Incorporating unobserved heterogeneity in Weibull survival models: A Bayesian approach

Catalina A. Vallejos

The Alan Turing Institute
British Library, 96 Euston Road, London, NW1 2DB, UK

Department of Statistical Science, University College London
1-19 Torrington Place, London, WC1E 7HB

Mark F.J. Steel*

Department of Statistics, University of Warwick
Coventry, CV4 7AL, UK

Abstract

Outlying observations and other forms of unobserved heterogeneity can distort inference for survival datasets. The family of Rate Mixtures of Weibull distributions includes subject-level frailty terms as a solution to this issue. With a parametric mixing distribution assigned to the frailties, this family generates flexible hazard functions. Covariates are introduced via an Accelerated Failure Time specification for which the interpretation of the regression coefficients does not depend on the choice of mixing distribution. A weakly informative prior is proposed by combining the structure of the Jeffreys prior with a proper prior on some model parameters. This improper prior is shown to lead to a proper posterior distribution under easily satisfied conditions. By eliciting the proper component of the prior through the coefficient of variation of the survival times, prior information is matched for different mixing distributions. Posterior infer-

[☆] Supplementary material is provided for this manuscript: all proofs of theoretical results (Section A); further details regarding the implementation of Bayesian inference (Section B); a simulation study (Section C) and code used throughout the case studies (Section D). Code to implement RMW-AFT regression models is provided as an R library freely available at `https://github.com/catavallejos/RMWreg`

[*] Corresponding author.

Email address: `M.F.Steel@stats.warwick.ac.uk` (Mark F.J. Steel)

ence on subject-level frailty terms is exploited as a tool for outlier detection. Finally, the proposed methodology is illustrated using two real datasets, one concerning bone marrow transplants and another on cerebral palsy.

Keywords: Survival analysis, Frailty model, Robust modelling, Outlier detection, Posterior existence

1. Introduction

Outlying observations and other forms of unobserved heterogeneity can distort inference for survival datasets. For instance, the popular Proportional Hazards (PH) assumption can be violated in the presence of unobserved confounders [1]. We explore the use of subject-level frailty terms as a natural solution to this critical issue, extending standard survival models through random effects, using an arbitrary (parametric) mixing distribution. These models can be represented as an infinite mixture of survival distributions with density

$$f(t_i|\psi,\theta) \equiv \int_{\mathbb{R}_+} f(t_i|\psi,\Lambda_i=\lambda_i)\, dP_{\Lambda_i}(\lambda_i|\theta), \qquad i \in \{1,\ldots,n\}, \tag{1}$$

where t_i is the observed time for subject i and the underlying $f(\cdot|\psi,\Lambda_i=\lambda_i)$ is a "standard" life-time density indexed by ψ and λ_i. In (1), λ_i is a subject-specific *frailty* and the spread of the mixing distribution $P_{\Lambda_i}(\cdot|\theta)$ controls the strength of the unobserved heterogeneity. Individual frailties are a powerful tool to robustify standard survival models in an intuitive manner. However, frailty models are also widely used in other contexts. For example, *shared frailty models* [2, 3, 4] assume common frailty values for groups of subjects to account correlations between clustered individuals (e.g. patients treated at the same hospital).

Varying the underlying model generates a wide class of distributions. Some examples explored in previous literature are mixtures of Birnbaum-Saunders distributions [5] and mixtures of log-normal distributions [6]. Here, we present the family of Rate Mixtures of Weibull (RMW) distributions, introducing the frailty via the rate parameter. This family accommodates flexible hazard shapes and contains i.a. the Lomax distribution, which is widely used as a heavy-tailed model. As an alternative to the mixed PH model, developed in econometrics [7], we introduce covariates via an Accelerated Failure Time (AFT) specification for which the interpretation of the regression coefficients is robust to the choice of mixing distribution. We derive a weakly informative improper prior distribution, combining the structure of the Jeffreys prior with a proper (informative) prior for some model parameters. The latter can be adapted to any mixing distribution by eliciting a unique prior on the coefficient of variation of the survival times. Mild and easily verified conditions for posterior existence are also derived and the appropriateness of different mixing distributions is assessed using standard Bayesian model comparison methods. Our modelling approach mitigates the effect of extreme observations and posterior inference on frailty terms leads to an intuitive outlier detection tool.

Section 2 introduces the RMW family, some of its properties and a regression model based on an AFT specification. Section 3 includes an extensive analysis of Bayesian inference for these regression models, allowing for right censored observations. In Section 4, our methods are illustrated using two real datasets, one concerning bone marrow transplants and another on cerebral palsy. Finally, Section 5 concludes.

Supplementary material is provided for this manuscript. This contains all proofs of theoretical results (Section A), further details regarding the implementation of Bayesian inference (Section B), a simulation study (Section C) and the code used throughout the case studies (Section D). The latter also includes traceplots and other convergence diagnostics for all the associated MCMC chains. Code to implement RMW-AFT regression models is provided as an R library freely available at `https://github.com/catavallejos/RMWreg`.

2. Mixtures of survival distributions

Survival models as in (1) are a simple and intuitive extension to standard survival models. In particular, inference is more robust to outlying observations, reducing the need of discarding anomalous records. In addition, if the underlying model is supported by theoretical or practical reasons, this intuition is preserved by the mixture. For example, if theory suggests that individuals have a constant hazard rate, an exponential model is appropriate. Using mixtures of exponential distributions leads to a decreasing hazard rate, yet does not contradict this theory. In such a case, the individual-specific hazards (λ_i) remain constant over time but high-risk subjects will tend to die earlier, so that a higher proportion of low-risk subjects is left to be observed at later times.

2.1. Rate Mixtures of Weibull distributions

The popularity of the Weibull distribution is partly explained by its flexibility, allowing for increasing, decreasing and non-monotonic hazard rates. However, if neglected, unobserved heterogeneity can lead to a biased estimation of the subject-level hazard rate [1]. Let T_i be a positive-valued random variable distributed as a Rate Mixture of Weibull (RMW) distributions. A hierarchical representation of this model, with $\alpha, \gamma > 0$ and $\theta \in \Theta$, is given by

$$T_i|\alpha, \gamma, \Lambda_i = \lambda_i \sim \text{Weibull}\left(\alpha\lambda_i, \gamma\right), \;\; \Lambda_i|\theta \sim P_{\Lambda_i}(\cdot|\theta) \;\text{ with support on } \mathbb{R}_+. \quad (2)$$

Denote this by $T_i \sim \text{RMW}_P(\alpha, \gamma, \theta)$. Alternatively, following (1), (2) can be re-written as

$$f(t_i|\alpha, \gamma, \theta) = \int_0^\infty \gamma\alpha\lambda_i t_i^{\gamma-1}\, e^{-\alpha\lambda_i t_i^\gamma}\, dP_{\Lambda_i}(\lambda_i|\theta), \quad (3)$$

If $\gamma \leq 1$, the hazard rate induced by the mixture decreases regardless of the mixing distribution [8]. For $\gamma > 1$, it has a more flexible shape and can accommodate non-monotonic behaviour. This formulation uses an arbitrary (parametric) mixing distribution. However, identifiability restrictions are required (Theorem 1). In particular, the use of (separate) unknown scale parameters for the mixing distribution is precluded. This is achieved by either fixing its scale parameter or by fixing $\text{E}(\Lambda_i|\theta) = 1$ (as in [7]). We use the latter for gamma mixing, as it leads to better properties of the MCMC sampler described in this article. For the other mixtures explored here, the sampler performs better if we fix the scale of the mixing distribution.

*Theorem 1. Let $T_i \sim RMW_P(\alpha, \gamma, \theta)$. (α, γ, θ) is identified by the distribution of T_i if and only if: **(i)** $E(\Lambda_i|\theta)$ is finite and **(ii)** (α, θ) is identified by the distribution of $\alpha\Lambda_i$.*

Proof. See Section A of the Supplementary Material. □

Special RMW cases appear in the existing literature, where often γ is fixed at 1 and the mixing parameters are gamma distributed [9, 10]. We refer to the case with $\gamma = 1$ as the Rate Mixtures of Exponentials (RME) family (denoted by $T_i \sim \text{RME}_P(\alpha, \theta)$). This case extends to the RMW family via a power transformation (if $T_i \sim \text{RME}_P(\alpha, \theta)$ then $T_i^{1/\gamma} \sim \text{RMW}_P(\alpha, \gamma, \theta)$).

Result 1. Provided all following expressions exist, the coefficient of variation (i.e. the ratio between standard deviation and expectation) of the survival distributions in (3) is

$$cv(\gamma, \theta) = \sqrt{\frac{\Gamma\left(1+2/\gamma\right)}{\Gamma^2\left(1+1/\gamma\right)} \underbrace{\frac{Var_{\Lambda_i}(\Lambda_i^{-1/\gamma}|\theta)}{E^2_{\Lambda_i}(\Lambda_i^{-1/\gamma}|\theta)}}_{(cv^*(\gamma,\theta))^2} + \underbrace{\frac{\left[\Gamma\left(1+2/\gamma\right) - \Gamma^2\left(1+1/\gamma\right)\right]}{\Gamma^2\left(1+1/\gamma\right)}}_{(cv^W(\gamma))^2}}. \quad (4)$$

Proof. See Section A of the Supplementary Material. □

The expression in (4) simplifies to $\sqrt{2\frac{\text{Var}_{\Lambda_i}(\Lambda_i^{-1}|\theta)}{\text{E}^2_{\Lambda_i}(\Lambda_i^{-1}|\theta)} + 1}$ when $\gamma = 1$. Result 1 indicates that $cv(\gamma, \theta)$ is an increasing function of $cv^*(\gamma, \theta)$, the coefficient of variation of $\Lambda_i^{-1/\gamma}$ given θ. In addition, for fixed γ, the coefficient of variation of the Weibull distribution $cv^W(\gamma)$ is a lower bound for $cv(\gamma, \theta)$ and they are equal if and only if $\Lambda_i \equiv \lambda_0$, for $i = 1, \dots, n$. Therefore, evidence of unobserved heterogeneity can be quantified using

$$R_{cv}(\gamma, \theta) = \frac{cv(\gamma, \theta)}{cv^W(\gamma)}, \quad (5)$$

i.e. the inflation induced in the coefficient of variation (w.r.t. a Weibull model with the same γ). If θ is such that $cv^*(\gamma, \theta) \approx 0$, then $R_{cv}(\gamma, \theta) \approx 1$ and the mixture reduces to the underlying Weibull model. If $\gamma \to 0$, $cv^W(\gamma)$ and, consequently, $cv(\gamma,\theta)$ become unbounded. If $\gamma = 1$, then $R_{cv}(\gamma,\theta) = cv(1,\theta)$. We restrict the range of (γ,θ) such that $cv(\gamma,\theta)$ is finite (this is not required when θ does not appear), facilitating the implementation of Bayesian inference (see Section 3.1).

