Objective Bayesian Survival Analysis Using Shape Mixtures of Log-Normal Distributions

Catalina A. Vallejos and Mark F.J. Steel *

Abstract

Survival models such as the Weibull or log-normal lead to inference that is not robust to the presence of outliers. They also assume that all heterogeneity between individuals can be modelled through covariates. This article considers the use of infinite mixtures of lifetime distributions as a solution for these two issues. This can be interpreted as the introduction of a random effect in the survival distribution. We introduce the family of Shape Mixtures of Log-Normal distributions, which covers a wide range of density and hazard functions. Bayesian inference under non-subjective priors based on the Jeffreys rule is examined and conditions for posterior propriety are established. The existence of the posterior distribution on the basis of a sample of point observations is not always guaranteed and a solution through set observations is implemented. In addition, a method for outlier detection based on the mixture structure is proposed. A simulation study illustrates the performance of our methods under different scenarios and an application to a real dataset is provided. Supplementary materials, which include R code, are available online.

*Catalina Vallejos is Ph.D student (Email: cata.vallejos.m@gmail.com) and Mark Steel is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: m.steel@warwick.ac.uk). The authors are grateful to three referees and an Associate Editor for insightful and constructive comments. Catalina Vallejos acknowledges research support from the University of Warwick and from the Department of Statistics of the Pontificia Universidad Católica de Chile.

Keywords: Jeffreys prior; Outlier detection; Posterior existence; Set observations

1. INTRODUCTION

Frequently, standard survival models do not accommodate all features of real applications. Datasets often exhibit more "rare" or "tail" observations than predicted by usual models. Hence, models such as Weibull or log-normal lead to inference that is not robust to the presence of outliers (Barros et al., 2008). A second, related, issue is the existence of specific individual factors that result in heterogeneity of the survival times which can not be captured by covariates (Marshall and Olkin, 2007). Therefore, the typical assumption that the survival times correspond to realizations of random variables $T_1, \ldots, T_n$ which have the same "thin tailed" distribution (possibly depending on a set of covariates) can be inappropriate. An example of such a dataset is the Veterans' Administration (VA) lung cancer data presented in Kalbfleisch and Prentice (2002), for which the previous literature found strong evidence of influential observations and unobserved heterogeneity related to outliers (*e.g.* Barros et al., 2008; Heritier et al., 2009).

We consider the use of infinite mixture of lifetime distributions as a solution for these issues. Mixture modeling can be interpreted as the introduction of a random effect on the survival distribution. This idea has been mentioned by previous authors (*e.g.* Marshall and Olkin, 2007), but is not yet much used in applied work. In particular, this article explores the Shape Mixtures of Log-Normal (SMLN) distributions for which the shape parameter is assigned a mixing distribution. This new class covers a wide range of shapes, in particular cases with fatter tail behaviour than the log-normal. It includes the already studied log-Student t, log-Laplace, log-exponential power and log-logistic distributions among others. This paper puts the earlier literature into a common (more general) framework and develops objective Bayesian inference methods, which should be attractive for practitioners. The proposed priors do not require the elicitation of hyper-parameters and can be used in the (frequently encountered) setting in which no reliable prior information is available. They also provide baseline comparison when such

information is available.

Section 2 introduces the use of mixture families of distributions with particular focus on the SMLN family. Covariates are introduced through an Accelerated Failure Times model. The interpretation of the regression coefficients is not affected by the mixing distribution. This is an advantage over proportional hazards models with frailty terms, in which the interpretation of the regression parameters is conditional to the random effect, and the proportional hazards property is not preserved after mixture (Wienke, 2010). Section 3 analyzes aspects of Bayesian inference for models in the SMLN family. This addresses the existence of censored observations. Non-subjective priors based on the Jeffreys rule are proposed and conditions for the propriety of the posterior distribution are provided. In addition, we highlight that the use of point observations can affect the existence of the posterior distribution and a solution through set observations is considered. Section 3 also discusses some implementation details and proposes a method for outlier detection that exploits the mixture structure. A simulation study in Section 4 illustrates the performance of the proposed framework under different scenarios. We also show that, even for small sample size or a high proportion of censoring, standard Bayesian model comparison criteria can successfully detect departures from the log-normal model. In Section 5 we apply our models to the VA lung cancer dataset. SMLN models fit these data better than the log-normal model and uncover strong evidence of the presence of heterogeneity that is not accounted for by the available covariates. Finally, Section 6 concludes. All proofs are contained in the Appendix without mention in the text.

2. MIXTURES OF LIFE DISTRIBUTIONS

Let $T_1, \ldots, T_n$ be the survival times of n independent individuals. Usually, $T_1, \ldots, T_n$ are assumed to have the same "thin-tailed" distribution such as a log-normal or a Weibull (possibly depending on a set of covariates). However, this is often not appropriate in the face of unobserved heterogeneity between the survival times. This heterogeneity can be interpreted as

Table 1: Some SMLN models. $f_{PS}(\cdot|\delta)$ denotes a positive stable density with parameter δ.

Distribution	Density $f(t_i)$	Mixing density
Log-Student t	$\frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\pi\sigma^2\nu}}\frac{1}{t_i}\left[1+\frac{(\log(t_i)-\mu)^2}{\sigma^2\nu}\right]^{-(\frac{\nu}{2}+\frac{1}{2})}, \nu>0$	Gamma($\nu/2$, $\nu/2$)
Log-Laplace	$\frac{1}{2\sigma}\frac{1}{t_i}\exp\left\{-\frac{\lvert\log(t_i)-\mu\rvert}{\sigma}\right\}$	Inv-Gamma(1,1/2)
Log-exponential power	$\frac{\alpha}{2\sigma\Gamma(\frac{1}{\alpha})}\frac{1}{t_i}\exp\left\{-\left(\frac{\lvert\log(t_i)-\mu\rvert}{\sigma}\right)^{\alpha}\right\}, \alpha\in(1,2)$	$\frac{\Gamma(3/2)}{\Gamma(1+1/\alpha)}\lambda_i^{-\frac{1}{2}}\times f_{PS}(\lambda_i\lvert\frac{\alpha}{2})$
Log-logistic	$\frac{1}{\sigma\, e^{\mu}}\frac{(t_i/\, e^{\mu})^{1/\sigma-1}}{[1+(t_i/\, e^{\mu})^{1/\sigma}]^2}$	$\lambda_i^{-2}\sum_{k=0}^{\infty}\binom{-2}{k}(1+k)\, e^{-\frac{(1+k)^2}{2\lambda_i}}$

unobserved individual effects, and can also be related to the presence of outlying observations. This paper considers the use of mixtures of life distributions in order to account for unobserved heterogeneity and add robustness to the presence of outliers. The distribution of T_i is defined as a mixture of life distributions, if and only if its density function is

$$f(t_i|\psi,\theta) \equiv \int_{\mathcal{L}} f(t_i|\psi,\Lambda_i=\lambda_i)\, dP_{\Lambda_i}(\lambda_i|\theta), \tag{1}$$

where $f(\cdot|\psi,\Lambda_i=\lambda_i)$ represents the density function associated with a lifetime distribution which depends on the values of ψ and λ_i (underlying distribution), and λ_i can be understood as a random effect associated with each individual (frailty term). The mixing distribution has cumulative distribution function $P_{\Lambda_i}(\cdot|\theta)$ with support $\mathcal{L}$ and depends on a parameter θ. If $\mathcal{L}$ is a finite set of values, the distribution of T_i is a finite mixture of life distributions. However, here we focus on the case in which Λ_i is a continuous positive random variable (usually $\mathcal{L}=\mathbb{R}_+$), in which case $f(\cdot|\psi,\theta)$ can be interpreted as an infinite mixture of densities.

The intuition behind the underlying model carries over to these mixtures. Conditional on the mixing parameters, the base model applies. Therefore, if there are any theoretical or practical reasons underpinning this model (without mixing), the same reasons hold for the mixture model in the presence of unobserved heterogeneity. We specifically focus here on mixtures generated from log-normal distributions. The log-normal distribution arises as the limiting distribution when additive cumulative damage is the cause of the death or failure. Using the

previous argument, mixtures generated from log-normal distributions can be justified in the same context. Besides mixtures of log-normal distributions, a large number of mixture families can be generated using (1). For example, Jewell (1982) explores mixtures of exponential and Weibull distributions. Barros et al. (2008) and Patriota (2012) consider an extended class of Birnbaum-Saunders distributions that can be represented as in (1). The latter family is motivated by models for crack extensions where the mixing distribution accounts for dependence between the cracks.

2.1 The family of Shape Mixtures of Log-Normals

Definition 1. *A random variable T_i has a distribution in the family of Shape Mixtures of Log-Normals (SMLN) if and only if its density can be represented as*

$$f(t_i|\mu,\sigma^2,\theta)=\int_{\mathcal{L}} f_{LN}\left(t_i|\mu,\frac{\sigma^2}{\lambda_i}\right)\, dP_{\Lambda_i}(\lambda_i|\theta), \quad t_i>0, \mu\in\mathbb{R}, \sigma^2>0, \theta\in\Theta, \tag{2}$$

where $f_{LN}(\cdot|\mu,\frac{\sigma^2}{\lambda_i})$ corresponds to the density of a log-normal distribution with parameters μ and σ^2/λ_i, and λ_i is a realized value of a random variable Λ_i which has distribution function $P_{\Lambda_i}(\cdot|\theta)$ defined on $\mathcal{L}\subseteq\mathbb{R}_+$ (possibly discrete). Denote $T_i\sim SMLN_P(\mu,\sigma^2,\theta)$. Alternatively, (2) can be expressed as a hierarchical representation which corresponds to

$$T_i|\mu,\sigma^2,\Lambda_i=\lambda_i\sim LN\left(\mu,\frac{\sigma^2}{\lambda_i}\right), \qquad \Lambda_i|\theta\sim P_{\Lambda_i}(\cdot|\theta). \tag{3}$$

The SMLN family can be interpreted as a mixture of log-normal distributions with random shape parameter or as the exponential transformation of a random variable distributed as a scale mixture of normals. This family includes a number of distributions that have been proposed in the context of survival analysis. Table 1 lists some of them. In particular, the log-Student t distribution was introduced by Hogg and Klugman (1983) and the log-Laplace appeared in Uppuluri (1981). The log-exponential power was proposed by Vianelli (1983) and used in Martín and Pérez (2009). The log-logistic distribution was introduced by Shah and Dave (1963) and is used regularly in survival analysis, hydrology and economics. This list can be increased

