# frailtyEM: An R Package for Estimating Semiparametric Shared Frailty Models

**Theodor Adrian Balan**
Leiden University Medical Center

**Hein Putter**
Leiden University Medical Center

**Abstract**

When analyzing correlated time to event data, shared frailty (random effect) models are particularly attractive. However, the estimation of such models has proved challenging. In semiparametric models, this is further complicated by the presence of the nonparametric baseline hazard. Although recent years have seen an increased availability of software for fitting frailty models, most software packages focus either on a small number of distributions of the random effect, or support only on a few data scenarios. **frailtyEM** is an R package that provides maximum likelihood estimation of semiparametric shared frailty models using the Expectation-Maximization algorithm. The implementation is consistent across several scenarios, including possibly left truncated clustered failures and recurrent events in both calendar time and gap time formulation. A large number of frailty distributions belonging to the Power Variance Function family are supported. Several methods facilitate access to predicted survival and cumulative hazard curves, both for an individual and on a population level. An extensive number of summary measures and statistical tests are also provided.

*Keywords*: shared frailty, EM algorithm, recurrent events, clustered failures, left truncation, survival analysis, R.

## 1. Introduction

Time-to-event data are very common in medical applications. Often, these data are characterized by incomplete observations. For example, the phenomenon of right censoring occurs when the actual event time is not observed, but the only thing that is known is that the event has not taken place by the end of follow-up. Sometimes, individuals enter the data set only if they have not experienced the event before a certain time point. This is known as left truncation, which, if not accounted for correctly, leads to bias. Regression models for such data have been developed in the field of survival analysis. The most popular is the Cox proportional hazards model (Cox 1972), which is semiparametric in nature: the effect of the covariates is assumed to be time-constant and fully parametric, while the time-dependent probability of observing an event arises from the nonparametric baseline hazard. Cox regression has been the standard in survival analysis for a few reasons. First, it does not require any a priori assumptions about the baseline hazard. Second, under the proportional hazards assumption, maximum likelihood estimation can be carried out efficiently using Cox's partial likelihood. Nowadays, such models may be estimated with most statistical software, such as R (R Core Team 2016) Stata (StataCorp 2017), SAS (SAS Institute Inc. 2003) or SPSS (IBM Corp 2016).

When individuals belong to clusters, or may experience recurrent events, the observations are correlated. In this case the Cox model is not appropriate for modeling individual risk. A natural extension is represented by random effect "shared frailty" models. Originating from the field of demographics (Vaupel, Manton, and Stallard 1979), these models traditionally assume that the proportional hazards model holds conditional on the frailty, a random effect that acts multiplicatively on the hazard. The variance of the frailty is usually indicative of the degree of heterogeneity in the data. This makes the choice of the random effect distribution relevant. However, the simplicity that made the Cox model so popular does not carry over to such models.

Arguably the most popular way of fitting semiparametric shared frailty models is via the penalized likelihood method (Therneau, Grambsch, and Pankratz 2003), available for the gamma and log-normal frailty distributions. This is the standard in the **survival** package (Therneau and Grambsch 2000; Therneau 2015a) in R, in the `PHREG` command in SAS and the `streg` procedure in Stata. This method has the advantage that it is generally fast and the Cox model is contained as a limiting case when the variance of the frailty is 0. However, this algorithm can not be used for estimating other frailty distributions or left-truncated data, and the provided standard errors are presented under the assumption that the estimated parameters of the frailty distribution are fixed. Log-normal frailty models may also be estimated in R via Laplace approximation in **coxme** (Therneau 2015b), h-likelihood in **frailtyHL** (Do Ha, Noh, and Lee 2012) or Monte Carlo Expectation-Maximization **phmm** (Donohue and Xu 2013; Vaida and Xu 2000; Donohue, Overholser, Xu, and Florin 2011). Parametric and spline based shared frailty models are implemented for the gamma and log-normal distributions in the **frailtypack** package (Rondeau, Mazroui, and Gonzalez 2012; Rondeau and Gonzalez 2005).

In Hougaard (2000), the Power Variance Function (PVF) family was proposed for modeling the frailty distribution. This family of frailty distributions includes the gamma, positive stable (PS), inverse Gaussian (IG) and compound Poisson distributions with mass at 0. Each choice of the distribution for the frailty implies a different marginal model, with some emphasizing early dependence of the observations (IG) and others late dependence (gamma). Of particular interest is the PS distribution: with assumed proportional hazards conditional on the frailty, the PS implies proportional hazards also unconditional on the frailty. This is unlike the other distributions which imply non-proportional hazards at the marginal level. Therefore, this is the only distribution where the potential violation of the proportional hazards is not confounded with a frailty effect.

The software implementation of the the PVF family of distributions so far been limited. At this time, two R packages incorporate a larger number of distributions from this family: the **frailtySurv** package (Monaco, Gorfine, and Hsu 2017; Gorfine, Zucker, and Hsu 2006) implements the above mentioned distributions except the PS via a pseudo full likelihood approach and the **parfm** package (Munda, Rotolo, Legrand *et al.* 2012) estimates fully parametric gamma, IG, PS and log-normal frailty models.

In this paper we present **frailtyEM** (Balan and Putter 2017), an R package which uses the general Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) for fitting semiparametric shared frailty models. This implementation comes to complete the landscape of packages that may be used for such models, with support for the whole PVF family of distributions for the scenarios of clustered failures, clustered failures with left truncation and recurrent events data. In the latter case, different time scales are supported, such

as calendar time (time since origin of the recurrent event process) and gap time (time since previous recurrent event). Point estimates for regression coefficients are provided with confidence intervals that take into account the estimation of the frailty distribution parameters, and plotting methods facilitate the visualization of both conditional and marginal survival or cumulative hazard curves with 95% confidence bands, marginal covariate effects, and empirical Bayes estimates of the random effects. A comparison with respect to functionality between **frailtyEM** and other R packages is provided in Table 1.

The rest of this paper is structured as follows. In Section 2 we present a brief overview the semiparametric shared frailty model, and the implications of left truncation. In Section 3 we discuss the estimation method and its implementation. In Section 4 we illustrate the usage of the functions from the **frailtyEM** package on three classical data sets available in R.

# 2. Model

## 2.1. Shared frailty models

In **frailtyEM**, the general framework is of $I$ clusters with $J_i$ individuals within cluster $i$, $i = 1, \ldots, I$. The event history of individual $j$ from cluster $i$ is represented by a counting process $N_{ij}$, with $N_{ij}(t)$ representing the number of events observed until time $t$. The "at-risk" process $Y_{ij}(t)$ is defined as 1 when individual $(ij)$ is under observation and 0 otherwise, and a vector of possibly time dependent covariates is denoted as $\mathbf{x}_{ij}(t)$.

The clustered failures scenario is represented when the $N_{ij}(t) \leq 1$ and $Y_{ij}(t) = 0$ after an event or right censoring. The data in cluster $i$ consists of $J_i$ possibly right censored survival times. If $N_{ij}(t)$ exceeds 1, the case of recurrent events is obtained. In this scenario, it is considered that each cluster contains only one individual ($J_i \equiv 1$, with the corresponding counting process $N_i$). Calendar time (also known as Andersen-Gill) models, when the time scale is "time since origin" and gap time models, where the time scale is "time since the previous event" are commonly employed (Cook and Lawless 2007). When subject $i$ is no longer under observation, the last time point is typically considered right censored.

The intensity of $N_{ij}$ (or hazard, in the clustered failure scenarios) is specified as

$$\lambda_{ij}(t|Z_i) = Y_{ij}(t) Z_i \exp(\beta^\top \mathbf{x}_{ij}(t)) \lambda_0(t) \tag{1}$$

where $Z_i$ is an unobserved random effect common to all observations from cluster $i$ (the "shared frailty"), $\beta$ a vector of unknown regression coefficients and $\lambda_0(t) \geq 0$ an unspecified baseline intensity function. It is assumed that the $Z_i$ are iid random variables with a distribution referred to as $Z$, and that event times are independent given $Z_i$. A stratified model (1) may also be specified by specifying different baseline intensities for different groups of observations. In this case, if individual $(i, j)$ belongs to strata $s$, $\lambda_0(t)$ is replaced by $\lambda_{0s}(t)$.