Heckman and Singer [11] remark that inference is sensitive to the mixing distribution and thus use non-parametric mixing. Non-parametric mixtures of Weibull distributions (mixing on both parameters) are studied by [12]. However, non-parametric mixing might not be appropriate for moderate sample sizes. To the best of our knowledge, the small sample properties of non-parametric frailty distributions together with parametric baseline models have not been systematically studied. However, among others, [13] and [14] have identified problems in the converse situation, where the baseline distribution is non-parametric (e.g. Cox proportional hazards) but the frailty distribution is within a parametric family. In particular, they found bias issues for maximum likelihood estimates with small sample sizes when combining a non-parametric baseline with a Gamma frailty. In the light of these results, we believe that using a parametric frailty distributions is a safer strategy when the sample size is small. Therefore, we opt for a fully parametric approach and the adequacy of a particular mixing distribution is evaluated using Bayesian model comparison tools. This is a compromise between the baseline model ($\Lambda_i \equiv \lambda_0$) and fully flexible non-parametric mixing.

The survival function of RMW random variables is the Laplace transform of the mixing density evaluated in αt_i^{γ} [15]. Hence, mixing densities with known Laplace transform, such as the Power Variance Function (PVF) family [16], are an attractive choice. The positive stable distribution is a limiting case of the PVF family [15] and the resultant model is the Weibull distribution itself. Other examples in this family are the gamma and the inverse Gaussian distributions. In particular, [10] gives an asymptotic argument for gamma mixing. If $\gamma = 1$, gamma$(\theta, 1)$ mixing generates the Lomax model [17], widely used as a heavy tailed distribution. Some mixing distributions (e.g. log-normal) do not lead to analytical expressions for the resulting density. In those cases, Bayesian inference can be conducted using data augmentation and the hierarchical representation (2).

Table 1: Examples in the RME family. $K_p(\cdot)$ denotes the modified Bessel function and $\Theta = (0, \infty)$, unless otherwise stated.

Mixing density	$\mathrm{E}(\Lambda_i \vert \theta)$	$f(t_i \vert \alpha, \theta)$	$h(t_i \vert \alpha, \theta)$
Exponential(1)	1	$\alpha(\alpha t_i + 1)^{-2}$	$\alpha(\alpha t_i + 1)^{-1}$
Gamma(θ, θ)	1	$\alpha([\alpha/\theta]\, t_i + 1)^{-(\theta+1)}$	$\alpha([\alpha/\theta]\, t_i + 1)^{-1}$
Inv-gamma($\theta, 1$)	$\frac{1}{\theta-1}$, $\theta > 1$	$\frac{2\alpha}{\Gamma(\theta)} K_{-(\theta-1)}(2\sqrt{\alpha t_i})(\alpha t_i)^{(\theta-1)/2}$	$\sqrt{\frac{\alpha}{t_i}} \frac{K_{-(\theta-1)}(2\sqrt{\alpha t_i})}{K_{-\theta}(2\sqrt{\alpha t_i})}$
Inv-Gauss($\theta, 1$)	θ	$\alpha\, e^{1/\theta} \left[\frac{1}{\theta^2} + 2\alpha t_i\right]^{-1/2} e^{-\left[\frac{1}{\theta^2} + 2\alpha t_i\right]^{1/2}}$	$\alpha \left[\frac{1}{\theta^2} + 2\alpha t_i\right]^{-1/2}$
Log-normal($0, \theta$)	$e^{\theta/2}$	$\frac{\alpha}{\sqrt{2\pi\theta}} \int_0^\infty e^{-\alpha\lambda_i t_i}\, e^{-\frac{(\log(\lambda_i))^2}{2\theta}}\, d\lambda_i$	No closed form

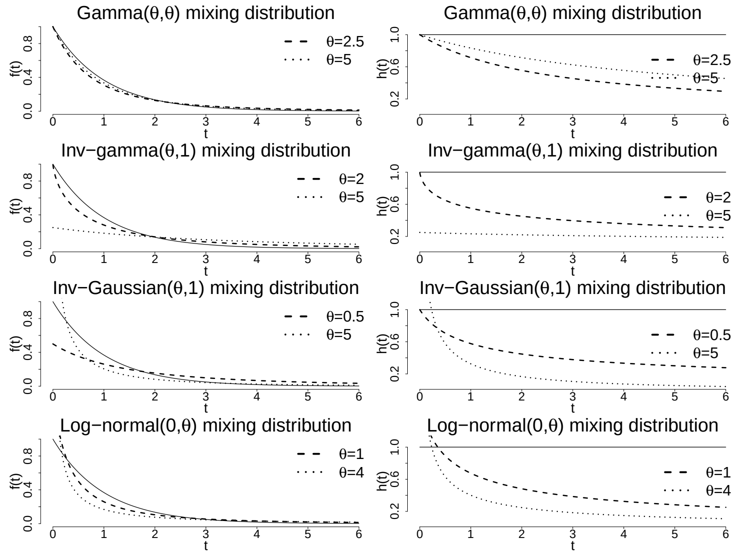

Figure 1: Density and hazard function (left and right panels, respectively) of some RME models ($\alpha = 1$). The solid line is the exponential(1) density (hazard).

Table 1 displays some RME examples and this list can be extended by selecting other mixing distributions. These examples generalize to the RMW case via the power transformation mentioned earlier. Figure 1 shows the corresponding RME densities for different values of θ. These are decreasing (like in the exponential case) but the tail behaviour is very flexible. Figure 1 also illustrates that the hazard function decreases over time but that its gradient varies among the different mixing distributions (see also [8]). Figure 2 illustrates the effect of a gamma(θ, θ) mixing (reparametrized version of the Lomax distribution) for RMW models. Whereas the shape of the density function was not greatly affected in this example, the effect on the hazard rate is more pronounced. For instance, while the hazard rate of the Weibull is an increasing function of t_i when $\gamma = 2$, the hazard of the mixture exhibits non-monotonic behaviour.

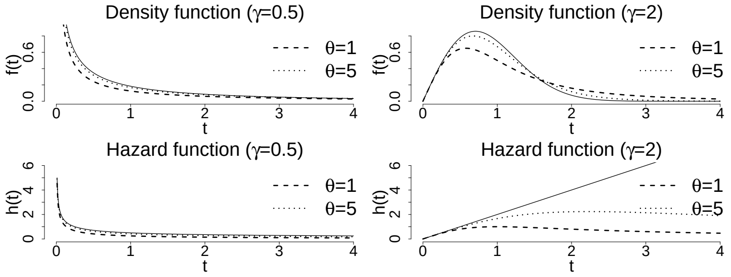

Figure 2: Some RMW models ($\alpha = 1$). The mixing distribution is gamma(θ, θ) (exponential(1) for $\theta = 1$). The solid line is the Weibull$(1, \gamma)$ density and hazard function.

2.2. A regression model for the RMW family

Let x_i be a vector of k covariates values for subject i and $\beta = (\beta_1, \dots, \beta_k)' \in \mathbb{R}^k$ be a vector of parameters. A Weibull regression can be equivalently written in terms of AFT and PH specifications, both broadly used in applied survival analysis. The RMW-AFT model is given by

$$T_i \sim RMW_P(\alpha_i, \gamma, \theta),\ \alpha_i = e^{-\gamma x_i'\beta},\ i = 1, \dots, n \text{ or equivalently,} \quad (6)$$

$$\log(T_i) = x_i'\beta + \log(\Lambda_i^{-1/\gamma} T_0), \text{ with } \Lambda_i|\theta \sim P_{\Lambda_i}(\theta), T_0|\gamma \sim \text{Weibull}(1, \gamma) (7)$$

The RMW-AFT is itself an AFT model with baseline survival function given by the distribution of $T_0' = \Lambda_i^{-1/\gamma} T_0$, where $T_0'|\theta \sim \text{RMW}_P(1, \gamma, \theta)$. In (7), e^{β_j}

represents proportional (marginal) changes in the scale of the survival times themselves, after a unit change in covariate j. This can be interpreted in terms of changes in the moments and percentiles of the survival distribution. In particular, we highlight the interpretation in terms of the median survival time, a quantify that is typically of interest in the context of survival analysis. For $\beta^* = -\gamma\beta$, (6) is equivalent to the RMW-PH model with hazard function

$$h(t_i|\beta^*, \gamma, \Lambda_i = \lambda_i; x_i) = \lambda_i \gamma t_i^{\gamma-1} \, e^{x_i'\beta^*}, \qquad \Lambda_i \sim P(\Lambda_i|\theta), \qquad i = 1, \ldots, n. \quad (8)$$

Such a model is also known as a mixed PH model which is popular in econometrics [7]. Even though the RMW-PH model is a mixture of PH models, the PH assumption is generally not preserved. Only the positive stable mixing distribution retains this property [15]. In the RMW-PH model, $e^{\beta_j^*}$ is interpreted as the proportional marginal change of the hazard rate after a unit change in covariate j at an individual level (conditional on λ_i). Unlike for the RMW-AFT model, this interpretation cannot be extended to the population level. While most of the earlier literature for unobserved heterogeneity is in terms of the PH model, here we present results in terms of the RMW-AFT presentation since the interpretation of the regression coefficients is clearer and the mixture model is still an AFT model.

