Ann Inst Stat Math (2019) 71:679–701
https://doi.org/10.1007/s10463-018-0658-9

Semiparametric generalized exponential frailty model for clustered survival data

Wagner Barreto-Souza[1] · **Vinícius Diniz Mayrink**[1]

Received: 11 October 2017 / Revised: 10 January 2018 / Published online: 16 March 2018

Abstract In this paper, we propose a novel and mathematically tractable frailty model for clustered survival data by assuming a generalized exponential (GE) distribution for the latent frailty effect. Both parametric and semiparametric versions of the GE frailty model are studied with main focus for the semiparametric case, where an EM-algorithm is proposed. Our EM-based estimation for the GE frailty model is simpler, faster and immune to a flat likelihood issue affecting, for example, the semiparametric gamma model, as illustrated in this paper through simulated and real data. We also show that the GE model is at least competitive with respect to the gamma frailty model under misspecification. A broad analysis is developed, with simulation results explored via Monte Carlo replications, to evaluate and compare models. A real application using a clustered kidney catheter data is considered to demonstrate the potential for practice of the GE frailty model.

Keywords Censored data · EM-algorithm · Flat likelihood · Gamma frailty model · Partial likelihood · Proportional hazards

1 Introduction

The semiparametric proportional hazards model by Cox (1972) is a widely used approach to deal with lifetime data. However, this model is not suitable for situations where the observations are correlated, which may occur when clusters or groups can be identified in the data configuration, or even when an important covariate

✉ Wagner Barreto-Souza
wagnerbs85@gmail.com; wagnerbs@est.ufmg.br

[1] Departamento de Estatística, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG 31270-901, Brazil

(possibly non-observed) is not included in the model. In these cases, the Cox model may underestimate the fixed covariate effects.

One way to handle this issue is to consider extensions of the Cox model by incorporating a latent random effect in the proportional hazards structure; this latent term is well known in the literature as the frailty. The study by Vaupel et al. (1979) is one of the first ones proposing a frailty model; in this case, the authors choose to work with the gamma distribution. Other important papers about the gamma frailty modeling are Nielsen et al. (1992), Klein (1992), Andersen et al. (1997), Parner (1998) and Therneau et al. (2003), where the first two propose an EM-algorithm for the semiparametric case. A modified version of this EM-algorithm is presented in Yu (2006) for situations involving a large number of clusters (groups) and distinct event times.

Many other models, based on different frailty distributions, have emerged in the literature as alternatives to the gamma model such as the log-normal (McGilchrist and Aisbett 1991; McGilchrist 1993), inverse-Gaussian (Hougaard 1984), positive stable (Hougaard 1986), power variance family (Crowder 1989; Hougaard et al. 1992) and compound Poisson (Aalen 1992) frailties; this last reference also deals with the cure fraction problem. More recently, Balakrishnan and Peng (2006) and Callegaro and Iacobelli (2012) have introduced the generalized gamma and log-skew normal frailty models, respectively. For an overview on frailty models, we suggest Hougaard (2000), Duchateau and Janssen (2008) and Wienke (2011). More recent papers on this topic are Ha et al. (2014), Enki et al. (2014), Christian et al. (2016) and Yavuz and Lambert (2016), just to name a few.

According to Hougaard (2000) there is not a unique family of frailty models having all desirable properties for inference. Choosing a model requires a detailed investigation of model properties for each distribution family, and this choice depends on the context of the problem related to the application. The author suggests that one should focus on finding the right tools for a given problem rather than using a single tool for all applications.

This is the spirit motivating the study developed in the present paper. The main aim here is to propose a novel frailty model to be considered by practitioners as an additional option for their survival data analyses exploring and comparing different well-known frailty distributions. In our proposed model, we assume a generalized exponential (GE) distribution for the frailty component; some references with details about the GE distribution are Gupta et al. (1998), Gupta and Kundu (1999) and Gupta and Kundu (2001). The GE distribution has been an interesting alternative to existing and well-established survival models in the literature such as gamma, Weibull, log-normal and inverse-Gaussian (IG) distributions, to name a few. As it will be shown along the paper, assuming a GE frailty determines a mathematically tractable and computationally appealing model; which in turn configures an alternative and strong competitor to other known frailty models.

The gamma frailty is perhaps the most popular model due to its analytical tractability. If we restrict our attention to the subclass of semiparametric frailty models, we can basically find two options: the gamma and the log-normal; the second one being computationally expensive since it is not tractable as the gamma case. The semiparametric GE frailty model arises as another alternative, challenging the gamma frailty model in many aspects; this is an important motivating point in this paper.

We summarize below the main contributions and advantages of the GE frailty over the popular gamma frailty model with respect to mathematical tractability and computational features.

- Mathematical tractability:
  - The GE frailty model also has a simple and explicit form for the likelihood function under the parametric and semiparametric approaches.
  - The conditional distributions of the frailty given the censoring indicator (distribution of frailties among the survivors and the frailty of individuals dying at time t) can also be determined explicitly.
  - As it will be shown later, the semiparametric gamma frailty model can suffer with a flat likelihood and this issue is not present in the GE case. We have observed this on both simulated and real data; see Figs. 3 and 7.
- Computational advantages of the semiparametric GE model:
  - It also provides a simple EM-algorithm, since all conditional expectations involved in the E-step are obtained in explicit form.
  - The estimate of the frailty parameter in each loop (M-step) of the EM-algorithm has a closed form, contrasting with a maximization procedure used in the semiparametric gamma model. Therefore, our model has lower computational cost.
  - In simulation studies, for both parametric and semiparametric configurations, we show that the GE model is at least competitive with respect to the gamma under misspecification.

The paper is organized in the following manner. Section 2 introduces some notations and the parametric GE frailty model. The semiparametric GE frailty (main focus of the paper) is presented in Sect. 3; which also discusses the estimation of parameters based on the partial likelihood, the EM-algorithm and how to get the estimates standard errors. In Sect. 4 we study the finite-sample behavior of the estimators for the GE model with respect to the gamma case under misspecification through a Monte Carlo simulation. Section 5 investigates the performance of the GE frailty model in a real data analysis. Concluding remarks and possible points for future research are presented in Sect. 6.

2 Model specification

In this section, we introduce the generalized exponential (in short GE) frailty modeling. The presentation begins with a short description of the GE distribution stating important results for our purposes.

Let Z be a random variable having the GE distribution, with scale and shape parameters $\gamma > 0$ and $\alpha > 0$, respectively. We denote $Z \sim \text{GE}(\gamma, \alpha)$ and the corresponding density function takes the form $f(z) = \gamma\alpha e^{-\gamma z}(1 - e^{-\gamma z})^{\alpha-1}$, for $z > 0$. For a standard GE random variable Z, that is assuming $\gamma = 1$, we denote $Z \sim \text{GE}(\alpha)$.

The associated Laplace transform $L(s) = E(e^{-sZ})$ is given by

$$L(s) = \frac{\Gamma(\alpha + 1)\Gamma(s/\gamma + 1)}{\Gamma(\alpha + s/\gamma + 1)}, \quad s > 0, \tag{1}$$

where $\Gamma(\tau) = \int_0^\infty x^{\tau-1}e^{-x}dx$, for $\tau > 0$, is the gamma function. The first two cumulants of Z are $E(Z) = \gamma^{-1}\{\Psi(\alpha+1) - \Psi(1)\}$ and $\text{Var}(Z) = \gamma^{-2}\{\Psi'(1) - \Psi'(\alpha+1)\}$, with $\Psi(\tau) = d\log\Gamma(\tau)/d\tau$ and $\Psi'(\tau) = d\Psi(\tau)/d\tau$ for $\tau > 0$ being the digamma and trigamma function, respectively.

The next step is to introduce the frailty model without covariates; the regression structure is discussed later. At this point, consider a subject with lifetime denoted by a random variable T whose hazard function, conditional on a latent variable Z, satisfies $\lambda(t|Z) = Z\lambda_0(t)$ for $t > 0$, where $\lambda_0(t)$ is a baseline hazard function and Z is the frailty of the individual. In this paper, we assume that Z follows a GE distribution. In order to avoid non-identifiability issues, we set $\gamma = 1$, i.e., $Z \sim \text{GE}(\alpha)$. The parameter α controls the heterogeneity.