We consider the general case where the $Z$ follows a distribution with positive support from the infinitely divisible family, i.e., they are i.i.d. realizations of a random variable described by the Laplace transform

$$\mathcal{L}_Z(c; \alpha, \gamma) \equiv \mathsf{E}\left[\exp(-Zc)\right] = \exp(-\alpha\psi(c; \gamma)) \tag{2}$$

with $\alpha > 0$ and $\gamma > 0$. This formulation includes several distributions, such as the gamma, positive stable, inverse Gaussian and compound Poisson distributions. This so-called power-

| | **frailtyEM** | **survival** | **coxme** | **frailtySurv** | **frailtyHL** | **frailtypack** | **parfm** | **phmm** |
|---|---|---|---|---|---|---|---|---|
| **Distributions** | | | | | | | | |
| Gamma | yes | yes | no | yes | no | yes | yes | no |
| Log-normal | no | yes | yes | yes | yes | yes | yes | yes |
| PS | yes | no | no | no | no | no | yes | no |
| IG | yes | no | no | yes | no | no | yes | no |
| Compound Poisson | yes | no | no | no | no | no | no | no |
| PVF | yes | no | no | yes | no | no | no | no |
| **Data** | | | | | | | | |
| Clustered failures | yes | yes | yes | yes | yes | yes | yes | yes |
| Recurrent events (AG) | yes | yes | yes | no | no | yes | no | no |
| Left truncation | yes | no | no | no | no | yes | yes | no |
| Correlated structure | no | no | yes | no | no | yes | no | yes |
| **Estimation** | | | | | | | | |
| Semiparametric | yes | yes | yes | yes | yes | no | no | yes |
| Posterior frailties | yes | yes | no | no | no | yes | no | no |
| Conditional $\Lambda_0$, $S_0$ | yes | limited | no | yes | no | yes | yes | no |
| Marginal $\Lambda_0$, $S_0$ | yes | no | no | no | no | no | no | no |

Table 1: Comparison of R packages for frailty models. Versions: **frailtyEM** 0.8.3, **survival** 2.40-1, **coxme** 2.2-5, **frailtyHL** 1.1, **frailtypack** 2.10.5, **parfm** 2.7.1, **phmm** 0.7-5.

variance-function (PVF) family of distributions have been extensively studied in Hougaard (2000). As detailed in Appendix A1, we assume that an identifiability constraint is imposed on the parameters $\alpha$ and $\gamma$ and that the distribution of $Z$ is indexed by a scalar parameter $\theta$.

## 2.2. Likelihood

Henceforth, we consider the problem of estimating $\beta$, $\lambda_0$ and $\theta$ via maximum likelihood. This is achieved by maximizing the marginal likelihood, based on the observed data and obtained by integrating over the random effect. For simplicity, we omit potential strata in this section. From model (1), the marginal likelihood is obtained as the product over clusters of expected marginal contributions, i.e.,

$$L(\theta,\beta,\lambda_0(\cdot)) = \prod_i \mathsf{E}_\theta \left[ \prod_j \int_0^\infty \left\{ Y_{ij}(t) Z \exp(\beta^\top \mathbf{x}_{ij}(t)\lambda_0(t) \right\}^{dN_{ij}(t)} \times \exp\left( -\sum_j \int_0^\infty Y_{ij}(t) Z \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(t)dt \right) \right]$$

The first part reduces to a product of contributions from the observed event times of the counting processes from cluster $i$. Denote the $k$-th observed time corresponding to the counting process $N_{ij}$ as $t_{ijk}$ and $\delta_{ijk} = 1$ if $t_{ijk}$ is an event time and 0 otherwise. Let $\tilde{\Lambda}_i = \sum_j \int_0^\infty Y_{ij}(t) \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(t)dt$ and $n_i = \sum_j \int_0^\infty Y_{ij}(t) dN_{ij}(t)$ the number of observed events in cluster $i$. The marginal likelihood can be written as

$$L(\theta,\beta,\lambda_0(\cdot)) = \prod_i \left[ \prod_j \prod_k \left\{ \exp(\beta^\top \mathbf{x}_{ij}(t_{ijk}))\lambda_0(t_{ijk}) \right\}^{\delta_{ijk}} \right] \mathsf{E}_\theta \left[ Z^{n_i} \exp(-Z\tilde{\Lambda}_i) \right]. \quad (3)$$

By using (2), the last term may be expressed in terms of the $n_i$-th derivative of the Laplace transform, i.e.

$$\mathsf{E}_\theta \left[ Z^{n_i} \exp(-Z\tilde{\Lambda}_i) \right] = (-1)^{n_i} \mathcal{L}_Z^{(n_i)}(\tilde{\Lambda}_i).$$

In **frailtyEM**, the Breslow estimator is employed for the baseline hazard, i.e., $\lambda_0(t) \equiv \lambda_{0t}$ for $t$ an event time, and 0 otherwise. This is equivalent with estimating $\int_0^t \lambda_0(s)ds$ as a step function with "jumps" of size $\lambda_{0t}$ at event times.

## 2.3. Ascertainment and left truncation

The problem of ascertainment with random effect time-to-event data is usually difficult. If $Z_i$ is the distribution of the frailty of cluster $i$ and $A_i$ denotes the event of selecting the observations in cluster $i$, the random effect distribution of cluster $i$ given the ascertainment is of the form $Z_i|A_i$. The Laplace transform of $Z_i|A_i$ follows from Bayes' rule as

$$\mathcal{L}_{Z_i|A_i}(c) = \frac{\mathsf{E}\left[\mathsf{P}(A_i|Z_i)\exp(-cZ_i)\right]}{\mathsf{E}\left[\mathsf{P}(A_i|Z_i)\right]}. \quad (4)$$

Expressing $\mathsf{P}(A_i|Z_i)$ depends on the type of the study at hand and on the way the data were collected.

In **frailtyEM** an option is included to deal with the scenario of left truncation for clustered failures. Consider that from a cluster of size $\tilde{J}_i$, $J_i \leq \tilde{J}_i$ individuals are are selected and $A_i$ is the event "selecting $J_i$ individuals with left truncation times $\mathbf{t}_{L,i} = \{t_{L,i1} \dots t_{L,iJ_i}\}$". Then $A_i$ can be expressed as

$$\mathsf{P}(A_i|Z_i) = \mathsf{P}(T_{i1} > t_{L,i1}, T_{i2} > t_{L,i2} ... T_{J_i} > t_{L,iJ_i}|Z_i).$$

A hidden assumption here is that the true cluster size $\tilde{J}_i$ does not depend on the frailty. For example, if a high frailty is associated with both a high rate of events and smaller cluster sizes, then the distribution of $\tilde{J}_i|Z$ must also be considered (Jensen, Brookmeyer, Aaby, and Andersen 2004).

Assume that, given $Z_i$, the left truncation times $\mathbf{t}_{L,i}$ are independent. In this case,

$$\mathsf{P}(A_i|Z_i) = \prod_{j=1}^{J_i} \exp\left(-Z_i \int_0^{t_{L,ij}} \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(t)dt\right). \tag{5}$$

A difficulty here is that the values of the covariate vector and of the baseline intensity must be known prior to the entry time in the study. Therefore, only cases when $\mathbf{x}_i$ is time constant are considered.

Denote $\tilde{\Lambda}_{L,i} = \sum_j \int_0^{t_{L,ij}} \exp(\beta^\top \mathbf{x}_{ij})\lambda_0(t)dt$. The marginal likelihood may be obtained from (3), (4) and (5) as

$$L(\theta, \beta, \lambda_0(\cdot)) = \prod_i \left[\prod_j \prod_k \left\{\exp(\beta^\top \mathbf{x}_{ij}(t_{ijk}))\lambda_0(t_{ijk})\right\}^{\delta_{ijk}}\right] \times \\ \times \frac{\mathsf{E}_\theta\left[Z^{n_i} \exp\left(-Z(\tilde{\Lambda}_{L,i} + \tilde{\Lambda}_i)\right)\right]}{\mathsf{E}_\theta\left[\exp(-Z\tilde{\Lambda}_{L,i})\right]}.$$

It can also be seen that, if the frailty distribution is degenerate and has no variability (i.e. $\mathsf{E}_\theta$ may be removed), then the contribution of $\tilde{\Lambda}_{L,i}$ cancels out. In particular, under left truncation, the Laplace distribution of $Z|A_i$ is given by

$$\mathcal{L}_{Z|A}(c) = \frac{\mathcal{L}(c + \tilde{\Lambda}_{L,i})}{\mathcal{L}(\tilde{\Lambda}_{L,i})}. \tag{6}$$

This distribution is often referred to as the frailty distribution of the survivors (Hougaard 2000). If $Z$ is from the PVF family, it can be shown that $Z|A$ is also in the PVF family. As a result, if $Z$ is gamma distributed, then also $Z|A$ is gamma distributed.

Note that, in general, the ascertainment scheme does not have a simple description and $P(A_i|Z_i)$ may or may not be available in closed form. For example, in family studies, the families may be selected only when a number of individuals live long enough (Rodríguez-Girondo, Deelen, Slagboom, and Houwing-Duistermaat 2016). In this case, (5) does not hold. In the case of registry data on recurrent events, individuals (clusters) may be selected only if they have at least one event during a certain time window (Balan, Jonker, Johannesma, and Putter 2016b). These specific cases are not currently accommodated by **frailtyEM**.

## 2.4. Analysis and quantities of interest

*Inference*

In **frailtyEM**, inference from the likelihood (3) is based on the non-parametric information matrix. This is obtained by treating each $\lambda_0(t) \equiv \lambda_{0t}$ as a finite-dimensional parameter. Even though its dimension grows with the number of event time points in the data, this has been shown to lead to consistent variance estimators (Andersen, Klein, Knudsen, and y Palacios 1997).

For assessing whether the frailty model is a better fit than the Cox proportional hazards model, the likelihood ratio test may be used. With the parametrizations described in Appendix A1, this is a problem of testing on the edge of the parameter space, and the test statistic under the null hypothesis follows asymptotically a mixture of $\chi^2(0)$ and $\chi^2(1)$ distribution (Zhi, Grambsch, and Eberly 2005). This test is provided as standard output in **frailtyEM**.

The Commenges-Andersen score test for heterogeneity Commenges and Andersen (1995) is implemented in **frailtyEM**. It may be applied to a proportional hazards model as fitted by the `coxph` function or automatically calculated when estimating a frailty model. If the null hypothesis of no unobserved heterogeneity is not rejected, it might be preferable to employ simpler Cox-type models.

