Comparison of Shared Gamma Frailty Models for Acute Leukemia Data.

David D. Hanagal and Richa Sharma

Technical Report No. 2012/1

Department of Statistics, University of Pune, Pune-411007, India.
Email: david_hanagal@yahoo.co.in; richa_stat@rediffmail.com
Comparison of Shared Gamma Frailty Models for Acute Leukemia Data.

David D. Hanagal and Richa Sharma

Department of Statistics, University of Pune, Pune-411007, India.
Email: david_hanagal@yahoo.co.in; richa_stat@rediffmail.com

Abstract

There are certain assumptions about the baseline distribution and distribution of frailty. Mostly Weibull distribution is considered for baseline hazard but we consider some different baseline distributions. Mostly assumption of gamma distribution is considered for frailty distribution. In this paper also, we consider gamma distribution as frailty distribution and two different baseline distributions namely, Gompertz and log-logistic distribution. With these two baseline distributions, we propose two different gamma shared frailty models. We fit these models to a real life bivariate survival data set of leukemia remission times (Freireich et al., 1963). Analysis is performed using Markov Chain Monte Carlo methods. Model comparison is made using Bayesian model selection criterion and a better model is suggested for the data.

Key words: Bayesian estimation, Censored sample, Gamma distribution, Gibbs sampling, Gompertz distribution, Log-logistic distribution, Markov Chain Monte Carlo (MCMC), Model selection criterion, Shared frailty, Survival times.

1 Introduction

A popular approach for analyzing survival data makes use of the Cox (1972) proportional-hazards model (see also Cox and Oakes, 1984). That model formulates the survival times when the subjects are independent. If some or all the subjects are related to each other, for example, if they come from the same family (related genetically), or if they share some
unobserved covariates (related by the same environmental exposure) or if the data may come from multiple recurrence times of a disease for the same patient (related by the same exposure to the unobserved physiological conditions of that patient). In these cases the survival times should not be assumed to be independent. To analyze such data, it is necessary to account for within-subject dependency in the multiple event times.

Modeling dependence in multivariate survival data has received considerable attention in recent literature. A key development in modeling such data is to consider the frailty models. The idea builds upon a familiar repeated measures trick. The event times are conditionally independent given the frailty, an individual random effect, see e.g., Clayton (1978), Oakes (1982) and Clayton and Cuzick (1985). These models formulate the variability of life times, coming from two distinct sources. The first source is natural variability and it is explained by the hazard function and the second is variability common to individuals of the same group or variability common to several events of an individual and it is explained by the frailty.

Models fitted to these survival data may involve parametric or non-parametric forms for the hazard function. This depends on whether the hazard function is defined (up to a small number of unknown parameters) as that of a known distribution, or whether it is completely undefined. In this paper, we shall be concerned only with parametric forms which fit well to acute leukemia data. Parametric survival models are regression models in which the distribution of the response is chosen to be consistent with what one would see if the response is time-to-failure.

In particular, the distribution of the response should have positive support. Examples of such distributions are the exponential, Weibull, log-normal, log-logistic, Gompertz, and the generalized gamma, among others. Here, we consider Gompertz distribution and log-logistic distribution as baseline distributions for frailty model. These two distributions fit to the remission times of the patients in acute leukemia data. We will choose a better baseline distribution between these two distributions for shared gamma frailty model. Both distributions can be used in different situations which we will study in detail in Section 4.

There are two important approaches in the field of multivariate data, the shared frailty model and the correlated frailty model. In a shared frailty model, the frailty is common to the individuals in the group, and is thus responsible for creating dependence. In the following, we will restrict our considerations to the bivariate case. Extensions to higher-dimensional
models are straightforward in the shared frailty approach. We are interested mainly on paired data with common shared frailty. The key assumption is that the dependence between two individual life time variables $T_1$ and $T_2$ is caused by the frailty representing unobserved common risk factors and conditional on frailty $T_1$ and $T_2$ are independent. Because the frailty is not observed, it is assumed to follow some distribution, typically a gamma distribution.

In this paper, we consider shared frailty model with gamma distribution as frailty distribution and Gompertz and log-logistic distribution as baseline distributions. The dependence in $T_1$ and $T_2$ is induced by gamma distributed frailty variable. After integrating out frailty, $T_1$ and $T_2$ have a bivariate distribution. Here, we are using Markov Chain Monte Carlo (MCMC) technique to estimate parameters involved in these two models. We apply our models and estimation procedure to bivariate survival data set of leukemia remission times first given by Freireich et al. (1963) and compare these two models using Bayesian comparison techniques such as AIC, AICc, BIC, DIC, and Bayes’ factor etc.

The remaining paper is organized as follows: In Section 2, we introduce the notion of shared frailty model with Laplace transformation followed by the Section 3 in which we give the introduction of gamma distribution with unconditional bivariate survival function evaluated at the cumulative baseline hazard. We introduce the Gompertz and log-logistic baseline model in Section 4. In this Section, we propose two gamma shared frailty models with two baseline distributions. The joint survival function of proposed distribution after integrating out frailty is also derived in this Section. Likelihood function with joint posterior density function of the parameters given the failure times is presented in Section 5. In Section 6, we discuss the issues of model adequacy and model choice criterion. Section 7 illustrates the methodology with acute leukemia data set. Finally, the paper ends with a discussion of our findings in Section 8.

2 General Shared Frailty Model

The shared gamma frailty model was suggested by Clayton (1978) for the analysis of the correlation between clustered survival times in genetic epidemiology. An advantage is that without covariates its mathematical properties are convenient for estimation (see Oakes, 1982, 1986). By measuring some potentially important covariates, we can examine the
influence of the covariates, and we can examine whether they explain the dependence, that is, whether the frailty has no effect (or more correctly, no variation), when the covariate is included in the model. The regression model is derived conditionally on the shared frailty \( U \).