3. Bayesian Inference for the RMW-AFT model

3.1. A prior distribution for the RMW-AFT model

First, we define a prior for the RME-AFT model ($\gamma = 1$). In the absence of prior information, a popular choice is a prior based on the Jeffreys rule, which require the Fisher information matrix.

Result 2. Let $T_1, \ldots, T_n$ be independent random variables distributed as in (6) with $\gamma = 1$ and define $X = (x_1 \cdots x_n)'$. The Fisher information matrix (FIM) corresponds to

$$I(\beta, \theta) = \begin{pmatrix} k_1(\theta) X'X & k_2(\theta) X'\mathbf{1}_n \\ k_2(\theta)\mathbf{1}_n' X & n k_3(\theta) \end{pmatrix}, \quad (9)$$

where $k_1(\theta), k_2(\theta)$ and $k_3(\theta)$ are functions of only θ (see Section A in the Supplementary Material) and $\mathbf{1}_n$ is a column vector of n ones.

Proof. See Section A of the Supplementary Material. □

In addition to the assumptions of Result 2, let us also assume that X has rank k and θ is a scalar parameter. The Jeffreys prior and the independence Jeffreys prior (which deals separately with the blocks for β and θ) for the RME-AFT model are then, respectively

$$\pi^J(\beta,\theta) \quad \propto \quad k_1^{k/2}(\theta)k_3^{1/2}(\theta)\left[1-\frac{k_2^2(\theta)}{nk_1(\theta)k_3(\theta)}\mathbf{1}_n' X(X'X)^{-1}X'\mathbf{1}_n\right]^{1/2}, \tag{10}$$

$$\pi^I(\beta,\theta) \quad \propto \quad k_3^{1/2}(\theta). \tag{11}$$

These two Jeffreys-style priors can be expressed as

$$\pi(\beta,\theta) \propto \pi(\theta), \tag{12}$$

where $\pi(\theta)$ only depends on θ. Although the result above gives a general structure for these priors, the actual expressions are not easily derived (even for simple mixing distributions). One alternative is to compute the FIM directly from the resultant density. For example, in the case of a gamma$(\theta, 1)$ mixing distribution the Jeffreys and independence Jeffreys priors are, respectively

$$\begin{aligned}\pi^J(\beta,\theta) \quad &\equiv \quad \pi^J(\theta) \propto \left[\frac{\theta}{\theta+2}\right]^{k/2}\frac{1}{\theta}\left[1-\frac{\theta(\theta+2)}{n(\theta+1)^2}\mathbf{1}_n' X(X'X)^{-1}X'\mathbf{1}_n\right]^{1/2} \text{ and}\\ \pi^I(\beta,\theta) \quad &\equiv \quad \pi^I(\theta) \propto \frac{1}{\theta}.\end{aligned} \tag{13}$$

Even though this is one of the simplest RME models, $\pi^J(\theta)$ is very involved, depending on k, n and X. For other mixtures, these priors become more complicated (already if we use gamma(θ,θ) mixing instead) and have no easy derivation. If the resultant distribution does not have a closed analytical form (e.g. with log-normal mixing), computing the FIM is very challenging. In addition, there is no guarantee of having a proper prior for θ when using an arbitrary mixing distribution. For instance, in the case above, $\pi^J(\theta)$ and $\pi^I(\theta)$ are not proper densities (both behave as $1/\theta$ for large θ). As the role of θ is specific to each mixture, improper priors for θ will not allow the comparison between models in the RME family using Bayes factors.

To overcome these issues, we propose a simplification of these Jeffreys-style priors. We keep the structure in (12) but assign a proper $\pi(\theta)$. To ensure the comparison between models is meaningful, $\pi(\theta)$ must reflect the same prior information regardless of the mixing distribution (i.e. the priors are "matched"). We achieve this by exploiting the relationship between θ and cv, the coefficient of variation of the survival times. A proper prior, common for all models, is then assigned to cv and denoted by $\pi^*(cv)$. As cv does not involve β (expression (4) does not involve α), $\pi^*(cv)$ only provides information about θ. Using (4), the functional relationship between cv and θ for some RME examples is derived (see Table 2). The inverse function of $cv(\theta)$ must exist ($cv(\theta)$ must be injective), yet an explicit expression is not required. Injectivity holds for all the examples in Table 2 ($cv(\theta)$ is a monotone function of θ), as illustrated by Figure 3. The induced prior for θ is then easily derived by a change of variable. When comparing a model with θ to models without θ, meaningful results derive from the fact that the prior on θ is reasonable. Two natural choices for $\pi^*(cv)$ are the truncated exponential and Pareto type I distributions (both on $(1, \infty)$) with hyper-parameters a and b, respectively. These priors cover a wide set of tails for cv. Smaller values of a and b assign larger probabilities to small values of cv (we restrict $b > 1$ in order to have a finite expectation for cv). These hyper-parameters can be elicited e.g. using the mean of cv. The expected values under these priors are $1 + 1/a$ and $b/(b-1)$ respectively, which are equal for $b = a + 1$. When the range of cv differs from $(1, \infty)$ (e.g. with an inverse gamma and inverse Gaussian mixing distribution), these priors can be adjusted by truncating $\pi^*(cv)$. If the values of a and b are such that the prior expectation of cv falls outside the range allowed by a specific model, the prior is deemed to be inconsistent with that model. For example, the RME model with inverse Gaussian mixing should be discarded a priori if $a < (\sqrt{5} - 1)^{-1}$ and $b < 1 + (\sqrt{5} - 1)^{-1}$.

For a general RMW-AFT model (unknown γ), the structure of the FIM is

Table 2: Relationship between cv and θ for some RME models. $\Theta = (0, \infty)$, unless otherwise stated.

Mixing density	Range of cv	$cv(\theta)$	$\left\lvert \frac{dcv(\theta)}{d\theta} \right\rvert$
Gamma(θ, θ), $\theta > 2$	$(1, \infty)$	$\sqrt{\frac{\theta}{\theta - 2}}$	$\theta^{-1/2}(\theta - 2)^{-3/2}$
Inverse-gamma$(\theta, 1)$	$(1, \sqrt{3})$	$\sqrt{\frac{\theta + 2}{\theta}}$	$\theta^{-3/2}(\theta + 2)^{-1/2}$
Inverse-Gaussian$(\theta, 1)$	$(1, \sqrt{5})$	$\sqrt{\frac{5\theta^2 + 4\theta + 1}{\theta^2 + 2\theta + 1}}$	$\frac{3\theta + 1}{(5\theta^2 + 4\theta + 1)^{1/2}(\theta + 1)^2}$
Log-normal$(0, \theta)$	$(1, \infty)$	$\sqrt{2\, e^{\theta} - 1}$	$e^{\theta}(2\, e^{\theta} - 1)^{-1/2}$

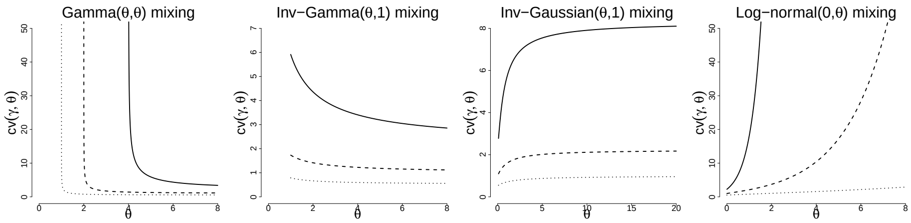

Figure 3: Relationship between (γ, θ) and cv for some RMW models. Solid, dashed and dotted lines are for $\gamma = 0.5, 1$ and 2, respectively. Thus, dashed lines indicate the relationship between θ and cv for RME models.

more involved than the one in (9). Thus, Jeffreys-style priors are not easy to obtain. As an alternative, we define

$$\pi(\beta, \gamma, \theta) \propto \pi(\gamma, \theta) \equiv \pi(\theta|\gamma)\pi(\gamma), \tag{14}$$

where $\pi(\theta|\gamma)$ and $\pi(\gamma)$ are proper density functions for θ (given γ) and γ, respectively. This extends the structure in (12) and again implies a flat prior for β. The product structure between β and (γ, θ) in (14) is reasonable in our RMW-AFT model where the interpretation of β does not depend on γ or θ. Conditional on γ, we define $\pi(\theta|\gamma)$ as in the RME-AFT case (via a prior for cv,

Table 3: $cv^*(\gamma, \theta)$ and its derivative w.r.t. θ for some RMW models. $\Theta = (0, \infty)$, unless otherwise stated and $\psi(\cdot)$ is the digamma function.