*Marginal and conditional quantities*

Several quantities are of interest in the context of frailty models. For a group of individuals with covariate vector $\mathbf{x}_{ij}(t)$ and frailty $Z_i$, the cumulative intensity (hazard) is defined as

$$\Lambda_{ij}(t|Z_i) = Z_i \int_0^t \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(s)ds. \tag{7}$$

The survival function for such individual is given by $S_{ij}(t|Z_i) = \exp\left(-\Lambda_{ij}(t|Z_i)\right)$. These quantities are *conditional* on the random effect $Z_i$.

The population-level, or *marginal* quantities may be obtained by integrating out the frailty from the conditional ones. The marginal survival is given by

$$S_{ij}(t) = \mathsf{E}_\theta\left[\exp(-\Lambda_{ij}(t|Z_i))\right] = \mathcal{L}_Z\left(\int_0^t \exp(\beta^\top \mathbf{x}_{ij}(t))\lambda_0(s)ds\right). \tag{8}$$

The marginal cumulative intensity is then given by $\Lambda_{ij}(t) = -\log S_{ij}(t)$. The "baseline" intensities or survival refer to an individual with $\mathbf{x}_{ij}(t) \equiv 0$.

In the simple case of only one binary covariate, we assume that there are two groups, the baseline with $x = 0$ and "treatment" group with $x = 1$. In this case, the estimated $\beta$ may be interpreted as the *conditional* intensity ratio (hazard ratio) between two individuals with the same frailty. Under a frailty model, the observed hazard ratio between these two groups is typically attenuated in time (Aalen, Borgan, and Gjessing 2008, ch. 6). This *marginal* intensity ratio is calculated as the ratio of the corresponding marginal cumulative intensities $\Lambda_{ij}(t)$.

Several measures of dependence are implemented in **frailtyEM**. The first is the variance of the estimated frailty distribution $Z$, which is useful for the gamma and the PVF family. The variance of $\log Z$ is also useful for the positive stable distribution for which the variance is

infinite. Other measures of association include Kendall's $\tau$ and the median concordance. A thorough discussion and comparison of these measures can be found in Hougaard (2000).

### 2.5. Goodness of fit

Given a large choice of distributions for the frailty, the question comes in selecting the most suitable one. A comparison of the PVF family of frailty distributions can be found in Hougaard (2000, ch. 7.8). In **frailtyEM**, all the frailty distributions depend on a positive parameter $\theta$ (see Appendix A1). Given that all the distributions are part of the same family (with gamma and positive stable being limiting cases in the PVF family), the likelihood of different models is comparable across distributions. This argument suggests that it makes sense, within the PVF family, to select the model with the distribution that has the highest likelihood.

An explicit assumption of model (1) is that the censoring is non-informative on the frailty. This assumption is usually difficult to test. In **frailtyEM**, a correlation score test is implemented for the gamma distribution, following Balan, Boonk, Vermeer, and Putter (2016a). This can also be used, for example, for testing whether a recurrent event event process and a terminal event are associated.

Martingale residuals have been used to assess goodness of fit in terms on functional form of the covariates (Therneau, Grambsch, and Fleming 1990; Lin, Wei, and Ying 1993). These are provided by the `residuals()` function. For Cox models, there are several methods for assessing the proportional hazards assumption (Therneau and Grambsch 2000, ch. 6). Graphical methods involve plotting estimated survival or cumulative intensity curves. The plotting capabilities of **frailtyEM** are discussed in Section 3.4. A second method is based on Schoenfeld residuals (Grambsch and Therneau 1994). In R, this is implemented for Cox models in the `cox.zph` function from the **survival** package. In **frailtyEM**, this is provided as part of the output and may be used to test whether the conditional proportional hazards model (1) holds. This is detailed in Section 3.5.

## 3. Estimation and implementation

### 3.1. Syntax

```
R> library("frailtyEM")
```

The main model fitting function in **frailtyEM** is `emfrail`:

```
R> emfrail(formula, data, distribution, control, ...)
```

The `formula` argument contains a `Surv` object as left hand side and a `+cluster()` statement on the right hand side, specifying the column of `data` that defines the different clusters (this is common to other packages such as **frailtypack**). This formulation, that is common to most survival analysis packages, allows for the representation of clustered failures with left truncation, recurrent events in both calendar time and gap time, time dependent covariates and discontinuous intervals at risk (Therneau and Grambsch 2000, ch. 3.7, ch. 8). Two

other statements may be used in the right hand side: `+strata()` for defining a column with a stratifying variable, and `+terminal()` for defining an event status column for dependent censoring (e.g. a terminal event in the case of recurrent events; this triggers the score test for dependent censoring described Section 2.5).

The `distribution` argument determines the frailty distribution. It may be generated by the `emfrail_dist()`:

```
R> str(emfrail_dist(dist = "gamma", theta = 2))
```

```
List of 5
 $ dist           : chr "gamma"
 $ theta          : num 2
 $ pvfm           : num -0.5
 $ left_truncation: logi FALSE
 $ basehaz        : chr "breslow"
 - attr(*, "class")= chr "emfrail_dist"
```

where `dist` may be one of `"gamma"`, `"stable"` or `"pvf"`. For `"pvf"`, the `m` parameter determines the precise distribution: for $m = -1/2$ for the IG, $m \in (-1, 0)$ for the so-called Hougaard distribution and $m > 0$ a compound Poisson distribution with mass at 0. The `theta` parameter determines the starting value of the optimization. The `left_truncation` argument, if `TRUE`, leads to the calculation described in Section 2.3. The `control` argument may be generated by the `emfrail_control()` function and regulates parameters regarding to the estimation.

### 3.2. Profile EM algorithm

In **frailtyEM**, a general full-likelihood estimation procedure is implemented for the gamma, positive stable and PVF frailty models, using a semi-parametric Breslow estimator for the baseline intensity. The goal is to find $\theta, \beta, \lambda_0(\cdot)$ that maximize $L(\theta, \beta, \lambda_0(\cdot))$ (3). This can be achieved in two steps, as

$$\max_{\theta,\beta,\lambda_0} L(\theta, \beta, \lambda_0) = \max_{\theta} \left\{ \max_{\beta,\lambda_0} L(\beta, \lambda_0 | \theta) \right\}$$

where $\widehat{L}(\theta) = \max_{\beta,\lambda_0} L(\beta, \lambda_0|\theta)$ is the profile likelihood of $\theta$. The profile EM algorithm refers to using a two-stage maximization procedure: the "inner problem" which involves calculating $\widehat{L}(\theta)$ (maximizing $L(\beta, \lambda_0|\theta)$ for fixed $\theta$ with the EM algorithm), and the "outer problem", maximizing the profile likelihood $\widehat{L}(\theta)$ over $\theta$.

**The inner problem** Maximizing the likelihood for fixed $\theta$ has been proposed for the gamma frailty in Nielsen, Gill, Andersen, and Sørensen (1992) and Klein (1992), and generalizations are discussed in Hougaard (2000). The crucial observation is that the E step involves calculating the empirical Bayes estimates of the frailties $\widehat{z}_i = \mathsf{E}[Z_i|data]$. This expectation is taken with respect to the "posterior" distribution of the random effect. This is detailed in Appendix A2. The M step involves estimating a proportional hazards model with the $\log \widehat{z}_i$ as offset for each cluster. This is done via the `agreg.fit()` function in the **survival** package,

which obtains estimates of $\beta$ via Cox's partial likelihood. Subsequently, $\lambda_0$ and $\tilde{\Lambda}_i$ (and $\tilde{\Lambda}_{L,i}$, in the case of left truncation) are calculated.

The EM algorithm is guaranteed to increase $L(\beta, \lambda_0|\theta)$ with every iteration and to converge to a local maximum. Convergence is achieved when the difference in $L(\beta, \lambda_0|\theta)$ between two consecutive iterations is smaller than $\varepsilon$.

**The outer problem** The "outer" problem involves maximizing $\widehat{L}(\theta)$. For this, a general purpose Newton-type algorithm is employed (`nlm` from the **stats** package).

### 3.3. Standard errors and confidence intervals

The non-parametric information matrix is not directly obtained by the estimation procedure described in Section 3.2. From the inner problem, the standard error of the estimates for $\beta$ and $\lambda_0(\cdot)$ are calculated with Louis' formula (Louis 1982), under the assumption that $\theta$ is fixed to the maximum likelihood estimate. The standard errors obtained in this way are included in the output as `se` and are comparable to the ones from other semi-parametric frailty models (**survival** or **coxme** packages) that assume that $\theta$ is fixed. However, this leads to an underestimate of the variability of $\beta$ and $\lambda_0(\cdot)$.

In **frailtyEM**, adjusted standard errors, presented in the column `adj se`, are calculated by "propagating" the uncertainty from the estimation of $\theta$ to $\beta, \lambda_0(\cdot)$. This is described in more detail in Appendix A3.

From the outer problem, standard errors for $\theta$ (more precisely, of $\log\theta$, since the maximization takes place on the log-scale for numerical stability) are directly obtained from the numeric Hessian calculated by `nlm`. The delta method, as implemented in the **msm** package (Jackson 2011), is employed for calculating the standard errors for $\theta$ and the measures of dependence that are detailed in Appendix A1.

