Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós Graduação em Estatística

Semiparametric Generalized Inverse-Gaussian Frailty Models

Luiza Sette Câmara Piancastelli

BELO HORIZONTE
MINAS GERAIS - BRAZIL
JULY 2019

Luiza Sette Câmara Piancastelli

Semiparametric Generalized Inverse-Gaussian Frailty Models

Dissertation presented to the Graduate Program in Statistics of the Institute of Exact Sciences of the Federal University of Minas Gerais (UFMG) as a partial requirement to obtain the Master's degree in Statistics.

Supervisor: Prof. PhD. Wagner Barreto de Souza
Co-Supervisor: Prof. PhD. Vinícius Diniz Mayrink

BELO HORIZONTE
MINAS GERAIS - BRAZIL
JULY 2019

Resumo

Neste trabalho, introduzimos um novo modelo de fragilidade para dados de sobrevivência agrupados usando a distribuição inversa-Gaussiana Generalizada (GIG) para a fragilidade. Assumir essa distribuição implica em um modelo flexível que é matematicamente vantajoso, uma vez que expressões fechadas estão disponíveis para as funções de sobrevivência e densidade incondicionais. As versões paramétrica e semiparamétrica do modelo de fragilidade GIG são apresentadas. Focamos na abordagem semiparamétrica que é baseada na distribuição exponencial por partes. Um algoritmo EM é proposto para estimar os parâmetros sob esta abordagem. A flexibilidade do modelo proposto vem da adoção de uma distribuição de fragilidade com dois parâmetros. Um deles determina a distribuição de fragilidade onde nosso interesse será ajustar os diferentes casos especiais da distribuição GIG, obtidos alterando-se o valor desse parâmetro. Esses casos especiais incluem as distribuições inversa-gaussiana, recíproca inversa-gaussiana, hiperbólica e hiperbólica positiva. Com isso, temos em mãos a flexibilidade de testar diferentes fragilidades, possibilitando acomodar estruturas de correlação distintas que poderiam não ser capturadas pelo ajuste de um único modelo. Apresentamos estudos de simulação sob as abordagens paramétrica e semiparamétrica. No estudo de simulação paramétrico, exploramos a estimação dos parâmetros sob tamanhos de amostra finitos e correta especificação do modelo. A comparação com outros modelos da literatura como os modelos de fragilidade gama e exponencial generalizada é feita sob a abordagem semiparamétrica, onde a fragilidade proposta mostra resultados competitivos sob falta de especificação. Ilustramos a aplicabilidade do modelo de fragilidade GIG através de dois exemplos de ajuste a dados reais. O primeiro consiste em dados obtidos pelo estudo Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [1] onde investigamos o efeito de duas variáveis genéticas no tempo de vida de crianças diagnosticadas com câncer de neuroblastoma. Para ilustrar a aplicação da metodologia proposta a dados de sobrevivência agrupados, incluímos também o ajuste ao conhecido conjunto de dados de cateter renal (kidney catheter). Nos exemplos de aplicação a dados reais comparamos o ajuste do modelo proposto com os dos modelos de fragilidade gama e exponencial generalizada sob as abordagens paramétrica e semiparamétrica. Através do conjunto de dados de câncer de neuroblastoma do estudo TARGET, foi possível mostrar que o modelo de fragilidade gama, sendo a escolha mais popular, sofre com problemas de convergência que os outros modelos não apresentaram. Além disso, neste exemplo, a fragilidade GIG provou ser a mais robusta quanto à especificação da função de risco base.

Palavras-chave: Inversa-Gaussiana Generalizada, semiparamétrico, modelo de fragilidade, algoritmo EM, exponencial por partes.

[1] Agradecemos o National Cancer Institute (Office of Cancer Genomics) por nos conceder permissão para usar os dados "TARGET Neuroblastoma Clinical data" para publicação.

Abstract

In this work we introduce a new frailty model for clustered survival data using the Generalized Inverse-Gaussian (GIG) distribution for the frailty. Assuming this distribution implies in a flexible model that is mathematically advantageous since closed expressions are avaliable for the unconditional survival and density functions. The parametric and semiparametric versions of the GIG frailty model are presented. We focus on the semiparametric approach that is based on the piecewise exponential distribution. An EM algorithm is proposed to estimate the parameters under this approach. The flexibility of the proposed model comes from working with a two-parameter frailty distribution. One of the parameters will determine the frailty distribution and our interest will be in adjusting the different special cases of the GIG distribution obtained by changing the value of this parameter. These include the Inverse-Gaussian, Reciprocal Inverse-Gaussian, Hyperbolic and Positive Hyperbolic distributions. With this, we have in hand the flexibility of testing different frailties, making it possible to accomodate distinct correlation structures that might not be captured by fitting a single model. We present simulation studies under both parametric and semiparametric approaches. In the parametric simulation study, we explore parameter estimation in finite samples sizes under correct model specification. A comparison to other models in the literature such as gamma and generalized exponential frailty models is made under the semiparametric approach where the proposed frailty shows competitive results under misspecification. We illustrate the applicability of the GIG frailty model through two real data examples. The first consists on data obtained from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [2] initiative, where we chose to investigate the effect of two genetic variables on the lifetime of children diagnosed with neuroblastoma cancer. To illustrate the application of the proposed methodology to clustered survival data, we also include the fit to the well known kidney catheter data set. In the real data examples we compared the fit of the proposed model with the fit of the gamma and generalized exponential frailty models under parametric and semiparametric approaches. Through the TARGET Neuroblastoma data set we were able to show that the gamma frailty model, being the most popular choice, suffers with convergence issues that the other models did not present. In addition, in this example, the GIG frailty proved to be the most robust regarding the specification of the baseline hazard function.

Keywords: Generalized Inverse-Gaussian, semiparametric, frailty model, EM algorithm, piecewise constant hazards.

[2]We thank the National Cancer Institute (Office of Cancer Genomics) for granting us permission to use the TARGET Neuroblastoma Clinical data for publication.

ABSTRACT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Ever since the introduction of the proportional hazards model by Cox [1972], models based on the hazard function have dominated the field of survival analysis. These accomodate well censoring and truncation which are key elements in time to event data. In addition, the hazard function has a useful interpretation as it specifies the instantaneous risk of failure that changes over time. However, it is not rare to be in the presence of correlated failure data, that arises, for example, when we have repeated measures of the same individual or even when some common traits such as biological or environmental factors are shared. In these situations, the proportional hazards model is not adequate since it implicitly assumes an homogeneous population. This is equivalent to say that given the observed covariates, the failure times of two different individuals are statistically independent, as mentioned by Wienke [2011]. This is a sensible assumption as, very often, important covariates may not be observed in the study, maybe because they are difficult to measure or even because the researcher did not know its importance in the first place. When ignoring the presence of the correlation structure or non observed risk factors, most often the effects of the covariates will be underestimated. Here we refer again to Wienke [2011] who comments that, typically, we observe an increase in the regression parameters estimates as well as in the standard errors when comparing a proportional hazards model to a frailty model with the same baseline hazard function.

Frailty models are a natural extension of the proportional hazards model that introduce a latent random component called frailty acting multiplicatively in the hazard function of an individual or a group of individuals. The univariate frailty models emerge as a way to account for unobserved sources of heterogeneity. This means that the variability is split in two parts: the fixed and random effects. The first accounts for the observed covariates and the former is a random unobserved component.

The multivariate frailty model, or shared frailty model, was introduced by Clayton [1978] motivated by the analysis of familial tendency in disease incidence. By making individuals in the same group share the frailty variable, positive dependence between those individuals is created. In this case, the random effect can be interpreted as the degree of association within the cluster, where in each risk level the lifetimes are independent.

In both univariate and multivariate contexts, fitting the unobserved risk components is of most im-

portance as it will allow us to properly evaluate the covariate effects. The presence of the frailty can even give different interpretations for the same problem. For example, a hazard function that is very high and then diminishes over time can represent an adaptation phenomenon or could also be explained by the existence of more susceptible individuals, that is, natural selection.

It is commonly assumed some distribution in $\mathbb{R}^+$ for the frailty variable, persuing desired mathematical properties, since the inferential aspects of the frailty models pose additional difficulties in comparison with usual mixed models due to censoring and truncation. The gamma distribution is the most used for this task. This was a model that became very popular because of its mathematical convenience and was explored by several authors such as Vaupel et al. [1979], Oakes [1982], Oakes [1986], Klein [1992] and Yashin et al. [1995].

Ever since, the need to model more complex dependence structures and also the search for more flexibility motivated the introduction of other frailty distributions. For instance, the inverse Gaussian frailty by Hougaard [1984] and the log-normal frailty by McGilchrist & Aisbett [1991]. The former was extended by Ripatti & Palmgren [2000] that proposed a model based on a multivariate log-normal distribution. In this model, estimation relies on a Laplace approximation of the likelihood function. Another relevant work in this area is Hougaard [1986a] that introduced the positive stable frailty. An important aspect of this model is that it is the only frailty distribution that preserves the proportional hazards condition after integrating out the latent component. It was later extended by Hougaard [1986b] with the power variance function distribution. The power variance function distribution is a three-parameter model that includes the gamma, inverse Gaussian and positive stable function as special cases. It is also worth mentioning the compound Poisson model by Aalen [1992] that allows the presence of a portion of individuals who will never experience the event of interest, that is, the model can handle cure fractions.

A growing interest in this area is on proposing semiparametric versions of frailty models. This approach allows us to estimate the regression effects without expliciting the form of the baseline hazard function. A parametric form for this function is often not easy to identify or to test for adequacy. One popular approach for doing this was the one introduced by Klein [1992] to develop a semiparametric version of the gamma frailty model. This approach is based on a modified EM algorithm that involves the Cox proportional hazards model. Other strategies that may be applied to develop semiparametric versions of frailty models are penalized partial likelihood functions introduced by Therneau et al. [2003], piecewise constant hazards with raising number of pieces and splines, for example. For references in piecewise constant hazards see Kim & Proschan [1991] that introduced the estimation of the survival function through the piecewise exponential estimator and Lawless & Zhan [1998] that studied the performance of the regression estimators and frailty coefficients using this method. The works by Liu & Huang [2008] and Feng et al. [2005], for instance, have applied the piecewise constant hazards to the context of frailty models. For a reference in the usage of splines in the estimation of the baseline hazard function we suggest Du & Ma [2010].

The semiparametric gamma frailty model is the most popular in practice. However, it has been

proved that more flexibility might be needed depending on the problem, motivating the introduction of new semiparametric frailty models. That because using the gamma distribution for the frailty has some implications that may not be desirable. For instance, Farrington et al. [2012] show that the relative frailty variance of this model is constant, meaning that the heterogeneity of the population remains constant over time, something that often may not be applicable.

It is worth mentioning Balakrishnan & Peng [2006] that introduced the generalized gamma frailty model pursuing to provide the extra flexibility that the popular gamma frailty model lacks. This is a model that includes an additional parameter and has the log-normal and Weibull frailty models as special cases. However, this generalization comes at the cost of handling intractable integrals in the likelihood function. The authors proposed to use numerical approximation through Monte Carlo simulation or by a quadrature method. Another model that was introduced with the same motivation is the log-skew-normal frailty by Callegaro & Iacobelli [2012]. This is a model that includes a parameter to control the skewness. The authors show that the dependence pattern can assume different shapes under the log-skew-normal frailty just by changing the values of the parameters. This overcomes a limitation of the gamma frailty model. However, the comparison of the proposed approach to those existing in the literature is not well explored showing no information on how the model behaves under misspecification. The same can be said about Wang & Klein [2012] when introducing the semiparametric additive inverse Gaussian frailty model, because the authors also did not explore the behavior of their methodology under misspecification. Instead, the focus of the additive inverse Gaussian frailty model is in allowing the presence of different sources of heterogeneity. With this, the frailty is the sum of two parts, one shared by all group members and an individual specific part, both following the inverse Gaussian distribution.

Another difficulty posed to the semiparametric gamma frailty model is that its version proposed by Klein [1992] is affected by a flat-likelihood issue that causes difficulties in the estimation of the frailty variance as shown by Barreto-Souza & Mayrink [2019]. To attack this problem, the authors introduced the generalized exponential (GE) frailty model that is immune to the flat-likelihood issue. A semiparametric version of the GE frailty model is avaliable and is also based on an EM algorithm. Although the proposed model shows to be competitive with respect to the gamma frailty model, it is most suitable for small clusters. That because the high-order derivatives of the Laplace transform are cumbersome in that case. Hence, there is some difficulty in appling the GE model to situations in which big clusters are handled.

Following these arguments, our goal is to introduce a class of semiparametric frailty models having advantages over existing ones. That because, as discussed, each model has its own advantages and limitations. In our work, we propose a frailty model based on the generalized inverse-Gaussian distribution. This is a two-parameter frailty distribution that contains the inverse-Gaussian as a particular case as well the special cases reciprocal inverse-Gaussin, hyperbolic and positive hyperbolic that will be presented later. Assuming this distribution for the frailty provides flexibility without compromising mathematical tractability, since all the main expressions have closed forms. We present the parametric and semiparametric versions of the proposed model, the latter being based on the piecewise exponential baseline hazard

function.

This work is organized as follows: In Section 2 we define our class of GIG frailty models and obtain basic results. Section 3 introduces the generalized inverse-Gaussian frailty model for clustered survival data and discusses estimation in the parametric case. The semiparametric version of the proposed model is based on an EM algorithm where the expressions of interest and detailed description are also given in Section 3. In Section 4, we present Monte Carlo simulation studies under the parametric and semiparametric approaches. Also in this section, we compare the performance of the GIG frailty model with other models in literature. In Section 5, we apply the developed methodology to real data sets illustrating the usefulness of the proposed model. In the first example we investigate the influence of two genetic variables in the lifetime of children diagnosed with a rare type of cancer called neuroblastoma. In this case, the frailty accounts for unobserved individual risk factors. These data were provided by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and are avaliable online. In addition, we present the application of the GIG frailty model to the well known kidney cathether data set, so that we illustrate the case of clustered data. With these examples, we are able to highlight problems that the popular gamma frailty model suffers and evidence the robustness of the proposed methodology.

CHAPTER 2

MODEL SPECIFICATION AND BASIC RESULTS

In this section our goal is to introduce the GIG frailty model. For that, we will present some results on the class of GIG distributions. The Generalized Inverse-Gaussian distribution with parameters $\lambda \in \mathbb{R}$, $a > 0$ and $b > 0$ has the following density

$$f_{\lambda,a,b}(x) = \frac{(a/b)^{\lambda/2}}{2K_\lambda(\sqrt{ab})} x^{\lambda-1} e^{-(ax+b/x)/2}, \quad x > 0, \tag{2.1}$$

where $K_\lambda(x) = \int_0^\infty u^{\lambda-1} \exp \frac{-t}{2}(u + 1/u)du$ denotes the third kind modified Bessel function with index λ.

Some basic properties of the GIG distribution are presented below.

The Laplace tranform associated to (2.1) is given by

$$L(t) = \frac{K_\lambda(\sqrt{(a+2t)b})}{K_\lambda(\sqrt{ab})} \left(\frac{a}{a+2t}\right)^{\lambda/2}, \quad t > 0. \tag{2.2}$$

From that, we have that the k-th order moments can be found through the following expression

$$E(X^k) = \frac{K_{\lambda+k}(\sqrt{ab})}{K_\lambda(\sqrt{ab})}(a/b)^{-k/2}.$$

Hence, the expected value and variance of the Generalized Inverse-Guassian distribution are given by

$$E(X) = \frac{K_{\lambda+1}(\sqrt{ab})}{K_\lambda(\sqrt{ab})}(a/b)^{-1/2} \tag{2.3}$$

and

$$Var(X) = \frac{K_{\lambda+2}(\sqrt{ab})}{K_\lambda(\sqrt{ab})}(a/b)^{-1} - \frac{K_{\lambda+1}(\sqrt{ab})^2}{K_\lambda(\sqrt{ab})^2}(a/b)^{-1}.$$

As will be presented in sequence, having a closed form for the Laplace transform is an advantage in the construction of frailty models. That because the marginal survival and density functions of the frailty model to be defined will be found through this expression. The fact that the Laplace transform of the gamma distribution has a closed form is one of the reasons for the popularity of the gamma frailty model.

Definition 1. *Assume Z to be an independent and identically distributed random variable following the GIG distribution and consider the time to event of an individual to be denoted by a random varible T where the latent effect Z acts multiplicatively in the baseline hazard function. Conditional on Z, the hazard function is given by $h(t|Z) = Zh_0(t)$. Here $h_0(t)$ is the baseline hazard function and Z is the frailty variable.*

A great motivation for choosing the GIG distribution for the frailty variable comes from the fact that for a fixed λ there are several special cases of this distribution. Hence, by changing λ value we can fit different frailty models to the data. A list of the known distributions obtained by changing λ is given in Table 2.1.

