atthd: An R package for accelerated failure time model using MCMC for high dimensional time-to-event data

Atanu Bhattacharjee
atanustat@gmail.com

University of Dundee  https://orcid.org/0000-0002-5757-5513

Gajendra K Vishwakarma
Indian Institute of Technology Dhanbad: Indian Institute of Technology

Pragya Kumari
Indian Institute of Technology Dhanbad: Indian Institute of Technology

Fatih Tank
Ankara University: Ankara Universitesi

Research Article

Keywords: Accelerated Failure Time model, Weibull, log-linear, log-logistic, high-dimensional data, survival analysis

Posted Date: August 9th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4681664/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
afthd: An R package for accelerated failure time model using MCMC for high dimensional time-to-event data

Atanu Bhattacharjee¹,*, Gajendra K. Vishwakarma², Pragya Kumari¹, and Fatih Tank³

¹Department of Mathematics and Computing, Indian Institute of Technology Dhanbad, Jharkhand -826004, India , pskkumari1990@gmail.com, statscurenetwork@gmail.com
²Division of Population Health and Genomics., Medical School, University of Dundee, United Kingdom , abhattacharjee001@dundee.ac.uk
³Department of Actuarial Sciences, Ankara University, Ankara, Turkey,, fatih.tank@ankara.edu.tr
*Corresponding authors

ABSTRACT

High-dimensional data, along with the time-to-event outcome analysis, is a challenging task. The accelerated failure time (AFT) model is an alternative to the Cox proportional hazard (CPH) model in survival analysis. However, there is a lack of available packages and functions to work on high-dimensional time-to-event data for the AFT model using Bayesian. We developed an R package ‘afthd’ that works with an advanced AFT model for high-dimensional time-to-event data under the Bayesian paradigm and also provides different diagnostics plots for univariate and multivariable Bayesian analysis. We attempted to present a computer code with open-source R software to work with high-dimensional data for the AFT model. The conventional AFT model has been extended for the Bayesian framework of log-normal, Weibull, and log-logistic AFT models for univariate and multivariable high-dimensional data contexts. The methodology is validated by performing the simulation technique, showing consistent results for parameters in all three types of parametric AFT models. The application part is also performed on two different real high-dimensional liver cancer datasets, which clearly reveals the proposed method’s significance by obtaining inferences for survival estimates for the disease. The developed package ‘afthd’ is competent to work with high-dimensional time-to-event data using the conventional AFT model and the Bayesian paradigm. Other aspects, like missing covariates in high-dimensional data and competing risks analysis, are also covered in this work.

keywords: Accelerated Failure Time model, Weibull, log-linear, log-logistic, high-dimensional data, survival analysis.

Introduction

In 2020, cancer was responsible for almost 10 million deaths and 19.3 million new cases globally, according to a study by Sung et al. (Sung et al., 2021). This highlights the need for continued improvement in cancer treatment and research to reduce mortality rates. Time-to-event survival analysis is commonly used to analyze mortality rates, with the outcome variable being the survival time until the event or last follow-up. Identifying prognostic biomarkers in cancer can help improve treatment strategies (Gajendra et al., 2021). There have been many efforts to identify biomarkers for cancer, but progress in time-to-event data methodology has been relatively limited.

The survival analysis is broadly framed into three categories: parametric, semiparametric, and non-parametric. If the pattern of survival time follows a particular distribution, it is considered parametric. Parametric models are more appropriate than semi-parametric and non-parametric models when the exact distribution of survival duration is known (David & Mitchel, 2012). They offer consistency with the theoretical survival function, complete specification with hazard and survival function, and the ability to make time-quantile predictions (Pochiraju & Seshadri, 2019). If there is no information about the distribution pattern, non-parametric models are suitable. Another alternative is the Cox proportional hazard (CPH) model (Cox.,1972), which is popular due to its minimal assumptions.

A log-linear regression model is the Accelerated Failure Time (AFT) model, which is an alternative to the CPH model for statistical modeling in time-to-event data (Wei., 1992). This model is mainly used for analyzing censored clinical trial data. The AFT model assumes that the predictor variable’s effect is accelerating or decelerating the life course of any disease(or interest of study) by some constant, whereas the CPH model assumes that they multiply the hazard by a constant (Prabhash et al., 2016). When the distribution pattern is known, the parametric form of the AFT model provides a robust statistical inference compared to the hazard function, particularly with the Bayesian approach (Saikia & Barman, 2017). The Bayesian approach is the alternative way to analyze the survival data, and it depends upon new knowledge or prior information of the experimental data (Wong et al., 2005). In this article,
we use Markov Chain Monte Carlo (MCMC) simulation to estimate the model’s parameter posteriorly. MCMC is a computational technique that is of great importance in Bayesian statistics, as it facilitates the estimation of intricate probability distributions, which are fundamental to Bayesian modeling and inference. This is a stochastic approach for producing posterior summaries and making predictions, and the process is done by simulating Markov chains from the posterior distribution of model parameters (Körper et al., 2015); Andrieu and Thomas (2008).

Working with high-dimensional survival data is a significant challenge in statistics, particularly with a small sample size (n) and a large number of covariates (p) (Zhao., 2018). Analyzing all covariates together is difficult in such cases, making variable selection techniques essential. Some of these techniques include LASSO (Tibshirani., 1996), elastic net (Zou & Hastie, 2005), Bayesian elastic net (Li & Lin, 2010), Bayesian variable selection in AFT (Zhang et al., 2018), weight function method for variable selection (Vishwakarma et al., 2021), feature selection method (Jovic et al., 2015), and feature selection using machine learning (Bhattacharjee et al., 2022).

There is extensive literature on the AFT model and its extensions. Although there are R packages to fit the AFT model, they cannot work with high-dimensional data using the AFT model along with Bayesian. For example, R package bayesSurv (Komarek et al., 2020) provides estimates of mixed-effects AFT model for censored data specification, package spBaysSurv (Zhou et al., 2017) also includes AFT model for spatial and non-spatial survival data, package RobustAFT (Marazzi & Muralti, 2011) is available for robust AFT regression for Gaussian and log-Weibull case, and function lss in package lss2 provides estimates for right-censored data in the AFT model which is based on the least-squares principle (Huang & Jin, 2007).

We propose a new methodological support package called afthd (Bhattacharjee et al., 2021) to fill the gap. This work provides methodologies to estimate parameters of the AFT model for high-dimensional data with different features, such as the conventional AFT model, the AFT model using MCMC, and data with competing risks for both univariate and multivariable analysis. The methodologies can handle missing values in covariates. Various estimates for augmented data are also presented through the depicted package ‘afthd’. The variable selection approach is also introduced in this work. Variable or efficient marker selection in any disease has been made before any predictive modeling (AFT model using Bayesian). We demonstrate the effectiveness of the proposed methodology on two real examples of high-dimensional cancer data utilizing the developed package ‘afthd’. The package is available on CRAN at https://cran.r-project.org/package=afthd. The article explains the methodology through two high-dimensional liver cancer data examples and validates this through simulations. The content of this article is organized as follows: Section 2 describes the model and methods used. Section 3 is there for the explanation of the ‘Program design and implementation’ of the package afthd (Bhattacharjee et al., 2021) using example head and neck cancer data. Section 4 provides the performance of the simulation study. Application of the package is provided in Section 5, utilizing two real high-dimensional gene expression datasets of liver cancer. The summary and discussion of the study are outlined in Section 6.