Two types of confidence intervals for $\theta$ (and for the frailty variance, which, in the cases where it exists, is $1/\theta$) are provided. The first are derived from symmetric confidence intervals on the log-scale. The resulting asymmetric confidence interval has been shown to provide good coverage (Balan *et al.* 2016b). The second, more computationally intensive, are referred to as "likelihood-based confidence intervals". Under the null hypothesis, the likelihood ratio test statistic follows a $\chi^2(0) + \chi^2(1)$ distribution. The critical value associated with this test statistic is approximately 1.92. Based on $\widehat{L}(\theta)$, a one-dimensional search is performed to find the confidence interval around the maximum likelihood estimate $\widehat{\theta}$ within which $\log\widehat{L}(\theta) \geq \log\widehat{L}(\widehat{\theta}) - 1.92$. The advantage of this type of confidence interval is that it is transformation invariant (with the same coverage for all derived dependence measures) and it has a 1-1 correspondence with the likelihood ratio test.

### 3.4. Methods

The `emfrail` function returns an object of class `emfrail` that is documented in `?emfrail`. Usual methods are associated with this class of objects: `print()`, `coef()`, `vcov()`, `residuals()`, `model.matrix()`, `model.frame()`, `logLik()`.

The `summary()` method returns an object of class `emfrail_summary()`, the printing of which contains general fit information, covariate estimates and distribution-specific measures of dependence and goodness of fit, discussed in Section 2.5. Arguments to `summary()` may be

used to show confidence intervals based on either the likelihood function or the delta method, as described in Section 3.3. Other arguments control the amount of information that is printed and may be used when less output is desirable.

The method for prediction of survival curves and cumulative intensity curves is implemented in `predict()`. Both conditional and marginal curves defined in Section 2.4 may be produced. The prediction is made for individuals with covariate values specified in a data frame (via the `newdata` argument) or for a fixed linear predictor (via the `lp` argument). For stratified models, the strata must also be specified. By default, the `predict` function creates predictions for each row of `newdata` or for each value of `lp` separately. With the `individual` argument, predicted curves may be produced for individuals with specific at-risk patterns (for example, if an individual is not at risk during a certain time frame), or for individuals with time dependent covariates.

After $\mathbf{x}_{ij}(t)$ is specified to `predict()`, $\Lambda_{ij}(t|Z=1)$ is calculated as in (7) and from this the other quantities are derived, including the conditional survival, the marginal survival (8) and the marginal cumulative intensity. Confidence bands are based on the asymptotic normality of the estimated $\lambda_0$, and are produced both adjusted and unadjusted for the uncertainty of $\theta$.

### 3.5. Plotting and additional features

Two plot methods are provided based on both **graphics** package via `plot()` and the **ggplot2** package, via `autoplot()`, both with identical syntax. Behind the scenes, they use calls to `predict()`. The `type` argument determines the type of plot:

- `type = "hist"` for a histogram of the posterior estimates of the frailties;
- `type = "pred"` for plotting marginal and conditional cumulative hazard or survival curves;
- `type = "hr"` for plotting marginal or conditional estimated hazard ratios between two groups of individuals. The marginal hazard ratio is determined as the ratio of the marginal intensities, as described in Section 2.4;
- `type = "frail"` for a scatter plot of the ordered posterior estimates of the frailties (also called a "caterpillar plot"). For the gamma distribution, quantiles of the posterior distribution are also included. Only available with the `autoplot()` method.

The Commenges-Andersen score test for heterogeneity is by default calculated every time `emfrail` is called and is part of the standard output. A separate function `ca_test()` is also provided, that may be used independently on Cox models produced by `coxph()` from the **survival** package.

While martingale residuals may be obtained with the `residuals()` method, the test for conditional proportional hazards, based on Schoenfeld residuals described in Section 2.5 may be accessed in the `$zph` field of the fit. This is an object of class `cox.zph` borrowed from the **survival** package and equivalent to calling `cox.zph` on a Cox model with the estimated log-frailties as offset. The structure and plot methods are described in `?cox.zph`.

An additional function is provided to calculate the marginal log-likelihood for a vector of values of $\theta$, `emfrail_pll()`, without actually performing the outer optimization. This may be

useful for visualizing the profile log-likelihood or when debugging (e.g., to see if the maximum likelihood estimate of $\theta$ lies on the boundary).

# 4. Illustration

The features of the package will now be illustrated with three well-known data sets available in R: The `CGD` data set (recurrent events, calendar time), the `kidney` data set (recurrent events, gap time) and the `rats` data set (clustered failures).

## 4.1. CGD

The data are from a placebo controlled trial of gamma interferon in chronic granulotomous disease (CGD) and is available in the **survival** package. It contains the time to recurrence of serious infections observed, from randomization until end of study for each patient (i.e. the time scale is calendar time). For the purpose of illustration, we will use `treat` (treatment or placebo) and `sex` (female or male) as covariates, although a larger number of variables are recorded in the data set.

```
R> data("cgd")
R> cgd <- cgd[c("tstart", "tstop", "status", "id", "sex", "treat")]
R> head(cgd)
```

```
  tstart tstop status id    sex   treat
1      0   219      1  1 female  rIFN-g
2    219   373      1  1 female  rIFN-g
3    373   414      0  1 female  rIFN-g
4      0     8      1  2   male placebo
5      8    26      1  2   male placebo
6     26   152      1  2   male placebo
```

A basic gamma frailty model can be fitted like this:

```
R> gam <- emfrail(Surv(tstart, tstop, status) ~ sex + treat + cluster(id),
+   data = cgd)
R> summary(gam)
```

```
Call:
emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
    cluster(id), data = cgd)

Regression coefficients:
              coef exp(coef) se(coef) adj. se      z    p
sexfemale   -0.227     0.797    0.396   0.396 -0.575 0.57
treatrIFN-g -1.052     0.349    0.310   0.310 -3.389 0.00
Estimated distribution: gamma / left truncation: FALSE
```

```
Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00172
no-frailty Log-likelihood: -331.997
Log-likelihood: -326.619
LRT: 1/2 * pchisq(10.8), p-val 0.00052

Frailty summary:
                   estimate lower 95% upper 95%
Var[Z]                0.821     0.231     1.854
Kendall's tau         0.291     0.104     0.481
Median concordance    0.289     0.101     0.491
E[logZ]              -0.464    -1.164    -0.120
Var[logZ]             1.241     0.260     4.341
theta                 1.218     0.539     4.326
Confidence intervals based on the likelihood function
```

The first two parts of this output, about regression coefficients and fit summary, exist regardless of the frailty distributions. The last part, "frailty summary", provides a different output according to the distribution.

Both the Commenges-Andersen test for heterogeneity and the one-sided likelihood ratio test deems the random effect highly significant. This is also suggested by the confidence interval for the frailty variance, which does not contain 0.

To illustrate the predicted cumulative hazard curves we take two individuals, one from the treatment arm and one from the placebo arm, both males:

```
R> library("ggplot2")
R> library("egg")
R> p1 <- autoplot(gam, type = "pred",
+                 newdata = data.frame(sex = "male", treat = "rIFN-g")) +
+   ggtitle("rIFN-g") +
+   ylim(c(0, 2)) +
+   guides(colour = FALSE)
R> p2 <- autoplot(gam, type = "pred",
+                 newdata = data.frame(sex = "male", treat = "placebo")) +
+   ggtitle("placebo") + ylim(c(0, 2))
R>
```

The two plots are shown in Figure 1.

The cumulative hazard in this case can be interpreted as the expected number of events at a certain time. It can be seen that the frailty "drags down" the marginal hazard. This is a well-known effect observed in frailty models, as described in Aalen *et al.* (2008, ch. 7). All prediction results could also be obtained directly:

```
R> dat_pred <- data.frame(sex = c("male", "male"),
+   treat = c("rIFN-g", "placebo"))
R> predict(gam, dat_pred)
```

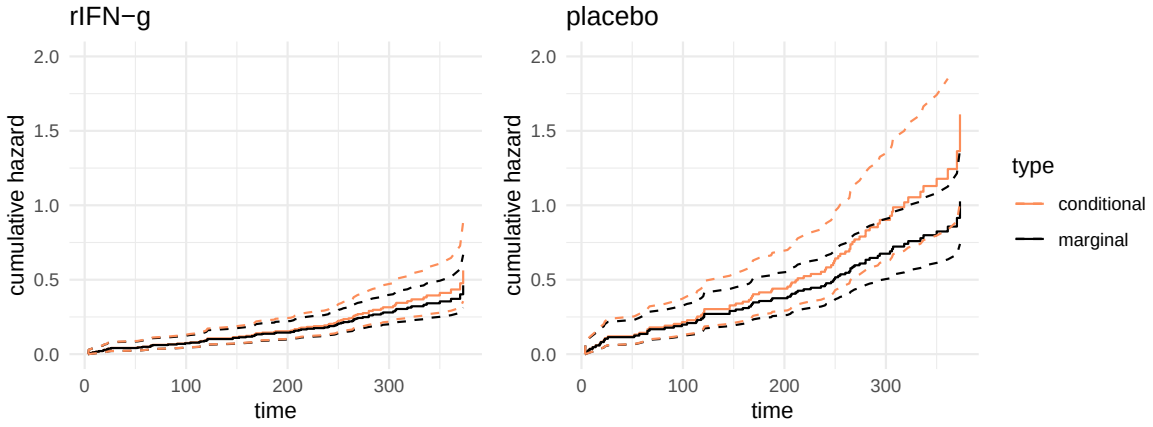



Figure 1: Predicted conditional and marginal cumulative hazards for males, one from the treatment arm and one from the placebo arm, as produced by `autplot()` with `type = "pred"`.

