MOX-Report No. 45/2017

# Nonparametric frailty Cox models for hierarchical time-to-event data

Gasperoni, F.; Ieva, F.; Paganoni, A.M.; Jackson C.H.; Sharples L.D.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

# NONPARAMETRIC FRAILTY COX MODELS FOR HIERARCHICAL TIME-TO-EVENT DATA

Francesca Gasperoni[1], Francesca Ieva[1], Anna Maria Paganoni[1], Christopher H. Jackson[2], Linda D. Sharples[3]

**1** MOX-Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Milano, Italy

**2** MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 0SR, U.K

**3** Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London WC1E 7HT, U.K

**Keywords**: Discrete frailty; Expectation-Maximization algorithm; Finite mixture model; Multi-level survival data; Time-to-event data.

**Abstract**

In this work we propose a novel model for dealing with hierarchical time-to-event data, which is a common structure in healthcare research field (i.e., healthcare providers, seen as groups of patients). The most common statistical model for dealing with this kind of data is the Cox proportional hazard model with shared frailty term, whose distribution has to be specified a priori.

The main objective of this work consists in overcoming this limit by avoiding any a priori hypothesis on the frailty distribution. In order to do it, we introduce a nonparametric discrete frailty, through which we are not just guaranteeing a very good level of flexibility, but we are also building a probabilistic clustering technique, which allows to detect a clustering structure of groups, where each cluster is named latent population.

A tailored Expectation-Maximization algorithm, combined with model selection techniques, is proposed for estimating model's parameters.

Beyond the new methodological contribution, we propose a useful tool for exploring big hierarchical time-to-event data, where it is very difficult to explain all the phenomenon variability through explanatory covariates. We show the power of this model by applying it to a clinical administrative database, where several information of patients suffering from Heart Failure is collected, like age, comorbidities, procedures etc. In this way, we are able to detect a latent clustering structure among healthcare providers.

## 1 Introduction

Time-to-event methods are used extensively in medical statistics with the Cox proportional hazards model providing both flexibility and tractability, and requiring only that the proportional hazards assumption is valid (Cox, 1972). Extensions to this model to allow for the common situation of clustering of individuals (or shared frailty), for example due to repeated assessments of patients within the same healthcare provider, have been developed (Hougaard, 1984, 1986a,b). Published examples include survival of children grouped as siblings (Guo and Rodriguez, 1992), survival of patients grouped in hospitals (Austin, 2017) and time to udder infection in cows, with the four mammary glands making up the udder grouped as individuals (Duchateau and Janssen, 2007). These examples rely on a parametric form for the frailty distribution such as the Gamma or Log-Normal. However, a nonparametric alternative is desirable, due to potential misspecification of the parametric form and as a method for detecting clusters of groups with similar frailties, which is the goal of this work. In particular, we propose an extension of the shared frailty Cox model for hierarchical time-to-event data, in which a nonparametric frailty is included. We describe an EM algorithm to fit the model and investigate the properties of the model.

The underlying clinical motivation of this work is the analysis of times to admission to healthcare provider (such as hospital, research center or nursing home) in Heart Failure (HF) patients in the Lombardia region of Italy. Specifically, we analyzed a dataset extracted from a clinical administrative database, which included dates of admission and discharge and corresponding patient age, gender, comorbidities and survival. As the healthcare path of patients may depend on the structure, we included healthcare providers as a frailty or random effect. We aimed to detect clusters of healthcare providers with similar outcomes, without choosing the number of clusters in advance, and without specifying a parametric form for the baseline survival distribution. Thus, we investigate hierarchical semi-parametric time-to-event models, in which groups of healthcare providers are clustered into an unknown number of sets, each with the same frailty. To the best of our knowledge, there is no literature regarding healthcare providers profiling through the analysis of time-to-event data. Statistical profiling of healthcare providers is typically based on multilevel logistic regression of binary outcomes on patient-level and structure-level covariates, for example Grieco et al. (2011) in a frequentist framework or Ohlssen et al. (2007) and Guglielmi et al. (2014) in a Bayesian framework. Graphical approaches, such as funnel plots, have been used for healthcare provider performance classification and outlier detection (Spiegelhalter, 2005; Ieva and Paganoni, 2015).

Austin (2017) reviews models for multilevel time-to-event data and available software for implementing them. In particular, Cox models with Gamma and Log-Normal frailty distributions are discussed. Methods and software for these distributions are well-established (e.g. Therneau and Grambsch, 2013; Therneau, 2014, 2015). Positive stable and power variance distributions are also feasible (Duchateau and Janssen, 2007; Wienke, 2010; Hougaard, 2012).

However, only a few publications have dealt with discrete frailties, mostly applying the frailty at the individual level (univariate) and using a parametric baseline. A Weibull baseline was used both by dos Santos et al. (1995) and Caroni et al. (2010), while a piecewise constant baseline was used by Guo and Rodriguez (1992). Both Guo and Rodriguez (1992) and dos Santos et al. (1995) included a nonparametric frailty, while Caroni et al. (2010) investigated Geometric, Poisson and Negative Binomial distributed frailties. However, only Guo and Rodriguez (1992) dealt with a shared frailty, dos Santos et al. (1995) and Caroni et al. (2010) used individual-specific (as opposed to group-specific) frailties. Li et al. (1998) used a Cox proportional hazard model with a Bernoulli distributed frailty, which can be viewed as a nonparametric frailty model with the number of clusters set equal to two. Sy and Taylor (2000) proposed an extension of cure models, originally proposed by Farewell (1982), to include a Cox proportional hazards model for non-cured individuals. In this case, there is a mixture of susceptible and nonsusceptible (cured) individuals, so, the frailty is not shared but is individual-specific. Nonparametric frailty models can be seen as a finite mixture of parametric survival models, as was suggested by Laird (1978) and Heckman and Singer (1982, 1984b). The models by Guo and Rodriguez (1992) and by Li et al. (1998) can be viewed as a mixture of populations, where each population is composed of groups (e.g. hospitals), while dos Santos et al. (1995) and Sy and Taylor (2000) describe a mixture of populations of individuals (e.g. patients). Mixtures of survival models have also been investigated in a Bayesian framework, though mostly with parametric survival and frailty distributions (Ibrahim et al., 2005). Manda (2011) used a Dirichlet process prior for the frailty term, which automatically detects clusters among groups, but with a parametric baseline hazard.

The limited use of nonparametric frailty terms is linked to two major issues: the identifiability of model parameters and the lack of available software. Elbers and Ridder (1982) gave an overview of the well-known identifiability problem in the case of a univariate frailty term and provided conditions to guarantee model identifiability (i.e., finite mean distribution of the frailty and presence of one covariate which assumes at least two values). Heckman and Singer (1984a) extended some of the previous results, relaxing the condition on the existence of the mean and adding constraints on the cumulative baseline hazard. Indeed, according to Elbers and Ridder (1982), the classical approach in the case of a Gamma or Log-Normal frailty distribution constrains the mean of the frailty distribution to be equal to 1. The same constraint was required by Guo and Rodriguez (1992) in their model, while Sy and Taylor (2000) put a constraint on the cumulative baseline hazard (known as zero-tail constraint).

The Gamma and Log-Normal are often preferred among parametric frailty distributions, due to their analytical tractability and the availability of the related software, for example the package *coxph* (Therneau and Grambsch, 2013; Therneau, 2014) for the Gamma and *coxme* (Therneau, 2015) for the Log-Normal in R software (R Development Core Team, 2016). Recently Positive Stable and Power Variance distributions have also become accessible in R through the package *frailtyEM* (Balan and Putter, 2017). Almost all this software is based on the Expectation-Maximization algorithm, EM (Dempster et al., 1977; Klein, 1992), which is used extensively for parameter estimation in the frequentist framework, since Cox proportional hazards models with a frailty term can be seen as an incomplete data problem, where the observable data are the times-to-event or the censoring times, and the frailty values are the unobservable data. It can be difficult to compute the observed information matrix (Efron and Hinkley, 1978) from the observable log-likelihood. Louis (1982) proposed an approximation of this matrix that can be computed within the EM steps from the complete log-likelihood, as well as an accelerated version of the EM algorithm. Louis's method was fully exploited by Guo and Rodriguez (1992), and was used just for computing the observed information by both Li et al. (1998) and Sy and Taylor (2000). In this work we take advantage of Louis's method for computing the observed information matrix.

Through the EM algorithm, we are able to estimate the baseline hazard function, the regression coefficients and the frailty values. Separate methods are needed to estimate the number of latent populations in the data. Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are both widely used for model selection in mixture models (McLachlan and Peel, 2004). Guo and Rodriguez (1992) tackled this problem by initially applying the EM assuming two populations and increasing the number until they found a group with no members, following Laird (1978), who proposed to choose the number of clusters to be the maximum, if the number of clusters increases in the algorithm (or minimum, if the number of clusters decreases in the algorithm) for which each population is estimated to have one member, which favours more complex models than AIC and BIC.

In this work, we extend the shared frailty Cox model to take into account a nonparametric frailty term in the context of grouped time-to-event data. This means that the frailty does not have a continuous distribution, but a discrete distribution with an unknown number of elements in its support. This choice leads not only to a very flexible model (no strong parametric assumptions are required for the frailty and the baseline hazard) but also to a probabilistic clustering technique, which can be useful for exploring heterogeneity in survival between groups. This is particularly useful in large routinely-collected datasets where the emphasis is on large numbers of individuals, rather than detailed and accurate records for large numbers of covariates. As such the methods can identify groups (of healthcare providers, say) that have similar results for further detailed study of the reasons for the observed similarity.
To ensure the frailties are identifiable, we focus on estimating the ratio of frailties between groups.

For selection of the number of clusters, we compare AIC, BIC and the approach of Laird (1978). Our novel EM algorithm for parameter estimation and automatic model selection is available as R code upon request.

Finally, the paper is organized as follows: in Section 2 we present the mathematical model, the proposed Expectation Maximization algorithm is described in Section 3; a simulation study provides insights into the scope and limitations of the model in Section 4; while in Section 5 the model is applied to the regional clinical administrative database. Section 6 provides discussion of the results and the future perspectives.