It is assumed that there is independence between the observations from different clusters. If the variation of the frailty variable is zero, this implies independence between event times in the clusters; otherwise, there is positive dependence between event times within a cluster. The shared frailty model abounds in the literature on frailty models and was extensively studied in the books by Hougaard (2000), Therneau and Grambsch (2000). A more detailed presentation of shared frailty models can be found in the books by Duchateau and Janssen (2008), Wienke (2011) and Hanagal (2011).

Suppose \( n \) individuals are observed for the study and let a bivariate random variable \((T_{i1}, T_{i2})\) be the first and second survival times of \( i^{th} \) individual \((i = 1, 2, 3, \ldots, n)\). Also suppose that there are \( p \) observed covariates collected in a vector \( X_i = (X_{i1}, \ldots, X_{ip})' \) for \( i^{th} \) individual where \( X_{ik} (k = 1, 2, 3, \ldots, p) \) represents the value of \( k^{th} \) observed covariate for \( i^{th} \) individual. Here we assume that the first and second survival time \( T_1 \) and \( T_2 \) for each cluster share the same value of the covariates. Assuming that the frailties are acting multiplicatively on the baseline hazard function and both the survival times of individuals \( T_1 \) and \( T_2 \) are conditionally independent for given frailty \( U_i = u_i \). The conditional hazard model for \( i^{th} \) cluster at \( j^{th} \) survival time \( t_{ij} > 0 \), for given frailty \( U_i = u_i \) has the form:

\[
    h(t_{ij}|U_i, X_i) = u_i h_0(t_{ij}) \exp(x_i'\beta); \quad i = 1, 2, \ldots, n; \quad j = 1, 2 \tag{2.1}
\]

where \( U_i \) is the unobserved (random) common risk factor shared by all subjects in cluster \( i \), \( h_0(t_{ij}) \) is the common baseline hazard function, \( X_i \) is a vector of observable covariates and \( \beta \) is a vector of unknown regression coefficients.

Here \( \exp(x_i'\beta) \) is the factor that gives the subject specific contribution to the hazard. Model (2.1) is called the shared frailty model because subjects in the same cluster share the same frailty factor. This model induces correlation between survival times of subjects within the same cluster. The value of the frailty \( U_i \) is common to the individuals in the group, and thus it is responsible for creating dependence. This dependence is always positive.
Under the assumption of independence, the conditional survival function in the bivariate case for given frailty $U_i = u_i$ at time $t_{i1} > 0$ and $t_{i2} > 0$ is,

$$S(t_{i1}, t_{i2} | U_i, X_i) = S(t_{i1} | U_i, X_i) S(t_{i2} | U_i, X_i) = \exp\left[-u_i \{H_{01}(t_{i1}) + H_{02}(t_{i2})\} \exp(x'_i \beta)\right]$$ (2.2)

where $H_{0j}(t_{ij})$ is the integrated baseline hazard of $T_{ij}, (i = 1, 2, ..., n; j = 1, 2)$. From this, we immediately derive the bivariate survival function by integrating out $U_i$ having the probability function $f(u_i)$, for $i^{th}$ individual.

$$S(t_{i1}, t_{i2} | X_i) = \int_{U_i} S(t_{i1}, t_{i2} | u_i, x_i) f(u_i) du_i = L_{U_i}\left[(H_{01}(t_{i1}) + H_{02}(t_{i2})) \exp(x'_i \beta)\right]$$ (2.3)

where $L_{U_i}(\cdot)$ is the Laplace transform of the distribution of $U_i$. Thus, the bivariate survivor function is easily expressed by means of the Laplace transform of the frailty distribution, evaluated at the total integrated conditional hazard.

### 3 Gamma Frailty

One important problem in the area of frailty model is the choice of the frailty distribution. We consider frailty distribution as gamma distribution because the gamma distribution fits very well to failure data from a computational and analytical point of view and it has closed form expression for Laplace transform. This model was suggested by Clayton (1978) and Oakes (1982) and hence the model is known as Clayton model or Clayton-Oakes model.

As the gamma variates are positive, it fits the non-negative criterion of frailties with no transformation. Here the cross ratio function (Clayton, 1978) is constant and is independent of the life times. The popularity of the model is due to the fact that the model functions are very easy to derive because of the simplicity of the derivatives of the Laplace transform.

Let a continuous random variable $U$ follows gamma distribution with shape parameter $\alpha$ and scale parameter $\kappa$ then density function and Laplace transform of $U$ is,

$$f(u) = \begin{cases} \\
\frac{\kappa^\alpha u^{\alpha-1} \exp(-\kappa u)}{\Gamma(\alpha)} & ; u > 0, \alpha > 0, \kappa > 0 \\
0 & ; otherwise. 
\end{cases}$$ (3.1)
and \( L_U(s) = E(e^{-us}) = (1 + \frac{s}{\kappa})^{-\alpha} \).

To make the model identifiable, although we consider two parameter gamma distribution, we restrict that expectation of the frailty equals 1 and variance be finite which implies that scale parameter = shape parameter, so that only one parameter needs to be estimated. Thus, the distribution of frailty \( U \) is the one parameter \((\kappa = \alpha = \theta^{-1})\) gamma distribution i.e. \( U \sim \text{Gamma}(\theta^{-1}, \theta^{-1}) \).