Mixing	$[cv^*(\gamma,\theta)]^2$	$\left\lvert \frac{d[cv^*(\gamma,\theta)]^2}{d\theta} \right\rvert$
Gamma(θ, θ), $\theta > \frac{2}{\gamma}$	$\frac{\Gamma(\theta)\Gamma(\theta-2/\gamma)}{\Gamma^2(\theta-1/\gamma)} - 1$	$\frac{\Gamma(\theta)\Gamma(\theta-2/\gamma)}{\Gamma^2(\theta-1/\gamma)} \left[\psi(\theta) + \psi(\theta - 2/\gamma) - 2\psi(\theta - 1/\gamma)\right]$
Inv-gamma$(\theta, 1)$	$\frac{\Gamma(\theta)\Gamma(\theta+2/\gamma)}{\Gamma^2(\theta+1/\gamma)} - 1$	$\frac{\Gamma(\theta)\Gamma(\theta+2/\gamma)}{\Gamma^2(\theta+1/\gamma)} \left[\psi(\theta) + \psi(\theta + 2/\gamma) - 2\psi(\theta + 1/\gamma)\right]$
Inv-Gaussian$(\theta, 1)$	$\sqrt{\frac{\theta\pi}{2}}\, e^{-\frac{1}{\theta}} \frac{K_{-(\frac{2}{\gamma}+\frac{1}{2})}(1/\theta)}{K^2_{-(\frac{1}{\gamma}+\frac{1}{2})}(1/\theta)} - 1$	$\sqrt{\frac{\pi}{2}} \frac{\theta^{-3/2}\, e^{-\frac{1}{\theta}}}{K^3_{-\left(\frac{1}{\gamma}+\frac{1}{2}\right)}(1/\theta)} \times$ $\left[K_{-\left(\frac{2}{\gamma}+\frac{1}{2}\right)}(1/\theta) K_{-\left(\frac{1}{\gamma}+\frac{1}{2}\right)}(1/\theta)\right.$ $+K_{-\left(\frac{1}{\gamma}+\frac{1}{2}\right)}(1/\theta) K_{-\left(\frac{2}{\gamma}-\frac{1}{2}\right)}(1/\theta)$ $\left.-2K_{-\left(\frac{2}{\gamma}+\frac{1}{2}\right)}(1/\theta) K_{-\left(\frac{1}{\gamma}-\frac{1}{2}\right)}(1/\theta)\right]$
Log-normal$(0, \theta)$	$e^{\theta/\gamma^2} - 1$	$\frac{1}{\gamma^2}\, e^{\theta/\gamma^2}$

$\pi^*(cv)$). Using $cv(\gamma, \theta)$ and $cv^*(\gamma, \theta)$ as defined in (4):

$$\begin{aligned} \pi(\theta|\gamma) &= \pi^*(cv(\gamma,\theta)) \left| \frac{dcv(\gamma,\theta)}{d\theta} \right|, \text{where} \\ & \qquad \frac{dcv(\gamma,\theta)}{d\theta} = \frac{\Gamma(1+2/\gamma)}{\Gamma^2(1+1/\gamma)} \frac{1}{2cv(\gamma,\theta)} \frac{d[cv^*(\gamma,\theta)]^2}{d\theta}. \qquad (15) \end{aligned}$$

Table 3 shows $[cv^*(\gamma, \theta)]^2$ and its partial derivative with respect to θ for the mixing distributions used in Table 2. Although some of these expressions are complicated, they can easily be evaluated numerically. Figure 3 shows the relationship between (γ, θ) and cv for some RMW models. As in the RME case, truncated exponential and Pareto type I priors for cv (given γ) are proposed. These are truncated to $(c_v^W(\gamma), \infty)$ (see (4)) but, as with RME models, some mixing distributions impose a finite upper bound for cv (e.g. $\left[\frac{\Gamma^2(1+2/\gamma)}{\Gamma^4(1+1/\gamma)} - 1\right]^{1/2}$ and $\left[\sqrt{\pi}\frac{\Gamma(1+2/\gamma)}{\Gamma^2(1+1/\gamma)}\frac{\Gamma(2/\gamma+1/2)}{\Gamma^2(1/\gamma+1/2)} - 1\right]^{1/2}$ for inverse gamma and inverse Gaussian mixing, respectively).

A proposal for $\pi(\gamma)$ is not trivial since a conjugate prior for γ in $(0, \infty)$ does not exist [18]. Here, a gamma prior is used for γ, with a range of hyperparameter values to asses the robustness of posterior inference. We recommend that users not choose hyper-parameters such that this prior is (nearly) flat, as this can lead to poor mixing of the MCMC algorithm described in Section 3.3; this is due to weak identifiability between γ and the intercept of the regression, when γ is close to 1 (see (8)).

3.2. The posterior

Censoring is a common feature in survival datasets, which must be taken into account. We assume noninformative censoring. However, as shown in Proposition 1 (see Supplementary Material, Section A), adding censored observations cannot destroy posterior propriety (and this applies to any survival model). In view of this result, Theorem 2 covers posterior propriety for the RMW-AFT model under the improper prior in (14) on the basis of the non-censored observations.

Theorem 2. Let $T_1, \ldots, T_n$ be the survival times of n independent individuals distributed as in (6). We observe survival times $t_1, \ldots, t_n$ and define $X = (x_1 \cdots x_n)'$. Assume that X has rank k and that the prior for (β, γ, θ) is proportional to $\pi(\gamma, \theta)$, which is a proper density function for (γ, θ). If $t_i \neq 0$ for all $i = 1, \ldots, n$, the posterior distribution of (β, γ, θ) is proper.

Proof. See Section A of the Supplementary Material. □

As discussed earlier, we use a proper prior for (γ, θ) so that Theorem 2 ensures a well-defined posterior if X has full rank and none of the observed survival times is equal to zero.

Posterior propriety can be precluded for particular samples of point observations, with zero Lebesgue measure [19]. However, this is not an issue for the RMW-AFT model. In this case, the posterior distribution is well-defined as long as there are no individuals for which $t_i = 0$. Whereas the latter is a reasonable assumption in most real applications, survival times can be recorded as zero due to rounding. In such a case, the zero point observation can be replaced by a set observation $(0, \epsilon)$, where ϵ stands for the minimum value that the recording mechanism detects (equivalent to a left censored observation on $(0, \epsilon)$).

3.3. Implementation

We assume right-censoring — common in survival datasets — and conduct Bayesian inference for the RMW-AFT model under the prior throughout this article. Mixing parameters are handled through data augmentation. An adaptive Metropolis-within-Gibbs sampler with Gaussian random walk proposals [20] is implemented. As the Weibull survival function has a known form, we do not use data augmentation for dealing with censored (and set) observations (as in [21, 12]). The full conditionals are

$$
\begin{aligned}
\pi(\beta_j|\beta_{-j},\gamma,\theta,\lambda,t,c) &\propto e^{-\gamma\beta_j\sum_{i=1}^n c_i x_{ij}}\, e^{-\sum_{i=1}^n \lambda_i (t_i\, e^{-x_i'\beta})^\gamma}, j=1,\ldots,k,\\
\pi(\gamma|\beta,\theta,\lambda,t,c) &\propto \gamma^{\sum_{i=1}^n c_i}\left[\prod_{i=1}^n t_i^{c_i}\right]^{\gamma-1} e^{-\gamma\sum_{i=1}^n c_i x_i'\beta}\times\\
&\qquad e^{-\sum_{i=1}^n \lambda_i (t_i\, e^{-x_i'\beta})^\gamma}\pi(\theta|\gamma)\pi(\gamma),\\
\pi(\theta|\beta,\gamma,\lambda,t,c) &\propto \prod_{i=1}^n dP(\lambda_i|\theta)\pi(\theta|\gamma),\\
\pi(\lambda_i|\beta,\gamma,\theta,\lambda_{-i},t,c) &\propto \lambda_i^{c_i}\, e^{-\lambda_i (t_i\, e^{-x_i'\beta})^\gamma}\, dP(\lambda_i|\theta), i=1,\ldots,n,
\end{aligned}
\tag{16}
$$

where $\beta_{-j}=(\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\beta_k)$, $\lambda_{-i}=(\lambda_1,\ldots,\lambda_{i-1},\lambda_{i+1},\lambda_n)$ and $c=(c_1,\ldots,c_n)'$ with c_i equal to 1 if the survival time for individual i is observed and 0 if it is censored. In general, Metropolis updates are required in all full conditionals. However, Gibbs steps can be used for some mixing distributions. For instance, the first four examples in Table 1, respectively, lead to gamma$(1+c_i, 1+(t_i\, e^{-x_i'\beta})^\gamma)$, gamma$(\theta+c_i,\theta+(t_i\, e^{-x_i'\beta})^\gamma)$, Generalized Inverse Gaussian$(-\theta+c_i, 2, 2(t_i\, e^{-x_i'\beta})^\gamma)$ and Generalized Inverse Gaussian$(c_i-1/2, 1, \theta^{-2}+2(t_i\, e^{-x_i'\beta})^\gamma)$ full conditionals for λ_i (the Generalized Inverse Gaussian is parametrized as in [22]).

We observed poor mixing of the chain for the log-normal$(0,\theta)$ mixture. This relates to a strong a priori correlation between γ and θ, which persists when not much can be learned about θ (as θ controls the tails of the distribution, this is especially problematic for small n and/or high proportion of censoring). We opt for a re-parametrization of this model from (θ,γ) to (θ^*,γ), where $\theta^*=\theta/\gamma^2$. As in the original parametrization, a prior for θ^* can be induced via a prior for cv (where $[cv^*(\gamma,\theta^*)]^2$ equals $e^{\theta^*}-1$). This new parametrisation is more orthogonal and substantially improves the mixing of the chain.

Code to implement RMW-AFT regression models is provided as an R library (see Section D of the Supplementary Material). Further details on the implementation can be found in Section B of the Supplementary Material.

3.4. Model comparison

The adequacy of different mixing distributions is evaluated using standard Bayesian model comparison criteria: Bayes factors (BF), conditional predictive ordinates (CPO) and pseudo Bayes factors (PsBF). The BF between two models is the ratio between the marginal likelihoods, which are computed using the method in [23]. Instead, CPO [24] is an indicator of predictive ability. For observation i, CPO_i is defined as

$$\text{CPO}_i = f(t_i|t_{-i}) = \left[\text{E}\left(\frac{1}{f(t_i|\beta,\gamma,\theta)}\right)\right]^{-1},\ t_{-i}=(t_1,\dots,t_{i-1},t_{i+1},\dots,t_n), \tag{17}$$

where the expectation is with respect to $\pi(\beta,\gamma,\theta|t)$ and $f(\cdot|t_{-i})$ is the predictive density given t_{-i}. If $c_i=0$, the survival function $S(\cdot|t_{-i})=1-F(\cdot|t_{-i})$ (where F is the CDF) is used instead of $f(\cdot|t_{-i})$ (as in [25]). Larger CPO values are preferred. We also use the pseudo marginal likelihood $\text{PsML}=\prod_{i=1}^{n}\text{CPO}_i$ [24]. Analogously to BF, PsBF are computed as the ratio between the PsML associated with two models. The performance of these model comparison criteria for RMW-AFT regression models has been assessed through simulations (see Section C in the Supplementary Material), which suggest they behave well.