## 2 Semiparametric Cox Model with a Nonparametric Frailty

Consider a random sample with a hierarchical structure, i.e, where each statistical unit belongs to one group. Define $T^*_{ij}$ as the survival time and $C_{ij}$ as the censoring time of subject $i$, $i = 1, ..., n_j$, in the $j$-th group, $j = 1, ..., J$. Let $\mathbf{X}_{ij} = (X_{ij1}, ..., X_{ijp})^T$ be the vector of covariates, assumed constant over time, for subject $i$ in group $j$. Then, we define $T_{ij} = min(T^*_{ij}, C_{ij})$, $t_{ij}$ its realization and $\delta_{ij} = \mathbf{1}_{(T^*_{ij} \leq C_{ij})}$. Let $\widetilde{\mathbf{w}}$ be the vector of shared random effects, and $\mathbf{w}$, $\mathbf{w} = \exp\{\widetilde{\mathbf{w}}\}$, be the vector of shared frailties (Rabe-Hesketh and Skrondal, 2008). In this work, we introduce a nonparametric frailty term, which can be modeled through a random variable with discrete distribution, with an unknown number of points in the support. In particular, we assume that each group $j$ can belong to one latent population $k$, $k = 1, ..., K$, with probability $\pi_k$. In this case, $w_1, ..., w_K$ are the points in the support of $w$, $K$ is the support's cardinality and $\mathbb{P}\{w = w_k\} = \pi_k$. In order to build the model, we introduce an auxiliary indicator random variable $z_{jk}$ which is equal to 1 if the $j$-th group belongs to the $k$-th population, so $z_{jk} \overset{i.i.d}{\sim} Bern(\pi_k)$. The requirement $\sum_{k=1}^{K} z_{jk} = 1$, for each $j$, is equivalent to the assumption that each group belongs to only one population. The vector $\mathbf{z}_j$ is distributed as a multinomial (Guo and Rodriguez, 1992). Note that there are two levels of clustering: the first one is known (i.e., healthcare providers as clusters of patients) and we refer to these clusters as groups, while the second level is the unknown clustering of healthcare providers that we want to detect and we refer to these clusters as latent populations.

The hazard function for individual $i$ in group $j$ is:

$$\lambda(t; \mathbf{X}_{ij}, w_k, z_{jk}) = \prod_{k=1}^{K} \Big[\lambda_0(t) w_k \exp\Big(\mathbf{X}_{ij}^T \boldsymbol{\beta}\Big)\Big]^{z_{jk}} \tag{2.1}$$

where $\lambda_0(t)$ represents the baseline hazard, $\boldsymbol{\beta}$ is the vector of regression coefficients and $w_k$ is the frailty term shared among groups of the same latent population $k$. Both the frailty and the baseline hazard are assumed to be nonparametric, which makes model (2.1) an extension of a proportional hazard Cox model. The observable data $\mathbf{Y}$ are made up of the set of $\mathbf{Y}_{ij} = \{T_{ij}, \delta_{ij}, \mathbf{X}_{ij}\}$ over all $i, j$. We define this as the "incomplete" data, while the "complete" data are the realizations of the vector $\{T_{ij}, \delta_{ij}, \mathbf{X}_{ij}, w_k, z_{jk}\}$. We also assume that censoring is noninformative, thus that $T^*_{ij}$ and $C_{ij}$ are conditionally independent, given $\mathbf{X}_{ij}$, $w_k$ and $z_{jk}$.
Starting from the hazard rate, we can to write down the full likelihood of our model for the complete data explicitly:

$$L_{full}(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{z}) = \prod_{k=1}^{K} \prod_{j=1}^{J} \pi_k^{z_{jk}} \cdot L_{full}^{jk}(\boldsymbol{\theta}; \mathbf{Y}_j|\mathbf{z}) \tag{2.2}$$

where

$$L^{jk}_{full}(\boldsymbol{\theta}; \mathbf{Y}_j|\mathbf{z}) = \prod_{i=1}^{n_j} \left\{ [\lambda_0(t_{ij}) w_k \exp\left(X^T_{ij}\beta\right)]^{\delta_{ij}} \cdot \exp\left[ -\Lambda_0(t_{ij}) w_k \exp\left(\mathbf{X}^T_{ij}\boldsymbol{\beta}\right)\right] \right\}^{z_{jk}} \tag{2.3}$$

and $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{w}, \lambda_0(t), \boldsymbol{\beta})$ is the vector of parameters, $\mathbf{z} := \{z_{jk}\}_{j=1:J}^{k=1:K}$ is the matrix of random vectors $\{z_j\}_{j=1:J}$ indicating membership of groups $j$ in populations $k$, and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the cumulative baseline hazard function.

This model can be interpreted as a shared frailty Cox model where the frailties are shared among latent populations, and also as a mixture model, where each component is a survival distribution, $\boldsymbol{\pi}$ is the vector of mixing proportions and $\mathbf{w}$ is the vector of component-specific frailties. Finally, the number of latent populations, $K$, can be considered as an unknown parameter, and the relative hazard between two individuals with the same covariate values but from different latent populations $k$ and $k^{\#}$ can be described by the *frailty ratio* $w_k/w_{k^{\#}}$. We note that the model as written is over-parameterised, since the same likelihood would result from multiplying $\lambda_0(t)$ by a constant $c$ while dividing all the $w_k$ by $c$, but identifiability is ensured within the estimation algorithm (Section 3.1).

# 3 Computation

## 3.1 A Tailored Expectation-Maximization Algorithm

We propose a novel Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to estimate $\theta$ for a given $K$. The algorithm iterates between two steps, Expectation and Maximization and, under regularity conditions, the algorithm is guaranteed to converge to a stationary point (Dempster et al., 1977; Laird, 1978; Vaida et al., 2000; Cortiñas Abrahantes and Burzykowski, 2005).

**E-step:** The full log-likelihood (2.2)-(2.3) can be decomposed into two parts, the first (3.1) depending on $\boldsymbol{\pi}$ and the second (3.2) depending on $\lambda_0(t), \boldsymbol{\beta}, \mathbf{w}$.

$$l_{full,1}(\boldsymbol{\pi}; \mathbf{Y}|\mathbf{z}) = \sum_{k=1}^{K}\sum_{j=1}^{J} z_{jk} \cdot \log(\pi_k). \tag{3.1}$$

$$\begin{aligned} l_{full,2}(\lambda_0(t), \boldsymbol{\beta}, \mathbf{w}; \mathbf{Y}|\mathbf{z}) = &\sum_{k=1}^{K}\sum_{j=1}^{J} z_{jk} \cdot \sum_{i=1}^{n_j} \delta_{ij}[\log(\lambda_0(t_{ij})) + \log(w_k) + \mathbf{X}^T_{ij}\boldsymbol{\beta}] - \\ &\Lambda_0(t_{ij}) w_k \exp\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\}. \end{aligned} \tag{3.2}$$

The Expectation step consists of computing:

$$Q(\boldsymbol{\theta}) = E_{\mathbf{z}|\hat{\boldsymbol{\theta}}}[l_{full}(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{z})] = E_{\mathbf{z}|\hat{\boldsymbol{\theta}}}[l_{full,1}(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{z})] + E_{\mathbf{z}|\hat{\boldsymbol{\theta}}}[l_{full,2}(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{z})] \tag{3.3}$$

the expectation over $\mathbf{z}$, given the current values of parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \hat{\lambda}_0(t), \hat{\boldsymbol{\beta}}, \hat{\mathbf{w}})$, of the full log-likelihood for the observed data $\mathbf{Y}$.

This reduces to the computation of $E[z_{jk}|\mathbf{Y}, \hat{\boldsymbol{\theta}}]$, which we then include in (3.1) and (3.2). $E[z_{jk}|\mathbf{Y}, \hat{\boldsymbol{\theta}}]$ can be derived in closed form using Bayes' theorem (3.4).

$$E[z_{jk}|\mathbf{Y}, \hat{\boldsymbol{\theta}}] = \frac{\pi_k \cdot \exp\left\{\sum_{i=1}^{n_j} \delta_{ij} \cdot \log(w_k) - \Lambda_0(t_{ij}) w_k \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\}\right\}}{\sum_{r\in\{1:K\}} \pi_r \exp\left\{\sum_{i=1}^{n_j} \delta_{ij} \cdot \log(w_r) - \Lambda_0(t_{ij}) w_r \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\}\right\}}. \tag{3.4}$$

For simplicity, we write $\alpha_{jk} = E[z_{jk}|\mathbf{Y}, \hat{\boldsymbol{\theta}}]$. Furthermore, we note that this step is similar to the posterior probability computation in general mixture models.

**M-step:** The Maximization step consists of maximizing $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. $Q(\boldsymbol{\theta})$ can be partitioned so that we can maximize $Q_1(\boldsymbol{\pi}) := E_{\mathbf{z}|\hat{\boldsymbol{\theta}}}[l_{full,1}|\mathbf{Y}, \hat{\boldsymbol{\theta}}]$ with respect to $\boldsymbol{\pi}$ and $Q_2(\lambda_0, \boldsymbol{\beta}, \mathbf{w}) := E_{\mathbf{z}|\hat{\boldsymbol{\theta}}}[l_{full,2}|\mathbf{Y}, \hat{\boldsymbol{\theta}}]$ with respect to $\lambda_0, \boldsymbol{\beta}, \mathbf{w}$ separately. The maximization of $Q_1(\boldsymbol{\pi})$ is a constrained optimization problem, since $\sum_{k=1}^K \pi_k$ is equal to 1, and we can solve it by applying the Lagrange multipliers technique (3.5).

$$\hat{\pi}_k = \frac{1}{J} \sum_{j=1}^J \alpha_{jk}. \tag{3.5}$$

The optimization of $Q_2(\lambda_0, \boldsymbol{\beta}, \mathbf{w})$ is not trivial, since we adopt a nonparametric baseline hazard. We note that $Q_2(\lambda_0, \boldsymbol{\beta}, \mathbf{w})$ is a weighted version of the log-likelihood in a Cox regression model with known offset. Following Johansen (1983), we adapt a profile log-likelihood approach for the estimation of the shared parametric frailty Cox model. Initially, we estimate the $\mathbf{w}$ fixing $\lambda_0, \boldsymbol{\beta}$, giving:

$$\hat{w}_k = \frac{\sum_{j=1}^J \alpha_{jk} \sum_{i=1}^{n_j} \delta_{ij}}{\sum_{j=1}^J \alpha_{jk} \sum_{i=1}^{n_j} \left\{\Lambda_0(t_{ij}) \cdot \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\}\right\}}. \tag{3.6}$$

By substituting these estimates in $Q_2$, we obtain:

$$Q_2(\lambda_0, \boldsymbol{\beta}, \hat{\boldsymbol{w}}) = \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} \cdot \sum_{i=1}^{n_j} \delta_{ij}[\log(\lambda_0(t_{ij})) + \log(\hat{w}_k) + \mathbf{X}_{ij}^T \boldsymbol{\beta}] - \Lambda_0(t_{ij}) \hat{w}_k \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\}. \tag{3.7}$$

We can rewrite $Q_2$ in the following form, recalling that $\sum_{k=1}^K \alpha_{jk} = 1$,

$$\begin{aligned} Q_2(\lambda_0, \boldsymbol{\beta}, \hat{\boldsymbol{w}}) = \sum_{j=1}^J \sum_{i=1}^{n_j} \delta_{ij} \log(\lambda_0(t_{ij})) + \delta_{ij} \left(\sum_{k=1}^K \alpha_{jk} \log(\hat{w}_k)\right) + \delta_{ij} \{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} - \\ \Lambda_0(t_{ij}) \left(\sum_{k=1}^K \alpha_{jk} \hat{w}_k\right) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\}. \end{aligned} \tag{3.8}$$

This is similar to the form of the log-likelihood in a Cox regression model with known offset $\log\left(\sum_{k=1}^K \alpha_{jk} \hat{w}_k\right)$. With arguments similar to Johansen (1983), it is possible to show that the estimate of the cumulative baseline that maximizes (3.7) is the following:

$$\hat{\Lambda}_0(t_{ij}) = \sum_{(fg):t_{fg}\le t_{ij}} \frac{d_{fg}}{\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}\hat{w}_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)} \tag{3.9}$$

where $d_{fg}$ is the total number of events happening at time $t_{fg}$ and $R(t_{fg})$ represents the set of patients who are at risk at time $t_{fg}$, which is the event time of patient $f$ in cluster $g$.