Using basic results on frailty models and the expression in (1), one can determine that the marginal survival $S(\cdot)$ and density function $f(\cdot)$ of T are, respectively,

$$S(t) = L(\Lambda_0(t)) = \frac{\Gamma(\alpha+1)\Gamma(\Lambda_0(t)+1)}{\Gamma(\alpha+\Lambda_0(t)+1)}$$

and

$$\begin{aligned} f(t) &= -\lambda_0(t)L'(\Lambda_0(t)) \\ &= \frac{\Gamma(\alpha+1)\lambda_0(t)\Gamma(\Lambda_0(t)+1)}{\Gamma(\alpha+\Lambda_0(t)+1)}\{\Psi(\alpha+\Lambda_0(t)+1) - \Psi(\Lambda_0(t)+1)\}, \end{aligned}$$

for $t > 0$, where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the cumulative hazard function, $L'(s) = dL(s)/ds$ and $\Psi(s)$ is the digamma function defined previously.

The hazard function of T can then be written as $\lambda(t) = \lambda_0(t)\{\Psi(\alpha+\Lambda_0(t)+1) - \Psi(\Lambda_0(t)+1)\}$, $t > 0$.

In the next step, we find the frailty distribution among the survivors and the deaths (or failures) at time t; these results are important for the implementation of the EM-algorithm. We begin by finding the conditional distribution of Z given $T > t$, that is the distribution of the frailty among the survivors at time t. The conditional density of $Z|T > t$ [for instance, see Eq. (3.8) from Wienke (2011)] is

$$\begin{aligned} f(z|T > t) &= \frac{f(z)S(t|z)}{S(t)} \\ &= \frac{1}{B(\alpha, 1+\Lambda_0(t))}\exp\left(-(\Lambda_0(t)+1)z\right)(1-\exp(-z))^{\alpha-1}, \quad z > 0, \end{aligned}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function, with $a, b > 0$. This density is related to the beta exponential (BE) distribution in Nadarajah and Kotz (2006). A random variable Y has a BE distribution, denoted by $Y \sim \text{BE}(\gamma, \beta, \alpha)$, if its density assumes the form $g(y) = \frac{\gamma}{B(\alpha,\beta)}\exp(-\gamma\beta y)(1-\exp(-\gamma y))^{\alpha-1}$, for $y > 0$, where $\gamma > 0$ is a scale parameter and $\beta, \alpha > 0$ are shape parameters. This provides $Z|T > t \sim \text{BE}(1, \Lambda_0(t)+1, \alpha)$. In particular, using the results of the

BE distribution given in Nadarajah and Kotz (2006), we have that $E(Z|T > t) = \Psi(\alpha + \Lambda_0(t) + 1) - \Psi(\Lambda_0(t) + 1)$.

Now, we find the distribution of the frailty given a failure at time t, that is the conditional distribution of $Z|T = t$. Using basic probability, we have that the conditional density of $Z|T = t$, denoted by $f(z|t)$, is given by

$$f(z|t) = \frac{z\Gamma(\alpha + \Lambda_0(t) + 1)}{\Gamma(\alpha)\Gamma(\Lambda_0(t) + 1)} \frac{\exp(-(\Lambda_0(t) + 1)z)\,(1 - \exp(-z))^{\alpha-1}}{\Psi(\alpha + \Lambda_0(t) + 1) - \Psi(\Lambda_0(t) + 1)}, \quad z > 0.$$

It can be shown that the conditional mean is

$$E(Z|T = t) = \Psi(\alpha + \Lambda_0(t) + 1) - \Psi(\Lambda_0(t) + 1) + \frac{\Psi'(\Lambda_0(t) + 1) - \Psi'(\alpha + \Lambda_0(t) + 1)}{\Psi(\alpha + \Lambda_0(t) + 1) - \Psi(\Lambda_0(t) + 1)},$$

where $\Psi'(s)$ is the trigamma function defined previously.

The next discussion explains how to obtain the log-likelihood function of the GE frailty model including a regression structure. Consider m clusters (or groups) with the i-th group containing n_i individuals, for $i = 1, \ldots, m$. The total sample size is $n = \sum_{i=1}^{m} n_i$. Let T_{ij}^0 and C_{ij} be the failure and censoring times for the individual (i, j) and $\mathbf{x}_{ij}$ be a $p \times 1$ associated covariate vector, for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. The random variable $T_{ij} = \min\{T_{ij}^0, C_{ij}\}$ is observable and $\delta_{ij} = I\{T_{ij}^0 \leq C_{ij}\}$ is the failure indicator; more generally, $I\{A\} = 1$ if the event A occurs (0 otherwise).

The i-th cluster is associated with a random variable Z_i, representing the frailty of that cluster, which induces dependence among the members. Let $Z_1, \ldots, Z_m$ be i.i.d. positive random variables with $Z_i \sim \text{GE}(\alpha)$. Two final assumptions are required to complete the model specification according Nielsen et al. (1992). The first one is that given Z_i, $\{(T_{ij}^0, C_{ij}),\ j = 1, \ldots, n_i\}$ are conditionally independent and both T_{ij}^0 and C_{ij} are independent, for $j = 1, \ldots, n_i$. The second assumption is that the censoring times within the cluster $\{C_{ij},\ j = 1, \ldots, n_i\}$ are non-informative about Z_i. Further, given Z_i, the conditional hazard function of T_{ij}^0 takes the form

$$\lambda(t_{ij}|Z_i) = Z_i\,\lambda_0(t_{ij})\,\exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta}), \quad t_{ij} > 0, \tag{2}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Here, $\lambda_0(\cdot)$ is a baseline hazard function as before and $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector associated to the covariates.

Assuming the structure (2) and using first equation of page 138 from Wienke (2011), we obtain that the joint survival function of $T_{i1}, \ldots, T_{in_i}$ is given by

$$S(t_{i1}, \ldots, t_{in_i}) = \Gamma(\alpha + 1)\zeta\left(\sum_{j=1}^{n_i} \Lambda_0(t_{ij}) e^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}} + 1, \alpha\right), \quad t_{ij} > 0, \quad j = 1, \ldots, n_i, \tag{3}$$

for $i = 1, \ldots, m$, where $\zeta(b, a) = \Gamma(b)/\Gamma(b + a)$ for $a, b > 0$. The joint density function associated to the joint survival function (3) is

$$f(t_{i1}, \ldots, t_{in_i}) = \Gamma(\alpha + 1) \prod_{j=1}^{n_i} \lambda_0(t_{ij}) \exp\left(\sum_{j=1}^{n_i} \mathbf{x}_{ij}^\top \boldsymbol{\beta}\right) \zeta^{(n_i)}\left(\sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \mathrm{e}^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}} + 1, \alpha\right),$$

for $t_{ij} > 0$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$, where $\zeta^{(0)}(b, a) \equiv \zeta(b, a)$ and $\zeta^{(k)}(b, a) = (-1)^k \partial^k \zeta(b, a)/\partial b^k$ for integer $k \geq 1$. We have in particular that $\zeta^{(1)}(b, a) = \zeta(b, a)\{\Psi(a + b) - \Psi(b)\}$ and $\zeta^{(2)}(b, a) = \zeta(b, a)\{(\Psi(b) - \Psi(b + a))^2 + \Psi'(b) - \Psi'(b + a)\}$. Analytical expressions for higher-order derivatives of $\zeta(a, b)$ can be obtained through programs such as Mathematica and Maple.

Denote $L(\boldsymbol{\theta})$ as the likelihood function, $\ell(\boldsymbol{\theta})$ the log-likelihood and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Lambda_0, \alpha)^\top$ the parameter vector. The likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int_0^\infty \prod_{j=1}^{n_i} \Big(z_i \lambda_0(t_{ij}) \times \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta})\Big)^{\delta_{ij}} \exp(-z_i \Lambda_0(t_{ij}) \mathrm{e}^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}}) \alpha \mathrm{e}^{-z_i} (1 - \mathrm{e}^{-z_i})^{\alpha - 1} \mathrm{d}z_i. \quad (4)$$

Using expression (4) and the integral

$$\int_0^\infty z^n \mathrm{e}^{-bz} (1 - \mathrm{e}^{-z})^{a-1} \mathrm{d}z = (-1)^n \frac{\partial^n}{\partial b^n} B(a, b) = \Gamma(\alpha) \zeta^{(n)}(b, a),$$

we obtain that the log-likelihood function assumes the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \left(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \log \lambda_0(t_{ij})\right) + m(\log \alpha - \log \Gamma(\alpha)) + \sum_{i=1}^{m} \log \left[\zeta^{(d_i)}\left(\sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \mathrm{e}^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}} + 1, \alpha\right)\right], \quad (5)$$

where $d_i = \sum_{j=1}^{n_i} \delta_{ij}$, for $i = 1, \ldots, m$.

In a parametric approach, the baseline hazard $\lambda_0(\cdot)$ must be specified. In this paper, we assume the Weibull baseline hazard function $\lambda_0(t) = \sigma \phi t^{\phi - 1}$ (thus $\Lambda_0(t) = \sigma t^\phi$) for $t > 0$, where $\sigma > 0$ and $\phi > 0$ are scale and shape parameters, respectively.

