mu_j, \sigma_j^2)$ is a cumulative distribution function of $\mathcal{N}_1(\mu_j, \sigma_j^2)$. The MCMC estimate of the predictive cdf is then given by $\hat{F}(t \mid \text{data}) = M^{-1} \sum_{m=1}^{M} F(t \mid \boldsymbol{\theta}^{(m)})$, 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 $\mathcal{N}_q(\boldsymbol{\gamma}^{(m)}, \mathbb{D}^{(m)})$. Predictive survivor or hazard curves can be obtained in an analogous manner.

Predictive emergence curves for the maxillary first premolar are shown in Figure 5. As a model check, Figure 5 shows also non-parametric estimates of the emergence curves computed separately in each group using the classical method of Turnbull (1976).

[Figure 5 about here.]

6.3 Inter-teeth relationship

Finally, Table 3 shows posterior summary statistics for variances and correlations of above defined tooth-specific linear combinations $d_{i,max4}$, $d_{i,man4}$, $d_{i,max5}$, $d_{i,man5}$ 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 means of all variance parameters in Table 3 are all about 0.2 which is much higher than the posterior mean of the variance of the error distribution which was equal to 0.01. Posterior means of all correlation parameters lie between 0.79 and 0.91.

[Table 3 about here.]

7 Discussion

We have proposed a Bayesian 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 credibility intervals of model parameters as well as predictive survivor or hazard curves.

Observe that the Bayesian approach is used here mainly for technical convenience. Indeed, in practice likelihood (2) 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. E.g., in multicenter studies, the vector $\boldsymbol{z}_{i,l} = (1, treatment_{i,l})'$ in the model formula (1) allows to consider not only the random center effect but also a random center-by-treatment interaction which can sometimes be substantial.

According to our best knowledge, no approach is available which tackles this complex data structure. With varying amount of effort some of the existing semi-parametric approaches, e.g. Pan and Louis (2000); Pan and Connett (2001), mentioned in Section 2, could be, of course, extended to handle regression with correlated interval-censored data.

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. E.g., 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 Komárek and

Lesaffre (2005) to handle also doubly-interval-censored data, i.e. the data where the response is given as the difference of two interval-censored observations.

Acknowledgments

The first author was financially supported by Research Grant DB/04/031, Katholieke Universiteit Leuven. The authors further acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs.

Data collection for the example in Section 6 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).

References

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.

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.

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

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.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–499.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, Boca Raton.

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.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics*, **56**, 1139–1144.

Green, P. J. (1995). Reversible jump Markov chain computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, **55**, 13–22.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition*. John Wiley & Sons, Chichester.

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.

Komárek, A. and Lesaffre, E. (2005). Bayesian semiparametric regression model for bivariate doubly-interval-censored data. In Francis, A. R., Matawie, K. M., Oshlack, A., and Smyth, G. K., editors, *Statistical Solutions to Modern Problems: Proceedings of the 20th International Workshop on Statistical Modelling*, pages 317–324, Sydney.

Komárek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *To appear in Journal of Computational and Graphical Statistics*.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics*, **53**, 1485–1494.

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.

Leroy, R., Bogaerts, K., Lesaffre, E., and Declerck, D. (2003). The emergence of permanent teeth in Flemish children (Belgium). *Community Dentistry and Oral Epidemiology*, **31**, 30–39.

Lesaffre, E., Komárek, A., and Declerck, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry. *To appear in Statistical Methods in Medical Research*.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York.

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 Louis, T. A. (2000). A linear mixed-effects model for multivariate censored data. *Biometrics*, **56**, 160–166.

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.

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.

Roeder, K. and Wasserman (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester.

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.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309–3324.

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.

Table 1: Simulation study. Results for the regression parameters: average estimate, mean squared error ($\times 10^{-4}$) in brackets.