Under the restriction, the corresponding density function and Laplace transformation of gamma distribution result in the following simplified form,

\[
 f(u) = \begin{cases} 
 \frac{u^{\theta-1} \exp(-u/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} & ; u > 0, \theta > 0 \\
 0 & ; \text{otherwise.}
\end{cases} 
\]  
(3.2)

and Laplace transform is

\[
 L_U(s) = (1 + s\theta)^{-1/\theta}. 
\]  
(3.3)

Note that there is heterogeneity if \( \theta > 0 \). So the large values of \( \theta \) reflect a greater degree of heterogeneity among groups and a stronger association within groups. Replacing Laplace transformation in equation (3.3) using (2.3), we get the unconditional bivariate survival function for \( i^{th} \) individual at time \( t_{i1} > 0 \) and \( t_{i2} > 0 \) as,

\[
 S_\theta(t_{i1}, t_{i2}) = \left[ 1 + \theta \{ H_{01}(t_{i1}) + H_{02}(t_{i2}) \} \exp(x_i' \beta) \right]^{-1/\theta} 
\]  
(3.4)

where \( H_{01}(t_{i1}) \) and \( H_{02}(t_{i2}) \) are cumulative baseline hazard functions of life time random variables \( T_{i1} \) and \( T_{i2} \) respectively.

According to different assumptions on baseline distributions we get different shared gamma frailty models. In this paper, we have considered two baseline distributions, First we consider Gompertz distribution because Gompertz distribution is one of the most important growth models. It has many applications in, for example, medical, biological and actuarial studies. Other baseline distribution that we consider is log-logistic distribution (LLD). The LLD provides a useful alternative to the Weibull distribution for the parametric modeling of survival data where the hazard rate is non-monotonic.
4 Baseline Distributions

In parametric proportional hazards model we assume a particular parametric function for the baseline hazard $h_0(t)$. Here we consider the hazard rate of Gompertz and log-logistic distribution.

4.1 Gompertz Baseline Model

Gompertz model, used most frequently by medical researchers and biologists in modeling the mortality ratio data, was formulated by Gompertz (1825). Gompertz distribution is a growth model and has been used in relation with tumor development. Ahuja and Nash (1979) showed that Gompertz distribution, with a simple conversion, related to some distributions in the Pearson distributions family. According to Jaheen (2003), Garg et al. (1970) obtained the maximum likelihood estimations of the Gompertz distribution parameters. Osman (1987), used a Gompertz distribution with two parameters, worked on the features of the distribution and offered that it should be used in modeling the lifespan data analyzing the survival ratio in heterogenic masses. It has been widely used, especially in actuarial and biological applications and in demography.

The Gompertz baseline hazard function corresponds to

$$h_0(t) = \lambda \exp(\gamma t)$$  \hspace{1cm} (4.1)

and the cumulative hazard function is

$$H_0(t) = \lambda \gamma^{-1}(\exp(\gamma t) - 1)$$  \hspace{1cm} (4.2)

with $\lambda > 0, \gamma \in R$. For $\gamma = 0$ the baseline hazard (4.1) reduces to the exponential hazard.

The corresponding survival function is

$$S_0(t) = \exp[-\lambda \gamma^{-1}(\exp(\gamma t) - 1)]$$  \hspace{1cm} (4.3)

We note that for $\gamma > 0$, $S_0(t)$ goes to zero for $t \to \infty$. With $\gamma < 0$, $S_0(t)$ goes to $0 < \exp(\lambda \gamma^{-1}) < 1$ for $t \to \infty$. Therefore the event never occurs for a proportion $\exp(\lambda \gamma^{-1})$ of the population. We therefore consider the case $\gamma > 0$. 

8
In this paper, the two-parameter Gompertz distribution is considered. Let us assume that
the independent random variables $T_1$ and $T_2$ have Gompertz distribution with parameters
$\lambda_1, \gamma_1$ and $\lambda_2, \gamma_2$ respectively. In short we say $\text{Gomp}(\lambda_j, \gamma_j), (j=1,2)$.

4.2 Log-logistic Baseline Model

The log-logistic distribution has a fairly flexible functional form, it is one of the parametric
survival time models in which the hazard rate may be decreasing, increasing, as well as
hump-shaped, that is it initially increases and then decreases. If the mortality ratio in a
life analysis slowly decreases after it reaches to a maximum point over a finite period, it is
suitable to use a non-monotonic failure rate distribution model on the survival data. An
example of this is given by Langlands et al. (1979) in a study of the curability of breast
cancer, where peak mortality occurred after about three years.

In cases where one comes across to censored data, using log-logistic distribution is math-
ematically more advantageous than other distributions. According to the study of Gupta
et al. (1999), the log-logistic distribution is proved to be suitable in analyzing survival
data conducted by Cox (1970), Cox and Oakes (1984), Bennet (1983) and O’Quigley and
Struthers (1982). Gupta et al. (1999) used log-logistic distribution in survival analysis on
lung cancer data in their studies. In their research, they estimated the point where the mor-
tality ratio reached maximum level. They estimated the parameters of the distribution using
the maximum likelihood estimate and bootstrap methods and they observed the proximity
of the results. Log-logistic analysis is used as a parametric model in survival analysis for
events whose rate increases initially and decreases later, for example mortality from cancer
following diagnosis or treatment. It has a fairly flexible functional form with two parameters,
denoted by $\log L(\gamma, \lambda)$.