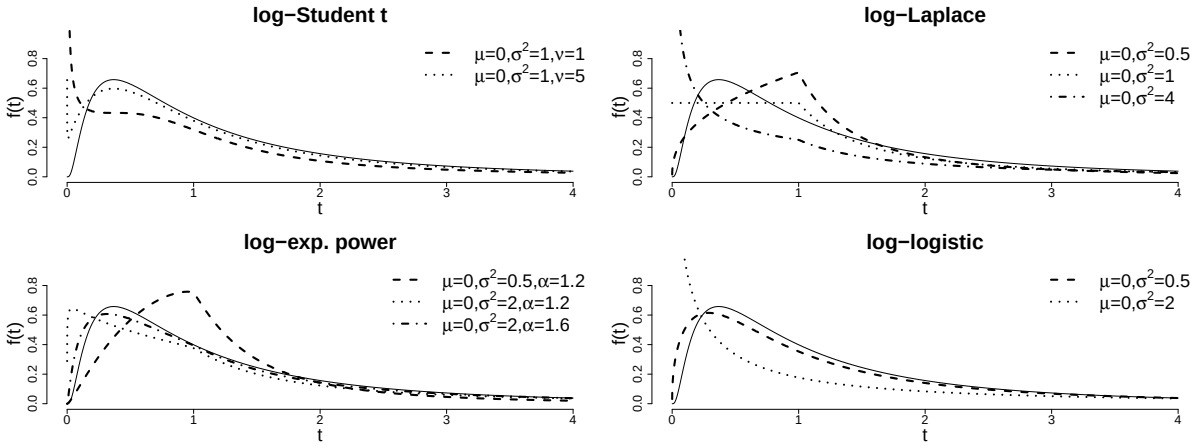

Figure 1: Density function of some SMLN models. Solid line is the log-normal$(0, 1)$ density.

by varying the mixing distribution. For example, all the mixing distributions used for scale mixtures of normals listed in Fernández and Steel (2000) can be used in this context. For identifiability reasons, the mixing distribution must not have separate unknown scale parameters (unknown scale parameters are allowed as long as they are linked to other features of the mixing distribution, *e.g.* $\Lambda_i \sim$ Gamma(θ,θ)). As illustrated in Figures 1 and 2, the SMLN family allows for a wide variety of shapes for the density and the hazard function. For example, while the hazard rate of the log-normal distribution has an increasing initial phase, the log-Laplace and log-logistic distributions produce a monotone decreasing hazard rate for some values of σ^2.

Whereas all positive moments exist for the log-normal, this is not necessarily the case for the shape mixtures: in particular, we can show that no positive moments exist for the log-Student t for any finite value of ν, and the log-Laplace and log-logistic models only allow for moments up to $1/\sigma$. The log-exponential power distribution with $\alpha > 1$ does possess all moments.

2.2 The AFT-SMLN model

An important aspect of survival modelling is the inclusion of covariates. Throughout, we will condition on the covariates which are assumed to be constant in time. An Accelerated Failure

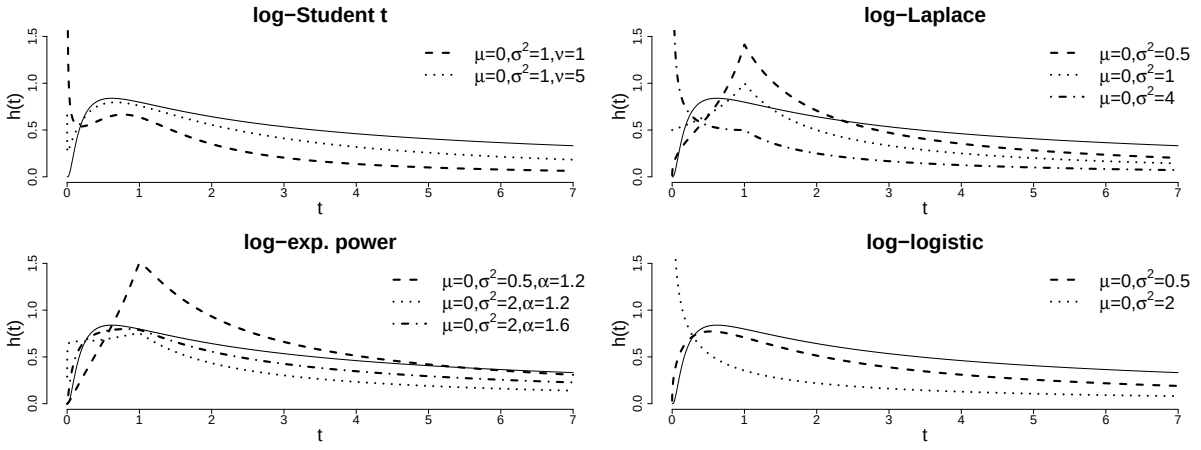

Figure 2: Hazard function of some SMLN models. Solid line is the log-normal(0,1) hazard rate.

Times (AFT) model is introduced. The AFT-SMLN model expresses the dependence between the covariates and the survival time by replacing the parameter μ with $x_i'\beta$, so that

$$T_i|x_i, \beta, \sigma^2, \theta \overset{ind}{\sim} SMLN_P(x_i'\beta, \sigma^2, \theta), \quad i = 1, 2, \ldots, n, \tag{4}$$

where x_i is a vector containing the value of k covariates associated with individual i and $\beta \in \mathbb{R}^k$ is a vector of parameters. This can also be interpreted as a linear regression model for the logarithm of the survival times with error term distributed as a scale mixture of normals. As the median of T_i in (4) is given by $e^{x_i'\beta}$, e^{β_j} is interpreted as the (proportional) marginal change of the median survival time as a consequence of a unitary change in covariate j. This interpretation is not affected by the mixing distribution.

3. BAYESIAN ANALYSIS OF THE AFT-SMLN MODEL

3.1 The prior

Bayesian inference will be conducted using *objective* priors that are generated by Jeffreys rule. This is one of the most common choices in the absence of prior information and has interesting

invariance and information-theoretic properties. The next theorem presents the Fisher information matrix for the AFT-SMLN model which is the basis for the Jeffreys-style priors.

Theorem 1. *Let $T_1, \ldots, T_n$ be independent random variables with T_i distributed according to (4), then its Fisher information matrix corresponds to*

$$I(\beta, \sigma^2, \theta) = \begin{pmatrix} \frac{1}{\sigma^2} k_1(\theta) \sum_{i=1}^n x_i x_i' & 0 & 0 \\ 0 & \frac{1}{\sigma^4} k_2(\theta) & \frac{1}{\sigma^2} k_3(\theta) \\ 0 & \frac{1}{\sigma^2} k_3(\theta) & k_4(\theta) \end{pmatrix}, \tag{5}$$

where $k_1(\theta), k_2(\theta), k_3(\theta)$ and $k_4(\theta)$ are functions depending only on θ.

The expressions involved in $k_1(\theta), k_2(\theta), k_3(\theta)$ and $k_4(\theta)$ are complicated (see the proof) and thus $I(\beta, \sigma^2, \theta)$ is not easily obtained from this theorem for any arbitrary mixing distribution. Indeed, it is usually more efficient to compute $I(\beta, \sigma^2, \theta)$ directly from $f(\cdot|\beta, \sigma^2, \theta)$. However, this structure facilitates a general representation of the Jeffreys-style priors:

Corollary 1. *Under the same assumptions as in Theorem 1 it follows that the Jeffreys, independence Jeffreys (which deals separately with the blocks for β and (σ^2, θ)) and independence I Jeffreys (which deals separately with β, σ^2 and θ) priors are respectively given by*

$$\pi^J(\beta, \sigma^2, \theta) \propto \frac{1}{(\sigma^2)^{1+\frac{k}{2}}} \sqrt{[k_1(\theta)]^k [k_2(\theta) k_4(\theta) - k_3^2(\theta)]}, \tag{6}$$

$$\pi^I(\beta, \sigma^2, \theta) \propto \frac{1}{\sigma^2} \sqrt{k_2(\theta) k_4(\theta) - k_3^2(\theta)}, \tag{7}$$

$$\pi^{II}(\beta, \sigma^2, \theta) \propto \frac{1}{\sigma^2} \sqrt{k_4(\theta)}. \tag{8}$$

The three non-subjective priors presented here can be written as

$$\pi(\beta, \sigma^2, \theta) \propto \frac{1}{(\sigma^2)^p} \pi(\theta), \tag{9}$$

where $\pi(\theta)$ is the factor of the prior that depends on θ. For the Jeffreys prior $p = 1 + (k/2)$ and $p = 1$ for the other two priors. If θ does not appear (*e.g.* log-normal, log-Laplace and log-logistic models) this prior simplifies to $\pi(\beta, \sigma^2) \propto (\sigma^2)^{-p}$.

Note that the result in Corollary 1 also specifies the prior for θ. The implied priors for the special cases of the log-Student t and the log-exponential power (derived directly from the specific likelihood functions) are explicitly presented in the proof of Theorem 3. In order to obtain meaningful Bayes factors between models, priors with a improper component $\pi(\theta)$ for θ are discarded. For the examples explored throughout this article, this argument discards the independence I Jeffreys prior for the log-Student t model.

3.2 The posterior distribution

The three priors presented in Corollary 1 do not correspond to proper probability distributions and therefore the propriety of the posterior distribution must be verified. At this stage we also introduce the existence of censoring (assumed to be non-informative). In the following, posterior propriety is verified for the AFT-SMLN model under the priors in Corollary 1.

Theorem 2. *Let $t_1, \ldots, t_n > 0$ be the survival times (possibly censored) of n independent individuals, realizations of random variables distributed as in (4). Assume the prior given in (9), with $\int_\Theta \pi(\theta)\, d\theta = 1$. Without loss of generality, assume that only the first n_o observations are uncensored. Define $X_o = (x_1, \ldots, x_{n_o})'$ and suppose that the rank of X_o is k.*

(i) For $p = 1$, a sufficient condition for posterior existence is $n_o > k$,

(ii) For $p = 1 + k/2$, a sufficient condition for the posterior propriety is $n_o > k$ and

$$\int_\Theta E(\Lambda_1^{-\frac{k}{2}}|\theta)\pi(\theta)\, d\theta < \infty. \tag{10}$$

Theorem 3. *Under the assumptions in Theorem 2 and provided that $n_o > k$, it follows that*

(i) For the log-Student t AFT model, the posterior is proper under the independence Jeffreys prior. However, the posterior does not exist for the Jeffreys prior.

(ii) For the log-Laplace AFT model, log-exponential power AFT model and log-logistic AFT model, the propriety of the posterior can be verified with any of the three proposed priors.

Theorem 3 tells us that the log-Student t AFT model does not lead to valid Bayesian inference in combination with the Jeffreys prior (the independence I Jeffreys prior was already discarded in Subsection 3.1). The other models can be combined with all priors considered here; of course, the absence of θ in the log-Laplace and log-logistic models implies that the independence Jeffreys and independence I Jeffreys priors coincide in those cases.

3.3 The problem with using point observations

Continuous sampling models assign zero probability to particular (point) values. In spite of this, the standard statistical analysis is based on point observations. This situation can cause problems in the context of Bayesian inference. The assessment of the propriety of the posterior distribution is usually conducted without taking into account events that have zero probability. Hence, the propriety of the posterior on the basis of a specific sample of point observations can be precluded. As argued in Fernández and Steel (1998), this issue introduces the risk of having senseless inference. The following theorem illustrates the problem of the use of point observations in the context of the AFT log-Student t model.