Setting	Bayesian mixture	Bayesian normal	ML, no random eff.
		$\beta = 0.4$	
Normal, small	0.3966 (26.3)	0.3973 (25.6)	0.3992 (34.7)
large	0.4016 (7.3)	0.4018 (7.3)	0.4022 (9.0)
Cauchy t_1, small	0.4122 (52.5)	0.3832 (68.2)	0.3783 (124.6)
large	0.3921 (13.2)	0.3608 (41.8)	0.3571 (50.7)
Student t_2, small	0.3933 (58.1)	0.3859 (49.6)	0.3823 (72.9)
large	0.3944 (11.5)	0.3794 (19.0)	0.3780 (21.6)
Extr. value, small	0.3928 (17.9)	0.3954 (20.9)	0.3952 (25.9)
large	0.4036 (4.3)	0.4035 (5.4)	0.4022 (6.8)
Mixture, small	0.3942 (17.6)	0.4324 (68.1)	0.4436 (127.3)
large	0.3997 (3.5)	0.4480 (45.6)	0.4505 (52.9)
		$\gamma = -0.8$	
Normal, small	-0.8128 (240.4)	-0.8105 (222.8)	-0.8121 (235.7)
large	-0.7981 (47.0)	-0.7983 (48.2)	-0.7982 (60.4)
Cauchy t_1, small	-0.7656 (512.7)	-0.7192 (716.4)	-0.7210 (704.0)
large	-0.8097 (107.2)	-0.7360 (234.5)	-0.7383 (238.9)
Student t_2, small	-0.7777 (479.0)	-0.7593 (401.3)	-0.7614 (415.5)
large	-0.7933 (99.9)	-0.7610 (123.7)	-0.7601 (132.1)
Extr. value, small	-0.8150 (191.1)	-0.8106 (192.4)	-0.8094 (202.3)
large	-0.7969 (47.4)	-0.7999 (56.6)	-0.8022 (66.9)
Mixture, small	-0.7868 (95.7)	-0.8693 (895.0)	-0.8635 (840.5)
large	-0.8040 (26.6)	-0.9264 (366.8)	-0.9227 (369.4)

Table 2: Signal Tandmobiel® data. Posterior medians, 95% equal-tail credibility intervals and Bayesian two-sided p-values for the effect of $dmft > 0$ for the two genders and different teeth.

maxilla 4		*maxilla* 5	
girl	*boy*	*girl*	*boy*
-0.0352	-0.0457	-0.0212	-0.0317
$(-0.0522,\ -0.0185)$	$(-0.0631,\ -0.0284)$	$(-0.0390,\ -0.0035)$	$(-0.0500,\ -0.0135)$
$p < 0.001$	$p < 0.001$	$p = 0.019$	$p = 0.001$
mandible 4		*mandible* 5	
girl	*boy*	*girl*	*boy*
-0.0098	-0.0201	0.0015	-0.0090
$(-0.0267,\ 0.0070)$	$(-0.0378,\ -0.0032)$	$(-0.0162,\ 0.0193)$	$(-0.0283,\ 0.0098)$
$p = 0.255$	$p = 0.021$	$p = 0.870$	$p = 0.353$

Table 3: Signal Tandmobiel® data. Posterior medians, 95% equal-tail credibility intervals for variances and correlations between tooth-specific linear combinations of random effects.

$\mathrm{var}(d_{max4})$	$\mathrm{var}(d_{man4})$	$\mathrm{var}(d_{max5})$	$\mathrm{var}(d_{man5})$
0.042	0.039	0.042	0.041
$(0.037,\ 0.047)$	$(0.035,\ 0.045)$	$(0.036,\ 0.049)$	$(0.035,\ 0.048)$
$\mathrm{cor}(d_{max4}, d_{man4})$	$\mathrm{cor}(d_{max4}, d_{max5})$	$\mathrm{cor}(d_{max4}, d_{man5})$	$\mathrm{cor}(d_{man4}, d_{max5})$
0.887	0.914	0.842	0.793
$(0.856,\ 0.914)$	$(0.887,\ 0.938)$	$(0.804,\ 0.874)$	$(0.749,\ 0.832)$
$\mathrm{cor}(d_{man4}, d_{man5})$	$\mathrm{cor}(d_{max5}, d_{man5})$		
0.895	0.847		
$(0.864,\ 0.923)$	$(0.810,\ 0.880)$		

Figure 1: DAG for the Bayesian AFT model.

