# Supplementary Material
## for "Incorporating unobserved heterogeneity in Weibull survival models: A Bayesian approach"

## A. Proofs and Proposition 1

*Proof of Theorem 1.* For (i): We follow the proof in [4], which assumed $\alpha = 1$. Using l'Hopital's rule twice, it follows that $\lim_{t \to 0} \frac{\log(-\log(S(t|\alpha,\gamma,\theta)))}{\log(t)} = \gamma$ if and only if $\mathrm{E}(\Lambda|\theta) < \infty$. For (ii): $S(t|\alpha,\gamma,\theta)$ is the Laplace transform of the density of $\alpha\Lambda$ evaluated at $t^\gamma$, which is unique. □

*Proof of Result 1.* $\mathrm{E}(T_i^r)$ exists if and only if $\int_0^\infty t_i^r f(t_i|\alpha,\gamma,\theta)\,dt_i < \infty$. Therefore, using Fubini's theorem and the $r$-th moment of the Weibull distribution it follows that $\mathrm{E}(T_i^r) = \Gamma\left(1 + r/\gamma\right)\alpha^{-r/\gamma}\mathrm{E}_{\Lambda_i}(\Lambda_i^{-r/\gamma}|\theta)$. An expression for $cv$ is a direct application of this result. □

*Proof of Result 2.* This proof consists in taking the expectation of minus the second derivatives of the likelihood for each observation and computing the FIM on the basis of the whole sample as the sum of the FIM for single observations. Define $Y_i = e^{-x_i'\beta}T_i$. The functions $k_1(\theta)$, $k_2(\theta)$ and $k_3(\theta)$ are

$$
\begin{aligned}
k_1(\theta) &= nE_{T_i}\left[\frac{\left[\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\left(1 - \lambda_i Y_i\right)\,dP(\lambda_i|\theta)\right]^2}{\left[\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\,dP(\lambda_i|\theta)\right]^2}\right] \\
&\quad -nE_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\left(1 - 3\lambda_i Y_i + Y_i^2\right)\,dP(\lambda_i|\theta)}{\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\,dP(\lambda_i|\theta)}\right]
\end{aligned}
\tag{S1}
$$

$$
\begin{aligned}
k_2(\theta) &= nE_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\left(1 - \lambda_i Y_i\right)\,dP(\lambda_i|\theta)\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\frac{d}{d\theta}\,dP(\lambda_i|\theta)}{\left[\int_{\mathcal{L}} \lambda_i\, e^{-\lambda_i Y_i}\,dP(\lambda_i|\theta)\right]^2}\right] \\
&\quad -nE_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-e^{-x_i'\beta}\lambda_i T_i}\left(1 - \lambda_i T_i\, e^{-x_i'\beta}\right)\frac{d}{d\theta}\,dP(\lambda_i|\theta)}{\int_{\mathcal{L}} \lambda_i\, e^{-e^{-x_i'\beta}\lambda_i T_i}\frac{d}{d\theta}\,dP(\lambda_i|\theta)}\right]
\end{aligned}
\tag{S2}
$$

$$
k_3(\theta) = nE_{T_i}\left[\frac{\left[\int_{\mathcal{L}} \lambda_i\, e^{-e^{-x_i'\beta}\lambda_i T_i}\frac{d}{d\theta}\,dP(\lambda_i|\theta)\right]^2}{\left[\int_{\mathcal{L}} \lambda_i\, e^{-e^{-x_i'\beta}\lambda_i T_i}\,dP(\lambda_i|\theta)\right]^2}\right] - nE_{T_i}\left[\frac{\int_{\mathcal{L}} \lambda_i\, e^{-e^{-x_i'\beta}\lambda_i T_i}\frac{d^2}{d\theta^2}\,dP(\lambda_i|\theta)}{\int_{\mathcal{L}} \lambda_i\, e^{-e^{-x_i'\beta}\lambda_i T_i}\,dP(\lambda_i|\theta)}\right].
\tag{S3}
$$

These functions do not depend on $\beta$ because all the terms inside the expectations depend on $T_i$ and $\beta$ only through $Y_i$ and the distribution of $Y_i$ does not depend on $\beta$ nor $i$. □

**Proposition 1** *Let $t = (t_1, \ldots, t_n)'$ be the recorded survival times of $n$ independent individuals, realizations of random variables distributed as in equation (6) (see main text). Without loss of generality, assume*

*only the first $n_o$ observations are uncensored ($n_o \leq n$) and let $t_o$ be the vector containing all uncensored observations. A sufficient condition for the existence of $\pi(\beta, \gamma, \theta|t)$ is the propriety of $\pi(\beta, \gamma, \theta|t_o)$.*

*Proof of Proposition 1.* The likelihood contribution of censored observations is $1 - F_T(t_i)$, $F_T(t_i)$ or $F_T(t_{i2}) - F_T(t_{i1})$ for right, left or interval censoring, respectively. These quantities are upper bounded by 1. So, under independence, the likelihood function is bounded by the likelihood of $t_o$. Hence, the marginal likelihood of the complete data is bounded by the marginal likelihood of these $n_o$ observations and the result follows. $\quad\square$

*Proof of Theorem 2.* The posterior distribution of $(\beta, \gamma, \theta)$ given the data is proper if and only if

$$\int_{\mathbb{R}^k} \int_0^\infty \int_\Theta \int_{\mathbb{R}_+^n} \gamma^n \prod_{i=1}^n t_i^{\gamma-1} \, e^{-\gamma \sum_{i=1}^n x_i'\beta} \prod_{i=1}^n \lambda_i \, e^{-\sum_{i=1}^n e^{-\gamma x_i'\beta} \lambda_i t_i^\gamma} \pi(\gamma, \theta) \prod_{i=1}^n dP(\lambda_i|\theta) \, d\theta \, d\gamma \, d\beta < \infty.$$

(S4)