Theorem 4. *Adopt the same assumptions as in Theorem 2 and assume that $n_o > k$. If the mixing distribution is Gamma($\nu/2$,$\nu/2$) and s ($k \leq s < n_o$) is defined as the largest number of uncensored observations that can be represented as an exact linear combination of their covariates (i.e. $\log(t_i) = x_i'\beta$ for some fixed β), a necessary condition for the propriety of the posterior distribution of (β, σ^2, ν) is*

$$\int_0^m \pi(\nu)\, d\nu = 0, \quad where \quad m = \frac{n_o - k + (2p-2)}{n_o - s} - 1. \tag{11}$$

This result indicates that it is possible to have samples of point observations for which no Bayesian inference can be conducted, unless $\pi(\nu)$ induces a positive lower bound for ν. For the log-Student t model we only use the independence Jeffreys prior, so that $p = 1$ and (11) is violated whenever $s > k$. When no covariates are taken into account ($k = 1$), s coincides with

the largest number of (uncensored) tied observations. Theorem 4 highlights the need for considering sets of zero Lebesgue measure when checking the propriety of the posterior distribution based on point observations.

In the context of scale mixture of normals, Fernández and Steel (1998, 1999) proposed the use of set observations as a solution to this problem. We now extend this to the SMLN family. The use of set observations is based on the fact that, in practice, it is impossible to record observations from continuous random variables with total precision and every observation can only be considered as a label of a set of positive Lebesgue measure. For instance, if the recorded value for the survival time is t_i, it really means that the actual survival time is between $t_i - \epsilon_l$ and $t_i + \epsilon_r$ where ϵ_l and ϵ_r are determined by the accuracy with which the data was recorded (*e.g.* if the data is recorded in integers, $\epsilon_l = \epsilon_r = 0.5$). This is equivalent to considering the observation as interval censored. As explained in Subsection 3.4, the implementation is done through data augmentation which does not involve a large increase of the computational cost. This procedure also naturally deals with left or right censored observations by taking, respectively, $(\epsilon_l, \epsilon_r) = (t_i, 0)$ or $(\epsilon_l, \epsilon_r) = (0, \infty)$. The following theorem indicates that the use of set observations can ensure a proper posterior distribution in situations where a particular sample of point observations might not.

Theorem 5. *Adopt the same assumptions as in Theorem 2 and assume that $n_o > k$. Replace the uncensored observations by set observations $t_\epsilon = \{(t_1 - \epsilon_l, t_1 + \epsilon_r), \ldots, (t_{n_o} - \epsilon_l, t_{n_o} + \epsilon_r)\}$ ($0 < \epsilon_l, \epsilon_r < \infty$). Define $E = (t_1 - \epsilon_l, t_1 + \epsilon_r) \times (t_2 - \epsilon_l, t_2 + \epsilon_r) \times \cdots \times (t_{n_o} - \epsilon_l, t_{n_o} + \epsilon_r)$. The posterior distribution of $(\beta, \sigma^2, \theta)$ given t_ϵ is proper if and only if the marginal likelihood under point observations is finite for any $t_o \in E$, except for a set of zero Lebesgue measure.*

3.4 Implementation

Bayesian inference was implemented through Markov chain Monte Carlo (MCMC) using the hierarchical representation (3) of the SMLN family and the data augmentation idea of Tanner

and Wong (1987). Throughout, we use the prior presented in (9). An Adaptive Metropolis-within-Gibbs sampling scheme with Gaussian Random Walk proposals is used (Roberts and Rosenthal, 2009). Both censored and set observations are accommodated through data augmentation (as in Fernández and Steel, 1999, 2000). This introduces an additional step in the sampler in which, given the current value of the parameters and mixing variables, point values of the survival times in line with the set observations are simulated. In this case, this can be easily done by sampling $\log(t_i)$ from a truncated normal distribution. Regardless of the mixing distribution and provided that $n > 2 - 2p$, the full conditionals for β and σ^2 are normal and inverted gamma distributions. However, the full conditionals for $\Lambda_1, \ldots, \Lambda_n$ and θ are generally not of a known form for arbitrary mixing distributions. For the Λ_i's we have $\pi(\lambda_1, \ldots, \lambda_n|\beta, \sigma^2, \theta, t) = \prod_{i=1}^n \pi(\lambda_i|\beta, \sigma^2, \theta, t)$ where $t = (t_1, \ldots, t_n)'$ are the simulated survival times and

$$\pi(\lambda_i|\beta, \sigma^2, \theta, t) \propto \lambda_i^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\lambda_i(\log(t_i) - x_i'\beta)^2\right\} dP(\lambda_i|\theta),\ i = 1, \ldots, n. \tag{12}$$

For the log-Student t and log-Laplace models, (12) has a known form. In the first case it is a Gamma$((\nu+1)/2$,$1/2\left[\{(\log(t_i) - x_i'\beta)^2/\sigma^2\} + \nu\right])$. In the second, it corresponds to an Inverse Gaussian$(\sigma/|\log(t_i) - x_i'\beta|$,1). If the mixing distribution has no closed form (as *e.g.* for the log-exponential power and log-logistic distributions), the acceptance probability in these Metropolis-Hastings steps is not easily computable. For the log-logistic model we implement the rejection sampling algorithm proposed in Holmes and Held (2006, p.163), using the fact that $(2\sqrt{\Lambda_i})^{-1}$ has the asymptotic Kolmogorov-Smirnov distribution. In the case of the log-exponential power model we adopt the mixture of uniforms representation used in Martín and Pérez (2009). This replaces the use of Λ_i by U_i $(i = 1, \ldots, n)$, with $U_i \overset{iid}{\sim}$ Gamma$(1 + 1/\alpha, 1)$ and $\log(T_i)|U_i = u_i, \beta, \sigma^2, \alpha \sim$ U$(x_i'\beta - \sigma u_i^{1/\alpha}, x_i'\beta + \sigma u_i^{1/\alpha})$. We restrict the range of α to $(1, 2)$, which is consistent with the SMLN representation. The case $\alpha = 1$ is excluded but is covered by the log-Laplace model. We decompose the posterior distribution $\pi(\beta, \sigma^2, \alpha, u_1, \ldots, u_n|t)$ as $\prod_{i=1}^n \pi(u_i|t, \beta, \sigma^2, \alpha) \times \pi(\beta, \sigma^2, \alpha|t)$ and the full conditionals for the U_i's are

$$\pi(u_i|t, \beta, \sigma^2, \alpha) \propto\ e^{-u_i},\ u_i > \left(\frac{|\log(t_i) - x_i'\beta|}{\sigma}\right)^{\alpha},\ i = 1, \ldots, n. \tag{13}$$

For $\pi(\beta, \sigma^2, \alpha|t)$, the full conditionals of β, σ^2 and α can easily be derived from the marginal likelihood (after integrating out the u_i's) of t given $(\beta, \sigma^2, \alpha)$. None of them has a known form and Adaptive Metropolis-Hastings steps are implemented. Additionally, for censored and set observations, $\log(t_i)$ is sampled from a truncated uniform distribution. In terms of setting up a sampler, it might be easier to simply start directly from the log-exponential power or log-logistic distribution, rather than its interpretation as a scale mixture. However, we would lose the inference on the mixing variables Λ_i (or U_i) which is particularly important in identifying outlying observations (Lange et al., 1989, Fernández and Steel, 1999). This is further discussed in Subsection 3.6. Also, the use of these mixing representations facilitates dealing with censored and set observations. See Supplementary material A for more details and R code.

3.5 Model comparison

We consider several standard model comparison criteria. Firstly, we use Bayes factors (BF), defined as the ratio between the marginal likelihoods of the models. Marginal likelihoods will be computed using the methodology proposed by Chib (1995) and Chib and Jeliazkov (2001). The latter was developed for a non-adaptive scheme. Using the stabilized proposal variances, we estimate the marginal likelihood from shorter nonadaptive chains for which the starting values are defined as the converged parameter values of the original chains. Secondly, the Deviance Information Criteria (DIC) developed by Spiegelhalter et al. (2002) is also provided. It is given by DIC $= \mathrm{E}(D(\beta, \sigma^2, \theta, y)|y) + p_D$ with $p_D = \mathrm{E}(D(\beta, \sigma^2, \theta, y)|y) - D(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}, y)$ (effective number of parameters), $D(\beta, \sigma^2, \theta, y) = -2\log(f(y|\beta, \sigma^2, \theta))$ (deviance function) and where $\hat{\beta}, \hat{\sigma^2}$ and $\hat{\theta}$ are the posterior medians of β, σ^2 and θ, respectively. This is computed using the marginal likelihood (integrating out the mixing parameters). Low DIC suggests a better model. In addition, models are compared by the quality of their predictions. We use the Conditional Predictive Ordinate (CPO) (Geisser and Eddy, 1979). For observation i, CPO$_i$ is defined as

$$\text{CPO}_i = f(t_i|t_{-i}) = \left[\mathrm{E}\left(\frac{1}{f(t_i|\beta, \sigma^2, \theta)}\right)\right]^{-1}, \qquad t_{-i} = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n), \quad (14)$$

where the expectation is with respect to $\pi(\beta, \sigma^2, \theta|t)$ and $f(\cdot|t_{-i})$ is the predictive density given t_{-i}. The density function $f(\cdot|t_{-i})$ is replaced by the survival function $S(\cdot|t_{-i})$ for right censored observations (Banerjee et al., 2007; Hanson, 2006). Larger values of CPO_i indicate better predictive accuracy for the observation i. Geisser and Eddy (1979) also proposes $\text{PsML} = \prod_{i=1}^{n} \text{CPO}_i$ as an estimator of the marginal likelihood (also called Pseudo Marginal Likelihood). Higher values of PsML indicate a better overall predictive performance of the model. Pseudo Bayes factors (PsBF) can be easily computed as ratios of PsML's.

3.6 Detection of influential observations and outliers

A robust model will have no (or few) influential observations. Influential observations can be detected using $K_i = \text{KL}(\pi(\beta, \sigma^2, \theta|t), \pi(\beta, \sigma^2, \theta|t_{-i}))$, where $\text{KL}(\cdot, \cdot)$ denotes the Kullback-Leibler divergence function (Peng and Dey, 1995; Cho et al., 2009). As suggested in McCulloch (1989), we transform K_i in terms of its calibration index $p_i = 0.5\left[1 + \sqrt{1 - \exp\{-2K_i\}}\right]$, $p_i \in [0.5, 1]$. In relation to the Kullback-Leibler divergence, the effect of removing observation i is equivalent to assigning probability p_i to an event which has true probability 0.5. A large value of p_i suggests that observation i is influential.