3.5. Detection of influential observations and outliers

A feature of models described by (1) is to reduce the number of influential observations. We illustrate this using the Kullback-Leibler divergence $\text{KL}_i=\text{KL}(\pi(\beta,\gamma,\theta|t),\pi(\beta,\gamma,\theta|t_{-i}))$[26]. As in [27], we use the calibration index $p_i=0.5\left[1+\sqrt{1-\exp\{-2\text{KL}_i\}}\right]$ ($p_i\in[0.5,1]$) as a criteria to characterise influential observations, where large value of p_i suggests that observation i is influential.

Intuitively, outliers relate to unusual values of the λ_i's [28]. Hence, outliers are identified using the posterior distribution of the λ_i's [6]. For each observation i, we compare the models $M_0 : \Lambda_i = \lambda_{ref}$ and $M_1 : \Lambda_i \neq \lambda_{ref}$ (all other $\Lambda_j, j \neq i$ free), where λ_{ref} is a reference value (specific to the mixing distribution). The BF in favour of M_0 versus M_1 is computed using a generalized Savage-Dickey density ratio [29], as

$$\begin{aligned} \mathrm{BF}_{01}^{(i)} &= \pi(\lambda_i|t,c)\mathrm{E}\left(\frac{1}{dP(\lambda_i|\theta)}\right)\bigg|_{\lambda_i=\lambda_{ref}} \\ &= \mathrm{E}\left(\frac{\pi(t_i|\beta,\gamma,\theta,\lambda_i,c_i)dP(\lambda_i|\theta)}{\pi(t_i|\beta,\gamma,\theta,c_i)}\right)\mathrm{E}\left(\frac{1}{dP(\lambda_i|\theta)}\right)\bigg|_{\lambda_i=\lambda_{ref}}, \end{aligned} \quad (18)$$

where the expectations are with respect to $\pi(\beta,\gamma,\theta|t,c)$ and $\pi(\theta|\Lambda_i = \lambda_{ref}, t, c)$, respectively. This is computationally intensive: for each $\mathrm{BF}_{01}^{(i)}$, we need to fit a sub-model fixing $\lambda_i = \lambda_{ref}$. However, if θ does not appear in the model, these fits are not required and (18) reduces to the usual Savage-Dickey density ratio

$$\mathrm{BF}_{01}^{(i)} = \frac{\pi(\lambda_i|t,c)}{dP(\lambda_i)}\bigg|_{\lambda_i=\lambda_{ref}} = \mathrm{E}\left(\frac{\pi(t_i|\beta,\gamma,\lambda_i,c_i)}{\pi(t_i|\beta,\gamma,c_i)}\right)\bigg|_{\lambda_i=\lambda_{ref}}, \quad (19)$$

where $\mathrm{E}(\cdot)$ is with respect to $\pi(\beta,\gamma|t,c)$. Here, $\pi(t_i|\beta,\gamma,\lambda_i,c_i)$ and $\pi(t_i|\beta,\gamma,c_i)$ are the conditional density (or survival if $c_i = 0$) functions of t_i when conditioning or not on λ_i, respectively.

This methodology relies on the choice of a reasonable λ_{ref}. In [6], $\lambda_{ref} = \mathrm{E}(\Lambda_i|\theta)$ was used arguing that, in the absence of unobserved heterogeneity, the posterior density of the frailty terms should behave as a Dirac function with a spike on $\mathrm{E}(\Lambda_i|\theta)$. In our context, this is always well-defined because $\mathrm{E}(\Lambda_i|\theta)$ is required to be finite for the identification of γ (see Theorem 1). Table 1 displays $\mathrm{E}(\Lambda_i|\theta)$ for the examples in this article. When θ is unknown, we replace it by its posterior median (based on the MCMC sample). However, in our context, the empirical evidence does not support the latter choice for the censored observations. Only a lower bound of the survival time is known for right-censored observations, which is highly informative for the λ_i's (as they are linked to the rate/scale of the underlying distribution). Hence, the posterior distributions of the λ_i's linked to right-censored observations are driven towards lower values. We propose to keep $\lambda^o_{ref} = \mathrm{E}(\Lambda_i|\theta)$ as the reference value for non-censored observations and adjust it by the effect of censoring for censored observations as follows:

$$\lambda^c_{ref} = C_i(\beta,\gamma,\theta)\lambda^o_{ref}, \text{ with } C_i(\beta,\gamma,\theta) = \frac{\mathrm{E}(\Lambda_i|t_i, c_i = 0, \beta, \gamma, \theta)}{\mathrm{E}(\Lambda_i|t_i, c_i = 1, \beta, \gamma, \theta)}. \tag{20}$$

For exponential mixing $C_i(\beta,\gamma,\theta) = 1/2$ and $C_i(\beta,\gamma,\theta) = \theta/(\theta+1)$ for gamma mixing (see the conditionals in (16)). In these cases $C_i(\beta,\gamma,\theta)$ does not depend on i, β or γ. Let $t^*_i = \left(t_i\, e^{-x_i'\beta}\right)^\gamma$ and $K_p(\cdot)$ be the modified Bessel function. If $\Lambda_i|\theta \sim$ inv-gamma$(\theta, 1)$ or $\Lambda_i|\theta \sim$ inv-Gaussian$(\theta, 1)$,

$$C_i(\beta,\gamma,\theta) = \frac{K^2_{-\theta+1}\left(2\sqrt{t^*_i}\right)}{K_{-\theta+2}\left(2\sqrt{t^*_i}\right) K_{-\theta}\left(2\sqrt{t^*_i}\right)} \text{ or} \tag{21}$$

$$C_i(\beta,\gamma,\theta) = \frac{K^2_{1/2}\left(\sqrt{2t^*_i + \theta^{-2}}\right)}{K_{-1/2}\left(\sqrt{2t^*_i + \theta^{-2}}\right) K_{3/2}\left(\sqrt{2t^*_i + \theta^{-2}}\right)}, \tag{22}$$

respectively. For log-normal mixing, $C_i(\beta,\gamma,\theta)$ has no closed form but can be estimated via numerical integration. The performance of this strategy has been validated using simulated datasets.

To illustrate our outlier detection method, Figure 4 displays $\mathrm{BF}^{(i)}_{01}$ as a function of a standardized observation z_i (horizontal line located are the threshold above which observations will be considered outliers [30]) . Following the structure in (7), this is defined in terms of $\log(t_i)$ minus its mean, divided by its standard deviation (given β, γ and θ). Let $\psi(\cdot)$ be the digamma function. As $\log(T_0) \sim$ Gumbel$(0, \gamma^{-1})$, we have

$$z_i = \gamma \left[\frac{\log(t_i) - x_i'\beta + \gamma^{-1}\left(\mathrm{E}_{\Lambda_i}(\log(\Lambda_i)|\theta) + \psi(1)\right)}{\sqrt{\mathrm{Var}_{\Lambda_i}(\log(\Lambda_i)|\theta) + \pi^2/6}} \right]. \tag{23}$$

In terms of z_i, $\mathrm{BF}^{(i)}_{01}$ does not depend on β nor γ (the full conditional of Λ_i depends on t_i only through t^*_i). Naturally, outliers relate to large values of $|z_i|$. The threshold on $|z_i|$ at which an observation is detected as outlier depends on θ. For example, the RMW model with gamma(θ,θ) mixing tends to the Weibull model as $\theta \to \infty$ and thus, the model with larger θ requires larger $|z_i|$ values to distinguish it from the Weibull. Finally, the correction factor $C_i(\beta,\gamma,\theta)$ leads to similar detection threshold (in terms of $|z_i|$) for censored and non-censored observations (see Figure 4).

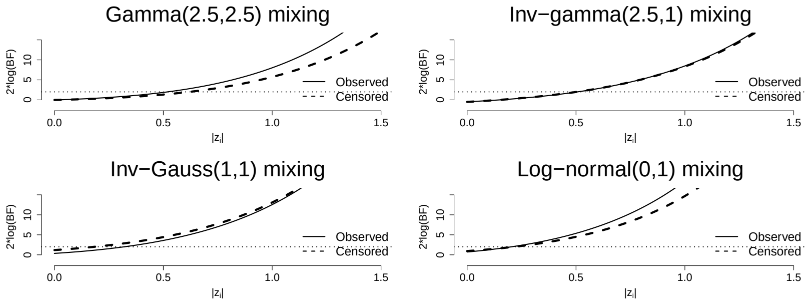

Figure 4: 2 × log-Bayes factor for outlier detection as a function of $|z|$ in RMW-AFT models. The horizontal line is the threshold above which observations will be considered outliers [30].

4. Applications

We illustrate the usage of RMW-AFT survival models using two case studies, both of them related to medical applications: one regarding bone marrow transplants and another on cerebral palsy. The main features of these datasets are summarized in Table 4. While these datasets differ in terms of sample size and the censoring percentage, there is an important common feature: only a small number of covariates has been recorded. As such, a substantial amount of unobserved heterogeneity is expected in both cases.

The R code used to perform these analyses is provided in Section D of the Supplementary Material.

4.1. Autologous and Allogeneic Bone Marrow Transplant [31]

The data contains post-surgery disease-free survival times (until relapse or death, in months) for 101 advanced acute myelogenous leukemia patients, including 51 right-censored observations. In the trial, 51 patients received an autologous bone marrow transplant, replacing the patient's marrow with their own marrow treated with high doses of chemotherapy. The remaining patients had

Table 4: Main features of the Autologous and Allogeneic Bone Marrow Transplant (AA) [31] and Cerebral Palsy (CP) [32, 33] datasets.