We use Fubini's theorem to change the order of the integrals. We integrate with respect to $\beta$, using a similar argument as in [6]. For $t_i > 0$ and any value of $(\beta, \gamma, \lambda_i) \in \mathbb{R}^k \times \mathbb{R}_+$, $f_{T_i}(t_i|\beta, \gamma, \lambda_i)$ is bounded by a finite constant. Therefore, the integral in (S4) has an upper bound proportional to

$$\int_{\mathbb{R}^k} \int_0^\infty \int_\Theta \int_{\mathbb{R}_+^k} \gamma^k \prod_{i \in I} t_i^{\gamma-1} \, e^{-\gamma \sum_{i \in I} x_i'\beta} \prod_{i \in I} \lambda_i \, e^{-\sum_{i \in I} e^{-\gamma x_i'\beta} \lambda_i t_i^\gamma} \pi(\gamma, \theta) \prod_{i \in I} dP(\lambda_i|\theta) \, d\theta \, d\gamma \, d\beta, \quad \text{(S5)}$$

for any $I = \{(i_1, \ldots, i_k) : 1 \leq i_1 < \cdots < i_k \leq n\}$. Define the transformation $U = g(\beta) = X^*\beta$, where $X^*$ is a $k \times k$ matrix containing the $i_1, \ldots, i_k$ rows of $X$. The rank of $X^*$ is $k$ since $X$ is of full rank. Therefore, $g(\cdot)$ is bijective with Jacobian $\det((X^*)^{-1})$ and (S5) is proportional to

$$\int_0^\infty \int_\Theta \int_{\mathbb{R}_+^k} \gamma^k \prod_{i=1}^k t_i^{\gamma-1} \prod_{i=1}^k \lambda_i \left[ \prod_{i=1}^k \int_{-\infty}^\infty e^{-\gamma u_i} \, e^{-e^{-\gamma u_i} \lambda_i t_i^\gamma} \, du_i \right] \prod_{i=1}^k \pi(\gamma, \theta) \, dP(\lambda_i|\theta) \, d\theta \, d\gamma. \quad \text{(S6)}$$

Using $w_i = e^{-\gamma u_i}$, the later integral simplifies to $\prod_{i=1}^k t_i^{-1} \int_0^\infty \int_\Theta \pi(\gamma, \theta) \, d\theta \, d\gamma$. Therefore, if $t_i \neq 0$ for all $i \in I$ and $\pi(\gamma, \theta)$ is a proper prior for $(\gamma, \theta)$, then the posterior distribution of $(\beta, \gamma, \theta)$ given the data exists. $\quad\square$

# B. Bayesian inference implementation details

Bayesian inference for the AFT-RMW model under the weakly informative priors presented in Section 3.1 is implemented via an Adaptive Metropolis-within-Gibbs algorithm with Gaussian Random Walk proposals [9]. We assume right-censoring, which is most common for survival data. Mixing parameters are handled through data augmentation [11]. Although the usual approach for dealing for censored (and set) observations is also by data augmentation, we do not use it for that because the Weibull survival function has a known simple form [5, 7]. Code to implement RMW-AFT regression models is provided as an R library. This includes the MCMC algorithm, criteria for Bayesian model comparison and a procedure for outlier detection. Because of space reasons, some details of the implementation were omitted in the article. These are explained below. Throughout, equation numbers refer to the main article.

## Update of $\gamma$

In principle, $\gamma$ can take any value in $(0, \infty)$. However, numerical problems are observed when very small values of $\gamma$ are proposed, in which case $cv^W(\gamma)$ becomes very large (in fact, $cv^W(\gamma) \to \infty$ when $\gamma \to 0$). As a solution, regardless of the mixing distribution, we truncate the range of $\gamma$ to $(0.06, \infty)$. This has no practical consequences as such small values of $\gamma$ are very rarely required for real datasets.

## Update of the mixing parameters

This implemented sampler involves the update of $\Lambda_1, \ldots, \Lambda_n$ at each step of the chain. This may be computationally inefficient (especially in situations in which sampling from the mixing variables is cumbersome). In order to avoid this problem, the $\Lambda_i$'s will be sampled only every $Q$ iterations of the chain. The value of $Q$ is chosen by considering the Effective Sample Size (ESS) of the chain and the CPU time required. Even though longer chains are required for $Q > 1$ (the mixing of the chains is affected), the reduction in term of CPU time can be substantial.

For the inverse gamma and inverse Gaussian mixing distributions, the full conditionals of the mixing parameters are given by Generalized Inverse Gaussian distributions. The algorithm proposed in [8] is used when sampling from these full conditionals.

## Detection of outliers and influential observations

For each observation, the models $M_0 : \lambda_i = \lambda_{ref}$ and $M_1 : \lambda_i \neq \lambda_{ref}$ are contrasted. Bayes factors between them can be computed as the generalized Savage-Dickey density ratio proposed by [12] and stated in (23). When the parameter $\theta$ does not appear in the model, this simplifies to the original version of the Savage-Dickey ratio in (24). However, if $\theta$ is unknown, the procedure is computationally intensive, since a reduced run of the MCMC algorithm (in which $\lambda_i$ is fixed at $\lambda_{ref}$) is required for each observation $i$. Nevertheless, as the $n$ runs are independent, the process can be easily parallelized. Our R functions for outlier detection receive $i$ as input in order to facilitate this. In a multi-core environment, each run can be sent to a different node (e.g. using the "parallel" library in R).

Evidence of influential observations is evaluated through the effect on the posterior distribution when delet-

ing one observation at the time. This evidence is quantified by means of the Kullback-Leiber divergence function $K_i = KL(\pi(\beta, \sigma^2, \theta|t), \pi(\beta, \sigma^2, \theta|t_{-i}))$. As explained in [3] it can easily be computed from MCMC output. Occasionally, numerical issues can lead to a negative estimate of $K_i$. In such a situation, a warning is printed.

## Marginal likelihood calculation

The marginal likelihood is defined as

$$m(t) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{\Theta} f(t|\beta, \sigma^2, \theta)\pi(\beta, \sigma^2, \theta) \, d\beta \, d\sigma^2 \, d\theta,$$

and $\log(m(t))$ is computed using the algorithm in [1] and [2].

## Convergence and mixing of the sampler

Convergence of the MCMC chains was never a problem with the burn-in used for the examples displayed in Section 4. Mixing is very good for models without an extra parameter $\theta$ in the mixing distribution (e.g. Weibull and RMW model with exponential(1) mixing). In the presence of an unknown $\theta$, reliable inference is obtained through the described algorithms, but the chains are mixing a bit less well for some of the parameters (higher autocorrelation and occasional excursions to extreme values), requiring MCMC run lengths of the order used throughout the case studies described in the manuscript.

# C. Simulation study

This section displays the results of a simulation study designed to illustrate the performance of the proposed methodology and to assess the effectiveness of the suggested Bayesian model comparison criteria.