In addition, the existence of outlying observations will be assessed using the posterior distribution of the mixing variables. Extreme values (with respect to a reference value, λ_{ref}) of the mixing variables are associated with outliers (see also West, 1984). Formally, evidence of outlying observations will be assessed by contrasting the models $M_0 : \Lambda_i = \lambda_{ref}$ versus $M_1 : \Lambda_i \neq \lambda_{ref}$ (with all other $\Lambda_j, j \neq i$ free). Evidence in favour of each of these models will be measured using Bayes factors, which can be computed as the generalized Savage-Dickey density ratio proposed in Verdinelli and Wasserman (1995). The evidence in favour of M_0 versus M_1 (*i.e.* against observation i being an outlier) is

$$\text{BF}_{01} = \pi(\lambda_i|t)\text{E}\left(\frac{1}{dP(\lambda_i|\theta)}\right)\Bigg|_{\lambda_i=\lambda_{ref}}, \tag{15}$$

where the expectation is with respect to $\pi(\theta|t, \Lambda_i = \lambda_{ref})$. When the parameter θ does not

appear in the model, this simplifies to the original Savage-Dickey density ratio

$$\text{BF}_{01} = \frac{\pi(\lambda_i|t)}{dP(\lambda_i)}\bigg|_{\lambda_i=\lambda_{ref}} = \text{E}\left(\frac{\pi(t_i|\beta,\sigma^2,\lambda_i)}{\pi(t_i|\beta,\sigma^2)}\right)\bigg|_{\lambda_i=\lambda_{ref}}, \tag{16}$$

where the expectation is with respect to $\pi(\beta,\sigma^2|t)$. The main challenge of this approach is the choice of λ_{ref}. Intuitively, if there is no unobserved heterogeneity, the posterior distribution of the mixing parameters should not be much affected by the data. Therefore, the mixing distribution (which can be interpreted as a prior distribution for Λ_i) will then be close to the posterior distribution of Λ_i. Following this intuition, we propose to use $E(\Lambda_i|\theta)$ (if it exists) as λ_{ref}. Using this rule, $\lambda_{ref} = 1$ for the log-Student t model. This choice was supported by our empirical examples. If $E(\Lambda_i|\theta)$ depends on θ (unknown), we suggest estimating it through the posterior median of θ. Examples for which $E(\Lambda_i|\theta)$ is not finite require a more detailed analysis. For example, the expectation of the mixing distribution that generates the log-Laplace and log-logistic distributions do not exist. For the log-Laplace model, simulated datasets indicate a large heterogeneity between the posterior distributions of the Λ_i's and the existence of a unique reference value is not clear (even in the absence of outlying observations). However, the average of the posterior medians of the mixing distribution is close to unity for all simulated data sets we have tried. In this calculation, we discard the lowest 25% of λ_i values in order to remove the influence of any possible outliers. Hence, we propose $\lambda_{ref} = 1$ for the log-Laplace model. In the log-logistic case, the posterior distributions of the Λ_i's behave as in the log-Student t case, where the reference value is clearer. We chose $\lambda_{ref} = 0.4$ for the log-logistic model, using the same argument as in the log-Laplace case. Figure 3 shows the performance of the reference values by plotting the Bayes factor in (15) for the log-Student t and in (16) for the log-Laplace and log-logistic models against a standardized log survival time z (given β, σ^2 and θ). This is defined as $\log(t)$ minus its mean, divided by its standard deviation (*i.e.* $\sigma\sqrt{\text{E}_\Lambda(1/\Lambda|\theta)}$). For the log-Student t, log-Laplace and log-logistic models, $z = \frac{\log(t)-x'\beta}{\sigma}\sqrt{\frac{\nu-2}{\nu}}$ (for $\nu > 2$), $z = \frac{\log(t)-x'\beta}{\sigma}\frac{1}{\sqrt{2}}$ and $z = \frac{\log(t)-x'\beta}{\sigma}\frac{\sqrt{3}}{\pi}$, respectively. As expected, large values of $|z|$ lead to evidence in favour of an outlier. The log-Student t model with very large number of degrees of freedom requires exceptionally large $|z|$ values to distinguish it from the log-normal case.

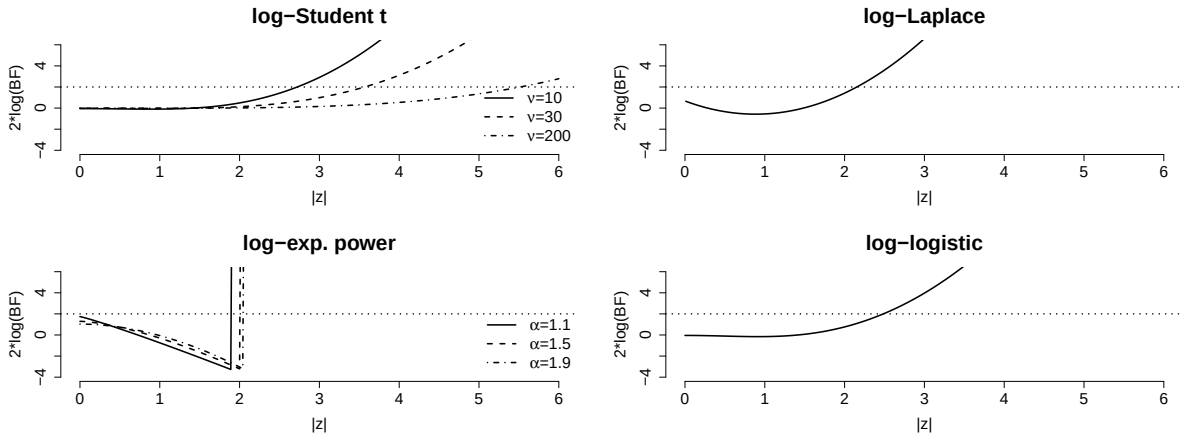

Figure 3: Bayes factor for outlier detection as a function of $|z|$. The log Bayes factor has been re-scaled by 2 in order to apply the interpretation rule proposed in Kass and Raftery (1995). The dotted horizontal line is the threshold above which observations will be considered outliers.

The log-exponential power model is a special case. As explained in Subsection 3.4, Bayesian inference for this model is implemented through a mixture of uniforms representation with mixing parameters denoted by U_i. The models for outlier detection in terms of U_i are $M_0 : U_i = u_{ref}$ versus $M_1 : U_i \neq u_{ref}$. The expectation of U_i given α is $1 + 1/\alpha$ and, according to the intuition presented previously, we might use this value as u_{ref}. With this rule, u_{ref} is a function of α which lies in $(1.5, 2)$. In practice, this choice detected large amounts of outliers (even for datasets generated from the log-normal model). We estimate $\pi(u_{ref}|t)$ by averaging $\pi(u_{ref}|t, \beta, \sigma^2, \alpha)$ in (13) using an MCMC sample from the posterior distribution of $(\beta, \sigma^2, \alpha)$. Hence, if the value of $(\beta, \sigma^2, \alpha)$ is such that $u_{ref} \leq \left(\frac{|\log(t_i) - x_i'\beta|}{\sigma}\right)^{\alpha}$, $\pi(u_{ref}|t, \beta, \sigma^2, \alpha)$ is equal to zero and the Bayes factor in favour of the observation i being an outlier is computed as infinity. Simulated datasets indicated that the means and medians of $\left(\frac{|\log(t_i) - x_i'\beta|}{\sigma}\right)^{\alpha}, i = 1, \ldots, n$, are around 0.6, regardless of the model from which the data was generated. We then adjust the reference value to $u_{ref} = 1 + 1/\alpha + 0.6$. This choice performed much better with simulated datasets (e.g. using log-normal data no outliers were detected). The resulting Bayes factors as a function of $z = \frac{\log(t) - x'\beta}{\sigma}\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$ (see Figure 3) are not much affected by the value of α.

For all models, moderate changes to the reference values do not have a large impact on the outlier detection curves in Figure 3.

4. SIMULATION STUDY

A simulation study is designed in order to evaluate the performance of the proposed mixture scheme versus a log-normal model (with no unobserved heterogeneity) under different scenarios. Two independent covariates, $x_1 \sim \text{Ber}(0.5)$ and $x_2 \sim \text{Unif}(0,1)$ are simulated, and an intercept is added ($k = 3$). Throughout, we use $\beta = (4, 0.5, -1)'$ and $\sigma^2 = 0.1$ (which are in the range of usual empirical values). Datasets are simulated from the following models: (i) log-normal, (ii) log-Student t with $\nu = 5$, (iii) log-Student t with $\nu = 20$, (iv) log-Laplace, (v) log-exponential power with $\alpha = 1.2$, (vi) log-exponential power with $\alpha = 1.8$ and (vii) log-logistic. Four different scenarios are defined through sample size ($n = 100, 500$) and percentage of censoring ($PC = 10\%, 70\%$). These rather small sample sizes are often observed in survival datasets. For each model, 100 independent datasets are simulated under each scenario. In all cases, survival times are rounded to integers in order to reflect the usual inaccuracy in the data recording process. Independent censoring times are sampled from a uniform distribution in $(0, C)$ where the value of C is tuned to control the percentage of censoring. Detailed results of the simulation study are displayed in supplementary material B.

For AFT models, β is usually the parameter of interest and its interpretation is not affected by our mixing scheme. We compare the performance of different SMLN-AFT models based on the posterior median of β. The choice between one of the three Jeffreys-rule based priors suggested in Subsection 3.1 is not very critical for the estimation of β as all priors produce very similar inference for the regression parameter. Of course, estimation is more accurate when the data provides more information, *i.e.* for $n = 500$ and $PC = 10\%$. There are no major differences between log-normal datasets and those generated by a SMLN model with weak unobserved heterogeneity (log-Student t with $\nu = 20$ and log-exponential power with

$\alpha = 1.8$). In such cases, the log-normal model correctly estimates β. Crucially, fitting SMLN models to log-normal datasets is harmless. The β estimates are concentrated around the true value, although they are slightly more spread out when using a log-Laplace model (which has a very dispersed mixing distribution). As expected, if the data display stronger unobserved heterogeneity, mixture models tend to outperform the log-normal one. For those cases, SMLN models produce more accurate estimates of β in terms of both bias and spread, especially under large amounts of censoring. This is even the case when using a different mixing distribution than the one that generated the data. These differences are largest for the log-Laplace datasets and diminish for milder cases of unobserved heterogeneity, like the log-logistic case.