Dataset	AA	CP
Sample size	101	1,549
Censoring	50.5%	84.4%
No. of covariates	1	2
	X_1: transplant type	X_1: No. of severe impairments (0, 1, 2, 3)
	(autologous, allogeneic)	X_2: birth weight (kg.)

an allogeneic bone transplant, using marrow extracted from a sibling. Besides the treatment type, no additional covariates were recorded. A brief descriptive summary of these data is provided in Section D of the Supplementary Material.

The standard graphical check of $\log(-\log(S(t)))$ versus t (not reported) suggests that the PH assumption does not hold. The data is first analyzed using exponential and Weibull AFT models (which have equivalent PH representations). If $\gamma \sim$ gamma(4,1) the BF in favour of the Weibull model with free γ (w.r.t. the exponential one) is 4.6, suggesting $\gamma \neq 1$. In line with this, the posterior median of γ is 0.69 (95% HPD: (0.54,0.87)). In addition, RME-AFT and RMW-AFT models with the mixing distributions in Table 1 are fitted using the priors proposed in Section 3.1. In contrast to the Weibull case, there is evidence in favour of $\gamma = 1$ in all the RMW-AFT regressions. For example, for the exponential(1) mixing and $\gamma \sim$ gamma(4,1), the BF in favour of the RME specification ($\gamma = 1$) vs. RMW is 12.7. In this case, the posterior median of γ is 0.86 (95% HPD: (0.65,1.06)). These opposite conclusions are linked to the fact that the Weibull model tends to underestimate γ in an attempt to capture the over-dispersion of the data (the cv of the Weibull is a decreasing function of γ). Based on this evidence, RME-AFT models are used for these data. We adopt E(cv) equal to 1.25, 1.5, 2, 5 and 10 (if there is no θ in the model, all these priors coincide). Large values of E(cv) are associated with stronger prior beliefs about the existence of unobserved heterogeneity (see (4)). Nevertheless, as explained in Section 3.1, if E(cv) is larger than $\sqrt{3}$, the model generated by inverse gamma mixing is not compatible with the prior beliefs. The same occurs for inverse Gaussian mixing when E(cv) $> \sqrt{5}$. The algorithm in Section 3.3 is implemented. For all models, the total number of iterations is 600,000. In the following, results are presented on the basis of 9,000 draws (after a burn-in of 25% and thinning). Graphical summaries, together with the convergence diagnostics proposed in [34] and [35] strongly suggest convergence for the chains (see Section D of the Supplementary Material).

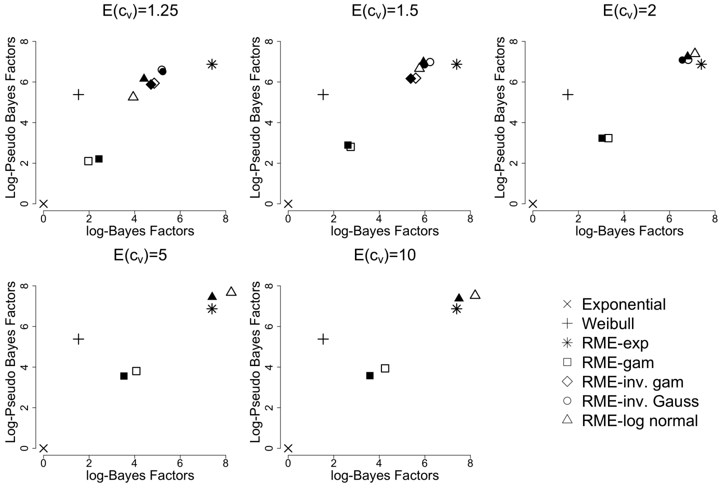

Figure 5: Autologous and allogeneic bone marrow transplant dataset. Model comparison in terms of Bayes factors (with respect to the exponential model) and pseudo Bayes factors for the mixing distributions presented in Table 1 using $\gamma \sim$ gamma(4,1). Unfilled and filled characters denote a truncated exponential and Pareto priors for cv, respectively.

The presence of unobserved heterogeneity is supported by the data. Figure 5 compares the fitted models in terms of BF and PsBF (w.r.t. the exponential model). For all priors considered, both criteria support all the mixture models in Table 1 over the exponential model. The Weibull model (which itself can be viewed as a mixture of exponentials provided $\gamma < 1$, see [9]) is also beaten in terms of BF, which are, of course, dependent on the prior on γ: for example, a gamma(1,1) prior leads to more support for the Weibull model while a

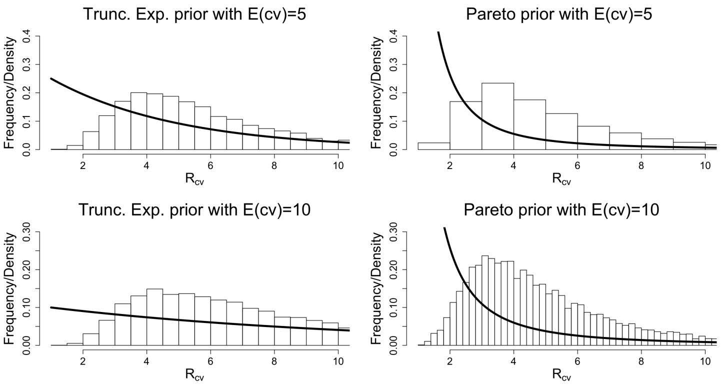

Figure 6: Autologous and allogeneic bone marrow transplant dataset. Comparison between the prior (continuous line) and the posterior distribution (histogram) of $R_{cv}(1, \theta)$ (which equals $cv(1, \theta)$) using a log-normal mixing distribution.

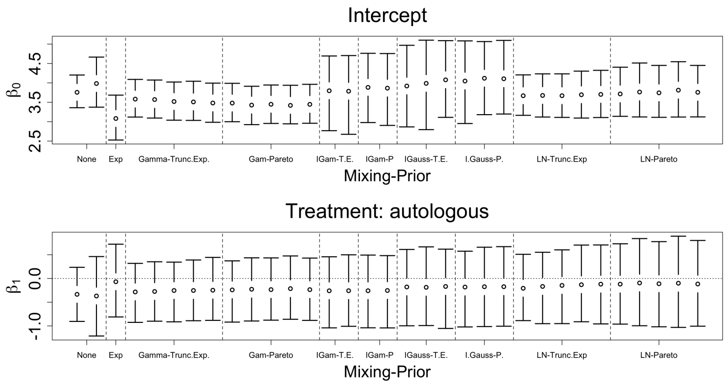

Figure 7: Posterior of the regression coefficients for the bone marrow transplant data using the RME-AFT model in (6) ($\gamma = 1$). The prior is (12) with (if appropriate) a Trunc-Exp and a Pareto prior for cv. The first two lines ("None") correspond respectively to the exponential and Weibull models without mixing. For models with θ, E(cv) is 1.25, 1.5, 2,5 and 10 from left to right (only 1.25 and 1.5 for inverse gamma; 1.25, 1.5 and 2 for inverse Gaussian mixing). For the Weibull model γ ~gamma(4,1). Vertical lines represent 95% HPD intervals and dots the posterior medians. β_0: Intercept, β_1: Treatment (autologous).

gamma(0.001,0.001) leads to slightly less support than for the exponential. The PsBF is a predictive criterion and is virtually unaffected by these changes in prior. The similarity of both criteria for the mixture models is indicative of the fact that priors are well-matched. Despite its simplicity, the exponential mixing receives most support overall. It is easy to implement (the full conditionals of the λ_i's have a known form) and does not require prior elicitation for θ. The log-normal mixing distribution has slightly more support for large $\mathrm{E}(cv)$, but rather less for small $\mathrm{E}(cv)$. Interestingly, the popular gamma mixing is the least preferred of all mixing distributions. Despite the small sample size, there is learning about R_{cv} (which equals cv here). Even though the truncated exponential and Pareto priors are concentrated around small values of R_{cv}, its posterior distribution is shifted to the right (Figure 6). This suggests the need for a mixture and is consistent with strong heterogeneity in the data that leads to support the exponential mixing model (infinite cv). Whereas the choice of a prior affects inference on R_{cv}, the posterior distribution of β (usually the parameter of interest) is more robust. The effect of the mixture models over β is illustrated in Figure 7. All models suggest that there is no substantive difference between the median survival times under both treatments. However, for all considered mixing distributions and priors, the effect of the treatment (β_1) is less pronounced than for the exponential and Weibull models without mixing. This discrepancy is among the largest when using the exponential mixing.

No influential observations are detected for any model considered, including the exponential and Weibull models without mixture (all p_i's are below 0.9). Figure 8(a) illustrates the posterior behaviour of the λ_i's for the RME model with exponential mixing. The outlier detection mechanism proposed in Section 2 does not detect outlying observations (see Figure 8(b)). So no single observation is identified as an outlier, yet there is ample evidence in favour of the exponential mixture model on the basis of the entire sample.