3 Semiparametric approach and EM-algorithm

This section presents the semiparametric version of the GE frailty model, where the baseline hazard function $\lambda_0(\cdot)$ does not need to be specified as opposed to the parametric case. In brief, we consider a discrete version of the function $\Lambda_0(t)$, being a step

function at the observed failure times, and then use an EM-algorithm for estimating parameters.

The complete data are represented by $(t_{ij}, \delta_{ij}, Z_i)$, for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where the pairs (t_{ij}, δ_{ij})'s are observable and the Z_i's are the latent frailties; define $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$. The complete likelihood function is $L_c(\boldsymbol{\theta}) = L_1(\boldsymbol{\beta}, \Lambda; \mathbf{Z})\, L_2(\alpha; \mathbf{Z})$, where we have defined $L_1(\boldsymbol{\beta}, \Lambda; \mathbf{Z}) = \prod_{i=1}^m \prod_{j=1}^{n_i} \left(z_i \lambda_0(t_{ij}) \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta})\right)^{\delta_{ij}} \exp(-z_i \Lambda_0(t_{ij}) \mathrm{e}^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}})$ and $L_2(\alpha; \mathbf{Z}) = \prod_{i=1}^m \alpha\, \mathrm{e}^{-z_i} (1 - \mathrm{e}^{-z_i})^{\alpha-1}$.

The complete log-likelihood is then $\ell_c(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\beta}, \Lambda; \mathbf{Z}) + \ell_2(\alpha; \mathbf{Z})$, with

$$\ell_1(\boldsymbol{\beta}, \Lambda; \mathbf{Z}) \equiv \log L_1(\boldsymbol{\beta}, \Lambda; \mathbf{Z}) \propto \sum_{i=1}^m \sum_{j=1}^{n_i} \delta_{ij} \left(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \log \lambda_0(t_{ij})\right) - \sum_{i=1}^m \sum_{j=1}^{n_i} Z_i \Lambda_0(t_{ij}) \mathrm{e}^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}} \tag{6}$$

and

$$\ell_2(\alpha; \mathbf{Z}) \equiv \log L_2(\alpha; \mathbf{Z}) \propto m \log \alpha + (\alpha - 1) \sum_{i=1}^m \log(1 - \mathrm{e}^{-Z_i}). \tag{7}$$

In a discrete version of the cumulative baseline hazard function, we replace $\Lambda_0(t)$ by $\Lambda_0^d(t) = \sum_{k:t_{(k)} \le t} \lambda_0\left(t_{(k)}\right)$, where $t_{(1)} < \cdots < t_{(q)}$ are the ordered distinct failure times t_{ij}'s (q is the number of distinct failure times). As a result, the first term of the log-likelihood ℓ_c in (6) becomes

$$\ell_1(\boldsymbol{\beta}, \Lambda^d; \mathbf{Z}) \propto \sum_{k=1}^q d_{(k)} \log \lambda_0(t_{(k)}) + \sum_{i=1}^m \sum_{j=1}^{n_i} \delta_{ij} \mathbf{x}_{ij}^\top \boldsymbol{\beta} - \sum_{k=1}^q \lambda_0(t_{(k)}) \sum_{i,j \in R(t_{(k)})} Z_i \mathrm{e}^{\mathbf{x}_{ij}^\top \boldsymbol{\beta}}, \tag{8}$$

where $R(t_{(k)}) = \{(i, j) : t_{ij} > t_{(k)}\}$ is the risk set at time $t_{(k)}$ and $d_{(k)}$ is the number of failures at $t_{(k)}$, for $k = 1, \ldots, q$.

The complete log-likelihood $\ell_c(\theta) = \ell_1(\boldsymbol{\beta}, \Lambda^d; \mathbf{Z}) + \ell_2(\alpha; \mathbf{Z})$ can now be used to build an EM-algorithm; ℓ_1 and ℓ_2 are presented in (8) and (7), respectively. First, we determine the Expectation step of the algorithm assuming Z_i's as latent random variables. The conditional density of Z_i given $(t_{ij}, \delta_{ij})_{j=1}^{n_i}$ is necessary for this task. It can be shown that the conditional density $f(z_i | t_{ij}, \delta_{ij},\ j = 1, \ldots, n_i) = f(t_{ij}, \delta_{ij},\ j = 1, \ldots, n_i | z_i) f(z_i) / f(t_{ij}, \delta_{ij},\ j = 1, \ldots, n_i)$ assumes the form

$$f(z_i|t_{ij},\delta_{ij},\ j=1,\ldots,n_i) = \frac{z_i^{d_i}(1-\mathrm{e}^{-z_i})^{\alpha-1}\exp\left\{-z_i\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})\mathrm{e}^{\mathbf{x}_{ij}^\top\boldsymbol{\beta}}+1\right)\right\}}{\Gamma(\alpha)\zeta^{(d_i)}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})\mathrm{e}^{\mathbf{x}_{ij}^\top\boldsymbol{\beta}}+1,\alpha\right)}, \quad z_i>0, \tag{9}$$

for $i=1,\ldots,m$.

Conditional on the observable data, the frailties $Z_1,\ldots,Z_n$ are independent random variables with density given in (9). The Q-function of the algorithm can be denoted by $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}) \equiv E(\ell_c(\boldsymbol{\theta})|(t_{ij},\delta_{ij}),\ i=1,\ldots,m,\ j=1,\ldots,n_i;\boldsymbol{\theta}^{(r)})$, with $\boldsymbol{\theta}^{(r)}$ being the estimate of $\boldsymbol{\theta}$ in the r-th step. In order to obtain the Q-function, the conditional expectations $E\left(Z_i|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right)$ and $E\left(\log(1-\mathrm{e}^{-Z_i})|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right)$ must be calculated. These expectations are presented in the next proposition and can be obtained after some algebra.

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

$$\omega_i(\boldsymbol{\theta}) \equiv E\left(Z_i\big|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right) = \frac{\zeta^{(d_i+1)}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^\top\boldsymbol{\beta}}+1,\alpha\right)}{\zeta^{(d_i)}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^\top\boldsymbol{\beta}}+1,\alpha\right)}$$

and

$$\begin{aligned}\kappa_i(\boldsymbol{\theta}) &\equiv -E\left(\log(1-e^{-Z_i})\big|(t_{ij},\delta_{ij})_{j=1}^{n_i}\right)\\ &= -\Psi(\alpha) - \frac{\chi^{(d_i)}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^\top\boldsymbol{\beta}}+1,\alpha\right)}{\zeta^{(d_i)}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^\top\boldsymbol{\beta}}+1,\alpha\right)},\end{aligned}$$

where $\chi^{(d)}(b,a)=\partial\zeta^{(d)}(b,a)/\partial a$, for $d\in\mathbb{N}$ and $a,b>0$.

Therefore the Q-function can expressed as $Q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(r)}) = Q_1(\boldsymbol{\beta},\Lambda_0^d;\widehat{\boldsymbol{\theta}}^{(r)}) + Q_2(\alpha;\widehat{\boldsymbol{\theta}}^{(r)})$, where

$$\begin{aligned}Q_1(\boldsymbol{\beta},\Lambda_0^d;\widehat{\boldsymbol{\theta}}^{(r)}) &\propto \sum_{k=1}^{q} d_{(k)}\log\lambda_0(t_{(k)}) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\mathbf{x}_{ij}^\top\boldsymbol{\beta}\\ &\quad -\sum_{k=1}^{q}\lambda_0(t_{(k)})\sum_{i,j\in R(t_{(k)})}\exp\left(\mathbf{x}_{ij}^\top\boldsymbol{\beta}+\log\omega_i(\widehat{\boldsymbol{\theta}}^{(r)})\right)\end{aligned}$$

and

$$Q_2(\alpha;\widehat{\boldsymbol{\theta}}^{(r)}) \propto m\log\alpha - (\alpha-1)\sum_{i=1}^{m}\kappa_i(\widehat{\boldsymbol{\theta}}^{(r)}). \tag{10}$$