Table 2.1: Known special cases of the GIG distribution for fixed λ

Configuration	**Distribution**
$\lambda = -1/2$	Inverse Gaussian
$\lambda = 1/2$	Reciprocal Inverse Gaussian
$\lambda = 0$	Hyperbolic
$\lambda = 1$	Positive Hyperbolic

In order to avoid non-identificability issues, it is sufficient to configure the parameterization of the GIG frailty model by one of the particular cases, which is going to be the inverse-Gaussian one. This means that when $\lambda = -1/2$, we have that $E(Z) = 1$ and the variance modeled by a single parameter being $Var(Z) = \alpha$. Hereby, our latent random variable, the frailty, follows a GIG(a $= 1/\alpha$, b $= 1/\alpha$, λ) distribution. By doing this, we avoid non-identificability issues and have an appealing interpretation for the parameter α in the inverse Gaussian case as the degree of non observed heterogeneity or association within cluster, as commonly pursued in frailty models. That, however, holds only when $\lambda = -0.5$ and the other particular cases require a transformation of α in order to obtain the frailty variance.

Now, we develop the theoretical results for the GIG frailty model based on the parametrization above mentioned. First, the GIG frailty model without covariates will be presented. Regression structure can easily be incorporated as discussed next. Through known results on frailty models which can be found, for example, in Wienke [2011] and assuming the GIG distribution with the aforementioned parameters, we can obtain the marginal survival function $S(\cdot)$ and density $f(\cdot)$ of T as follows. In line with the Definition 1, we denote by $h_0(t)$ the baseline hazard function and $H_0(t) = \int_0^t h_0(u)du$ the cumulative baseline hazard function.

The marginal survival function is given by

$$S(t) = L(H_0(t)) = \left(\frac{1/\alpha}{1/\alpha + 2H_0(t)}\right)^{\lambda/2} \frac{K_\lambda(\sqrt{\alpha^{-1}(1/\alpha + 2H_0(t))})}{K_\lambda(1/\alpha)}, \quad t > 0. \tag{2.4}$$

The marginal density function can be found using the relation $f(t) = -h_0(t)L'(H_0(t))$. This quantity

depends on the derivative of $K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}+2t)})$ that is given by the expression

$$\frac{\partial}{\partial t}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}+2t)}) = \frac{-1}{2\alpha}\sqrt{\frac{1}{\alpha^{-1}(\alpha^{-1}+2t)}}\left(K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2t)}) + K_{\lambda-1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2t)})\right),$$

with this, we have the following marginal density function

$$\begin{aligned} f(t) &= \frac{h_0(t)(\alpha^{-1})^{\lambda/2}}{K_\lambda(\alpha^{-1})}\left\{\frac{\lambda K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}{(\alpha^{-1}+2H_0(t))^{\frac{\lambda}{2}+1}} + \right. \\ &\left. \frac{\sqrt{\alpha^{-1}}\left(K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))}) + K_{\lambda-1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})\right)}{2((\alpha^{-1}+2H_0(t)))^{\frac{\lambda+1}{2}}}\right\}, \quad t>0. \end{aligned}$$

We can simplify the expression above by applying the following recurrence identity of the Bessel function: $K_v(z) = \frac{z}{2v}(K_{v+1}(z) - K_{v-1}(z))$. With this, we obtain

$$f(t) = \frac{h_0(t)\alpha^{-(\lambda+1)/2}}{(\alpha^{-1}+2H_0(t))^{\frac{\lambda+1}{2}}}\frac{K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}{K_\lambda(\alpha^{-1})}, \quad t>0.$$

Next, we find the frailty distribution among the survivors and among the failures at time t. These results are important to the EM algorithm that will be discussed in Section 3.2. The condional density of Z given $T>t$, that is, the distribution of the frailty among the survivals at time t, can be found using $f(z|T>t) = f(z)S(t|z)/S(t)$ and is given by

$$f(z|T>t) = \frac{z^{\lambda-1}}{2K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}e^{\frac{-1}{2}(z(\alpha^{-1}+2H_0(t))+\frac{1}{z\alpha})}\left(\frac{\alpha^{-1}+2H_0(t)}{\alpha^{-1}}\right)^{\lambda/2}, \quad z>0.$$

Therefore, $Z|T>t \sim \text{GIG}(\alpha^{-1}+2H_0(t), \alpha^{-1}, \lambda)$. As argued by Hougaard [1995], if the frailty distribution is in the natural exponential family, the frailty distribution between the survivals will be in the same family. And according to Jørgensen [1982] the GIG distribution is a full exponential family of order three. The conditional mean of $Z|T>t$ is given by

$$E(Z|T>t) = \frac{K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}{K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}\left(\frac{\alpha^{-1}+2H_0(t)}{\alpha^{-1}}\right)^{-1/2}.$$

Using the expression found for the marginal density function and some fundamental results in survival analysis, we can calculate the distribution of the failures at time t as $f(z|t) = f(t|z)f(z)/f(t)$. After some algebra we get that this density is given by

$$f(z|t) = \frac{z^{\lambda}}{2K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}e^{\frac{-1}{2}(z(\alpha^{-1}+2H_0(t))+\frac{1}{z\alpha})}\left(\frac{\alpha^{-1}+2H_0(t)}{\alpha^{-1}}\right)^{\frac{\lambda+1}{2}}, \quad z>0.$$

The frailty distribution of failures of time t is a $\text{GIG}(\alpha^{-1}+2H_0(t), \alpha^{-1}, \lambda+1)$. The conditional mean in

this case is

$$E(Z|T=t) = \frac{K_{\lambda+2}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}{K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}\left(\frac{\alpha^{-1}+2H_0(t)}{\alpha^{-1}}\right)^{-1/2}.$$

We can define $\delta = 1$ if $T = t$ (0 if $T > 0$) so we have that $f(z|\delta) \sim \text{GIG}(\alpha^{-1}+2H_0(t), \alpha^{-1}, \lambda+\delta)$. Hence, the conditional mean of $Z|\delta$ is given by

$$E(Z|\delta) = \frac{K_{\lambda+1+\delta}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}{K_{\lambda+\delta}(\sqrt{\alpha^{-1}(\alpha^{-1}+2H_0(t))})}\left(\frac{\alpha^{-1}+2H_0(t)}{\alpha^{-1}}\right)^{-1/2}.$$

Next, we discuss another way of characterizing the frailty distribution that was introduced by Farrington et al. [2012], which is the relative frailty variance (hereby RFV). The RFV is a measure of how the heterogeneity of the population evolves over time. This function can be used as a way to compare patterns of dependence among different frailty models and is obtained through the Laplace transform of the frailty distribution. Let $J(s) = \log L(-s)$ where $L(\cdot)$ is the Laplace transform and μ is the expected value of the frailty distribution given in (2.3), then $\text{RFV}(s) = \frac{J''(-s/\mu)}{J'(-s/\mu)^2}$. The expressions required to calculate the RFV for the GIG frailty model are given by

$$\begin{aligned}\frac{\partial J(s)}{\partial s} &= \frac{\partial \log L(-s)}{\partial s} = \frac{\alpha^{-1}}{2(\alpha^{-1}(\alpha^{-1}-2s))^{3/2}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})} \times \\ &\{\alpha^{-1}(\alpha^{-1}-2s)K_{\lambda-1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) \\ &+2\lambda\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) \\ &+\alpha^{-1}(\alpha^{-1}-2s)K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})\}\end{aligned} \tag{2.5}$$

and

$$
\begin{aligned}
\frac{\partial J(s)^2}{\partial s^2} \quad = \quad & \frac{\partial \log L(-s)^2}{\partial s^2} = \frac{1}{4(\alpha^{-1}-2s)^2\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2} \times \\
& \{-(\alpha^{-1}(\alpha^{-1}-2s))^{3/2}K_{\lambda-1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2 + 2\alpha^{-1}(\alpha^{-1}-2s) \times \\
& [K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) - \sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})] \times \\
& K_{\lambda-1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) - 4\alpha^{-1}s\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2 + \\
& 8\lambda\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2 + \\
& 2\alpha^{-2}\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2 + \\
& 2\alpha^{-1}s\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2 - \\
& \alpha^{-2}\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})^2 + \\
& (\alpha^{-1}(\alpha^{-1}-2s))^{3/2}K_{\lambda-2}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) + \\
& 2\alpha^{-2}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) - \\
& 4\alpha^{-1}sK_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})K_{\lambda+1}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) - \\
& 2\alpha^{-1}s\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})K_{\lambda+2}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}) + \\
& \alpha^{-2}\sqrt{\alpha^{-1}(\alpha^{-1}-2s)}K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})K_{\lambda+2}(\sqrt{\alpha^{-1}(\alpha^{-1}-2s)})\}
\end{aligned} \tag{2.6}
$$

We are now able to calculate the RFV of the proposed frailty model. Our interest lies on exploring how the evolution of RFV occurs over time. In multivariate frailty models, that is, in the presence of data with cluster structure, it is a tool to understand if the dependence between the individuals of the same group increases, decreases or remains constant over time. We consider that in practical situations having this type of information about the models can help the researcher to select the model that best suits a particular problem. Let us explore this later, after we have discussed some particular cases of the GIG frailty model. These particular cases are important for our purpose that is fitting the GIG frailty model for fixed λ. The proposed methodology is not restricted to these λ values, but we are considering the four special cases in sequence in the simulation studies and applications.

Example 1. *(Inverse Gaussian frailty model)*
Suppose Z to be a GIG random variable with density function in (2.1). When $\lambda = -1/2$ we have that $K_{-1/2}(z) = \frac{\sqrt{\frac{\pi}{2}}e^{-x}}{\sqrt{z}}$, so Z is an inverse Gaussian (hereafter IG) random variable with density function given as follows

$$
f(z) = \sqrt{\frac{b}{2\pi\alpha}}\exp\{\alpha^{-1}\}z^{-3/2}\exp\left\{-\frac{1}{2\alpha}(z+\frac{1}{z})\right\}, \quad z>0.
$$

Example 2. *(Reciprocal Inverse Gaussian frailty model)*
The Reciprocal Inverse Gaussian (denoted here by RIG) case is obtained by taking $\lambda = 1/2$. Here we have $K_{1/2}(z) = K_{-1/2}(z)$, since the modified Bessel function of the third kind is symmetric. In this case

the density function of Z is given by

$$f(z) = \sqrt{\frac{1}{2\pi\alpha}} \exp\{\alpha^{-1}\} z^{-1/2} \exp\left\{-\frac{1}{2\alpha}(z + \frac{1}{z})\right\}, \quad z > 0.$$

Example 3. *(Hyperbolic frailty model)*
Another special case of the GIG distribution is obtained when $\lambda = 0$ and is called the Hyperbolic distribution. The Hyperbolic Generalized distributions were introduced by Barndorff-Nielsen [1977] and were called Hyperbolic distributions since the graph of the logarithm of the probability function takes the form of a hyperbola. The class of Generalized Hyperbolic distributions includes the Hyperbolic, Scaled Laplace and Positive Hyperbolic distributions by changing parameter specification. The Hyperbolic frailty (that will be denoted by HYP) has the following density function

$$f(z) = \frac{z^{-1}}{2K_0(\sqrt{\alpha^{-1}})} \exp\left\{-\frac{1}{2\alpha}(z + \frac{1}{z})\right\}, \quad z > 0.$$

Example 4. *(Positive Hyperbolic frailty model)*
Our last special case is called the Positive Hyperbolic (for short PHYP) and corresponds to the GIG distribution with $\lambda = 1$. The density function a Positive Hyperbolic random variable is

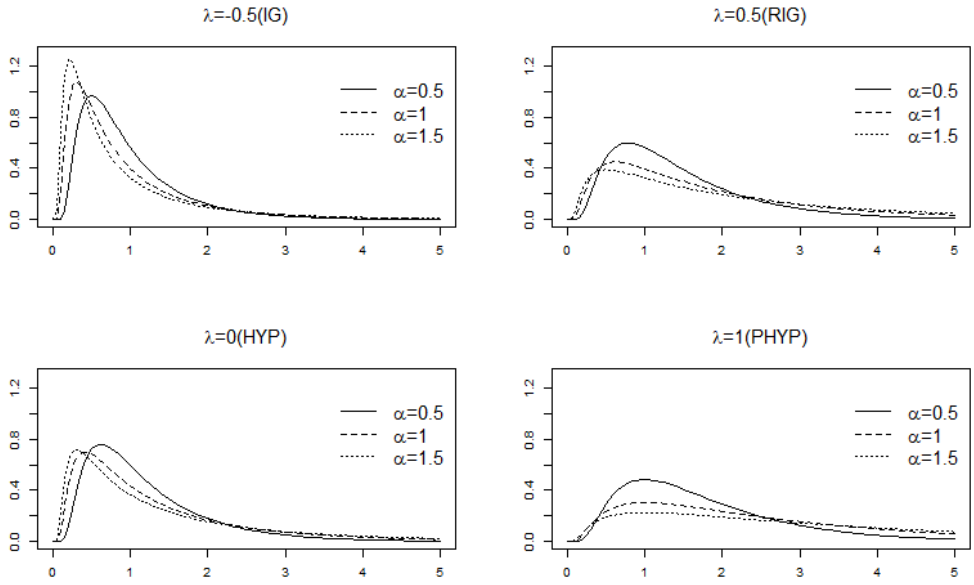
$$f(z) = \frac{1}{2K_1(\sqrt{\alpha^{-1}})} \exp\left\{-\frac{1}{2\alpha}(z + \frac{1}{z})\right\}, \quad z > 0.$$

Now that we have become familiar the known distributions for fixed λ, we are going to explore some aspects of these distributions. We will compare the form of the marginal density and hazard function of T and the RFV evolution for the four values of λ that refer to the examples above.

The densities of the IG, RIG, HYP and PHYP distributions under different values of α can be found in Figure 2.1. Hougaard [1995] states that the tails of the frailty distribution determine the type of dependence within the clusters, in a way that a large right tail leads to strong dependence initially while a large left tail leads to strong late dependence. In such a matter, it is convenient that the distributions that can be adjusted through the GIG frailty model provide alternatives with stronger right tails (RIG and PHYP) as well as stronger left tails (IG and HYP). From Figure 2.1 we also conclude, as expected, that the higher the value of α, the more dispersed becomes the distribution.

In Figure 2.2 we investigate how the choice of λ affects the form of the marginal density function and marginal hazard function. For that we assume the baseline hazard to follow a Weibull distribution (that is $h_0(t) = \sigma\gamma t^{\gamma-1}$ and $H_0(t) = \sigma t^{\gamma}$) with parameters $\sigma = 0.25$ and $\gamma = 2$. In the right panel we can see that positive values of λ imply greater risk in the initial times than negative values of λ as the curves, from top to bottom, correspond to the values 1, 0.5, 0 and -0.5. As for the left panel, we can observe that the marginal density function under negative values of λ has stronger right tails than under positive values of this parameter.

Figure 2.1: Density function of IG, RIG, HYP and PHYP distributions under some values of α where line type indicates different specifications.



As mentioned previously, the RFV is a way of comparing dependece patterns within frailty distributions and we will do that using the expressions in Equations (2.5) and (2.6). Through the graph of the RFV(s) versus s, we can obtain information on how the heterogeneity or dependence evolves over time in each frailty distribution. This means that exploring this relationship may help giving the researcher a better understanding on what the choice of a particular frailty distribution entails for the fit.

In Figure 2.3, we present the evolution of the RFV(s) function for the IG, RIG, HYP and PHYP frailties. For the four distributions to be compared we find for each of them the root of the equation $RFV(0) - 0.7 = 0$. That is, the value of α for which RFV(0) is equal to 0.7. In this way we can better compare how their evolution differs since they have the same starting point.

The results of Figure 2.3 evidence similarity within the GIG family regarding the RVF evolution. All special cases have shown decreasing RFV over time. This means that in GIG frailty models individuals who survive tend to be less heterogeneous (or less dependent) over time. What differentiates between the distributions under comparison is the speed in which this happens. This can be observed in the curvature of the graph of $RFV(s)$ versus s. The more pronounced curvature is due to the inverse Gaussian model, consequently in this frailty distribution, the decrease in heterogeneity (or dependence) occurs more quickly. In the PHYP distribution, we notice a much more gradual decrease. The HYP and RIG models have intermediate behavior regarding the evolution of the relative frailty variance.

Having introduced the GIG frailty model and its particular cases, in the next section we extend the model presented to the multivariate case, that is, the presence of clustered data.

Figure 2.2: Marginal density and marginal hazard function considering a Weibull($\sigma = 0.25$, $\gamma = 2$) baseline hazard, $\alpha = 1$ and different values of λ.