Model and Methods

Variable Selection

Data with multiple dimensions, or high-dimensional data, often have a larger number of predictor variables (covariates) than sample size n. In that case, interpreting data with all covariates is difficult. Hence, to address this, reducing unwanted covariates is necessary. Regression models use ordinary least squares or Maximum Likelihood Estimate (MLE) methods for parameter estimation, which can lead to overfitting if inclined too heavily towards the training data. Regularization helps to prevent overfitting by adding a penalty term to the error function, tuning the model, and reducing error. There are three types of regularization: L1 or LASSO, L2 or ridge, and elastic net regularization. The response variable $y_j \in \mathbb{R}$ and predictor variables $x_j \in \mathbb{R}^p$, where $j = 1, 2, ..., n$, sample size or the number of covariates is denoted by $p$ and $n$ is the number of individuals, can be expressed as a linear regression model:

$$y_j = x_j^\prime \beta + \varepsilon$$

(1)

Where, $\beta = \{\beta_1, \beta_2, ..., \beta_{p+1}\}$, $x_j = \{1, x_{j1}, x_{j2}, ..., x_{jp}\}$, such that $\beta_1$ is intercept and $\beta_2, \beta_3, ..., \beta_{p+1}$ are regression coefficients corresponding to the covariates $\{x_1, x_2, ..., x_p\}$, and $\varepsilon$ is the error term associated in the prediction. Hence the sum of squared residuals of equation (1), say $Y_e$ can be expressed as

$$Y_e = \sum_{j=1}^{n} (y_j - x_j^\prime \hat{\beta})^2$$

LASSO (Least Absolute Shrinkage and Selection Operator) or L1 regularization is achieved by adding a penalty
term $\lambda \sum_i |\beta_i|$ to $Y_e$, where $\lambda$ is constant, $i = 1, 2, ..., p$ and $\beta$ is the parameter estimate. The Lasso penalized function $R_L$ is expressed as

$$R_L = Y_e + \lambda \sum_i |\beta_i| ; \quad i = 1, 2, ..., p$$ (2)

If $\lambda = 0$, equation (2) reduces to the ordinary least squares method. Ridge or L2 regularization is achieved by adding a penalty term $\lambda \sum_i \beta_i^2$ to $Y_e$. The ridge penalized function $R_R$ can be described as

$$R_R = Y_e + \lambda \sum_i \beta_i^2 ; \quad i = 1, 2, ..., p$$ (3)

Again, if $\lambda = 0$, equation (3) becomes the ordinary least squares method. If both penalty terms of L1 and L2 regularization are incorporated with the sum of squared residuals $Y_e$, then it becomes elastic net regularization. Elastic net penalized function $R_{en}$ can be expressed as

$$R_{en} = \frac{Y_e}{2n} + \lambda \left[ \alpha \sum_i |\beta_i| + \frac{(1-\alpha)}{2} \sum_i \beta_i^2 \right] ; \quad i = 1, 2, ..., p$$ (4)

where $\lambda \geq 0$ is the complexity parameter and $\alpha \in [0, 1]$. If $\alpha = 0$, equation (4) acts as Ridge, $\alpha = 1$ acts as LASSO, and $0 < \alpha < 1$ represents the Elastic Net regularization (Hastie & Qian, 2016, Z).

**Accelerated failure time models**

This scenario is presented for high-dimensional time-to-event data with $n$ individuals (sample size) and $p$ covariates. All covariates are denoted as a vector $x : \{x_1, x_2, ..., x_p\}$. The Log-linear regression model for survival time $T_1, T_2, ..., T_n$ is taken as the AFT model and expressed as

$$log(T_i) = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + ... + \beta_p x_{i(p-1)} + \beta_{p+1} x_{ip} + \sigma \epsilon_i$$
$$= x_i^T \beta + \sigma \epsilon_i ; \quad i = 1, 2, ..., n$$ (5)

The term $log(T_i)$ denotes log-transformed survival time, $\beta_1$ is intercept, $(\beta_2, \beta_3, ..., \beta_{p+1})$ are regression coefficients corresponding to covariates $(x_1, x_2, ..., x_p)$ respectively, $\sigma = 1/\sqrt{\pi}$ is scale parameter, $\epsilon_i \overset{iid}{\sim} F_\epsilon(.)$. Where $F_\epsilon$ is known as cumulative density function (CDF), defined on the real line for corresponding density $f_\epsilon$ and hazard $h_\epsilon$ (Christensen et al., 2011). It is shown in Table 1 for log-normal, Weibull, and log-logistic distribution in the AFT model. When equation (5) holds, it is expressed as

$$T_i \overset{ind}{\sim} AFT(F_\epsilon, \beta, \tau | x_i)$$ (6)

AFT model is based on the assumption that covariates act proportionally (multiplicatively) concerning the survival time with the assumption

$$s(t|x) = s_0(\exp(\beta'x)t) ; \quad t \geq 0$$ (7)

where $s(t|x)$ is the survival function at the time $t$ for covariate $x$, $s_0(\exp(\beta'x)t)$ be the baseline survival function at time $t$, and $\exp(\beta'x)$ is the acceleration factor. The covariate effect is said to be decelerated if $\exp(\beta'x) > 1$, and the covariate effect is said to be accelerated if $\exp(\beta'x) < 1$.

When baseline distribution in equation (5) is Normal, Logistic, or Extreme value then corresponding to these, is represented by the AFT model for Log-normal, Log-logistic, or Weibull distribution, respectively.

Using model (5), the survival time $T$ is expressed as
\[ S(t|x, \beta, \tau) = P_r(T > t|x, \beta, \tau) \]
\[ = P_r \left[ \frac{\log(T) - x' \beta}{\sigma} > \frac{\log(t) - x' \beta}{\sigma} \right] | x, \beta, \tau \]
\[ = P_r \left[ \epsilon > \frac{\log(t) - x' \beta}{\sigma} \right] | x, \beta, \tau = S_\epsilon(t) \]
\[ = 1 - F_\epsilon \left[ \frac{\log(t) - x' \beta}{\sigma} \right] \quad (8) \]

The corresponding probability density function is derived from this as
\[ f(t|x, \beta, \tau) = \frac{1}{t \sigma} f_\epsilon \left[ \frac{\log(t) - x' \beta}{\sigma} \right] \quad (9) \]

All stated equations in this section work for multivariable in high-dimensional data, i.e., for vector \( x \), a set of covariates \( x_1, x_2, ..., x_p \). Similarly, all these equations also work for a single covariate (univariate) when \( p \) becomes 1.

**Log normal AFT model**

Now, for fitting the high-dimensional data as a log-normal AFT model for multivariable analysis, \( \epsilon_i \) of equation (5) has a standard normal distribution, and hence \( T_i \) is distributed as log-normal. From equation (9), its density function comes out to be
\[ f(t|x, \beta, \tau) = \frac{1}{\sqrt{2\pi} t \sigma} \exp \left[ -\frac{1}{2\sigma^2} (\log(t) - x' \beta)^2 \right] ; \ t > 0 \quad (10) \]

Furthermore, it can be represented as \( T \sim LN(x' \beta, \sigma^2) \), i.e., when \( \log(T) \sim N(\mu, \sigma^2) \) for \( T > 0 \), where \( \mu = x' \beta \).

The survival function of the normal distribution is \( S(\epsilon) = 1 - \Phi(\epsilon) \quad (11) \)
and the distribution function of normal distribution and its cumulative hazard functions are
\[ \Phi(\epsilon) = \frac{\log t - x' \beta}{\sigma} \quad (12) \]
\[ H_\epsilon(\epsilon) = -\log \{1 - \Phi(\epsilon)\} \quad (13) \]

In this way, the log-normal AFT model is derived for multivariable. It has been used in the functions lgnbymv() and lgnbyuni() for multivariable and univariate posterior estimates, respectively.

**Weibull AFT model**

If \( \epsilon_i \) of equation (5) has extreme value distribution, then \( T_i \) follows Weibull distribution, and in this case, equation (5) will be denoted as the Weibull AFT model. i.e., \( T \sim Weib(\sqrt{\tau}, e^{-x' \beta \sqrt{\tau}}) \). Its pdf and cdf have been shown in Table 1. Exponential distribution is a particular case of Weibull. When \( \tau = 1 \) i.e., \( \sigma^2 = 1 \) then \( T \sim Weib(1, e^{-x' \beta}) \) is equivalent to exponential distribution.