Moving to the M-step of the algorithm, consider the previous Q-function and note that $(\boldsymbol{\beta}, \Lambda_0^d)$ and α can be estimated independently in each iteration. Starting by estimating $(\boldsymbol{\beta}, \Lambda_0^d)$, we fix $\boldsymbol{\beta}$ and take $\partial Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(r)})/\partial\lambda_0(t_{(k)}) = 0$; this implies that $\partial Q_1(\boldsymbol{\beta}, \Lambda_0^d; \widehat{\boldsymbol{\theta}}^{(r)})/\partial\lambda_0(t_{(k)}) = 0$. Hence, we get

$$\widetilde{\lambda}_0(t_{(k)}) = \frac{d_{(k)}}{\sum_{i,j\in R(t_{(k)})} \exp\left(\mathbf{x}_{ij}^\top\boldsymbol{\beta} + \log\omega_i\left(\widehat{\boldsymbol{\theta}}^{(r)}\right)\right)}, \tag{11}$$

for $k = 1, \ldots, q$. Replacing $\lambda_0(t_{(k)})$ by (11) determines that the Q_1-function, now depending only on $\boldsymbol{\beta}$, is

$$Q_1(\boldsymbol{\beta}; \widehat{\boldsymbol{\theta}}^{(r)}) \propto -\sum_{k=1}^{q} d_{(k)} \log\left(\sum_{i,j\in R(t_{(k)})} \exp\left(\mathbf{x}_{ij}^\top\boldsymbol{\beta} + \log\omega_i\left(\widehat{\boldsymbol{\theta}}^{(r)}\right)\right)\right) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \delta_{ij}\mathbf{x}_{ij}^\top\boldsymbol{\beta}.$$

The above profile Q_1-likelihood function has the usual form of the partial likelihood function of the Cox model with the inclusion of the offset $\log\omega_i(\widehat{\boldsymbol{\theta}}^{(r)})$. Therefore, the estimation of $\boldsymbol{\beta}$ in each step of the EM-algorithm can be performed through the Cox approach. The same holds for the estimation of the parameters $\lambda_0(t_{(1)}), \ldots, \lambda_0(t_{(q)})$. Once $\widehat{\boldsymbol{\beta}}^{(r+1)}$ (the parameter vector $\boldsymbol{\beta}$ estimated in the $(r+1)$-th step of the EM-algorithm) is obtained, the parameter $\lambda_0(t_{(k)})$ can be estimated as in (11) by replacing $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\beta}}^{(r+1)}$, this yields $\widehat{\lambda}_0(t_{(k)})^{(r+1)}$ for $k = 1, \ldots, q$. This estimation procedure can be applied using existing computational packages in the literature to fit the Cox model.

The nonparametric estimator of the baseline cumulative hazard $\Lambda_0(\cdot)$, in the $(r+1)$-th step of the EM-algorithm, is given by

$$\widehat{\Lambda}_0(t)^{(r+1)} = \sum_{k:t_{(k)}\le t} \frac{d_{(k)}}{\sum_{i,j\in R(t_{(k)})} \exp\left(\mathbf{x}_{ij}^\top\widehat{\boldsymbol{\beta}}^{(r+1)} + \log\omega_i(\widehat{\boldsymbol{\theta}}^{(r)})\right)}, \quad t > 0. \tag{12}$$

The estimate of the parameter α in the $(r+1)$-th step of the algorithm is obtained by solving the equation $\partial Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(r)})/\partial\alpha = 0$, which in the GE case is $\partial Q_2(\alpha; \widehat{\boldsymbol{\theta}}^{(r)})/\partial\alpha = 0$. This provides

$$\widehat{\alpha}^{(r+1)} = m \Big/ \sum_{i=1}^{m} \kappa_i\left(\widehat{\boldsymbol{\theta}}^{(r)}\right). \tag{13}$$

This is an important result in the comparison with the well-known semiparametric gamma frailty model. The estimation of α (related to the frailty distribution) is obtained

in a closed form in each step of the EM-algorithm for the GE model. The corresponding procedure in the semiparametric gamma frailty model requires the use of maximization routines such as the Newton-Raphson method (Yu 2006) `R` via `optim`.

In summary, the EM-algorithm for the semiparametric GE model has four main steps:

1. Input a starting value $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$. The Cox regression model might be used for initial guesses of $\boldsymbol{\beta}$ and $\Lambda_0(\cdot)$. Consider $\omega_i(\widehat{\boldsymbol{\theta}}^{(0)}) = 1$ and $\widehat{\alpha}^{(0)} = 1$.
2. (E-step) Update the Q-function using $\boldsymbol{\theta}^{(r)}$ through the conditional expectations given in Proposition 1, where $\boldsymbol{\theta}^{(r)}$ is the estimate of $\boldsymbol{\theta}$ in the step r.
3. (M-step) Find $\widehat{\boldsymbol{\beta}}^{(r+1)}$ and $\widehat{\Lambda}_0^{(r+1)}$ by fitting a Cox regression model with offset $\log \omega_i(\widehat{\boldsymbol{\theta}}^{(r)})$ and compute $\widehat{\alpha}^{(r+1)}$ using (13).
4. Verify a convergence criterion, e.g., $\max\{||Q(\widehat{\boldsymbol{\theta}}^{(r+1)}; \widehat{\boldsymbol{\theta}}^{(r)}) - Q(\widehat{\boldsymbol{\theta}}^{(r)}; \widehat{\boldsymbol{\theta}}^{(r)})||, ||\widehat{\boldsymbol{\theta}}^{(r+1)} - \widehat{\boldsymbol{\theta}}^{(r)}||\} < \epsilon$, for some $\epsilon > 0$. If this criterion is satisfied, set $\widehat{\boldsymbol{\theta}}^{(r+1)}$ as the estimate of $\boldsymbol{\theta}$; otherwise, update $\widehat{\boldsymbol{\theta}}^{(r)}$ with $\widehat{\boldsymbol{\theta}}^{(r+1)}$ and return to Step 2.

In order to obtain the standard errors of the parameters estimates, we proceed as indicated in Klein (1992). The information matrix is $I(\boldsymbol{\beta}, \alpha) = -\partial^2 \ell(\boldsymbol{\beta}, \alpha)/\partial(\boldsymbol{\beta}, \alpha)\partial(\boldsymbol{\beta}, \alpha)^\top$, where ℓ is the observed log-likelihood given in (5) with $\lambda_0(\cdot)$ and $\Lambda_0(\cdot)$ replaced by (11) and (12), respectively. This matrix can be obtained numerically and we do not present an explicit form for it here to be concise. Once the EM-algorithm convergence is obtained after r steps, the estimated information matrix $I(\widehat{\boldsymbol{\beta}}^{(r)}, \widehat{\alpha}^{(r)})$ is calculated.

4 Monte Carlo simulation

In this section, we develop a simulation study to evaluate the performance of the parametric and semiparametric versions of the proposed GE frailty model. The results are compared to those obtained from other frailty models well known in the literature. The Monte Carlo strategy is considered to avoid drawing conclusions from a single sample and to allow a broader analysis investigating the inherit bias and the variability associated with the estimators. First, we explore the parametric configuration assuming the Weibull distribution to represent the behavior of the baseline hazard function. Next, we study the semiparametric version, which provides a more attractive analysis without the strong restriction of choosing a distribution to model the baseline hazard.

In order to generate the data, we consider the following steps:

1. Set the real values for β, σ, ϕ and α. The first three parameters may differ between the two mechanisms generating the failure and censoring time points.
2. Generate the covariates. We choose to work with two covariates obtained from the Bernoulli (0.5). The vector β is 2×1 and the intercept β_0 is not included to prevent identifiability issues with σ in the analysis of the parametric version (as usual in frailty models).
3. Generate the frailties from one of the following distributions: GE (shape $= \alpha$, scale $= 1$), gamma (shape $= 1/\alpha$, rate $= 1/\alpha$) or IG (mean $= 1$, shape $= 1/\alpha$). In addition, the log-normal (LN) distribution (mean $= 1$, variance $= e^{\alpha} - 1$) is also considered for the semiparametric case, as suggested by a referee. Using different

frailty distributions to simulate data is useful to verify how well the distinct models can handle a correct or wrong frailty configuration.

4. Generate a failure time t_{ij} by inverting the cumulative distribution function (cdf) of the Cox regression model with Weibull baseline hazard.
5. The censoring time c_{ij} is generated by inverting the cdf of a similar Cox regression with Weibull baseline hazard (as mentioned in Step 1, β, σ and ϕ may differ here from those used in Step 4).
6. Let $y_{ij} = \min\{t_{ij}, c_{ij}\}$ and set $\delta_{ij} = 1$, if $t_{ij} < c_{ij}$ ($\delta_{ij} = 0$ otherwise).

The steps 4–6 should be considered for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. Step 2 is applied a single time to generate all samples for the Monte Carlo procedure, i.e., the matrix of covariates is kept fixed along the Monte Carlo simulation.

All functions and programs to generate the data and to fit the different frailty models were implemented through the `R` (R Core Team 2016) programming language. The `R` package `statmod` (Giner and Smyth 2016) is used to sample from the IG distribution. In the Monte Carlo studies, we have considered the `R` package `snowfall` (Knaus 2015) for faster results through parallel computing.