The distribution imposes the following functional forms on the density, survival, hazard
and cumulative hazard function:
probability density function

\[ f_0(t) = \frac{(\frac{\gamma}{\lambda})^{\gamma-1}}{[1+(\frac{t}{\lambda})^{\gamma}]} \quad (\gamma > 0, \lambda > 0) \]  (4.4)

survival function

\[ S_0(t) = \left[ 1 + \left( \frac{t}{\lambda} \right)^{\gamma} \right]^{-1} \]  (4.5)

hazard function

\[ h_0(t) = \frac{(\frac{\gamma}{\lambda})^{\gamma-1}}{1+(\frac{t}{\lambda})^{\gamma}} \]  (4.6)

cumulative hazard function

\[ H_0(t) = \ln \left[ 1 + \left( \frac{t}{\lambda} \right)^{\gamma} \right] \]  (4.7)

The fact that the cumulative distribution function can be written in closed form is particularly useful for analysis of survival data with censoring (Bennett, 1983). The log-logistic distribution is very similar in shape to the log-normal distribution, but is more suitable for use in the analysis of survival data.

In this paper, the two-parameter log-logistic distribution is also considered. Let us assume that the independent random variables \( T_1 \) and \( T_2 \) have log-logistic distribution with parameters \( \gamma_1, \lambda_1 \) and \( \gamma_2, \lambda_2 \) respectively. In short we say \( LogL(\gamma_j, \lambda_j), (j = 1, 2) \).

4.3 Proposed Models

Substituting cumulative hazard function for Gompertz and log-logistic distribution in equation (3.4), we get the unconditional bivariate survival functions as,

\[ S_\theta(t_{i1}, t_{i2}|X_i) = [1 + \theta \exp(x_i^t \beta)\{\lambda_1 \gamma_1^{-1}(\exp(\gamma_1 t_{i1}) - 1) + \lambda_2 \gamma_2^{-1}(\exp(\gamma_2 t_{i2}) - 1)\}]^{-1/\theta} \]  (4.8)

and

\[ S_\theta(t_{i1}, t_{i2} | X_i) = \left[ 1 + \theta \left\{ \ln \left( 1 + \left( \frac{t_{i1}}{\lambda_1} \right)^{\gamma_1} \right) + \ln \left( 1 + \left( \frac{t_{i2}}{\lambda_2} \right)^{\gamma_2} \right) \right\} \exp(x_i^t \beta) \right]^{-1/\theta} \]  (4.9)

Here onwards we call equations (4.8) and (4.9) as Model-I and Model-II respectively.

Once we have unconditional survival function of bivariate random variable \( (T_{i1}, T_{i2}) \) we can obtain likelihood function and estimate the parameters of the model. Here onwards we represent \( S_\theta(t_{i1}, t_{i2}|X_i) \) as \( S_\theta(t_{i1}, t_{i2}) \).
5 Bayesian Estimation of Parameters with Likelihood Specification

Suppose there are $n$ individuals under study whose first and second observed failure times are represented by $(t_{i1}, t_{i2})$. Let $w_i$ be the observed censoring time for $i^{th}$ individual ($i = 1, 2, 3, \ldots, n$) for first and second recurrence times. We consider censoring time ($W$) is univariate random right censoring type for both failure times $T_1$ and $T_2$. We use censoring scheme as given by Hanagal (1992a, 1992b). Also we assume independence between censoring scheme and life times of individuals. One of the following censoring situations can happen for each data point $(t_{i1}, t_{i2})$.

\[
(T_{i1}, T_{i2}) = \begin{cases} 
(t_{i1}, t_{i2}), & \text{if } \max(t_{i1}, t_{i2}) < w_i \\
(t_{i1}, w_i), & \text{if } t_{i1} < w_i < t_{i2} \\
(w_i, t_{i2}), & \text{if } t_{i2} < w_i < t_{i1} \\
(w_i, w_i), & \text{if } w_i < \min(t_{i1}, t_{i2}) 
\end{cases}
\] (5.1)

Let $n_1$, $n_2$, $n_3$ and $n_4$ denote the random number of observations observed to fall in the range $t_{i1} \leq w_i, t_{i2} \leq w_i$; $t_{i1} \leq w_i, t_{i2} > w_i$; $t_{i1} > w_i, t_{i2} \leq w_i$ and $t_{i1} > w_i, t_{i2} > w_i$ respectively. Discarding factors which do not contain any of the parameters, we want to estimate the parameters in the proposed model. Now the contribution of the $j^{th}$ individual in the $i^{th}$ pair of the conditional likelihood of data given the parameters, based on the survival functions (4.8) and (4.9) is given by

\[
L(t_{1i}, t_{2i}) = \left( \prod_{i=1}^{n_1} f_{i1} \right) \left( \prod_{i=1}^{n_2} f_{i2} \right) \left( \prod_{i=1}^{n_3} f_{i3} \right) \left( \prod_{i=1}^{n_4} F_{i} \right)
\] (5.2)
where $\zeta$ is the vector of baseline parameters, frailty parameter and regression coefficients and

$$
\begin{align*}
    f_{i1} &= \frac{\partial^2 S_\theta(t_{i1}, t_{i2})}{\partial t_{i1} \partial t_{i2}} \\
    &= (1 + \theta) h_{01}(t_{i1}) h_{02}(t_{i2}) S_\theta(t_{i1}, t_{i2})^{(1+2\theta)} \exp(2x'_i\beta), \quad \max(t_{i1}, t_{i2}) < w_i \\
    f_{i2} &= -\frac{\partial S_\theta(t_{i1}, w_i)}{\partial t_{i1}} \\
    &= h_{01}(t_{i1}) S_\theta(t_{i1}, w_i)^{(1+\theta)} \exp(x'_i\beta), \quad t_{i1} < w_i < t_{i2} \\
    f_{i3} &= -\frac{\partial S_\theta(w_i, t_{i2})}{\partial t_{i2}} \\
    &= h_{02}(t_{i2}) S_\theta(w_i, t_{i2})^{(1+\theta)} \exp(x'_i\beta), \quad t_{i2} < w_i < t_{i1} \\
    \mathcal{F}_i &= S_\theta(w_i, w_i), \quad w_i < \min(t_{i1}, t_{i2})
\end{align*}
$$

Substituting hazard functions $h_{01}(t_{i1})$, $h_{02}(t_{i2})$ and survival function $S(., .)$ defined in Section 4, for both proposed models we get the likelihood function given by equation (5.2). $f_{i1}$ is the pdf with respect to Lebesgue measure in $R^2$ and $f_{i2}$ and $f_{i3}$ are the pdf with respect to Lebesgue measure in $R^1$ in their respective regions.