For a hypothetical individual that changes treatment from placebo to rIFN-g at time 200, predictions may also be obtained:

```
R> dat_pred_b <- data.frame(sex = c("male", "male"),
+   treat = c("placebo", "rIFN-g"),
+   tstart = c(0, 200), tstop = c(200, Inf))
R> p <- autoplot(gam, type = "pred", newdata = dat_pred_b, individual = TRUE) +
+   ggtitle("change placebo to rIFN-g at time 200")
R>
```

This plot is shown in Figure 2.

A positive stable frailty model can also be fitted by specifying the `distribution` argument.

```
R> stab <- emfrail(Surv(tstart, tstop, status) ~ sex + treat + cluster(id),
+   data = cgd,
+   distribution = emfrail_dist(dist = "stable"))
R> summary(stab)

Call:
emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
    cluster(id), data = cgd, distribution = emfrail_dist(dist = "stable"))

Regression coefficients:
              coef exp(coef) se(coef) adj. se      z    p
sexfemale   -0.137     0.872    0.407   0.407 -0.337 0.74
treatrIFN-g -1.085     0.338    0.332   0.336 -3.230 0.00
Estimated distribution: stable / left truncation: FALSE
```

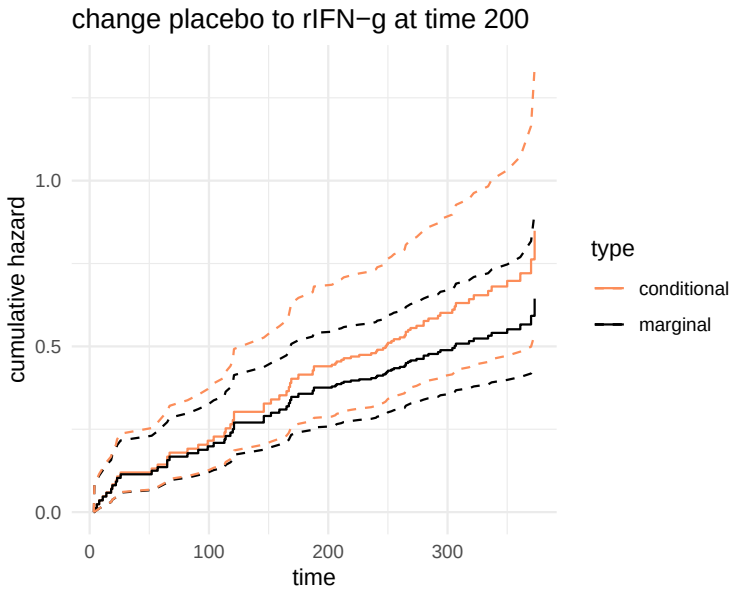



Figure 2: Predicted conditional and marginal cumulative hazards for a male that switches treatment from placebo to rIFN-g at time 200 as produced by `autoplot()` with `type = "pred"`

```
Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00172
no-frailty Log-likelihood: -331.997
Log-likelihood: -329.39
LRT: 1/2 * pchisq(5.21), p-val 0.0112

Frailty summary:
                   estimate lower 95% upper 95%
Kendall's tau         0.104     0.011     0.236
Median concordance    0.102     0.011     0.233
E[logZ]               0.067     0.006     0.179
Var[logZ]             0.406     0.037     1.176
Attenuation           0.896     0.764     0.989
theta                 8.572     3.232    90.316
Confidence intervals based on the likelihood function
```

The coefficient estimates are similar to those of the gamma frailty fit. The "Frailty summary" part is quite different. For the positive stable distribution, the variance is not defined. However, Kendall's $\tau$ is easily obtained, and in this case it is smaller than in the gamma frailty model. Unlike the gamma or PVF distributions, the positive stable frailty predicts a marginal model with proportional hazards where the marginal hazard ratios are an attenuated version of the conditional hazard ratios shown in the output. The calculations are detailed in Appendix A1.

The conditional and marginal hazard ratios from different distributions can also be visualized

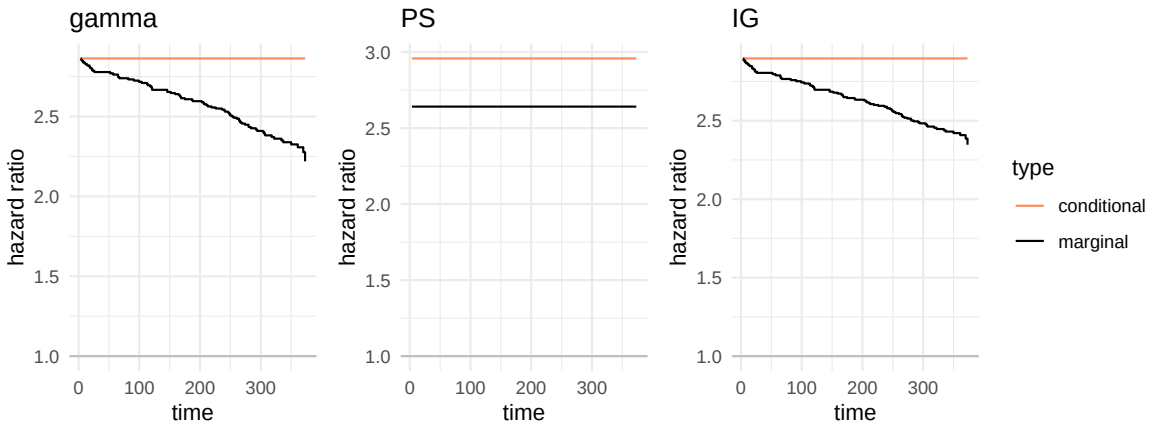



Figure 3: Conditional and marginal hazard ratio between two males from the placebo and rIFN-g treatment arms from the gamma, PS and IG frailty models as produced by `autoplot()` with `type = "hr"`.

easily. We also fitted an IG frailty model on the same data, and plots of the hazard ratio between two males from different treatment arms created below are shown in Figure 3.

```
R> ig <- emfrail(Surv(tstart, tstop, status) ~ sex + treat + cluster(id),
+   data = cgd,
+   distribution = emfrail_dist(dist = "pvf"))
R> newdata <- data.frame(treat = c("placebo", "rIFN-g"),
+   sex = c("male", "male"))
R> pl1 <- autoplot(gam, type = "hr", newdata = newdata) +
+   ggtitle("gamma") +
+   guides(colour = FALSE)
R> pl2 <- autoplot(stab, type = "hr", newdata = newdata) +
+   ggtitle("PS") +
+   guides(colour = FALSE)
R> pl3 <- autoplot(ig, type = "hr", newdata = newdata) +
+   ggtitle("IG")
R> pp <- ggarrange(pl1, pl2, pl3, nrow = 1)
```

While all models shrink the hazard ratio towards 1, it can be seen that this effect is slightly more pronounced for the gamma than for the IG, while the PS exhibits a constant "average" shrinkage. This type of behaviour from the PS is often seen as a strength of the model (Hougaard 2000).

## 4.2. Kidney

The `kidney` data set is also available in the **survival** package. The data, presented originally

in McGilchrist and Aisbett (1991), contains the time to infection for kidney patients using a portable dialysis equipment. The infection may occur at the insertion of the catheter and at that point, the catheter must be removed, the infection cleared up, and the catheter reinserted. Each of the 38 patients has exactly 2 observations, representing recurrence times from insertion until the next infection (i.e. the time scale is gap time). There are 3 covariates: sex, age and disease (a factor with 4 levels). A data analysis based on frailty models is described in Therneau and Grambsch (2000, ch. 9.5.2). For the purpose of illustration, we do not include the `disease` variable here.

```
R> data("kidney")
R> kidney <- kidney[c("time", "status", "id", "age", "sex" )]
R> kidney$sex <- ifelse(kidney$sex == 1, "male", "female")
R> head(kidney)
```

```
  time status id age    sex
1    8      1  1  28   male
2   16      1  1  28   male
3   23      1  2  48 female
4   13      0  2  48 female
5   22      1  3  32   male
6   28      1  3  32   male
```

```
R> zph_t = emfrail_control(zph = TRUE)
R> m_gam <- emfrail(Surv(time, status) ~ age + sex + cluster(id),
+   data = kidney, control = zph_t)
R> m_ps <- emfrail(Surv(time, status) ~ age + sex + cluster(id),
+   data = kidney,
+   distribution = emfrail_dist("stable"),
+   control = zph_t)
```

Therneau and Grambsch discuss the gamma fit conclude that an outlier case is at the source of the frailty effect. We omit the frailty part of the output; the estimated frailty variance is 0.397 with a 95% likelihood based confidence interval of $(0.04, 1.03)$ and therefore significantly different from 0.

```
R> summary(m_gam, print_opts = list(frailty = FALSE))
```

```
Call:
emfrail(formula = Surv(time, status) ~ age + sex + cluster(id),
    data = kidney, control = zph_t)

Regression coefficients:
           coef exp(coef) se(coef) adj. se       z    p
age     0.00544   1.00545  0.01158 0.01170 0.46481 0.64
sexmale 1.55284   4.72487  0.44518 0.49952 3.10868 0.00
Estimated distribution: gamma / left truncation: FALSE
```

```
Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00238
no-frailty Log-likelihood: -184.657
Log-likelihood: -182.053
LRT: 1/2 * pchisq(5.21), p-val 0.0112
```