4.1 Results of the parametric model

The Monte Carlo simulation, presented in this section, is set to explore 1000 data sets generated through the exact same configuration. The chosen real values for the failure time generator are: $\beta = (1.5, -1)^\top$, $\sigma = 0.25$ and $\phi = 2$. The censoring time generator takes into account: $\beta = (0, 0)^\top$, $\sigma = 0.05$ and $\phi = 2$. Given the assumption of non-informative censoring of this study, we emphasize that our focus is to evaluate the behavior of the estimators with respect to the real values reported for the failure generating procedure only. The parameter α is set to be 1.5 in the frailty distribution. All data sets have the same number of clusters (scenarios: $m = 50$, 100 and 200) with $n_i = 2$ observations each.

The maximum likelihood estimation is considered in a comparative analysis involving three different frailty models to fit the simulated data. The log-likelihood expression of the parametric GE model, shown in (5), is maximized through the `R` general purpose optimization command `optim`. The parametric gamma and inverse-Gaussian (IG) frailty models (with Weibull baseline hazard function) are widely used in applications focused on the classical approach for inference and available in `R` through the package `parfm` (Munda et al. 2012). In all cases, the maximization is developed using the BFGS method (Fletcher 2000). The algorithm initial values for all parameters (except α, starting at 1) are obtained by fitting the parametric proportional hazards model without the frailty term; this is done through the function `phreg` from the `R` package `eha` (Brostrom 2016).

The boxplots in Fig. 1 represent the distributions of the percentages of censored observations in the 1000 samples generated for the Monte Carlo simulations. Note that all graphs are below the 50% horizontal level, suggesting that the chosen configuration of true values of the parameters leads to a higher proportion of failure times in all samples. The larger the number of clusters, the lower the variability expressed through the boxplot size.

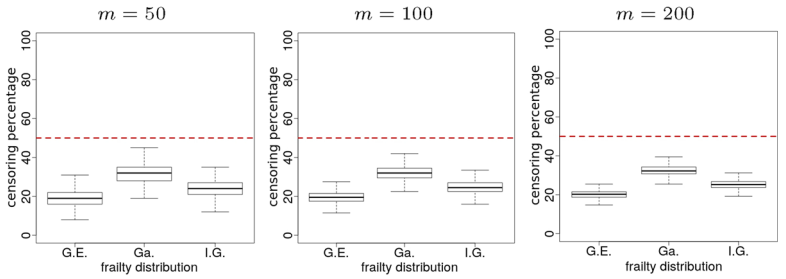

Fig. 1 Percentage of censored observations in the simulated data sets evaluated in the Monte Carlo simulation. The dashed horizontal line indicates the 50% level

The following analyses are entirely focused on the parameters β_1, β_2 and α, which are present in both parametric and semiparametric versions of the studied models. The results for the Weibull parameters ϕ and σ (estimated only in the parametric version) are not reported here since their study is not critical for the model comparisons.

Figure 2 shows the behavior of the 1000 Monte Carlo estimates for β_1, β_2 and the frailty variance. Each panel with three graphs represents a combination involving the number of clusters ($m = 50$, 100 or 200) and the true frailty distribution used to generate the data. The bottom of the graphs identifies the parametric frailty model and the horizontal dashed lines indicate the true values of the parameters. For comparison reasons, the analysis related to α is developed in terms of frailty variance, since α has a different meaning in the GE model. Note that α is the variance of the gamma and IG frailty distributions chosen to generate the data (in both cases the expected value is 1). On the other hand, the variance of the chosen GE frailty distribution is $\Psi'(1) - \Psi'(\alpha + 1)$ and the expected value is $\Psi(\alpha + 1) - \Psi(1)$. For comparison of the frailty variances, note that model (2) is equivalent to $\lambda(t_{ij}|Z_i) = Z_i^* \lambda_0^*(t_{ij}) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})$, with $Z_i^* \equiv Z_i / E(Z_i)$ (mean equal to 1) and $\lambda_0^*(t_{ij}) \equiv \lambda(t_{ij}) E(Z_i)$. Therefore, the comparison of frailty variances must consider the transformation for the GE case: $\text{Var}(Z_i^*) = [\Psi'(1) - \Psi'(\alpha + 1)]/[\Psi(\alpha + 1) - \Psi(1)]^2$.

The overall picture clearly suggests two expected results: (i) the variability of the estimates decreases when m increases and (ii) the medians are close to the real values when fitting a model assuming the correct frailty distribution. In terms of variability, the three fitted models tend to provide similar Monte Carlo standard errors for the coefficients, when exploring the same data set. This comment is not valid for the frailty variance estimates (higher Monte Carlo variability for the IG model).

The behavior of the gamma and GE models are quite similar in any configuration; they tend to fit well data sets generated with gamma or GE frailties, but some bias is observed for an IG misspecification. In contrast, the IG model usually provides biased estimates under misspecification; this is severe for all scenarios, except when estimating β_1 and β_2 using GE data.

Table 1 reinforces the conclusions taken from Fig. 2. As can be seen, the gamma and GE models have similar Monte Carlo means for the coefficients and frailty variances in

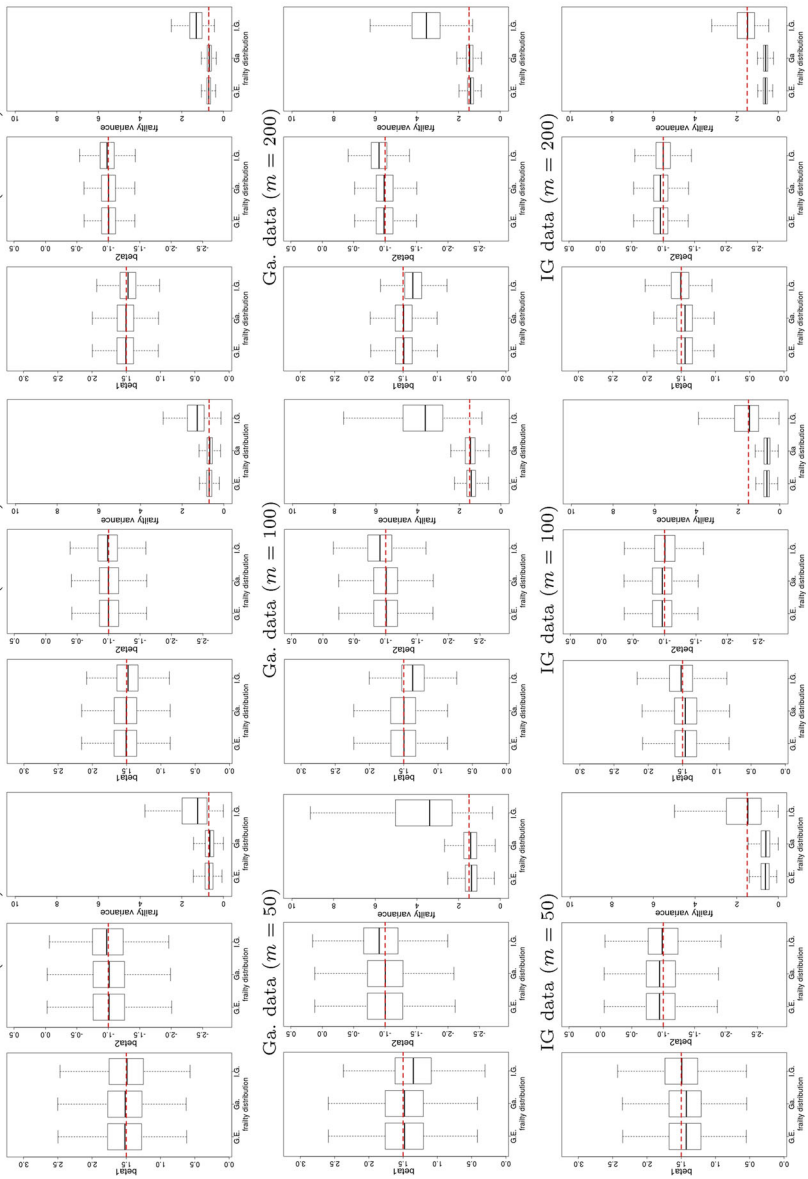

Fig. 2 Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance; columns (number of clusters) and rows (real frailty distribution). The horizontal dashed line indicates the true value of the parameter. The real values of the frailty variances are: 0.7 (GE) and 1.5 (gamma and IG)