In the Bayesian framework, the parameters of the model are viewed as random variables with some distribution known as prior distribution. To apply MCMC methods, we assume that, conditional on explanatory variables and on the entire set of parameters, observations are independent and prior distributions for all parameters are mutually independent. Given the distribution (5.2) and the priors, all full conditional distributions of the parameters can be calculated. These full conditional distributions are used in a Gibbs sampling procedure. The distribution of a parameter can be updated by combining its prior distribution and the likelihood function, called as posterior density of a parameter. So, if $L(Y \mid \alpha)$ is likelihood function and $p(\alpha)$ is prior density of a parameter $\alpha$ then posterior density function of the same parameter $\pi(\alpha \mid Y)$ is given by, $\pi(\alpha \mid Y) \propto L(Y \mid \alpha)p(\alpha)$. In our case the joint
posterior density function of parameters for given failure times is given by,

$$\pi(\lambda_1, \gamma_1, \lambda_2, \gamma_2, \theta, \beta | t_1, t_2) \propto L(t_1, t_2 | \lambda_1, \gamma_1, \lambda_2, \gamma_2, \theta, \beta) g_1(\lambda_1) g_2(\gamma_1) g_3(\lambda_2) g_4(\gamma_2) g_5(\theta) \prod_{k=0}^{p} p_k(\beta_k)$$

(5.3)

where $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)'$, $g_i(.)$ ($i = 1, 2, \ldots, 5$) indicates the prior density function which is gamma distribution with known hyper parameters of corresponding argument for baseline parameters and frailty variance; $p_k(.)$ is prior density function for regression coefficient $\beta_k$ which is normal with known hyper parameters and likelihood function $L(.)$ is given by equation (5.2). Here, we assume that all the parameters are independently distributed. In our case full conditional distributions are not easy to integrate out. So, full conditional distributions are obtained by considering that they are proportional to the joint distribution of the parameters of the model. We have full conditional distribution of the parameter $\lambda_1$ as,

$$\pi_1(\lambda_1 | \gamma_1, \lambda_2, \gamma_2, \theta, \beta) \propto L(\lambda_1, \gamma_1, \lambda_2, \gamma_2, \theta, \beta) \cdot g_1(\lambda_1)$$

$$\pi_1(\lambda_1 | \gamma_1, \lambda_2, \gamma_2, \theta, \beta) \approx L(\lambda_1, \gamma_1, \lambda_2, \gamma_2, \theta, \beta) \cdot g_1(\lambda_1)$$

Similarly full conditional distributions for other parameters can be obtained.

6 Model Comparison

Model comparison and selection are among the most common problems of statistical practice, with numerous procedures for choosing among a set of models proposed in the literature (see, for example, Kadane and Lazar (2001) and Rao and Wu (2001) for recent reviews). Most (but not all) selection methods are defined in terms of an appropriate information criterion, a mechanism that uses data to give each candidate model a certain score; this then leads to a fully ranked list of candidate models, from the best to the worst. Here, we use these criterion to determine the best model in a class of frailty models. Bayesian model comparison is commonly performed by computing posterior model probabilities. It is the most common method of Bayesian model assessment. In order to compare proposed models we use Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and Deviance Information Criterion (DIC).

Akaike (1973) advocated that, given a class of competing models for a data set, one choose
the model that minimizes

\[ AIC = D(\hat{\theta}) + 2p \]  

(6.1)

where \( p \) represents number of parameters of the model. \( D(\hat{\theta}) \) represents an estimate of the deviance evaluated at the posterior mean, \( \hat{\theta} = E(\theta \mid data) \). The deviance is defined by, \( D(\theta) = -2 \cdot \log L(\theta) \), where \( \theta \) is a vector of unknown parameters of the model and \( L(\theta) \) is the likelihood function of the model.

AICc is AIC with a correction for finite sample sizes:

\[ AICc = AIC + \frac{2p(p + 1)}{n - p - 1} \]  

(6.2)

where \( n \) represents number of sample points.

The AIC penalizes the number of parameters less strongly than the Bayesian information criterion (BIC), which was independently developed by Schwarz (1978). Shibata (1976) and Katz (1981) also have shown that the AIC tends to overestimate the number of parameters needed, even asymptotically. The Schwarz criterion indicates that the model with the highest posterior probability is the one that minimizes

\[ BIC = D(\hat{\theta}) + p \cdot \log(n) \]  

(6.3)

The major benefit of the BIC approximation is that it includes the BIC penalty for the number of parameters being estimated. The model with the smallest BIC value is chosen as the best model.

DIC, a generalization of AIC, is introduced by Spiegelhalter et al.(2002) and is defined as;

\[ DIC = D(\hat{\theta}) + 2 \cdot p_D \]  

(6.4)

where \( p_D \) is the difference between the posterior mean of the deviance and the deviance of the posterior mean of parameters of interest, that is, \( p_D = \overline{D} - D(\hat{\theta}) \), where \( \overline{D} = E(D(\theta) \mid data) \). Models with smaller values of the DIC are preferred.