However, the LRT is not significant for the positive stable frailty model (which does not have a defined frailty variance, for comparison). Furthermore, the estimated regression coefficients are different.

```
R> summary(m_ps, print_opts = list(frailty = FALSE))

Call:
emfrail(formula = Surv(time, status) ~ age + sex + cluster(id),
    data = kidney, distribution = emfrail_dist("stable"), control = zph_t)

Regression coefficients:
           coef exp(coef) se(coef)       z    p
age     0.00218   1.00218  0.00922 0.23649 0.81
sexmale 0.82100   2.27278  0.29873 2.74830 0.01
Estimated distribution: stable / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.00238
no-frailty Log-likelihood: -184.657
Log-likelihood: -184.657
LRT: 1/2 * pchisq(0), p-val>0.5
```

The test for proportional hazards described in Section 2.5 reveals an insight into how the two models work. The gamma frailty model specifies conditional proportional hazards and marginal non-proportional hazards, while the positive stable model specifies proportional hazards at both levels.

```
R> m_gam$zph

         chisq df     p
age     0.0599  1 0.807
sexmale 2.7504  1 0.097
GLOBAL  2.9649  2 0.227

R> m_ps$zph

        chisq df       p
age       0.1  1 0.75147
sexmale  11.1  1 0.00088
GLOBAL   12.2  2 0.00230
```

Therefore, the gamma frailty model appears to explain the marginal non-proportionality, while the positive stable frailty model does not. Such a phenomenon may be observed if, for example, the PS marginal model is a bad fit for the data. Further research is being carried out on this topic (Balan and Putter Forthcoming).

### 4.3. Rats data

These is an example of clustered failure data from Mantel, Bohidar, and Ciminera (1977) Three rats were chosen from each of 100 litters, one of which was treated with a drug (`rx = 1`) and the rest with placebo (`rx = 0`), and then all followed for tumor incidence. The data are also available in the **survival** package.

```
R> data("rats")
R> head(rats)
```

```
  litter rx time status sex
1      1  1  101      0   f
2      1  0   49      1   f
3      1  0  104      0   f
4      2  1   91      0   m
5      2  0  104      0   m
6      2  0  102      0   m
```

While often used to illustrate frailty models, the gamma frailty fit shows a relatively large, yet not significant frailty variance

```
R> summary(emfrail(Surv(time, status) ~ rx + sex + cluster(litter),
+                  data = rats))
```

```
Call:
emfrail(formula = Surv(time, status) ~ rx + sex + cluster(litter),
    data = rats)

Regression coefficients:
        coef exp(coef) se(coef) adj. se       z    p
rx    0.7873    2.1974   0.3135  0.3135  2.5112 0.01
sexm -3.1341    0.0435   0.7385  0.7409 -4.2298 0.00
Estimated distribution: gamma / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.201
no-frailty Log-likelihood: -200.426
Log-likelihood: -199.73
LRT: 1/2 * pchisq(1.39), p-val 0.119

Frailty summary:
                   estimate lower 95% upper 95%
```

```
Var[Z]                0.445     0.000     1.678
Kendall's tau         0.182     0.000     0.456
Median concordance    0.179     0.000     0.464
E[logZ]              -0.239    -1.038     0.000
Var[logZ]             0.559     0.000     3.678
theta                 2.245     0.596       Inf
Confidence intervals based on the likelihood function
```

The `Surv` object takes two arguments here: time of event and status. This implicitly assumes that each row of the data (in this case, each rat) is under follow-up from time 0 to `time`. This is very similar to the representation of the recurrent events in gap-time, where each recurrent event episode is "at risk" from time 0 (time since the previous event).

We artificially simulated left truncation from an exponential distribution with mean 50, which is now an entry time to the study. The rats with a follow-up smaller than the entry time are removed.

```
R> set.seed(1)
R> rats$tstart <- rexp(nrow(rats), rate = 1/50)
R> rats_lt <- rats[rats$tstart < rats$time, ]
```

The first model, `m1`, is what happens if left truncation is completely ignored. Each rat is assumed to have been at risk from time 0, which is not the case.

```
R> m1 <-
+   emfrail(Surv(time, status) ~ rx  + cluster(litter),
+           data = rats_lt)
```

The second model, `m2`, is what happens when the at-risk indicator is correctly adjusted for, with the entry time also present. Refering back to Section2.3, this is equivalent to considering $P(Z)$ instead of $P(Z|A)$.

```
R> m2 <-
+   emfrail(Surv(tstart, time, status) ~ rx + sex + cluster(litter),
+           data = rats_lt)
```

As may be seen from equation (6), this is correct only if there is in fact no left truncation, or if there is no variability in $Z$ (i.e. $Z$ is degenerate at 1). Therefore, this formulation is correct, for example, when the `Surv` object represents recurrent events in calendar time, as is the case in Section 4.1. This is, for example, what is returned by the frailty models in the **survival** package.

The third model, `m3`, specifies the correct time at risk but also the fact that the distribution of the frailty must be taken conditional on the entry time. Under this (artificial) left truncation problem, this would be the correct way of analyzing this data.

```
R> m3 <-
+   emfrail(Surv(tstart, time, status) ~ rx + sex + cluster(litter),
+           data = rats_lt,
+           distribution = emfrail_dist(left_truncation = TRUE))
```

In this case, the output shows little difference between models. This is because the frailty, even in the complete data set, is not significant. In this case, the frailty distribution is also not significant in either `m2` or `m3` and they lead to estimates very close to each other. In a limited unpublished simulation study, we have seen that applying the correction in `m3` leads to approximately unbiased estimates of the regression coefficients, unlike `m1` or `m2`.

## 5. Conclusion

In the current landscape for modeling random effects in survival analysis, **frailtyEM** is a contribution that focuses on implementing classical methodology in an efficient way with a wide variety of frailty distributions. We have shown that the EM based approach has certain advantages in the context of frailty models. First of all, it is semiparametric, which means that it is a direct extension of the Cox proportional hazards model. In this way, classical results from semiparametric frailty models (for example, based on the data sets in Section 4) can be replicated and further insight may be obtained by fitting models with different frailty distributions. Until now, the Commenges-Andersen test, positive stable and PVF family, have not all been implemented in a consistent way in an R package. Another advantage of the EM algorithm is that, by its nature, it is a full maximum likelihood approach, and the estimators have well known desirable asymptotic properties.

To our knowledge, no other statistical package provides similar capabilities for visualizing conditional and marginal survival curves, or the marginal effect of covariates. Since this is implemented across a large number of distributions, this might come to the aid of both applied and theoretical research into shared frailty models. While the question of model selection with different random effect distributions is still an open one, the functions included **frailtyEM** may be useful for further research in this direction.

Evaluating goodness of fit for shared frailty models is still a complicated issue, particularly in semiparametric models. However, tests based on martingale residuals, such as that of Commenges and Rondeau (2000), should be now possible by extrating the necessary quantities from an `emfrail` fit.

Regarding the left truncation implementation in **frailtyEM**, it is very similar to that from the **parfm** package. However, performing of a larger simulation study to assess the effects of left truncation in clustered failure data with semiparametric frailty models is now possible. In a limited simulation study we have seen that correctly accounting for this phenomenon leads to unbiased estimates. The scenario of time dependent covariates and left truncation is not supported at this time. This is because this would require also specifying values of these covariates from time 0 to the left truncation time, which would likely involve some speculation.

Technically, extending the package to other distributions is possible, as long as their Laplace transform and the corresponding derivatives may be specified in closed form. An interesting extension would be to choose discrete distributions from the infinitely divisible family for the random effect, such as the Poisson distribution. The newest features will be implemented in the development version of the package at `https://github.com/tbalan/frailtyEM`.

## Appendix A1: Results for the Laplace transforms

We consider distributions from the infinitely divisible family Ash (1972, ch 8.5) with the Laplace transform

$$\mathcal{L}_Y(c) = \exp(-\alpha\psi(c;\gamma)).$$

We now consider how $\alpha$ and $\gamma$ can be represented as a function of a positive parameter $\theta$.

**The gamma distribution** For $Y$ a gamma distributed random variable, $\psi(c;\gamma) = \log(\gamma + c) - \log(\gamma)$, the derivatives of which are

$$\psi^{(k)}(c;\gamma) = (-1)^{k-1}(k-1)!(\gamma+c)^{-k}.$$

For identifiability, the restriction $\mathsf{E}Y = 1$ is imposed; this leads to $\alpha = \gamma$. The distribution is parametrized with $\theta > 0$, $\theta = \alpha = \gamma$. The variance of $Y$ is $\mathsf{VAR}Y = \theta^{-1}$. Kendall's $\tau$ is then $\tau = \frac{1}{1+2\theta}$ and the median concordance is $\kappa = 4\left(2^{1+1/\theta} - 1\right)^{-\theta} - 1$. Furthermore, $\mathsf{E}\log Y = \psi(\theta) - \log\theta$ and $\mathsf{VAR}\log Y = \psi'(\theta)$ where $\psi$ and $\psi'$ are the digamma and trigamma functions.

