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

Catalina A. Vallejos and Mark F. J. Steel

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 modeled through covariates. This article considers the use of infinite mixtures of lifetime distributions as a solution for these 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 nonsubjective 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, we propose a method for outlier detection based on the mixture structure. A simulation study illustrates the performance of our methods under different scenarios and an application to a real dataset is provided. Supplementary materials for the article, which include R code, are available online.

KEY WORDS: 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, Paula, and Leiva 2008). A second, related issue is the existence of specific individual factors that result in heterogeneity of the survival times, which cannot be captured by covariates (Marshall and Olkin 2007). Therefore, the typical assumption that the survival times correspond to realizations of random variables $T_1, \ldots, T_n$ that 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, Paula, and Leiva 2008; Heritier et al. 2009).

We consider the use of infinite mixtures 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 behavior than the log-normal. It includes the already studied log-Student $t$, log-Laplace, log-exponential power, and log-logistic distributions among others. This article 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 hyperparameters 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 (AFT) 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 on 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. Nonsubjective 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 unobserved individual effects, and can also be related to the presence of outlying observations. This article considers the use of mixtures of life distributions 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

$$f(t_i | \psi, \theta) = \int_{\mathcal{L}} f(t_i | \psi, \Lambda_i = \lambda_i) d P_{\Lambda_i}(\lambda_i | \theta),$$

where $f(\cdot | \psi, \Lambda_i = \lambda_i)$ represents the density function associated with a lifetime distribution, which depends on the values of $\psi$ and $\lambda_i$ (underlying distribution), and $\lambda_i$ 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 $\theta$. 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 where $\Lambda_i$ 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) explored mixtures of exponential and Weibull distributions. Barros, Paula, and Leiva (2008) and Patriota (2012) considered 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

$$f(t_i | \mu, \sigma^2, \theta) = \int_{\mathcal{L}} f_{LN}(t_i | \mu, \sigma^2) d P_{\Lambda_i}(\lambda_i | \theta),$$

$$t_i > 0, \mu \in \mathbb{R}, \sigma^2 > 0, \theta \in \Theta,$$

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

$$T_i | \mu, \sigma^2, \lambda_i \sim LN \left( \mu, \frac{\sigma^2}{\lambda_i} \right),$$

$$\Lambda_i | \theta \sim P_{\Lambda_i}(\cdot | \theta).$$

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 the log-Laplace appeared in Uppuluri (1981). The log-exponential power was proposed by Vianelli (1983) and 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 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, for example, $\Lambda_i \sim Gamma(\theta, \theta)$). 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 $\sigma^2$.

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 $\nu$, and the log-Laplace and log-logistic models only allow for moments up to $1/\sigma^2$. The log-exponential power distribution with $\alpha > 1$ does possess all moments.

### 2.2 The AFT-SMLN Model

An important aspect of survival modeling is the inclusion of covariates. Throughout, we will condition on the covariates that are assumed to be constant in time. An AFT model is introduced. The AFT-SMLN model expresses the dependence between the covariates and the survival time by replacing the parameter $\mu$ with $x' \beta$, so that

$$ T_i | x_i, \beta, \sigma^2, \theta \overset{i.i.d.}{\sim} \text{SMLN}_P(x_i' \beta, \sigma^2, \theta), \quad i = 1, 2, \ldots, n, \tag{4} $$

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 $e^{x_i' \beta}$, $e^{\beta j}$ 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-theoretical 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

$$ I(\beta, \sigma^2, \theta) = \begin{pmatrix} \frac{1}{\sigma^2} k_1(\theta) \sum_{i=1}^n x_i x_i' & 0 & 0 \\ 0 & \frac{1}{\sigma^2} k_2(\theta) & \frac{1}{\sigma^2} k_3(\theta) \\ 0 & \frac{1}{\sigma^2} k_3(\theta) & k_4(\theta) \end{pmatrix}, \tag{5} $$
where \( k_1(\theta), k_2(\theta), k_3(\theta), \) and \( k_4(\theta) \) are functions depending only on \( \theta \).

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 \( \beta \) and \( (\sigma^2, \theta) \)), and independence I Jeffreys’ (which deals separately with \( \beta, \sigma^2, \) and \( \theta \)) priors are, respectively, given by

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

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

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

The three nonsubjective 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 \( \theta \). For the Jeffreys’ prior, \( p = 1 + (k/2) \) and \( p = 1 \) for the other two priors. If \( \theta \) 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 \( \theta \). 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. To obtain meaningful Bayes factors (BFs) between models, priors with an improper component \( \pi(\theta) \) for \( \theta \) 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 noninformative). 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 \( n_0 \) observations are uncensored. Define \( X_0 = (x_1, \ldots, x_{n_0})' \) and suppose that the rank of \( X_0 \) is \( k \).

(i) For \( p = 1 \), a sufficient condition for posterior existence is \( n_0 > k \).