To keep the interpretation of regression coefficients simpler in functions of the depicted package, we are using a modified extreme value distribution with a median 0. Hence, we redefine the extreme value cumulative distribution function as
\[ F_\epsilon(x) = 1 - \exp[-\log(2)e^x] \quad (14) \]
So we are using Weibull distribution that leads to AFT model when

$$T \sim Weib(\sqrt{\tau}, \log(2) \exp(-x'\beta\sqrt{\tau}))$$

(15)

We use these formulations to estimate the regression coefficient, scale parameter, shape parameter, and survival time. (Liu & Lim, 2018) has also used the Weibull AFT regression model to estimate survival time. A similar approach is used for univariate estimation.

**Log-logistic AFT model**

Considering model (5) i.e., \(T_i \sim AFT(F_\epsilon, \beta, \tau | x_i)\) with logistic error distribution. If baseline distribution is logistic, then \(T_i\) follows log-logistic distribution, and equation (5) is said to be a Log-logistic AFT model. Here it is expressed as

$$T_{\text{ind}} \sim \text{LogLogis}(x'\beta, \sqrt{\tau})$$

which is equivalent to

$$\log(T) \overset{iid}{\sim} \text{Logis}(x'\beta, \sqrt{\tau})$$

and corresponding cdf (using the arguments in equation (8)) is

$$F_\epsilon[(\log(t) - x'\beta)\sqrt{\tau}] = 1 - \frac{1}{1 + \exp[(\log(t) - x'\beta)\sqrt{\tau}]}$$

$$= 1 - S(t|x, \beta, \tau)$$

(16)

The log-logistic AFT model uses these functions to estimate parameters for multivariable and univariate covariates. We use these forms of AFT, i.e., log-normal, Weibull, and log-logistic, to analyze the high-dimensional data for multivariable and univariate cases using MCMC with variable selection technique.

**AFT model with smooth time functions**

Functions have been designed for the AFT model without Bayesian, i.e., a conventional approach using the smooth time functions. Generalized survival models deliver a general and flexible approach to analyze the clinical trial data. Survival function \(S(t|x_i)\) for covariates \(x_i\) and to time \(t\) modelled as an AFT model by

$$S(t|x_i) = S_0(t \exp(-\eta(x_i, t; \beta)))$$

(17)

Where \(\beta\) is the regression parameter and \(\eta\) is a function of \(t\) and covariates \(x\). If AFT is a model with a time-constant accelerated factor, in that case, \(\eta\) is the linear predictor. The baseline survival function is modeled as

$$S_0(t) = \exp(-\exp(\eta_0(\log(t); \beta_0)))$$

(18)

Where \(\beta\) is the regression parameter and \(\eta\) is a function of \(t\) and covariates \(x\). If AFT is a model with a time-constant accelerated factor, in that case, \(\eta\) is the linear predictor. The baseline survival function is modeled as

$$S_0(t) = \exp(-\exp(\eta_0(\log(t); \beta_0)))$$

(19)

Where \(\eta_0\) is a linear predictor, the combined regression model is defined as

$$S(t|x) = \exp(-\exp(\eta_0(\log(t) - \eta(x, t; \beta); \beta_0)))$$

(20)

Corresponding to it, the hazard function can be calculated as

$$h(t|x) = \frac{\partial}{\partial t}[-\log(S(t|x))]$$

$$= \exp[\eta_0\{\log(t) - \eta(x, t; \beta); \beta_0]\} \eta_0'\{\log(t) - \eta(x, t; \beta); \beta_0\}$$

$$+ \eta_0\left\{\frac{1}{t} - \eta'(x, t, \beta)\right\}$$

(21)
Where \( \eta_0 \) is the linear predictor and modeled using natural splines, and \( \eta \) can be freely modeled, provided this linear predictor is a smooth function of time. For design matrix \( X(t,x) \), linear predictor in the model using as \( \eta(x, t; \beta) = X(x, t)\beta \). Construction of the linear predictor considers flexibility in it, such that time effects are twice differential and smooth, allowing possible interaction between covariates and time \((\text{Clements et al., 2021})\). The estimation of the model is being done using maximum likelihood estimates (MLE) as we are interested in a fully parametric model.

**Bayesian Inference**

We have used the same prior for all three analyses using log-linear, Weibull, and log-logistic of the AFT model. For estimating the unknown parameters \( \beta \) and \( \tau \) (where, \( \sigma = 1/\sqrt{\tau} \)) of the AFT model provided in equation (5), we use the MCMC posterior simulation technique to get Bayesian inference. Our approach, which we are using in the AFT model with Bayesian, is based on conditional median priors \((\text{Bedrick et al., 2000})\). Here, the prior specification on which it is made is a collection of median responses, each corresponding to a particular covariate combination \((\text{Christensen et al., 2011})\). We need a joint prior such as \( p(\beta, \tau) \) that induces a distribution on regression coefficient. Here, we preferred to choose standard diffuse prior such as \( \tau \sim \text{Gamma}(0.001, 0.001) \).

Our preferred informative priors satisfy the assumptions of AFT model with Bayesian, is based on conditional median priors \((\text{Christensen et al., 2011})\). We need a joint prior such as \( p(\beta, \tau) \) and \( p(\tau) \) that induces a distribution on regression coefficient. Here, we preferred to choose standard diffuse prior such as \( \tau \sim \text{Gamma}(0.001, 0.001) \). It is reasonable to assume that all \( \beta \) and \( \tau \) are to be modeled, provided this assumption aligns with the notion that the observation of \( m \) can be considered independent.

Hence, if we specify \( p_m(.) = \prod_{i=1}^{p} p_{\tilde{m}_i}(.) \) as prior which we set on the vector of medians then

\[
p(\beta) \propto \prod_{i=1}^{p} p_{\tilde{m}_i}(e^{x_i'\beta})e^{x_i'\beta}
\]

and Jacobian of the transformation is

\[
\left| \frac{d}{d\beta} e^{X\beta} \right| = \left| \tilde{x}_i e^{x_i'\beta} : i = 1, \ldots, p \right| = \left| \text{Diag}\{e^{\tilde{x}_i'\beta}\} \tilde{X} \right| = \prod_{i=1}^{p} \exp(\tilde{x}_i'\beta) | \tilde{X} | \propto \prod_{i=1}^{p} \exp(\tilde{x}_i'\beta)
\]

Here, \( \tilde{m} \) is working as the scale of the data, and only we are required to specify one prior on \( \tilde{m} \). This same prior can be obtained on \( \beta \) for baseline distribution with a median of 0. It is a good way to make a single elicitation irrespective of which or how many such AFT models will be considered. So we preferred to define prior for \( \beta \) as \( \text{norm}(0, 0.000001) \) and \( \text{Gamma}(0.001, 0.001) \) prior for \( \tau \) for each, which is an approximation of \( p(\tau) = 1/\tau \), is taken as a choice to simple use \((\text{Christensen et al., 2011})\). The given dataset exhibits a mean value of 1, a mode value of 0, and a variance of 1000. The priors in the functions utilized to derive posterior estimates are provided below.

\[
\beta \sim \text{norm}(0, 0.000001)
\]

\[
\tau \sim \text{Gamma}(0.001, 0.001)
\]

**Model selection criteria** There are many varieties for model selection in Bayesian inference. We are using deviance information criterion (DIC) as a model selection approach, proposed by \((\text{Spiegelhalter et al., 2002})\). The DIC
delivers an assessment about a penalty of model complexity and fitting of the model. The deviance statistic for different cases of the parametric AFT model has been provided using the Bayesian paradigm, indicating which model is more appropriate. For model selection purposes, we have developed a function aftbybmv() in presenting the R package. It estimates parameters and deviance information for each model, such as log-normal, Weibull, and log-logistic AFT models, and displays the one having minimum deviance. More minor, the deviance for any model indicates a better-fitting model. All the functions available in package ‘afthd’, running with Bayesian for univariate or multivariable analysis, provide deviance information.