Including (3.9) in $Q_2(\lambda_0, \boldsymbol{\beta}, \hat{\boldsymbol{w}})$, we obtain the profile log-likelihood as a function of only $\boldsymbol{\beta}$:

$$l_{profile}(\boldsymbol{\beta}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\delta_{ij}[\mathbf{X}_{ij}^T\boldsymbol{\beta} - \log\sum_{rs\in R(t_{ij})}\left(\sum_{k=1}^{K}\alpha_{sk}\hat{w}_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)] \tag{3.10}$$

Since (3.10) is of the form of the usual partial log-likelihood in the Cox model with known offsets, standard software can be used to obtain the maximal $\hat{\boldsymbol{\beta}}$.

The profile likelihood method also ensures identifiability between $\lambda_0(t)$ and the $w_k$, since at each step of the algorithm, one is estimated conditionally on the current value of the other. We observed better convergence from leaving the $w_k$ unconstrained, compared to applying a constraint such as $w_1 = 1$ before running the EM algorithm. However, for interpretability we divide the estimates of the set $w_k$'s by the lowest value of $w_k$.

In order to address the identifiability issue, we tried to include the frailty ratio directly in the model:

$$\lambda(t; \mathbf{X}_{ij}, w_k, z_{jk}) = \prod_{k=1}^{K}\left[\lambda_0(t)\frac{w_k}{w_{k-1}}\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right]^{z_{jk}} \tag{3.11}$$

where $w_0$ is assumed to be equal to $w_1$, so the ratio associated to $z_{j1}$ is equal to 1. This hazard leads to the following full log-likelihood, which is the analogous of (3.2):

$$\begin{aligned} l_{full,2}(\lambda_0(t), \boldsymbol{\beta}, \mathbf{w}; \mathbf{Y}|\mathbf{z}) = &\sum_{k=1}^{K}\sum_{j=1}^{J} z_{jk}\cdot\sum_{i=1}^{n_j}\delta_{ij}[\log(\lambda_0(t_{ij})) + \log\left(\frac{w_k}{w_{k-1}}\right) + \mathbf{X}_{ij}^T\boldsymbol{\beta}] - \\ &\Lambda_0(t_{ij})\frac{w_k}{w_{k-1}}\exp\{\mathbf{X}_{ij}^T\boldsymbol{\beta}\}. \end{aligned} \tag{3.12}$$

If we differentiate with respect to the ratio $w_k/w_{k-1}$, we obtain that the ratio's estimate is equal to the singular $w_k$'s estimate in (3.6) and this is due to the fact that it is not possible to identify the $w_1$.

$$\widehat{\left(\frac{w_k}{w_{k-1}}\right)} = \frac{\sum_{j=1}^{J}\alpha_{jk}\sum_{i=1}^{n_j}\delta_{ij}}{\sum_{j=1}^{J}\alpha_{jk}\sum_{i=1}^{n_j}\left\{\Lambda_0(t_{ij})\cdot\exp\{\mathbf{X}_{ij}^T\boldsymbol{\beta}\}\right\}}. \tag{3.13}$$

Moreover, we tried to fix an order constraint in order to avoid label switching. We defined a new variable $b_k = \log\left(\frac{w_k}{w_{k-1}} - 1\right)$, which means that $\frac{w_k}{w_{k-1}} = \exp\{b_k\} + 1$ and that $w_k > w_{k-1}$. We wrote (3.12) as follows:

$$\begin{aligned} l_{full,2}(\lambda_0(t), \boldsymbol{\beta}, \mathbf{w}; \mathbf{Y}|\mathbf{z}) = &\sum_{k=1}^{K}\sum_{j=1}^{J} z_{jk}\cdot\sum_{i=1}^{n_j}\delta_{ij}[\log(\lambda_0(t_{ij})) + \log(\exp\{b_k\}+1) + \mathbf{X}_{ij}^T\boldsymbol{\beta}] - \\ &\Lambda_0(t_{ij})(\exp\{b_k\}+1)\exp\{\mathbf{X}_{ij}^T\boldsymbol{\beta}\}. \end{aligned} \tag{3.14}$$

By differentiating with respect to $b_k$ we obtain:

$$\hat{b}_k = \log\left\{\widehat{\left(\frac{w_k}{w_{k-1}}\right)} - 1\right\} \tag{3.15}$$

The obtained result is coherent with the fact that the derivative of the log-likelihood with respect to the ratio is analytically solvable, and the same holds for its transformation $b_k$. So, it is not possible to introduce an order constraint for the proposed model.

## 3.2 Estimation of the Standard Errors

In the case of the Cox model with shared frailty terms, it is not possible to compute the variance-covariance matrix directly from the marginal log-likelihood, but it is possible to derive it from the observed information matrix, $\mathbf{I}(\boldsymbol{\theta})^{-1}$, (Klein, 1992), and this has been shown to be a consistent estimator (Parner et al., 1998). The observed information matrix can be written as:

$$\mathbf{I}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \tag{3.16}$$

where $l(\boldsymbol{\theta})$ is the observable log-likelihood:

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \delta_{ij} \log\Big(\lambda_0(t_{ij}) \exp\Big(\mathbf{X}_{ij}^T \boldsymbol{\beta}\Big)\Big) + \log\left(\sum_{k=1}^{K} \pi_k w_k^{D_j} \cdot \exp \sum_{i=1}^{n_j} \Big[-\Lambda_0(t_{ij}) w_k \exp\Big(\mathbf{X}_{ij}^T \boldsymbol{\beta}\Big)\Big]\right)$$

where $D_j$ is the total number of events in cluster $j$, $D_j = \sum_{i=1}^{n_j} \delta_{ij}$. Note that this is obtained by integrating the full likelihood over the random variable $\mathbf{z}$. For further information about the derivation of the observable log-likelihood and the elements of the observed information matrix, see Appendix A.

This asymptotic estimate of the covariance matrix can be computed once the parameters are estimated from the EM algorithm. A more computationally convenient approximation that exploits the EM framework was proposed by Louis (1982) together with a method to accelerate the algorithm, and a proof of quadratic convergence near the maximum likelihood estimate. Louis (1982) states that the $j^{th}$ component of the observed information matrix $\mathbf{I}$ can be written as:

$$\mathbf{I}^j = \mathbb{E}[B_j] - \mathbb{E}[S_j S_j^T] + S_j^\star S_j^{\star T} \tag{3.17}$$

where $S$ and $S^\star$ are the gradient vectors of the full log-likelihood and the observable log-likelihood respectively, while $B$ is the negative second derivative matrix of the full log-likelihood (see Appendix A for element-wise computation).

In this work we implement both methods in order to compare them and we provide also a third estimate for the observed information matrix, by using numerical methods (Gilbert and Varadhan, 2016) to obtain the first and second derivatives of the full log-likelihood.

In this work we estimate the frailties separately due to superior convergence; however, because we are interested in the ratios of frailties, we estimate the standard errors related to the ratios through the following formula:

$$Var(\hat{w}_k/\hat{w}_1) = \left(\frac{\mu_{\hat{w}_k}}{\mu_{\hat{w}_1}}\right)^2 \cdot \left[\frac{\sigma^2_{\hat{w}_1}}{\mu^2_{\hat{w}_1}} + \frac{\sigma^2_{\hat{w}_k}}{\mu^2_{\hat{w}_k}} - \frac{2Cov(\hat{w}_1, \hat{w}_k)}{\mu_{\hat{w}_1}\mu_{\hat{w}_k}}\right] \tag{3.18}$$

which can be derived by using the first and second order Taylor expansions.

### 3.3 Selection of the number of latent populations

Since it is impossible to estimate $K$ using a log-likelihood maximization argument (Figueiredo and Jain, 2002), we estimate $\boldsymbol{\theta}$ for each potential $K$, and compute a model selection criterion such as AIC, BIC, or search for the optimal $K$ using the the approach proposed by Laird (1978). For all the computations, we used the R software (R Development Core Team, 2016) developing an R code available upon request.

## 4 Simulation study

A simulation study was conducted to evaluate the performance of the estimators obtained with the algorithm described in Section 3. We simulated 100 datasets, each with $J = 100$ groups (e.g., healthcare providers) and $n_j = 50$ statistical units (e.g., patients) per group, giving a total of $5,000$ records in each dataset. For all simulations, we set the covariate-related log hazard ratio $\boldsymbol{\beta} = 0.4$, and define the baseline cumulative hazard so that $\Lambda_0^{-1}(t) = 0.01 \cdot t^{1.9}$ in order to mimic the dataset that motivated this work. The aim of the simulation was to estimate how well the algorithm estimates the true frailty ratios $\mathbf{w}/w_1$, mixing proportions $\boldsymbol{\pi}$ and number of latent populations $K$ for various values of these parameters of interest.

(i) Firstly we focus on $\boldsymbol{\pi}$, by setting $K = 2$ and $w_2/w_1 = 1.71$, and run 9 scenarios with $\pi_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The results are shown in Appendix B. In general, we observe estimates closer to the true values when the mixing proportion, $\pi_1$, is closer to 0.5, i.e., when there is a relatively large amount of data in all mixture components.

(ii) We focus on $w_2/w_1$, by setting $K = 2$, $\boldsymbol{\pi} = [0.3, 0.7]$ and run 7 scenarios, with $w_2/w_1 \in \{1.14, 1.29, 1.43, 1.57, 1.71, 2, 3\}$. We assume that frailty ratios smaller than 1.1 are not of practical interest. Conversely, we assume that for frailty ratios bigger than about 3, the presence of two latent populations can be identified easily by exploratory analysis, e.g. plotting a set of survival curves by group. The results are shown in Appendix C. The estimates of all parameters become more accurate as the frailty ratio increases, thus the contrast between latent populations becomes larger. In particular, the true number of latent populations $K = 2$ is detected for values of $w_2/w_1$ of around 1.6 and higher.

(iii) We focus on $K$, which leads to a complex pattern of simulations since varying $K$ changes the length of the vectors $\boldsymbol{\pi}$ and of $\mathbf{w}/w_1$. We tested $K \in \{2, 3, 4\}$, $\pi \in \{(0.4, 0.6), (0.3, 0.2, 0.5), (0.15, 0.25, 0.3, 0.3)\}$ and $\mathbf{w}/w_1 \in \{(1.5), (1.5, 2.5), (1.5, 2.5, 4)\}$ respectively. In our applications (5) we did not detect more than 4 populations with any method of model selection. The results are shown in Appendix D. The frailty ratios and mixing proportions are estimated accurately for all values of $K$. However, the three model selection methods produce different estimates of $K$, with BIC recovering the true values more often, and AIC and the method of Laird (1978) tending to estimate higher values.