Covariates were simulated by randomly sub-sampling the Cerebral Palsy dataset described in Section 4.2 of the manuscript. Due to the small proportion of children where 1 and 2 impairments were observed, these categories were merged into a single one for the purpose of the simulation (this is to avoid sub-sampled datasets that contain a covariate with constant value across all observations). Regression coefficients were defined as the posterior estimates obtained for the full Cerebral Palsy dataset under a Weibull-AFT regression model (i.e. $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (3.56, 3.32, 1.56, -0.07)'$). Datasets were simulated by random sampling from the following models: (i) Weibull-AFT, (ii) RMW-AFT with log-normal(0,2) mixing and (iii) RMW-AFT with exponential(1) mixing to cover scenarios with increasing amounts of unobserved heterogeneity. Throughout, we used $\gamma = 1$ (which is the most challenging case for inference), but similar results were observed for other values of $\gamma$ (not shown).

Four different scenarios are defined as a combination of sample size ($n = 100, 500$) and percentage of censoring (high: 85%-95%; low: 35%-45%). 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, 20 independent datasets are simulated under each scenario. The Weibull-AFT and RMW-AFT models (with exponential, Gamma and log-normal mixing) were fitted to each dataset. These mixing distributions were chosen to cover the models used to generate the data, also including the Gamma mixing distribution as it is a popular choice in the context of frailty models.

MCMC chains are run for 100,000 iterations with a burn-in period of 25,000 and thinning period equal to 10 (i.e. we use 7,500 draws for the results presented here). Results of these simulations are displayed in terms of the performance of the adopted Bayesian model comparison criteria and the bias observed in the estimation of model parameters.

## C.1. Model comparison

Here, we evaluate the performance of the suggested Bayesian model comparison criteria. Since BF and PsBF estimates were highly correlated (see Figure 10 in the manuscript), only PsBF results are shown. In all cases, PsBF is calculated with respect to the Weibull-AFT model (no mixing) so that, in log scale, positive values of PsML indicate that a RMW-AFT model (presence of unobserved heterogeneity) is preferred.

As additional information, the results based on the Deviance Information Criteria (DIC) [10] are also displayed. The DIC provides a fast tool for model comparison. It is defined as $\text{DIC} \equiv \text{E}(\text{D}(\beta, \gamma, \theta, t)|t) + p_D$, where $\text{D}(\beta, \gamma, \theta, t) = -2\log(f(t|\beta, \gamma, \theta))$ (deviance function) and $p_D = \text{E}(\text{D}(\beta, \gamma, \theta, t)|t) - \text{D}(\hat{\beta}, \hat{\gamma}, \hat{\theta}, t)$ (effective number of parameters) with $\hat{\beta}, \hat{\gamma}$ and $\hat{\theta}$ being the posterior medians of $\beta, \gamma$ and $\theta$, respectively. DIC

is computed using the marginal model (after integrating the $\lambda_i$'s), and lower values suggest better models. Results are displayed in terms of the DIC difference with respect to the Weibull-AFT model (no mixing). Hence, negative values of this difference indicate that a RMW-AFT model (presence of unobserved heterogeneity) is preferred.

Figures S1, S2, S3 and S4 display the results of these comparisons for the four simulation scenarios described above in terms of sample size and censoring percentage.

For datasets generated using the Weibull-AFT model with no mixing (no unobserved heterogeneity), we observe that RMW-AFT models indexed by a free parameter $\theta$ can accommodate the data by adjusting the value of this parameter. This leads to similar PsML and DIC values across all models (yet the Weibull-AFT model with no mixing will still be preferred as a more parsimonious choice). In contrast, the RMW-AFT model with exponential(1) mixing is less flexible and does not provide a good fit for data with little or no unobserved heterogeneity. This is more pronounced when the data are more informative (either the sample size is large or the censoring percentage is small). These results suggest that the proposed Bayesian model comparison criteria do not incorrectly prefer the mixture models in the absence of unobserved heterogeneity.

If the data is generated by a model that includes an unobserved heterogeneity term (i.e. log-normal(0,2) and exponential(1) mixing), the evidence in favour of the mixture models strongly depends upon the sample size: only marginal evidence for $n = 100$ but strong evidence for $n = 500$ (particularly when the censoring percentage is low). However, while the information contained in the data is enough to distinguish between the cases with and without unobserved heterogeneity, selecting the correct mixing distribution is more challenging (at least for the simulated sample sizes and censoring percentages).


## C.2. Parameter estimation

This section summarises the results of the simulations in terms of the bias observed in the estimation of model parameters. For each parameter, the bias is estimated as the difference between its posterior median and the true value used to generate the data. These results are summarized in Figures S5, S6, S7 and S8.

Our simulations suggest some bias in the estimation of the regression coefficients when the sample size is small. However, as it can be expected, this effect is attenuated for larger sample sizes (e.g. for $n = 500$) and as the censoring becomes less important.

Overall, our findings are in line with what is observed for the Cerebral Palsy dataset. In particular, we found that the bias observed in the estimation of the regression coefficients (excluding the intercept term) is somewhat robust to the choice of mixing distribution. This is similar to the small differences observed between the regression coefficients estimated under each mixture for the Cerebral Palsy data and can be interpreted as an analogy of what occurs in linear regression models with heteroscedastic errors, where least square estimates remain unbiased when fitting a homoscedastic model (in fact, RMW-AFT models can be interpreted as a heteroscedastic linear regression model for the log-survival times).

A different result is observed for the estimation of the intercept parameter $\beta_0$ and Weibull shape parameter

$\gamma$. Firstly, we observe that biases on the estimation of $\gamma$ tend to be compensated by biases in the estimation of $\beta_0$. This is not surprisingly due to the weak identifiability that occurs between the intercept and $\gamma$ when $\gamma \approx 1$ (see last paragraph in Section 3.1). We also note that the choice of mixing distribution substantially affects the amount of the bias that is observed for these parameters. This is critical, as a biased estimate of $\gamma$ can lead to a wrong interpretation of the subject-level hazard rate.

In particular, our simulations suggest that, in order to attenuate this bias, it is critical for the mixing distribution to be flexible enough and accommodate the strength of unobserved heterogeneity that generated the data. For example, models that are indexed by a free parameter $\theta$ (e.g. Gamma and log-normal mixing) can accommodate data generated by a Weibull-AFT model by adjusting the value of $\theta$. Instead, the exponential mixing distribution is less flexible and does not perform well for data simulated using a Weibull-AFT model. Similarly, since the exponential mixing distribution is linked to very strong levels of unobserved heterogeneity (e.g. leading to an infinite value for $cv$ for $\gamma = 1$), models that only allow for moderate or no unobserved heterogeneity (i.e. Gamma, log-normal and no mixing) lead to higher bias when estimating the parameters.