**Competing Risk method**

Since we work with cancer patients, any individual can die due to other causes except cancer. If the patient dies due to different reasons, it is known as competing risk (CR). Now, datasets having a death column need to be revisited by putting death=0 (alive or censored), 1 (died), and 2 (death due to other causes). We provide the provision to work with the AFT model in CR data using a Bayesian parametric system. The parametric approach is expanded in multivariable (wbyscrkm) and univariate (wbyscrku) setup in high-dimensional data in the Weibull AFT model.

**Missing data imputation**

Complications frequently arise in survival data analysis as a result of the existence of missing values within covariates. Missing values in the data can arise as a result of either record loss or experimental design. When a variable in a dataset is not measured for inadvertent or deliberate reasons, or when participants in a study fail to attend the scheduled follow-up or do not respond to specific questions. This study is applicable for analyzing partially observed data by addressing the issue of missing information by substituting the empirical mean of the available observations. This study employs an imputation technique to address the issue of missing values in covariates. In this context, the imputation method involves replacing missing values for a specific covariate with the mean value of that covariate, utilizing the available non-missing responses. Before implementing estimation procedures, the issue of missingness in covariates is addressed through diligent data monitoring. The functions within the illustrated R package offer support for managing missing data in the covariates as well as providing posterior and maximum likelihood estimates for various accelerated failure time (AFT) models.

**Program design and implementation**

We presented the R package afthd to estimate the parameters in the context of survival modeling for the parametric AFT model. It is prepared with fourteen core functions, and those are listed in Table 2. All the functions compatible to work with Bayesian are using the R2jags package. It makes fast execution of operations even at high iterations of MCMC. Functions pvaft and rglaf are working for estimates of parametric AFT model with smooth time function with conventional approach or without MCMC (uses packages rstpm2 and glmnet), other functions offer estimates for AFT model using MCMC. Table 2 provides a concise overview of the arrangement and brief descriptions of the functions from the depicted package. The package contains ‘Head and neck’ cancer data named ‘hdata’, through which we will provide a comprehensive overview by utilizing all the functions inbuilt in afthd software package. Each function provides execution time or run time for every implementation. The functions that operate in conjunction with MCMC methodology yield posterior estimations as well as various diagnostic visualizations, such as density plots, trace plots, and autocorrelation plots, for each parameter derived from the MCMC process.

The functions within the afthd package have been developed with distinct objectives within the framework of the AFT model. All functions of the package can handle data containing missing values in the covariates. The syntax or command for these functions is elucidated in this section. Some common arguments used in functions such as $m$ and $n$ denote the inclusion of covariates in the study available from the $m$th column to the $n$th column number of the high-dimensional data. ‘STime’ represents the individual’s survival duration in the data. ‘Event’ (1 represents the occurrence of the event and 0 for censored) is operated in all functions to provide patients’ death or event status in the data. The term ‘data’ refers to high-dimensional gene expression data that includes information on event status, survival time, and a set of covariates. The variables ‘nc’ and ‘ni’ represent the quantities of Markov chains and MCMC iterations, respectively. In the function, the variable ‘alpha’ represents a selected value within the interval $[0, 1]$ that serves to determine the regularisation method. The parameter ‘alpha’ is assigned a value of 1 for Lasso regularisation, 0 for Ridge regularisation, and a value between 0 and 1 is selected for the elastic net regularisation technique.

*Implementation in R*
<table>
<thead>
<tr>
<th>Function</th>
<th>Used method or function</th>
<th>Variable taken as</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pvaft</td>
<td><code>aft()</code> from <code>rstpm2</code></td>
<td>univariate</td>
<td>Estimates of covariates (together) with restriction on p value with parametric AFT model with smooth time functions without MCMC.</td>
</tr>
<tr>
<td>rgl aft</td>
<td><code>cv.glmnet()</code> from <code>glmnet</code>, <code>aft()</code> from <code>rstpm2</code></td>
<td>univariate</td>
<td>Estimates of selected variable using regularization (lasso, elastic net or ridge) in parametric AFT model using smooth time functions without MCMC.</td>
</tr>
<tr>
<td>wbysuni</td>
<td>AFT with Weibull dist</td>
<td>univariate</td>
<td>Posterior estimates of AFT model with Weibull distribution using MCMC (for all chosen covariates).</td>
</tr>
<tr>
<td>wbysmv</td>
<td>AFT with Weibull dist</td>
<td>multivariate</td>
<td>Posterior multivariate estimates of chosen covariates (maximum 5 at a time) for AFT model with Weibull distribution using MCMC.</td>
</tr>
<tr>
<td>lgnbyuni</td>
<td>AFT with log-normal</td>
<td>univariate</td>
<td>Posterior estimates of chosen covariates (all together) for AFT model with log normal distribution using MCMC.</td>
</tr>
<tr>
<td>lgnbymv</td>
<td>AFT with log-normal</td>
<td>multivariate</td>
<td>Posterior estimate of chosen covariates (max 5 at once) for AFT model with log normal distribution using MCMC.</td>
</tr>
<tr>
<td>lgstbyuni</td>
<td>AFT with log-logistic</td>
<td>univariate</td>
<td>Posterior estimates (all chosen together covariates) of AFT model with log logistic distribution using MCMC.</td>
</tr>
<tr>
<td>lgstbymv</td>
<td>AFT with log-logistic</td>
<td>multivariate</td>
<td>Posterior estimates of chosen covariates (max 5 at once) for AFT model with log logistic distribution using MCMC.</td>
</tr>
<tr>
<td>aftbybmv</td>
<td>Weibull, log-normal, log-logistic</td>
<td>multivariate</td>
<td>Posterior parametric estimate of AFT model with minimum deviance (DIC) among Weibull, log-normal and log-logistic distributions.</td>
</tr>
<tr>
<td>rglwbysu</td>
<td><code>cv.glmnet</code> from <code>glmnet</code>, Weibull dist</td>
<td>univariate</td>
<td>Bayesian univariate estimates (all selected together) of AFT model for selected covariates using regularization method.</td>
</tr>
<tr>
<td>rglwbysm</td>
<td><code>cv.glmnet</code> from <code>glmnet</code>, Weibull</td>
<td>multivariate</td>
<td>Posterior estimates of AFT model for selected covariates using regularization method (Estimates for maximum first 5 selected covariates).</td>
</tr>
<tr>
<td>wbyscrku</td>
<td>Weibull dist, Competing risk</td>
<td>univariate</td>
<td>Posterior estimates (for all chosen covariates) with competing risk using AFT model of Weibull distribution.</td>
</tr>
<tr>
<td>wbyscrkm</td>
<td>Weibull distribution, Competing risk</td>
<td>multivariate</td>
<td>Posterior estimates (max. 5 covariates at a time) with competing risk using AFT model of Weibull distribution.</td>
</tr>
<tr>
<td>wbyAgmv</td>
<td>Weibull dist, Augmented data</td>
<td>multivariate</td>
<td>Posterior multivariate estimate of AFT model with Weibull distribution using MCMC that supports augmented data.</td>
</tr>
</tbody>
</table>

Short descriptions and illustrations of all the functions and their interpretation are provided in this section. We are using the head and neck cancer dataset from the package `afthd`. The dataset contains the survival time of 565 patients, their death status, and the set of covariates. This dataset is named `hdata`, which will be loaded using the package.

```r
> library(afthd)
> data("hdata", package="afthd")
```