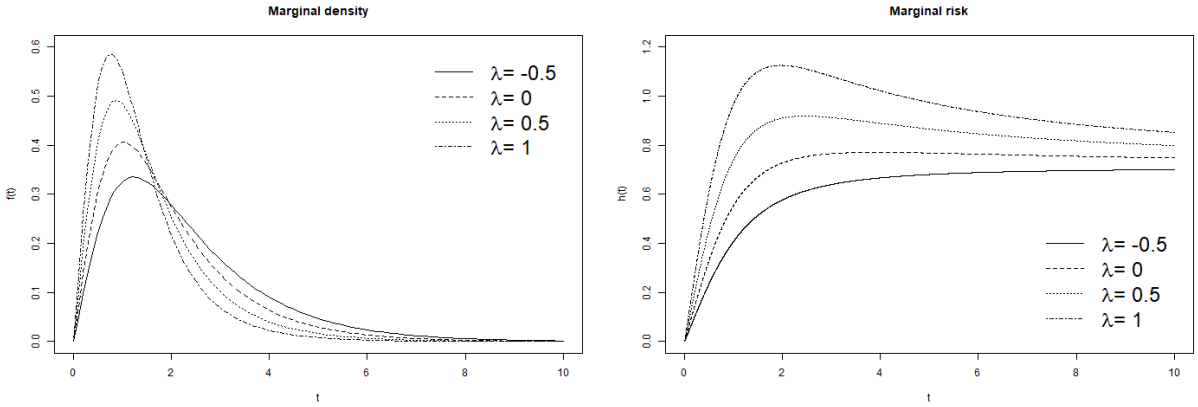
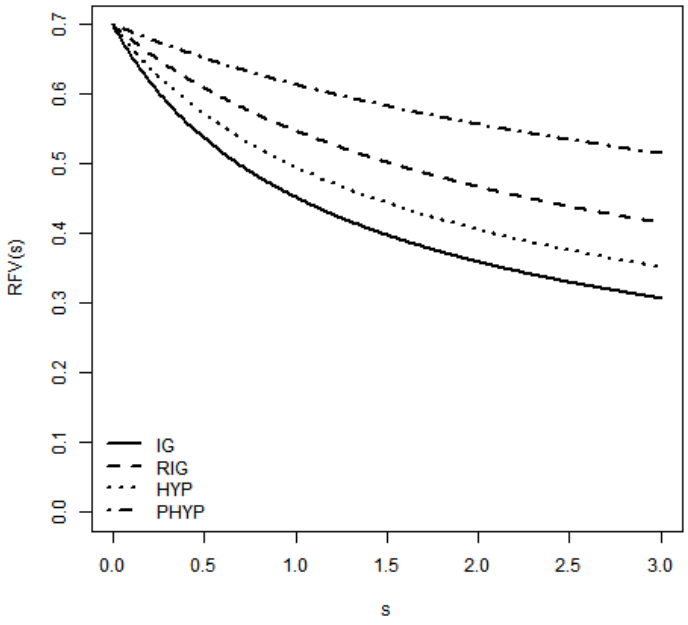


Figure 2.3: Relative frailty variance evolution for IG, RIG, HYP and PHYP frailties with α such that RFV(0) = 0.7 in each case.



CHAPTER 3

GENERALIZED INVERSE GAUSSIAN FRAILTY MODEL FOR CLUSTERED DATA

In this section we discuss the GIG frailty model including a regression structure and extending to the multivariate approach. In the univariate approach each individual has its own frailty value and the latent variable is interpreted as the degree of unobserved heterogenity. The univariate frailty models are applicable, for example, when important covariates are missing from the analysis. In the shared frailty models, individuals in a group, or cluster, share the same frailty. This creates positive dependence between individuals belonging to the same group. The shared frailty model is reduced to the univariate approach when groups are formed by one observation.

Assume m clusters with the i-th cluster having n_i individuals, for $i = 1, ...m$. Here T_{ij}^0 and C_{ij} denote the failure and censoring times for the individual $j = 1..., n_i$ in the i-th cluster. The total sample size is $n = \sum_{i=1}^{m} n_i$. In addition let $T_{ij} = \min\{T_{ij}^0, C_{ij}\}$ for $i = 1, ...m$ and $j = 1, ..., n_i$ be the observed random variables and $\delta_{ij} = I\{T_{ij}^0 \leq C_{ij}\}$ the failure indicator. Naturally, the frailty Z_i is associated to the i-th cluster. To complete model specification, we make the assumptions that given Z_i, $\{(T_{ij}^0, C_{ij}), j = 1, ..., n_i\}$ are conditionally independent and that T_{ij}^0 and $, C_{ij}$ are independent for all j. Another assumption that we will rely on is that the censoring times within the cluster $\{C_{ij}, j = 1, ..., n_i\}$ are non-informative with respect to Z_i.

Conditional to the frailty, the model has the same structure as the proportional hazards model by Cox [1972], with the inclusion of the frailty term Z_i acting multiplicatively in the baseline hazard function, as follows

$$h(t_{ij}|Z_i) = Z_i h_0(t_{ij}) \exp(x_{ij}^\top \beta), \quad t_{ij} > 0 \quad \forall \quad i, j.$$

3.1 Parametric GIG frailty model

In this section we will present the joint expressions of the clusters unconditional to the frailty. These expressions will then be used to construct the likelihood function where maximum likelihood estimation can be performed to find the parameters estimates under the parametric approach.

The joint survival function of a cluster can be found by using that $S(t_{i1},...,t_{in_i}) = L_{Z_i}\left(\sum_{j=1}^{n_i} H_0(t_{ij})e^{x_{ij}^\top\beta}\right)$. When accounting for Expression (2.2) we have that the joint survival function associated to the i-th cluster is given by

$$S(t_{i1},...,t_{in_i}) = L_{Z_i}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{x_{ij}^\top\beta}\right) = \left(\frac{\alpha^{-1}}{\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta}}\right)^{\lambda/2}\times \frac{K_\lambda(\sqrt{\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta})})}{K_\lambda(\alpha^{-1})}. \quad (3.1)$$

Next, in order to find an explicit and simple form for the joint density function we need the following Lemma. The proof of this Lemma can be found in the Appendix.

Lemma 1. *Let* $\zeta(x) = \frac{1}{x^{\phi/2}}K_\phi(\sqrt{x})$, *we have that* $\frac{\partial}{\partial x^k}\zeta(x) = \left(-\frac{1}{2}\right)^k \frac{K_{\phi+k}(\sqrt{x})}{x^{(\phi+k)/2}}$.

Using this result, we have that the joint density associated to the survival function in (3.1) is

$$f(t_{i1},...,t_{in_i}) = \frac{\alpha^{-\lambda}}{K_\lambda(\alpha^{-1})}(2/\alpha)^{n_i}\prod_{j=1}^{n_i}h_0(t_{ij})e^{x_{ij}^\top\beta}\chi^{(n_i)}\left(\alpha^{-1}(\alpha^{-1}+2H_0(t_{ij})e^{x_{ij}^\top\beta})\right),$$

where $\chi^{(k)}(x) = \frac{\partial}{\partial x^k}\frac{(-1)^k}{x^{\phi/2}}K_\phi(\sqrt{x})$. The result given in Lemma 1 provides an expression for $f(t_{i1},...,t_{in_i})$ no matter the size of the cluster. This is an advantage with respect to the GE model for example, because the high-order derivatives of the Laplace transform do not have a simple form there.

Let θ be the parameter vector and $L(\theta)$ denote the likelihood function. Here, θ is given by $(\beta, H_0, \alpha)^\top$, as we are considering fixed λ. $L(\theta)$ is given by

$$L(\theta) = \prod_{i=1}^{m}\int_0^\infty \prod_{j=1}^{n_i}(z_i h_0(t_{ij})e^{x_{ij}^\top\beta})^{\delta_{ij}}\exp\left(-z_i H_0(t_{ij})e^{x_{ij}^\top\beta}\right)f(z_i)dz_i,$$

where we are integrating with respect to the latent frailties so that $L(\theta)$ does not depend on unobserved quantities. To solve this, we can use the integral that comes from (2.1):

$$\int_0^\infty x^{\lambda-1}e^{\left\{\frac{-1}{2}(ax+b/x)\right\}}dx = 2K_\lambda(\sqrt{ab})\left(\frac{b}{a}\right)^{\lambda/2},$$

and conclude that the observed likelihood function is given by

$$L(\theta) = \prod_{i=1}^{m}\frac{1}{K_\lambda(\alpha^{-1})}\Psi_{\lambda+\sum_{j=1}^{n_i}\delta_{ij}}\left(\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta})\right)\alpha^{-(\sum_{j=1}^{n_i}\delta_{ij})+\lambda}\prod_{j=1}^{n_i}\left(h_0(t_{ij})e^{x_{ij}^\top\beta}\right)^{\delta_{ij}}, \quad (3.2)$$

where $\Psi_\lambda(x) = \frac{K_\lambda(\sqrt{x})}{x^{\lambda/2}}$.

The associated log-likelihood denoted by $\ell(\theta)$ is given as follows:

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{m} \log \Psi_{\lambda+\sum_{j=1}^{n_i}\delta_{ij}}\left(\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i} H_0(t_{ij})e^{x_{ij}^\top\beta})\right) - \log\alpha\left(\sum_{j=1}^{n_i}\delta_{ij}+\lambda\right) \\ &+ \sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\left(x_{ij}^\top\beta + \log h_0(t_{ij})\right) - m\log K_\lambda(\alpha^{-1}).\end{aligned} \tag{3.3}$$

In order to fit a parametric GIG frailty model, we specify the baseline hazard function $h_0(\cdot)$ and then obtain parameter estimates via maximum likelihood, that is maximizing (3.3) using numerical optimization methods, for example some are avaliable in R (R Core Team [2019]) `optim` such as BFGS and Nelder-Mead for example. In this work we will assume a with Weibull baseline hazard, that is obtained by using that $h_0(t) = \sigma\gamma t^{\gamma-1}$ and $H_0(t) = \sigma t^\gamma$.

3.2 Semiparametric GIG frailty model

When no parametric form is desired for the baseline hazard function, a semiparametric approach can be used. In this section we introduce the semiparametric generalized inverse-Gaussian frailty model based on an EM algorithm and piecewise exponential hazards.

Our choice of methodology follows Lawless & Zhan [1998] who argues that the use of the piecewise-constant hazard function avoids many problems associated with non- and semiparametric methods for incomplete survival data, but still provides a high degree of robustness. Naturally, using piecewise constant hazards offers greater flexibility than making a parametric assumption about the form of the hazard function.

In frailty models this methodology was also adopted by Liu & Huang [2008] who proposed a Gaussian quadrature estimation method for the gamma and normal frailties. Some motivations for using this method, according to the authors, were the simplicity of implementation, accuracy of the estimates and avaliability of the standard errors. Feng et al. [2005] also highlights the flexibility provided by the usage of piecewise constant hazards, that according to them is quite general and can approximate various shapes of the baseline hazard function.

The baseline hazard function of a piecewise exponential distribution is given by

$$h_0(t) = \eta_l, t^{(l-1)} \leq t < t^{(l)}, \quad l = 1, ..., k+1 \quad , \tag{3.4}$$

where $t^{(l)}$ denotes the l-th ordered time and $0 = t^{(0)} < \min(t_i, i = 1, ..., n) \leq t^{(1)} < ... < t^{(k)} \leq \max(t_i, i = 1, ..., n) = t^{(k+1)}$ are the prespecified cutpoints. This implies that the cumulative baseline function is $H(t|\eta) = \sum_{j=1}^{i}\eta_j[min(t, t^{(j+1)}) - t^{(j)}]$ if $t^{(i)} \leq t < t^{(i+1)}$ for $i = 0, ..., k$. Hence, specifying k change points means separating the data into $k+1$ groups, so there are $k+1$ failure rates to be estimated in the model.

When Kim & Proschan [1991] introduced the estimation of the survival function through the piecewise exponential estimator, they considered as many partitions as the number of observed failures. However, in this matter we will follow Lawless & Zhan [1998] that showed through simulation studies that frailty models based on the piecewise constant hazards often result in excelent estimation of regression and frailty coefficients when 8-10 pieces are adopted. Next, we will proceed to the estimation of parameters via an EM (Expectation-Maximization) algorithm [Dempster et al., 1977].

The EM algorithm is a popular method to deal with the presence of non-observed data in the likelihood function and has been used in the literature of frailty models by Klein [1992], Wang & Klein [2012], Callegaro & Iacobelli [2012] and Barreto-Souza & Mayrink [2019] to name a few. This algorithm consists on alternating between two parts. In the Expectation (E) step, we compute the conditional expectation of the log-likelihood given the observed data, evaluated at the current estimate of the parameters. This is named Q-function. It also includes a Maximization (M) step, in which we maximize the Q-function. The new estimates obtained in the M step are used to update the Q-function and we iterate between these steps until some convergence criteria is reached.

The complete data set is given by $(t_{ij}, \delta_{ij}, Z_i)$, for $i = 1, ..., m$ and $j = 1, ..., n_i$. We observe the pairs (t_{ij}, δ_{ij}) and Z_i are the latent random effects. The complete likelihood is given by

$$L_c(\theta) = \prod_{i=1}^{m}\prod_{j=1}^{n_i}(z_i h_0(t_{ij})e^{x_{ij}^\top\beta})^{\delta_{ij}} \exp\left(-z_i H_0(t_{ij})e^{x_{ij}^\top\beta}\right) \frac{1}{2K_\lambda(\alpha^{-1})} z_i^{\lambda-1} \exp\left\{-\frac{1}{2\alpha}\left(z_i + \frac{1}{z_i}\right)\right\},$$

This likelihood function can be written as the product of two terms, i.e. $L_c(\theta) = L_1(\beta, H)L_2(\alpha)$ where

$$L_1(\beta, H) = \prod_{i=1}^{m}\prod_{j=1}^{n_i}(z_i h_0(t_{ij})e^{x_{ij}^\top\beta})^{\delta_{ij}} \exp\left(-z_i H_0(t_{ij})e^{x_{ij}^\top\beta}\right),$$

and

$$L_2(\alpha, \lambda) = \prod_{i=1}^{m} \frac{1}{2K_\lambda(\alpha^{-1})} z_i^{\lambda-1} \exp\left\{-\frac{1}{2\alpha}\left(z_i + \frac{1}{z_i}\right)\right\}.$$

As a consequence, the associated complete log-likelihood can be written as $\ell_c(\theta) = \ell_1(\beta, H) + \ell_2(\alpha)$ where the first term is given by

$$\ell_1(\beta, H) \propto \sum_{i=1}^{m}\sum_{j=i}^{n_i} \delta_{ij}\left(x_{ij}^\top\beta + \log h_0(t_{ij})\right) - \sum_{i=1}^{m}\sum_{j=i}^{n_i} Z_i H_0(t_{ij})e^{x_{ij}^\top\beta},$$

and the second term is

$$\ell_2(\alpha, \lambda; Z) = -m\log 2K_\lambda(\alpha^{-1}) + (\lambda - 1)\sum_{i=1}^{m} log(z_i) - \frac{1}{2\alpha}\sum_{i=1}^{m}(z_i + 1/z_i).$$

In order to determine the expectation step of the EM algorithm, we need to find the conditional density of Z_i given the observed data $\{t_{ij}, \delta_{ij}\}_{j=1}^{n_i}$. Using basic probability it can be shown that this density function is given by $f(z_i|t_{ij}, \delta_{ij}, j = 1, ..., n_i) = f(t_{ij}, \delta_{ij}, j = 1, ..., n_i|z_i)f(z_i)/f(t_{ij}, \delta_{ij}, j = 1..., n_i)$.

More specifically, we have that

$$f(z_i|t_{ij},\delta_{ij},j=1,...,n_i) \propto z_i^{\lambda+\sum_{j=1}^{n_i}\delta_{ij}-1}\exp\left\{-\frac{1}{2}\left[z_i\left(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta}\right)+\frac{1}{\alpha z_i}\right]\right\}.$$

Therefore, it can be observed that the conditional density of Z_i given the observed data $\{t_{ij},\delta_{ij}\}_{j=1}^{n_i}$ follows a GIG$(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta},\alpha^{-1},\lambda+\sum_{j=1}^{n_i}\delta_{ij})$ distribution for $i=1,...,m$. Conditional on the observed data, the frailties $Z_1,...,Z_m$ are independent GIG distributed random variables with the mentioned parameters.

The Q-function is the conditional expected value of the complete likelihood given the observed data at the current parameter estimates, that is, $Q(\theta,\theta^{(r)}) \equiv E\left(\ell_c(\theta)|(t_{ij},\delta_{ij}),i=1,...,m,j=1...,n_i;\theta^{(r)}\right)$ where $\theta^{(r)}$ denotes the estimated vector of parameters θ in the r-th step. The Q-function depends on the expectations $E\left(Z_i|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right)\equiv\omega_i(\theta)$, $E\left(1/Z_i|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right)\equiv\kappa_i(\theta)$ and $E\left(\log(Z_i)|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right)\equiv\nu_i(\theta)$. These expectations are presented in the next proposition.

Proposition 1. *(E-step of the EM algorithm) For $i=1,...,m$, we have that*

$$\omega_i(\theta)=E\left(Z_i|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right) = \frac{K_{\lambda+\sum_{j=1}^{n_i}\delta_{ij}+1}\left(\sqrt{\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta})}\right)}{K_{\lambda+\sum_{j=1}^{n_i}\delta_{ij}}\left(\sqrt{\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta})}\right)}\times \left(\frac{\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta}}{\alpha^{-1}}\right)^{-1/2},$$

that is obtained from (2.3).