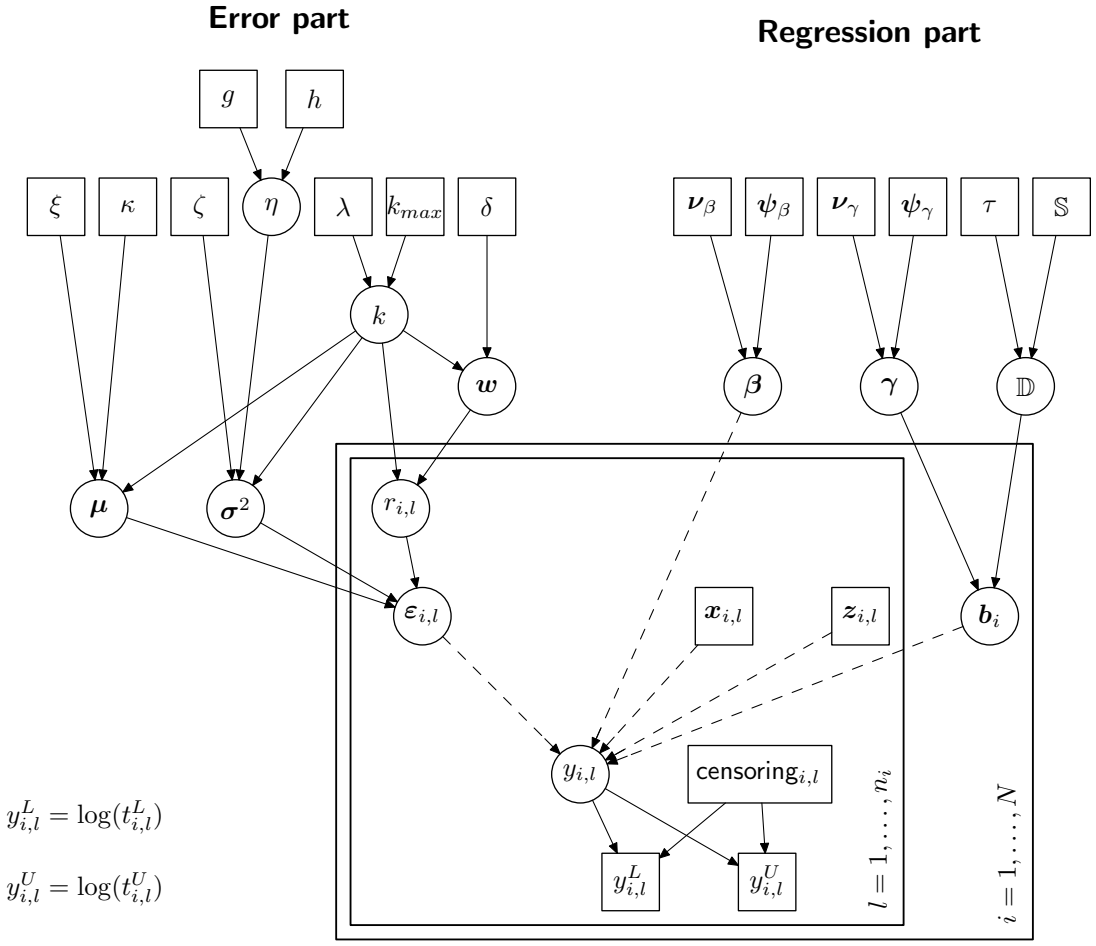

Figure 2: Simulation study – hazard functions. Results for a small sample size ($N = 50$, $n_i = 5$) and Cauchy and Student t_2 error distributions. Left column: estimates based on the Bayesian mixture model, right column: estimates based on the Bayesian normal model. Solid line: Estimate of the hazard function for $x = 8.13$ – median value and $z = 0$, gray region: simulation based 95% point-wise confidence interval, dashed line: true hazard function.