Next, the function `pvaft` is used to obtain the univariate estimate for the parametric AFT model using the smooth time function. It uses the conventional approach for providing estimates for selected covariates from column number `m` to `n` of the dataset. Here, argument `p` is used to restrict the p-value for selecting covariates. i.e., only those covariates will be picked, whose p-value is less than or equal to `p`. By default, it takes `p = 1`, meaning all covariates from `m` to `n` will be selected and their estimates obtained. For example, we choose covariates from column number `m = 13` to `n = 100`. Suppose we want only those covariates whose p-value is less than or equal to 0.01, hence taking
\( p = 0.01 \), then function \texttt{pvaft} results as below.

\begin{verbatim}
> pvaft(m=13,n=50, STime="os", Event="death", p=0.01, data=hdata)

Execution time: 1.317075
                 Estimate    std.Error    z.value   p_value
ZMAT1          0.1113351  0.03822706  2.912470  0.003585831
\end{verbatim}

This function obtains a data frame containing estimates of the selected covariates’ regression coefficients. The next function, \texttt{rglaft} is also developed for the AFT model using smooth time functions. Here, feature selection is first performed, and then the estimation of parameters using the AFT model. Feature selection is made in two steps. In the first step, covariates are selected using the regularization technique, and after that, based on the significance of covariates in the AFT model in the second step. Here, in regularization, ‘cox’ is chosen as a family. We believe that the Cox Model is widely used and expected in a large. So, initial variable selection is conducted by Cox, and then applied AFT model towards generating statistical inference. Consequently, we use the Cox model as the filtration method. This function provides univariate estimates for the selected covariates using the conventional approach. To implement this function \texttt{rglaft}, we choose covariates from column number \( m = 8 \) to \( n = 45 \), and \( \alpha = 1 \) to choose LASSO regularization as a variable selection technique in the first step.

\begin{verbatim}
> set.seed(1000)
> head(rglaft(m=8, n=45, STime="os", Event="death", alpha=1,
+ data=hdata))

Execution time: 2.685153
                 Estimate    Std..Error    z.value   Pr.z.
ZMAT1          0.11133515  0.03822706  2.912470  0.003585831
GAL3ST1        -0.13089781  0.05452930 -2.400504  0.016372503
LRP2           -0.09360935  0.04790015 -1.954260  0.050670451
NME5           -0.08829066  0.04911814 -1.797516  0.072253697
XIST           -0.02664405  0.01584490 -1.681553  0.092655478
C2orf72        -0.06179485  0.03962097 -1.559650  0.118842635
\end{verbatim}

It obtains the result as a similar pattern of \texttt{pvaft}. Hence, all the notations have the same meaning. Here, one thing can be noticed, in both functions, \texttt{pvaft} and \texttt{rglaft}, covariates are arranged in increasing order of their \( p \)-value in the outcome. It means upper covariates are more significant than lower ones.

**AFT model using MCMC**

Functions to estimate parametric AFT models using MCMC have been developed for log-normal, Weibull, and log-logistic distributions. All functions available in the package for univariate estimates display results for all chosen covariates at a glance. Whereas existing operations for multivariable estimation display estimates for the maximum first five selected covariates. All functions support missing values in covariates. Now, two functions are available in the package to analyze the data using the log-normal AFT model with MCMC. Function \texttt{lgnbyuni} is for univariate estimation and \texttt{lgnbymv} is for multivariable analysis. Execution of both procedures has been shown below for chosen covariates from column \( m \) to \( n \), and their outcomes can be displayed by calling the assigned objects \texttt{lgn1} and \texttt{lgn2}.

\begin{verbatim}
> lgn1<lgnbyuni(m=10,n=11, STime="os", Event="death", nc=4, ni=1000,
+ data=hdata)
> lgn2<lgnbymv(m=10,n=12, STime= "os", Event="death", nc=4, ni=1000,
+ data=hdata)

In the outcome, \texttt{lgnbyuni} provides estimates of the regression coefficient for listed covariates in the first column of the data frame. Whereas \texttt{lgnbymv} delivers estimates, in which \( \beta_1 \) is for intercept and other \( \beta \)’s for regression coefficients corresponding to covariates (in the order of selected covariates). Here, \( sigma \) is the scale parameter of the distribution, \( Rhat \) provides the convergence diagnostic, \( n.eff \) gives the efficient number of samples, and deviance obtains deviance Information criteria (DIC) of the model. Credible intervals and standard error are also available in
the result.

For estimating the parameters using the Weibull AFT model with MCMC, function `wbysuni` is developed for univariate estimation, and `wbysmv` is for multivariable. For example, we have chosen covariates from \( m = 15 \) to \( n = 16 \) to execute both these functions, which is shown below. Results have been assigned in objects `wb1` and `wb2`, which can be seen by calling it. The result of both functions contains a data frame of estimates of regression coefficients and their credible intervals.

```r
> wb1<-wbysuni(m=15, n=16, STime= "os", Event= "death", nc=4, ni=1000, +    data=hdata)
> wb2<-wbysmv(m=15, n=16, STime= "os", Event= "death", nc=4, ni=1000, +    data=hdata)
```

For analyzing the high-dimensional data using a log-logistic AFT model with MCMC, `lgstbyuni` is available in the package for univariate estimation and `lgstbymv` for multivariable. For example, choosing covariates from column number \( m = 55 \) to \( n = 56 \) to understand the working feature of these functions. Output in a data frame can be displayed by calling assigned objects `lgst1` and `lgst2` for `lgstbyuni` and `lgstbymv`, respectively. All the variables in output have the same meaning as explained before.

```r
> lgst1<-lgstbyuni(m=55,n=56, STime="os", Event="death", nc=3, +    ni=1000, data=hdata)
> lgst2<-lgstbymv(m=55, n=56, STime="os", Event="death", nc=3, +    ni=1000, data=hdata)
```

**Model selection in AFTs using MCMC**

A function `aftbybmv` is developed for multivariable posterior estimates using an appropriate model among Weibull, log-normal, and log-logistic (which has minimum deviance) AFT models. For example, we choose covariates as column numbers \( m = 10 \) to \( n = 11 \) from the head and neck cancer dataset available in the package. Execution of this function can be done as done below. Its result is stored in the object `aft` and this will be displayed by calling it. The result shows the log-logistic AFT model fits better than other models for this example.

```r
> aft<-aftbybmv(m=10, n=11, STime= "os", Event= "death", nc=3, +    ni=1000, data=hdata)
```

**Weibull AFT model with different specifications**

Functions `rglwbysu` and `rglwbysm` are available in the package, to provide univariate and multivariable posterior estimates, respectively. Both functions provide estimates using the Weibull AFT model for those covariates selected by regularization technique from chosen covariates \( m \) to \( n \). Here, for example, initially, we select covariates as column numbers \( m = 10 \) to \( n = 22 \) and choose \( \alpha = 1 \), i.e., LASSO regularization technique for selection of covariates. Objects `rglw1` and `rglw2` are assigned as a result of both functions. This implementation is shown below.

```r
> set.seed(1000)
> rglw1<-rglwbysu(m=10, n=22, STime= "os", Event= "death", nc=3, +    ni=1000, alpha=1, data=hdata)
> rglw2<-rglwbysm(m=10, n=22, STime= "os", Event= "death", nc=3, +    ni=1000, alpha=1, data=hdata)
```