The Bayesian model comparison criteria described in Subsection 3.5 are applied to each dataset in order to assess their effectiveness. Here we focus on the Jeffreys and independence Jeffreys priors (both types of independence Jeffreys priors lead to similar results). The performance of BF is better (and more in line with the other criteria) under the independence Jeffreys prior, except for the log-logistic data. Under the Jeffreys prior and with log-normal data, BF assigns relatively little support to the log-normal model when $n = 100$ (especially with high PC). For $k = 3$, the Jeffreys prior favours small values of σ^2, much more than the independence Jeffreys (the difference increases with k). When the dataset provides little information (small n and/or large PC), the prior has a strong influence on posterior inference. We might, thus, underestimate σ^2 and the fitted log-normal model will have too small a spread to accommodate the data, even though they were generated by the log-normal model. Predictive criteria are less affected by this. Overall, DIC, BF and PsBF point in the same direction, largely successfully detecting the presence and absence of unobserved heterogeneity. However, very mild forms of unobserved heterogeneity (log-Student t with $\nu = 20$, log-exponential power with $\alpha = 1.8$) are often indistinguishable from the log-normal model. Stronger unobserved heterogeneity is more easily detected (even when $n = 100$ and $PC = 70\%$). Jointly, these criteria successfully indicate the existence of unobserved heterogeneity. Even in the worst scenario, the log-normal model is correctly detected more than 60% of the time if we use the independence Jeffreys prior.

Distinguishing between the different mixing distributions is more demanding, but can be achieved for large sample sizes. The best results are observed for the independence Jeffreys prior. In this case, we correctly classify data generated by the log-Laplace model in at least 60% of the cases when $n = 100$ and at least 82% of the cases for $n = 500$. With log-logistic datasets, the right model is detected in at least 70% of the simulations with $n = 500$ under either prior. The rate of correct detection is lower for the log-Student t and log-exponential power models, for which an extra parameter needs to be estimated. The DIC and PsBF criteria do best overall: under both priors they correctly identify models with moderate or strong heterogeneity on the basis of 500 observations with low censoring in at least 57% of the cases.

5. APPLICATION: THE VA LUNG CANCER TRIAL

The dataset (presented in Kalbfleisch and Prentice, 2002) relates to a trial in which a therapy (standard or test chemotherapy) was randomly applied to 137 patients who were diagnosed with inoperable lung cancer. The survival times of the patients were measured in days since treatment and the following covariates were used: the treatment that is applied to the patient (0: standard, 1: test); the histological type of the tumor (squamous, small cell, adeno, large cell); a continuous index representing the status of the patient at the moment of the treatment (the higher the index, the better the patient's condition); the time between the diagnosis and the treatment (in months); age (in years); and a binary indicator of prior therapy (0: no, 1: yes). The data contain 9 right censored observations. This dataset has been previously analyzed from a frequentist point of view using traditional models such as the Cox, Weibull, log-normal and log-logistic regressions (see Lee and Wang, 2003; Heritier et al., 2009). These models all suggest that the status of the patient at the moment of treatment and the histological type of the tumor are the relevant explanatory variables for the survival time. Nevertheless, evidence of influential observations has been found. Barros et al. (2008) illustrated that the inference produced by a log-Birnbaum-Saunders model is greatly modified when dropping observations 77, 85 and 100.

They proposed a log-Birnbaum-Saunders Student t distribution as a more robust alternative for this dataset because it allows for fatter tails and accommodates heterogeneity in the data. This distribution can also be represented through a mixture family as in (1), so our methodology could be also extended to include this distribution. Heritier et al. (2009) detected observations 17 and 44 as influential when fitting a Cox proportional hazard model and proposed the use of an adaptive robust estimator instead.

Table 2 presents a summary of the posterior distribution of (β, σ^2) when a log-normal AFT model is fitted. This is based on 10,000 draws, recorded from a total of 400,000 iterations with a burn-in period of 200,000 and a thinning of 20. The use of different starting points and the usual convergence criteria strongly suggest convergence of the chains (see supplementary material C). The Jeffreys and the independence Jeffreys prior produced similar results. Bayesian inference was conducted on the basis of point and set observations, using $\epsilon_l = \epsilon_r = 0.5$ for uncensored observations. For this model, inference on point and set observations is quite similar. The use of point observations does not produce problems for the log-normal model. However, we know that set observations avoid potential problems with the inference for other models (see Subsection 3.3), so we will focus on the analysis with set observations in the rest of the paper. Results suggest that the main covariate effects are due to the tumour type and patient status, and are roughly in line with the maximum likelihood results for the log-Birnbaum Saunders AFT model in Barros et al. (2008), although the effect of the test treatment is less clearly negative.

We then use the AFT-SMLN model (4) with the continuous mixing distributions presented in Table 1. Bayesian inference is conducted under the Jeffreys-type priors introduced in Corollary 1. We adopted the same total number of iterations, burn in and thinning as for the log-normal model. We compare the models through Bayes factors, Pseudo Bayes factors, DIC and the CPO predictive performance, summarized in Figure 4 and Table 3. Clearly, all these criteria provide evidence in favour of mixture models. For the log-Student t model, this evidence is also supported by the fact that inference on ν favours relative small values. Similarly, the log-exponential power model suggests values of α far from 2. From plots (unreported) of the prior

Table 2: Summary of the posterior distribution of the parameters of the log-normal AFT model. β_0: Intercept, β_1: Treat (test), β_2: Type (squamous), β_3: Type (small cell), β_4: Type (adeno), β_5: Status, β_6: Time from diagnosis, β_7: Age, β_8: Prior therapy (yes).

	Jeffreys prior				Independence Jeffreys prior			
	Point Observations		Set Observations		Point Observations		Set Observations	
	Median	HPD 95%	Median	HPD 95%	Median	HPD 95%	Median	HPD 95%
β_0	1.82	[0.50, 3.08]	1.79	[0.47, 3.10]	1.82	[0.42, 3.14]	1.80	[0.42, 3.09]
β_1	-0.17	[-0.53, 0.22]	-0.17	[-0.56, 0.21]	-0.17	[-0.56, 0.21]	-0.17	[-0.57, 0.22]
β_2	-0.12	[-0.65, 0.46]	-0.12	[-0.67, 0.45]	-0.11	[-0.68, 0.46]	-0.12	[-0.71, 0.45]
β_3	-0.73	[-1.28,-0.22]	-0.72	[-1.26,-0.19]	-0.73	[-1.29,-0.21]	-0.73	[-1.27,-0.19]
β_4	-0.77	[-1.32,-0.16]	-0.77	[-1.37,-0.18]	-0.78	[-1.37,-0.13]	-0.77	[-1.37,-0.15]
β_5	0.04	[0.03, 0.05]	0.04	[0.03, 0.05]	0.04	[0.03, 0.05]	0.04	[0.03, 0.05]
β_6	0.00	[-0.02, 0.02]	0.00	[-0.02, 0.02]	0.00	[-0.02, 0.02]	0.00	[-0.02, 0.02]
β_7	0.01	[-0.01, 0.03]	0.01	[0.00, 0.03]	0.01	[0.00, 0.03]	0.01	[-0.01, 0.03]
β_8	-0.11	[-0.54, 0.34]	-0.11	[-0.57, 0.34]	-0.11	[-0.59, 0.35]	-0.11	[-0.57, 0.37]
σ^2	1.12	[0.87, 1.42]	1.13	[0.88, 1.43]	1.20	[0.90, 1.53]	1.21	[0.92, 1.54]

and posterior distributions it is clear they differ and that the latter is strongly driven by the data itself. Overall, the log-logistic model seems the best candidate for fitting this dataset. This is in line with the results in Lee and Wang (2003) in which, using a maximum likelihood approach, the log-logistic model is preferred to the log-normal and other standard models.

The choice of prior and mixing distribution is not too critical for the inference about β (only results under the independence Jeffreys prior are reported). For mixture models, the posterior distribution of β is somewhat different from that for the log-normal model (see Table 4). In particular, the effect of the test treatment is less pronounced. The results on β are relatively close to the classical ones reported in Barros et al. (2008) using the log-Birnbaum Saunders Student t model and to the ones in Lee and Wang (2003) using the log-logistic model. Estimates of σ^2 cannot be compared because it has a different interpretation for each model.

Table 3 also indicates that the number of influential observations is smaller for the SMLN

Table 3: DIC, the fraction of observations with better CPO performance than the log-normal model, and the number of influential observations.

Prior	Model	DIC	CPO better than LN	No. obs. $p_i \geq 0.8$	No. obs. $p_i \geq 0.9$
	Log-normal	1449.01	-	6	2
	Log-Student t	-	-	-	-
Jeffreys	Log-Laplace	1444.18	52%	2	1
	Log-exp. power	1444.00	54%	3	1
	Log-logistic	1444.14	66%	3	1
	Log-normal	1449.56	-	6	2
	Log-Student t	1445.86	64%	3	1
Ind. Jeffreys	Log-Laplace	1444.37	53%	1	1
	Log-exp. power	1444.79	55%	3	1
	Log-logistic	1444.49	66%	3	1
	Log-normal	1449.56	-	6	2
	Log-Student t	-	-	-	-
Ind. I Jeffreys	Log-Laplace	1444.37	53%	1	1
	Log-exp. power	1444.81	55%	3	1
	Log-logistic	1444.49	66%	3	1

models than for the log normal model, which is consistent with the superior ability of the SMLN models to accommodate unusual observations. Disregarding the prior, observations 12, 77, 85, 95, 100 and 106 are detected as influential observations for the log-normal model (with no mixture) when using the threshold $p_i \geq 0.8$ (in fact, $p_i \geq 0.9$ for observations 85 and 106). In contrast, for all the mixture models, observations 77, 95 and 100 do not appear as influential observations (both thresholds). Despite of the mixture, observation 106 is always considered as influential ($p_i \geq 0.9$). Observations 12 and 85 are pointed as influential by some of the mixtures. These results are roughly in line with the results in Barros et al. (2008), where observations 77, 85 and 100 were also identified as (strong) influential observations when fitting a log-Birnbaum-Saunders model. Using a log-Birnbaum-Saunders-t model they also label observations 12, 77, 95 and 106 as (mild) influential observations.

The posterior distributions of the mixing parameters (not reported) vary substantially between the patients, suggesting heterogeneity in the data. Figure 5 formalizes this by presenting

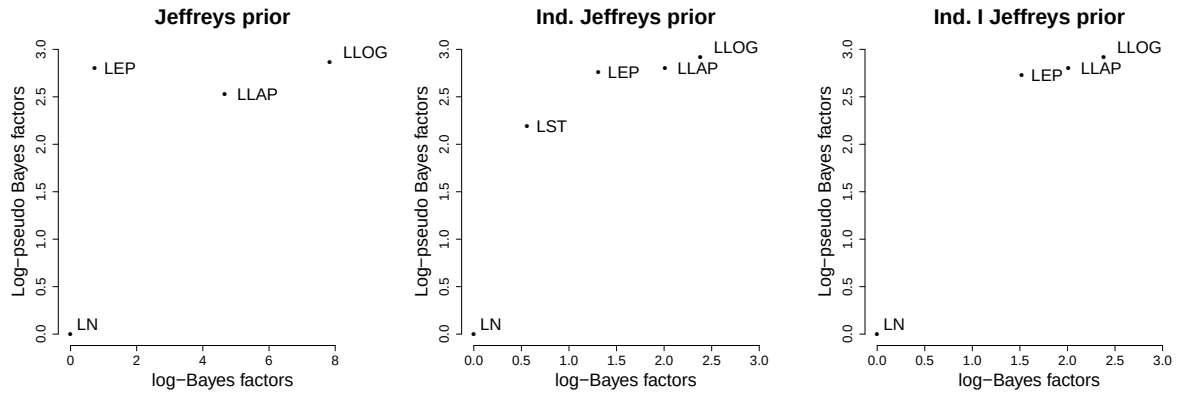

Figure 4: Bayes factors and pseudo Bayes factors of each model with respect to the log-normal one.