4.2. Cerebral palsy [32, 33]

These data contain records of 1,549 children affected by cerebral palsy and born during the period 1966-1984 in the area of the Mersey Region Health Authority. The time to follow-up (survival or censoring) is recorded in years since birth. We use the amount of severe impairments (ambulation, manual

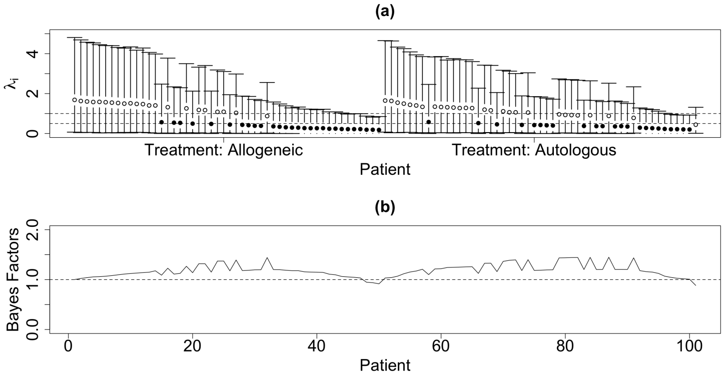

Figure 8: Autologous and allogeneic bone marrow transplant dataset. (a) 95% HPD interval of the λ_i's for the exponential mixing distribution. Horizontal lines at $\lambda^{o}_{ref} = 1$ and $\lambda^{c}_{ref} = 1/2$. Circles located at posterior medians (filled circles for censored observations). Observations are grouped by treatment and displayed in ascending order of the t_i's. (b) Bayes factors in favour of the model $M_1 : \Lambda_i \neq \lambda_{ref}$ versus $M_0 : \Lambda_i = \lambda_{ref}$.

dexterity and mental ability) and the birth weight (in kg.) as predictors for the time to death [33]. The deaths of 242 patients were observed by the end of the study leaving the survival times of the remaining 1,307 patients as right-censored (very large proportion of censoring: 84.4%). A brief descriptive summary of these data is provided in Section D of the Supplementary Material.

The data are analysed using the RMW-AFT model defined in (6) with the mixing distributions in Table 1. For comparison, a Weibull regression is also fitted. For models without θ (i.e. Weibull and RMW with exponential(1) mixing), the total number of MCMC iterations is 600,000. We doubled the iterations for the remaining models, whose chains mix less rapidly. In all cases, results are shown based on 9,000 draws (after a 25% burn-in and thinning). Convergence diagnostics, including graphical summaries and formal tests [34, 35] suggest the chains have converged (see Supplementary Material, Section D).

Figure 9 summarises marginal posterior inference. Throughout, results are fairly insensitive to the choice of prior for γ (three different gamma priors), the form of the prior for cv (truncated exponential or Pareto) and its corresponding prior mean (1.5 or 5). The main differences relate to whether mixing is used or not. The bottom panel shows that, in all cases, γ is estimated to be larger than 1. This suggest a non-monotone hazard shape (in line with the results in [33]). Like in the previous application, the Weibull model tends to underestimate γ in order to accommodate the variability in the data. This result is in line with the simulations described in Section C of the Supplementary Material.

In the AFT specification we use, e^{β_j} can be interpreted as proportional changes of the median survival time, regardless of the mixture. Figure 9 shows that mixture models estimate a similar effect of the covariates. The effect of no impairments (β_1) is less strong than in the Weibull model, where the median survival time is increased by a median factor of approximately $e^{3.3} \approx 27$ for children with no impairments (w.r.t. those with 3 impairments). Under the mixture models, the same factor is roughly $e^{3.1} \approx 22$. We note that these small differences in the estimation of the regression coefficients are coherent with the simulation results shown in Section C of the Supplementary Material.

Figure 10 shows that the mixture models provide a better fit for the data and lead to better predictions. In fact, for all priors considered, all the mixture models have a better performance in terms of BF and PsBF (and thus PsML). Again, both criteria are very close, strongly suggesting the existence of unobserved heterogeneity. This evidence is also supported by Table 5, where the posterior distribution of R_{cv} is concentrated away from one (results with a Pareto prior on cv are very similar). However, since R_{cv} measures the ratio of cv in RMW models versus the Weibull model assuming a common value of γ, whereas γ is estimated to be smaller for the Weibull model, R_{cv} somewhat overestimates the actual ratio of cv's between the models. For example, for gamma(θ, θ) mixing with a truncated exponential prior for cv, $d_1 = 1$, $d_2 = 4$ and $\mathrm{E}(cv) = 1.5$, the actual ratio is estimated as 2.08 (while the posterior median of R_{cv} is 2.39).

Overall, the exponential mixing provides the best results in terms of BF and PsBF. The latter model is also the simplest model to elicit (there is no θ) and is computationally attractive. Figure 11 presents results for the exponential mixture model on the outlier detection procedure of Section 2, which does not

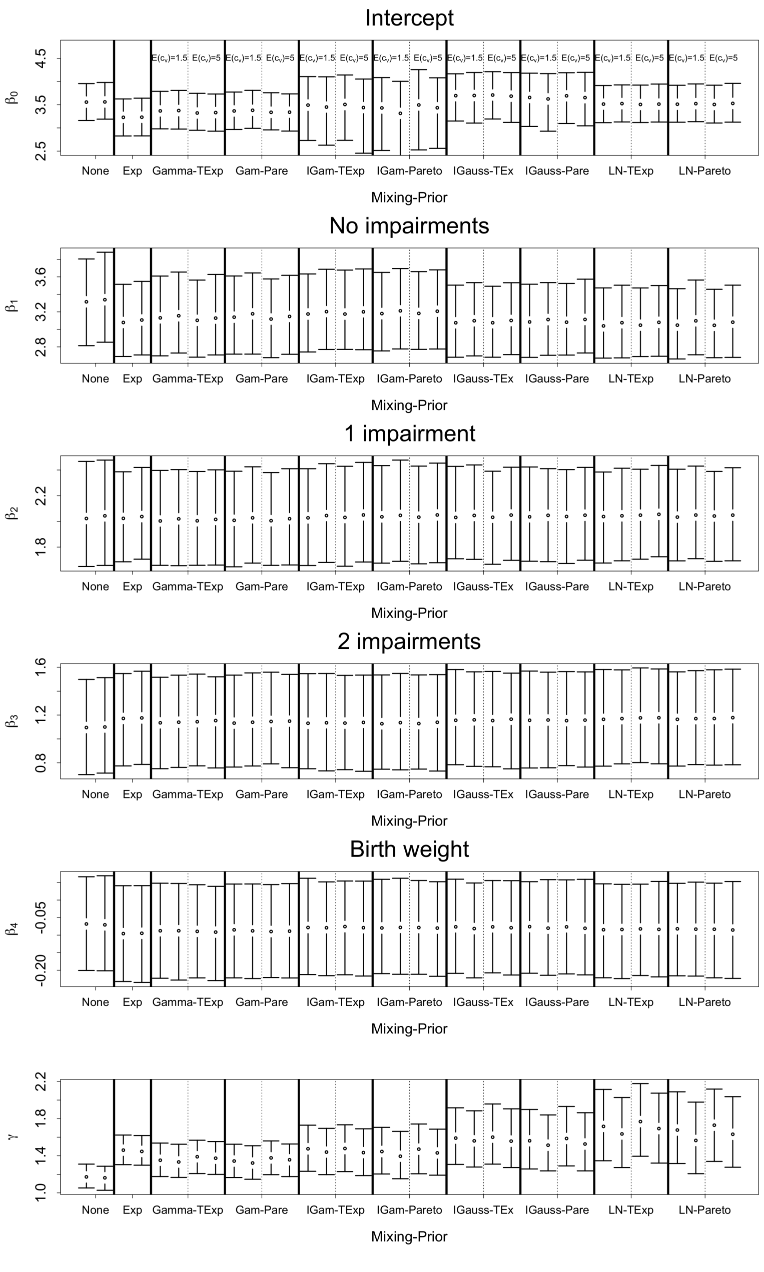

Figure 9: Posterior results for the cerebral palsy data using different distributions in the RMW-AFT family in (6). The prior is (14) with a gamma prior for γ and (if appropriate) a Trunc-Exp(a) and a Pareto(b) prior for cv. Vertical lines represent 95% HPD intervals and dots are the posterior medians. From left to right, we use a gamma(4,1) and gamma(1,1) prior for γ. Values of E(cv) are displayed in the top panel. β_0: intercept, β_1: no impairments, β_2: 1 impairment, β_3: 2 impairments, β_4: birth weight. Bottom panel is for γ.

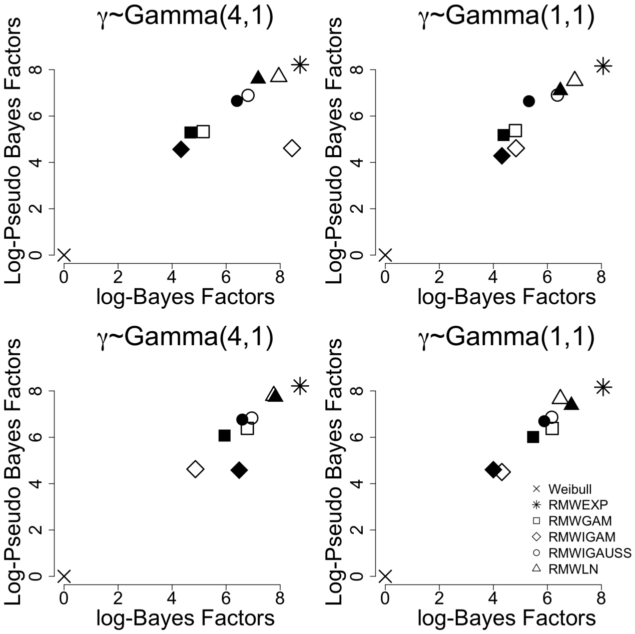

Figure 10: Cerebral palsy dataset. Model comparison in terms of Bayes factors and pseudo Bayes factors (with respect to the Weibull model) for the mixing distributions presented in Table 1. Unfilled and filled characters denote a truncated exponential and Pareto prior for cv, respectively. Upper panels use $\mathrm{E}(cv) = 1.5$. Lower panels use $\mathrm{E}(cv) = 5$. Legend is displayed in the last panel.

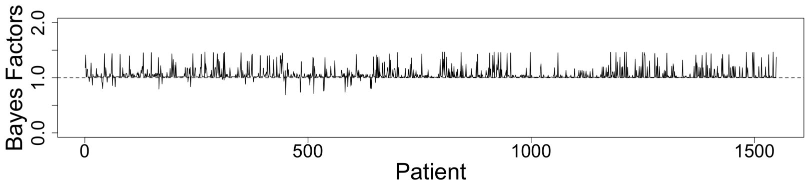

Figure 11: Cerebral palsy dataset. Bayes factors for the RMW-AFT model with the exponential mixture in favour of the hypothesis $H_1 : \lambda_i \neq \lambda_{ref}$, with $\lambda_{ref}^{o} = 1$ and $\lambda_{ref}^{c} = 1/2$ under a gamma(4,1) prior for γ.

detect any outlying observations. Again, we have strong evidence of unobserved heterogeneity in the sample, which provides strong support for mixture models, but there are no particular single observations that could be considered clear outliers.