AIC theoretically favours more complex models than BIC, however they produced the same estimate for the number of latent populations in the majority of the scenarios. As discussed by, e.g. Burnham and Anderson (2003), BIC would be preferred if we believe there is a low-dimensional "true" clustering structure which would not change with the amount of data, whereas AIC is preferred if we expect more latent populations (with more weakly-contrasting frailties) to be revealed as the dataset becomes bigger. We would expect the method of Laird (1978) to behave in a similar way to AIC, though we are unaware of any formal comparison.

Overall, the algorithm performs well, especially when there is a moderately large contrast between the frailty in different latent populations, and there is sufficient information in all latent populations. Thus, more clearly-defined latent population structures are revealed more easily.

# 5 An application to healthcare structures admission for patients with heart failure

The nonparametric frailty Cox model was applied to administrative data from patients with heart failure treated in Lombardia Region, Italy. Heart failure is the most common cause of hospitalization in Western countries for people more than 65 years old, with a 5 year risk of death similar or worse than that observed after a diagnosis of cancer (Frigerio et al., 2017). Moreover, heart failure is a chronic disease and has a substantial economic impact on health services (Corrao et al., 2014).

While some work has investigated the full history of healthcare structure admission and death for heart failure patients through multi-state models (Gasperoni et al., 2017), an event of particular interest is the second admission, as a marker of success of the initial treatment and possible future health care use.

Patient outcomes may be influenced by characteristics of the healthcare provider that may be difficult to measure or observe, such as infrastructure, extent or expertise of staff, efficiency or case mix. These unobserved covariates may cause over-dispersion. The time to readmission is thought to be particularly related to healthcare provider policies, and the risk of readmission is high. In contrast, times to first discharge or death are thought to be primarily related to illness severity. By assuming a shared frailty between patients admitted to the same healthcare provider, we investigate how the time to second admission is associated with the healthcare provider. Furthermore, we use the nonparametric frailty distribution to detect clusters of healthcare providers with similar outcomes, which may reflect unobserved structure-level predictors.

The original database is composed of $338,861$ admission records for a total of $210,917$ patients with a first diagnosis of Heart Failure between 2005 and 2012, in the Lombardia Region in Italy. The outcome is defined as the time between the first discharge and the second admission, excluding those patients who died during the first treatment and between the first discharge and the second admission. We select only those healthcare providers with more than 20 patients. The final dataset consists of $25,621$ patients admitted for the first time between 2006 and 2007, from 124 healthcare providers. The selected population have an average age of 73 years (SD 12) and 52% male. 41% have three or more comorbidities, which include renal disease, tumours and diabetes. 20% of the patients underwent one or more (up to 5) procedures, including coronary artery bypass graft surgery (CABG), percutaneous transluminal coronary angioplasty (PTCA), or insertion of an implantable cardioverter defibrillator (ICD).

We applied the nonparametric frailty Cox model, described in Section 2, with four individual-level predictors: age, gender, a binary indicator of the presence of three or more comorbidities, and the (continuous) total number of procedures. We fitted models with values of $K$ ranging from 1 to 5. The AIC and BIC were optimised by a model with $K = 2$ latent populations, while the criterion of Laird (1978), as expected, suggested a greater number $K = 4$. In cases where the true $K \leq 4$ and between-population frailty ratios are $> 1.1$ our simulation suggested that AIC and BIC estimate $K$ more accurately.

We illustrate the variability among healthcare providers in times to second admission as 124 structure-specific Kaplan-Meier curves in Figure 1. The curves are coloured according to the classification of healthcare providers from the model with $K = 2$. The blue curves represent the latent population of healthcare providers with $\hat{w}_2 = 1.39$ times the hazard of readmission relative

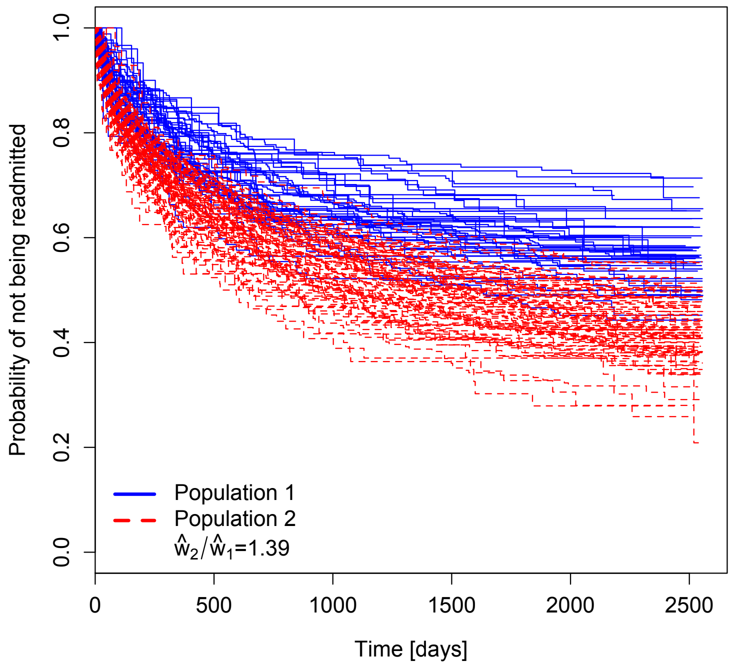


Figure 1: 124 structure-specific Kaplan-Meier curves colored by membership of two latent populations from the frailty model for time to readmission.

to $\hat{w}_1$ , and the model estimates a probability of $\hat{\pi}_2 = 0.67$ that a healthcare provider belongs to this population. All estimates are reported in Table 1.

Older people have a higher risk of being readmitted (hazard ratio, HR, $e^{0.04} = 1.04$ per year of age), as do men (HR $e^{0.27} = 1.31$), people having three or more comorbidities (HR $e^{0.35} = 1.42$) and people having fewer medical or surgical procedures (HR $e^{-0.14} = 0.87$ per procedure). The relationship between fewer procedures and risk of readmission may seem counter-intuitive but it reflects the fact that people undergoing procedures are younger on average (with mean age 68.5 (SD 11.7), compared to 74.3 (11.7) for people who do not) and there may be some collinearity between age and number of procedures. Moreover, the procedure should have successfully treated the underlying heart disease, thereby reducing the need for readmission. The estimates and standard errors for the same covariate effects from a standard Cox model without frailties are almost identical (Table 1).

We then sought to describe the latent population structure, indicated by the model with $K = 2$, in terms of characteristics of the healthcare providers that are recorded in the database. Healthcare providers belonging to the population with higher risk of readmission, on average, had a higher number of patients and a higher percentage of in-structure deaths per year, although the percentages of surgical and complex cases were similar between the two latent populations, Table 2. Comparing the type of institution, we found that medical institutions belonged to the higher-frailty population more often, while nursing homes and public IRCCS (research center institutes) tended to belong to the lower-frailty population, Figure 2.

Table 1: Estimates of Cox model with a nonparametric frailty term and a classical Cox model

| Parameters | Cox with nonparametric frailty | | | | Cox model | |
|---|---|---|---|---|---|---|
| | Estimates | Standard errors | | | Estimates | Standard errors |
| | | Louis | Exact | Numerical | | (exact) |
| $\pi_1$ | 0.33 | 0.0596 | 0.0600 | 0.0596 | - | - |
| $\pi_2$ | 0.67 | - | - | - | - | - |
| $w_2/w_1$ | 1.39 | 0.0420 | 0.0420 | 0.0420 | - | - |
| Log hazard ratios for covariates | | | | | | |
| 1 year of age | 0.04 | 0.0009 | 0.0009 | 0.0009 | 0.04 | 0.0009 |
| Male | 0.27 | 0.0181 | 0.0181 | 0.0181 | 0.28 | 0.0180 |
| 3 or more comorbidities | 0.35 | 0.0175 | 0.0175 | 0.0175 | 0.35 | 0.0174 |
| Number of procedures | -0.14 | 0.0125 | 0.0125 | 0.0125 | -0.13 | 0.0124 |

Table 2: Healthcare providers' profiles

| | Latent population 1 | Latent population 2 |
|---|---|---|
| Average number of patients (s.d.) | 7,071.8 (5,121.1) | 11,965.2 (9894.0) |
| Average % of in-structure death (s.d.) | 2.8 (2.1) | 3.5 (1.5) |
| Average % of surgical cases (s.d.) | 35.9 (21.2) | 30.7 (12.2) |
| Average % of complex cases (s.d.) | 13.8 (5.7) | 14.3 (3.4) |

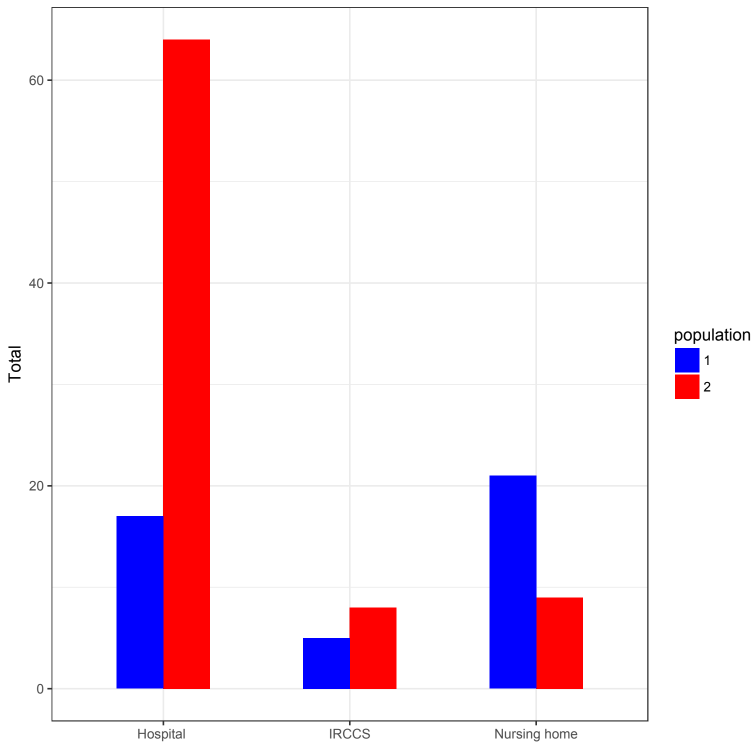


Figure 2: Healthcare providers structures in the two latent populations. Red bars are related to structures that belong to the second latent population, while blue bars to the first one.