Two functions, `wbyscrku` (for univariate) and `wbyscrkm` (for multivariable), have been developed and are currently accessible within the package for the purpose of analyzing high-dimensional data with competing risks. Both functions utilize the Weibull Accelerated Failure Time (AFT) model within a Bayesian framework to generate posterior estimates. For example, we choose covariates from column numbers \( m \) to \( n \) in the dataset `hdata` to implement both functions. The results of both functions are stored in objects `wcrk1` and `wcrk2`, and the outcome will be displayed by calling them. The event status in these functions is classified as 0 for censored, 1 for death due to study of interest, and 2 for the event’s occurrence due to other causes. Consequently, the outcome provides separate estimates for these two distinct scenarios. One for event status (0,1) and another for (0,2). Here, 0 (censored) is taken as a reference.
In data analysis, data augmentation is a technique that is used to artificially increase the amount of data or the sample size by adding slightly modified copies of already existing data. There is a function named `wbyAgmv` in the package, which provides multivariable posterior estimates for covariates as well as an estimate of survival time after $t$ days/months/years (same unit as survival time) for augmented data (i.e., augmented rows are already added in last rows of the dataset). Here, one more advantage is, that the estimate of survival time can also be estimated only for the desired number of individuals present in the row numbers $p$ to $q$ of the dataset. For example, we choose covariates from column $m = 10$ to $n = 12$, and we choose rows from $p=560$ to $q=565$ to estimate survival time for individuals, available in row numbers 560 to 565. It is executed as shown below, and its result is stored in the object `aug`. It can be displayed by calling it.

```r
> aug<-wbyAgmv(m=10, n=12, p=563, q=565, t=200, STime="os", + Event="death", nc=3, ni=1000, hdata)
```

In the output section, ‘STime’ provides the estimated value of survival time ‘os’ of data for chosen individuals from row number $p$ to $q$. Overall ‘S’ in output provides an overall estimate of survival time ‘os’ in data for all individuals (i.e., for `nrow(data)`). Other variables in the result have the same meaning as explained before.

**Simulation study**

In this section, a simulation study is conducted to assess the performance of the proposed methodology. Here, the simulation study is done by estimating the posterior estimates with varying priors. This procedure has been implemented on the first real TCGA data of liver cancer, discussed in the next section. In the study, a total of 419 individuals were observed, out of which 164 individuals experienced mortality. This corresponds to a censoring rate of 60.9% in the available data.

We have taken the priors for $\beta$ and $\sigma$ as explained in subsection 2.3 and fixed at $\beta \sim \text{norm}(0,0.000001)$ and $\sigma \sim \text{Gamma}(0.001,0.001)$ in models. Now, to verify the accuracy of models, we obtained the posterior estimates for high-dimensional liver cancer data with varying priors as shown in Table 3 (with varying $\beta$ and fixed $\sigma$) and Table 4 (with varying $\sigma$ and fixed $\beta$).

For these Bayesian estimates, we run the code with 3 MCMC chains and 10,000 iterations. From these obtained posterior estimates, we can easily see the consistency of estimates of regression coefficients ($\beta_2$ to $\beta_6$) in all three distributions, log-normal, Weibull, and log-logistic of AFT model, and Figure 1 provides its easy visualization. In the figure, visualization of posterior estimates is provided with one of $\sigma$ and $\beta$ is fixed, and the other varies. The first column of the figure is for Weibull distribution, the second is for log-normal, and the third column is for log-logistic. The first and second rows of the figure have fixed $\sigma$ and varying $\beta$ at SD and mean of its normal distribution, respectively. Similarly, the third and fourth rows have fixed $\beta$ and varying $\sigma$ for parameters $b$ and $a$ of its gamma distribution, respectively. Hence, it can be concluded that these posterior estimates with varying priors on the same distribution give a consistent result, and hence, the AFT model with these three distributions provides adequate performance.

**Analysis on two real gene expression datasets**

In this section, we implement the proposed methodology in a unidirectional manner, starting with selecting biomarkers and proceeding to estimate univariate and multivariable posterior estimates using the Weibull AFT model. The analysis will be conducted on two authentic gene expression datasets related to liver cancer, which have been sourced from the National Cancer Institute’s The Cancer Genome Atlas (TCGA) database. This resource offers comprehensive datasets encompassing clinical information and gene expression profiles derived from primary tumor samples across various types of cancer. The first time-to-event dataset consists of 1881 markers, and the second dataset comprises 18526 markers. The following steps are taken using the developed R package `afthd` to analyze these two time-to-event data.

**Step 1:** First, we make variable (here, covariates are biomarkers) selection using regularization technique and estimates using the conventional approach of AFT model using single function `rglaft()`.
Table 3. Estimation of coefficient of regression in simulated study with fix $\tau \sim \text{Gamma}(0.001,0.001)$ and varying $\beta$, $\beta \sim \text{Normal}(\mu,\sigma)$

<table>
<thead>
<tr>
<th></th>
<th>Weibull</th>
<th>Log-normal</th>
<th>Log-logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean(SD)</td>
<td>Mean(SD)</td>
<td>Mean(SD)</td>
<td>Mean(SD)</td>
</tr>
<tr>
<td>$\mu = 0, \sigma = 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7.4(0.09)</td>
<td>7.44(0.11)</td>
<td>7.38(0.1)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.26(0.1)</td>
<td>-0.27(0.11)</td>
<td>-0.26(0.11)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.19(0.08)</td>
<td>-0.23(0.1)</td>
<td>-0.21(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.22(0.06)</td>
<td>-0.29(0.09)</td>
<td>-0.26(0.08)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.25(0.1)</td>
<td>-0.23(0.11)</td>
<td>-0.25(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.21(0.08)</td>
<td>-0.26(0.09)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2593(15.5)</td>
<td>2621(14.5)</td>
<td>659(16)</td>
</tr>
<tr>
<td>$\mu = 0, \sigma = 10^{-3}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>8.00(0.09)</td>
<td>7.42(0.12)</td>
<td>7.38(0.1)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.26(0.09)</td>
<td>-0.26(0.11)</td>
<td>-0.27(0.12)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.18(0.08)</td>
<td>-0.23(0.1)</td>
<td>-0.22(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.22(0.07)</td>
<td>-0.3(0.09)</td>
<td>-0.26(0.08)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.23(0.09)</td>
<td>-0.23(0.1)</td>
<td>-0.26(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.20(0.07)</td>
<td>-0.27(0.09)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2593(16.4)</td>
<td>2623(15.5)</td>
<td>660(15.8)</td>
</tr>
<tr>
<td>$\mu = 0, \sigma = 10^{-2}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7.39(0.1)</td>
<td>7.45(0.12)</td>
<td>7.38(0.1)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.26(0.1)</td>
<td>-0.27(0.11)</td>
<td>-0.27(0.11)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.18(0.07)</td>
<td>-0.23(0.1)</td>
<td>-0.22(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.22(0.06)</td>
<td>-0.29(0.07)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.24(0.1)</td>
<td>-0.24(0.11)</td>
<td>-0.25(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.21(0.07)</td>
<td>-0.26(0.09)</td>
<td>-0.27(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2594(15.3)</td>
<td>2623(16.2)</td>
<td>661(16)</td>
</tr>
<tr>
<td>$\mu = 0, \sigma = 10^{-1}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7.39(0.1)</td>
<td>7.43(0.11)</td>
<td>7.38(0.11)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.26(0.09)</td>
<td>-0.26(0.11)</td>
<td>-0.27(0.11)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.18(0.07)</td>
<td>-0.22(0.01)</td>
<td>-0.22(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.22(0.06)</td>
<td>-0.29(0.07)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.24(0.1)</td>
<td>-0.24(0.11)</td>
<td>-0.25(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.21(0.07)</td>
<td>-0.26(0.09)</td>
<td>-0.27(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2594(17)</td>
<td>2620(16)</td>
<td>661(16)</td>
</tr>
<tr>
<td>$\mu = 0, \sigma = 10^{-1}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7.37(0.1)</td>
<td>7.45(0.11)</td>
<td>7.38(0.11)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.25(0.09)</td>
<td>-0.26(0.11)</td>
<td>-0.27(0.11)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.18(0.07)</td>
<td>0.23(0.097)</td>
<td>-0.22(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.22(0.06)</td>
<td>-0.3(0.074)</td>
<td>-0.26(0.08)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.24(0.1)</td>
<td>-0.22(0.109)</td>
<td>-0.25(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.21(0.07)</td>
<td>-0.26(0.088)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2592(15.9)</td>
<td>2627(15.8)</td>
<td>660(16.3)</td>
</tr>
<tr>
<td>$\mu = 0, \sigma = 10^{-1}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7.38(0.09)</td>
<td>7.43(0.111)</td>
<td>7.37(0.1)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.27(0.09)</td>
<td>-0.27(0.11)</td>
<td>-0.27(0.11)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.18(0.07)</td>
<td>-0.23(0.097)</td>
<td>-0.22(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.21(0.06)</td>
<td>-0.3(0.074)</td>
<td>-0.26(0.08)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.24(0.1)</td>
<td>-0.23(0.108)</td>
<td>-0.25(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.20(0.07)</td>
<td>-0.26(0.087)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2592(15.8)</td>
<td>2620(15.4)</td>
<td>659(15.8)</td>
</tr>
</tbody>
</table>
Table 4. Estimation of regression coefficient in simulated studies for fixed $\beta \sim \text{Normal}(0,0.000001)$ and varying $\tau$, if $\tau \sim \text{Gamma}(a,b)$