Table 1 Empirical mean and mean square error (in parentheses) for β_1, β_2 and the frailty variance. Columns indicate the real frailty distribution and rows represent the fitted model. The real values of the frailty variance are: 0.7 (GE) and 1.5 (gamma and IG)

	m	Model	GE data	Ga. data	IG data
β_1	50	GE	1.537 (0.140)	1.494 (0.167)	1.454 (0.132)
		Ga.	1.535 (0.140)	1.496 (0.168)	1.454 (0.133)
		IG	1.501 (0.135)	1.379 (0.183)	1.517 (0.138)
	100	GE	1.522 (0.059)	1.506 (0.072)	1.458 (0.058)
		Ga.	1.520 (0.059)	1.508 (0.073)	1.458 (0.059)
		IG	1.489 (0.055)	1.372 (0.083)	1.522 (0.061)
	200	GE	1.513 (0.031)	1.491 (0.034)	1.453 (0.031)
		Ga.	1.511 (0.031)	1.493 (0.034)	1.454 (0.031)
		IG	1.473 (0.030)	1.356 (0.053)	1.517 (0.032)
β_2	50	GE	− 1.019 (0.138)	− 1.014 (0.171)	− 0.969 (0.134)
		Ga.	− 1.017 (0.138)	− 1.015 (0.172)	− 0.969 (0.135)
		IG	− 0.999 (0.144)	− 0.939 (0.183)	− 1.008 (0.141)
	100	GE	− 1.008 (0.058)	− 0.999 (0.082)	− 0.966 (0.057)
		Ga.	− 1.006 (0.058)	− 1.001 (0.082)	− 0.967 (0.058)
		IG	− 0.984 (0.059)	− 0.910 (0.086)	− 1.010 (0.061)
	200	GE	− 1.006 (0.026)	− 0.993 (0.034)	− 0.965 (0.029)
		Ga.	− 1.004 (0.026)	− 0.994 (0.035)	− 0.965 (0.030)
		IG	− 0.981 (0.026)	− 0.904 (0.042)	− 1.009 (0.030)
Var.	50	GE	0.705 (0.080)	1.413 (0.684)	0.651 (0.088)
		Ga.	0.685 (0.092)	1.458 (0.789)	0.636 (0.103)
		IG	1.839 (37.07)	4.509 (77.86)	2.226 (17.31)
	100	GE	0.703 (0.037)	1.435 (0.626)	0.632 (0.046)
		Ga.	0.680 (0.043)	1.482 (0.717)	0.615 (0.054)
		IG	1.427 (1.065)	3.969 (14.12)	1.737 (2.380)
	200	GE	0.709 (0.019)	1.429 (0.570)	0.636 (0.025)
		Ga.	0.685 (0.023)	1.474 (0.647)	0.618 (0.031)
		IG	1.366 (0.661)	3.684 (10.10)	1.627 (1.353)

all scenarios. Comparing all three models, the IG case shows significantly larger bias for the frailty variance. In terms of mean square error (MSE), the GE model provides the lowest values in most scenarios; in particular, the MSE of the gamma model is never smaller than the one reported for the corresponding GE case, even when the data comes from the gamma model.

4.2 Results of the semiparametric model

This section explores the results obtained via the EM-algorithm implemented to fit the proposed semiparametric GE frailty model. Consider again the same configura-

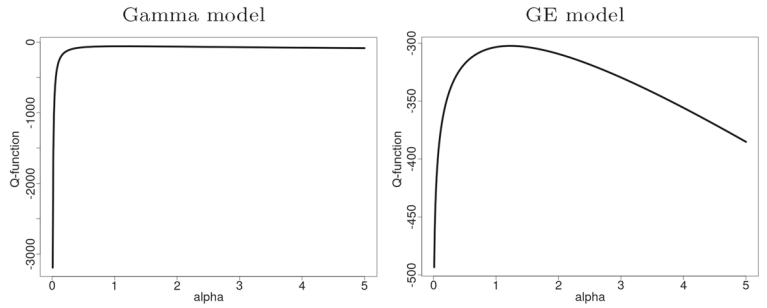

Fig. 3 Shape of the Q-function to be maximized in the EM-algorithm to estimate α. Curves built using a data set generated with gamma frailty and $m = 50$ groups

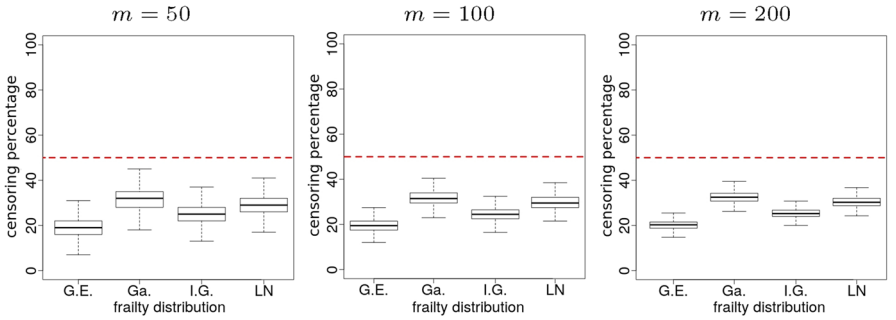

Fig. 4 Percentage of censored observations in the simulated data sets evaluated in the Monte Carlo simulation. The dashed horizontal line indicates the 50% level

tion to generate data and the same scenarios evaluated in the previous section. The present analysis compares the GE results with those from the semiparametric version of the gamma frailty model available in the literature (Klein 1992). In the convergence criterion of the EM-algorithm, we set $\epsilon = 10^{-5}$ for both models.

Two main aspects must be clarified with respect to the study developed here: (i) the IG model is not investigated since its semiparametric version is not yet proposed or well explored in the literature, (ii) in order to provide a fair comparison, only the non-penalized version of the gamma frailty model is considered in our study; this version is based on the EM-algorithm by Klein (1992). A penalized semiparametric gamma model with an EM-algorithm is available though the `R` package `survival` (Therneau and Grambsch 2000; Therneau 2015). We believe that it would be more appropriate to confront any penalized semiparametric frailty model from the literature with a penalized version of the semiparametric GE model. This topic is left for future work since the present paper is focused on exploring the non-penalized case.

Figure 3 presents the behavior of the Q-function to be maximized during the EM-algorithm to estimate α. In the GE model, consider the expression in (10); for the gamma case, see Klein (1992). The analysis here is based on the same data set, generated with $m = 50$ clusters and assuming the gamma distribution for the frailties.

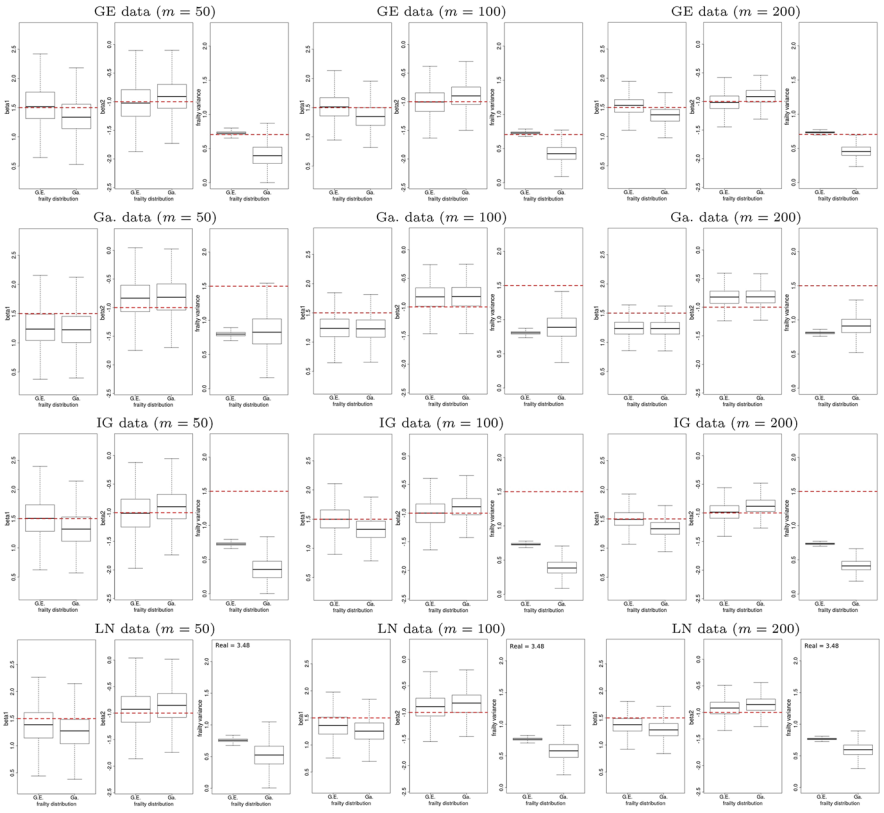

Fig. 5 Boxplots of the maximum likelihood estimates obtained in the Monte Carlo simulation for β_1, β_2 and the frailty variance; columns (number of clusters) and rows (real frailty distribution). In the fourth row, the Log-normal (LN) distribution is assumed for the frailties. The horizontal dashed line indicates the true value of the parameter. The real values of the frailty variances are: 0.7 (GE), 1.5 (gamma and IG) and 3.48 (LN)

As can be seen in this figure, the Q-function related to the gamma frailty model is definitely flat, which implies in difficulties to estimate α. In contrast, the Q-function associated with the GE case has an evident concave shape where a maximum (with a closed form) can be easily identified; see (13).