We then extended the frailty model to include two structure-specific covariates, describing the type of healthcare provider with three categories, as in Figure 2, and the percentage of admissions in which the patient died. The optimal model according to AIC and BIC still has $K = 2$, with relative frailty $w_2/w_1 = 1.42$ between the two populations, and an estimated probability of $\pi_2 = 0.76$ that the healthcare provider belongs to the group with the higher frailty. Thus these two covariates can characterize the two latent populations only partially, and the remaining clustering pattern probably depends on unobserved characteristics of the healthcare providers. The nonparametric frailty model therefore serves as a starting point for further investigation of the effect of healthcare providers and their characteristics on patient outcomes.

# 6 Discussion

In this paper, we propose a new model that deals with hierarchical time-to-event data and tackles two issues: extending the classical Cox proportional hazard model and detecting a clustering structure among groups by including a shared nonparametric frailty term. Classical approaches for hierarchical time-to-event data use proportional hazard models with a parametric shared frailty, however, the most appropriate parametric frailty distribution will not always be clear (Austin, 2017) and the data may not fit any standard parametric family. Having a discrete frailty distribution, together with an unspecified baseline hazard, leads to a novel and very flexible model for grouped survival data.

Moreover, we are able to detect clustering at the second level of a hierarchy of time-to-event data. In published literature, healthcare providers (or specifically hospitals) clustering is usually investigated by applying a logistic regression where the covariates are patient or structure specific (Ohlssen et al., 2007; Grieco et al., 2011). These models have two limitations: first, the covariates are chosen a priori; second, the time-to-event data are reduced to a single binary variable representing incidence (or not) of an event of interest. Through our model, we can identify the existence and nature of a clustering structure, without defining a priori a set of covariates that describe the investigators' opinions about the performance of the healthcare providers. A further strength of the proposed model is that it may be used to detect clusters of individuals, as well as clusters of groups, since the frailties may be group-related or individual-related. Here we defined a model with a shared frailty term for groups of healthcare providers, but individual frailty models are simply specific cases of shared frailty models where each group is composed of a single healthcare provider. This application would be of interest when patient clusters are suspected but the available covariates are not sufficient to describe the full variability. Additionally, also detecting more complex hierarchical structures (e.g., patients grouped in structures grouped in regions) may be of interest. In this case, we could consider an extension to nested frailty models, in a frequentist framework, or we could consider Bayesian methods, that would express the uncertainty in the clustering structure more easily. These models could also be fitted in standard Bayesian software using Markov Chain Monte Carlo algorithms, although they may take a longer time to converge, especially for big databases.

Usually, the software used to estimate the parameters of proportional hazard models with shared frailties relies on some version of the EM algorithm. In this work, we proposed a EM algorithm that was designed for our model. Other techniques, besides the EM algorithm, have been explored in the literature for specific models: for example, the penalized partial likelihood approach (Therneau et al., 2003) has been applied for Gamma-distributed frailties (with the same results as the EM) or Log-Normal-distributed frailties (with similar results to the EM), while Gauss-Hermite quadrature was applied by Crowther et al. (2014) for a parametric proportional hazards model. Li et al. (1998) proposed Monte Carlo EM (MCEM), in which the expectation step is computed through a Monte Carlo simulation. Extension of our work to investigate alternative implementation methods that could speed up the procedure would be worthwhile. More efficient algorithms would be particularly important for analysis of very large databases, such as the administrative clinical database that motivated this work. Such administrative databases are emerging as powerful tools for addressing questions in epidemiology and other medical research; the need of rigorously defined models and reliable methods for preliminary analysis is clear. The proposed model, which makes few assumptions about the baseline hazard or frailty distribution, represents a step in this direction. Further extension of this model to a realistic but more complex framework, such as multiple events, would also be a natural next step.

## References


Peter C. Austin. A tutorial on multilevel survival analysis: Methods, models and applications. *International Statistical Review*, 2017. ISSN 1751-5823. doi: 10.1111/insr.12214. URL http://dx.doi.org/10.1111/insr.12214. doi:10.1111/insr.12214.

Theodor Adrian Balan and Hein Putter. *frailtyEM: Fitting Frailty Models with the EM Algorithm*, 2017. URL https://CRAN.R-project.org/package=frailtyEM. R package version 0.5.4.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2003.

Chrys Caroni, Martin Crowder, and Alan Kimber. Proportional hazards models with discrete frailty. *Lifetime Data Analysis*, 16(3):374–384, 2010. ISSN 1572-9249. doi: 10.1007/s10985-010-9151-3. URL http://dx.doi.org/10.1007/s10985-010-9151-3.

Giovanni Corrao, Arianna Ghirardi, Buthaina Ibrahim, Luca Merlino, and Aldo Pietro Maggioni. Burden of new hospitalization for heart failure: a population-based investigation from italy. *European Journal of Heart Failure*, 16(7):729–736, 2014. ISSN 1879-0844. doi: 10.1002/ejhf.105. URL http://dx.doi.org/10.1002/ejhf.105.

Jos Cortiñas Abrahantes and Tomasz Burzykowski. A version of the em algorithm for proportional hazard model with random effects. *Biometrical Journal*, 47(6):847–862, 2005. ISSN 1521-4036. doi: 10.1002/bimj.200410141. URL http://dx.doi.org/10.1002/bimj.200410141.

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

Michael J Crowther, Maxime P Look, and Richard D Riley. Multilevel mixed effects parametric survival models using adaptive gauss–hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics In Medicine*, 33(22):3844–3858, 2014.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, Series B*, 39:1–38, 1977.

Dirley M dos Santos, Richard B Davies, and Brian Francis. Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer. *Journal of statistical planning and inference*, 47(1-2):111–127, 1995.

Luc Duchateau and Paul Janssen. *The frailty model*. Springer Science & Business Media, 2007.

Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65:457–487, 1978.

Chris Elbers and Geert Ridder. True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, 49(3):403–409, 1982.

V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046, 1982. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529885.

Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

Maria Frigerio, Cristina Mazzali, Anna Maria Paganoni, Francesca Ieva, Pietro Barbieri, Mauro Maistrello, Ornella Agostoni, Cristina Masella, Simonetta Scalvini, et al. Trends in heart failure hospitalizations, patient characteristics, in-hospital and 1-year mortality: A population study, from 2000 to 2012 in lombardy. *International Journal of Cardiology*, 236:310–314, 2017.

Francesca Gasperoni, Francesca Ieva, Giulia Barbati, Arjuna Scagnetto, Annamaria Iorio, Gianfranco Sinagra, and Andrea Di Lenarda. Multi-state modelling of heart failure care path: A population-based investigation from italy. *PLOS ONE*, 12(6):1–15, 06 2017. doi: 10.1371/journal.pone.0179176. URL https://doi.org/10.1371/journal.pone.0179176.

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

Niccolò Grieco, Francesca Ieva, and Anna Maria Paganoni. Performance assessment using mixed effects models: a case study on coronary patient care. *IMA Journal of Management Mathematics*, 23(2):117–131, 2011.

Alessandra Guglielmi, Francesca Ieva, Anna M Paganoni, Fabrizio Ruggeri, and Jacopo Soriano. Semiparametric bayesian models for clustering and classification in the presence of unbalanced in-hospital survival. *Journal of the Royal Statistical Society, Series C*, 63(1):25–46, 2014.

Guang Guo and German Rodriguez. Estimating a multivariate proportional hazards model for clustered data using the em algorithm, with an application to child survival in guatemala. *Journal of the American Statistical Association*, 87(420):969–976, 1992.

James J. Heckman and Burton Singer. Population heterogeneity in demographic models. In Kenneth C. Land and Andrei Rogers, editors, *Multidimensional Mathematical Demography*, pages 567–599. Academic Press, New York, NY, 1982.

James J. Heckman and Burton Singer. The identifiability of the proportional hazard model. *The Review of Economic Studies*, 51(2):231–241, 1984a.

James J. Heckman and Burton Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, 52(2):271–320, 1984b.

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

Philip Hougaard. A class of multivariate failure time distributions. *Biometrika*, 73:671–678, 1986a.

Philip Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986b.

Philip Hougaard. *Analysis of multivariate survival data.* New York, NY: Springer Science & Business Media, 2012.

Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian survival analysis.* Hoboken, NJ: John Wiley & Sons, 2005.

Francesca Ieva and Anna Maria Paganoni. Detecting and visualizing outliers in provider profiling via funnel plots and mixed effect models. *Health care management science*, 18(2):166–172, 2015.

Søren Johansen. An extension of Cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, 51(2):165–174, 1983.

John P Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3):795–806, 1992.

Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

Hongzhe Li, Elizabeth A Thompson, and Ellen M Wijsman. Semiparametric estimation of major gene effects for age of onset. *Genetic epidemiology*, 15(3):279–298, 1998.

Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

Samuel OM Manda. A nonparametric frailty model for clustered survival data. *Communications in StatisticsTheory and Methods*, 40(5):863–875, 2011.

Geoffrey McLachlan and David Peel. *Finite mixture models*. Hoboken, NJ: John Wiley & Sons, 2004.

David I Ohlssen, Linda D Sharples, and David J Spiegelhalter. Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons. *Statistics in medicine*, 26(9):2088–2112, 2007.

Erik Parner et al. Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, 26(1):183–214, 1998.

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

Sophia Rabe-Hesketh and Anders Skrondal. *Multilevel and longitudinal modeling using Stata*. College Station, TX: STATA press, 2008.

David J Spiegelhalter. Funnel plots for comparing institutional performance. *Statistics in medicine*, 24(8):1185–1202, 2005.

Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

Terry Therneau. *A package for survival analysis in S*, 2014. URL http://CRAN. R-project. org/package= survival. R package version 2.37-4.

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

Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. New York, NY: Springer Science & Business Media, 2013.

Terry M Therneau, Patricia M Grambsch, and V Shane Pankratz. Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1):156–175, 2003.

Florin Vaida, Ronghui Xu, et al. Proportional hazards model with random effects. *Statistics in medicine*, 19(24):3309–3324, 2000.

Andreas Wienke. *Frailty models in survival analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

# Supplementary Materials

# Appendix A

In this section we compute the observed information matrix and we compute it in two ways: evaluating the derivatives of the observable loglikelihood and with the Louis method.

### Hessian of the observable loglikelihood

First of all, we write the observable likelihood, which is obtained by integrating out the random variable $\mathbf{z}$:

$$
\begin{aligned}
l(\boldsymbol{\theta}; data) = &\log\left(\prod_{j=1}^{J}\sum_{k=1}^{K}\pi_k\prod_{i=1}^{n_j}\left\{[\lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)]^{\delta_{ij}}\cdot\exp\left[-\Lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right]\right\}\right) = \\
&\sum_{j=1}^{J} log\left(\sum_{k=1}^{K}\pi_k\prod_{i=1}^{n_j}\left\{[\lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)]^{\delta_{ij}}\cdot\exp\left[-\Lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right]\right\}\right) = \\
&\sum_{j=1}^{J} log\left[\prod_{i=1}^{n_j}[\lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)]^{\delta_{ij}}\cdot\left(\sum_{k=1}^{K}\pi_k w_k^{D_j}\cdot\prod_{i=1}^{n_j}\exp\left[-\Lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right]\right)\right] = \\
&\sum_{j=1}^{J}\sum_{i=1}^{n_j}\delta_{ij}\log\left(\lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right) + \log\left(\sum_{k=1}^{K}\pi_k w_k^{D_j}\cdot\exp\sum_{i=1}^{n_j}\left[-\Lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right]\right) = \\
&\sum_{j=1}^{J} l_1^j + l_2^j
\end{aligned}
\tag{6.1}
$$

where $D_j$ is the total events happened in group $j$, $D_j = \sum_{i=1}^{n_j}\delta_{ij}$.

$$
l_1^j = \sum_{i=1}^{n_j}\delta_{ij}\log\left(\lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right) \tag{6.2}
$$

$$
l_2^j = \log\left(\sum_{k=1}^{K}\pi_k w_k^{D_j}\cdot\exp\sum_{i=1}^{n_j}\left[-\Lambda_0(t_{ij})w_k\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right]\right) \tag{6.3}
$$

To compute the second derivatives with respect to all parameters, recall the definition of the baseline and cumulative baseline hazard estimates and the related derivatives:

$$\lambda_0(t_{fg}) = \frac{d_{fg}}{\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)} \tag{6.4a}$$

$$\lambda_{0\alpha}(t_{fg}) = \frac{\partial\lambda_0(t_{fg})}{\partial\boldsymbol{\beta}_\alpha} = \frac{-d_{fg}\cdot\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\mathbf{X}_{rs\alpha}}{\left(\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\right)^2} \tag{6.4b}$$

$$\lambda_{0\alpha\gamma}(t_{fg}) = \frac{\partial^2\lambda_0(t_{fg})}{\partial\boldsymbol{\beta}_\alpha\partial\boldsymbol{\beta}_\gamma} = \frac{-d_{fg}\cdot\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\mathbf{X}_{rs\alpha}\mathbf{X}_{rs\gamma}}{\left(\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\right)^2} \tag{6.4c}$$

$$+\frac{2\cdot d_{fg}\left(\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\mathbf{X}_{rs\alpha}\right)\left(\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\mathbf{X}_{rs\gamma}\right)}{\left(\sum_{rs\in R(t_{fg})}\left(\sum_{k=1}^{K}\alpha_{sk}w_k\right)\exp\left(\mathbf{X}_{rs}^T\boldsymbol{\beta}\right)\right)^3}$$

$$\Lambda_{0\alpha}(t_{ij}) = \sum_{fg:t_{fg}\le t_{ij}}\frac{\partial\lambda_0(t_{fg})}{\partial\boldsymbol{\beta}_\alpha} \tag{6.4d}$$

$$\Lambda_{0\alpha\gamma}(t_{ij}) = \sum_{fg:t_{fg}\le t_{ij}}\frac{\partial^2\lambda_0(t_{fg})}{\partial\boldsymbol{\beta}_\alpha\partial\boldsymbol{\beta}_\gamma} \tag{6.4e}$$

where $d_{fg}$ is the total number of events recorded at time $t_{fg}$.

$$\frac{\partial^2 l_1^j}{\partial\boldsymbol{\beta}_\alpha\partial\boldsymbol{\beta}_\gamma} = \sum_{i=1}^{n_j}\delta_{ij}\left\{\frac{\lambda_{0\alpha\gamma}(t_{ij})\lambda_0(t_{ij}) - \lambda_{0\alpha}(t_{ij})\lambda_{0\gamma}(t_{ij})}{\lambda_0(t_{ij})^2}\right\},\quad \alpha,\gamma = 1:p \tag{6.5}$$

where $p$ is the total number of covariates.

$$\frac{\partial l_2^j}{\partial\pi_g} = \frac{w_g^{D_j}\exp\{-w_g\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\} - w_K^{D_j}\exp\{-w_K\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\}}{\sum_{k=1}^{K}\pi_k w_k^{D_j}\cdot\exp\{-w_k\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\}},$$
$$g = 1:(K-1) \tag{6.6}$$

$$\frac{\partial l_2^j}{\partial w_q} = \frac{\pi_q w_q^{D_j-1}\exp\{-w_q\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\}\left(D_j - w_q\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\right)}{\sum_{k=1}^{K}\pi_k w_k^{D_j}\cdot\exp\{-w_k\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\}},$$
$$q = 1:K \tag{6.7}$$

$$\frac{\partial l_2^j}{\partial\boldsymbol{\beta}_\alpha} = \frac{-\sum_{k=1}^{K}\pi_k w_k^{D_j+1}\exp\{-w_k\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\}\sum_{i=1}^{n_j}(\Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij})X_{ij\alpha})\exp\{\mathbf{X}_{ij}^T\boldsymbol{\beta}\}}{\sum_{k=1}^{K}\pi_k w_k^{D_j}\cdot\exp\{-w_k\sum_{i=1}^{n_j}\Lambda_0(t_{ij})\exp\left(\mathbf{X}_{ij}^T\boldsymbol{\beta}\right)\}},$$
$$\alpha = 1:p \tag{6.8}$$

$$\frac{\partial^2 l_2^j}{\partial \pi_g \partial \pi_l} = \frac{-\left(w_g^{D_j} \exp\{-w_g \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} - w_K^{D_j} \exp\{-w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}\right)}{\left(\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}\right)^2} \cdot$$
$$\cdot \left(w_l^{D_j} \exp\{-w_l \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} - w_K^{D_j} \exp\{-w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}\right),$$
$$g, l = 1 : (K-1) \tag{6.9}$$

$$\frac{\partial^2 l_2^j}{\partial w_q \partial w_r} = \mathbf{1}_{\{r=q\}} \left\{ \frac{\pi_q w_q^{D_j - 2} \exp\{-w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}}{\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}} \cdot \right.$$
$$\left[\left(D_j - 1 - w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right)\left(D_j - w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right) - \right.$$
$$\left.\left. w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right]\right\}$$
$$- \frac{\pi_g \pi_l (w_q w_r)^{D_j - 1} \exp\{-(w_q + w_r) \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}}{\left(\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}\right)^2} \cdot$$
$$\cdot \left(D_j - w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right)\left(D_j - w_r \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right), \quad q, r = 1 : K \tag{6.10}$$

$$\frac{\partial^2 l_2^j}{\partial \boldsymbol{\beta}_\alpha \partial \boldsymbol{\beta}_\gamma} = \frac{-\sum_{k=1}^K \pi_k w_k^{D_j + 1} \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}}{\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}} \cdot$$
$$\cdot \left\{ -w_k \sum_{i=1}^{n_j} (\Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} \cdot \sum_{i=1}^{n_j} (\Lambda_{0\gamma}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\gamma}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} + \right.$$
$$\left. + \sum_{i=1}^{n_j} (\Lambda_{0\alpha\gamma}(t_{ij}) + \Lambda_{0\gamma}(t_{ij}) X_{ij\alpha} + \Lambda_{0\alpha}(t_{ij}) X_{ij\gamma} + \Lambda_0(t_{ij}) X_{ij\gamma} X_{ij\alpha}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} \right\} +$$
$$- \frac{(\sum_{k=1}^K \pi_k w_k^{D_j + 1} \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\})^2}{\left(\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}\right)^2} \cdot$$
$$\cdot \sum_{i=1}^{n_j} (\Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} \cdot \sum_{i=1}^{n_j} (\Lambda_{0\gamma}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\gamma}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\},$$
$$\alpha, \gamma = 1 : p \tag{6.11}$$

$$\frac{\partial^2 l_2^j}{\partial \pi_g \partial w_q} = \mathbf{1}_{\{q=g\}} \left\{ \frac{w_g^{D_j-1} \exp\{-w_g \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} \left(D_j - w_g \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right)}{\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}} \right\} +$$
$$+ \mathbf{1}_{\{q=K\}} \left\{ - \frac{w_K^{D_j-1} \exp\{-w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} \left(D_j - w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right)}{\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}} \right\} +$$
$$- \frac{\left(w_g^{D_j} \exp\left\{-w_g \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right\} - w_K^{D_j} \exp\left\{-w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right\}\right)}{\left(\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp \sum_{i=1}^{n_j} \left[-\Lambda_0(t_{ij}) w_k \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right]\right)^2} \cdot$$
$$\cdot \left(\pi_q w_q^{D_j-1} \exp\{-w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} \left(D_j - w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\right)\right),$$
$$g = 1 : (K-1), q = 1 : K \tag{6.12}$$

$$\frac{\partial^2 l_2^j}{\partial \pi_g \partial \boldsymbol{\beta}_\alpha} = \sum_{i=1}^{n_j} (\Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} \cdot$$
$$\left\{ \frac{w_K^{D_j+1} \exp\{-w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} - w_g^{D_j+1} \exp\{-w_g \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}}{\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}} + \right.$$
$$- \frac{w_K^{D_j} \exp\{-w_K \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} - w_g^{D_j} \exp\{-w_g \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}}{(\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\})^2}$$
$$\left. \cdot \sum_{k=1}^K \pi_k w_k^{D_j+1} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} \right\}, \quad g = 1 : (K-1), \alpha = 1 : p \tag{6.13}$$

$$\frac{\partial^2 l_2^j}{\partial w_q \partial \boldsymbol{\beta}_\alpha} = \pi_q w_q^{D_j-1} \exp\{-w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\} \sum_{i=1}^{n_j} (\Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha}) \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} \cdot$$
$$\left\{ \frac{-w_q(D_j + 1 - w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right))}{\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}} + \right.$$
$$\left. + \frac{(D_j - w_q \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)) \sum_{k=1}^K \pi_k w_k^{D_j+1} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\}}{(\sum_{k=1}^K \pi_k w_k^{D_j} \cdot \exp\{-w_k \sum_{i=1}^{n_j} \Lambda_0(t_{ij}) \exp\left(\mathbf{X}_{ij}^T \boldsymbol{\beta}\right)\})^2} \right\},$$
$$q = 1 : K, \alpha = 1 : p \tag{6.14}$$

## Louis method

We start from the $j^{th}$ component of the full loglikelihood, written in Eq. (6.15).

$$l^j_{full}(\boldsymbol{\theta}; data|\mathbf{z}) = \sum_{k=1}^{K} z_{jk} \cdot \left[ \log(\pi_k) + \left\{ \sum_{i=1}^{n_j} \delta_{ij}[\log(\lambda_0(t_{ij})) + \log(w_k) + \mathbf{X}_{ij}^T \boldsymbol{\beta}] - \Lambda_0(t_{ij}) w_k \exp\{\mathbf{X}_{ij}^T \boldsymbol{\beta}\} \right\} \right]. \tag{6.15}$$

Louis states that the $j^{th}$ component of the observed information matrix $\mathbf{I}$ can be written as follows:

$$\mathbf{I}^j = \mathbf{E}[B_j(T_{ij}, \delta_{ij}, z_{jk})] - \mathbf{E}[S_j(T_{ij}, \delta_{ij}, z_{jk}) S_j(T_{ij}, \delta_{ij}, z_{jk})^T] + S_j(T_{ij}, \delta_{ij})^\star S_j(T_{ij}, \delta_{ij})^{\star T} \tag{6.16}$$

The final observed information matrix is the sum of all $j^{th}$ elements: $\mathbf{I} = \sum_{j=1}^{J} \mathbf{I}^j$.

We define $S$ and $S^\star$ as the gradient vectors of the full loglikelihood and the observable loglikelihood, see (6.17) and (6.18), while $B$ and $B^\star$ are the negative second derivative matrices of full loglikelihood and the observable loglikelihood, see (6.20). We remind that the unique observable variables are $(T_{ij}, \delta_{ij})$, while the complete set of variables is $(T_{ij}, \delta_{ij}, z_{jk})$.

$$S_j^T(T_{ij}, \delta_{ij}, z_{jk}) = \nabla l^j_{full} = \left[ \frac{\partial l^j_{full}}{\partial \pi_1}, \dots, \frac{\partial l^j_{full}}{\partial \pi_{K-1}}, \frac{\partial l^j_{full}}{\partial w_1}, \dots, \frac{\partial l^j_{full}}{\partial w_K}, \frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_1}, \dots, \frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_p} \right]^T \tag{6.17}$$

$$S_j^\star(T_{ij}, \delta_{ij}) = \mathbf{E}[S_j(T_{ij}, \delta_{ij}, z_{jk})] \tag{6.18}$$

$$\mathbf{E}[S_j(T_{ij}, \delta_{ij}, z_{jk}) S_j(T_{ij}, \delta_{ij}, z_{jk})^T] = \left[ \begin{array}{c|c|c} \mathbf{E}[\nabla_\pi l^j_{full} \nabla^T_\pi l^j_{full}] & \mathbf{E}[\nabla_\pi l^j_{full} \nabla^T_w l^j_{full}] & \mathbf{E}[\nabla_\pi l^j_{full} \nabla^T_{\boldsymbol{\beta}} l^j_{full}] \\ \hline \mathbf{E}[\nabla_w l^j_{full} \nabla^T_\pi l^j_{full}] & \mathbf{E}[\nabla_w l^j_{full} \nabla^T_w l^j_{full}] & \mathbf{E}[\nabla_w l^j_{full} \nabla^T_{\boldsymbol{\beta}} l^j_{full}] \\ \hline \mathbf{E}[\nabla_{\boldsymbol{\beta}} l^j_{full} \nabla^T_\pi l^j_{full}] & \mathbf{E}[\nabla_{\boldsymbol{\beta}} l^j_{full} \nabla^T_w l^j_{full}] & \mathbf{E}[\nabla_{\boldsymbol{\beta}} l^j_{full} \nabla^T_{\boldsymbol{\beta}} l^j_{full}] \end{array} \right] \tag{6.19}$$

$$B_j(T_{ij}, \delta_{ij}, z_{jk}) =$$
$$\begin{pmatrix}
-\frac{\partial^2 l^j_{full}}{\partial \pi_1^2} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial \pi_1 \partial \pi_{K-1}} & & & & & & \\
\vdots & \ddots & \vdots & & & & & & \\
-\frac{\partial^2 l^j_{full}}{\partial \pi_{K-1} \partial \pi_1} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial \pi_{K-1}^2} & & & & & & \\
& & & -\frac{\partial^2 l^j_{full}}{\partial w_1^2} & & & -\frac{\partial^2 l^j_{full}}{\partial w_1 \partial \boldsymbol{\beta}_1} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial w_1 \partial \boldsymbol{\beta}_p} \\
& & & & \ddots & & \vdots & \vdots & \vdots \\
& & & & & -\frac{\partial^2 l^j_{full}}{\partial w_K^2} & -\frac{\partial^2 l^j_{full}}{\partial w_K \partial \boldsymbol{\beta}_1} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial w_K \partial \boldsymbol{\beta}_p} \\
& & & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_1 \partial w_1} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_1 \partial w_k} & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_1^2} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_p} \\
& & & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
& & & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_p \partial w_1} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_p \partial w_K} & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_p \partial \boldsymbol{\beta}_1} & \cdots & -\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_p^2}
\end{pmatrix} \tag{6.20}$$

The components of $S$ are computed in (6.21), (6.22) and (6.23).

$$\frac{\partial l^j_{full}}{\partial \pi_g} = \frac{z_{jg}}{\pi_g} - \frac{z_{jK}}{\pi_K}, \qquad g = 1:(K-1) \tag{6.21}$$

$$\frac{\partial l^j_{full}}{\partial w_q} = z_{jq} \sum_{i=1}^{n_j} \frac{\delta_{ij}}{w_q} - \Lambda_0(t_{ij}) \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\}, \quad q = 1:K \tag{6.22}$$

$$\frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_\alpha} = \sum_{k=1}^{K} z_{jk} \sum_{i=1}^{n_j} \delta_{ij} \left\{ \frac{\lambda_{0\alpha}(t_{ij})}{\lambda_0(t_{ij})} + X^T_{ij\alpha} \right\} - w_k \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\} \left\{ \Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha} \right\}, \quad \alpha = 1:p \tag{6.23}$$

The components of $\mathbf{E}[S_j(T_{ij}, \delta_{ij}, z_{jk}) S_j(T_{ij}, \delta_{ij}, z_{jk})^T]$ are:

$$\mathbf{E}\left[\frac{\partial l^j_{full}}{\partial \pi_g} \frac{\partial l^j_{full}}{\partial \pi_l}\right] = \frac{\alpha_{jg}}{\pi_g^2} \mathbf{1}_{\{g=l\}} + \frac{\alpha_{jK}}{\pi_K^2}, \qquad g, l = 1:(K-1) \tag{6.24}$$

$$\mathbf{E}\left[\frac{\partial l^j_{full}}{\partial w_q} \frac{\partial l^j_{full}}{\partial w_r}\right] = \mathbf{1}_{\{q=r\}} \cdot \alpha_{jq} \left( \sum_{i=1}^{n_j} \frac{\delta_{ij}}{w_q} - \Lambda_0(t_{ij}) \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\} \right)^2, \quad q, r = 1:K \tag{6.25}$$

$$\begin{aligned}
\mathbf{E}\left[\frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_\alpha} \frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_\gamma}\right] =& \sum_{k=1}^{K} \alpha_{jk} \left( \sum_{i=1}^{n_j} \delta_{ij} \left\{ \frac{\lambda_{0\alpha}(t_{ij})}{\lambda_0(t_{ij})} + X^T_{ij\alpha} \right\} - w_k \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\} \left\{ \Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha} \right\} \right) \cdot \\
& \left( \sum_{i=1}^{n_j} \delta_{ij} \left\{ \frac{\lambda_{0\gamma}(t_{ij})}{\lambda_0(t_{ij})} + X^T_{ij\gamma} \right\} - w_k \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\} \left\{ \Lambda_{0\gamma}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\gamma} \right\} \right), \\
& \alpha, \gamma = 1:p
\end{aligned} \tag{6.26}$$

$$\mathbf{E}\left[\frac{\partial l^j_{full}}{\partial \pi_g} \frac{\partial l^j_{full}}{\partial w_q}\right] = \begin{cases} \frac{\alpha_{jg}}{\pi_g} \sum_{i=1}^{n_j} \frac{\delta_{ij}}{w_g} - \Lambda_0(t_{ij}) \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\}, \text{if} \quad q = g \\ -\frac{\alpha_{jK}}{\pi_K} \sum_{i=1}^{n_j} \frac{\delta_{ij}}{w_K} - \Lambda_0(t_{ij}) \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\}, \text{if} \quad q = K \end{cases} \tag{6.27}$$

$$\begin{aligned}
\mathbf{E}\left[\frac{\partial l^j_{full}}{\partial \pi_g} \frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_\alpha}\right] =& \frac{\alpha_{jg}}{\pi_g} \sum_{i=1}^{n_j} \delta_{ij} \left\{ \frac{\lambda_{0\alpha}(t_{ij})}{\lambda_0(t_{ij})} + X^T_{ij\alpha} \right\} - w_g \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\} \left\{ \Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha} \right\} \\
& - \frac{\alpha_{jK}}{\pi_K} \sum_{i=1}^{n_j} \delta_{ij} \left\{ \frac{\lambda_{0\alpha}(t_{ij})}{\lambda_0(t_{ij})} + X^T_{ij\alpha} \right\} - w_K \exp\Big\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\Big\} \left\{ \Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij}) X_{ij\alpha} \right\}, \\
& g = 1:(K-1), \alpha = 1:p
\end{aligned} \tag{6.28}$$

$$\mathbf{E}\left[\frac{\partial l^j_{full}}{\partial w_q}\frac{\partial l^j_{full}}{\partial \boldsymbol{\beta}_\alpha}\right] = \alpha_{jg}\sum_{i=1}^{n_j}\delta_{ij}\left\{\frac{\lambda_{0\alpha}(t_{ij})}{\lambda_0(t_{ij})} + X^T_{ij\alpha}\right\} - w_q \exp\left\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\right\}\left\{\Lambda_{0\alpha}(t_{ij}) + \Lambda_0(t_{ij})X_{ij\alpha}\right\} \cdot$$
$$\sum_{i=1}^{n_j}\frac{\delta_{ij}}{w_q} - \Lambda_0(t_{ij})\exp\left\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\right\}, \quad q = 1:K, \alpha = 1:p \tag{6.29}$$

The components of $B$ are computed in (6.30), (6.31), (6.32) and (6.33).

$$\frac{\partial^2 l^j_{full}}{\partial \pi_g \partial \pi_l} = -\frac{z_{jg}}{\pi_g^2}\mathbf{1}_{\{g=l\}} - \frac{z_{jK}}{\pi_K^2}, \qquad g, l = 1, .., K-1 \tag{6.30}$$

$$\frac{\partial^2 l^j_{full}}{\partial w_q^2} = -\, z_{jq}\sum_{i=1}^{n_j}\frac{\delta_{ij}}{w_q^2}, \quad q = 1:K \tag{6.31}$$

$$\frac{\partial^2 l^j_{full}}{\partial \boldsymbol{\beta}_\alpha \partial \boldsymbol{\beta}_\gamma} = \sum_{k=1}^{K} z_{jk}\sum_{i=1}^{n_j}\delta_{ij}\left\{\frac{\lambda_{0\alpha\gamma}(t_{ij})\lambda_0(t_{ij}) - \lambda_{0\alpha}(t_{ij})\lambda_{0\gamma}(t_{ij})}{\lambda_0(t_{ij})^2}\right\} -$$
$$w_k \exp\left\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\right\}\left\{\Lambda_{0\alpha\gamma}(t_{ij}) + \Lambda_{0\alpha}(t_{ij})X_{ij\gamma} + \Lambda_{0\gamma}(t_{ij})X_{ij\alpha} + \Lambda_0(t_{ij})X_{ij\alpha}X_{ij\gamma}\right\},$$
$$\alpha, \gamma = 1:p \tag{6.32}$$

$$\frac{\partial^2 l^j_{full}}{\partial w_q \partial \boldsymbol{\beta}_\alpha} = z_{jq}\sum_{i=1}^{n_j} - \left(\Lambda_0(t_{ij})X_{ij\alpha} + \Lambda_{0\alpha}(t_{ij})\right)\exp\left\{\mathbf{X}^T_{ij}\boldsymbol{\beta}\right\}, \quad q = 1:K, \alpha = 1:p \tag{6.33}$$