<table>
<thead>
<tr>
<th></th>
<th>Weibull Mean(SD)</th>
<th>Log-normal Mean(SD)</th>
<th>Log-logistic Mean(SD)</th>
<th>Weibull Mean(SD)</th>
<th>Log-normal Mean(SD)</th>
<th>Log-logistic Mean(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$(a = 0.001, b = 0.002)$</td>
<td>$(a = 0.002, b = 0.001)$</td>
<td>$(a = 0.001, b = 0.003)$</td>
<td>$(a = 0.004, b = 0.001)$</td>
<td>$(a = 0.005, b = 0.001)$</td>
<td>$(a = 0.001, b = 0.005)$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>7.38(0.09)</td>
<td>7.44(0.12)</td>
<td>7.38(0.1)</td>
<td>7.38(0.09)</td>
<td>7.45(0.13)</td>
<td>7.39(0.11)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.26(0.1)</td>
<td>-0.27(0.11)</td>
<td>-0.27(0.12)</td>
<td>-0.24(0.1)</td>
<td>-0.27(0.1)</td>
<td>-0.27(0.1)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.18(0.08)</td>
<td>-0.22(0.1)</td>
<td>-0.22(0.09)</td>
<td>-0.19(0.08)</td>
<td>-0.22(0.1)</td>
<td>-0.22(0.09)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.22(0.06)</td>
<td>-0.3(0.08)</td>
<td>-0.26(0.08)</td>
<td>-0.21(0.06)</td>
<td>-0.3(0.09)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.24(0.09)</td>
<td>-0.23(0.1)</td>
<td>-0.25(0.1)</td>
<td>-0.24(0.09)</td>
<td>-0.24(0.1)</td>
<td>-0.25(0.1)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.21(0.07)</td>
<td>-0.27(0.09)</td>
<td>-0.27(0.09)</td>
<td>-0.21(0.07)</td>
<td>-0.26(0.09)</td>
<td>-0.26(0.09)</td>
</tr>
<tr>
<td>deviance</td>
<td>2593(15.9)</td>
<td>2622(16.2)</td>
<td>662(16.5)</td>
<td>2592(15.4)</td>
<td>2624(16)</td>
<td>661(15)</td>
</tr>
</tbody>
</table>

Step 2: Estimates for these selected covariates will be estimated using (i) the Bayesian AFT model for univariate and (ii) the Bayesian AFT model for multivariable.

There are three key points that should be considered in order to streamline procedures. First, the parameters of the parametric AFT model will be estimated using the conventional approach. This estimation is performed for the selected covariates, incorporating regularization techniques.

The desired outcome can be achieved by utilizing a single function $\text{rglaf}$, which is capable of performing all the necessary procedures. Second, the process of estimating the parameters of the Weibull AFT model using the Bayesian approach for selected covariates is accomplished through the utilization of regularization. This procedure is facilitated by a single function called $\text{wbysu}$. The third point for multivariable estimation of the parameters of the AFT model with Bayesian for selected covariates (the first five covariates) using regularization is accomplished by the single function $\text{wbysm}$. Both functions, $\text{wbysu}$ and $\text{wbysm}$, employ a variable selection step utilizing a one-step regularization technique.

In this section, a two-step variable selection technique is employed for feature selection in order to analyze both real datasets. The first step uses the LASSO method, and the second employs the MLE method. It is worth noting that the covariates in the output of the $\text{rglaf}$ are arranged in ascending order of $p$-value. Therefore, following the selection of covariates, we obtain posterior estimates for both univariate and multivariable models using the functions $\text{wbysuni}$ and $\text{wbysm}$, respectively. In the package $\text{afthd}$, there exists an additional benefit pertaining to the functionality of univariate estimation. It provides estimates for all covariates at a glance and does not need to run the code for all the covariates individually.
In the datasets pertaining to liver cancer, there are a total of 1881 and 18526 covariates, respectively. In the initial stage, a covariate selection is performed to identify biomarkers. In the process of variable selection, it is feasible to include more than five covariates. It is widely recognized that multivariable estimates involving a high number of covariates are insufficient to explain the parameters effectively. Therefore, in order to determine the most significant five covariates, we utilized the two-step feature selection technique to narrow down the list of significant covariates. Subsequently, we employ a set of selected covariates, including survival time and event status, within various functions to achieve a favorable outcome. The flowchart in Figure 2 provides a concise and visually accessible representation of the concept being discussed.

**Example 1: Liver Cancer Data I**

The first TCGA data on liver cancer consists of 1881 markers (gene expressions as covariates), events of interest as death, and the survival time as Overall Survival (OS). The dataset comprises 419 individuals, of which 164 experienced death during follow-up, and the remaining are censored. We analyze this data, referred to as *dt1* using the AFT model. We are first estimating with the conventional approach of the parametric AFT model with the variable selection regularization technique LASSO. It is obtained in Table 5: Part 1 by a specified command within the table. In this context, the covariates that exhibit a strong association with the given response variable (event) are identified. These covariates are shown in the first column; here, the uppermost covariates are highly significant. Different estimates of the regression coefficient for selected covariates are shown in other columns. If the estimated coefficients of the selected covariates (biomarkers) in column 2 are negative, it shows that the individual’s survival duration decreases as the value of the continuous-valued covariate increases, i.e., the higher the values of covariates, the greater the risk of the event’s occurrence. If the sign of the estimate is positive, it indicates that the increment in the covariate’s value increases the individual’s survival rate, i.e., decreases the death (event) risk.

**Table 5. Estimates of AFT model for 1st dataset**

<table>
<thead>
<tr>
<th>Part 1: Estimates of AFT model using conventional approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>head(rglaft(4, 1884, STime=&quot;os&quot;, Event=&quot;death&quot;, alpha=1, data=dt1))</td>
</tr>
<tr>
<td>Coef</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>hsa.mir.149</td>
</tr>
<tr>
<td>hsa.mir.7.3</td>
</tr>
<tr>
<td>hsa.mir.4792</td>
</tr>
<tr>
<td>hsa.mir.212</td>
</tr>
<tr>
<td>hsa.mir.4771.2</td>
</tr>
<tr>
<td>hsa.mir.6728</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 2: Posterior univariate estimate using Weibull AFT model</th>
</tr>
</thead>
<tbody>
<tr>
<td>wbysuni(m=3, n=7, STime = &quot;os&quot;, Event = &quot;death&quot;, 3, 10000, data=d22)</td>
</tr>
<tr>
<td>Coef</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>hsa.mir.149</td>
</tr>
<tr>
<td>hsa.mir.7.3</td>
</tr>
<tr>
<td>hsa.mir.4792</td>
</tr>
<tr>
<td>hsa.mir.212</td>
</tr>
<tr>
<td>hsa.mir.4771.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 3: Posterior multivariable estimate using Weibull AFT model</th>
</tr>
</thead>
<tbody>
<tr>
<td>wbysmv(m=3, n=7, STime = &quot;os&quot;, Event = &quot;death&quot;, 3, 10000, data=d22)</td>
</tr>
<tr>
<td>Coef</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>hsa.mir.149</td>
</tr>
<tr>
<td>hsa.mir.7.3</td>
</tr>
<tr>
<td>hsa.mir.4792</td>
</tr>
<tr>
<td>hsa.mir.212</td>
</tr>
<tr>
<td>hsa.mir.4771.2</td>
</tr>
</tbody>
</table>