Figure 4 shows that the percentages of censored observations are below the 50% level for all generated samples. This configuration is very similar to the one investigated in the previous section for the parametric models.

Figure 5 illustrates with boxplots the behavior of the 1000 Monte Carlo estimates in each scenario; as expected, the variability exhibited by the graphs reduces as m increases. This figure also suggests that the GE model works better than the gamma for estimating both regression coefficients when the data is built with GE, IG and LN frailties. The performance (for β_1 and β_2) is quite similar to the gamma case under gamma frailty data; both semiparametric models indicate bias here. In terms of frailty variance, the GE model provides smaller Monte Carlo standard errors and it estimates

Table 2 Empirical mean and mean square error (in parentheses) for β_1, β_2 and the frailty variance. Columns indicate the real frailty distribution and rows represent the fitted model. The real values of the frailty variance are: 0.7 (GE), 1.5 (gamma and IG) and 3.48 (LN)

	m	Model	GE data	Ga. data	IG data	LN data
β_1	50	GE	1.534 (0.114)	1.264 (0.163)	1.512 (0.114)	1.386 (0.132)
		Ga.	1.353 (0.118)	1.238 (0.171)	1.325 (0.124)	1.271 (0.160)
	100	GE	1.518 (0.056)	1.243 (0.119)	1.503 (0.055)	1.366 (0.076)
		Ga.	1.350 (0.072)	1.233 (0.123)	1.329 (0.075)	1.266 (0.106)
	200	GE	1.528 (0.027)	1.244 (0.089)	1.500 (0.027)	1.374 (0.042)
		Ga.	1.368 (0.040)	1.240 (0.091)	1.337 (0.049)	1.280 (0.072)
β_2	50	GE	− 1.028 (0.115)	− 0.835 (0.145)	− 1.014 (0.129)	− 0.936 (0.134)
		Ga.	− 0.915 (0.103)	− 0.817 (0.147)	− 0.897 (0.115)	− 0.860 (0.135)
	100	GE	− 1.008 (0.057)	− 0.831 (0.086)	− 1.003 (0.060)	− 0.901 (0.072)
		Ga.	− 0.898 (0.058)	− 0.822 (0.090)	− 0.889 (0.060)	− 0.835 (0.083)
	200	GE	− 1.016 (0.026)	− 0.827 (0.054)	− 0.983 (0.028)	− 0.915 (0.034)
		Ga.	− 0.913 (0.029)	− 0.823 (0.055)	− 0.879 (0.038)	− 0.852 (0.047)
Var.	50	GE	0.727 (0.001)	0.799 (0.010)	0.729 (0.001)	0.758 (0.071)
		Ga.	0.404 (0.122)	0.842 (0.092)	0.357 (0.154)	0.530 (0.289)
	100	GE	0.731 (0.001)	0.807 (0.011)	0.732 (0.001)	0.764 (0.067)
		Ga.	0.434 (0.089)	0.896 (0.076)	0.389 (0.114)	0.581 (0.220)
	200	GE	0.734 (0.001)	0.810 (0.012)	0.736 (0.001)	0.767 (0.066)
		Ga.	0.458 (0.069)	0.916 (0.066)	0.416 (0.092)	0.596 (0.195)

well this quantity under GE data; the gamma model clearly underestimates the frailty variance in all scenarios. Results for the frailty variance are quite similar in the IG and LN frailty scenarios. This is expected due the great similarity between these two distributions; see page 96 of Wienke (2011). We believe that the gamma frailty model does not perform well in this case due to the flatness behavior of the Q-function as illustrated in Fig. 3.

Table 2 shows the empirical means and MSEs calculated for the Monte Carlo replications in each scenario. As can be seen, the results from both models tend to differ mainly for the frailty variance estimation. The MSEs of the GE model are almost always lower than those from the gamma model; the only exceptions are found for β_2 in the scenarios with $m = 50$ groups (GE and IG data).

The semiparametric version of the LN frailty model [see, for example, Zeng et al. (2008)] was not investigated in this paper, since its implementation is not available through any `R` package or for download from a repository. The implementation of this model is not straightforward and would be computationally demanding, configuring a point that is outside the scope of our paper which is focused on the comparison of the GE model with the most popular competitor (the gamma). The comparison including the LN model is left for future work.

5 Empirical illustration

The main goal of this section is to present a real data application exploring the parametric and semiparametric versions of the proposed GE frailty model. The analysis also involves a comparison with the gamma (parametric and semiparametric cases) and the IG modeling (parametric case). We choose to work with the kidney catheter data in McGilchrist and Aisbett (1991), which is a data set often used to illustrate survival models with random effects; data available through the `R` package `survival` (Therneau 2015). The response is the time to infection from the insertion of a catheter in a patient using portable dialysis equipment. The time of first and second infections are registered for each patient; there are 38 subjects and thus 76 time observations for the analysis (18 of them are right-censored). After the occurrence (or censoring) of the first event, enough time is allowed to cure the infection before initiating the second insertion. The data set includes an indicator variable identifying the event status (1 for infection, 0 for censoring) and three covariates: gender, age in years and disease type (with four categories). Following other studies in the literature [for instance, see Ibrahim et al. (2001)], the covariate "disease type" is not considered in our analyses.

Table 3 contains the estimates of the parameters and the corresponding standard errors for the parametric and semiparametric GE and gamma frailty models; again, only the parametric version of the IG model is evaluated. The semiparametric gamma (non-penalized) and GE models required 106 and 40 iterations of the EM-algorithm, respectively, to reach convergence. The coefficients estimates exhibited here resemble those reported in the literature; see for example Ibrahim et al. (2001) for results in a Bayesian context. In the parametric case, the GE and gamma models tend to provide similar results. Some differences can be noted when comparing the corresponding parametric and semiparametric estimates. As an example, both versions of the GE model produce similar results (suggesting robustness), whereas this behavior is not observed for the gamma modeling. In the penalized semiparametric gamma case, the standard error for $\widehat{\alpha}$ is missing since the `R` package `survival` does not estimate this quantity.

The values reported in Table 3 do not provide a clear picture of which model is the most suitable for this real data set. In order to assess the goodness of fit, we choose to investigate the deviance residuals, whose calculation is a transformation to deal with the typical skewness of the martigale residuals. Deviance and martingale residuals have been widely used in the literature to verify the adequacy of a Cox proportional hazards model fit. The reader should refer to Therneau et al. (1990) for further details.

A residual sum of squares (RSS) tells how much of the variation in the data is not explained by the model. In essence, it is the sum of the squared differences between the actual and the predicted response; the greater the RSS, the poorer is the model fit. Figure 6 shows the behavior of the cumulative RSS across the kidney data observations; this analysis is focused on the semiparametric models. As can be seen, the solid line is almost always below the others suggesting a lower RSS for the GE model fit. The non-penalized and penalized versions of the gamma model provide similar RSS and thus indicates their equivalence in terms of model adequacy.

Figure 7 exhibits the Q-function being maximized during the EM-algorithm for the semiparametric frailty gamma and GE models in the present kidney catheter data

Table 3 Kidney catheter data analysis

Model	β_{age}	β_{sex}	Frailty var.	α	σ	ϕ
Par. GE	0.0067 (0.0117)	− 1.9621 (0.5424)	0.5999	1.8197 (0.8800)	0.0093 (0.0059)	1.2285 (0.1498)
Par. Ga	0.0071 (0.0124)	− 1.9116 (0.5394)	0.5102	0.5102 (0.2572)	0.0129 (0.0105)	1.2155 (0.1591)
Par. IG	0.0056 (0.0124)	− 1.4809 (0.4318)	0.6774	0.6774 (0.5404)	0.0135 (0.0113)	1.1451 (0.1475)
Semi. GE	0.0067 (0.0134)	− 1.8438 (0.4756)	0.7121	1.4805 (0.3177)	−	−
Semi Ga.	0.0044 (0.0111)	− 1.4827 (0.4132)	0.3341	0.3341 (0.1454)	−	−
Semi. Pen. Ga.	0.0052 (0.0119)	− 1.5875 (0.4605)	0.4120	0.4120 (−)	−	−

Estimates of the parameters for the parametric and semiparametric models assuming GE and gamma frailties with standard errors in parentheses. We here also report the frailty variances and the parametric IG and the penalized semiparametric gamma fitted models