**The positive stable distribution** For $Y$ a positive stable random variable, $\psi(c;\gamma) = c^\gamma$ with $\gamma \in (0,1)$, the derivatives of which are

$$\psi^{(k)}(c;\gamma) = \frac{\Gamma(k-\beta)}{\Gamma(1-\gamma)}(-1)^{k-1}c^{\gamma-1}.$$

For identifiability, the restriction $\alpha = 1$ is made; $\mathsf{E}Y$ is undefined and $\mathsf{VAR}Y = \infty$. The distribution is parametrized with $\theta > 0$, $\gamma = \frac{\theta}{\theta+1}$.

Kendall's $\tau$ is then $\tau = 1 - \frac{\theta}{\theta+1}$ and the median concordance is $\kappa = 2^{2-2^{\frac{\theta}{\theta+1}}} - 1$. Furthermore,

$$\mathsf{E}\log Y = -\left(\left\{\frac{\theta}{1+\theta}\right\}^{-1} - 1\right)\psi(1)$$

and

$$\mathsf{VAR}\log Y = \left(\left\{\frac{\theta}{1+\theta}\right\}^{-2} - 1\right)\psi'(1)$$

.

In the case of the PS distribution, the marginal hazard ratio is an attenuated version of the conditional hazard ratio. If the conditional log-hazard ratio is $\beta$, the marginal hazard ratio is equal to $\beta\frac{\theta}{\theta+1}$.

**The PVF distributions** For $Y$ a PVF distribution with fixed parameter $m \in \mathbb{R}$, $m > -1$ and $m \neq 0$,

$$\psi(c;\gamma) = \mathrm{sign}(m)(1 - \gamma^m(\gamma+c)^{-m})$$

where $\mathrm{sign}(\cdot)$ denotes the sign. This is the same parametrizaion as in Aalen *et al.* (2008). The derivatives of $\psi$ are

$$\psi^{(k)}(c;\gamma) = \mathrm{sign}(m)(-\gamma)^m(\gamma+c)^{-m-k}(-1)^{k+1}\frac{\Gamma(m+k)}{\Gamma(m)}.$$

The expectation of this distribution can be calculated as minus the first derivative of the Laplace transform calculated in 0, i.e.,

$$\mathsf{E}Y = \alpha\psi'(0;\gamma)\mathcal{L}(0;\alpha,\gamma) = \frac{\alpha}{\gamma}m.$$

The second moment of the distribution can be calculated as the second derivative of the Laplace transform at 0,

$$\mathsf{E}Y^2 = \alpha^2\psi'^2(0) - \alpha\psi''(0) = \frac{\alpha^2}{\gamma^2}m^2 + \frac{\alpha}{\gamma^2}m(m+1).$$

For identifiability, we set $\mathsf{E}Y = 1$. The distribution is parametrized through a parameter $\theta > 0$ which is determined by $\gamma = (m+1)\theta$ and $\alpha = \text{sign}(m)\frac{m+1}{m}\theta$. This results in $\mathsf{VAR}Y = \theta^{-1}$.

A slightly different parametrization is presented in Hougaard (2000), dependent on the parameter $\eta_H$. The correspondence is obtained by setting $\eta_H = (m+1)\theta$.

The PVF family of distributions includes the gamma as a limiting case when $m \to 0$. When $\gamma \to 0$ the positive stable distribution is obtained. When $m = -1$ the distribution is degenerate, and with $m = 1$ a non-central gamma distribution is obtained. Of special interest is the case $m = -0.5$, when the inverse Gaussian distribution is obtained. With $m > 0$, the distribution is compound Poisson with mass at 0. In this case, $P(Y = 0) = \exp(-\frac{m+1}{m}\theta)$.

For $m < 0$, closed forms for Kendall's $\tau$ and median concordance are given in Hougaard (2000, Section 7.5).

### Left truncation

To determine the Laplace transform under left truncation, we determine $\tilde{\psi}$ from (4) and (5).

For the gamma distribution, we have

$$\tilde{\psi}(c;\gamma,\Lambda_L) = \log(\gamma + \Lambda_L + c) - \log(\gamma + \Lambda_L)$$

which implies that the frailty of the survivors is still gamma distributed, but with a change in the parameter $\gamma$.

For the positive stable we have

$$\tilde{\psi}(c;\gamma,\Lambda_L) = (c + \Lambda_L)^\gamma - \Lambda_L^\gamma,$$

which is not a positive stable distribution any more.

For the PVF distributions, we have

$$\tilde{\psi}(c;\gamma,\Lambda_L) = \text{sign}(m)\left(\gamma^m(\gamma+\Lambda_L)^{-m} - (\gamma+\Lambda_L)^m(\gamma+\Lambda_L+c)^{-m}\right),$$

which is not PVF any more (however, it stays in the same "infinitely divisible" family).

### Closed forms

The gamma distribution leads to a Laplace transform for which the derivatives can be calculated in closed form. It can be seen that

$$\mathcal{L}(c;\alpha,\gamma) = \gamma^\alpha(\gamma+c)^{-\alpha}.$$

The $k$-th derivative of this expression is

$$\mathcal{L}^{(k)}(c;\alpha,\gamma) = \gamma^{\alpha}(\gamma+c)^{-\gamma-k}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}.$$

This can be exploited also in the case of left truncation, since the gamma frailty is preserved, as shown in the previous section.

The inverse gaussian distribution is obtained when the PVF parameter is $m = -\frac{1}{2}$. Under the current parametrization, we have $\beta = \theta/2$ and $\alpha = \theta$. In this case, the Laplace transform is

$$\mathcal{L}(c;\theta) = \exp\left\{\theta\left(1-\sqrt{1+2c/\theta}\right)\right\}.$$

The $k$-th derivative of this can be written as

$$\mathcal{L}^{(k)}(c;\theta) = (-1)^k\left(\frac{2}{\theta}c+1\right)^{-k/2}\frac{\mathcal{K}_{k-1/2}\left(\sqrt{2\theta\left(c+\frac{\theta}{2}\right)}\right)}{\mathcal{K}_{1/2}\left(\sqrt{2\theta\left(c+\frac{\theta}{2}\right)}\right)}$$

where $\mathcal{K}$ is the modified Bessel function of the second kind.

The `emfrail()` uses the closed form formulas when possible, by default.

## Appendix A2: The E step

For the E step $\beta$ and $\lambda_0$ are fixed, either at their initial values or at the values from the previous M step. Let $n_i = \sum_{j,k}\delta_{ijk}$ be the number of events in cluster $i$. The conditional distribution of $Z_i$ given the data is described by the Laplace transform

$$\mathcal{L}(c) = \frac{\mathsf{E}\left[Z_i^{n_i}\exp(-Z_i\tilde{\Lambda}_i)\exp(-Z_ic)\right]}{\mathsf{E}\left[Z_i^{n_i}\exp(-Z_i\tilde{\Lambda}_i)\right]} = \frac{\mathcal{L}^{(n_i)}(c+\tilde{\Lambda}_i)}{\mathcal{L}^{(n_i)}(\tilde{\Lambda}_i)}. \tag{9}$$

The E step reduces to calculating the expectation of this distribution, i.e. the derivative of (9) in 0:

$$\widehat{z}_i = -\frac{\mathcal{L}^{(n_i+1)}(\tilde{\Lambda}_i)}{\mathcal{L}^{(n_i)}(\tilde{\Lambda}_i)}. \tag{10}$$

The marginal (log-)likelihood is also calculated at this point to keep track of convergence of the EM algorithm. It can be seen that (3) involves the denominator of (9) in addition to a straight-forward expression of $\beta$ and $\lambda_0$.

The E step is generally the expensive operation of the EM algorithm. In a few scenarios (10) may be expressed in a closed form: for the gamma and the inverse gaussian distributions. In these scenarios, the E step is calculated with the `fast_estep()` routine. For all other cases, the E step is calculated via a recursive algorithm with an internal routine which is described here. For easing the computational burden, this is implemented in C++ and is interfaced with R via the **Rcpp** library (Eddelbuettel and François 2011; Eddelbuettel 2013).

As shown in (9), the calculation of the E step for the general case involves taking derivatives of Laplace transforms of the form

$$\mathcal{L}(c) = \exp(g(c))$$

where for simplicity we denote $g(c) = -\alpha\psi(c;\gamma)$. The expression for the $k$-th derivative of $\mathcal{L}(c)$ can be obtained with a classical calculus result, di Bruno's formula, i.e.,

$$\mathcal{L}^{(n)}(c) = \sum_{\mathbf{m}\in\mathcal{M}_n} \frac{n!}{m_1!m_2!...m_n!} \prod_{j=1}^{n} \left(\frac{g^{(j)}(c)}{j!}\right)^{m_j} \mathcal{L}(c), \tag{11}$$

where $\mathcal{M}_n = \{(m_1,...,m_n) | \sum_{j=1}^n j \times m_j = n\}$. For example, for $n = 3$,

$$\mathcal{M}_3 = \{(3,0,0),\ (1,1,0),\ (0,0,1)\}\,.$$

This corresponds to the "partitions of the integer" 3, i.e., all the integers that sum up to 3:

$$\{(1,1,1),\ (1,2,0), (3,0,0)\}\,.$$

We implemented a recursive algorithm in C++ which resides in the `emfrail_estep.cpp` which loops through these partitions, calculates the corresponding derivatives of $\psi$ and the coefficients.