(ii) For \( p = 1 + k/2 \), a sufficient condition for the posterior propriety is \( n_0 > k \) and

\[
\int_\Theta E(A_1^{-1/2} | \theta) \pi(\theta) d\theta < \infty. \tag{10}
\]

**Theorem 3.** Under the assumptions in Theorem 2 and provided that \( n_0 > k \), 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 Section 3.1). The other models can be combined with all priors considered here; of course, the absence of \( \theta \) 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 \( n_0 > k \). If the mixing distribution is Gamma\((v/2, v/2)\) and \( s (k \leq s < n_0) \) 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 \( \beta \)), a necessary condition for the propriety of the posterior distribution of \((\beta, \sigma^2, v)\) is

\[
\int_0^m \pi(v) dv = 0, \quad \text{where} \quad m = \frac{n_0 - k + (2p - 2)}{n_0 - s} - 1. \tag{11}
\]

This result indicates that it is possible to have samples of point observations for which no Bayesian inference can be conducted, unless \( \pi(v) \) induces a positive lower bound for \( v \). 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 \( t_i - \epsilon_i \) and \( t_i + \epsilon_i \), where \( \epsilon_i \) and \( \epsilon_i \) are determined by the accuracy with which the data were recorded (e.g., if the data are recorded in integers, \( \epsilon_i = 0.5 \)). This is equivalent to considering the observation as interval censored. As explained in Section 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, \( (\epsilon_i, \epsilon_i) = (t_i, 0) \) or \( (\epsilon_i, \epsilon_i) = (0, \infty) \). 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_k = ((t_i - \epsilon_i, (t_i + \epsilon_i), \ldots, (t_n - \epsilon_i, t_n + \epsilon_i)) (0 < \epsilon_i, \epsilon_i < 0). Define \( E = (t_i - \epsilon_i, t_i + \epsilon_i) \times (t_2 - \epsilon_i, t_2 + \epsilon_i) \times \cdots \times (t_n - \epsilon_i, t_n + \epsilon_i). The posterior distribution of \((\beta, \sigma^2, \theta) given t_k \) is proper if and only if the marginal likelihood under point observations is bounded from above for any \( t_k \in E \), 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 parameter 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 \( \beta \) and \( \sigma^2 \) are normal and inverted gamma distributions. However, the full conditionals for \( \Lambda_1, \ldots, \Lambda_n \) and \( \theta \) are generally not of a known form for arbitrary mixing distributions. For the \( \Lambda_i 's \), we have \( \pi(\lambda_1, \ldots, \lambda_n | \beta, \sigma^2, \theta, t) = \prod_{i=1}^n \pi(\lambda_i | \beta, \sigma^2, \theta, t_i) \), where \( t = (t_1, \ldots, t_n) \) are the simulated survival times and

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

For the log-Student \( t \) and log-Laplace models, Equation (12) has a known form. In the first case, it is a Gamma\((v + 1)/2, v/2\) \[
\left\{ \frac{(\log(t_i) - x' \beta)^2}{\sigma^2} + v \right\} \]. In the second, it corresponds to an Inverse Gaussian\((\sigma^2/\log(t_i) - x' \beta), 1\). 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 \( (2\sqrt{\alpha})^{-1} \) has the asymptotic Kolmogorov–Smirnov distribution. In the case of the log-exponential power model, we adopt the mixture of uniform representations used in Martín and Pérez (2009). This replaces the use of \( \Lambda_i \) by \( U_i (i = 1, \ldots, n) \), with \( U_i \sim \text{Gamma}(1 + 1/\alpha, 1) \) and \( \log(T_i) U_i = u_i, \beta, \sigma^2, \alpha \sim U(x' \beta - \sigma u_i/\alpha, x' \beta + \sigma u_i/\alpha) \). We restrict the range of \( \alpha \) 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 \( \prod_{i=1}^n \pi(u_i | \beta, \sigma^2, \alpha) \times \pi(\beta, \sigma^2, \alpha | t) \) and the full conditionals for the \( U_i 's \) are

\[
\pi(u_i | \beta, \sigma^2, \alpha) \propto e^{-u_i}, u_i > \left( \frac{\log(t_i) - x' \beta}{\sigma} \right)^{\alpha}, i = 1, \ldots, n. \tag{13}
\]

For \( \pi(\beta, \sigma^2, \alpha | t) \), the full conditionals of \( \beta, \sigma^2, \) and \( \alpha \) 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 \( \Lambda_i \) (or \( U_i \)), which is particularly important in identifying outlying observations (Lange et al. 1989; Fernández and Steel 1999). This is further discussed in Section 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. First, we use BF's, 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 later was developed for a nonadaptive 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. Second, the deviance information criteria (DIC) developed by Spiegelhalter et al. (2002) is also provided. It is given by \( \text{DIC} = E(D(\beta, \sigma^2, \theta, y)) + p_D \) with \( p_D = E(D(\beta, \sigma^2, \theta, y)) - D(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}, y) \) (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 \( \beta, \sigma^2, \) and \( \theta \), 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

\[
CPO_i = f(t_i | t_{-i}) = \left[ \mathbb{E}\left( f(t_i, \beta, \sigma^2, \theta) \right) \right]^{-1},
\]

\( t_{-i} = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) \),

(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 (Hansoon 2006; Banerjee et al. 2007). Larger values of CPO \( i \) indicate better predictive accuracy for the observation \( i \). Geisser and Eddy (1979) also proposed PsML = \( \prod_{i=1}^n \) 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 (PsBFs) can be easily computed as ratios of PsMLs.

3.6 Detection of Influential Observations and Outliers

A robust model will have no (or few) influential observations. Influential observations can be detected using \( K_i = KL(\pi(\beta, \sigma^2, \theta | t), \pi(\beta, \sigma^2, \theta | t_{-i})) \), where \( KL(\cdot, \cdot) \) 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, \( \lambda_{\text{ref}} \)) of the mixing variables are associated with outliers (see also West 1984). Formally, evidence of outlying observations will be assessed by contrasting the models \( M_0 : \Lambda_i = \lambda_{\text{ref}} \) versus \( M_1 : \Lambda_i \neq \lambda_{\text{ref}} \) (with all other \( \Lambda_j, j \neq i \) fixed). Evidence in favor of each of these models will be measured using BF's, which can be computed as the generalized Savage-Dickey density ratio proposed in Verdinelli and Wasserman (1995). The evidence in favor of \( M_0 \) versus \( M_1 \) (i.e., against observation \( i \) being an outlier) is

\[
BF_{01} = \int \pi(\lambda_i | t) \frac{dP(\lambda_i | \theta)}{dP(\lambda_{\text{ref}} | \theta)} \bigg|_{\lambda_i = \lambda_{\text{ref}}} ,
\]

(15)

where the expectation is with respect to \( \pi(\theta | t, \Lambda_i = \lambda_{\text{ref}}) \). When the parameter \( \theta \) does not appear in the model, this simplifies to the original Savage-Dickey density ratio

\[
BF_{01} = \int \frac{dP(\lambda_i | \theta)}{dP(\lambda_{\text{ref}} | \theta)} \bigg|_{\lambda_i = \lambda_{\text{ref}}} ,
\]

(16)

where the expectation is with respect to \( \pi(\beta, \sigma^2 | t) \). The main challenge of this approach is the choice of \( \lambda_{\text{ref}} \). 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 \( \Lambda_i \)) will then be close to the posterior distribution of \( \Lambda_i \). Following this intuition, we propose to use \( E(\Lambda_i | \theta) \) (if it exists) as \( \lambda_{\text{ref}} \). Using this rule, \( \lambda_{\text{ref}} = 1 \) for the log-Student t model. This choice was supported by our empirical examples. If \( E(\Lambda_i | \theta) \) depends on \( \theta \) (unknown), we suggest estimating it through the posterior median of \( \theta \). Examples for which \( E(\Lambda_i | \theta) \) is not finite require a more detailed analysis. For example, the expectations of the mixing distributions that generate 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 \( \Lambda_i \)'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 datasets we have tried. In this calculation, we discard the lowest 25% of \( \lambda_i \) values to remove the influence of any possible outliers. Hence, we propose \( \lambda_{\text{ref}} = 1 \) for the log-Laplace model. In the log-logistic case, the posterior distributions of the \( \Lambda_i \)'s behave as in the log-Student t case, where the reference value is clearer. We chose \( \lambda_{\text{ref}} = 0.4 \) for the log-logistic model, using the same argument as in the log-Laplace case. Figure 3 shows the

![Figure 3](image-url)

Figure 3. Bayes factor for outlier detection as a function of \(|z|\). The log Bayes factor has been rescaled by 2 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.
performance of the reference values by plotting the inverse of the BF in (15) for the log-Student and in (16) for the log-Laplace and log-logistic models against a standardized log survival time \( z \) (given \( \beta, \sigma^2 \), and \( \theta \)). This is defined as \( \log(t) \) minus its mean, divided by its standard deviation (i.e., \( \sigma \sqrt{\text{Var}(1/\Lambda|\beta)} \)). For the log-Student \( t \), log-Laplace, and log-logistic models,

\[
\begin{align*}
\sigma^{-2} & = \left( \frac{2 \alpha - 1}{\sqrt{12}} \right) \left( \frac{1}{\gamma} \right) \left( \frac{1}{\beta} \right), \quad \text{for } \nu > 2, \\
\sigma^{-2} & = \left( \frac{8}{\sqrt{3\alpha}} \right) \left( \frac{1}{\gamma} \right) \left( \frac{1}{\beta} \right), \quad \text{for } \nu = 2, \\
\sigma^{-2} & = \left( \frac{6}{\pi} \right) \left( \frac{1}{\gamma} \right) \left( \frac{1}{\beta} \right), \quad \text{for } \nu = 1.
\end{align*}
\]

\( z \) lead to evidence in favor 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.

The log-exponential power model is a special case. As explained in Section 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 \( M_0 : U_i = u_{\text{ref}} \) versus \( M_1 : U_i \neq u_{\text{ref}} \). The expectation of \( U_i \) given \( \alpha \) is \( 1 + 1/\alpha \) and, according to the intuition presented previously, we might use this value as \( u_{\text{ref}} \). With this rule, \( u_{\text{ref}} \) is a function of \( \alpha \), 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 \( \pi(u_{\text{ref}}) \) by averaging \( \pi(u_{\text{ref}}|t, \beta, \sigma^2, \alpha) \) 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 \( u_{\text{ref}} \leq \left( \frac{\log(t) - \gamma \beta}{\sigma} \right)^{\alpha} \), \( \pi(u_{\text{ref}}|t, \beta, \sigma^2, \alpha) \) is equal to zero and the BF in favor of the observation \( i \) being an outlier is computed as infinity. Simulated datasets indicated that the means and medians of \( \frac{\log(t) - \gamma \beta}{\sigma} \), \( i = 1, \ldots, n \), are around 0.6, regardless of the model from which the data were generated. We then adjust the reference value to \( u_{\text{ref}} = 1 + 1/\alpha + 0.6 \). This choice performed much better with simulated datasets (e.g., using log-normal data no outliers were detected). The resulting BFAs as a function of \( z = \frac{\log(t) - \gamma \beta}{\sigma} \sqrt{\frac{1/k}{\text{Var}(1/\Lambda|\beta)}} \) (see Figure 3) are not much affected by the value of \( \alpha \).

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 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 \( \text{PC} = 10\%, 70\% \). 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 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, \( \beta \) 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 \( \beta \). The choice between one of the three Jeffreys’-rule-based priors suggested in Section 3.1 is not very critical for the estimation of \( \beta \) as all priors produce very similar inference for the regression parameter. Of course, estimation is more accurate when the data provide more information, that is, for \( n = 500 \) and \( \text{PC} = 10\% \). There are no major differences between log-normal datasets and those generated by an 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 \( \beta \). Crucially, fitting SMLN models to log-normal datasets is harmless. The \( \beta \) 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 strong unobserved heterogeneity, mixture models tend to outperform the log-normal one. For those cases, SMLN models produce more accurate estimates of \( \beta \) 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 Section 3.5 are applied to each dataset 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 favors small values of \( \sigma^2 \), 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 \( \sigma^2 \) 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 \( \text{PC} = 70\% \)). 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 nine 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, Paula, and Leiva (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 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, σ²) 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 suggests convergence of the chains (see supplementary material C). The Jeffreys’ and the independence Jeffreys’ priors produced similar results. Bayesian inference was conducted on the basis of point and set observations, using ε₁ = ε₂ = 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 Section 3.3), so we will focus on the analysis with set observations in the rest of the article. Results suggest that the main covariate effects are due to the tumor type and patient status, and are roughly in line with the maximum likelihood results for the log-Birnbaum-Saunders AFT model in Barros, Paula, and Leiva (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 BF s, PsBF s, DIC, and the CPO predictive performance, summarized in Figure 4 and Table 3. Clearly, all these criteria provide evidence in favor of mixture models. For the log-Student t model, this evidence is also supported by the fact that inference on ν favors relative small values. Similarly, the log-exponential power model suggests values of α far from 2. From plots (unreported) of the prior 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 $\beta$ (only results under the independence Jeffreys’ prior are reported). For mixture models, the posterior distribution of $\beta$ 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 $\beta$ are relatively close to the classical ones reported in Barros, Paula, and Leiva (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 $\sigma^2$ 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 models than for the log-normal model, which is consistent with the superior ability of the SMLN models to accommodate unusual observations. For all priors considered, 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). For any of the mixtures, observation 106 is always considered as influential ($p_i \geq 0.9$). Observations 12 and 85 are flagged up as influential by some of the mixtures. These results are roughly in line with the results in Barros, Paula, and Leiva (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 the BF in favor 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 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.

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. $p_i \geq 0.9$ |
|------------------|---------------|---------|-------------------|--------------------------|--------------------------|
| Jeffreys’        | Log-normal    | 1449.01 | —                 | 6                        | 2                        |
|                  | Log-Student $t$ | —       | —                 | —                        | —                        |
|                  | Log-Laplace    | 1444.18 | 52%               | 2                        | 1                        |
|                  | Log-exp. power | 1444.00 | 54%               | 3                        | 1                        |
|                  | Log-logistic   | 1444.14 | 66%               | 3                        | 1                        |
| Ind. Jeffreys’   | Log-normal    | 1449.56 | —                 | 6                        | 2                        |
|                  | Log-Student $t$ | 1445.86 | 64%               | 3                        | 1                        |
|                  | Log-Laplace    | 1444.37 | 53%               | 1                        | 1                        |
|                  | Log-exp. power | 1444.79 | 55%               | 3                        | 1                        |
|                  | Log-logistic   | 1444.49 | 66%               | 3                        | 1                        |
| Ind. I Jeffreys’ | Log-normal    | 1449.56 | —                 | 6                        | 2                        |
|                  | Log-Student $t$ | —       | —                 | —                        | —                        |
|                  | Log-Laplace    | 1444.37 | 53%               | 1                        | 1                        |
|                  | Log-exp. power | 1444.81 | 55%               | 3                        | 1                        |
|                  | Log-logistic   | 1444.49 | 66%               | 3                        | 1                        |
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 with 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 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.

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 modeling, we recommend AFT regressions instead of the well-known proportional hazard models. Mixtures of AFT models provide a clearer interpretation of the regression parameters.

Table 4. Summary of the posterior distribution under the independence Jeffreys’ prior for various SMLN models on the basis of set observations. $\beta_0$: Intercept, $\beta_1$: Treat (test), $\beta_2$: Type (squamous), $\beta_3$: Type (small cell), $\beta_4$: Type (adeno), $\beta_5$: Status, $\beta_6$: Time from diagnosis, $\beta_7$: Age, $\beta_8$: Prior therapy (yes)

| Observation | Log-Student $t$ | Log-Laplace | Log-exp. power | Log-logistic |
|-------------|-----------------|-------------|----------------|--------------|
| Median      | HPD 95%         | Median      | HPD 95%        | Median      |
| $\beta_0$   | 2.09 [0.76, 3.38] | 2.08 [0.82, 3.36] | 2.00 [0.71, 3.27] | 2.06 [0.73, 3.35] |
| $\beta_1$   | -0.08 [-0.46, 0.27] | -0.06 [-0.42, 0.28] | -0.09 [-0.47, 0.26] | -0.09 [-0.48, 0.26] |
| $\beta_2$   | 0.02 [-0.52, 0.57] | -0.03 [-0.56, 0.52] | -0.04 [-0.61, 0.52] | -0.01 [-0.56, 0.55] |
| $\beta_3$   | -0.73 [-1.25, -0.24] | -0.73 [-1.22, -0.24] | -0.73 [-1.22, -0.20] | -0.73 [-1.22, -0.20] |
| $\beta_4$   | -0.75 [-1.27, -0.21] | -0.67 [-1.17, -0.17] | -0.71 [-1.22, -0.16] | -0.77 [-1.27, -0.21] |
| $\beta_5$   | 0.04 [0.03, 0.05] | 0.04 [0.03, 0.04] | 0.04 [0.03, 0.05] | 0.04 [0.03, 0.05] |
| $\beta_6$   | 0.00 [-0.02, 0.02] | 0.01 [-0.02, 0.02] | 0.00 [-0.02, 0.02] | 0.00 [-0.02, 0.02] |
| $\beta_7$   | 0.01 [-0.01, 0.03] | 0.01 [-0.01, 0.02] | 0.01 [-0.01, 0.03] | 0.01 [-0.01, 0.03] |
| $\beta_8$   | -0.09 [-0.53, 0.32] | -0.11 [-0.50, 0.33] | -0.10 [-0.52, 0.33] | -0.10 [-0.53, 0.35] |
| $\sigma^2$  | 0.80 [0.47, 1.19] | 0.69 [0.47, 0.95] | 1.26 [0.57, 2.19] | 0.36 [0.26, 0.48] |
| $\theta$    | 5.22 [1.68,17.60] | —           | —              | —           |

Figure 5. $2 \times \log(BF)$ in favor of $M_1: \Lambda_i \neq \lambda_{ref}(U_i \neq u_{ref})$ versus $M_0: \Lambda_i = \lambda_{ref}(U_i = u_{ref})$, using independence Jeffreys’ prior. Horizontal lines reflect the interpretation rule of Kass and Raftery (1995).
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. Section 3.2 provides conditions for the existence of the posterior distribution based on an arbitrary mixing distribution. In particular, Theorem 3 provides 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 an MCMC sampling scheme. We recommend the use of set observations is explored. The use of set observations is provided some extra guidance and results for specific mixing.

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. Section 3.2 provides conditions for the existence of the posterior distribution based on an arbitrary mixing distribution. In particular, Theorem 3 provides 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 an MCMC sampling scheme. We recommend the use of set observations is explored. The use of set observations is provided some extra guidance and results for specific mixing.

Vallejos and Steel: Objective Bayesian Survival Analysis Using SMLN Distributions 707

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. Section 3.2 provides conditions for the existence of the posterior distribution based on an arbitrary mixing distribution. In particular, Theorem 3 provides 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 an MCMC sampling scheme. We recommend the use of set observations is explored. The use of set observations is provided some extra guidance and results for specific mixing.

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. Section 3.2 provides conditions for the existence of the posterior distribution based on an arbitrary mixing distribution. In particular, Theorem 3 provides 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 an MCMC sampling scheme. We recommend the use of set observations is explored. The use of set observations is provided some extra guidance and results for specific mixing.

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

$$
k_1(\theta) = nE_{T_i} \left[ \left( \log(t_i) - x_i' \beta \right)^2 \right] \times \left[ E_{H_i} \left( \Lambda_i f_{LN} \left( t_i \left| x_i', \sigma_i^2 \right. \right) \right) \right]^{-1}, \tag{A.1}
$$

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

$$
k_3(\theta) = \frac{n}{2} E_{T_i} \left[ \left( \log(t_i) - x_i' \beta \right)^{2} \right] \times \left[ E_{H_i} \left( \Lambda_i f_{LN} \left( t_i \left| x_i', \sigma_i^2 \right. \right) \right) \right]^{-1}.
$$

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

Corollary 1. The proof follows directly from Theorem 1 using the structure of the determinant of the Fisher information matrix and its submatrices.

Theorem 2. Define $t_0 = (t_1, \ldots, t_n)'$ and $D_n = \text{diag}(\lambda_1, \ldots, \lambda_n)$. 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 censored observations ($f_{T_r}(t_r)$). Therefore, a sufficient condition for existence of the posterior distribution of $(\beta, \sigma^2, \theta)$ is $f_T(t_r) < \infty$. After some algebraic manipulation, $f_T(t_r)$ is equal to

$$
\int_{\mathbb{R}^k} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \prod_{i=1}^{n_0} k_i^2 \frac{1}{(2\pi \sigma^2)^k} e^{-\frac{1}{2} \left( \lambda_i (\beta - a)^T (\beta - a) + \lambda_i^2 (D_i)_{ii} \right)}
\times \pi(\theta) \prod_{i=1}^{n_0} dP(\Lambda_i|\theta) d\theta d\sigma^2 d\beta,
$$

where $A = X_0' D_0 X_0$, $a = A^{-1} X_0' D_0 y_0$, $S^2(D_0, y_0) = y_0' D_0 y_0 - y_0' D_0 X_0 (X_0' D_0 X_0)^{-1} X_0' D_0 y_0$, and $y_0 = (\log(t_1), \ldots, \log(t_n))'$. Provided that $t_i \neq 0$ for all $i \in \{1, \ldots, n_0\}$, using Fubini’s theorem for the integral (A.5) and integrating first with respect to $\beta$, we have

$$
f_{T_r}(t_r) \propto \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \prod_{i=1}^{n_0} k_i^2 \frac{1}{\sqrt{\det(X_0' D_0 X_0)}} e^{-\frac{1}{2} \left( \lambda_i (\beta - a)^T (\beta - a) + \lambda_i^2 (D_i)_{ii} \right)}.$$
\[ f_{\theta_i}(t_i) \propto \int_{\mathbb{R}^{+}} \int_{\mathbb{R}^{+}} \cdots \int_{\mathbb{R}^{+}} \prod_{i=1}^{k} \psi^{\frac{\lambda_i}{2}} \frac{\prod_{i=1}^{k} \lambda_i}{\psi_{\lambda} + \psi_{\lambda_i}} \prod_{i=1}^{k} dP(\lambda_i | \theta) \, d\theta, \]  
(A.8)

where

\[ \prod_{i=1}^{k} \lambda_{mk} \equiv \max \left\{ \prod_{i=1}^{k} \lambda_i : \det (x_1, \ldots, x_k) \neq 0, l_1, \ldots, l_k \in \{1, \ldots, n\} \right\}. \]  
(A.9)

\[ \prod_{i=1}^{k} \lambda_{mk} \equiv \max \left\{ \prod_{i=1}^{k} \lambda_i : \det \left(\frac{x_{ki}}{\log(t_i)}, \ldots, \frac{x_{k+1,i}}{\log(t_{k+1,i})}\right) \neq 0, l_1, \ldots, l_{k+1} \in \{1, \ldots, n\} \right\}. \]  
(A.10)

(i) For \( p = 1 \). Barring a set of zero Lebesgue measure, \( \lambda_{mk+1} = \max \{\lambda_i : i \not\in \{m_1, \ldots, m_n\}\} \). Hence, (A.8) is bounded above by \( f_{\theta_i} \pi(\theta) d\theta = 1 \). If \( n_0 > k \), the posterior exists.

(ii) For \( p = 1 + k/2 \). By the same argument, (A.8) is bounded above by \( f_{\theta_i} \pi(\theta) d\theta = 1 \). \( f_{\theta_i} \pi(\theta) d\theta \) is \( \Lambda_{mk+1} \). Using the density of the first-order statistic it follows that \( f_{\theta_i} \pi(\theta) d\theta \leq \Lambda_{mk+1} \). Hence, as \( \lambda_{mk+1} 's \) are iid, the results holds.

Theorem 3.

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

\[ \left( \begin{array}{cc} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{array} \right) = \left( \begin{array}{cc} 0 & 0 \\ 0 & 0 \end{array} \right) \left( \begin{array}{cc} \psi(\frac{v}{2}) - \psi(\frac{v+1}{2}) \\ \frac{v}{v+3} - \frac{2(v+3)}{v(v+1)^2} \end{array} \right). \]  
(A.11)

Therefore, the components depending on \( v \) of the Jeffreys’, independence Jeffreys’, and independence I Jeffreys’ priors are, respectively,

\[ \pi'(v) \propto \left( \frac{v+1}{v+3} \right)^{1/2} \left( \frac{v}{v+3} \right)^{1/2} \left( \frac{v}{v+3} \right)^{1/2} \psi(\frac{v}{2}) - \psi(\frac{v+1}{2}) - \frac{2(v+3)}{v(v+1)^2}. \]  
(A.12)
also proper. In fact, for the log-Laplace model, $\Lambda_{-1}^{-1}$ is Gamma distributed and all its positive moments are finite. For the log-logistic model, it can be shown that $\Omega_2 = \sqrt{1/(4\lambda_0)}$ has an asymptotic Kolmogorov distribution with density function $g(o_1) = 8o_1^{-1/2}(1-o_1)^{1/2}e^{-2o_1^2}$, for $o_1 > 0$. Therefore, for $k > -2$, it follows that

$$E(\Lambda_{-1}^{k/2}) = 2^{k+1} \sum_{i=1}^{\infty} (-1)^{i} \frac{1}{i^{k/2}} \cdot$$

$$\times \int_{0}^{\infty} o_1^{k+1} e^{-2o_1^2} d o_1$$

$$= 2^{k+2} \sum_{i=1}^{\infty} (-1)^{i+1} 2^{i} \int_{0}^{\infty} \eta^{k/2} e^{-2\eta} d \eta$$

$$= 2^{k+2} \Gamma(1 + k/2) \sum_{i=1}^{\infty} (-1)^{i+1} \frac{1}{i^{k/2}} < \infty.$$  \hspace{1cm} (A.20)

For the log-exponential power model, the Fisher information matrix was derived by Martin and Perez (2009) and is given by

$$\pi'(\alpha) = \left[ \frac{\alpha(\alpha - 1)\Gamma(1 - 1/\alpha)}{\Gamma(1/\alpha)} \right]^{1/2} \times \frac{1}{\alpha} \sqrt{\frac{1}{1 + 1/\alpha}} \psi \left( \frac{1 + 1/\alpha}{\alpha} \right) - 1.$$  \hspace{1cm} (A.21)

$$\pi'(\sigma) = \frac{1}{\alpha} \sqrt{\frac{1}{1 + 1/\alpha}} \psi \left( \frac{1 + 1/\alpha}{\alpha} \right) - 1.$$  \hspace{1cm} (A.22)

$$\pi''(\alpha) = \frac{1}{\alpha^2} \left[ \frac{1}{1 + 1/\alpha} \psi \left( \frac{1 + 1/\alpha}{\alpha} \right) \right]^{2} - 1.$$  \hspace{1cm} (A.23)

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

$$E \left( \Lambda_{-1}^{-1} | \alpha \right) = \frac{\Gamma(3/2)}{\Gamma(1 + 1/\alpha)} E \left( W^{+1/2} | \alpha \right)$$

$$\times \frac{1}{\alpha} \sqrt{\frac{1}{1 + 1/\alpha}} \psi \left( \frac{1 + 1/\alpha}{\alpha} \right) + 1 + \psi \left( \frac{1 + 1/\alpha}{\alpha} \right) - 1.$$  \hspace{1cm} (A.24)

where $W \sim \text{Weibull}(\alpha/2, 1)$ and $Z \sim \text{Exponential}(1)$. The latter uses the lemma in Meintanis (1998), which states that a Weibull$(\alpha, 1)$ 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, $\lambda_{s+1}$ (defined in (A.10)) corresponds to $\lambda_{s+1}$, which represents the $(n_s - s)$th order statistic of $\lambda_1, \ldots, \lambda_n$. The rest of the proof is obtained by iteratively integrating (A.8), using the inequality in (A.15). After integrating with respect to $n_s = s - 1$ smallest $\lambda$, Equation (A.8) has a lower bound given by

$$\int_{0}^{\infty} \int_{x_1 < \cdots < x_{n_s+1}} \frac{n_s^{-1}}{\Gamma \left( \frac{n_s}{2} \right)} \prod_{i=1}^{n_s} dP(\lambda_i | y) \pi(\nu) d\nu,$$  \hspace{1cm} (A.25)

where $a = -x_{n_s+1} - \sum_{i=1}^{n_s} x_i + \frac{1}{\alpha} = \frac{1}{\alpha}$. Note that when integrating with respect to $\lambda_{s+1}$, we need $\alpha > 0$ to have a finite integral in (A.25). Hence, the propriety of the posterior distribution requires $v > \frac{a}{a-k+2(k-1)} = 1$.

**Theorem 5**. Define $I(s) = \int_{x_1 < \cdots < x_{n_s+1}} \int_{\beta, \sigma} f_{\sigma}(s) \beta, \sigma \pi(\beta, \sigma) d\beta d\sigma$. Based on the sample size $s$, the posterior distribution exists if and only if $\int_{x} I(s) ds$ is finite. As $E$ is bounded by $\int_{x} I(s) ds$ is finite as long as $I(s)$ is bounded from above except on a set of zero Lebesgue measure.

**SUPPLEMENTARY MATERIALS**

**Supplementary material A**: Implementation details and description of R-code.

**Supplementary material B**: Detailed results of the simulation study.

**Supplementary material C**: Convergence and mixing diagnostics for MCMC chains in the VA lung cancer data application.

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

[Received January 2013. Revised January 2014.]

**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. [702]

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. [697,698,704,705]

Chib, S. (1995), “Marginal Likelihood From the Gibbs Output,” *Journal of the American Statistical Association*, 90, 1315–1321. [701]

Chib, S., and Jeliazkov, I. (2001), “Marginal Likelihood From the Metropolis-Hastings Output,” *Journal of the American Statistical Association*, 96, 270–281. [701]

Cho, H., Ibrahim, J. G., Sinha, D., and Zhu, H. (2009), “Bayesian Case Influence Diagnostics for Survival Models,” *Biometrics*, 65, 116–124. [702]

Fernández, C., and Steel, M. (1998), “On the Dangers of Modelling Through Continuous Distribution: A Bayesian Perspective,” in *Bayesian Statistics*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 213–238. [700,701]

(1999), “Multivariate Student-t Regression Models: Pitfalls and Inference,” *Biometrika*, 86, 153–167. [701,708,709]

(2000), “Bayesian Regression Analysis With Scale Mixtures of Normals,” *Econometric Theory*, 16, 80–101. [698,701,708]

Fonseca, T., Ferreira, M., and Migon, H. (2008), “Objective Bayesian Analysis for the Student-t Regression Model,” *Biometrika*, 95, 325–333. [708]

Geisser, S., and Eddy, W. (1979), “A Predictive Approach to Model Selection,” *Journal of the American Statistical Association*, 74, 153–160. [702]
Hanson, T. (2006), “Inference for Mixtures of Finite Polya Tree Models,” Journal of the American Statistical Association, 101, 1548–1565. [702]
Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M. (2009), Robust Methods in Biostatistics, Wiley Series in Probability and Statistics, Chichester, UK: Wiley. [697,707]
Hogg, R. V., and Klugman, S. A. (1983), “On the Estimation of Long Tailed Skewed Distributions With Actuarial Applications,” Journal of Econometrics, 23, 91–102. [698]
Holmes, C., and Held, L. (2006), “Bayesian Auxiliary Variable Models for Binary and Multinomial Regression,” Bayesian Analysis, 1, 145–168. [701]
Jewell, N. (1982), “Mixtures of Exponential Distributions,” The Annals of Statistics, 10, 479–484. [698]
Kalbfleisch, J., and Prentice, R. (2002), The Statistical Analysis of Failure Time Data (2nd ed.), Hoboken, NJ: Wiley. [697,704,707]
Kaas, R., and Raftery, A. (1995), “Bayes Factors,” Journal of the American Statistical Association, 90, 773–795. [702,706]
Kass, R., and Raftery, A. (1995), “Bayes Factors,” Journal of the American Statistical Association, 90, 773–795. [702,706]
Lange, K., Little, R., and Taylor, J. (1989), “Robust Statistical Modeling Using the t Distribution,” Journal of the American Statistical Association, 84, 881–896. [701]
Lee, E., and Wang, J. (2003), Statistical Methods for Survival Data Analysis (3rd ed.), Hoboken, NJ: Wiley. [704,705]
Marshall, A. W., and Olkin, I. (2007), Life Distributions, New York: Springer. [697]
Martín, J., and Pérez, C. (2009), “Bayesian Analysis of a Generalized Lognormal Distribution,” Computational Statistics and Data Analysis, 53, 1377–1387. [698,701,709]
McCulloch, R. E. (1989), “Local Model Influence,” Journal of the American Statistical Association, 84, 473–478. [702]
Meintanis, S. (1998), “Moment-Type Estimation for Positive Stable Laws With Applications,” IAENG International Journal of Applied Mathematics, 38, 26–29. [709]
Patriota, A. (2012), “On Scale-Mixture Birnbaum-Saunders Distributions,” Journal of Statistical Planning and Inference, 142, 2221–2226. [698]
Peng, F., and Dey, D. K. (1995), “Bayesian Analysis of Outlier Problems Using Divergence Measures,” The Canadian Journal of Statistics, 23, 199–213. [702]
Roberts, G., and Rosenthal, J. (2009), “Examples of Adaptive MCMC,” Journal of Computational and Graphical Statistics, 18, 349–367. [701]
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, 399–420. [707]
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. [698]
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, Series B, 64, 583–640. [701]
Tanner, M., and Wong, W. (1987), “The Calculation of Posterior Distributions by Data Augmentation,” Journal of the American Statistical Association, 82, 528–540. [701]
Uppuluri, V. (1981), “Some Properties of Log-Laplace Distribution,” in Statistical Distributions in Scientific Work 4, eds. C. Taillie, G. P. Patil, and B. A. Baldessari, Dordrecht: Reidel, pp. 105–110. [698]
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. [702]
Vianelli, L. (1983), “The Family of Normal and Lognormal Distributions of Order r,” Metron, 41, 3–10. [698]
West, M. (1984), “Outlier Models and Prior Distributions in Bayesian Linear Regression,” Journal of the Royal Statistical Society, Series B, 46, 431–439. [702]
Wienke, A. (2010), Frailty Models in Survival Analysis, Boca Raton, FL: Chapman & Hall/CRC. [697]