In order to calculate $\kappa_i(\theta)$ we can identify the kernel of a GIG distribution and find that this conditional expectation is given by

$$\kappa_i(\theta)=E\left(1/Z_i|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right) = \frac{K_{\lambda+\sum_{j=1}^{n_i}\delta_{ij}-1}\left(\sqrt{\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta})}\right)}{K_{\lambda+\sum_{j=1}^{n_i}\delta_{ij}}\left(\sqrt{\alpha^{-1}(\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta})}\right)}\times \left(\frac{\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta}}{\alpha^{-1}}\right)^{1/2}.$$

Our strategy to find $\nu_i(\theta)$ consists on calculating $M_{\log Z_i}(t)$, the moment generating function (MGF) of $\log Z_i$ and evaluate its derivative at $t=0$. For the sake of simplifing the demostration we will use the following notation for the parameters:

$$a^*=\alpha^{-1}+2\sum_{j=1}^{n_i}H_0(t_{ij})e^{x_{ij}^\top\beta},\quad b^*=\alpha^{-1}\ \text{and}\ \lambda^*=\lambda+\sum_{j=1}^{n_i}\delta_{ij}.$$

The MGF of $\log Z_i$ *is*

$$M_{\log Z_i}(t) = E\left(e^{\log(z_i)t}\right) = \left(\frac{a^*}{b^*}\right)^{\lambda^*/2} \frac{1}{2K_{\lambda^*}(\sqrt{a^*b^*})} \int_0^\infty z_i^{\lambda^*+t-1} \exp\left\{-\frac{1}{2}\left(a^* z_i + \frac{b^*}{z_i}\right)\right\} dz_i.$$

Solving this, we obtain

$$M_{\log Z_i}(t) = \left(\frac{a^*}{b^*}\right)^{-t/2} \frac{K_{\lambda^*+t}(\sqrt{a^*b^*})}{K_{\lambda^*}(\sqrt{a^*b^*})}.$$

The derivative of the expression above depends on the derivative of the Bessel function in terms of its index, that does not have a closed form. Finally, $E\left(\log(Z_i)|(t_{ij}, \delta_{ij})_{j=1}^{n_i}\right)$ *is given by*

$$\nu_i(\theta) = E\left(\log(Z_i)|(t_{ij}, \delta_{ij})_{j=1}^{n_i}\right) = -\frac{1}{2}\log\left(\frac{a^*}{b^*}\right) + \frac{\frac{\partial}{\partial t}K_{\lambda^*+t}(\sqrt{a^*b^*})\big|_{t=0}}{K_{\lambda^*}(\sqrt{a^*b^*})}.$$

Although an analytical expression was found for $\nu_i(\theta)$, it depends on the derivative of the Bessel function with respect to its index, which can be numerically unstable. This is discussed by Palmer et al. [2016], who presents an alternative form for this expectation. We will work with the expression provided in this reference which gives us that a good approximation for $\nu_i(\theta)$ is

$$\nu_i(\theta) = E\left(\log(Z_i)|(t_{ij}, \delta_{ij})_{j=1}^{n_i}\right) \approx \epsilon^{-1}\left(\left(\frac{a^*}{b^*}\right)^{\epsilon} \frac{K_{\lambda^*+\epsilon}(\sqrt{a^*b^*})}{K_{\lambda^*}(\sqrt{a^*b^*})} - 1\right),$$

when taking a sufficiently small $\epsilon > 0$. We will work with $\epsilon = 10^{-6}$ as suggested by the authors.

Now we are ready to obtain the Q-function, that is $Q(\theta, \theta^{(r)}) = Q_1(\beta, H_0(t); \theta^{(r)}) + Q_2(\alpha; \theta^{(r)})$, where

$$Q_1(\beta, H_0(t); \theta^{(r)}) \propto \sum_{i=1}^m \sum_{j=i}^{n_i} \delta_{ij}\left(x_{ij}^\top \beta + \log h_0(t_{ij})\right) - \sum_{i=1}^m \sum_{j=i}^{n_i} \omega_i(\theta^{(r)}) H_0(t_{ij}) e^{x_{ij}^\top \beta},$$

and

$$Q_2(\alpha; \theta^{(r)}) \propto -m\log(K_\lambda(\alpha^{-1})) + (\lambda - 1)\sum_{i=1}^m \nu_i(\theta^{(r)}) - \frac{1}{2\alpha}\sum_{i=1}^m (\omega_i(\theta^{(r)}) + \kappa_i(\theta^{(r)})).$$

By assuming the piecewise hazard function for $h_0(t)$ in (3.4) and providing its cutpoints, we have all components needed to find the estimates of the parameters. Using the EM algorithm instead of maximizing the observed log-likelihood function in (3.3) with piecewise constant hazards, allows us to have simpler expressions to maximize. In addition, we are able to do this separately for (β, η) and α. The EM algorithm for the GIG frailty model will be carried as descripted in Algorithm 1.

There are two procedures available to find the standard errors of the estimates. The first one consists on the method indicated in Klein [1992] that relies on the observed Fisher Information matrix, that

Algorithm 1 EM-algorithm for the semiparametric GIG frailty model

1: Provide initial guesses for the parameters. Denoting $\theta^{(r)}$ as the estimate of the set of parameters at step r, $\beta^{(0)}$ will be those obtained through the fit of the proportional hazards model in R package `survival` (Therneau [2015]). We set $\alpha^{(0)} = 1$ and $\eta_k^{(0)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}/\sum_{i=1}^{m}\sum_{j=1}^{n_i} t_{ij}$ if t_{ij} is a failure time for $t^{(l)} \leq t_{ij} < t^{(l+1)}$ where $t^{(l)}$ represents the l-th change point, for $l = 1, ..., k$. The last expression corresponds to the maximum likelihood estimator of the rate of an exponential distribution in each given interval considering only the failure times within that interval.

2: Update the Q-function using the expressions found for $\omega_i(\theta)$, $\kappa_i(\theta)$ and $\nu_i(\theta)$ and $\theta = \widehat{\theta}^{(r)}$.

3: Use the expressions of $h_0(t)$ and $H_0(t)$, corresponding to a piecewise exponential distribution, and numerically obtain the maximum likelihood estimates of β and η by maximizing Q_1. For this task, we chose to work with the `optim` function in R using BFGS maximization routine (Fletcher, R. [2000]). The new estimates of the parameters are denoted by $\widehat{\beta}^{(r+1)}$ and $\widehat{\eta}^{(r+1)}$.

4: Obtain the maximum likelihood estimate of α by maximize Q_2 numerically. Again, we use R `optim` command and the BFGS maximization procedure. The new estimate is denoted by $\widehat{\alpha}^{(r+1)}$.

5: Verify a convergence criterion. For example, $||\theta^{(r+1)} - \theta^{(r)}|| < \epsilon$ for some prestablished $\epsilon > 0$. If the convergence criterion is satisfied, then $\theta^{(r+1)}$ are the final parameter estimates. Otherwise, we update $\theta^{(r)}$ with $\theta^{(r+1)}$ and return to step 2.

is, $I(\beta, \alpha, \eta) = -\dfrac{\partial^2 \ell(\beta, \alpha, \eta)}{\partial(\beta, \alpha, \eta)\partial(\beta, \alpha, \eta)^\top}$, where ℓ is the observed log-likelihood given in (3.2). In order to use this method, an option would be to obtain this matrix numerically since the analytical calculations would be very cumbersome. After having the final estimates of the algorithm, the standard errors of the parameters are given by the diagonal elements of $I(\widehat{\beta}, \widehat{\alpha}, \widehat{\eta})$. Another way of approaching the standard errors is through bootstrap resampling (Efron, B. [1979]).

CHAPTER 4

SIMULATION STUDIES

In this section we present simulation studies with the goal of evaluating the performance of the estimates produced by the proposed model under correct model specification and under misspecification. First, we will explore the parametric version of the GIG frailty model specifying a Weibull baseline hazard function. We evaluate parameter estimation under correct model specification and finite sample sizes by running a Monte Carlo study with 1000 replicates. This is presented in Subsection 4.1. Our main interest is in the semiparametric version of the GIG frailty model, which will be explored in Subsection 4.2. We will present scenarios in which the data are generated with gamma, generalized exponential (GE), log-normal and inverse Gaussian frailties. We fit to each data set the IG, RIG, HYP and PHYP frailty models with different numbers of change points for the piecewise constant hazards. In addition, we fit the semiparametric versions of the gamma and GE frailty models avaliable in literature. In the semiparametric simulation studies we also consider 1000 Monte Carlo replicas in each scenario.

4.1 Parametric simulation study

In this subsection we explore the behavior of the estimators under finite samples. We evaluated the performance of the proposed model under the parametric approach by assuming a Weibull baseline hazard function with the parametrization previously mentioned. The simulation study was conducted using 1000 Monte Carlo replicas and we assessed the number of cluster of $m = 50, 100$ and 200 with $n_i = 2 (i = 1, ..., m)$ observations per cluster in all cases. In the parametric approach, we chose to explore the performance of the correct specified models. A simulation study under misspecification will be presented in the next section for the semiparametric version.

We generated independent random covariates from a Bernoulli($p = 0.5$) and Uniform($-1, 1$) distributions with true values of the fixed effects being $(\beta_1, \beta_2) = (1.5, -1.0)$. True value of α was set to 0.7. Given the frailties, the survival times were generated from a Weibull distribution with $\gamma = 2$ and $\sigma = 0.25$. The censoring times were generated independently from a Weibull distribution with parameters $\gamma = 2$ and $\sigma = 0.05$. Thus, we have a censoring rate around 30%. We explore parameter estimation for the complete set of parameters $\theta = (\beta_1, \beta_2, \sigma, \gamma, \alpha)^\top$ using the IG, RIG, HYP and PHYP frailty models.

Table 4.1: Empirical mean and root mean square error (RMSE) of the parameter estimates under the parametric approach considering a Weibull baseline hazard.

Parameter	Model	m=50 Mean	RMSE	m=100 Mean	RMSE	m=200 Mean	RMSE
$\beta_1 = 1.5$	IG	1.5664	0.0664	1.5322	0.0322	1.5159	0.0159
	RIG	1.5566	0.0291	1.5325	0.0325	1.5092	0.0092
	HYP	1.5678	0.0678	1.5167	0.0167	1.5172	0.0172
	PHYP	1.5489	0.0489	1.5180	0.0180	1.5194	0.0194
$\beta_2 = -1$	IG	−1.0474	0.0474	−0.9992	0.0008	−1.0072	0.0072
	RIG	−1.0291	0.0291	−1.0160	0.0160	−1.0127	0.0127
	HYP	−1.0561	0.0561	−1.0125	0.0125	−1.0125	0.0125
	PHYP	−1.0296	0.0296	−1.0083	0.0083	−1.0114	0.0114
$\sigma = 0.25$	IG	0.2700	0.0200	0.2563	0.0063	0.2518	0.0018
	RIG	0.2507	0.0007	0.2513	0.0013	0.2542	0.0042
	HYP	0.2565	0.0065	0.2542	0.0042	0.2510	0.0010
	PHYP	0.2616	0.0116	0.2570	0.0070	0.2504	0.0004
$\gamma = 2$	IG	2.0912	0.0912	2.0378	0.0368	2.0214	0.0214
	RIG	2.0850	0.0850	2.0290	0.0290	2.0135	0.0135
	HYP	2.0851	0.0851	2.0327	0.0327	2.0218	0.0218
	PHYP	2.0660	0.0660	2.0246	0.0246	2.0205	0.0205
$\alpha = 0.7$	IG	1.0491	0.3491	0.8298	0.1298	0.7487	0.0487
	RIG	0.9573	0.2573	0.8087	0.1087	0.7306	0.0306
	HYP	0.9277	0.2277	0.7871	0.0871	0.7470	0.0470
	PHYP	1.0605	0.3605	0.8046	0.1046	0.7717	0.0717

Table 4.1 contains the empirical means and root mean square errors (RMSE) of the parameter estimates for this simulation study. Figures 4.1 - 4.5 illustrate with boxplots the estimates under the four proposed frailty models so that we have a better picture on how the estimates are distributed.

In Table 4.1 we note that the covariate effects are well estimated even for the smallest sample size in all frailty models. The same can be said about the parameters of the baseline hazard function, where we have observed that γ and σ are estimated with low bias in all sample sizes. In Figure 4.5, we observed that the estimation of α occurs with greater variability when in smaller sample sizes and presents slight bias, as reported in Table 4.1. However, as expected, bias and RMSE reduce when increasing sample size and the estimates of α, when $m = 200$, are reasonably close to the true value.

There was not a relevant difference between the models under correct specification in the estimation of the parameters. The variability of the estimates under the four models considered also seem to be similar when looking at the boxplots in Figures 4.1 - 4.5.

4.2 Semiparametric simulation study

In this section we analyze the four special cases of the GIG frailty under model misspecification using the semiparametric approach discussed in subsection 3.2. For that we will consider data sets generated with gamma, generalized exponential (GE) and log-normal frailties and sample sizes of $m = 200$ and $m = 500$ with clusters formed by $n_i = 2$ individuals each. In addition, we explore λ misspecification by generating data with inverse Gaussian frailty and fitting all special cases. In this last scenario,

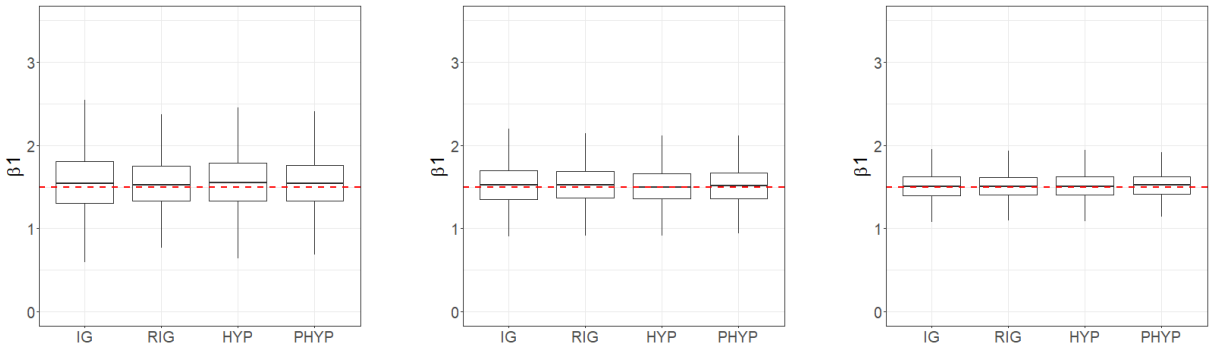


Figure 4.1: Boxplots of the estimates obtained in the Monte Carlo simulation for β_1. Each boxplot corresponds to a frailty distribution as indicated in the horizontal axis. The columns represent the sample sizes of $m = 50$ (on the left), $m = 100$ (in the middle) and $m = 200$ (on the right). The horizontal dashed line indicates the true value of the parameter.

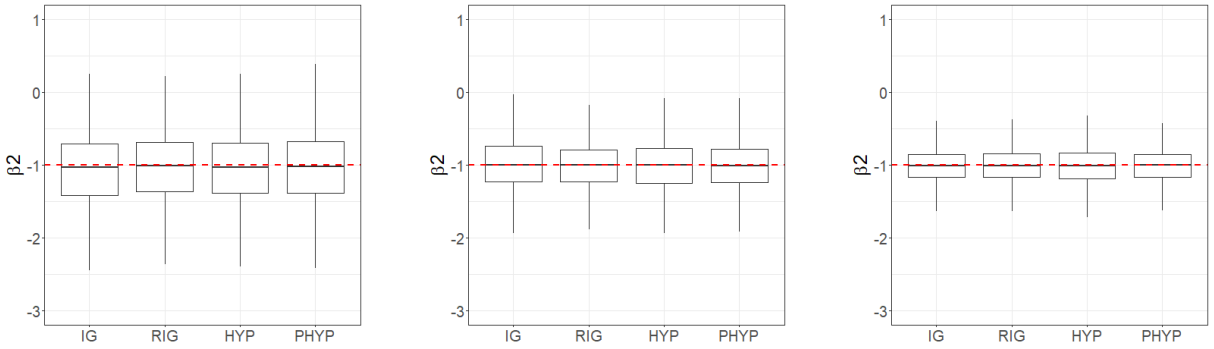


Figure 4.2: Boxplots of the estimates obtained in the Monte Carlo simulation for β_2. Each boxplot corresponds to a frailty distribution as indicated in the horizontal axis. The columns represent the sample sizes of $m = 50$ (on the left), $m = 100$ (in the middle) and $m = 200$ (on the right). The horizontal dashed line indicates the true value of the parameter.

