frailtyEM: an R Package for Estimating Semiparametric Shared Frailty Models

Theodor Adrian Balan
Leiden University Medical Center

Hein Putter
Leiden University Medical Center

Abstract

The software support for fitting so-called "frailty" (random effect) models for time-to-event data has grown considerably in the previous years. Such models are attractive to use when modeling recurrent event data or clustered failures. The usual problem specific to mixed models, which is integrating over the random effects, is further complicated by the presence of a non-parametric "baseline" intensity function. So far, the support for such semi-parametric models was limited, both in terms of the choice for the random effect distribution and in terms of the type of data that the model can be fitted on. We propose a new R package that estimates shared frailty models using the full likelihood, based on the Expectation-Maximization algorithm. The software supports a large number of distributions for the random effect from the Power Variance Family (PVF). Left truncated clustered failures and recurrent events in Andersen-Gill or gaptime formulation are also supported, and conditional and marginal estimates of the survival and cumulative hazard are provided.

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

1. Introduction

Time-to-event data is very common in medical applications. Often, these data are marked by incomplete observations. For example, the phenomena 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 semi-parametric in nature: the effect of the covariates is assumed to be time-constant and fully parametric, while the time-dependency arises from the non-parametric baseline hazard. Cox regression has been the standard in survival analysis for a few reasons: the non-parametric baseline is the best one can do if this function is not known in advance, and the estimation is not computationally intensive. For a long time, this has been implemented in all major statistical software, such as R (R Core Team 2016) (from S-PLUS times), Stata, SAS or SPSS.

When individuals belong to clusters, or may experience recurrent events, the observations are correlated, and in this case the Cox model is not appropriate. Random effect "shared frailty"

models have been developed for dealing with such situations. 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. The most popular distribution for the random effect is the gamma distribution, chosen mostly for computational convenience. A variety of distributions that have desirable properties have been proposed; see Hougaard (2000) for a comprehensive overview. These include the gamma, positive stable (PS), inverse gaussian (IG) and the Power Variance Family (PVF), which includes compound Poisson distributions with mass at 0.

Several surprising results have been demonstrated with frailty models. For example, if proportional hazards are assumed *conditional* on the frailty, then this assumption does not hold on the *marginal* level, for most distributions except the PS. In this case, the model fits proportional hazards on both conditional and marginal levels, with the marginal effect always being an attenuated version of the conditional. In fact, the choice of the distribution for the frailty implies a different marginal model, with some emphasizing early depence of the observations (IG) and others the late dependence (gamma). Therefore, it is of great interest to be able to compare a number of different distributions of the random effect.

For the Cox model, the computational advantage comes from the fact that the semi-parametric (infinitely dimensional) baseline hazard is not directly estimated, and this is due to the proportional hazards assumption. This simplicity does not carry over to shared frailty models. In this paper we present **frailtyEM** (Balan and Putter 2017), an R package which uses the general Expectation-Maximization (EM) algorithm for fitting shared frailty models. This implementation comes to complete the landscape of packages that may be used for such models. At the time of writing this manuscript, in R, semi-parametric shared frailty models can also be fitted in other ways. The first is via a penalized likelihood method with the **survival** package for the gamma and log-normal distributions (Therneau and Grambsch 2000; Therneau 2015a) and **coxme** package for the log-normal distribution (Therneau 2015b). The second way is via h-likelihood with the **frailtyHL** (Do Ha, Noh, and Lee 2012) package, and the third way is via a pseudo full likelihood approach **frailtySurv** package (Monaco, Gorfine, and Hsu 2017; Gorfine, Zucker, and Hsu 2006). Finally, a Monte Carlo EM-type estimation is available in the **phmm** Donohue and Xu (2013); Vaida and Xu (2000); Donohue, Overholser, Xu, and Florin (2011). Several other options are available for parametric modeling: **parfm** (Munda, Rotolo, Legrand *et al.* 2012) for fully parametric models and **frailtypack** (Rondeau, Mazroui, and Gonzalez 2012; Rondeau and Gonzalez 2005) for models where the baseline is fully parametric or spline-approximated. The latter is comparable in flexibility with semi-parametric models, where the baseline hazard is unspecified.

The **frailtyEM** package estimates semi-parametric shared frailty models that may be used for recurrent events data in Andersen-Gill and gaptime formulation, clustered failures and clustered failures with left truncation. The supported family of distributions for the random effect includes gamma, IG, PS and the PVF family. The results of the estimation can be easily visualized. Point estimates for regression coefficients are provided with confidence intervals that take into account the estimation of the frailty distribution, and plot methods may be used to visualize both conditional and marginal survival and 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

semi-parametric shared frailty model, and the implications of left truncation. In Section 3 we discuss the estimation method used and how this is implemented. In Section 4 we illustrate the usage of the functions from the **frailtyEM** package on two classical data sets available in R.

2. Model

We consider the following scenario: there are I clusters and J_i individuals in cluster i. The outcome from each individual is represented by a realization of a counting process N_{ij}. We consider that the intensity of N_{ij} takes the form

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

where $Y_{ij}(t)$ indicates whether N_{ij} is under observation at time t, Z_i is an unobserved random effect common to all individuals from cluster i (the "shared frailty"), $\mathbf{x}_{ij}(t)$ a vector of possibly time-dependent covariates, β a vector of unknown regression coefficients and $\lambda_0(t) > 0$ an unspecified baseline intensity function. We assume that event times are independent given $Z_i = z_i$. We consider the general case where the z_i 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

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

with $\alpha > 0$ and $\gamma > 0$. This formulation includes several distributions, such as the gamma, PS, IG, PVF. These distributions have been extensively studied in Hougaard (2000). Denote $\theta = (\alpha, \gamma)$ as the parameter vector that describes the distribution. The parametrizations used are described in Appendix A1.