Table 4: Summary of the posterior distribution under the independence Jeffreys prior for various SMLN models on the basis of set observations. β_0: Intercept, β_1: Treat (test), β_2: Type (squamous), β_3: Type (small cell), β_4: Type (adeno), β_5: Status, β_6: Time from diagnosis, β_7: Age, β_8: Prior therapy (yes).

	Log-Student t		Log-Laplace		Log-exp. power		Log-logistic	
	Median	HPD 95%	Median	HPD 95%	Median	HPD95%	Median	HPD95%
β_0	2.09	[0.76, 3.38]	2.08	[0.82, 3.36]	2.00	[0.71, 3.27]	2.06	[0.73, 3.35]
β_1	-0.08	[-0.46, 0.27]	-0.06	[-0.42, 0.28]	-0.09	[-0.47, 0.26]	-0.09	[-0.48, 0.26]
β_2	0.02	[-0.52, 0.57]	-0.03	[-0.56, 0.52]	-0.04	[-0.61, 0.52]	-0.01	[-0.56, 0.55]
β_3	-0.73	[-1.25,-0.24]	-0.73	[-1.22,-0.24]	-0.73	[-1.22,-0.20]	-0.73	[-1.22,-0.20]
β_4	-0.75	[-1.27,-0.21]	-0.67	[-1.17,-0.17]	-0.71	[-1.22,-0.16]	-0.77	[-1.27,-0.21]
β_5	0.04	[0.03, 0.05]	0.04	[0.03, 0.04]	0.04	[0.03, 0.05]	0.04	[0.03, 0.05]
β_6	0.00	[-0.02, 0.02]	0.01	[-0.02, 0.02]	0.00	[-0.02, 0.02]	0.00	[-0.02, 0.02]
β_7	0.01	[-0.01, 0.03]	0.01	[-0.01, 0.02]	0.01	[-0.01, 0.03]	0.01	[-0.01, 0.03]
β_8	-0.09	[-0.53, 0.32]	-0.11	[-0.50, 0.33]	-0.10	[-0.52, 0.33]	-0.10	[-0.53, 0.35]
σ^2	0.80	[0.47, 1.19]	0.69	[0.47, 0.95]	1.26	[0.57, 2.19]	0.36	[0.26, 0.48]
θ	5.22	[1.68,17.60]	-	-	1.30	[1.00, 1.74]	-	-

the Bayes factor in favour of being an outlier for each of the 137 observations. There is clear evidence for the existence of outlying observations under the suggested priors for all models. Although all priors present similar results, this evidence is slightly stronger for the Jeffreys prior. The choice of the mixture model does not greatly affect the conclusions. The analysis

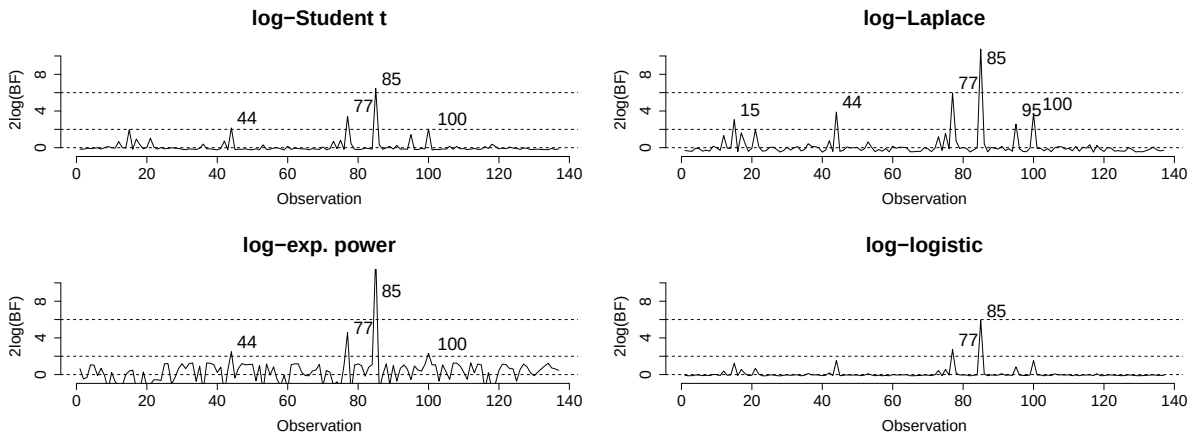

Figure 5: $2 \times \log(\text{BF})$ in favour of $H_1 : \lambda_i \neq \lambda_{ref}(u_i \neq u_{ref})$ versus $H_0 : \lambda_i = \lambda_{ref}(u_i = u_{ref})$, using independence Jeffreys prior. Horizontal lines reflect the interpretation rule of Kass and Raftery (1995).

suggests that, regardless of the prior, observations 77 and 85 are very clear outliers. Patients 77 and 85 had an uncensored survival time of 1 day (the lowest value observed in the dataset), were under the standard treatment and had a squamous type of tumor. Different models might detect different outliers. In fact, observations 15, 44, 95 and 100 are also detected as outlying observations only for some of these models under both types of independence Jeffreys prior. Observations 17, 21 and 75 are added to this list under the Jeffreys prior. With the exception of patient 21, they all correspond to uncensored observations. While observations 15, 95 and 100 have a small survival time (8, 2 and 11 days respectively), the survival times associated to patients 17, 21, 44 and 75 are larger (384, 123, 392 and 991 days respectively). In particular, the survival time of patient 75 is the second largest of the survival times of patients with squamous type of tumor (observation 70 has the largest survival time, but it is explained by a very good patient's status at treatment time) Similarly, the survival times of patients 17 and 44 are the largest survival times for patients that had the same type of tumor (small cell). None of the patients detected as possible outliers had tumors type adeno or large cell. Additionally, patient 95 has a considerably larger number of months from diagnosis than other patients with the same type of tumor (small cell).

6. CONCLUDING REMARKS

We recommend the use of mixtures of life distributions as a convenient framework for survival analysis, particularly when standard models such as the Weibull or log-normal are not able to capture some features of the data. This approach intuitively leads to flexible distributions on the basis of a known distribution by mixing over a parameter. This can also be interpreted through random effects or frailty terms. These mixture families can accommodate unobserved heterogeneity or outlying observations. In particular, the SMLN family is proposed. This family of mixtures of life distributions is based on the log-normal model and allows us to fit data with a variety of tail behavior. The mixing is applied to the shape parameter of the log-normal distribution and the resulting distribution is quite flexible and can be adjusted by choosing the mixing distribution. This makes this family applicable in a wide range of situations. Under mixture modelling, we recommend AFT regressions instead of the well-known proportional hazards models. Mixtures of AFT models provide a clearer interpretation of the regression parameters, which does not depend on the mixing distribution. In addition, the estimation of the regression coefficients is not much affected by the choice of mixing distribution. The latter is illustrated with both simulated and real data.

We consider objective Bayesian inference under the AFT-SMLN model. We propose three different Jeffreys-type priors. These priors are improper and therefore the propriety of the posterior distribution needs to be verified. Subsection 3.2 provides conditions for the existence of the posterior distribution based on an arbitrary mixing distribution. In particular, Theorem 3 provide some extra guidance and results for specific mixing distributions. In addition, the problem associated with the use of point observations is explored. The use of set observations is considered as a solution, which can easily be implemented in an MCMC sampling scheme. We recommend the use of set observations throughout. Set observations might also be helpful in other contexts. For example, the issues of the Cox proportional hazard model with ties in the data are well known. Heritier et al. (2009) ignored ties when analyzing the real dataset used

here, but that strategy might lead to serious loss of information if applied routinely. Different methods have been proposed for dealing with ties in the Cox regression model (see Kalbfleisch and Prentice, 2002, p. 104), but they might lead to biased estimations (Scheike and Sun, 2007). Set observations are a natural solution that takes into account the imprecision with which the data was recorded.

We also propose a methodology for outlier detection that is based on the mixing structure. Outliers are associated with extreme values of their corresponding mixing variable and the evidence for outlying observations is formalized by means of Bayes factors. We provide recommendations for the (critical) choice of a reference value.

A simulation study shown that standard Bayesian model comparison criteria can fairly easily identify the need of incorporating unobserved heterogeneity to the model, even with rather small sample sizes and a considerable amount of censoring. Ignoring unobserved heterogeneity can lead to biased or less precise inference for the regression parameters, whereas inference with SMLN models works well even in the absence of unobserved heterogeneity. The best results in terms of identifying the correct model are obtained for the independence Jeffreys prior and the model selection criteria DIC and PsBF. Our methodology was also applied to the VA lung cancer data, for which previous studies have found evidence of influential observations. For these data, we uncover strong evidence of unobserved heterogeneity which is mostly driven by outlying observations.

The mixing framework proposed here can be used with any proper mixing distribution, whether the induced survival time distribution has a closed-form density function or not. The proposed MCMC inference scheme does not rely on a closed form expression for the survival density with the mixing variables integrated out, so our Bayesian inference can be used much more widely than in the examples illustrated here. The challenge then may be that the Jeffreys-type prior for the parameter(s) of the mixing distribution needs to be derived from the expression in Theorem 1 (rather than from the integrated survival density) and this may not be trivial in general.