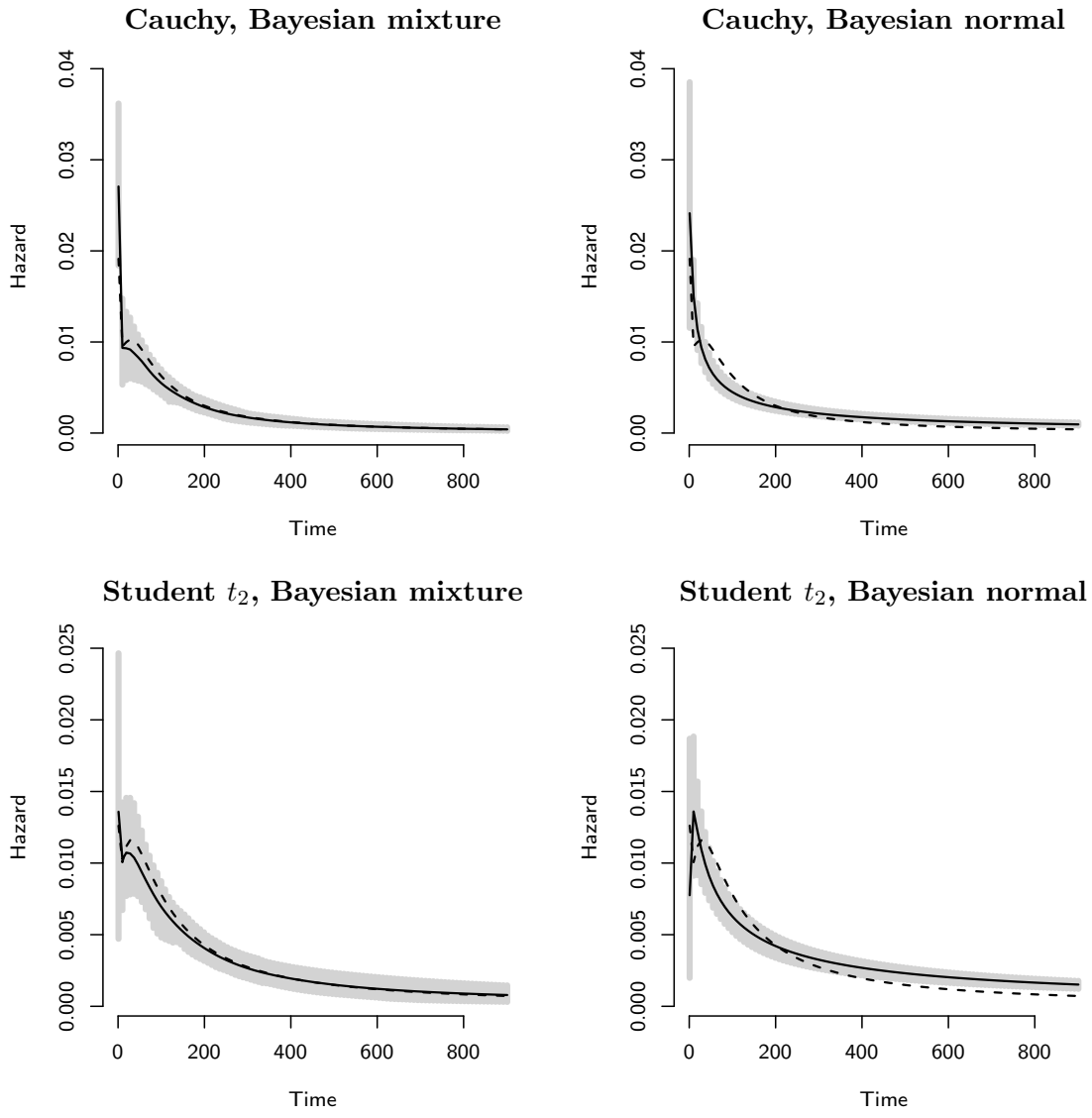

Figure 3: Simulation study – survivor functions. Results for a small sample size ($N = 50$, $n_i = 5$) and extreme value and normal mixture error distributions. Left column: estimates based on the Bayesian mixture model, right column: estimates based on the Bayesian normal model. Solid line: estimate of the survivor function for $x = 8.13$ – median value and $z = 0$, gray region: simulation based 95% point-wise confidence interval, dashed line: true survivor function.