## D. Software and Example Results

Code to implement RMW-AFT regression models is provided as an R library freely available at **URL suppressed for reasons of anonymity**. Additionally, the R code used to generate the results displayed in Section 4 of the manuscript is provided in the additional file "SuppMaterialSectionD.html". The latter also includes numerical and graphical summaries of the analysed datasets, together with traceplots and other convergence diagnostics for the associated MCMC chains. Please note that the vignette in the additional html file does not work using Internet Explorer, but works well for other browsers such as Mozilla Firefox, Google Chrome and Safari.
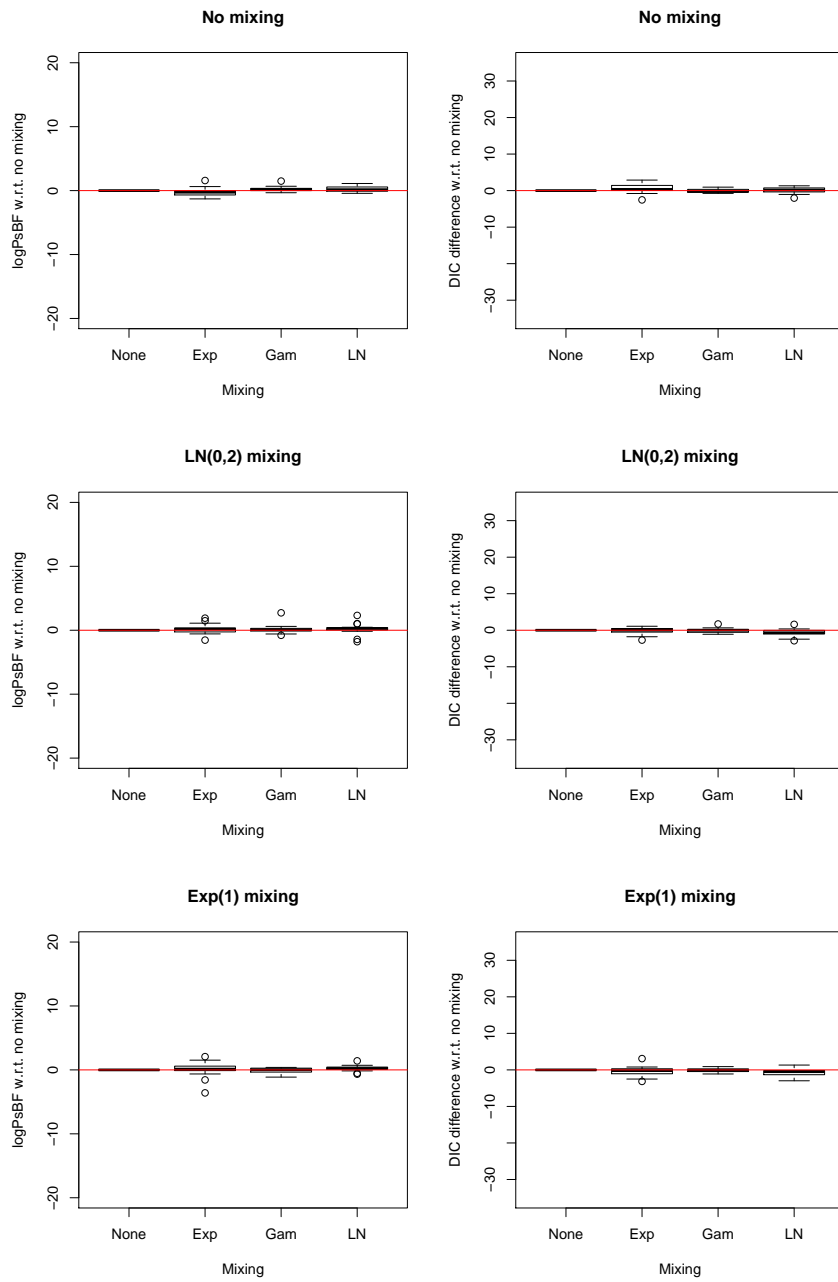
Figure S1: Simulation results for sample size $n = 100$, high censoring and $\gamma = 1$. Performance of PsML and DIC as model comparison criteria. Panel titles indicate the true model used to simulate the data.
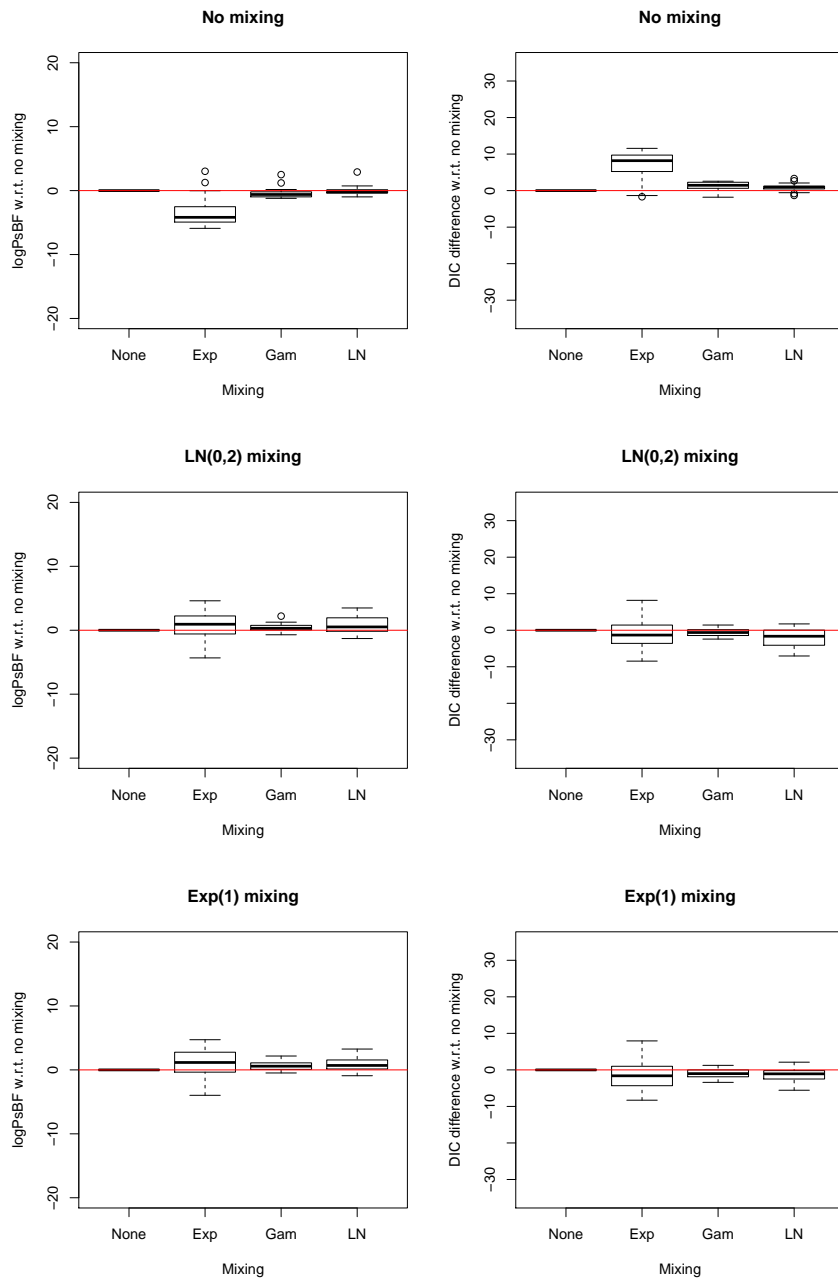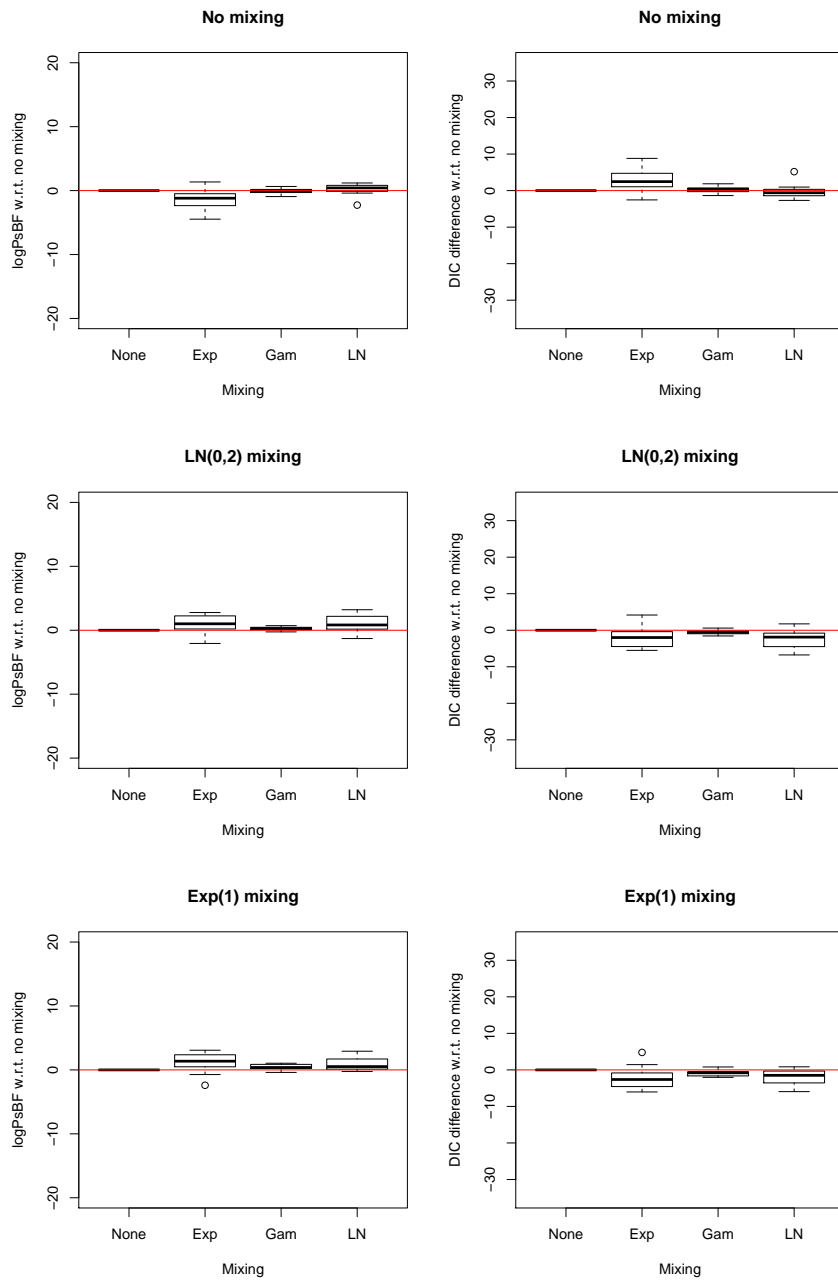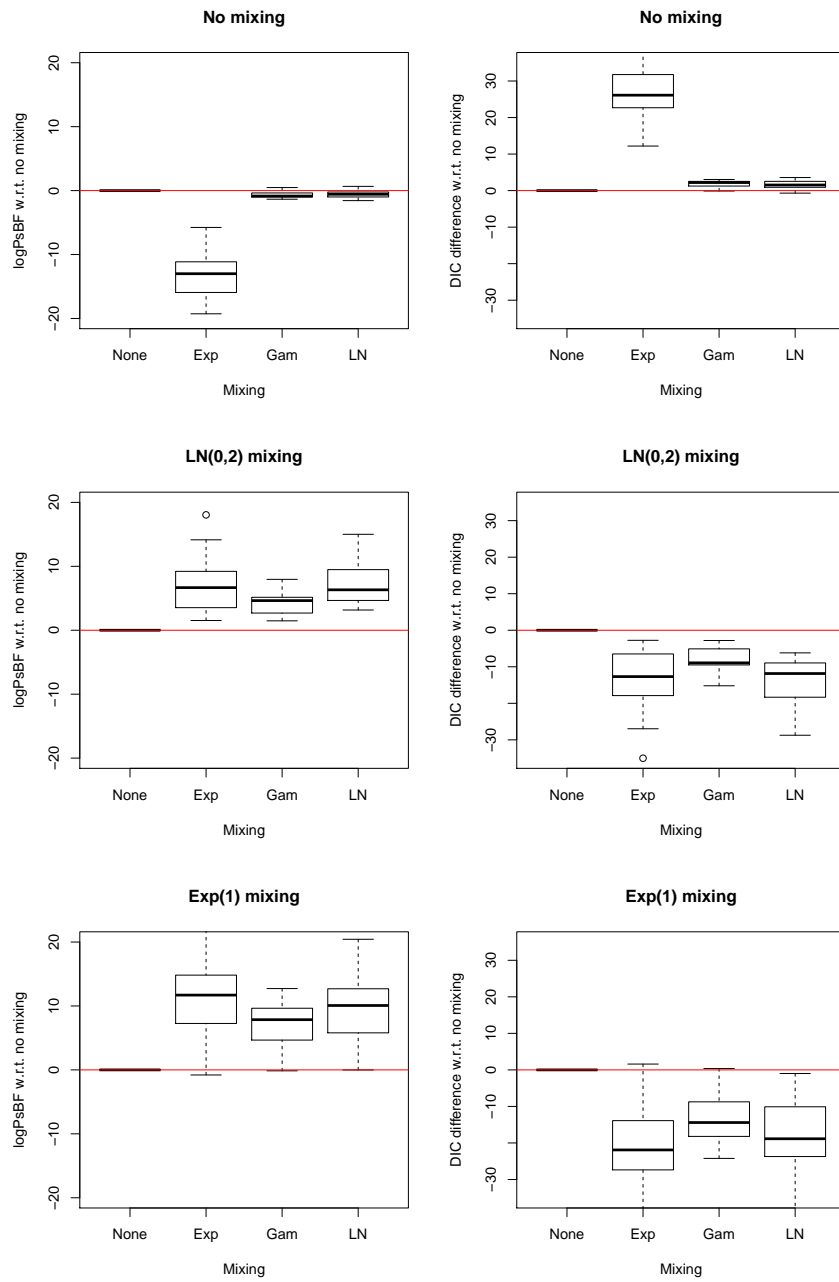
Figure S2: Simulation results for sample size $n = 100$, low censoring and $\gamma = 1$. Performance of PsML and DIC as model comparison criteria. Panel titles indicate the true model used to simulate the data.