## Appendix A3: Standard errors

Considering the vector of parameters $\eta = (\beta, \lambda_0(\cdot))$, and consider that for a given $\theta$, $\eta_\theta$ is the maximizer of the "inner problem" described in Section (3.2), i.e. $\eta(\theta) = \text{argmax}_\eta L(\eta|\theta)$. Further, for a given $\theta$, the variance-covariance matrix $\mathsf{VAR}(\eta(\theta))$ is obtained with Louis' formula (Louis 1982). The restulting standard errors for $\eta$ are underestimated because they do not factor in the uncertainty in estimating $\theta$, as is noted also in Therneau and Grambsch (2000, sec. 9.5). Below is the sketch of how this is addressed in **frailtyEM**, following Hougaard (2000, Appendix B.3).

Let $\hat\theta$ be the maximum likelihood estimate with variance $\mathsf{VAR}(\hat\theta)$ and standard error $\text{se}(\hat\theta)$, which are given by the maximizer from the "outer problem". The correct information matrix for inference on $\eta$ is a "perturbed" version of $\mathsf{VAR}(\eta(\hat\theta))$, namely

$$\mathsf{VAR}(\eta(\hat\theta)) + \left(\frac{d\eta}{d\theta}\right) \mathsf{VAR}(\hat\theta) \left(\frac{d\eta}{d\theta}\right)^\top.$$

Here, $d\eta/d\theta$ may be approximated as $(\eta^+ - \eta^-)/\text{se}(\hat\theta)$ where $\eta^+ = \eta(\hat\theta + \text{se}(\hat\theta)/2)$ and $\eta^- = \eta(\hat\theta - \text{se}(\hat\theta)/2)$. In `emfrail`, this whole calculation takes place for $\log\theta$ for computational stability, and to avoid the edge problem when $\theta$ is close to 0.

Confidence intervals for the conditional cumulative hazard are obtained from the part of the variance-covariance matrix corresponding to $\lambda_0(\cdot)$, and confidence intervals for $\Lambda_0(t) = \sum_{s\le t}\lambda_0(t)$ are obtained with the usual formula. For confidence intervals, the delta method is used to calculate a symmetric confidence interval for $\log\Lambda_0(t)$ for all $t$, which is then exponentiated.

## References

Aalen O, Borgan O, Gjessing H (2008). *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag New York. doi:10.1007/978-0-387-68560-1.

Andersen PK, Klein JP, Knudsen KM, y Palacios RT (1997). "Estimation of variance in Cox's regression model with shared gamma frailties." *Biometrics*, pp. 1475–1484.

Ash RP (1972). *Real Analysis and Probability*. Academic press.

Balan TA, Boonk SE, Vermeer MH, Putter H (2016a). "Score Test for Association Between Recurrent Events and a Terminal Event." *Statistics in Medicine*, **35**(18), 3037–3048. doi:10.1002/sim.6913.

Balan TA, Jonker MA, Johannesma PC, Putter H (2016b). "Ascertainment Correction in Frailty Models for Recurrent Events Data." *Statistics in Medicine*, **35**(23), 4183–4201. doi:10.1002/sim.6968.

Balan TA, Putter H (2017). ***frailtyEM**: Fitting Frailty Models with the EM Algorithm*. R package version 0.5.4, URL https://CRAN.R-project.org/package=frailtyEM.

Balan TA, Putter H (Forthcoming). *Non-proportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference?*

Commenges D, Andersen PK (1995). "Score Test of Homogeneity for Survival Data." *Lifetime Data Analysis*, **1**(2), 145–156. doi:10.1007/BF00985764.

Commenges D, Rondeau V (2000). "Standardized martingale residuals applied to grouped left truncated observations of dementia cases." *Lifetime Data Analysis*, **6**(3), 229–235.

Cook RJ, Lawless J (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media.

Cox DR (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society B*, **34**(2), 187–220. ISSN 00359246. URL https://www.jstor.org/stable/2985181.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, pp. 1–38.

Do Ha I, Noh M, Lee Y (2012). "**frailtyHL**: A Package for Fitting Frailty Models with h-likelihood." *R Journal*, **4**(2), 28–36.

Donohue MC, Overholser R, Xu R, Florin V (2011). "Conditional Akaike Information under Generalized Linear and Proportional Hazards Mixed Models." *Biometrika*, (98, 3), 685–700. doi:10.1093/biomet/asr023.

Donohue MC, Xu R (2013). ***phmm**: Proportional Hazards Mixed-effects Models*. R package version 0.7-5.

Eddelbuettel D (2013). *Seamless R and C++ Integration with Rcpp*. Springer-Verlag New York. doi:10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7.

Eddelbuettel D, François R (2011). "**Rcpp**: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08. URL https://www.jstatsoft.org/v40/i08/.

Gorfine M, Zucker DM, Hsu L (2006). "Prospective Survival Analysis with a General Semiparametric Shared Frailty Model: A Pseudo Full Likelihood Approach." *Biometrika*, pp. 735–741.

Grambsch PM, Therneau TM (1994). "Proportional hazards tests and diagnostics based on weighted residuals." *Biometrika*, **81**(3), 515–526.

Hougaard P (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York. doi:10.1007/978-1-4612-1304-8.

IBM Corp (2016). *IBM SPSS Statistics for Windows, Version 24.0*. IBM Corp, Armonk, NY. URL https://www.ibm.com/analytics/spss-statistics-software/.

Jackson CH (2011). "Multi-State Models for Panel Data: The **msm** Package for R." *Journal of Statistical Software*, **38**(8), 1–29. doi:10.18637/jss.v038.i08. URL https://www.jstatsoft.org/v38/i08/.

Jensen H, Brookmeyer R, Aaby P, Andersen PK (2004). *Shared frailty model for left-truncated multivariate survival data*. Department of Biostatistics, University of Copenhagen.

Klein JP (1992). "Semiparametric Estimation of Random Effects using the Cox Model based on the EM Algorithm." *Biometrics*, pp. 795–806.

Lin DY, Wei LJ, Ying Z (1993). "Checking the Cox model with cumulative sums of martingale-based residuals." *Biometrika*, **80**(3), 557–572.

Louis TA (1982). "Finding the Observed Information Matrix When Using the EM Algorithm." *Journal of the Royal Statistical Society B*, pp. 226–233.

Mantel N, Bohidar NR, Ciminera JL (1977). "Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information." *Cancer Research*, **37**(11), 3863–3868.

McGilchrist C, Aisbett C (1991). "Regression with Frailty in Survival Analysis." *Biometrics*, pp. 461–466. doi:10.2307/2532138.

Monaco JV, Gorfine M, Hsu L (2017). ***frailtySurv**: General Semiparametric Shared Frailty Model*. R package version 1.3.2, URL https://CRAN.R-project.org/package=frailtySurv.

Munda M, Rotolo F, Legrand C, *et al.* (2012). "**parfm**: Parametric Frailty Models in R." *Journal of Statistical Software*, **51**(1), 1–20. doi:10.18637/jss.v051.i11.

Nielsen GG, Gill RD, Andersen PK, Sørensen TI (1992). "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models." *Scandinavian Journal of Statistics*, pp. 25–43.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rodríguez-Girondo M, Deelen J, Slagboom EP, Houwing-Duistermaat JJ (2016). "Survival Analysis with Delayed Entry in Selected Families with Application to Human Longevity." *Statistical Methods in Medical Research*, p. 0962280216648356.

Rondeau V, Gonzalez JR (2005). "**frailtypack**: A computer program for the analysis of correlated failure time data using penalized likelihood estimation." *Computer Methods and Programs in Biomedicine*, **80**(2), 154–164. doi:10.1016/j.cmpb.2005.06.010.

Rondeau V, Mazroui Y, Gonzalez JR (2012). "**frailtypack**: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation." *Journal of Statistical Software*, **47**(4), 1–28. doi:10.18637/jss.v047.i04. URL https://www.jstatsoft.org/v47/i04/.

SAS Institute Inc (2003). *SAS/STAT Software, Version 9.1*. Cary, NC. URL https://www.sas.com/.

StataCorp (2017). *Stata Statistical Software: Release 15*. StataCorp LLC, College Station, TX. URL https://www.stata.com/.

Therneau TM (2015a). *A Package for Survival Analysis in S*. Version 2.38, URL https://CRAN.R-project.org/package=survival.

Therneau TM (2015b). ***coxme**: Mixed Effects Cox Models*. R package version 2.2-5, URL https://CRAN.R-project.org/package=coxme.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, New York. ISBN 0-387-98784-3. doi:10.1007/978-1-4757-3294-8.

Therneau TM, Grambsch PM, Fleming TR (1990). "Martingale-based residuals for survival models." *Biometrika*, **77**(1), 147–160.

Therneau TM, Grambsch PM, Pankratz VS (2003). "Penalized Survival Models and Frailty." *Journal of Computational and Graphical Statistics*, **12**(1), 156–175. ISSN 10618600. doi:10.1198/1061860031365. URL https://www.jstor.org/stable/1391074.

Vaida F, Xu R (2000). "Proportional Hazards Model with Random Effects." *Statistics in Medicine*, (19), 3309–3324.

Vaupel JW, Manton KG, Stallard E (1979). "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography*, **16**(3), 439–454. doi:10.2307/2061224.

Zhi X, Grambsch PM, Eberly LE (2005). "Likelihood Ratio Test for the Variance Component in a Semi-Parametric Shared Gamma Frailty Model." *Research Report 2005-5*.

## Affiliation:

Theodor Adrian Balan
Department of Biomedical Data Sciences
Leiden University Medical Center
2300 RC Leiden, The Netherlands
E-mail: t.a.balan@lumc.nlhttps://tbalan.github.io/