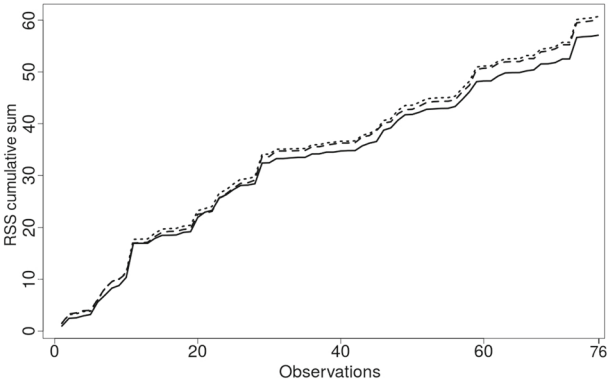

Fig. 6 Cumulative residuals (deviance) sum of squares for the semiparametric models; GE (solid line), non-penalized gamma (dashed line) and penalized gamma (dotted line)

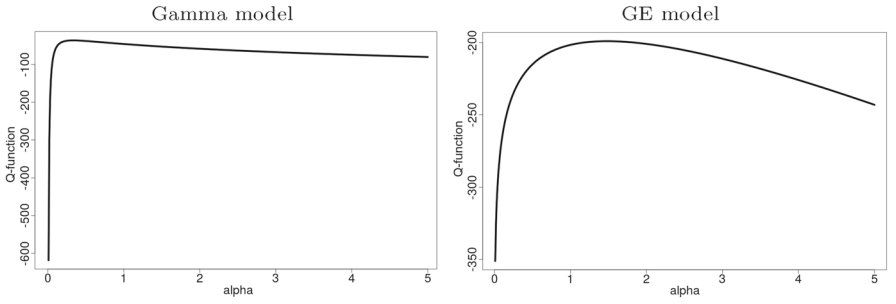

Fig. 7 Shape of the Q-function to be maximized in the EM-algorithm to estimate α. Curves built using the kidney catheter data

application. Note that, again, the corresponding Q-function is flat for the gamma model and it is not flat for the GE case. This result reinforces the discussion for Fig. 3 (in the simulated data analysis) and confirms that the maximization to find α is indeed problematic for the gamma frailty model.

6 Concluding remarks and future research

The main goal of this paper was to introduce a new tractable and computationally attractive frailty model that can be seen as an additional option (having important features) to be explored, in practical situations, jointly with the existing frailty models in the literature. In the proposed model, the frailty follows a generalized exponential distribution; this lifetime distribution has been an alternative to other survival models such as: gamma, inverse-Gaussian and Weibull. The advantages of the semiparametric GE model over the gamma frailty model were emphasized along the study.

Results obtained through simulated data were explored in a Monte Carlo setup involving 1000 replications. Three models were compared (GE, gamma and IG) assuming nine different data configurations (varying the sample size and the true frailty

distribution). The results suggested, for both parametric and semiparametric cases, a good performance of the GE model under GE data and at least a similar performance, with respect to the gamma model, under misspecification (non-GE data). The simulation study, was also important to illustrate the fact that the semiparametric GE model does not suffer with a flat likelihood issue; such issue compromises the estimates of the frailty variance in the gamma model. The estimator for α has an explicit form in each step of the EM-algorithm for the GE model. In contrast, the gamma case requires a maximization of a function with a severe flat shape depending on the data.

The final study was devoted to investigate a real data set related to a kidney catheter experiment known in the literature. The estimates of the regression coefficients are, in general, relatively close to those reported in other works. Comparing the parametric and semiparametric estimates, the results are more consistent for the GE case than those for the gamma model, which can be a consequence of the mentioned flat behavior. In terms of model fit, the residual sum of squares (based on deviance residuals, i.e., transformed martingale residuals) are lower for the semiparametric GE case; this suggests, if not a significantly better adjustment, a similar performance compared to the widely used semiparametric gamma model.

Again, following the ideas in Hougaard (2000), this paper is not intended to promote the GE frailty model as the best choice for all situations. The practitioner/researcher should always evaluate different models, with the proposed GE model being an additional alternative shown to be tractable, computationally attractive and behaving well under different data scenarios.

Possible points for future research are extensions of the GE frailty model in the following directions: (a) GE frailty model with cure fraction; (b) time-varying GE frailty model; (c) multivariate GE frailty model; and further (d) penalized likelihood to improve the estimation of the frailty variance parameter. Some of these points are currently under investigation.

Acknowledgements The authors would like to thank the Associate Editor and two anonymous referees for their constructive comments and suggestions. The authors also acknowledge the financial support from Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG/Brazil) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil).

References

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

Andersen, P. K., Klein, J. P., Knudsen, K., Palacios, R. T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, *53*, 1475–1484.

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

Brostrom, G. (2016). *eha: Event history analysis*. R package version 2.4-4. https://CRAN.R-project.org/package=eha. Accessed March 2018.

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

Christian, N. J., Ha, I. D., Jeong, J. H. (2016). Hierarchical likelihood inference on clustered competing risks data. *Statistics in Medicine*, *35*, 251–267.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, *34*, 187–220.

Crowder, M. (1989). A multivariate distribution with Weibull connections. *Journal of the Royal Statistical Society B*, *51*, 93–107.
Duchateau, L., Janssen, P. (2008). *The frailty model. Springer series in statistics*. New York: Springer.
Enki, D. G., Noufaily, A., Farrington, C. P. (2014). A time-varying shared frailty model with application to infectious diseases. *The Annals of Applied Statistics*, *8*, 430–447.
Fletcher, R. (2000). *Practical methods of optimization* (2nd ed.). New York: Wiley.
Giner, G., Smyth, G. K. (2016). statmod: Probability calculations for the inverse Gaussian distribution. *R Journal*, *8*(1), 339–351.
Gupta, R. C., Gupta, P. L., Gupta, R. D. (1998). Modeling failure time data by Lehman alternatives. *Communications in Statistics: Theory and Methods*, *27*, 887–904.
Gupta, R. D., Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, *41*, 173–188.
Gupta, R. D., Kundu, D. (2001). Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal*, *43*, 117–130.
Ha, I. D., Pan, J., Oh, S., Lee, Y. (2014). Variable selection in general frailty models using penalized h-likelihood. *Journal of Computational and Graphical Statistics*, *23*, 1044–1060.
Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, *71*, 75–83.
Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika*, *73*, 671–678.
Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer. Springer series. in Statistics.
Hougaard, P., Harvald, B., Holm, N. V. (1992). Measuring the similarities between the lifetimes of adult danish twins born 1881–1930. *Journal of the American Statistical Association*, *87*, 17–24.
Ibrahim, J. G., Chen, M. H., Sinha, D. (2001). *Bayesian survival analysis. Springer series in statistics*. New York: Springer.
Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, *48*, 795–806.
Knaus, J. (2015). *Snowfall: Easier cluster computing (based on snow)*. R package version 1.84-6.1. https://CRAN.R-project.org/package=snowfall. Accessed Mar 2018.
McGilchrist, C. A. (1993). REML estimation for survival models with frailty. *Biometrics*, *49*, 221–225.
McGilchrist, C. A., Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, *49*, 461–466.
Munda, M., Rotolo, F., Legrand, C. (2012). Parfm: Parametric frailty models in R. *Journal of Statistical Software*, *51*(11), 1–20.
Nadarajah, S., Kotz, S. (2006). The beta exponential distribution. *Reliability Engineering and System Safety*, *91*, 689–697.
Nielsen, G. G., Gill, R. D., Andersen, P. K., Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, *19*, 25–44.
Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, *26*, 181–214.
R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed Mar 2018.
Therneau, T. (2015). *A package for survival analysis in S*. R package version 2.38. https://CRAN.R-project.org/package=survival. Accessed Mar 2018.
Therneau, T. M., Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer.
Therneau, T. M., Grambsch, P. M., Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160.
Therneau, T. M., Grambsch, P. M., Pankratz, V. S. (2003). Penalized survival models. *Journal of Computational and Graphical Statistics*, *12*(1), 156–175.
Vaupel, J., Manton, K., Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*, 439–454.
Wienke, A. (2011). *Frailty models in survival analysis. CRC biostatistics series*. New York: Chapman and Hall.
Yavuz, A. C., Lambert, P. (2016). Semi-parametric frailty model for clustered interval-censored data. *Statistical Modelling*, *16*, 360–391.

Yu, B. (2006). Estimation of shared gamma frailty models by a modified EM algorithm. *Computational Statistics and Data Analysis*, *50*, 463–474.

Zeng, D., Lin, D. Y., Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica*, *18*, 355–377.