2.1. Likelihood

The maximum likelihood problem is to maximize the marginal likelihood, based only on the observed data. This is obtained by integrating over the random effects. With the specification (1), the marginal likelihood is obtained as the product over clusters of expected marginal contributions, i.e.,

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

To make the connection with the data representation, we consider that (i, j, k) refers to the k-th observation from the j-th individual in the i-th cluster. Thus, t_{ijk} is the event or censoring time and $\delta_{ijk} = dN_{ij}(t_{ijk})$ is the event indicator for (i, j, k). We write the value of the covariate vector for this observation as $\mathbf{x}_{ijk}$. In the most basic case of clustered failures, $k \equiv 1$, while in the case of recurrent events $j \equiv 1$. More observations for one individual may also arise in the case of clustered failures when the covariates are time-dependent, and the individual

	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								
semi-parametric	yes	yes	yes	yes	yes	no	no	yes
posterior frailties	yes	yes	no	no	no	yes	no	no
conditional Λ_0, S_0	yes	limited	no	yes	no	yes	yes	no
marginal Λ_0, S_0	yes	no	no	no	no	no	no	no

Table 1: Comparsion of R packages for frailty models. Versions: **frailtyEM** 0.4.8, **survival** 2.40-1, **coxme** 2.2-5, **frailtyHL** 1.1, **frailtypack** 2.10.5, **parfm** 2.7.1, **phmm** 0.7-5

is artificially censored at the time when the value of the covariates changes. Nevertheless, the (i, j, k) pair refers to a certain cluster, individual, and period of time where the covariate vector does not change.

The baseline cumulative hazard for this observation is denoted as $\Lambda_{0,ijk}$. Also, let $\tilde{\Lambda}_i = \sum_{jk} \exp(\beta' \mathbf{x}_{ijk}) \Lambda_{0,ijk}$. The marginal likelihood can be written as

$$L(\theta, \beta, \lambda_0(\cdot)) = \prod_i \mathrm{E}_\theta \left[\prod_j \left\{ \prod_k (Z_i \exp(\beta^\top \mathbf{x}_{ijk}) \lambda_0(t_k))^{\delta_{ijk}} \right\} \exp(-z_i \tilde{\Lambda}_i) \right].$$

We consider the Breslow estimator for the baseline hazard, i.e., $\lambda_0(t) \equiv \lambda_{0t}$ for t an event time, and 0 otherwise. By using (2), the marginal likelihood can be rewritten as

$$L(\theta, \beta, \lambda_0(\cdot)) = \prod_i \left[\prod_j \left\{ \prod_k (\exp(\beta^\top \mathbf{x}_{ijk}) \lambda_0(t_k))^{\delta_{ijk}} \right\} (-1)^{n_i} \mathcal{L}_Z^{(n_i)}(\tilde{\Lambda}_i) \right], \tag{3}$$

where $\mathcal{L}_Z^{(k)}$ is the k-th derivative of the Laplace transform and n_i is the total number of events in cluster i.

2.2. Ascertainment and left truncation

The problem of ascertainment with random effect time-to-event data is usually a difficult one. Consider that the event of observing the cluster i in the data set is A_i. Thus, the distribution of the random effect in cluster i is described by the Laplace transform of $Z_i|A_i$, which follows from Bayes' rule as

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

Expressing $P(A_i|Z = z)$ 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 classical scenario of left truncation, i.e., where

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

Assume that, given z_i, the left truncation times $\mathbf{t}_{L,i}$ are independent and the cluster size is not informative. In this case,

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

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. To assign a value for $\mathbf{x}$ before the entry time is speculative. Therefore, we only consider this case when $\mathbf{x}_i$ is time constant.

With the previous notation, denote the risk accumulated before each of the entry times of cluster i as

$$\tilde{\Lambda}_{L,i} = \sum_j \exp(\beta^\top \mathbf{x}_{ij}) \Lambda_{0L,ij}$$

where $\Lambda_{0L,ij} = \int_0^{t_{L,ij}} \lambda_0(t)dt$. Then, it follows from (2), (4) and (5) that the Laplace transform can be written as

$$\mathcal{L}_{Z|A_i}(c;\alpha,\gamma) = \frac{\exp(-\alpha\psi(c+\tilde{\Lambda}_{L,i};\gamma))}{\exp(-\alpha\psi(\tilde{\Lambda}_{L,i};\gamma))} = \exp(-\alpha\tilde{\psi}(c;\tilde{\Lambda}_{L,i},\gamma)) \tag{6}$$

where $\tilde{\psi}(c;\Lambda_{L,i},\gamma) = \psi(c+\Lambda_{L,i};\gamma) - \psi(\Lambda_{L,i};\gamma)$. Thus, the random effect stays in the same infinitely divisible family of distributions under this ascertainment scheme.

Note that, in general, the ascertainment scheme does not have a simple description and $P(A_i|Z_i = z_i)$ 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**.

2.3. Goodness of fit and measures of dependence

A reasonable question when fitting random effect models is whether there is evidence for heterogeneity. To answer this *a priori* , the Commenges-Andersen score test may be used (Commenges and Andersen 1995). This test is referred in **frailtyEM** as the Commenges-Andersen test, and is performed before the actual estimation of the model (1). This test does not depend on the frailty distribution. The user may opt to skip this, or to just perform the test without fitting the shared frailty model in the `.control` argument of `emfrail()`.