The Bayes factor \( B_{jk} \) for a model \( M_j \) against \( M_k \) for given data \( D = (t_{i1}; t_{i2}); i = 1, 2, \ldots, n \) is

\[ B_{jk} = \frac{P(D|M_j)}{P(D|M_k)} \]  

(6.5)
where

\[
P(D|M_k) = \int P(D|M_k, \theta_k) \pi(\theta_k|M_k) d\theta_k
\]  \hspace{1cm} (6.6)

where \(\theta_k\) is the parameter vector under model \(M_k\); \(\pi(\theta|M_k)\) is prior density. Raftery (1994), following Jeffreys (1961), proposes the rules of thumb for interpreting twice the logarithm of the Bayes factor. For two models of substantive interest, \(M_j\) and \(M_k\), twice the log of the Bayes factor is approximately equal to the difference in their BIC approximations.

7 Analysis of Acute Leukemia Data

We illustrate the proposed model with one well-known example. The proposed methods are applied to the data set of acute leukemia remission times of patients, given by Freireich et al. (1963). The data are reported in Hougaard (2000). We examine the effect of a clinical trial of a drug 6-mercaptopurine (6-MP) versus a placebo in 42 children with acute leukemia.

Here we demonstrate the method using the well-known leukemia data, consisted of 21 pairs matched of leukemia patients, analyzed by Cox (1972), Hougaard, (2000) (subsection 1.5.4), Ibrahim et al., (2001) (example 3.4), Spiegelhalter et al., (2004), among others. The leukemia remission times data, first given by Freireich et al. (1963), have been cited many times, without reference to their origin from a paired study.

The random variable of interest consists of remission times (in weeks) of the patients assigned to treatment with a 6-MP drug or a placebo during remission maintenance therapy. After having been judged to be in a state of partial or complete remission for the primary treatment with prednisone, a patient was paired with a second patient in the same state. One randomly chosen patient in each pair received the maintenance treatment 6-MP and the other a placebo. It was assumed that deaths at a given time always preceded censoring at the same time, and other ties were broken by randomization. Success (failure) was defined to occur in the \(i^{th}\) pair if the time from remission to relapse or censoring for the patient on 6-MP (placebo) exceeded the time to relapse for the patient on placebo (6-MP). The trial was stopped once the number of successes or failures had reached significance. These data have been used in many articles, but in most of them neglecting the pairing. Out of 21 patients in treatment group, 9 failed during the study period and 12 were censored. In contrast, none
of the data are censored in placebo group; that is, all 21 patients in the placebo group went out of remission during the study period. The data set contains a single covariate $$x$$ with value 0 or 1 indicating remission status (0=partial, 1=complete).

First we check goodness-of-fit of the data for both Gompertz and log-logistic baseline distributions and then apply the Bayesian estimation procedure. In the analysis of this study, we have used the R package. To check goodness-of-fit of data set, firstly we consider graphical procedure to check appropriateness of the model. To assess model graphically, we plot the graph of \( \ln[1 - \{\gamma \ln(S(t))/\lambda\}] \) versus \( t \) for Model-I and \( \ln[(1 - S(t))/S(t)] \) versus \( \ln(t) \) for Model-II. If the resulted plot is roughly a straight line then we can say that the underline baseline model is appropriate. The graph of \( \ln[1 - \{\gamma \ln(S(t))/\lambda\}] \) versus \( t \) for Gompertz baseline distribution for placebo and 6-MP treatment using R-program are shown in the Figure 1(a) and 1(b) respectively and the graph of \( \ln[(1 - S(t))/S(t)] \) versus \( \ln(t) \) for log-logistic baseline distribution for placebo and 6-MP treatment are shown in the Figure 3(a) and 3(b) respectively. From the figure, we can observe that many of the points are nearly on the straight line.

We have considered two statistical tests also viz. Kolmogorov-Smirnov (K-S) test and Hollander & Proschan’s (HP) test (proposed by Hollander and Proschan (1979)) for testing composite goodness-of-fit hypotheses to the Model-I and Model-II. Test procedure is developed for testing goodness-of-fit with data subjected to random right censoring. Here, we assume that survival times of individuals are conditionally independent so we apply K-S test and HP test to the survival times of both the individuals separately. The p-values of K-S and HP test for both baseline distributions are presented in Table 1 which are very large approximately close to 1. Although the p-value of HP test for T6 (time to relapse for 6-MP patients) is not enough large as compared to the same for TP (time to relapse for placebo patients) for Model-I and Model-II but that may be due to censoring because there is no censored observation in placebo group. Thus, from different goodness-of-fit graphs (see Figures 1(a), 1(b) and 3(a), 3(b)) and p-values of K-S and HP test (see Table 1) we can say that there is no statistical evidence to reject the hypothesis that data are from Gompertz distribution and log-logistic distribution.

Figure 2(a), 2(b) and 4(a), 4(b) which plots the Kaplan-Meier estimates and hypothesized theoretical distribution \( S_0(t) \) for Model-I and Model-II respectively, demonstrates a close
agreement between the two i.e. non-parametric and parametric survival curves. The sample mean of the relapse time for the 6-MP drug treated patients is 17.095 weeks while the same for the placebo treated patient is 8.667 weeks. This may indicate a possible 6-MP drug effect. It clearly conveys the effectiveness of the drug in maintaining remission.