APPENDIX: PROOFS

Theorem 1. Taking the negative expectation of the second derivatives of the log likelihood, the expressions $k_1(\theta)$, $k_2(\theta)$, $k_3(\theta)$ and $k_4(\theta)$ are given by

$$k_1(\theta) = nE_{T_i}\left(\left[\frac{\log(t_i)-x_i'\beta}{\sigma}\right]^2\left[\frac{E_{\Lambda_i}\left(\Lambda_i f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\Lambda_i}\right)\right)}{f(t_i)}\right]^2\right), \tag{17}$$

$$k_2(\theta) = \frac{n}{4}\left[E_{T_i}\left(\left[\frac{\log(t_i)-x_i'\beta}{\sigma}\right]^4\left[\frac{E_{\Lambda_i}\left(\Lambda_i f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\Lambda_i}\right)\right)}{f(t_i)}\right]^2\right)-1\right], \tag{18}$$

$$k_3(\theta) = \frac{n}{2}E_{T_i}\left(\frac{\left[\frac{\log(t_i)-x_i'\beta}{\sigma}\right]^2 E_{\Lambda_i}\left(\Lambda_i f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\Lambda_i}\right)\right)}{f^2(t_i)}\int_0^\infty f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\lambda_i}\right)\frac{d}{d\theta}dP_{\Lambda_i}(\lambda_i|\theta)\right)$$
$$-\frac{1}{2}\sum_{i=1}^{n}\int_0^\infty \frac{d}{d\theta}dP_{\Lambda_i}(\lambda_i|\theta), \tag{19}$$

$$k_4(\theta) = nE_{T_i}\left(\left[\frac{\int_0^\infty f_{LN}\left(t_i|x_i'\beta,\frac{\sigma^2}{\lambda_i}\right)\frac{d}{d\theta}dP_{\Lambda_i}(\lambda_i|\theta)}{f(t_i)}\right]^2\right)-\sum_{i=1}^{n}\int_0^\infty \frac{d^2}{d\theta^2}dP_{\Lambda_i}(\lambda_i|\theta). \tag{20}$$

□

Corollary 1. The proof follows directly from Theorem 1 using the structure of the determinant of the Fisher information matrix and its sub-matrices. □

Theorem 2. Define $t_o = (t_1,\ldots,t_{n_o})'$ and $D_o = \text{diag}(\lambda_1,\ldots,\lambda_{n_o})$. The contribution of the censored observations to the likelihood function is a factor in $[0,1]$. Hence, the marginal likelihood of the complete sample ($f_T(t)$) is bounded above by the marginal likelihood of the non-censored observations ($f_{T_o}(t_o)$). Therefore, a sufficient condition for existence of the posterior distribution of (β,σ^2,θ) is $f_{T_o}(t_o)<\infty$. After some algebraic manipulation, $f_{T_o}(t_o)$ is equal to

$$\int_{\mathbb{R}^k}\int_{\mathbb{R}^+}\int_{\Theta}\int_{\mathbf{R}^{+n_o}}\frac{\prod_{i=1}^{n_o}\lambda_i^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{n_o}{2}}\prod_{i=1}^{n_o}t_i}e^{-\frac{1}{2\sigma^2}\left[(\beta-a)'A(\beta-a)+S^2(D_o,y_o)\right]}\frac{\pi(\theta)}{(\sigma^2)^p}\prod_{i=1}^{n_o}dP(\lambda_i|\theta)\,d\beta\,d\sigma^2\,d\theta, \tag{21}$$

where $A = X_o' D_o X_o$, $a = A^{-1} X_o' D_o y_o$, $S^2(D_o, y_o) = y_o' D_o y_o - y_o' D_o X_o (X_o' D_o X_o)^{-1} X_o' D_o y_o$ and $y_o = (\log(t_1), \ldots, \log(t_{n_o}))'$. Provided that $t_i \neq 0$ for all $i \in \{1, \ldots, n_o\}$, using Fubini's theorem for the integral (21) and integrating first with respect to β, we have

$$f_{T_o}(t_o) \propto \int_{\mathbb{R}^+} \int_{\Theta} \int_{\mathbb{R}^{+n_o}} (\sigma^2)^{-\frac{n_o+2p-k}{2}} \frac{\prod_{i=1}^{n_o} \lambda_i^{\frac{1}{2}}}{\sqrt{\det(X_o' D_o X_o)}} e^{-\frac{S^2(D_o, y_o)}{2\sigma^2}} \pi(\theta) \prod_{i=1}^{n_o} dP(\lambda_i|\theta)\, d\theta\, d\sigma^2. \tag{22}$$

After integrating with respect to σ^2, it follows that

$$f_{T_o}(t_o) \propto \int_{\Theta} \int_{\mathbb{R}^{+n_o}} \prod_{i=1}^{n_o} \lambda_i^{\frac{1}{2}} (\det(X_o' D_o X_o))^{-\frac{1}{2}} [S^2(D_o, y_o)]^{-\frac{n_o+2p-k-2}{2}} \pi(\theta) \prod_{i=1}^{n_o} dP(\lambda_i|\theta)\, d\theta, \tag{23}$$

as long as $n_o + 2p - k - 2 > 0$ and $S^2(D_o, y_o) > 0$. If $n_o > k$ we know that $S^2(D_o, y_o) > 0$ a.s. and the first condition is certainly satisfied when $p \geq 1$. Analogously to Lemma 1 in Fernández and Steel (1999), $f_{T_o}(t_o)$ has upper and lower bounds proportional to

$$\int_{\Theta} \int_{0<\lambda_1<\cdots<\lambda_{n_o}<\infty} \prod_{i \notin \{m_1,\ldots,m_k\}} \lambda_i^{\frac{1}{2}} \lambda_{m_{k+1}}^{-\frac{n_o+2p-k-2}{2}} \pi(\theta) \prod_{i=1}^{n_o} dP(\lambda_i|\theta)\, d\theta, \tag{24}$$

where

$$\prod_{i=1}^{k} \lambda_{m_i} \equiv \max \left\{ \prod_{i=1}^{k} \lambda_{l_i} : \det(x_{l_1} \cdots x_{l_k}) \neq 0, l_1, \ldots l_k \in \{1, \ldots, n_o\} \right\}, \tag{25}$$

$$\prod_{i=1}^{k+1} \lambda_{m_i} \equiv \max \left\{ \prod_{i=1}^{k+1} \lambda_{l_i} : \det \begin{pmatrix} x_{l_1} & \cdots & x_{l_{k+1}} \\ \log(t_{l_1}) & \cdots & \log(t_{l_{k+1}}) \end{pmatrix} \neq 0, l_1, \ldots l_k \in \{1, \ldots, n_o\} \right\} \tag{26}$$

(i) For $p = 1$. Barring a set of zero Lebesgue measure, $\lambda_{m_{k+1}} = \max\{\lambda_i : i \notin \{m_1, \ldots, m_k\}\}$. Hence, (24) is bounded above by $\int_{\Theta} \pi(\theta)\, d\theta = 1$. If $n_o > k$, the posterior exists.

(ii) For $p = 1 + k/2$. By the same argument, (24) is bounded above by $\int_{\Theta} E(\Lambda_{m_{k+1}}^{-\frac{k}{2}}|\theta)\pi(\theta)\, d\theta$. However, $E(\Lambda_{m_{k+1}}^{-\frac{k}{2}}|\theta) \leq E(\Lambda_{(1)}^{-\frac{k}{2}}|\theta)$ where $\Lambda_{(1)} = \min\{\Lambda_1, \ldots, \Lambda_{n_o}\}$. Using the density of the first order statistic it follows that $E(\Lambda_{(1)}^{-\frac{k}{2}}|\theta) \leq n_o E(\Lambda_i^{-\frac{k}{2}}|\theta)\ \forall i = 1, \ldots, n_o$ and hence, as the Λ_i's are iid, the results holds.

□

Theorem 3. (i) It can be shown that the Fisher information matrix corresponds to

$$\begin{pmatrix} \frac{1}{\sigma^2}\frac{\nu+1}{\nu+3}\sum_{i=1}^{n} x_i x_i' & 0 & 0 \\ 0 & \frac{n}{2\sigma^4}\frac{\nu}{\nu+3} & -\frac{n}{\sigma^2}\frac{1}{(\nu+1)(\nu+3)} \\ 0 & -\frac{n}{\sigma^2}\frac{1}{(\nu+1)(\nu+3)} & \frac{n}{4}\left[\Psi'(\frac{\nu}{2}) - \Psi'(\frac{\nu+1}{2}) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}\right] \end{pmatrix}. \quad (27)$$

Therefore, the components depending on ν of the Jeffreys, independence Jeffreys and independence I Jeffreys prior are, respectively

$$\pi^J(\nu) \propto \left(\frac{\nu+1}{\nu+3}\right)^{k/2} \sqrt{\frac{\nu}{\nu+3}} \sqrt{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}}, \quad (28)$$

$$\pi^I(\nu) \propto \sqrt{\frac{\nu}{\nu+3}} \sqrt{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}}, \quad (29)$$

$$\pi^{II}(\nu) \propto \sqrt{\Psi'\left(\frac{\nu}{2}\right) - \Psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+5)}{\nu(\nu+1)(\nu+3)}}. \quad (30)$$

It can be shown that $\pi^J(\nu)$ and $\pi^I(\nu)$ are proper priors for ν (Corollary 1 in Fonseca et al., 2008). However, $\pi^{II}(\nu)$ is not (it behaves as ν^{-1} when $\nu \to 0$). Hence, as mentioned in Subsection 3.1, the independence I prior is discarded for the log-Student t model.

Theorem 2 part (i) implies the propriety of the posterior distribution for the independence Jeffreys prior. Theorem 2 cannot be used in order to conclude about the posterior existence under the Jeffreys prior (the condition in part (ii) is not satisfied because $\mathrm{E}(\Lambda_1^{-k/2}|\nu)$ does not exist for $\nu < k$). However, upper and lower bounds for the integral in (24) can be found using the inequality (Fernández and Steel, 1999, 2000)

$$\frac{\lambda_{i+1}^v}{v} e^{-r\lambda_{i+1}} \leq \int_0^{\lambda_{i+1}} \lambda_i^{v-1} e^{-r\lambda_i} d\lambda_i \leq \frac{\lambda_{i+1}^v}{v}, \qquad r, v > 0. \quad (31)$$

The integral in (31) is not finite for $v \leq 0$. Barring a set of zero Lebesgue measure, $\lambda_{m_{k+1}} = \lambda_{(n-k)}$, where $\lambda_{(n-k)}$ is the $(n-k)$-th order statistic of $\lambda_1, \ldots, \lambda_n$. After integrating with respect to the $n-k-1$ smallest λ_i's, (24) has the lower bound

$$\int_0^\infty \int_{\Lambda^*} \left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{n-k} \frac{\left[\frac{\nu+1}{2}\right]^{-(n-k-1)}}{(n-k-1)!} \lambda_{(n-k)}^{c-1} e^{-\frac{(n-k)\nu}{2}\lambda_{(n-k)}} d\lambda_{(n-k)} \left[\prod_{i=n-k+1}^{n} dP(\lambda_{(i)}|\nu)\right] \pi(\nu)\, d\nu, \quad (32)$$

where $\Lambda^* = \{(\lambda_{(n-k)}, \ldots, \lambda_{(n)}) : 0 < \lambda_{(n-k)} < \cdots < \lambda_{(n)} < \infty\}$ and $c = -\frac{n+2p-k-3}{2} + \frac{\nu}{2} + \frac{(n-k-1)(\nu+1)}{2} = \frac{\nu(n-k)+2-2p}{2}$. When integrating with respect to $\lambda_{(n-k)}$ we need $c > 0$ in order to have a finite integral in (41). Hence, the propriety of the posterior distribution requires $\nu > \frac{2p-2}{n-k}$.