After fitting the model, the likelihood ratio test may be used to assess whether the model with the frailty is a better fit than a model without frailty. In this case, the null model is the model without frailty. With the parametrizations described in Appendix A1, this test lies at 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).

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.

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 associtaion include Kendall's τ and the median concordance. A thorough discussion and comparison of these measures can be found in Hougaard (2000).

3. Estimation

frailtyEM implements a general full-likelihood estimation procedure for the gamma, positive stable and PVF frailty models, based on a profile likelihood method and making use of the expectation-maximization (EM) algorithm Dempster, Laird, and Rubin (1977).

For fixed parameters of the frailty distribution θ, we define the profile maximum likelihood

$$\widehat{L}(\theta) = \max_{\beta,\lambda_0} L(\beta, \lambda_0|\theta).$$

For each θ, denote $\hat{\beta}(\theta)$ and $\hat{\lambda}_0(\theta)$ the value of the parameters that maximize $L(\beta, \lambda_0|\theta)$. A first observation is that, if $\hat{\theta}$ maximizes $L(\theta)$, then $(\hat{\theta}, \hat{\beta}(\hat{\theta}), \hat{\lambda}_0(\hat{\theta}))$ maximize $L(\theta, \beta, \lambda_0)$. Thus, we split the problem of maximizing the likelihood into two: obtaining $\hat{\beta}(\theta), \hat{\lambda}_0(\theta)$ for a fixed θ (the "inner problem") and maximizng $L(\theta)$ over θ (the "outer problem").

3.1. The inner problem

For the inner problem the EM algorithm is used. This has been first proposed for the gamma frailty model in Nielsen, Gill, Andersen, and Sørensen (1992) and Klein (1992), and a generalization is discussed in Hougaard (2000).

Most ideas from Nielsen *et al.* (1992) are used here. The crucial observations are that the E step involves calculating the empirical Bayes estimates of the frailties $\widehat{z}_i = \mathrm{E}[Z_i|data]$. The expectation is taken with respect to the "posterior" distribution of the random effect. Afterwards, the M step is essentially a proportional hazards model with the $\log \widehat{z}_i$ as offset for each cluster.

The E step For the E step β and λ_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 has Laplace transform

$$\mathcal{L}(c) = \frac{\mathrm{E}\left[Z_i^{n_i} \exp(-Z_i\tilde{\Lambda}_i)\exp(-Z_i c)\right]}{\mathrm{E}\left[Z_i^{n_i}\exp(-Z_i\tilde{\Lambda}_i)\right]} = \frac{\mathcal{L}^{(n_i)}(c+\tilde{\Lambda}_i)}{\mathcal{L}^{(n_i)}(\tilde{\Lambda}_i)}. \tag{7}$$

The E step reduces to calculating the derivative of (7) in 0, i.e.,

$$\widehat{z}_i = -\frac{\mathcal{L}^{(n_i+1)}(\tilde{\Lambda}_i)}{\mathcal{L}^{(n_i)}(\tilde{\Lambda}_i)}. \tag{8}$$

The marginal (log-)likelihood is also calculated at this point, $L_\theta(\beta, \lambda_0)$ to keep track of convergence. It can be seen that (3) involved only the denominator of (7) in addition to a straight forward expression of β and λ_0.

The E step is generally the expensive operation of the EM algorithm. In very few scenarios can (8) 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 `estep()`, which is described in Appendix A2. For efficiency and speed, this function was written in C++ and is interfaced with R via the **Rcpp** library (Eddelbuettel and François 2011; Eddelbuettel 2013).

The M step With the same argument as made in Nielsen *et al.* (1992), the M step is equivalent to a regular proportional hazards model with $\log \widehat{z}_i$ added as an offset for all the cases in z_i. This is done via the `agreg.fit()` function in the **survival** package. Estimates of β are directly obtained from this, while estimates for λ_0 and the subsequent calculations of $\tilde{\Lambda}_i$

(and, eventually $\tilde{\Lambda}_{L,i}$, in the case of left truncation) require a careful calculation of subjects at risk and the respective linear predictors at every event and entry time point. The ordering required for determining these "at risk" sets is cached in `emfrail()`.

The EM algorithm stops after the marginal log-likelihood has converged, i.e., when difference in $\widehat{L}(\theta)$ is smaller than ε between two consecutive iterations. The value of ε can be set with the `.control` argument of `emfrail`.

3.2. Outer problem

The "outer" problem refers to finding $\widehat{\theta}$ which maximizes the profile likelihood $\widehat{L}(\theta)$. The resulting $\widehat{\theta}$ is the maximum likelihood estimator and the maximum likelihood is obtained at $\widehat{L}(\widehat{\theta})$. For the infinitely divisible distributions in **frailtyEM**, θ is one dimensional.

In the maximization procedure, θ is introduced on the log-scale. That is for numerical stability (the interval for searching for the maximum likelihood is "stretched" to the real numbers) and for asymptotic normality. Although maximum likelihood estimates are asymptotically normal, the likelihood is likely to be skewed, especially if the maximum likelihood is close to the edge of the parameter space. In this case, the standard error of the estimate may be difficult to interpret for constructing confidence intervals. It has been shown that symmetric confidence intervals of $\log\theta$, translated to the scale of θ, provide good coverage (Balan *et al.* 2016b).

Several parameters may be used to regulate the outer optimization via the `.control` argument. These can be found in the documentation of the `emfrail_control()` function.