In the analysis, we have used the R program. Given the model assumptions, this program performs the Gibbs sampler by simulating from the full conditional distributions. We run two parallel chains for both models using two sets of prior distributions with the different starting points using Metropolis-Hastings algorithm within Gibbs sampler based on normal transition kernels. We implemented 95,000 iterations of the algorithm and described the first 3,500 and 2,000 iterations as a burn-in for Model-I and Model-II respectively. We have taken the independent prior as $\lambda_i \sim \Gamma(10^{-4}, 10^{-4})$, $\gamma_i \sim \Gamma(10^{-4}, 10^{-4})$, $\beta \sim N(0, 10^5)$ and $\theta \sim Gamma(10^{-4}, 10^{-4})$ for Model-I and $\lambda_i \sim \Gamma(1, 10^{-4})$, $\gamma_i \sim \Gamma(1, 10^{-4})$, $\beta \sim N(0, 10^5)$ and $\theta \sim Gamma(10^{-4}, 10^{-4})$ for Model-II. Here, results for chain I and chain II are similar so we present results for only one chain (i.e. chain I). Trace plots for all the parameters show zigzag pattern which indicates that parameters move and mix more freely. Thus, it seems that the Markov chain has reached the stationary state. Due to lack of space we are not presenting graphs for time series (or trace) plot and sample autocorrelation plot with thinning for the parameters.

Tables 4 and 5 present posterior summary along with Gelman-Rubin convergence statistic values and Geweke test statistic values with corresponding p-values for Model I and II respectively. Monitoring convergence of the chains has been done via the Brooks and Gelman (1998) convergence-diagnostic. Hence, once convergence has been achieved, 91,500 and 93,000 observations, for Model-I and Model-II respectively, are taken from each chain after the burn-in period. On inspection of the Brooks and Gelmans diagnostic, we find the BGR (Brooks and Gelman Ratio) convergent to one, this shows that the convergence for the coefficient of regression $\beta$, the variance of frailty $\theta$ and other parameters has been obtained. Also, the Geweke test statistic values are quite small and corresponding p-values are large enough to say the chains attain stationary distribution.

The autocorrelation of parameters become almost negligible after the defined lag, given in Tables 4 and 5 for both models. There is little difference in the posterior estimates of baseline parameters, presented in Tables 4 and 5 for both the models. The posterior estimate
of $\theta = 0.3536$ for Model-I and $\theta = 0.5231$ for Model-II, shows that there exists significant heterogeneity in population of patients even though each patient share the same value of the covariate. The posterior mean of $\beta$ is 0.3417 with 95\% credible interval (-0.5689, 1.3067) for Model-I. Thus, patients had no significant effect due to a partial or complete remission of their leukemia for Model-I. In other words, we can say that there is no risk of acute leukemia relapse for both partial or complete remission patients, according to the results obtained from Gompertz distribution. But the posterior mean of $\beta$ for model-II is $-1.7867$ with 95\% credible interval (-2.8118, -0.8658), i.e. significant effect. Thus, patients who had a complete remission of their leukemia have significant effect of induced treatment. In other words, we can say that there is lower risk of acute leukemia relapse for complete remission patients, according to the results obtained from log-logistic distribution. Since the results obtained from both the models are contradictory in interpreting the significance of covariate effect. Thus, we conclude the final result, that is if remission status has significant effect or not, on the basis of comparison of both the models.

To compare both models we firstly use AIC, AICc, BIC, DIC and log-likelihood values which are given in Table 2. As it is clear from the Table 2, the difference between AIC, AICc, BIC, DIC and log-likelihood values for Model-I and Model-II is significant, so AIC, AICc, BIC, DIC and log-likelihood values are worthy to take decision between the Model-I and Model-II. According to these criterion, Model-I is preferred. To take decision about better model between Model-I and Model-II, we can use Bayes factor also. From the Table 3, which represents Bayes factor for models we can observe that, between Model-I and Model-II, there is very strong positive evidence against Model-II, so Model-I is better than Model-II. Since model-I is fitting well to the acute leukemia data, we can conclude that patients had no significant effect due to a partial or complete remission of their leukemia and there is no change in the risk of acute leukemia relapse either the patients have partial or complete remission. Thus, Model-I that is gamma frailty regression model with Gompertz as baseline distribution is better model than gamma frailty regression model with log-logistic baseline for modeling acute leukemia data.
8 Conclusions

The present study focuses on parametric models, which implies parametric specification of the baseline hazard and the distribution of the frailty. This paper presents some results for the shared frailty models. In this paper, we have considered two failure times by allowing for potential dependence in the random quantities corresponding to each failure time which is induced by frailty. In the literature, gamma distribution is mostly used as frailty distribution because of its simplicity. So, our aim of this study is to suggest a better shared gamma frailty model to analyze acute leukemia data among two frailty models with Gompertz and log-logistic as baseline distribution. A vast literature on human mortality suggests the use of the Gompertz baseline hazard rate, instead of Weibull, to describe the mortality and also to model the risk of disease. The LLD provides a useful alternative to the Weibull distribution for the parametric modeling of survival data where the hazard rate is non-monotonic. Thus the main objective of this study was to find a better alternative of Weibull distribution.

We have discussed the Bayesian estimation procedure including Gibbs sampling for computing the estimation of the unknown parameters for real data example of Acute Leukemia Study. We have discussed two statistical tests for testing goodness-of-fit to Gompertz distribution and log-logistic distribution. Both tests give the statistical evidence of not rejecting the hypotheses that data follows Gompertz baseline distribution as well as log-logistic baseline distribution.

We have run two parallel chains from different starting points and considered the “burn-in” interval for each chain. We have provided 95,000 iterations to perform the analysis. We have clearly written the steps involved in the iteration procedure. The quality of convergence was checked by Gelman-Rubin statistics (see Brooks and Gelman, 1998). The values of the Gelman-Rubin statistics in this case are quite close to one. Thus, the sample can be considered to have arisen from stationary distribution.