The Bayesian univariate estimates for selected covariates in the first dataset, using the Weibull AFT model, are
presented in Part 2 of Table 5. These estimates were obtained by considering the biomarkers selected in Part 1 of Table 5 and assigning them to the dataset named d22. In this particular instance, we have selected five covariates for estimation purposes. The obtained estimates of the regression coefficient for all five covariates are negative. It indicates that as the values of the selected gene expressions increase, the survival of the individual decreases, i.e., the number of certain deaths increases. The multivariable posterior estimates of the Weibull AFT model for the selected covariates employing a two-step feature selection procedure (using Table 5: Part 1) are computed using the wbysmve function.

We got the five most significant and efficient liver cancer biomarkers (using two-step variable selection) are hsa.mir.149, hsa.mir.7.3, hsa.mir.4792, hsa.mir.212 and hsa.mir.4771.2, which are found using the first TCGA data of liver cancer. Posterior multivariable Estimates for these biomarkers are provided in Table 5: Part 3. These covariates have a negative estimate (column named, ‘coef’) similar to univariate estimates (Table 5: Part 2) with some variations in values. Moreover, the interpretation of this data can be concluded as, if any patient with liver cancer has large values for these markers hsa.mir.149, hsa.mir.7.3, hsa.mir.4792, hsa.mir.212 and hsa.mir.4771.2, it means they have a low chance of survival and a higher chance of reaching death. Different diagnostic plots, such as density, trace, and autocorrelation plots for each parameter, are also obtained for visualized inspection in this posterior estimate process. The trace plot explains the convergence of MCMC chains for each parameter. Here, all three different colors represent different chains (since we chose three number of chains). In Figure 3, different diagnostics for each parameter are shown, obtained using the multivariable AFT model with Weibull distribution for first liver cancer data (Estimates are in Table 5: part 3).

**Example 2: Liver Cancer Data II**

There are 18526 biomarkers (as covariates) in the second TCGA data of liver cancer of sample size 216. In this time-to-event data, the event is specified as death and survival time as OS. Out of 216 individuals, 66 reached the event during the follow-up study, and the remaining were censored. Analysis of this data (named as dt2) is done with the AFT model. First analysis and estimation with the conventional approach of parametric AFT model with variable selection LASSO for data with 18526 covariates are obtained in Table 6: Part 1.

The highly efficient and most significant covariates (biomarkers) selected from the second liver cancer dataset are gene expressions ILMN_1794643, ILMN_1699496, ILMN_1795183, ILMN_1693598, and ILMN_2070052. These covariates were chosen using a two-step variable selection approach. Let ‘dd2’ be the data that contains these variables and demographic variables. Now its univariate (using wbysuni()) and multivariable (using wbysmve()) posterior estimates of the Weibull AFT model are shown in Table 6: Part 2 and Table 6: Part 3 respectively. In all three parts of Table 6, We can clearly see that the regression coefficient (i.e., beta) is positive for covariate ILMN_1794643 and negative for other covariates ILMN_1699496, ILMN_1795183, ILMN_1693598, and ILMN_2070052. Hence, we conclude that as the value of the covariate ILMN_1794643 increases, the survival of liver cancer patients increases, i.e., death risk decreases. Patients with larger values of other covariates ILMN_1699496, ILMN_1795183, ILMN_1693598, and ILMN_2070052 have a lower chance of survival, meaning a higher risk of death. These estimates facilitate the early prediction of patients’ disease conditions. Therefore, diagnosis can be conducted accordingly. Different diagnostics plots (density plot, trace plot, autocorrelation plot) for each parameter of the multivariable AFT model with Weibull distribution have been obtained and depicted in Figure 4 for the second liver cancer dataset (estimates obtained in Table 6: part 3). We can easily conclude that MCMC chains converge for each parameter at 10000 iterations using trace plots obtained in Figure 4.

**Summary and discussion**

This article describes our developed R package *afthd*. It provides user-friendly implementations for the AFT model with various specifications like analyzing the survival data with missing values, competing risk, selection of appropriate model, and data with augmented entries. There was a gap in analyzing the high-dimensional data with the Bayesian AFT model. The proposed method provides a combined technique for selecting covariates and applying the AFT model (with Weibull, log-normal, and log-logistic distribution) with the Bayesian setting, in which inference is based on MCMC simulation. This formulation is for univariate and multivariable analysis in high-dimensional data. This technique allows for the identification of the set of imperative regression parameters, their credible intervals, and estimated survival time. This package *afthd* is based on a parametric AFT model for Bayesian and conventional interface on high-dimensional settings for multivariable and univariate estimates. In univariate estimations, a good thing can be seen as estimates (output) for all chosen covariates come at a glance. Some functions of ‘afthd’ offer estimates of parameters with Bayesian for different distributions in the
parametric AFT model, and some functions provide estimates for the conventional approach of the AFT model with smooth time functions associated with variable selection using regularization. We have given step-by-step guidelines to use the functions on a high-dimensional dataset (head and neck cancer data) available in this `afthd` package to fit all possible models and on two real high-dimensional data (two liver cancer sides). All functions of this package are capable of working with missing covariates by imputation. In this package, two functions examine the data with competing risk models, and also, there is a function that supports the augmented data and provides multivariable analysis. A short time is required to fit the AFT models. This `afthd` package uses `R2jags`, `rstpm2`, `mcmcplots` and `glmnet` R packages to implement its functions. This `afthd` package allows users to make changes in data to monitor with minimal coding effort when there are missing values in covariates, and it is easy to use.

Acknowledgements
The authors are deeply indebted to the editor-in-chief and learned reviewers for their valuable comments, which improved the original manuscript’s contents and presentation. Vishwakarma is thankful to the Science and Engineering Research Board, Government of India, for providing research Grant No. SIR/2022/000769 to carry out this research work.

Financial disclosure
Not Applicable.

Conflict of Interest:
The authors declare that there is no conflict of interest regarding the publication of this manuscript.
Ethics declarations

Conflict of interest
The author declares that they have no known competing financial interests.

Ethical approval
Ethical and professional standards have been met. No animals or humans are used in this study.

Informed consent
This study did not involve human participants. Therefore, informed consent was not applicable.

References


Figure 1. Consistency of posterior estimates for regression coefficient using the AFT models with different values of priors.
High-dimensional gene expression dataset, number of genes = 1, 2, ..., i, ..., n

- Variable selection using glmnet
  - n1 covariates selected at first stage; n1 < n
  - Variable selection using AFT model
    - n2 covariates selected at second stage; n2 < n1

Analysis after variable selection

- Univariate posterior estimates for all selected covariates at 2nd stage
  - Log-normal AFT model
  - Log-logistic AFT model
  - Weibull AFT model

- Multivariate posterior estimates for 5 most significant selected covariates at 2nd stage
  - Log-normal AFT model
  - Log-logistic AFT model
  - Weibull AFT model

Figure 2. A flowchart for the procedure of the Bayesian AFT model in the code.
Figure 3. MCMC plot for multivariable AFT model with Weibull distribution for first liver cancer data (Estimates are in Table 5: part 3).
Figure 4. MCMC plot for multivariable AFT model with Weibull distribution for second liver cancer data (Estimates are in Table 6: part 3).