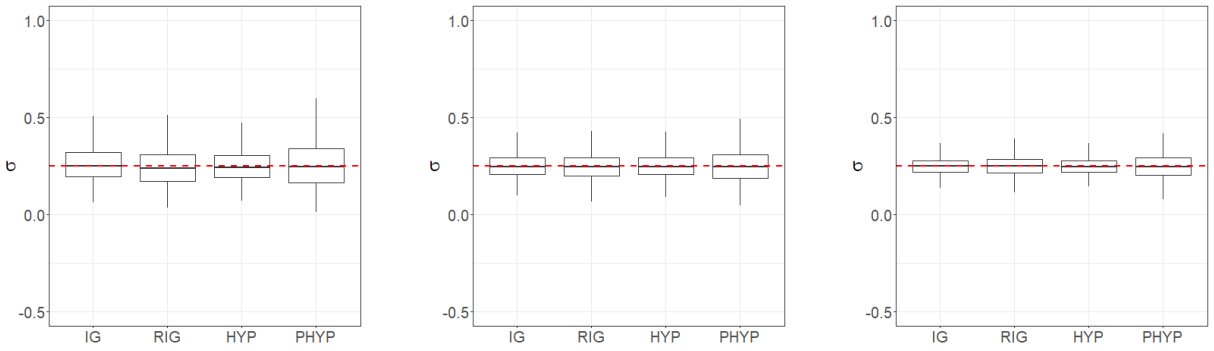


Figure 4.3: Boxplots of the estimates obtained in the Monte Carlo simulation for σ. Each boxplot corresponds to a frailty distribution as indicated in the horizontal axis. The columns represent the sample sizes of $m = 50$ (on the left), $m = 100$ (in the middle) and $m = 200$ (on the right). The horizontal dashed line indicates the true value of the parameter.

we also aim to explore the behavior of the models under a larger cluster size. Thus, in this part we consider $m = 20$ and $m = 100$ with $n_i = 10$ for all i, mantaining the total sample size. Analogous to the parametric simulation, given the frailties, the failure times T_{ij}^0 were generated from a Weibull distribution with $\gamma = 2$ and $\sigma = 0.25$ and the censoring times C_{ij} were generated, independently of T_{ij}^0, following

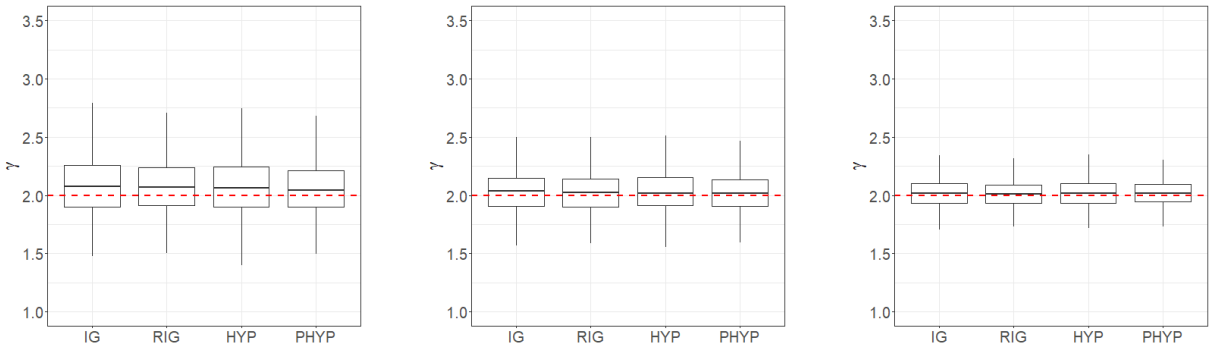


Figure 4.4: Boxplots of the estimates obtained in the Monte Carlo simulation for γ. Each boxplot corresponds to a frailty distribution as indicated in the horizontal axis. The columns represent the sample sizes of $m = 50$ (on the left), $m = 100$ (in the middle) and $m = 200$ (on the right). The horizontal dashed line indicates the true value of the parameter.

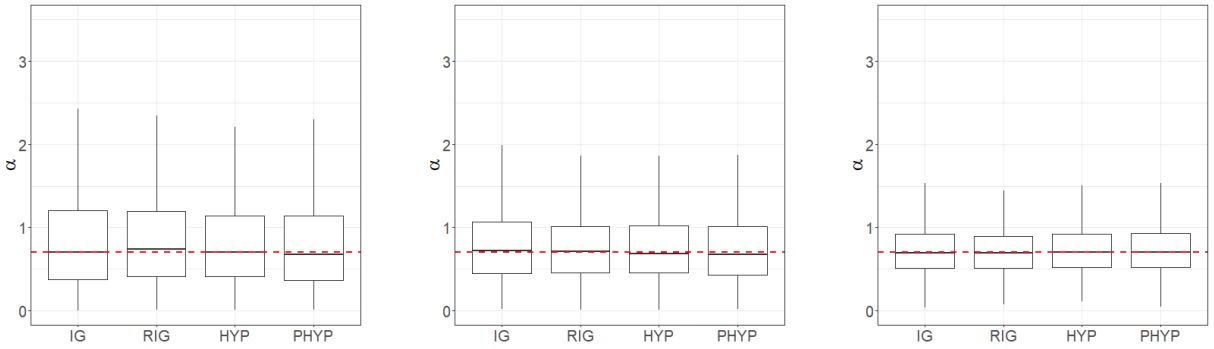


Figure 4.5: Boxplots of the estimates obtained in the Monte Carlo simulation for α. Each boxplot corresponds to a frailty distribution as indicated in the horizontal axis. The columns represent the sample sizes of $m = 50$ (on the left), $m = 100$ (in the middle) and $m = 200$ (on the right). The horizontal dashed line indicates the true value of the parameter.

a Weibull distribution with parameters $\gamma = 2$ and $\sigma = 0.05$. The observed data are $\min\{T_{ij}^0, C_{ij}\}$ and $\delta_{ij} = I\{T_{ij}^0 \leq C_{ij}\}$. With this configuration we have a percentage of approximately 30% censored observations.

We generated independent random covariates from the Bernoulli($p = 0.5$) and Uniform$(-1, 1)$ distributions with true values of their effects being $(\beta_1, \beta_2) = (1.5, -1.0)$. True value of α was set to 1. We evaluated parameter estimation for the complete set of parameters $\theta = (\beta_1, \beta_2, \alpha)^\top$ under the IG, RIG, HYP and PHYP frailty models. Fitting the IG, RIG, HYP and PHYP models is done considering 5 and 10 change points for the piecewise baseline hazard function, that is, $k = 5$ and $k = 10$. In the first three scenarios, the fits of the gamma and GE models are also included. In the last one, we do not report the GE model as it is most suitable for small clusters. The semiparametric versions of the gamma and GE models are based on the Cox partial likelihood function and therefore do not require cut point specification.

The parameter α represents the frailty variance only in the gamma and IG cases. Hence, in each case an appropriate transformation of this parameter is calculated so that we obtain the frailty variance. This comparison is done as indicated by Barreto-Souza & Mayrink [2019] that states that the model given

Table 4.2: Empirical mean and root mean square error (RMSE) of β_1, β_2 and the frailty variance for gamma distributed frailty data with sample size $m = 200$. Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is also 1.

		β_1		β_2		Var	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.288	0.212	-0.857	0.143	0.592	0.408
Semiparametric GE	-	1.396	0.104	-0.927	0.073	0.767	0.233
Semiparametric - IG	$k = 5$	1.307	0.193	-0.862	0.138	1.541	0.541
	$k = 10$	1.361	0.139	-0.900	0.100	1.866	0.866
Semiparametric - RIG	$k = 5$	1.408	0.092	-0.927	0.073	1.153	0.153
	$k = 10$	1.482	0.018	-0.981	0.019	1.272	0.272
Semiparametric - HYP	$k = 5$	1.365	0.135	-0.898	0.102	1.388	0.388
	$k = 10$	1.434	0.066	-0.946	0.054	1.608	0.608
Semiparametric - PHYP	$k = 5$	1.383	0.117	-0.914	0.086	0.865	0.135
	$k = 10$	1.427	0.073	-0.949	0.051	0.894	0.106

by $h(t_{ij}|Z_i) = Z_i h_0(t_{ij}) \exp(x_{ij}^\top \beta)$ is equivalent to $h(t_{ij}|Z_i) = Z_i^* h_0^*(t_{ij}) \exp(x_{ij}^\top \beta)$ with $Z_i^* = Z_i/E(Z_i)$ having mean 1, and $h_0^*(t_{ij}) = h_0(t_{ij})E(Z_i)$. In other words, the comparison of the frailty variance should be done through the transformation $\text{Var}(Z_i^*) = \text{Var}(Z_i)/E(Z_i)^2$. The proper transformation for each model is done so that they are comparable and is reported in the column named "Var".

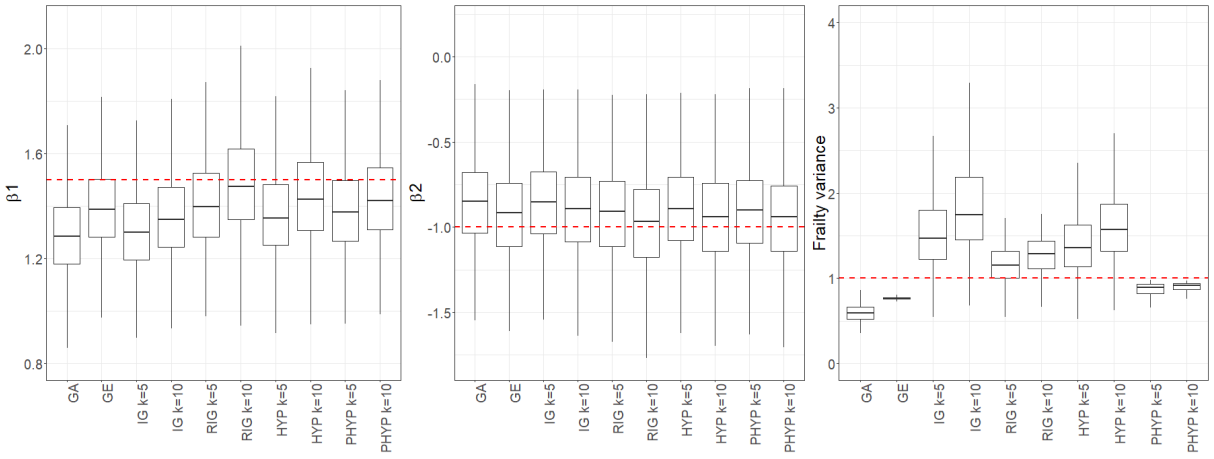


Figure 4.6: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a gamma distribution with $m = 200$. The horizontal dashed line indicates the real value of the parameter.

In Table 4.2 we present the empirical mean and root mean square error (RMSE) of 1000 Monte Carlo replicas when data are generated with gamma frailty and sample size is equal to $m = 200$ with $n_i = 2$ for all i, totalizing 400 observations. We estimated the effects of the two covariates considered and the frailty variance through the gamma, GE and GIG fits. For the GIG model, we included the four special cases under investigation and the number of change points of $k = 5$ and $k = 10$.

We observed that the correct model clearly underestimated the frailty variance. This is something that was also observed by Barreto-Souza & Mayrink [2019] in their simulation studies. The authors

Table 4.3: Empirical mean and root mean square error (RMSE) of β_1. β_2 and the frailty variance for gamma distributed frailty data with sample size $m = 500$. Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is also 1.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.290	0.210	−0.863	0.137	0.609	0.391
Semiparametric GE	-	1.393	0.107	−0.929	0.071	0.769	0.231
Semiparametric - IG	$k = 5$	1.294	0.206	−0.860	0.140	1.480	0.480
	$k = 10$	1.345	0.155	−0.897	0.103	1.782	0.782
Semiparametric - RIG	$k = 5$	1.395	0.105	−0.923	0.077	1.149	0.149
	$k = 10$	1.471	0.029	−0.978	0.022	1.275	0.275
Semiparametric - HYP	$k = 5$	1.350	0.150	−0.894	0.106	1.351	0.351
	$k = 10$	1.418	0.082	−0.942	0.058	1.570	0.570
Semiparametric - PHYP	$k = 5$	1.381	0.119	−0.915	0.085	0.882	0.118
	$k = 10$	1.425	0.075	−0.951	0.049	0.909	0.091

pointed to the fact that the difficulty in estimating α in the gamma frailty model is due to the flat shape of its Q-function. We can note clear advantage of the PHYP, RIG and GE models in estimating this quantity, as they returned smaller RMSEs than the true model. Although the gamma model is the correctly specified model in this scenario, it presented the greatest bias in the estimation of fixed effects as well. The lowest RMSEs in the estimation of β_1 and β_2 are related to GIG frailties.

The boxplots of the estimates in Figure 4.6 confirm our statement that gamma model had the poorest performance in estimating β_1 and β_2 as well as the frailty variance, since its boxplots are farthest from the true values. The boxplots of frailty variance estimates show us that between the GIG frailty special cases, the PHYP is the one showing lowest variability. The gamma and GE models also show low variability but higher bias.

As for the number of cutpoints k, we observed a reduction in the RMSEs of β_1 and β_2 in all cases, but only in the PHYP model it implied in an improvement in the estimation of the frailty variance too. The other three models showed an increase in the frailty variance bias when increasing k.

In general, we can observe the same behavior when we increase the total sample size from $n = 400$ to $n = 1000$. Again, each clusters is formed by two individuals. There is not a significant difference between the two sample sizes considered. The decrease in the RMSE of the frailty variance estimates in the gamma and GE cases is very small, which gives evidence that increasing the sample size did not help much to decrease the bias. In the same way, for the IG, RIG and HYP models a similar result occurs.

In the PHYP model, the frailty variance was estimated with low bias when $m = 200$ unlike the other models under comparison. Thus, as expected, this is also the model that performed best when $m = 500$. Again, for both fits with $k = 5$ and $k = 10$, the mean estimation of the frailty variance in this model was closer to the real value than the correctly specified one. Something we can observe by comparing the two sample sizes tested is that there is evidence that the number of change points equal to 10 is more appropriate when $m = 500$. This can be said based on fixing the number of change points to 5 and

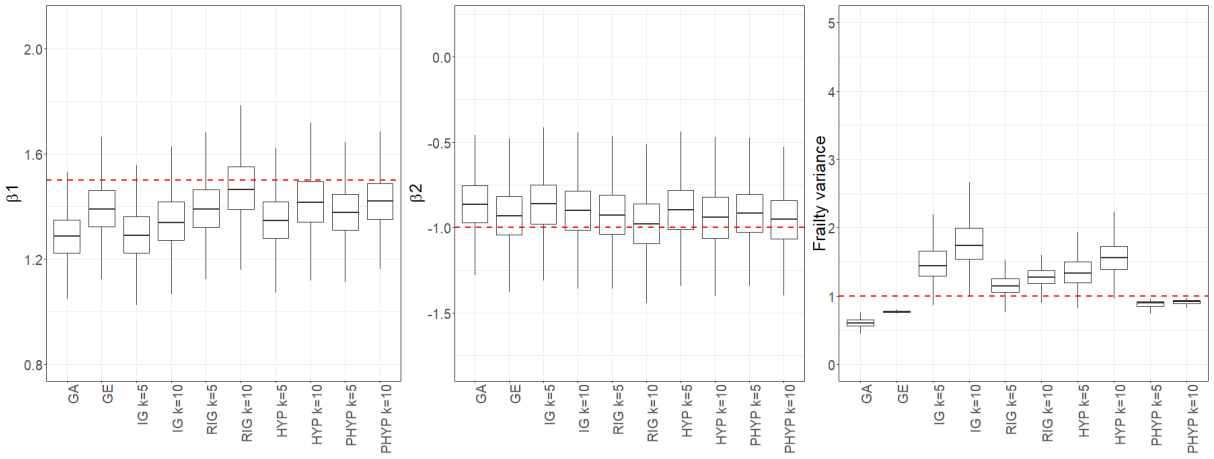


Figure 4.7: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a gamma distribution with $m = 500$. The horizontal dashed line indicates the real value of the parameter.

Table 4.4: Empirical mean and root mean square error (RMSE) of β_1, β_2 and the frailty variance for GE distributed frailty data with sample size $m = 200$. Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is also 1.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.304	0.196	−0.869	0.131	0.645	0.355
Semiparametric GE	-	1.391	0.109	−0.924	0.076	0.770	0.230
Semiparametric - IG	$k = 5$	1.335	0.166	−0.885	0.115	1.602	0.602
	$k = 10$	1.407	0.093	−0.938	0.062	2.064	1.064
Semiparametric - RIG	$k = 5$	1.423	0.077	−0.938	0.062	1.163	0.163
	$k = 10$	1.483	0.017	−0.982	0.018	1.257	0.257
Semiparametric - HYP	$k = 5$	1.388	0.112	−0.917	0.083	1.407	0.407
	$k = 10$	1.448	0.052	−0.960	0.040	1.595	0.595
Semiparametric - PHYP	$k = 5$	1.393	0.107	−0.921	0.079	0.876	0.124
	$k = 10$	1.426	0.074	−0.949	0.051	0.895	0.105

comparing the average estimates of sample size $m = 200$ to those of $m = 500$, where we can see equal or slightly worse estimates of the covariate effects. This behavior is not expected when the sample size is increased. For example, the estimates of β_1 in the IG, RIG, HYP and PHYP were 1.307, 1.408, 1.350 and 1.383 when $m = 200$, whereas for $m = 500$ these estimates were respectively 1.294, 1.395, 1.350 and 1.381. We also tested the number of cut points of 20 but did not observe a significant difference in comparison to the use of 10 cut points, which is in agreement with Lawless & Zhan [1998].