From the posterior estimates of data set we can conclude that treatment from the drug 6-mercaptopurine (6-MP) is more effective than placebo. The type of remission status is significant in reducing the risk of acute leukemia relapse for the patients. Also, there is significant heterogeneity among the patients. The acute leukemia data analysis indicates that the performance of Bayesian estimation method is quite satisfactory.
We have compared the models using Bayesian model comparison criterion such as, log-likelihood, AIC, AICc, BIC, DIC, Bayes factor and prediction model choice criterion. From the value of all these criterion we conclude that the Gompertz distribution provide a suitable choice for the life time model as compared to log-logistic distribution. Thus, through our study we have suggested a better gamma frailty model to fit acute leukemia data.

References


Appendix A: Figures and Summary Tables

Figure 1: Graphs for goodness-of-fit for Acute Leukemia Data for Gompertz distribution

Figure 2: Comparison of Non Parametric Survival Function with Gompertz Survival Function for Acute Leukemia Data
Figure 3: Graphs for goodness-of-fit for Acute Leukemia Data for log-logistic distribution

(a) Placebo treatment

(b) 6-MP drug treatment

Figure 4: Comparison of Non Parametric Survival Function with Log-logistic Survival Function for Acute Leukemia Data

(a) Placebo treatment

(b) 6-MP drug treatment
Table 1: p-values for Kolmogorov-Smirnov Test Statistic and Hollander & Proschan’s Test Statistic for Acute Leukemia Data Set.

<table>
<thead>
<tr>
<th>Test</th>
<th>Model − I</th>
<th>Model − II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP T6</td>
<td>TP T6</td>
</tr>
<tr>
<td>K-S</td>
<td>0.995 0.99</td>
<td>0.9625 0.9634</td>
</tr>
<tr>
<td>HP</td>
<td>0.9764 0.1746</td>
<td>1.0000 0.1662</td>
</tr>
</tbody>
</table>

Table 2: AIC, BIC and DIC values for Models I and II fitted to Acute Leukemia Data Set.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
<th>DIC</th>
<th>log − likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>228.8574</td>
<td>234.8574</td>
<td>235.1246</td>
<td>224.8326</td>
<td>-108.4287</td>
</tr>
<tr>
<td>Model II</td>
<td>325.3373</td>
<td>331.3373</td>
<td>331.6044</td>
<td>321.4197</td>
<td>-156.6686</td>
</tr>
</tbody>
</table>

Table 3: Bayes factor values and decision for Models fitted to Acute Leukemia Data Set.

<table>
<thead>
<tr>
<th>numerator model against denominator model</th>
<th>2log_e(Buv)</th>
<th>range</th>
<th>Evidence against model in denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>I against II</td>
<td>97.19116</td>
<td>≥ 10</td>
<td>Very Strong</td>
</tr>
</tbody>
</table>
Table 4: Posterior Summary for Acute Leukemia Data Set for Model-I.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>BGR stat. value</th>
<th>Geweke test stat. value</th>
<th>ACF</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Credible Interval Lower limit</th>
<th>Credible Interval Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>1.0012</td>
<td>0.0189</td>
<td>0.5076</td>
<td>300</td>
<td>0.0393</td>
<td>0.0154</td>
<td>0.0143</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>1.0516</td>
<td>-0.0114</td>
<td>0.4954</td>
<td>100</td>
<td>0.0138</td>
<td>0.0084</td>
<td>0.0026</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>1.0037</td>
<td>-0.0287</td>
<td>0.4885</td>
<td>250</td>
<td>0.1214</td>
<td>0.0447</td>
<td>0.0440</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>1.0087</td>
<td>0.0005</td>
<td>0.5002</td>
<td>200</td>
<td>0.0383</td>
<td>0.0262</td>
<td>0.0028</td>
</tr>
<tr>
<td>$\theta$</td>
<td>1.0013</td>
<td>-0.0223</td>
<td>0.4911</td>
<td>250</td>
<td>0.3536</td>
<td>0.2793</td>
<td>0.0548</td>
</tr>
<tr>
<td>$\beta$</td>
<td>1.0513</td>
<td>0.0113</td>
<td>0.5045</td>
<td>150</td>
<td>0.3417</td>
<td>0.4772</td>
<td>-0.5689</td>
</tr>
</tbody>
</table>

Table 5: Posterior Summary for Acute Leukemia Data Set for Model-II.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>BGR stat. value</th>
<th>Geweke test stat. value</th>
<th>ACF</th>
<th>Estimate</th>
<th>S.E.</th>
<th>Credible Interval Lower limit</th>
<th>Credible Interval Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>1.0006</td>
<td>0.0153</td>
<td>0.5061</td>
<td>80</td>
<td>0.0593</td>
<td>0.0093</td>
<td>0.0344</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>1.0047</td>
<td>-0.0071</td>
<td>0.4971</td>
<td>25</td>
<td>0.0514</td>
<td>0.0143</td>
<td>0.0189</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>1.0139</td>
<td>0.0016</td>
<td>0.5006</td>
<td>170</td>
<td>1.2241</td>
<td>0.5075</td>
<td>0.5627</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>1.0072</td>
<td>0.0023</td>
<td>0.5009</td>
<td>180</td>
<td>0.5302</td>
<td>0.2472</td>
<td>0.2161</td>
</tr>
<tr>
<td>$\theta$</td>
<td>1.0291</td>
<td>0.0034</td>
<td>0.5013</td>
<td>230</td>
<td>0.5231</td>
<td>0.2864</td>
<td>0.0806</td>
</tr>
<tr>
<td>$\beta$</td>
<td>1.0175</td>
<td>-0.0041</td>
<td>0.4983</td>
<td>160</td>
<td>-1.7867</td>
<td>0.4899</td>
<td>-2.8118</td>
</tr>
</tbody>
</table>