9

Figure S3: Simulation results for sample size $n = 500$, high censoring and $\gamma = 1$. Performance of PsML and DIC as model comparison criteria. Panel titles indicate the true model used to simulate the data.

Figure S4: Simulation results for sample size $n = 500$, low censoring and $\gamma = 1$. Performance of PsML and DIC as model comparison criteria. Panel titles indicate the true model used to simulate the data.
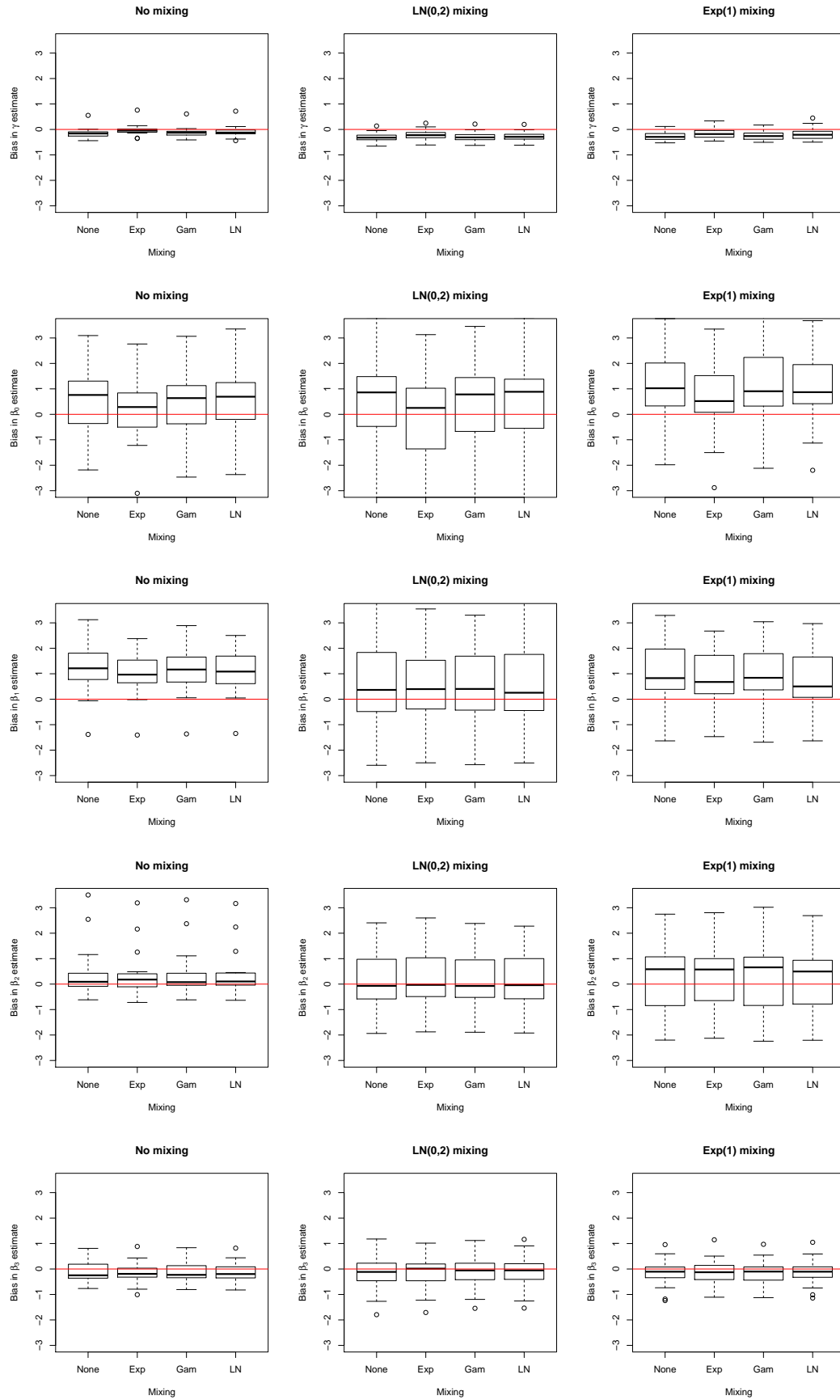
Figure S5: Simulation results for sample size $n = 100$ and high censoring. From top to bottom, panels display the bias on the estimation of $\gamma$, $\beta_0$, $\beta_1$ and $\beta_2$. Posterior estimates are defined as the posterior median of each parameter. Panel titles indicate the true model used to simulate the data.
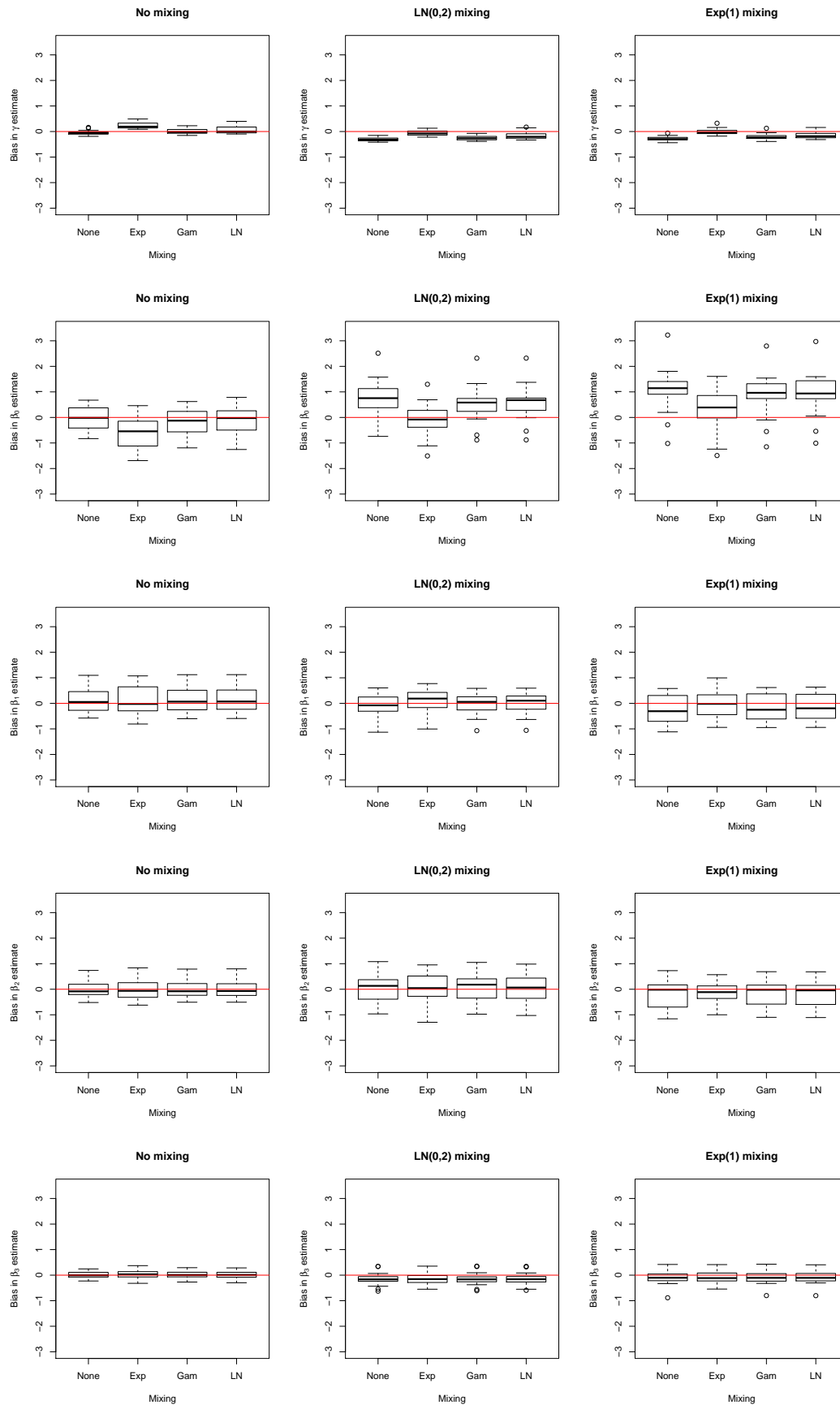