3.3. Standard errors

After the maximizer has converged and the outer maximization is finished and $\hat{\theta}$ has been obtained, the Hessian is approximated numerically with the functions available in the **numDeriv** package (Gilbert and Varadhan 2016). By inverting this value, the variance of $\widehat{\log\theta}$ is obtained. A symmetric 95% confidence interval is built on that scale, and translated for all the other derived quantities described in 2.3. Furthermore, the delta method is used to provide standard errors for these parameters as well, by using the `deltamethod()` implementation in the **msm** (Jackson 2011) package.

A more precise yet computationally intensive method for quantifying the uncertainty in $\widehat{\log\theta}$ or θ is through likelihood-based confidence intervals. This requires finding the θ values for which the difference between the maximum likelihood and the specific profile maximum likelihood values at θ equals a critical value, calculated from the $\chi^2(1)$ distribution, and is discussed in Appendix A2. This can be achieved with a root-finding routine such as the `uniroot()` function in the **stats** package.

The standard error of the estimates for β and $\lambda_0(\cdot)$ are calculated with Louis' formula (Louis 1982), for θ fixed to the maximum likelihood estimate. The resulting information matrix leads to an underestimate of the standard errors, because it does not account for the uncertainty in estimating θ. These standard errors are reported by the **survival** package for example, although Therneau and Grambsch (2000) recommend using the bootstrap for more precision. In **frailtyEM**, adjusted standard errors are obtained by calcuating the information matrix for β and λ_0 also at $\hat{\theta} \pm \varepsilon$. This is described in more detail in Appendix A3.

3.4. Output, summary and prediction

The return object type is `emfrail`, which is essentially a list that contains the results of the "outer" maximization, the results of the "inner" maximization at this estimate, and a few other fields which are used for different methods. The object type is documented in `?emfrail`. Some options to obtain only part of this object as an output are available via the `.control` argument.

By itself, an `emfrail` object prints the call, a summary of "outer" optimization, the estimates of the covariates and the *p*value of the Commenges-Andersen test. A more user-readable summary of an `emfrail` object is provided by the `summary.emfrail()` method. This returns an object of the class `emfrail_summary` that contains general fit information, covariate estimates and several distribution-specific measures of fit and dispersion described in Section 2.3. Arguments to `summary.emfrail()` may be used to show confidence intervals either likelihood based or delta method based, as described in Section 3.3.

A method for predicting cumulative hazard and survival curves, both conditional and marginal, exists in `predict.emfrail()`. Confidence bands are based on the asymptotic normality of the estimated λ_0, and available both for adjusted and un-adjusted for the uncertainty of θ. The user can specify which quantities to obtain and values of the linear predictor at which to calculate these curves. The function returns a data frame from which several plots can be easily created.

A few simple plot functions have been created for convenience, both using the base plot engines from the **graphics** package and the **ggplot2** package. An overview of the available plots may be found in `?plot_emfrail` and `?ggplot_emfrail`. These include `plot_pred()` for plotting marginal and conditional cumulative hazard or survival curves, `plot_hr()` for plotting marginal and conditional estimated hazard ratios, and `hist_frail()` for a histogram of the posterior estimates of the frailties. The same plots may be obtained with the **ggplot2** engine by adding the prefix `gg` to these functions. Furthermore, a scatter plot of the posterior estimates of the frailties may be obtained with `ggplot_frail()`, which also includes quantiles of the posterior distribution in the case of the gamma distribution.

An additional function is provided to calculate the marginal log-likelihood for a vector of values of θ, `emfrail_pll()`, without actually performing the outer optimizaion. This may be useful for visualizing the profile log-likelihood or when debugging (e.g., to see if the maximum likelihood estimate of θ lies on the boundary).

4. Illustration

The package is loaded in the usual way,

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

The features of the package will now be illustrated with two well-known data sets available in R.

4.1. CGD

The data are from a placebo controlled trial of gamma interferon in chronic granulotomous

disease (CGD). It contains the time to recurrence of serious infections observed, from randomizatio until end of study for each patient.

```
> data("cgd")
```

The variables of interest here are: `tstart`, `tstop` and `status` determine the outcome. The individual is identified by the variable `id`. For the purpose of illustration, we will use `treat` (treatment or placebo) and `sex` (female or male) as covariates, althought a larger number of variables are recorded in the data set.

A basic `emfrail` model can be fitted like this:

```
> m1 <- emfrail(.data = cgd,
+               .formula = Surv(tstart, tstop, status) ~ sex + treat + cluster(id))
```

The arguments of `emfrail` visible above are `.data` and `.formula`. The `.control` and `.distribution` are taken as defaults; for the latter, that is the gamma frailty distribution. The `.formula` argument contains a `Surv` object at the left hand side and a `+cluster()` statement on the right hand side (essentially as `+frailty()` in `coxph`). The `.distribution` and `.control` arguments must be objects of the type `emfrail_distribution` and `emfrail_control`, which are created by calls to functions with the same names. For example, the default choice for the distribution is:

```
> str(emfrail_distribution())
```

```
List of 4
 $ dist           : chr "gamma"
 $ theta          : num 2
 $ pvfm           : num -0.5
 $ left_truncation: logi FALSE
 - attr(*, "class")= chr "emfrail_distribution"
```

The `emfrail_distribution` objects have 4 fields: `dist` describes the distribution of the frailty (here, a gamma distribution), `theta` is the frailty parameter and the starting value for the optimization. The parametrizations are described in Appendix A1. The `pvfm` field only plays a role when `dist=="pvf"`, and describes which PVF family distribution should be used (default is -0.5, corresponding to the IG). Finally, `left_truncation` is a logical variable, on whether to treat an observation as left truncated or not. For example, in the case of recurrent events in Andersen-Gill format, this should be `FALSE`, because the "entry" time does not refer to ascertainment, and the frailty must not be taken conditional on not having had an event before that time point. The adjustment that is applied if left truncation is present is described in Section 2.2.