As a consequence, the posterior distribution of (β, σ^2, ν) is not proper if $p > 1$ and the range of ν is $(0, \infty)$. In particular, the Jeffreys-rule prior (for which $p = 1 + k/2$) does not lead to a proper posterior distribution and Bayesian inference is thus precluded with this prior. Incorporating censored observations does not help, as the posterior distribution is still not well defined. For example, under right censoring, the marginal likelihood can be expressed as

$$f_T(t) = \int_{\mathcal{T}^*} \int_{\mathbb{R}^k} \int_{\mathbb{R}^+} \int_{\Theta} \left[\prod_{i=1}^{n} f_{T_i}(t_i^* | \beta, \sigma^2, \theta) \right] \pi(\beta, \sigma^2, \theta)\, d\beta\, d\sigma^2\, d\theta\, dt^* \equiv \int_{\mathcal{T}^*} f_T^*(t^*)\, dt^*, \quad (33)$$

where $\mathcal{T}^* = t_1 \times \cdots \times t_{n_o} \times (t_{n_o+1}, \infty) \times \cdots (t_n, \infty)$ and $f_T^*(t^*)$ is an auxiliary marginal likelihood that treats censored observations as if they were non-censored. For any $t^* \in \mathcal{T}^*$, $f_T^*(t^*)$ is not finite. Therefore, we conclude that $f_T(t)$ is not finite and the posterior based on the complete sample is not well-defined under the Jeffreys prior.

(ii) As the parameter θ is not required for the log-Laplace model, the independence Jeffreys and independence I Jeffreys coincide. Theorem 2 part (i) indicates that the posterior is proper under these priors. In both cases, $E(\Lambda_1^{-\frac{k}{2}})$ is finite and therefore the posterior under the Jeffreys prior is also proper. In fact, for the log-Laplace model, Λ_1^{-1} is Gamma distributed and all its positive moments are finite. For the log-logistic model, it can be shown that $\Omega_i = \sqrt{1/(4\Lambda_i)}$ has an Asymptotic Kolmogorov distribution with density function $g(\omega_i) = 8\omega_i \sum_{s=1}^{\infty} (-1)^{s+1} s^2\, e^{-2s^2\omega_i^2}$, for $\omega_i > 0$. Therefore, for $k > -2$, it follows that

$$\begin{aligned} E(\Lambda_1^{-k/2}) &= 2^{k+3} \sum_{s=1}^{\infty} (-1)^{s+1} s^2 \int_0^{\infty} \omega_1^{k+1}\, e^{-2s^2\omega_1^2}\, d\omega_1 && (34) \\ &= 2^{k+2} \sum_{s=1}^{\infty} (-1)^{s+1} s^2 \int_0^{\infty} \eta^{k/2}\, e^{-2s^2\eta}\, d\eta && (35) \\ &= 2^{k/2+1} \Gamma(1 + k/2) \sum_{s=1}^{\infty} (-1)^{s+1} \frac{1}{s^k} < \infty. && (36) \end{aligned}$$

For the log-exponential power model, the Fisher Information matrix was derived by Martín and Pérez (2009) and is given by

$$\begin{pmatrix} \frac{\alpha(\alpha-1)\Gamma(1-\frac{1}{\alpha})}{\sigma^2\Gamma(\frac{1}{\alpha})}\sum_{i=1}^{n} x_i x_i' & 0 & 0 \\ 0 & \frac{n\alpha}{\sigma^2} & -\frac{n(1+\Psi(1+\frac{1}{\alpha}))}{\sigma\alpha} \\ 0 & -\frac{n(1+\Psi(1+\frac{1}{\alpha}))}{\sigma\alpha} & \frac{n}{\alpha^3}\left[(1+\frac{1}{\alpha})\Psi'(1+\frac{1}{\alpha})+(1+\Psi(1+\frac{1}{\alpha}))^2-1\right] \end{pmatrix}.$$

Therefore, the components depending on α of the Jeffreys, independence Jeffreys and independence I Jeffreys prior are, respectively

$$\pi^J(\alpha) = \left[\frac{\alpha(\alpha-1)\Gamma(1-1/\alpha)}{\Gamma(1/\alpha)}\right]^{k/2} \frac{1}{\alpha}\sqrt{\left(1+\frac{1}{\alpha}\right)\Psi'\left(1+\frac{1}{\alpha}\right)-1}, \quad (37)$$

$$\pi^I(\alpha) = \frac{1}{\alpha}\sqrt{\left(1+\frac{1}{\alpha}\right)\Psi'\left(1+\frac{1}{\alpha}\right)-1}, \quad (38)$$

$$\pi^{II}(\alpha) = \frac{1}{\alpha^{\frac{3}{2}}}\sqrt{\left(1+\frac{1}{\alpha}\right)\Psi'\left(1+\frac{1}{\alpha}\right)+\left[1+\Psi\left(1+\frac{1}{\alpha}\right)\right]^2-1}. \quad (39)$$

As the previous components are bounded continuous functions of α in $(1,2)$, they are proper priors for α. Theorem 2 part (i) implies the propriety of the posterior distribution under the independence Jeffreys and independence I Jeffreys prior. The propriety of the posterior under the Jeffreys prior can be verified using Theorem 2 part (ii) because $E(\Lambda_1^{-\frac{k}{2}}|\alpha)$ is a continuous bounded function for $\alpha \in (1,2)$. In fact,

$$E(\Lambda_1^{-\frac{k}{2}}|\alpha) = \frac{\Gamma(3/2)}{\Gamma(1+1/\alpha)}\frac{E(W^{\frac{k+1}{2}}|\alpha)}{E(Z^{\frac{k+1}{2}}|\alpha)} = \frac{\Gamma(3/2)}{\Gamma(1+1/\alpha)}\frac{\Gamma((k+1)/\alpha+1)}{\Gamma((k+3)/2)}, \quad (40)$$

where $W \sim$ Weibull$(\alpha/2,1)$ and $Z \sim$ Exponential(1). The latter uses the lemma in Meintanis (1998) which states that a Weibull$(a,1)$ random variable can be represented as the ratio of an Exponential(1) and an independent positive stable(a) random variable.

□

Theorem 4 (Based on Fernández and Steel, 1999). If s is the largest number of observations that can be written as an exact linear combination of their covariates, $\lambda_{m_{k+1}}$ (defined in (26)) corresponds to $\lambda_{(n_o-s)}$, which represent the (n_o-s)-th order statistic of $\lambda_1,\ldots,\lambda_{n_o}$. The rest

of the proof is obtained by iteratively integrating (24), using the inequality in (31). After integrating with respect to the $n_o - s - 1$ smallest λ's, (24) has a lower bound given by

$$\int_0^\infty \int_{0<\lambda_{(n_o-s)}<\cdots<\lambda_{(n_o)}<\infty} \left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{n_o-s} \frac{\left[\frac{\nu+1}{2}\right]^{-(n_o-s-1)}}{(n_o-s-1)!} \lambda_{(n_o-s)}^{a-1} e^{-\frac{(n_o-s)\nu}{2}\lambda_{(n_o-s)}} \prod_{i=n_o-s+1}^{n_o} dP(\lambda_{(i)}|\nu)\pi(\nu)\, d\nu. \tag{41}$$

Where $a = -\frac{n_o+2p-k-3}{2} + \frac{\nu}{2} + \frac{(n_o-s-1)(\nu+1)}{2}$. Note that when integrating with respect to $\lambda_{(n_o-s)}$, we need $a > 0$ in order to have a finite integral in (41). Hence, the propriety of the posterior distribution requires $\nu > \frac{n_o-k+(2p-2)}{n_o-s} - 1$. □

Theorem 5. Define $I(s) = \int_{\mathbb{R}^k} \int_0^\infty \int_\Theta f_{T_o}(s|\beta, \sigma^2, \theta)\pi(\beta, \sigma^2, \theta)\, d\beta\, d\sigma^2\, d\theta$. Based on the sample t_ϵ, the posterior distribution exists if and only if $\int_E I(s)\, ds$ is finite. As E is bounded $\int_E I(s)\, ds$ is bounded as long as $I(\cdot)$ is finite except on a set of zero Lebesgue measure. □

SUPPLEMENTARY MATERIALS

Supplementary material A: Implementation details and description of R-code. (pdf file)

Supplementary material B: Detailed results of the simulation study. (pdf file)

Supplementary material C: Convergence and mixing diagnostics for MCMC chains in the VA lung cancer data application. (pdf file)

R-code: Code for the MCMC algorithm and model comparison methods described in the article. It also includes the VA lung cancer dataset. (zip file)

References

Banerjee, T., Chen, M., Dey, D., and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime data analysis*, 13:241–260.

Barros, M., Paula, G., and Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime Data Analysis*, 14:316–332.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321.

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

Cho, H., Ibrahim, J. G., Sinha, D., and Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, 65:116–124.

Fernández, C. and Steel, M. (1998). 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, pages 213–238.

Fernández, C. and Steel, M. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, 86:153–167.

Fernández, C. and Steel, M. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16:80–101.

Fonseca, T., Ferreira, M., and Migon, H. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika*, 95:325–333.

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

Hanson, T. (2006). Inference for mixtures of finite polya tree models. *Journal of the American Statistical Association*, 101:1548–1565.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M. (2009). *Robust Methods in Biostatistics*. Wiley Series in Probability and Statistics. Wiley.

Hogg, R. V. and Klugman, S. A. (1983). On the estimation of long tailed skewed distributios with actuarial applications. *Journal of Econometrics*, 23:91–102.

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168.

Jewell, N. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, 10:479–484.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, 2nd edition.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Lange, K., Little, R., and Taylor, J. (1989). Robust statistical modelling using the t distribution. *Journal of the American Statistical Association*, 84:881–896.

Lee, E. and Wang, J. (2003). *Statistical methods for survival data analysis*. Wiley, 3rd edition.

Marshall, A. W. and Olkin, I. (2007). *Life Distributions*. Springer.

Martín, J. and Pérez, C. (2009). Bayesian analysis of a generalized lognormal distribution. *Computational Statistics and Data Analysis*, 53:1377–1387.

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

Meintanis, S. (1998). Moment-type estimation for positive stable laws with applications. *IAENG International Journal of Applied Mathematics*, 38:26–29.

Patriota, A. (2012). On scale-mixture Birnbaum-Saunders distributions. *Journal of Statistical Planning and Inference*, 142:2221–2226.

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

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

Scheike, T. H. and Sun, Y. (2007). Maximum likelihood estimation for tied survival data under Cox regression model via EM-algorithm. *Lifetime data analysis*, 13(3):399–420.

Shah, B. and Dave, P. (1963). A note on log-logistic distribution. *Journal of the M.S. University of Baroda (Science Number)*, 12:15–20.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B*, 64:583–640.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540.

Uppuluri, V. (1981). Some properties of log-laplace distribution. *Statistical Distributions in Scientific Work 4,* Taillie, C., Patil, G. P., Baldessari, B. A. (eds.), Reidel, pages 105–110.

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

Vianelli, S. (1983). The family of normal and lognormal distributions of order r. *Metron*, 41:3–10.

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

Wienke, A. (2010). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC.