Supplementary material B: Simulation study
“Objective Bayesian Survival Analysis Using Shape Mixtures of Log-Normal Distributions”
C.A. Vallejos and M.F.J. Steel

This document displays the results of the simulation study described in Section 4 of the paper. The objective is to illustrate the performance of the proposed methodology and to assess the effectiveness of the suggested Bayesian model comparison criteria.

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 as a combination of sample size ($n = 100, 500$) and percentage of censoring ($PC = 10\%, 70\%$). Independent censoring times are sampled from a uniform distribution in $(0, C)$ where the value of C is tuned in order to control the percentage of censoring. 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. The log-normal and the mixture models introduced in Table 1 of the paper are fitted to each dataset. We use set observations with $\epsilon_l = \epsilon_r = 0.5$ for non-censored observations. MCMC chains are run for 300,000 iterations with a burn-in period of 75,000 and thinning period equal to 50 (*i.e.* we use 4,500 draws for the results presented here).

For AFT models, β is usually the parameter of interest. Its interpretation is not affected by our mixing scheme. We compare the performance of different SMLN-AFT models based on the posterior median $\hat{\beta}$. Under each scenario, the values of $\hat{\beta}$ over the 100 simulations are displayed. The value of β used to generate the data is indicated by a horizontal line. The Bayesian model comparison criteria described in Section 3.5 of the paper are applied to each dataset. We report the number of times in which each model was chosen using the deviance information criteria (DIC), Bayes factors (BF) and pseudo Bayes factors (PsBF).

Estimation of β under the independence Jeffreys prior

Figure 1: Boxplot of $\hat{\beta}$ for data generated using a log-normal model.

Figure 2: Boxplot of $\hat{\beta}$ for data generated using a log-Student t model ($\nu = 5$).

Figure 3: Boxplot of $\hat{\beta}$ for data generated using a log-Student t model ($\nu = 20$).

Figure 4: Boxplot of $\hat{\beta}$ for data generated using a log-Laplace model.

Figure 5: Boxplot of $\hat{\beta}$ for data generated using a log-exp. power model ($\alpha = 1.2$).

Figure 6: Boxplot of $\hat{\beta}$ for data generated using a log-exp. power model ($\alpha = 1.8$).

Figure 7: Boxplot of $\hat{\beta}$ for data generated using a log-logistic model.

Bayesian model comparison criteria for the Jeffreys prior

Figure 8: Distribution of Bayesian model choice under the Jeffreys prior. From darkest to lightest (and left to right): log-normal, log-Laplace, log-exp. power and log-logistic.

Bayesian model comparison criteria for the independence Jeffreys prior

Figure 9: Distribution of Bayesian model choice under the independence Jeffreys prior. From darkest to lightest (and left to right): log-normal, log-Student t , log-Laplace, log-exp. power and log-logistic.

Table 1: Percentage of correct classification.

Simulated model	Jeffreys prior						independence Jeffreys prior						
	PC=10%			PC=70%			PC=10%			PC=70%			
	DIC	BF	PsBF	DIC	BF	PsBF	DIC	BF	PsBF	DIC	BF	PsBF	
$n = 100$	log-normal	73	17	71	69	0	61	75	71	72	65	62	60
	log-Student t ($\nu = 5$)	-	-	-	-	-	-	11	0	10	5	1	6
	log-Student t ($\nu = 20$)	-	-	-	-	-	-	0	0	0	0	0	1
	log-Laplace	75	58	73	66	38	65	72	73	68	62	62	60
	log-exp. power ($\alpha = 1.2$)	1	0	3	2	0	4	0	1	2	2	12	6
	log-exp. power ($\alpha = 1.8$)	2	0	3	4	0	3	5	4	7	2	13	1
	log-logistic	72	98	77	27	94	33	42	51	50	28	31	33
$n = 500$	log-normal	94	86	90	88	48	88	91	96	87	88	88	86
	log-Student t ($\nu = 5$)	-	-	-	-	-	-	61	20	62	23	8	26
	log-Student t ($\nu = 20$)	-	-	-	-	-	-	13	0	16	1	0	2
	log-Laplace	91	96	90	85	83	87	89	94	87	82	85	82
	log-exp. power ($\alpha = 1.2$)	62	1	62	19	0	20	57	44	60	12	19	17
	log-exp. power ($\alpha = 1.8$)	23	0	26	1	0	3	21	8	22	1	5	3
	log-logistic	81	96	80	74	91	76	74	85	72	71	74	70

The choice between one of the three Jeffreys-rule based priors suggested in the article is not too critical when making inference about β . For each data-generating model, all priors produced similar estimates of the regression parameters when fitting the same model. Figures 1-7 show the posterior median $\hat{\beta}$ across simulations, adopting the independence Jeffreys prior (which is the only prior that produces a valid posterior for all our examples). Of course, the most accurate estimations arise when the data provides more information, *i.e.* $n = 500$ 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 . In addition, 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 and can accommodate log-normal tails less well). As expected, if the data generating mechanism involves 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 when $PC = 70\%$. 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.

Figures 8 and 9 summarize the results from the model comparison criteria described in Subsection 3.5. Both types of independence Jeffreys priors produce similar results as they only differ for the log-exponential power model. Hence, we only display results under the Jeffreys and independence Jeffreys priors. 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 for the higher percentage of censoring). 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 little spread to

accommodate the data, even though they were generated by the same 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 $PC = 70\%$). In any case, jointly, these criteria provide a confident assessment of 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 a rather difficult task but it can be achieved for large sample sizes. The percentages of correct classification under each scenario are shown in Table 1. 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. For example, the log-Laplace model is a frequent choice for log-exponential power data with $\alpha = 1.2$. The DIC and PsBF criteria do best overall: under both priors they lead to correct classification of models with moderate or strong heterogeneity on the basis of 500 observations with low censoring in at least 57% of the cases.