The `emfrail` object may be accessed with the `summary()` method:

```
> sm1 <- summary(m1, lik_ci = TRUE)
> sm1
```

```
Call:
emfrail(.data = cgd, .formula = Surv(tstart, tstop, status) ~
```

```
    sex + treat + cluster(id))

Regression coefficients:
               coef exp(coef) se(coef) adjusted se        z      p
sexfemale   -0.22750   0.79652  0.39565     0.39580 -0.57500 0.5653
treatrIFN-g -1.05208   0.34921  0.31037     0.31042 -3.38973 0.0007
Estimated distribution: gamma / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.0221
(marginal) no-frailty Log-likelihood: -331.997
(marginal) Log-likelihood: -326.619
LRT: 1/2 * pchisq(10.8), p-val 0.00052

Frailty summary:
theta = 1.218 (0.59) / 95% CI: [0.539, 4.326]
variance = 0.821 / 95% CI: [0.231, 1.855]
Kendall's tau: 0.291 / 95% CI: [0.104, 0.481]
Median concordance: 0.289 / 95% CI: [0.101, 0.491]
E[log Z]: -0.464 / 95% CI: [-1.165, -0.12]
Var[log Z]: 1.241 / 95% CI: [0.26, 4.346]
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 useful output according to the distribution. The calculations behind this section are described for each distribution in Appendix A1. Since only $\log\theta$ is actually estimated in the "outer" step, the delta method is employed to obtain standard errors for all derived quantities. The confidence intervals may be obtained either likelihood-based or delta method-based, see Appendix A3 for details. The delta method based confidence intervals are shown if `lik_ci = FALSE`. We found the profile likelihood confidence intervals more reliable, especially when the parameter estimates approach the edges of the parameter space.

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 is far from 0.

The results are almost identical to a gamma frailty fit from `coxph`. The marginal log-likelihood in the `emfrail` object is slightly higher, that is because the estimation of the parameters of the frailty distribution is more precise. In addition, `emfrail` also provides a 95% confidence interval for the frailty variance.

```
> m_cph <- coxph(Surv(tstart, tstop, status) ~ sex + treat + frailty(id),
+             data = cgd,
+             ties = "breslow")
> m_cph

Call:
coxph(formula = Surv(tstart, tstop, status) ~ sex + treat + frailty(id),
```

```
    data = cgd, ties = "breslow")

              coef se(coef)    se2  Chisq   DF       p
sexfemale   -0.227    0.396  0.330  0.330  1.0 0.56590
treatrIFN-g -1.051    0.308  0.264 11.673  1.0 0.00063
frailty(id)                       56.160 37.4 0.02495

Iterations: 6 outer, 27 Newton-Raphson
     Variance of random effect= 0.822   I-likelihood = -326.6
Degrees of freedom for terms=  0.7  0.7 37.4
Likelihood ratio test=98.8  on 38.8 df, p=3.98e-07  n= 203
```

The empirical Bayes frailty estimates are also identical for the two ways of fitting the model, as seen in Figure 4.1.

To illustrate the predicted cumulative hazard curves we take two individuals, one from the treatment arm and one from the placebo arm, both males. The two are shown in Figure 4.1

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, Borgan, and Gjessing (2008, ch. 7).

A similar model can be fitted with the positive stable distribution:

```
> m2 <- emfrail(.data = cgd,
+               .formula = Surv(tstart, tstop, status) ~ treat + sex + cluster(id),
+               .distribution = emfrail_distribution(dist = "stable"))
> summary(m2)

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

Regression coefficients:
               coef exp(coef) se(coef) adjusted se        z      p
treatrIFN-g -1.08462   0.33803  0.33188     0.33583 -3.26806 0.0011
sexfemale   -0.13710   0.87188  0.40689     0.40692 -0.33694 0.7362
Estimated distribution: stable / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.0221
(marginal) no-frailty Log-likelihood: -331.997
(marginal) Log-likelihood: -329.39
LRT: 1/2 * pchisq(5.21), p-val 0.0112

Frailty summary:
theta = 8.572 (5.41) / 95% CI: [3.232, 90.316]
Kendall's tau: 0.104 / 95% CI: [0.011, 0.236]
```