Figure S6: Simulation results for sample size $n = 100$ and low censoring. From top to bottom, panels display the bias on the estimation of $\gamma$, $\beta_0$, $\beta_1$ and $\beta_2$. Posterior estimates are defined as the posterior median of each parameter. Panel titles indicate the true model used to simulate the data.
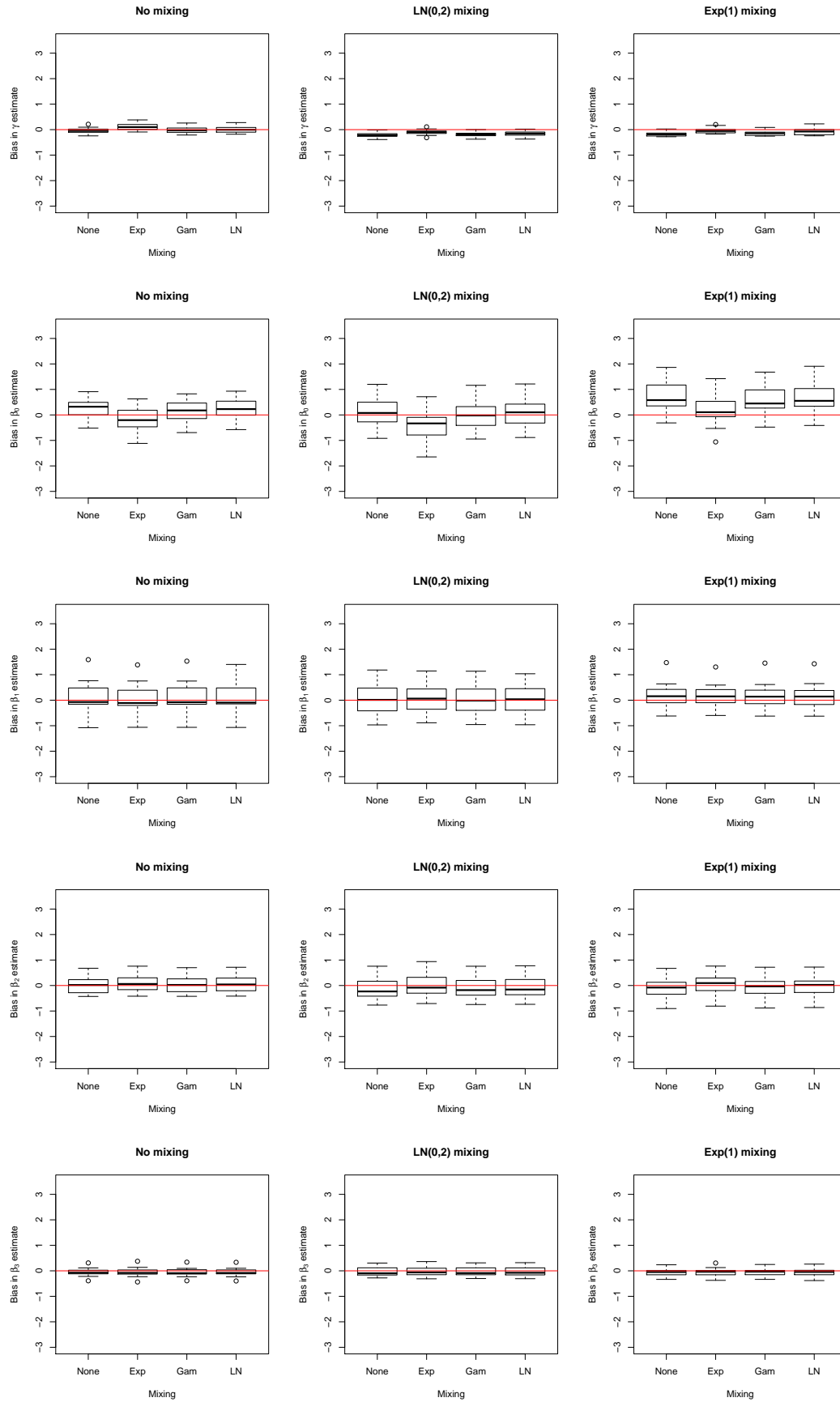
Figure S7: Simulation results for sample size $n = 500$ and high censoring. From top to bottom, panels display the bias on the estimation of $\gamma$, $\beta_0$, $\beta_1$ and $\beta_2$. Posterior estimates are defined as the posterior median of each parameter. Panel titles indicate the true model used to simulate the data.
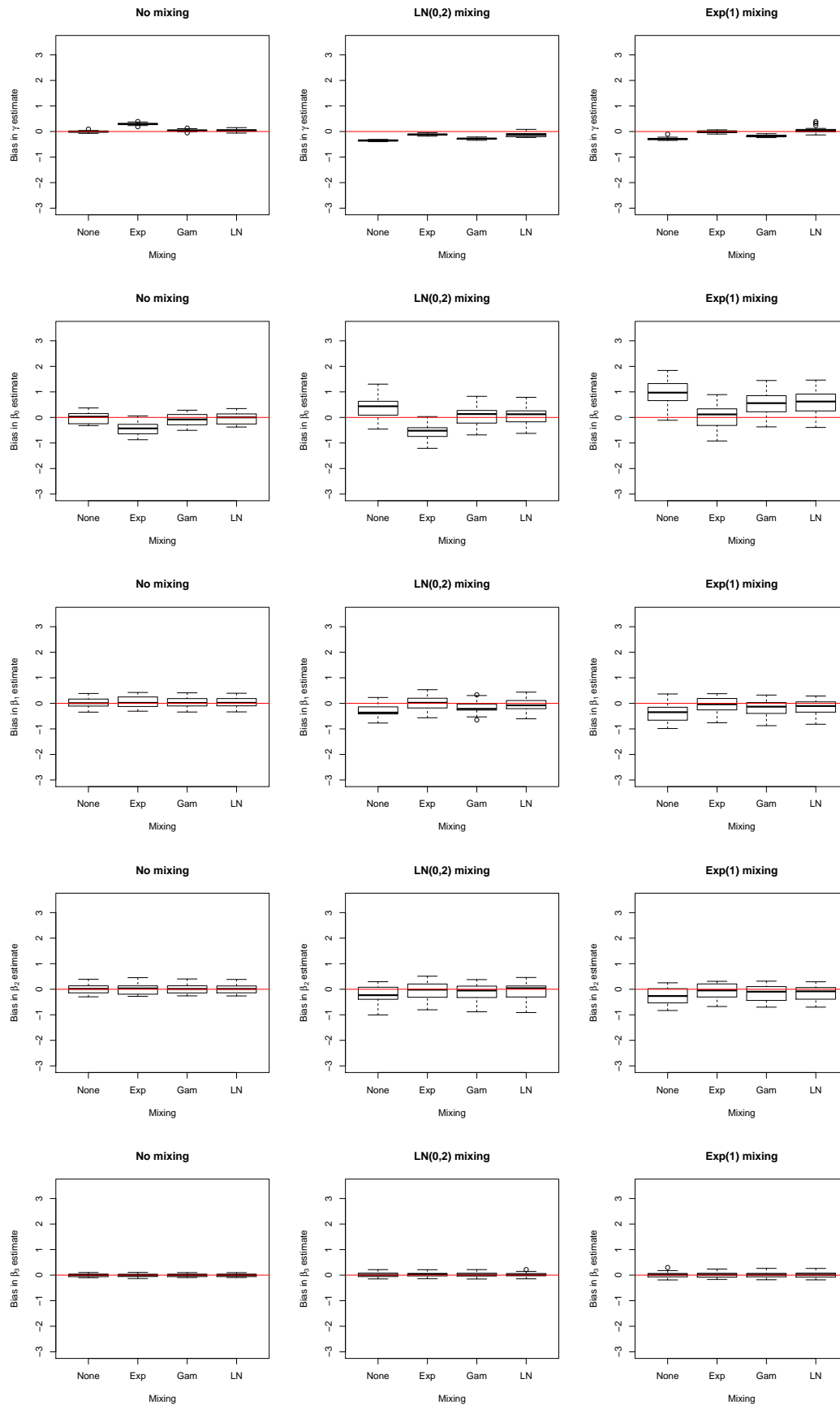
Figure S8: Simulation results for sample size $n = 500$ and low censoring. From top to bottom, panels display the bias on the estimation of $\gamma$, $\beta_0$, $\beta_1$ and $\beta_2$. Posterior estimates are defined as the posterior median of each parameter. Panel titles indicate the true model used to simulate the data.

# References

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

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

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

[4] B.E. Honoré. Simple estimation of a duration model with unobserved heterogeneity. *Econometrica*, 58:453–473, 1990.

[5] J.G. Ibrahim, M-H. Chen, and D. Singha. *Bayesian Survival Analysis*. Springer, 2001.

[6] S.W. Kim and J.G. Ibrahim. On Bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis*, 6:331–341, 2000.

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

[8] J. Leydold and W. Hörmann. Generating generalized inverse Gaussian random variates by fast inversion. *Computational Statistics & Data Analysis*, 55(1):213–217, 2011.

[9] G.O. Roberts and J.S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367, 2009.

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

[11] M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

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