When the data are generated with GE frailty, we can see in Table 4.4 that all models considered estimate well the effects of the covariates, among which the gamma model again displays the higher RMSEs in the estimation of β_1 and β_2. The correct specified model underestimates the frailty variance, as does the gamma model. Meanwhile, we observe in the IG frailty a considerable overestimation of this quantity. However, the PHYP and RIG frailties perform well in this scenario. The first one showed

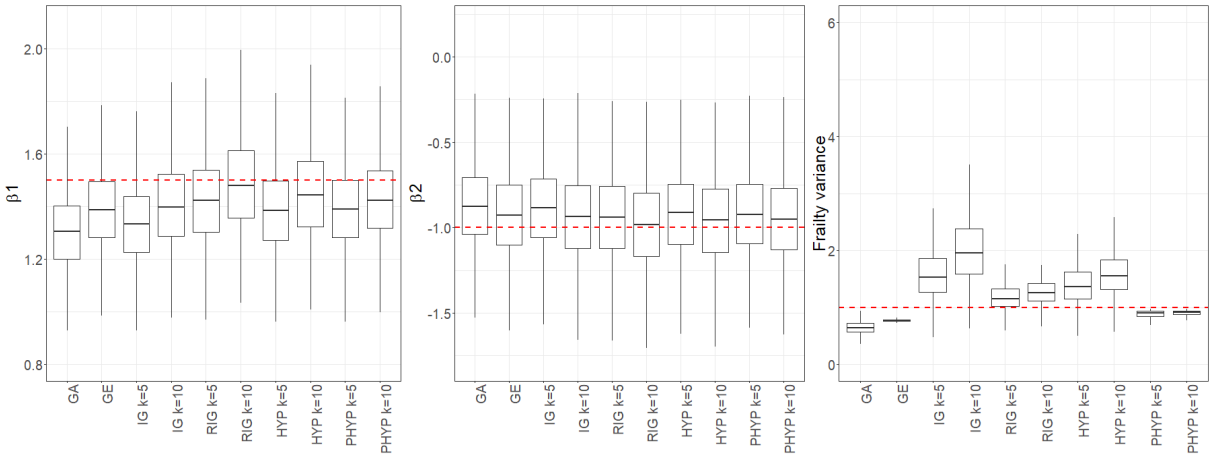


Figure 4.8: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1 , β_2 and the frailty variance when data is generated from a generalized exponential distribution with $m = 200$. The horizontal dashed line indicates the real value of the parameter that are 1.5, -1 and 1 respectively.

Table 4.5: Empirical mean and root mean square error (RMSE) of β_1, β_2 and the frailty variance for GE distributed frailty data with sample size $m = 500$. Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is also 1.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.311	0.189	−0.876	0.124	0.655	0.345
Semiparametric GE	-	1.394	0.106	−0.928	0.072	0.771	0.229
Semiparametric - IG	$k = 5$	1.397	0.103	−0.925	0.075	3.683	2.683
	$k = 10$	1.382	0.118	−0.921	0.079	1.847	0.847
Semiparametric - RIG	$k = 5$	1.417	0.083	−0.936	0.064	1.159	0.159
	$k = 10$	1.479	0.021	−0.982	0.018	1.259	0.259
Semiparametric - HYP	$k = 5$	1.382	0.118	−0.914	0.086	1.376	0.376
	$k = 10$	1.443	0.057	−0.957	0.043	1.565	0.565
Semiparametric - PHYP	$k = 5$	1.397	0.103	−0.925	0.075	0.892	0.108
	$k = 10$	1.431	0.069	−0.954	0.046	0.910	0.090

smaller bias than the correct model for both fits with $k = 5$ and $k = 10$.

In the bloxplots presented in Figure 4.8, we can see that the estimates of β_1 and β_2 are concentrated close to the true values in all models under investigation with exception of the gamma one. However, the most striking difference between the models is obviously in the estimation of the frailty variance, that varies greatly among them. The gamma, GE and PHYP models show the lowest variabilities, while the largest comes from the IG model that also has the largest RMSE in this scenario. The boxplots confirm that the PHYP and RIG frailties were responsible for the best estimates of the frailty variance under GE generated data.

Once more, increasing sample size did not reduce much the RMSE under the gamma and GE models. For the GIG frailty models, an increase in RMSE was observed again when the number of cut points is equal to 5. The IG case was the most sensitive to this, since the average estimation of the frailty variance

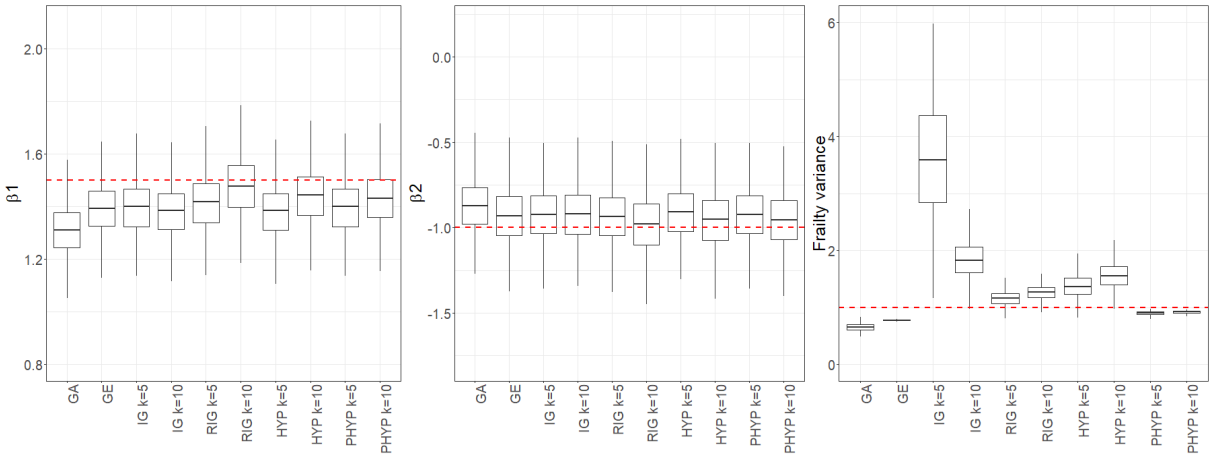


Figure 4.9: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a generalized exponential distribution with $m = 500$. The horizontal dashed line indicates the real value of the parameter.

differs significantly between the two values of k tested. This supports our previous assertion that 10 change points seems to be more appropriate when $m = 500$.

Just as when the data were generated from a model with gamma frailty, the PHYP model had the smallest RMSE in the estimation of the frailty variance. It consists on the smaller RMSE among all the models under investigation, including the correctly specified one. With this, we conclude that the PHYP model is quite competitive under misspecification.

In Table 4.6 we present the simulation study when the frailty follows a log-normal distribution and the total sample size is equal to 400. Figure 4.10 contains the corresponding boxplots. We highlight that unlike the gamma and GE models, where $\alpha = 1$ implies in a frailty variance of 1, in the log-normal frailty when α is set to 1 the variance is approximately 1.718. We refer again to Barreto-Souza & Mayrink [2019], that noted in their work that the gamma and GE models considerably underestimate the frailty variance when the the frailty distribution is log-normal. Here we can see that poorest performance is due to the gamma model. Not only it displays the highest bias in the estimation on the frailty variance, but β_1 and β_2 also show RMSEs that are considerably higher than the competing models.

In this scenario, the model that best estimated the frailty variance was the inverse Gaussian frailty model. With respect to this quantity, when $k = 10$, this model presents a significantly lower RMSE than all competing models. This also shows in the boxplots of the estimates given in Figure 4.10. The gamma case, which is the most commonly used in practice, estimates very poorly the true frailty distribution is log-normal.

Table 4.7 presents the results of the simulation study with log-normal frailty data when total sample size is 1000. The boxplots in Figure 4.11 illustrate these results where, as expected, the variability exhibited by the graphs reduces as sample size increases.

We can see that gamma model is not competitive in this scenario, displaying the highest RMSEs in

Table 4.6: Empirical mean and root mean square error (RMSE) of β_1, β_2 and the frailty variance for LN distributed frailty data with sample size $m = 200$. Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is approximately 1.718.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.335	0.165	−0.887	0.113	0.442	1.276
Semiparametric GE	-	1.493	0.007	−0.988	0.012	0.739	0.979
Semiparametric - IG	$k = 5$	1.400	0.100	−0.926	0.074	1.124	0.594
	$k = 10$	1.464	0.036	−0.971	0.029	1.421	0.297
Semiparametric - RIG	$k = 5$	1.407	0.093	−0.928	0.072	0.782	0.936
	$k = 10$	1.466	0.034	−0.972	0.028	0.884	0.834
Semiparametric - HYP	$k = 5$	1.413	0.087	−0.933	0.067	0.947	0.771
	$k = 10$	1.478	0.022	−0.979	0.021	1.121	0.597
Semiparametric - PHYP	$k = 5$	1.390	0.110	−0.918	0.082	0.648	1.070
	$k = 10$	1.439	0.061	−0.955	0.045	0.704	1.014

Table 4.7: Empirical mean and root mean square error (RMSE) of β_1. β_2 and the frailty variance for LN distributed frailty data with sample size $m = 500$. Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is approximately 1.718.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.332	0.168	−0.891	0.109	0.454	1.264
Semiparametric GE	-	1.485	0.015	−0.988	0.012	0.741	0.977
Semiparametric - IG	$k = 5$	1.385	0.115	−0.921	0.079	1.080	0.638
	$k = 10$	1.451	0.049	−0.967	0.033	1.366	0.352
Semiparametric - RIG	$k = 5$	1.390	0.110	−0.922	0.078	0.774	0.944
	$k = 10$	1.451	0.049	−0.965	0.035	0.880	0.837
Semiparametric - HYP	$k = 5$	1.397	0.103	−0.928	0.072	0.926	0.792
	$k = 10$	1.463	0.037	−0.974	0.026	1.100	0.618
Semiparametric - PHYP	$k = 5$	1.377	0.123	−0.914	0.086	0.652	1.066
	$k = 10$	1.428	0.072	−0.952	0.048	0.713	1.005

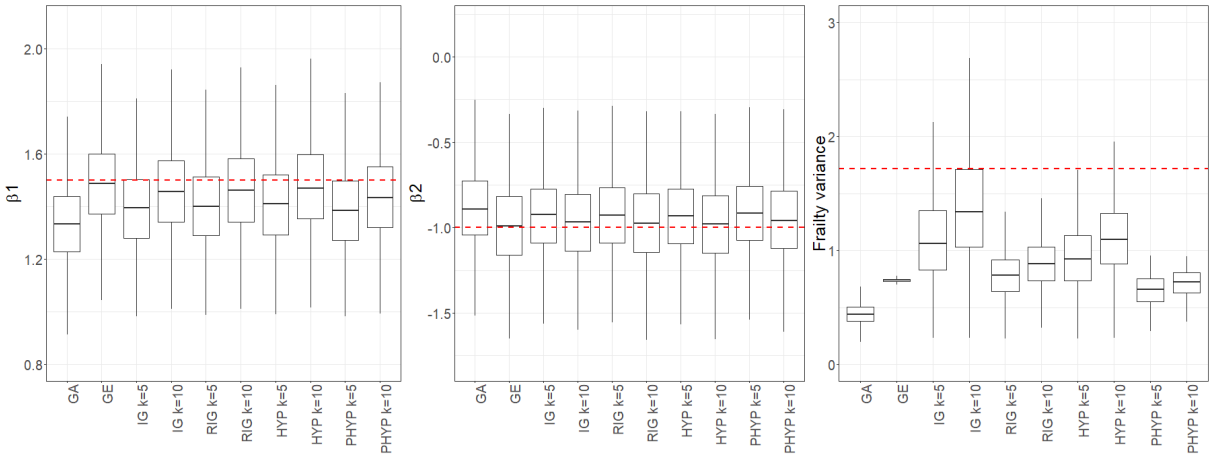


Figure 4.10: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a log-normal distribution with $m = 200$. The horizontal dashed line indicates the real value of the parameter.

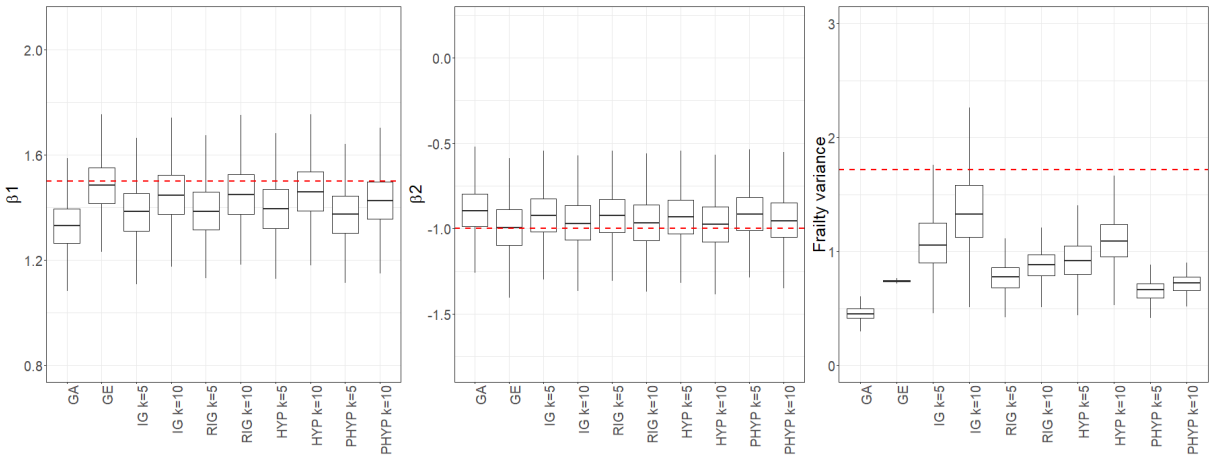


Figure 4.11: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a log-normal distribution with $m = 500$. The horizontal dashed line indicates the real value of the parameter.

the estimation of all quantities compared. Furthermore, the IG frailty continued to show great advantage in the estimation of the frailty variance comparing to the other models when using 10 change points.

Our last scenario consists on data generated with inverse Gaussian frailty, one of our particular cases. We have two main goals here, (i): analyzing what happens to the other particular cases under misspecification of λ and (ii): exploring how the gamma frailty model behaves under larger cluster sizes. For this task, we will consider the total sample sizes of 400 and 1000 as before, but $n_i = 10$ for all i, hence $m = 20$ and $m = 100$. The fit of the GE frailty will not be presented as this model is most suitable for small clusters.

Table 4.8 contains the mean and RMSE of the estimates obtained in 1000 Monte Carlo replicas. We can see that, as expected, the true model estimates well the fixed effects and the frailty variance, unlike what happened with the gamma model. Comparing among the particular cases of the GIG frailty, we see that the choice of λ does not affect the estimation of the covariate effects, it only interferes in the frailty

Table 4.8: Empirical mean and root mean square error (RMSE) of β_1, β_2 and the frailty variance for IG distributed frailty data with total sample size equal to 400 ($m = 20$ with $n_i = 10$ for all i). Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is also 1.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.471	0.029	−0.980	0.020	0.552	0.448
Semiparametric - IG	$k = 5$	1.450	0.050	−0.962	0.038	0.878	0.122
	$k = 10$	1.487	0.014	−0.992	0.022	0.978	0.022
Semiparametric - RIG	$k = 5$	1.449	0.051	−0.961	0.039	0.657	0.343
	$k = 10$	1.484	0.016	−0.990	0.010	0.702	0.304
Semiparametric - HYP	$k = 5$	1.452	0.048	−0.962	0.038	0.755	0.245
	$k = 10$	1.488	0.013	−0.992	0.008	0.822	0.182
Semiparametric - PHYP	$k = 5$	1.444	0.056	−0.958	0.042	0.573	0.427
	$k = 10$	1.478	0.023	−0.986	0.015	0.603	0.406

variance. In relation to this quantity, the model that presents smaller RMSE other than the true model, is the HYP frailty. This behavior is also antecipated since the λ value corresponding to this special case is zero, being the closest to the real λ (-0.5: Inverse-Gaussian) of three models where λ is misspecificated. We observe that the further from the true value of λ, the greater becomes the underestimation of the frailty variance. Even so, it is noted that all GIG frailties estimated this quantity with smaller RMSEs than the gamma model.

Figure 4.12 contains the boxplots of the estimates obtained in the simulation study where the true frailty is inverse Gaussian frailty and $n = 400$. We can see that, for all the models tested, the boxplots of β_1 and β_2 are concentrated close to the true values. The difference between the fits lies in the estimation of the frailty variance, where the boxplots referring to the gamma and PHYP models are those that are farthest from the true value. In the correct specified model, we can note that the increase in the number of cut points improved the estimation of the frailty variance.