Table 5: Cerebral palsy dataset. Posterior medians and 95% HPD intervals of $R_{cv}(\gamma, \theta)$ (as in (5)) for RMW-AFT models under a gamma(d_1, d_2) prior for γ and a truncated exponential prior for cv.

		$d_1 = 4, d_2 = 1$		$d_1 = d_2 = 1$	
E(cv)	Mixing	Med.	95% HPD	Med.	95% HPD
1.5	Gam(θ, θ)	2.39	[1.23, 4.42]	2.32	[1.19, 4.24]
	Inv-gam$(\theta, 1)$	1.42	[1.24, 1.55]	1.41	[1.21, 1.55]
	Inv-Gauss$(\theta, 1)$	1.67	[1.47, 1.83]	1.66	[1.45, 1.83]
	Log-norm$(0, \theta)$	2.38	[1.53, 3.17]	2.24	[1.46, 3.13]
5	Gam(θ, θ)	6.99	[1.59,19.87]	6.94	[1.45,19.79]
	Inv-gam$(\theta, 1)$	1.43	[1.24, 1.56]	1.40	[1.19, 1.55]
	Inv-Gauss$(\theta, 1)$	1.68	[1.47, 1.85]	1.66	[1.44, 1.84]
	Log-norm$(0, \theta)$	2.56	[1.69, 3.39]	2.43	[1.61, 3.29]

5. Conclusion

Mixtures of life distribution are proposed in order to account for unobserved heterogeneity in survival models. In particular, the family generated by mixtures of Weibull distributions with random rate parameter is explored in detail (and its special case of rate mixtures of exponentials). These mixtures are shown to induce a larger coefficient of variation than the original Weibull distribution and more flexible hazard functions.

Instead of the usual mixed PH scheme adopted in this context, covariates are added via an AFT specification. As an advantage, the marginal model retains the AFT structure and the interpretation of the covariate effects is invariant to the mixing distribution. This allows comparison between estimates based on different RMW-AFT models (with any mixing) and those produced by any other AFT model (in particular the Weibull AFT model). The mixing representation facilitates the choice of a prior distribution. We opt for a prior that is inspired by the Jeffreys rule, with a product structure comprising an (improper) flat prior for the regression coefficients and a proper component for the remaining parameters. In view of the clear interpretation of the covariate effects, this product structure seems a reasonable assumption. The proper part of the prior is elicited via the coefficient of variation of the survival times. Priors for different mixing distributions are matched by a common prior on this coefficient of variation, so that models can be meaningfully compared through Bayes factors. We derive simple (and easily satisfied) conditions for posterior propriety. In addition, we show that adding censored observations cannot destroy the existence of the posterior distribution.

Mixture models diminish the effect that anomalous observations have on posterior inference. Nonetheless, it might be of interest to identify any outlying observations driving the unobserved heterogeneity. An outlier detection method is designed, which exploits the mixing structure and compares individual frailties with a reference level. The comparison is formalized by means of Bayes factors. Choosing a reference value is crucial. A general recommendation is presented, including a correction factor for censored observations.

Both analysed datasets provide strong evidence for unobserved heterogeneity, shown not to be a consequence of a small number of specific outliers. Mixture models are supported by the data in terms of Bayes factors and predictive performance. In particular, the use of an exponential mixture distribution (for which the coefficient of variation for the survival times does not exist) leads to the overall best results in both applications. Our simulations suggest this is a reflection of the strong unobserved heterogeneity that is present in the analysed datasets (not surprising in light of the small number of covariates recorded in both cases). More flexible mixing distributions, such as the ones indexed by a free parameter θ in Table 1 can provide a better fit in other situations. Therefore, we would recommend practitioners to investigate the performance of multiple mixing distributions (e.g. through the Bayesian model comparison criteria discussed here) rather than fixing the mixing distribution a priori.

Acknowledgements

We are grateful to P.O.D. Pharoah and Jane Hutton for access to the cerebral palsy dataset, and thank Jane Hutton, an Associate Editor and three reviewers for insightful and constructive comments.

During most of this research, Catalina Vallejos was a Ph.D. student at the University of Warwick. She acknowledges funding from the University of Warwick and the Pontificia Universidad Católica de Chile. No conflicts of interest are declared.

References

[1] Y. Omori, R. Johnson, The influence of random effects on the unconditional hazard rate and survival functions, Biometrika 80 (1993) 910–914.

[2] P. Hougaard, Frailty models for survival data, Lifetime Data Analysis 1 (1995) 255–273.

[3] L. Duchateau, P. Janssen, The Frailty Model, Springer, 2008.

[4] P. Du, S. Ma, Frailty model with spline estimated nonparametric hazard function, Statistica Sinica (2010) 561–580.

[5] N. Balakrishnan, V. Leiva, A. Sanhueza, F. E. V. Labra, Estimation in the Birnbaum-Saunders distribution based on scale-mixture of normals and the EM-algorithm, Sort: Statistics and Operations Research Transactions 33 (2009) 171–192.

[6] C. A. Vallejos, M. F. J. Steel, Objective bayesian survival analysis using shape mixtures of log-normal distributions, Journal of the American Statistical Association 110 (510) (2015) 697–710.

[7] J. Heckman, B. Singer, The identifiability of the proportional hazards model, The Review of Economic Studies 51 (1984) 231–241.

[8] A. W. Marshall, I. Olkin, Life Distributions, Springer, 2007.

[9] N. Jewell, Mixtures of exponential distributions, The Annals of Statistics 10 (1982) 479–484.

[10] J. Abbring, G. Van Den Berg, The unobserved heterogeneity distribution in duration analysis, Biometrika 94 (2007) 87–99.

[11] J. Heckman, B. Singer, A method for minimizing the impact of distributional assumptions in econometric models for duration data, Econometrica 52 (1984) 271–320.

[12] A. Kottas, Nonparametric Bayesian survival analysis using mixtures of Weibull distributions, Journal of Statistical Planning and Inference 136 (2006) 578–596.

[13] I. D. Ha, M. Noh, Y. Lee, Bias reduction of likelihood estimators in semiparametric frailty models, Scandinavian Journal of Statistics 37 (2) (2010) 307–320.

[14] P. Barker, R. Henderson, Small sample bias in the gamma frailty model for univariate survival, Lifetime data analysis 11 (2) (2005) 265–284.

[15] A. Wienke, Frailty Models in Survival Analysis, Chapman & Hall/CRC, 2010.

[16] S. Wasinrat, W. Bodhisuwan, P. Zeephongsekul, A. Thongtheeraparp, A mixture of Weibull hazard rate with a Power Variance Function frailty, Journal of Applied Sciences 13 (2013) 103–110.

[17] K. Lomax, Business failures: Another example of the analysis of failure data, Journal of the American Statistical Association.

[18] R. Soland, Bayesian analysis of the Weibull process with unknown scale and shape parameters, IEEE Transactions on Reliability 18 (1969) 181–184.

[19] C. Fernández, M. Steel, On the dangers of modelling through continuous distribution: A Bayesian perspective, Bayesian Statistics 6, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.), Oxford University Press (1998) 213–238.

[20] G. Roberts, J. Rosenthal, Examples of adaptive MCMC, Journal of Computational and Graphical Statistics 18 (2009) 349–367.

[21] J. Ibrahim, M.-H. Chen, D. Singha, Bayesian Survival Analysis, Springer, 2001.

[22] L. Devroye, Non-Uniform Random Variable Generation, Springer, 1986.

[23] S. Chib, I. Jeliazkov, Marginal likelihood from the Metropolis-Hastings output, Journal of the American Statistical Association 96 (2001) 270–281.

[24] S. Geisser, W. Eddy, A predictive approach to model selection, Journal of the American Statistical Association 74 (1979) 153–160.

[25] T. Banerjee, M. Chen, D. Dey, S. Kim, Bayesian analysis of generalized odds-rate hazards models for survival data, Lifetime Data Analysis 13 (2007) 241–260.

[26] F. Peng, D. K. Dey, Bayesian analysis of outlier problems using divergence measures, The Canadian Journal of Statistics 23 (1995) 199–213.

[27] R. E. McCulloch, Local model influence, Journal of the American Statistical Association 84 (1989) 473–478.

[28] M. West, Outlier models and prior distributions in Bayesian linear regression, Journal of the Royal Statistical Society, B 46 (1984) 431–439.

[29] I. Verdinelli, L. Wasserman, Computing Bayes factors by using a generalization of the Savage-Dickey density ratio, Journal of the American Statistical Association 90 (1995) 614–618.

[30] R. Kass, A. Raftery, Bayes factors, Journal of the American Statistical Association 90 (1995) 773–795.

[31] J. P. Klein, M. L. Moeschberger, Survival Analysis: techniques for censored and truncated data, 1st Edition, Springer, 1997.

[32] J. Hutton, T. Cooke, P. Pharoah, Life expectancy in children with cerebral palsy, British Medical Journal 309 (1994) 431–435.

[33] G. Kwong, J. Hutton, Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy, Journal of the Royal Statistical Society: Series C (Applied Statistics) 52 (2003) 153–168.

[34] J. Geweke, Evaluating the accuracy of sampling-based approaches to calculating posterior moments, Bayesian Statistics 4, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.), Oxford University Press, Oxford, UK. (1992) 169–193.

[35] P. Heidelberger, P. D. Welch, Simulation run length control in the presence of an initial transient, Operations Research 31 (1983) 1109–1144.