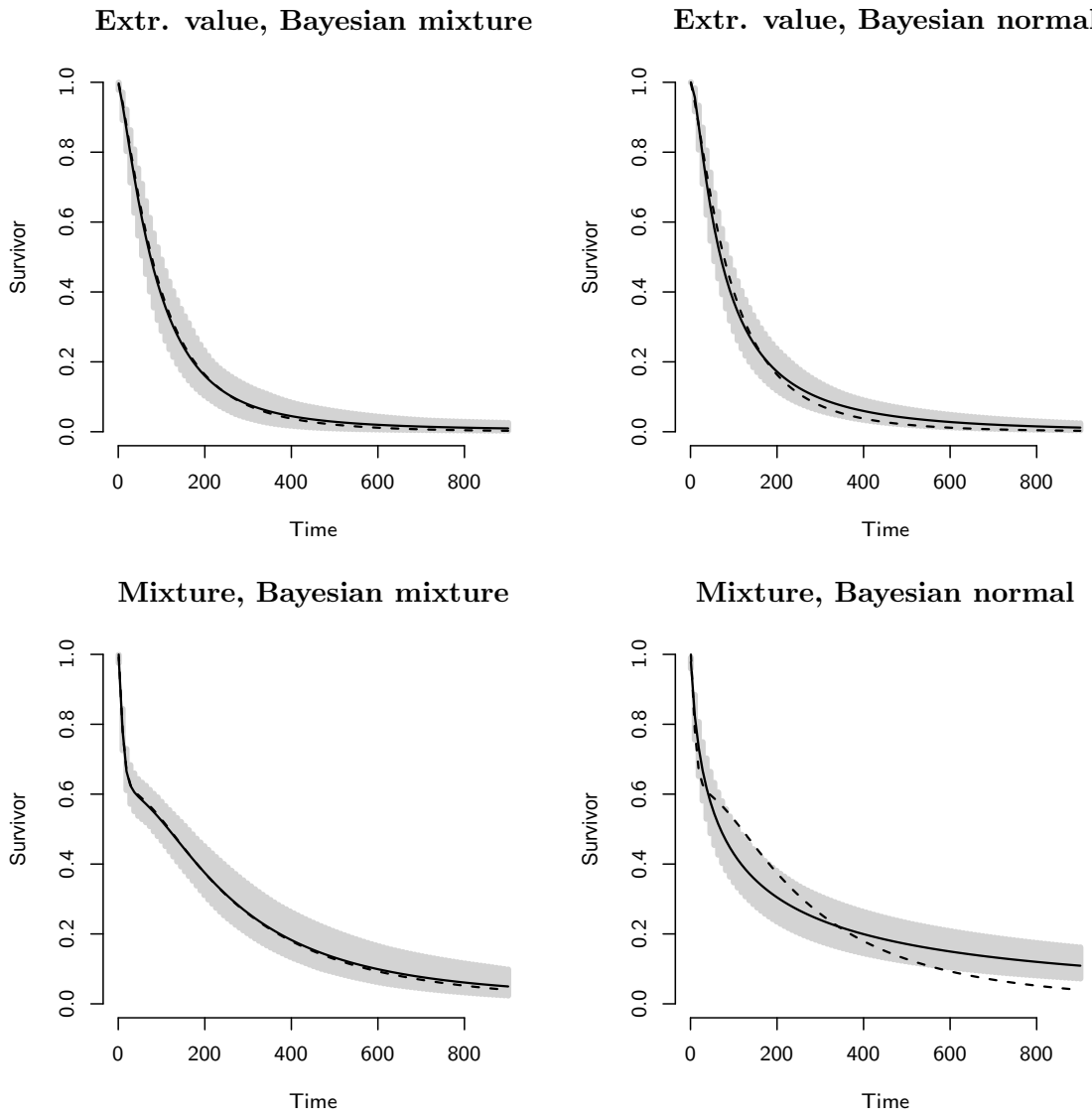

Figure 4: Simulation study – error densities. Results for both sample sizes and extreme value and normal mixture error distributions, estimates based on the Bayesian mixture model. Solid line: estimate of the standardized error density, gray region: simulation based 95% point-wise confidence interval, dashed line: true standardized error density.