```
> plot(exp(m_cph$frail),
+      sm1$frail$z,
+      xlab = "frailty estimates (coxph)",
+      ylab = "frailty estimates (emfrail)")
> abline(0,1)
```

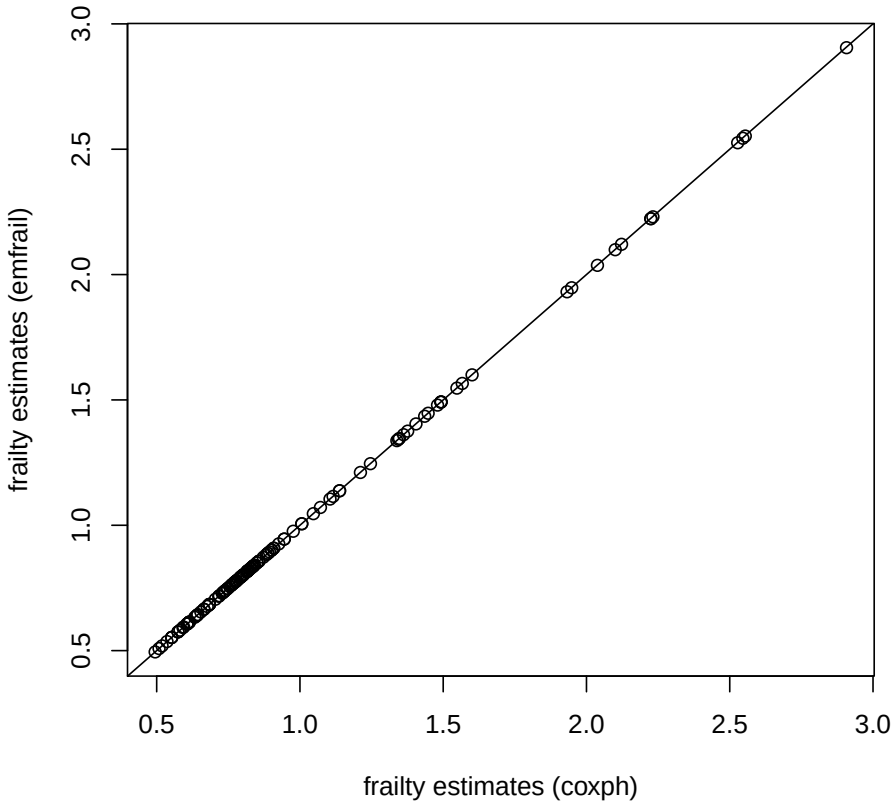

Figure 1: Scatterplot of the empirical Bayes frailty estimates from `emfrail()` versus those from `coxph()`

```
> library("ggplot2")
> p1 <- ggplot_pred(m1,
+           newdata = data.frame(sex = "male", treat = "rIFN-g")) +
+   ggtitle("rIFN-g") + ylim(c(0, 2)) +
+   theme_minimal()
> p2 <- ggplot_pred(m1,
+           newdata = data.frame(sex = "male", treat = "placebo")) +
+     ggtitle("placebo") + ylim(c(0, 2)) +
+     theme_minimal()
> gridExtra::grid.arrange(p1, p2, nrow = 1)
```

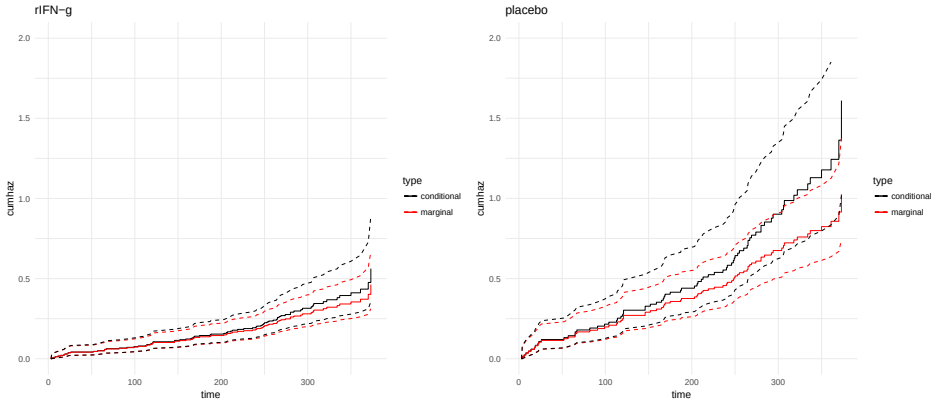

Figure 2: Predicted conditional and marginal cumulative hazards for males, one from the treatment arm and one from the placebo arm

```
> pl1 <- ggplot_hr(m1,
+         newdata = data.frame(treat = c("placebo", "rIFN-g"),
+                              sex = c("male", "male"))) +
+   ggtitle("gamma") +
+   theme_minimal()
> pl2 <- ggplot_hr(m2,
+         newdata = data.frame(treat = c("placebo", "rIFN-g"),
+                              sex = c("male", "male"))) +
+   ggtitle("stable") +
+   theme_minimal()
> gridExtra::grid.arrange(pl1, pl2, nrow = 1)
```

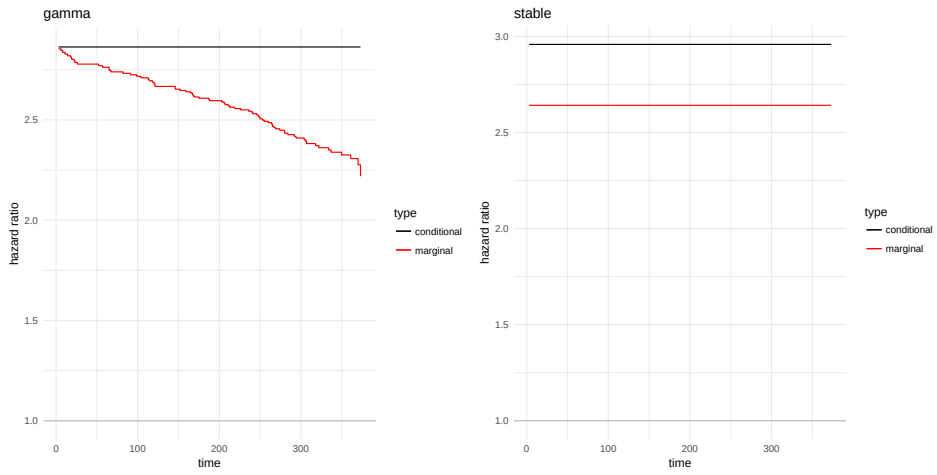

Figure 3: Conditional and marginal hazard ratio from the gamma and the positive stable frailty models

```
Median concordance: 0.102 / 95% CI: [0.011, 0.233]
E[log Z]: 0.067 / 95% CI: [0.006, 0.179]
Var[log Z]: 0.406 / 95% CI: [0.037, 1.176]
Attenuation factor: 0.896 / 95% CI: [0.764, 0.989]
Confidence intervals based on the likelihood function
```

The coefficient estimates are similar to those of `m1`. The "Frailty summary" part is quite different though. The positive stable distribution has infinite expectation. However, Kendall's τ 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 attenuation factor, shown in the output, is calculated as described in Appendix A1. This is discussed at length in Hougaard (2000) and it can be easily visualized with `emfrail`, as shown in Figure 3.

The plot shows that the marginal hazard ratio of the gamma frailty model is not time-constant, while the one from the positive stable frailty model is. This is discussed in Aalen *et al.* (2008, ch. 7). In Hougaard (2000) this is seen as a strength of the positive stable frailty model.

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. 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). The authors note that, when `disease` is included in the model, a gamma frailty model offers no evidence of heterogeneity. However, when `disease` is removed from the model, then there seems to be moderate evidence for heterogeneity. This is an example where the frailty may be interpreted as a missing covariate.

```
> data(kidney)
> kidney$sex <- ifelse(kidney$sex == 1, "male", "female")
> m_gam <- emfrail(.data = kidney,
+                  .formula = Surv(time, status) ~ age + sex + cluster(id))
> summary(m_gam)

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

Regression coefficients:
             coef exp(coef)  se(coef) adjusted se         z      p
age     0.0054372 1.0054520 0.0115813   0.0116976 0.4694817 0.6387
sexmale 1.5528412 4.7248755 0.4451769   0.4995172 3.4881440 0.0005
Estimated distribution: gamma / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.0245
(marginal) no-frailty Log-likelihood: -184.657
(marginal) Log-likelihood: -182.053
LRT: 1/2 * pchisq(5.21), p-val 0.0112

Frailty summary:
theta = 2.517 (1.49) / 95% CI: [0.97, 21.802]
variance = 0.397 / 95% CI: [0.046, 1.031]
Kendall's tau: 0.166 / 95% CI: [0.022, 0.34]
Median concordance: 0.162 / 95% CI: [0.022, 0.34]
E[log Z]: -0.212 / 95% CI: [-0.597, -0.023]
Var[log Z]: 0.486 / 95% CI: [0.047, 1.72]
Confidence intervals based on the likelihood function
```

Therneau and Grambsch discuss these models and they conclude that an outlier case is at the source of the frailty effect. With the **frailtyEM** package, the positive stable frailty model

may also be fitted. Unlike the gamma frailty model, the positive stable does not attempt to "correct" non-proportional hazards.

```
> m_stab <- emfrail(.data = kidney,
+                   .formula = Surv(time, status) ~ age + sex + cluster(id),
+                   .distribution = emfrail_distribution(dist = "stable"))
> summary(m_stab)

Call:
emfrail(.data = kidney, .formula = Surv(time, status) ~ age +
    sex + cluster(id), .distribution = emfrail_distribution(dist = "stable"))

Regression coefficients:
             coef exp(coef)  se(coef) adjusted se         z      p
age     0.0021816 1.0021839 0.0092248   0.0092248 0.2364892 0.8131
sexmale 0.8209988 2.2727687 0.2987240   0.2987245 2.7483521 0.0060
Estimated distribution: stable / left truncation: FALSE

Fit summary:
Commenges-Andersen test for heterogeneity: p-val  0.0245
(marginal) no-frailty Log-likelihood: -184.657
(marginal) Log-likelihood: -184.657
LRT: 1/2 * pchisq(-1.96e-05), p-val 0.5

Frailty summary:
theta = 105683.7 (33775246) / 95% CI: [2.879, Inf]
Kendall's tau: 0 / 95% CI: [0, 0.258]
Median concordance: 0 / 95% CI: [0, 0.255]
E[log Z]: 0 / 95% CI: [0, 0.2]
Var[log Z]: 0 / 95% CI: [0, 1.341]
Attenuation factor: 1 / 95% CI: [0.742, 1]
Confidence intervals based on the likelihood function
```

The Commenges-Andersen test for heterogeneity shows the same evidence as before, as it does not depend on the frailty distribution. However, the positive stable parameter lies at the edge of the parameter space (θ is between 0 and 1 for the PS distribution). Therefore, the LRT is not significant. The major difference with the gamma frailty fit is that the regression coefficient for sex is much smaller. To untangle this effect, one can check the (marginal) proportional hazards assumption. This reveals that `sex` has a significant non-proportional effect on the hazards:

```
> zph1 <- cox.zph(coxph(Surv(time, status) ~ age + sex + cluster(id), data = kidney))
> zph1

            rho   chisq        p
age      0.0214  0.0231 8.79e-01
sexmale -0.4390 29.2598 6.33e-08
GLOBAL       NA 29.3325 4.27e-07
```

In small samples, the gamma frailty model implicitly fits a marginal non-proportional hazards model, and in this case it succeeds. The PS distribution fits proportional hazards both conditionally and marginally, and in this case it fails. To untangle this effect, we can perform a proportional hazards test with the log-estimated frailties as an offset:

```
> s_gam <- summary(m_gam)
> off_z <- log(s_gam$frail$z)[match(kidney$id, s_gam$frail$id)]
> zph2 <- cox.zph(coxph(Surv(time, status) ~
+                         age + sex + offset(off_z) + cluster(id),
+                       data = kidney))
> zph2

            rho   chisq     p
age     -0.0145 0.00427 0.948
sexmale -0.2170 1.39043 0.238
GLOBAL      NA 1.41146 0.494
```

In this case, this is evidence that the gamma frailty corrects for proportionality rather than heterogeneity.

5. Conclusion

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 an 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.

Several options not discussed in this paper include the left truncation adjustment. There is no available data set to illustrate this option, however the peroforming of a larger simulation study to assess the effects of left truncation in clustered failure data is now possible.

Other extensions of this software are possible, since all that is needed is to specify the Laplace transform and the corresponding derivatives for the E step. 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/teddybalan/frailtyEM.

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. This comes to aid researches, as well as clinicians, facilitating the analysis of present and future studies.

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 α and γ can be represented as a function of a positive parameter θ.

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 $\mathrm{E}Y = 1$ is imposed; this leads to $\alpha = \gamma$. The distribution is parametrized with $\theta > 0$, $\theta = \alpha = \gamma$. The variance of Y is $\mathrm{Var}Y = \theta^{-1}$. Kendall's τ 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, $\mathrm{E}\log Y = \psi(\theta) - \log\theta$ and $\mathrm{Var}\log Y = \psi'(\theta)$ where ψ and ψ' 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; $\mathrm{E}Y$ is undefined and $\mathrm{Var}Y = \infty$. The distribution is parametrized with $\theta > 0$, $\gamma = \frac{\theta}{\theta+1}$.

Kendall's τ is then $\tau = 1 - \frac{\theta}{\theta+1}$ and the median concordance is $\kappa = 2^{2-2^{\frac{\theta}{\theta+1}}} - 1$. Furthermore, $\mathrm{E}\log Y = -\left(\left\{\frac{\theta}{1+\theta}\right\}^{-1} - 1\right)\psi(1)$ and $\mathrm{Var}\log Y = \left(\left\{\frac{\theta}{1+\theta}\right\}^{-2} - 1\right)\psi'(1)$.
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 β, 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 ψ 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.,

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

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

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

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

A slightly different parametrization is presented in Hougaard (2000), dependent on the parameter η_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 τ 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 (6).

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 γ.

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

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

which is not PVF any more.

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: A general E step

As shown in (7), 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.,

$$\mathcal{L}^{(n)}(c) = \sum_{\mathbf{m}\in\mathcal{M}_n} \frac{n!}{m_1!m_2!...m_n!} \prod_{j=1}^{n} \left(\frac{g^{(j)}(c)}{j!}\right)^{m_j} \mathcal{L}(c), \tag{9}$$

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 ψ and the coefficients.

Appendix A3: Standard errors

The outer maximization of $\widehat{L}(\theta)$ is carried out on the log-scale, as described in section 3, and the numeric hessian is used to obtain $\mathrm{Var}(\widehat{\theta})$. Afterwards, the delta method is employed to derive standard errors for θ and the other functionals of θ described in Appendix A1.

However, the standard error is not very meaningul for parameters with skewed distributions. Confidence intervals are constructed in two ways.

The first type of confidence intervals provided by **frailtyEM** are based on the the asymptotic normality of $\widehat{\log \theta}$, by constructing a 95% symmetric confidence interval on the log-scale, and then translating it to the other functionals of θ.

The second type are 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. Using the root-finding algorithm implemented in the `uniroot()` function in the **stats** package, a confidence interval is obtained from the values of θ with the property that $\widehat{L}(\theta) \geq \widehat{L}(\widehat{\theta}) - 1.92$. This confidence interval is then translated to the functionals of θ.

The likelihood-based confidence intervals are the default in `emfrail()` because the coverage is guaranteed to be the same for all transformations of θ.

Considering the vector of parameters $\eta = (\beta, \lambda_0(\cdot))$, the information matrix for (θ, η) can be written as follows:

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{\theta,\theta} & \mathcal{I}_{\theta,\eta} \\ \mathcal{I}_{\eta,\theta} & \mathcal{I}_{\eta,\eta} \end{bmatrix}.$$

The part corresponding to η, $\mathcal{I}_{\eta,\eta}$ is calculated using Louis' formula, which has been commonly employed to obtain this quantity from EM algorithms Louis (1982). This is done under the assumption of θ fixed to the maximum likelihood estimate $\hat{\theta}$. This leads to an underestimate of the standard errors, as is noted also in Therneau and Grambsch (2000, sec. 9.5). The calculation of the variance-covariance matrix $\mathcal{I}^{-1}$ in this case involves approximating $\mathcal{I}_{\eta,\theta}$ and adjusting $\mathcal{I}_{\eta,\eta}$, as described in Hougaard (2000, Appendix B.3) and Putter and Van Houwelingen (2015).

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 $\Lambda_0(t) = \sum_{s \leq t} \lambda_0(t)$ 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.

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.

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

Cox DR (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society B*, **34**(2), 187–220. ISSN 00359246. URL http://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 http://www.jstatsoft.org/v40/i08/.

Gilbert P, Varadhan R (2016). ***numDeriv**: Accurate Numerical Derivatives*. R package version 2016.8-1, URL https://CRAN.R-project.org/package=numDeriv.

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.

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

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 http://www.jstatsoft.org/v38/i08/.

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

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

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.

Putter H, Van Houwelingen HC (2015). "Dynamic Frailty Models Based on Compound Birth–Death Processes." *Biostatistics*, **16**(3), 550–564. doi:10.1093/biostatistics/kxv002.

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 http://www.jstatsoft.org/v47/i04/.

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.

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 Medical Statistics and Bioinformatics
Leiden University Medical Center
2300 RC Leiden, The Netherlands
E-mail: t.a.balan@lumc.nl