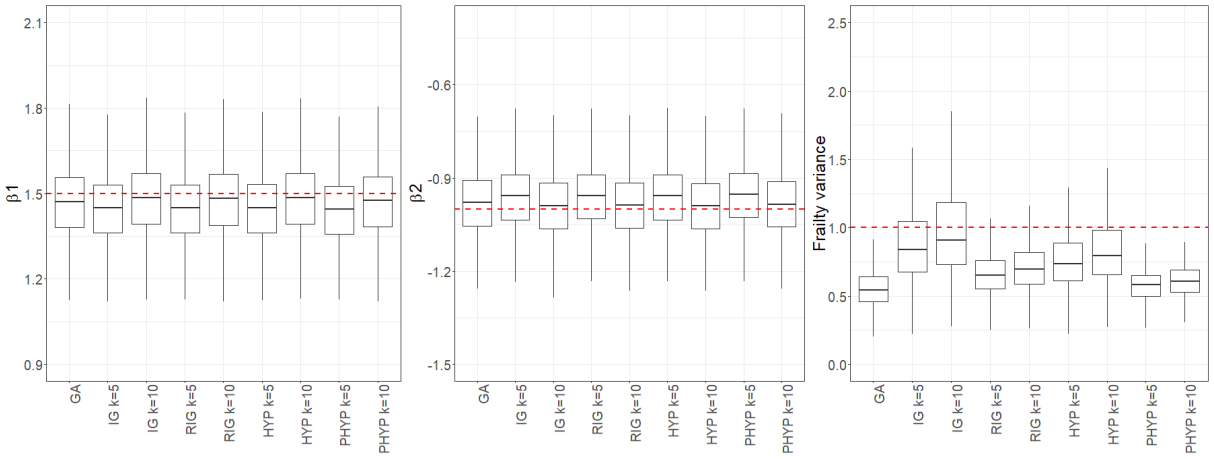


Figure 4.12: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a inverse Gaussian distribution with $m = 20$ and $n_i = 10$ for all i. The horizontal dashed line indicates the real value of the parameter.

Table 4.9: Empirical mean and root mean square error (RMSE) of β_1, β_2 and the frailty variance for IG distributed frailty data with total sample size equal to 1000 ($m = 100$ with $n_i = 10$ for all i). Rows represent the fitted model. The real values of the parameters are $\beta_1 = 1.5$, $\beta_2 = -1$ and $\alpha = 1$ so real frailty variance is also 1.

		β_1		β_2		**Var**	
Model	**Cut points**	**Mean**	**RMSE**	**Mean**	**RMSE**	**Mean**	**RMSE**
Semiparametric Gamma	-	1.466	0.034	−0.980	0.020	0.557	0.443
Semiparametric - IG	$k = 5$	1.438	0.062	−0.958	0.042	0.867	0.133
	$k = 10$	1.480	0.020	−0.987	0.013	0.957	0.043
Semiparametric - RIG	$k = 5$	1.436	0.064	−0.956	0.044	0.658	0.342
	$k = 10$	1.477	0.023	−0.985	0.015	0.702	0.298
Semiparametric - HYP	$k = 5$	1.439	0.061	−0.958	0.042	0.753	0.247
	$k = 10$	1.480	0.020	−0.987	0.013	0.815	0.185
Semiparametric - PHYP	$k = 5$	1.432	0.068	−0.953	0.047	0.579	0.421
	$k = 10$	1.471	0.029	−0.981	0.019	0.608	0.392

Table 4.9 contains the results of the same simulation study when the total sample size is 1000. As expected, all models continue to estimate well the fixed effects. The gamma model seemns to estimate better the fixed effects under larger cluster sizes, but it would be necessary to increase the size of clusters under the same frailty distribution in order to make this conclusion. It is noted that increasing sample size did not help decreasing bias of the frailty variance estimation in any case. The boxplots of the estimates in Figure 4.13 show that the increase in sample size caused the variability of the estimates to decrease, something that we can conclude by comparing Figures 4.12 and 4.13, that are on the same scale.

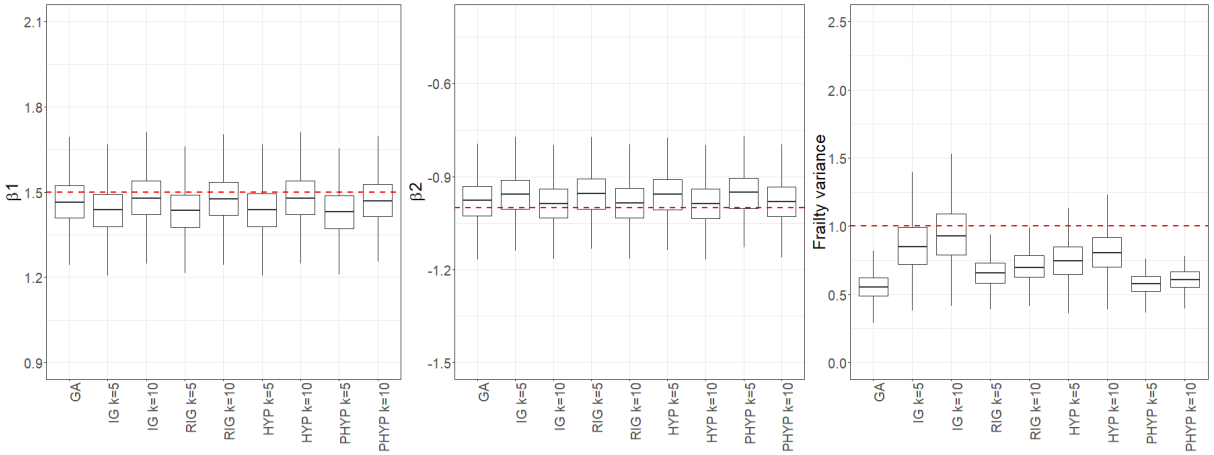


Figure 4.13: Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance when data is generated from a inverse Gaussian distribution with $m = 100$ and $n_i = 10$ for all i. The horizontal dashed line indicates the real value of the parameter.

4.2.1 Conclusion of Simulation Studies

In this section we presented simulations under the parametric and semiparametric approaches, exploring the behavior of the proposed model under the correct model specification and under misspecification.

In Subsection 4.1, under a Weibull baseline hazard function, we explored the behavior of the estimates of the parameter set $\theta = (\beta_1, \beta_2, \sigma, \gamma, \alpha)^\top$ under the four particular cases of the GIG model. We concluded that all IG, RIG, HYP and PHYP frailties estimate well the complete set of parameters under correct model specification. In Subsection 4.2, using piecewise constant hazards, we compared the semiparametric version of the proposed model with the semiparametric versions available in the literature of the gamma and GE frailty models. Through the simulation studies shown in this section, we were able to evaluate the behavior of the GIG frailty model under model misspecification since we fit it to gamma, GE and log-normal generated data. The proposed model proved to be robust under misspecification as in the gamma and GE scenarios there was a particular case that performed better than the correct specified model. It was also possible to obtain a particular case that performed satisfactorily when the real frailty was log-normal, scenario in which the competitive models did not yield good results. At last, we explored the fit of the GIG special cases under misspecification of λ concluding that, although it affects the estimatation of the frailty variance, it does not influence the estimation of the fixed effects.

We also conclude that for all GIG frailties in all scenarios, the increase in the number of change points implied an increase in the magnitude of the estimates of β_1, β_2 and the frailty variance. This meant an improvement in the estimates of the fixed effects but this did not always imply an improvement in the estimation of the frailty variance, since an increase in bias was observed in some cases. A key point here is that the estimation of the fixed effects is not as affected by the choice of the model as is the frailty variance. However, in frailty models it is crucial to obtain reliable estimates of the former quantity since it represents either the degree of non-observed heterogeneity or cluster association, something that is important to measure in practical problems.

We complete by highlighting that, as expected, there is no single model that is appropriate in all situations. It was observed that the model that performed best when the data was generated with gamma and GE frailties differed from the one that returned the best estimates under log-normal frailty data. In the first two scenarios, we pointed advantages towards the PHYP model, while in the latter, the IG model proved to be the most competitive. Thus, it seems advantageous to have in hand the possibility of easily fitting different frailty distributions to the same problem, something that is offered by the proposed model. The frailties obtained by fitting the special cases of the GIG distribution showed different behavior in this simulation study. This evidences that they can capture distinct dependence structures that were not well modeled by fitting the competing models. We emphasize that in the gamma, GE and log-normal scenarios there was one of the special cases of the GIG frailty that performed better than the correctly specified model, evidencing its robustness. In addition, in the last scenario considered, the model also showed excelent performance when correctly specified, something that is not reached by the gamma model.

CHAPTER 5

APPLICATIONS

In this section our goal is to illustrate the utility of the proposed model with the application to real problems and to compare the estimates obtained using different frailty models avaliable. We will present the fit of the four special cases of the GIG frailty model as well as the gamma and generalized exponential (denoted by GE) frailty models under the parametric and semiparametric approaches.

In the parametric approach we will work with a Weibull distributed baseline hazard function with the parameterization previously mentioned. As for the semiparametric approach, the gamma and GE models used here are based on the Cox partial likelihood function, while the GIG frailty model is based on the piecewise exponential distribution. In this aspect, we chose to work with equally spaced intervals and an specification of the number of change points so that we have a reasonable amount of information to estimate the failure rate within each interval. All cases are based on an EM algorithm, where initial guesses were given equivalently for all fits as previously described in Algorithm 1. Also, the convergence parameter ϵ is set to 10^{-5} in all models.

5.1 TARGET-Neuroblastoma Clinical Data

Our first application consists on data obtained from the TARGET (Therapeutic Applicable Research to Generate Effective Treatments) initiative that aims to determine the genetic factors that lead to the emergence and progression hard-to-treat cancers in children. More specifically, we will work with clinical data from the rare cancer named neuroblastoma where our response variable is the time (in months) from diagnosis to the last follow-up or death of the patient. The neuroblastoma is a type of cancer that originates in primitive forms of the nerve cells of the sympathetic nervous system. Most neuroblastomas develop in the adrenal glands, which lie on top of the kidneys. We chose to work with the subset of patients classified as high risk consisting of 315 observations (where 178 events are recorded and 137 are censored) and investigate the effect of two covariates on the lifetime of the subjects. These covariates were selected after a preliminary analysis and are: (i) MYCN Status, corresponds to MYCN gene amplification status and categorizes tumors as amplified or non-amplified, and (ii): Ploidy (DNA Ploidy Analysis by Flow Cytometry Result Category) this is a categorical value based on the DNA content of the tumor cell

population compared to normal diploid cells based on flow cytometry. The categories of covariate Ploidy are diploid or hyperploid.

Our preliminary analysis showed that the Cox model was not adequate in this application. This can be early identified since it is observed that the curves in the Kaplan-Meier estimation of $S(t)$ are not proportional over time, as shown in Figure 5.1. This occurs for the levels of both covariates MYCN and Ploidy. We apply a frailty model since it can deal with the non-proportionality seen in the distance between the curves. We consider one individual per cluster, so that the frailty represents non-observed risk factors. The inclusion of the frailty will allow us to correctly evaluate the covariate effects in the group of high risk patients.

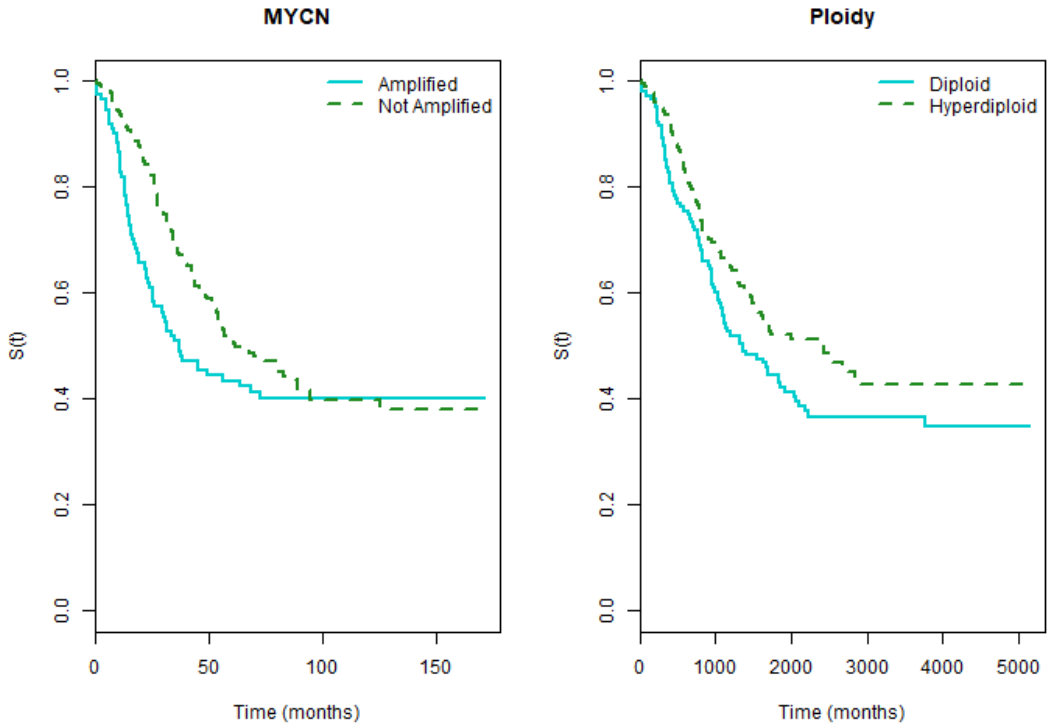


Figure 5.1: Kaplan-Meier estimates of the survival function of the variables MYCN and Ploidy for the TARGET Neuroblastoma clinical data set.

In Table 5.1 we report the estimates obtained under the parametric approach using a Weibull baseline hazard function where rows correspond to the fit of the different frailty models. Here the gamma estimates are not reported because this model presented convergence issue, more specifically in the estimation of α. We point out that for this task, both the maximization of the likelihood function that we implemented and the fit made using the `parfm` (Munda et al [2012]) package presented the same problem and did not return the estimates. We report this error below. This application evidences that the problem of estimating the frailty parameter using a gamma frailty model occurs even when considering a parametric assumption of the baseline hazard function, and confirms that this issue, evidently, happens in practical situations.

```
Error in optim(par = theta0, fn = loglikGA, y = y, delta = delta, x = x,  :
```

Table 5.1: Estimates of the parameters and standard errors (in parenthesis) for the TARGET Neuroblastoma data set under the parametric approach considering a Weibull baseline hazard function with parameters σ and γ. Rows correspond to fitting the GIG special cases and the generalized exponential (GE) frailty models.

	β_{MYCN}	β_{ploidy}	Var	α	σ	γ
Par. IG	0.3529 (0.2307)	0.3410 (0.2230)	3.3398	3.3398 (1.9558)	0.0058 (0.0017)	1.2396 (0.1142)
Par. RIG	0.3174 (0.2025)	0.2537 (0.1947)	0.9815	1.5622 (0.5000)	0.0035 (0.0006)	1.0527 (0.0439)
Par. HYP	0.3705 (0.2240)	0.3418 (0.2156)	1.9696	2.8251 (0.8813)	0.0028 (0.0004)	1.2134 (0.0472)
Par. PHYP	0.2401 (0.1874)	0.2460 (0.1789)	0.5706	0.8545 (0.3442)	0.0050 (0.0013)	0.9578 (0.0475)
Par. GE	0.9767 (0.3149)	0.3962 (0.2990)	3.2384	0.2836 (0.0517)	0.0031 (0.0004)	1.4721 (0.1252)

Table 5.2: Estimates of the parameters and standard errors in parenthesis for the TARGET Neuroblastoma data application under the semiparametric approach. Rows correspond to fitting the GIG special cases and the generalized exponential (GE) frailty models.

	β_{MYCN}	β_{ploidy}	Var	α
Semi. IG (k=5)	0.3834 (0.2451)	0.3467 (0.2449)	1.9579	1.9579 (0.8748)
Semi. IG (k=10)	0.4142 (0.2465)	0.3529 (0.2465)	2.1382	2.1382 (0.9661)
Semi. RIG (k=5)	0.3258 (0.2296)	0.2941 (0.2296)	0.7795	1.0604 (0.3821)
Semi. RIG (k=10)	0.3505 (0.2374)	0.3201 (0.2374)	0.6257	0.7724 (0.1391)
Semi. HYP (k=5)	0.3812 (0.2533)	0.3102 (0.2533)	2.1659	3.2633 (1.7049)
Semi. HYP (k=10)	0.4382 (0.2576)	0.3420 (0.2576)	1.8615	2.6003 (0.9407)
Semi. PHYP (k=5)	0.2919 (0.2196)	0.2793 (0.2196)	0.3603	0.4255 (0.0459)
Semi. PHYP (k=10)	0.3092 (0.2229)	0.2874 (0.2229)	0.3719	0.4436 (0.0306)
Semi. GE	0.4799 (0.2130)	0.3230 (0.2001)	0.5263	1.2049 (0.1316)

```
non-finite finite-difference value [3]
```

Comparing the GIG fits to the GE fit, we can notice significant differences. The effect of β_{mycn} is much larger under the GE frailty model than in all other models considered. The frailty variance and $\widehat{\beta}_{ploidy}$ estimated by the GE model resemble those of the particular case IG, these being the two models under which the frailty variance has larger magnitude.