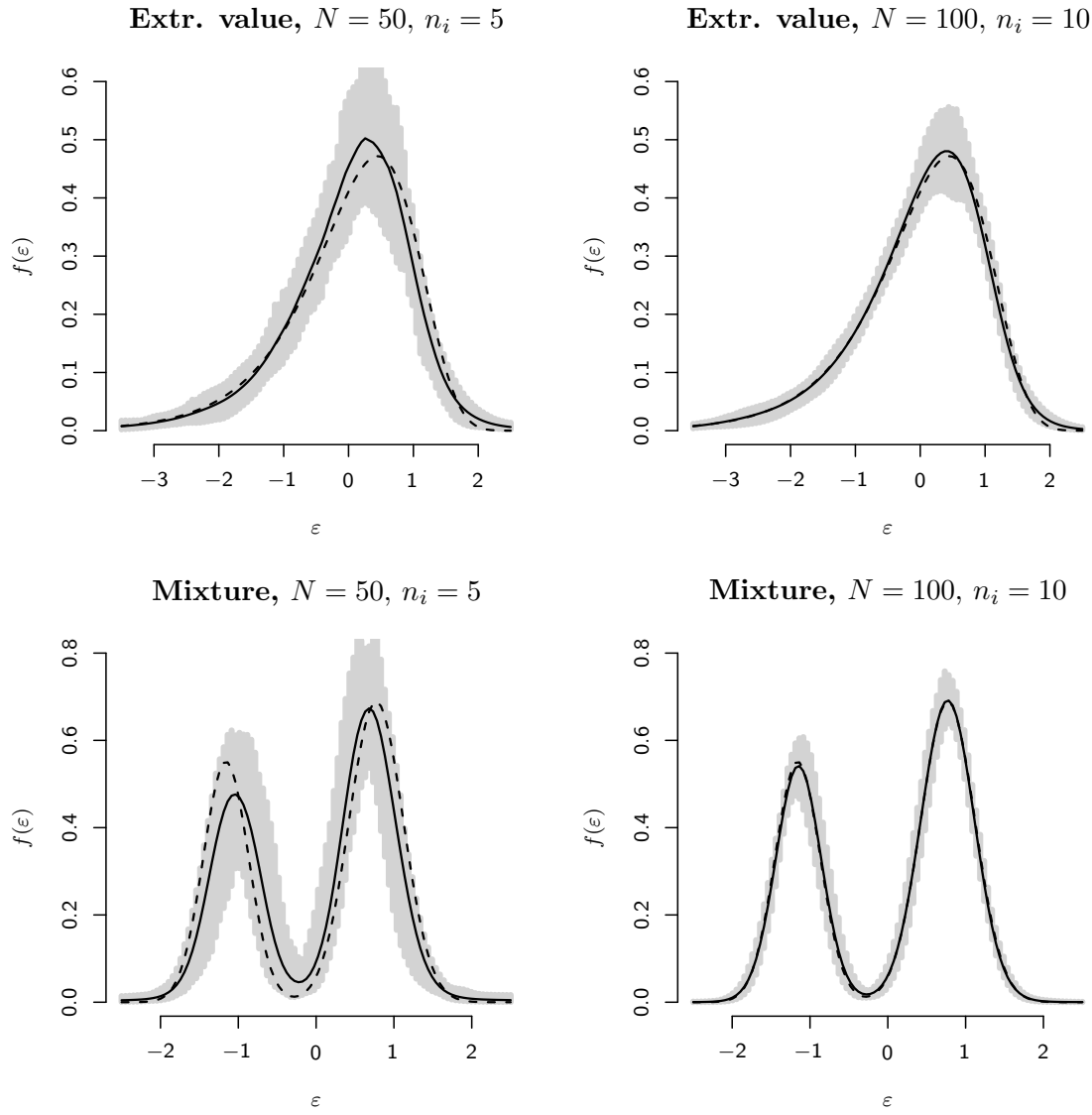

Figure 5: Signal Tandmobiel® data. Predictive emergence curves (solid lines) compared to the non-parametric estimate of Turnbull (dashed lines) for maxillary first premolars.