Table 5.2 contains the fits of the semiparametric versions of the GIG and GE frailty models to the TARGET Neuroblastoma clinical data. The estimates of the penalized version of the semiparametric gamma frailty model available in the R package `survival` are not reported because this model also presented convergence problems, that we report below.

```
Warning message: In coxpenal.fit(X, Y, strats, offset, init = init, control,
weights = weights,  : Inner loop failed to converge for iterations 4 5
```

We fit the semiparametric GIG frailty models considering 5 and 10 cut points in the piecewise constant hazards and estimate the standard errors by running 1000 bootstrap replicas. In Table 5.2 we can note that $\widehat{\beta}_{mycn}$ in the semiparametric GE model differed significantly from the one obtained under the parametric approach (0.4799 and 0.9767 respectively), whereas the magnitude of the difference between the parametric and semiparametric approaches of the GIG frailties were smaller. The same comment applies to the frailty variance, that in the GE model went from 3.2384 to 0.5263 which is the largest

Table 5.3: Estimates of the parameters and standard errors (in parenthesis) for the kidney catheter data set under the parametric approach considering a Weibull baseline hazard function with parameters σ and γ. Rows correspond to fitting the GIG special cases, generalized exponential (GE) and gamma frailty models.

	β_{age}	β_{sex}	Var	α	σ	γ
Par. IG	0.0053 (0.0117)	−1.4786 (0.4306)	0.6752	0.6752 (0.5340)	0.0136 (0.0100)	1.1454 (0.1418)
Par. RIG	0.0043 (0.0113)	−1.6308 (0.4880)	0.6398	0.7962 (0.6153)	0.0084 (0.0060)	1.1624 (0.1358)
Par. HYP	0.0050 (0.0116)	−1.5542 (0.4571)	0.6754	0.7309 (0.5628)	0.0106 (0.0080)	1.1571 (0.1430)
Par. PHYP	0.0042 (0.0111)	−1.7173 (0.5157)	0.6178	0.9959 (0.7646)	0.0056 (0.0032)	1.1797 (0.1164)
Par. Gamma	0.0071 (0.0124)	−1.9116 (0.5394)	0.5102	0.5102 (0.2572)	0.0129 (0.0105)	1.2155 (0.1591)
Par. GE	0.0067 (0.0117)	−1.9621 (0.5424)	0.5999	1.8197 (0.8800)	0.0093 (0.0059)	1.2285 (0.1498)

variation observed among the competitive models. This suggests robustness of the proposed model on the specification of the baseline hazard function.

5.2 Kidney Catheter Data

The second application refers to the kidney catheter data of McGilchrist & Aisbett [1991] that we present in order to illustrate the multivariate version of the proposed model. This data set is widely used to illustrate shared frailty models and is available in the R `survival` package. The response variable is the time to infection since the insertion of a catheter in a patient who is carrying a portable dialysis equipment. The time to the first and second infections of each patient are recorded. There are 38 individuals constituting 38 clusters of size 2. We have a total of 76 observations in which 18 of them are right-censored. After the first event (or censoring) a cure time of the infection is allowed until the second measurement is made. We will consider in our analysis the covariates gender and age in years.

In Table 5.3, we report the estimates of the parameters under the parametric approach with IG, RIG, HYP, PHYP, gamma and GE frailties. This table contains the estimates of the covariate effects gender and age, as well as the parameters σ, γ and α. The former, in all cases, is the parameter of the frailty distribution. However, it does not correspond directly to the frailty variance in all models. Therefore, they are not directly comparable without a proper transformation. Hence, here we are going to use the same transformation that was discussed in Section 4 that is reported in the columns named "Var". We can see that the estimates of β_{age} and β_{sex} obtained through the gamma and GE models are more similar to each other than to those estimated using GIG frailties. However, the frailty variance estimated by the fit of the GE model is closer to the GIG ones. In the fit of the gamma model, the frailty variance has the smallest magnitude among the models compared.

Estimates for the same data set under the semiparametric approach can be found in Table 5.4. We report the results obtained with the fit of the semiparametric GE model based on the Cox partial likelihood function, and also report the fit of the gamma frailty model available in the `survival` package which relies on the penalized likelihood approach. The standard error of the estimate of α, however, is not available in the package, hence not reported. Here we fit the GIG frailties with 3 and 5 change points in the piecewise constant hazards as the data set contains only 76 observations. Likewise observed in the

Table 5.4: Estimates of the parameters and standard errors (in parenthesis) for the kidney catheter data set under the semiparametric approach. Rows correspond to fitting the GIG special cases, the generalized exponential (GE) and gamma frailty models.

	β_{age}	β_{sex}	Var	α
Semi. IG (k=3)	0.0038 (0.0114)	−1.3387 (0.4122)	0.3893	0.3893 (0.3457)
Semi. IG (k=5)	0.0031 (0.0116)	−1.3817 (0.4209)	0.4306	0.4306 (0.3930)
Semi. RIG (k=3)	0.0043 (0.0119)	−1.5772 (0.4725)	0.4830	0.5532 (0.4957)
Semi. RIG (k=5)	0.0038 (0.0121)	−1.5773 (0.4813)	0.5094	0.5910 (0.5532)
Semi. HYP (k=3)	0.0039 (0.0116)	−1.4282 (0.4329)	0.4246	0.4412 (0.3832)
Semi. HYP (k=5)	0.0033 (0.0118)	−1.4823 (0.4454)	0.4699	0.4916 (0.4376)
Semi. PHYP (k=3)	0.0042 (0.0114)	−1.2617 (0.4483)	0.3765	0.4510 (0.4776)
Semi. PHYP (k=5)	0.0036 (0.0116)	−1.3045 (0.4597)	0.4054	0.4987 (0.5467)
Semi. Pen. Gamma	0.0052 (0.0119)	−1.5875 (0.4605)	0.4120	0.4120 (-)
Semi. GE	0.0067 (0.0134)	−1.8438 (0.4756)	0.7121	1.4805 (0.3177)

simulation studies, increasing the value of k implied in an increase in the magnitude of the fixed effects in all cases. We can see a difference among the particular cases but, in general, they are closer to those obtained by the gamma fit than the GE one. This can be said both with respect to ($\widehat{\beta}_{age}$, $\widehat{\beta}_{sex}$) as well as the frailty variance, which is significantly larger under GE frailty.

Although it is relevant to compare the magnitude of the estimates, none of these values gives us an idea of which model is best fitted to the data. There is not much work developed in verifying model adequacy of frailty models and we would need to find some appropriate method to allows us deciding on the most suitable model.

The real data applications illustrated the practical utility of the proposed model in its univariate and multivariate versions and allowed us to observe some important points. The gamma frailty model, being the most used in practice, presented convergence problems in both the parametric and semiparametric approaches in the first application. This and other problems presented by the gamma frailty, in our view, justify the introduction of more flexible and robust models. In addition, the GIG frailties proved to be robust regarding the specification of the baseline hazard function, since the estimates under the two approaches are consistent, something that did not occur in all models compared.

CHAPTER 6

CONCLUSION

The GIG frailty model proposed in this work proved to be a mathematically tractable, flexible and robust model. The flexibility is obtained through the additional parameter λ which we considered fixed. Particular values of λ correspond to the known particular cases of the GIG distribution. We investigated the performance of the IG, RIG, HYP and PHYP frailties under the correct specification of the model and under misspecification of the frailty distribution. In the simulation studies, we observed that when the data were generated with gamma and GE frailties, it was possible to identify a GIG frailty that performed better than the correctly specified model. In these two scenarios the most competitive frailty was the PHYP special case. When the true frailty distribution is log-normal, the gamma and GE frailty models suffer in the estimation of the frailty variance, showing high bias in the estimation of this quantity. However, it was possible to obtain a satisfactory result with our model, this time under the IG frailty. In addition to being robust under misspecification, the proposed model, as expected, performs well when correctly specified.

Through the TARGET Neuroblastoma application we showed that the gamma frailty model, the most common choice, can be problematic in practical situations. It presented convergence issues under the parametric and semiparametric approaches. The fits were obtained through the popular R (R Core Team [2019]) packages used for this task, that are respectively, the `parfm` (Munda et al [2012]) and `survival` (Therneau [2015]) packages. Another important point that this application allowed us to observe was the robustness of the GIG frailty model regarding the specification of the baseline hazard function. This was evidenced by the consistency of the estimates obtained under the parametric and semiparametric approaches, something that was not observed in all models under comparison. By fitting the GIG frailty models to the kidney catheter data, we were able to illustrate the application of the proposed methodology to clustered survival data.

APPENDIX

Proof of Lemma 1. *: For $k = 1$: It can be shown that the derivative of the Bessel function with respect to its argument is*

$$\frac{\partial}{\partial x} K_\phi(\sqrt{x}) = -\frac{1}{4\sqrt{x}} \left(K_{\phi+1}(\sqrt{x}) + K_{\phi-1}(\sqrt{x}) \right).$$

Using this expression we have that

$$\frac{\partial}{\partial x} \zeta(x) = -\frac{1}{2} \left(\phi \frac{K_\phi(\sqrt{x})}{x^{(\phi/2+1)}} + \frac{K_{\phi+1}(\sqrt{x}) + K_{\phi-1}(\sqrt{x})}{2x^{(\phi+1)/2}} \right).$$

At this point, we apply the recurrence identity on the Bessel function mentioned previously,

$$K_v(z) = \frac{z}{2v} \left(K_{v+1}(z) - K_{v-1}(z) \right), \tag{6.1}$$

and the former derivative simplifies to

$$\frac{\partial}{\partial x} \zeta(x) = -\frac{1}{2} \frac{K_{\phi+1}(\sqrt{x})}{x^{(\phi+1)/2}}.$$

Using this, we continue the demonstration by finding the second derivative of $\zeta(x)$ with respect to x, given by

$$\frac{\partial^2}{\partial x^2} \zeta(x) = -\frac{1}{2} \left(-\frac{K_{\phi+1}(\sqrt{x})(\phi+1)}{2x^{(\phi+3)/2}} - \frac{K_{\phi+2}(\sqrt{x}) + K_\phi(\sqrt{x})}{4x^{(\phi/2+1)}} \right).$$

The second derivative simplifies to the following after using (6.1)

$$\frac{\partial^2}{\partial x^2} \zeta(x) = \frac{1}{4} \frac{K_{\phi+2}(\sqrt{x})}{x^{(\phi+2)/2}}.$$

To conclude the proof by induction, we now assume as true the case $k-1$ and use it to prove the k-th order expression. If this is satisfied, then the result is true for all k. If the $\{k-1\}th$ derivative of $\zeta(x)$ in terms of x is given by

$$\frac{\partial^{k-1}}{\partial x^{k-1}} \zeta(x) = \left(-\frac{1}{2} \right)^{k-1} \frac{K_{\phi+k-1}(\sqrt{x})}{x^{(\phi+k-1)/2}},$$

then the k-th order derivative is

$$\frac{\partial^{k-1}}{\partial x^{k-1}}\zeta(x) = \left(-\frac{1}{2}\right)^{k-1}\left\{\frac{-K_{\phi+k-1}(\sqrt{x})(\phi+k-1)}{2x^{(\phi+k+1)/2}} - \frac{K_{\phi+k}(\sqrt{x})+K_{\phi+k-2}(\sqrt{x})}{4x^{(\phi+k)/2}}\right\}.$$

Here we use again (6.1), that gives us $K_{\phi+k-1}(\sqrt{x}) = \frac{\sqrt{x}}{2(\phi+k-1)}[K_{\phi+k}(\sqrt{x}) - K_{\phi+k-2}(\sqrt{x})]$ *and get that* $\frac{\partial^k}{\partial x^k}\zeta(x)$ *is*

$$\frac{\partial^k}{\partial x^k}\zeta(x) = \left(-\frac{1}{2}\right)^k \frac{K_{\phi+k}(\sqrt{x})}{x^{(\phi+k)/2}}.$$

This completes the proof of Lemma 1. □

BIBLIOGRAPHY

Aalen, O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine.* **7**(11), 1121–1137.

Aalen, O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability.* **2**(4), 951–972.

Balakrishnan, N. & Peng, Y. (2006). Generalized Gamma frailty model. *Statistics in Medicine.* **25**(16), 2797–2816.

Barndorff-Nielsen, O. (1977). Exponentially Decreasing Distributions for the Logarithm of Particle Size. *Proceedings of The Royal Society A.* **353**(1674), 401–419

Barreto-Souza, W. & Mayrink, V.D. (2019). Semiparametric generalized exponential frailty model for clustered survival data. *Annals of the Institute of Statistical Mathematics.* **71**(3), 679–701.

Callegaro, A. & Iacobelli, S. (2012). The Cox shared frailty model with log-skew-normal frailties. *Statistical Modelling.* **12**(5), 399–418.

Clayton, D.(1978). A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika.* **65**(1), 141–151.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B.* **34**(2), 187–220.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B.* **39**(1), 1–38.

Du, P., & Ma, S. (2010). Frailty Model with Spline Estimated Nonparametrix Hazard Function. *Statistica Sinica.* **20**(2), 561–580.

Efron, Bradley. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics.* **7**(1), 1–26.

Farrington, C., Unkel, S. & Anaya-Izquierdo, K. (2012). The relative frailty variance and shared frailty models. *Journal of the Royal Statistical Society. Series B.* **74**(4), 673-696.

Feng S., Wolfe R., Port F. (2005). Frailty survival model analysis of the National Deceased Donor Kidney Transplant Dataset using Poisson variance structures. *Journal of the American Statistical Association.* **100**(471), 728–735.

Fletcher, R. (2000). Practical methods of optimization (2nd ed.). New York: Wiley.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika.* **71**(1), 75–83.

Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika.* **73**(2), 387–396.

Hougaard, P. (1986b). A class of multivariate failure time distributions. *Biometrika.* **73**(3), 671–678.

Hougaard, P. (1995). Frailty Models for Survival Data. *Lifetime Data Analysis.* **1**(3), 255–273.

Jørgensen, B. (1982). Statistical properties of the generalized inverse gaussian distribution. *Lecture Notes in Statistics.* Heidelberg: Springer.

Kim, J. S. & Proschan, F. (1991). Piecewise Exponential Estimation of the Survival Function. *IEEE Translations on Reliability.* **40**(2), 134–139.

Klein, J.P. (1992). Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. *Biometrics.* **48**(3), 795–806.

Lawless JF, Zhan M. (1998). Analysis of interval-grouped recurrent event data using piecewise constant rate function. *Canadian Journal of Statistics.* **26**(4), 549–565.

Liu, L., Huang, X. (2008). The use of Gaussian quadrature for estimation in frailty proportional hazard models. *Statistics in Medicine.* **27**(14), 2665–2683.

McGilchrist, C.A. & Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics.* **47**(2), 461–466.

Munda, M., Rotolo, F. & Legrand, C. (2012). Parametric Frailty Models in R. package version 2.7.6. https://cran.r-project.org/web/packages/parfm. Accessed Jun 2019.

Oakes, D. (1982). A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society. Series B.* **44**(3), 414–422.

Oakes, D. (1986). Semiparametric Inference in a Model for Association in Bivariate Survival Data. *Biometrika.* **73**(2), 353–361.

Palmer, J., Kreutz-Delgado, K. & Makeig, S. (2016). An EM algorithm for maximum likelihood estimation of Barndorff-Nielsen's generalized hyperbolic distribution. *Conference: 2016 IEEE Statistical Signal Processing Workshop (SSP)* **1-4**, 10.1109/SSP.2016.7939245.

R Core Team (2019). R: A language and environment for statistical computing.. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed Jun 2019.

Ripatti, S. & Palmgren, J. (2000). Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood. *Biometrics.* **56**(4), 1016–1022.

Therneau, T.M., Grambsch, P.M. & Pankratz, V.S. (2003). Penalized survival models. *Journal of Computational and Graphical Statistics.* **12**(1), 156–175.

Therneau, T. (2015). A package for survival analysis in S. R package version 2.43.3. https://CRAN.R-project.org/package=survival. Accessed Jun 2019.

Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography.* **16**(3), 439–454.

Wang, H. & Klein, J.P. (2012). Semiparametric Estimation for the Additive Inverse Gaussian Frailty Model. *Communications in Statistics - Theory and Methods.* **41**(12), 2269–2278.

Wienke, A. (2011). Frailty Models in Survival Analysis. New York: Chapman and Hall/CRC Biostatistics Series.

Yashin, A., Vaupel, J.W. & Iachine, I. (1995). Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate Data. *Mathematical Population Studies.* **5**(2), 145–159.