# Appendix B

This appendix shows the results of the first simulation study with $K = 2$ hidden populations, a constant frailty ratio of $w_2/w_1 = 1.71$ and the proportion $\pi_1$ belonging to the first population varied in 9 scenarios (Table 3).

| | $\pi_1$ | $w_1$ | $w_2$ | ratio |
|---|---|---|---|---|
| 1 | 0.10 | 0.70 | 1.20 | 1.71 |
| 2 | 0.20 | 0.70 | 1.20 | 1.71 |
| 3 | 0.30 | 0.70 | 1.20 | 1.71 |
| 4 | 0.40 | 0.70 | 1.20 | 1.71 |
| 5 | 0.50 | 0.70 | 1.20 | 1.71 |
| 6 | 0.60 | 0.70 | 1.20 | 1.71 |
| 7 | 0.70 | 0.70 | 1.20 | 1.71 |
| 8 | 0.80 | 0.70 | 1.20 | 1.71 |
| 9 | 0.90 | 0.70 | 1.20 | 1.71 |

Table 3: First simulation study

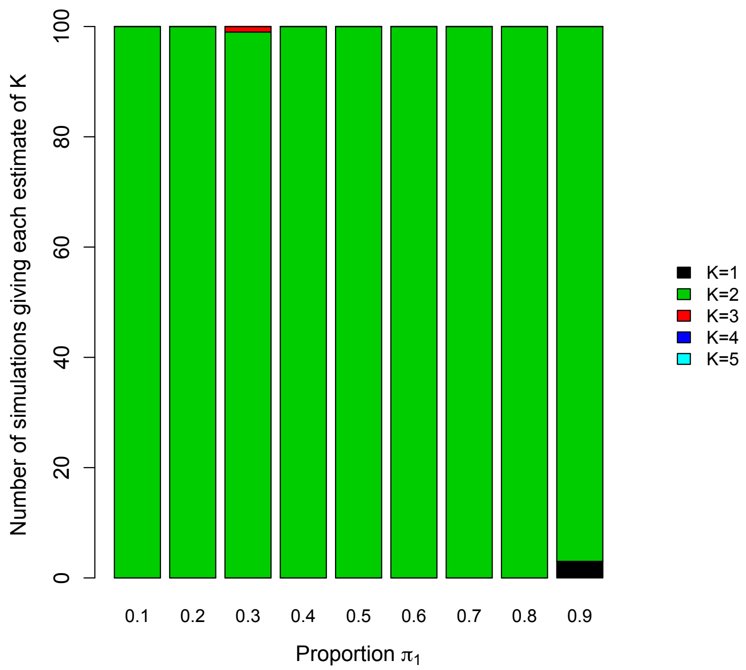


Figure 3: Estimates of number of latent populations, $K$, from model with minimum AIC, fixed frailty ratio and nine alternative values of $\pi_1$. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the nine scenarios (from the lowest $\pi_1$ on the left to the highest $\pi_1$ on the right). $\pi_1$ is associated with the lowest frailty ($w_1 = 0.7$). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$.

Fig. 3, 4 and 5, present the resulting estimates of $K$, the number of latent populations, using three alternative methods of model selection. Each bar represents one of the nine scenarios with alternative values of $\pi_1$. The proportion of the 100 simulations estimating each value of $K$ from 1 to 5 is indicated by stacked bars of different colours. The majority of simulations estimate the correct value of $K = 2$ for all three model selection methods, otherwise for AIC and BIC $K = 1$ is the next most common estimate, and, for the method of Laird (1978), $K = 3$.

Fig. 6 shows that the mixing proportion $\pi_1$ is well estimated in all scenarios. Fig. 7 shows that the frailty ratio of 1.71 tends to be estimated more accurately when the mixing proportion is closer to 0.5.

# Appendix C

This appendix shows the results of the second simulation study with $K = 2$, the proportion of groups with the lower frailty value fixed at $\pi_1 = 0.3$, and the frailty ratio $w_2/w_1$ varying in 7 scenarios from 1.14 to 3.00 (Table 4).

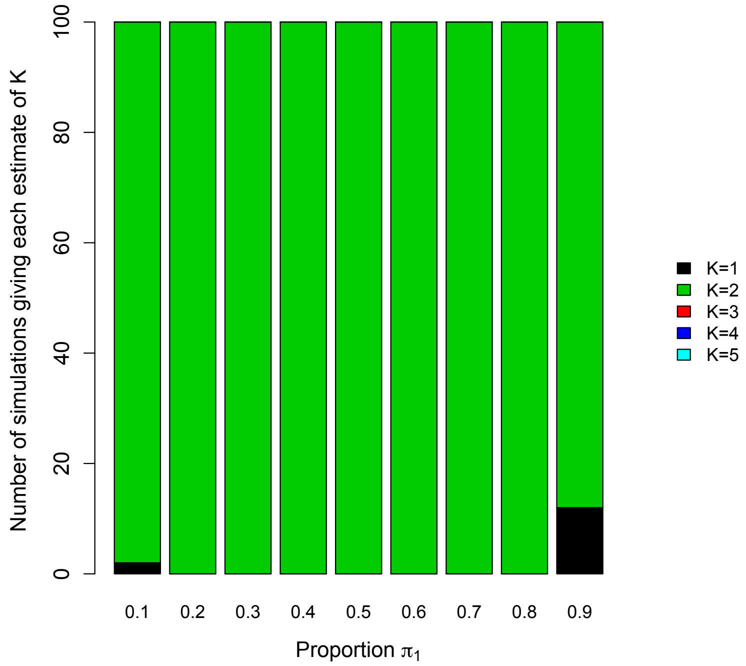


Figure 4: Estimates of number of latent populations, $K$, from model with minimum BIC, fixed frailty ratio and nine alternative values of $\pi_1$. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the nine scenarios (from the lowest $\pi_1$ on the left to the highest $\pi_1$ on the right). $\pi_1$ is associated with the lowest frailty ($w_1 = 0.7$). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$.

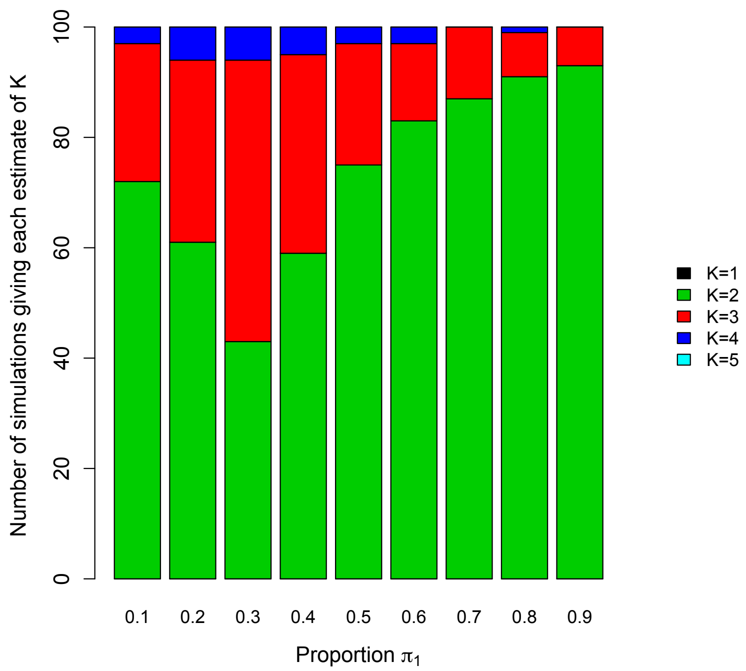


Figure 5: Estimates of number of latent populations, $K$, up to Laird (1978) criterium, fixed frailty ratio and nine alternative values of $\pi_1$. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the nine scenarios (from the lowest $\pi_1$ on the left to the highest $\pi_1$ on the right). $\pi_1$ is associated with the lowest frailty ($w_1 = 0.7$). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$.

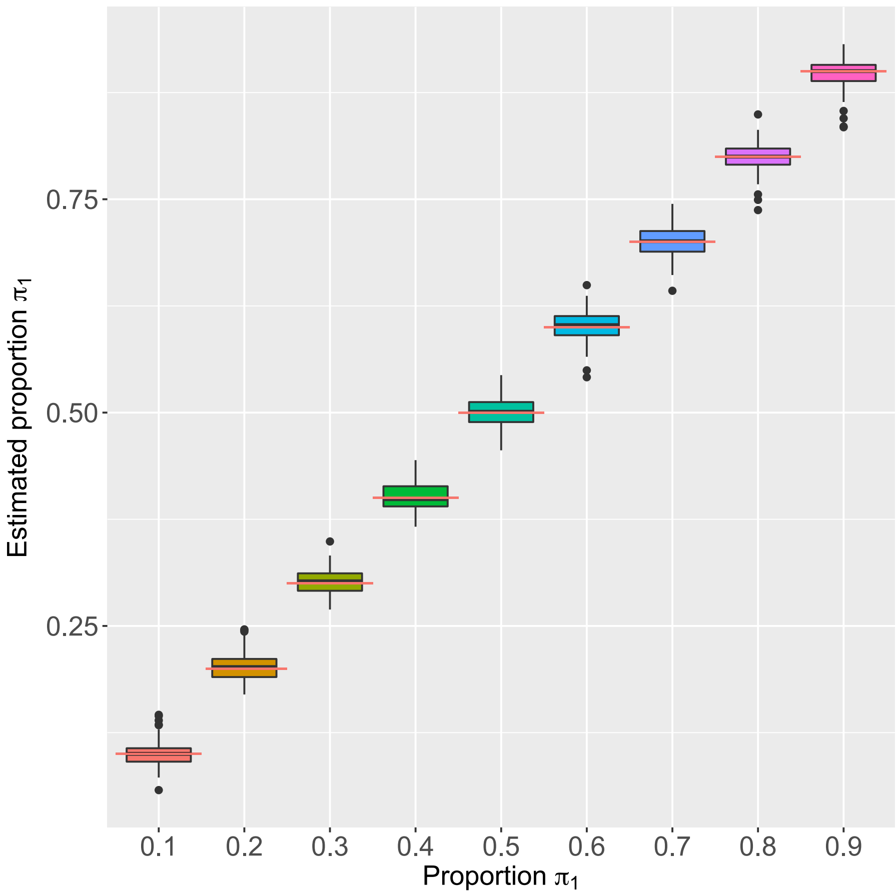


Figure 6: Estimates of $\pi_1$, fixed frailty ratio and nine alternative values of $\pi_1$. We represent the nine boxplot (median and quantiles) of the maximum likelihood estimators for $\pi_1$ over all 100 simulations, for each case. The red lines represent the real values.

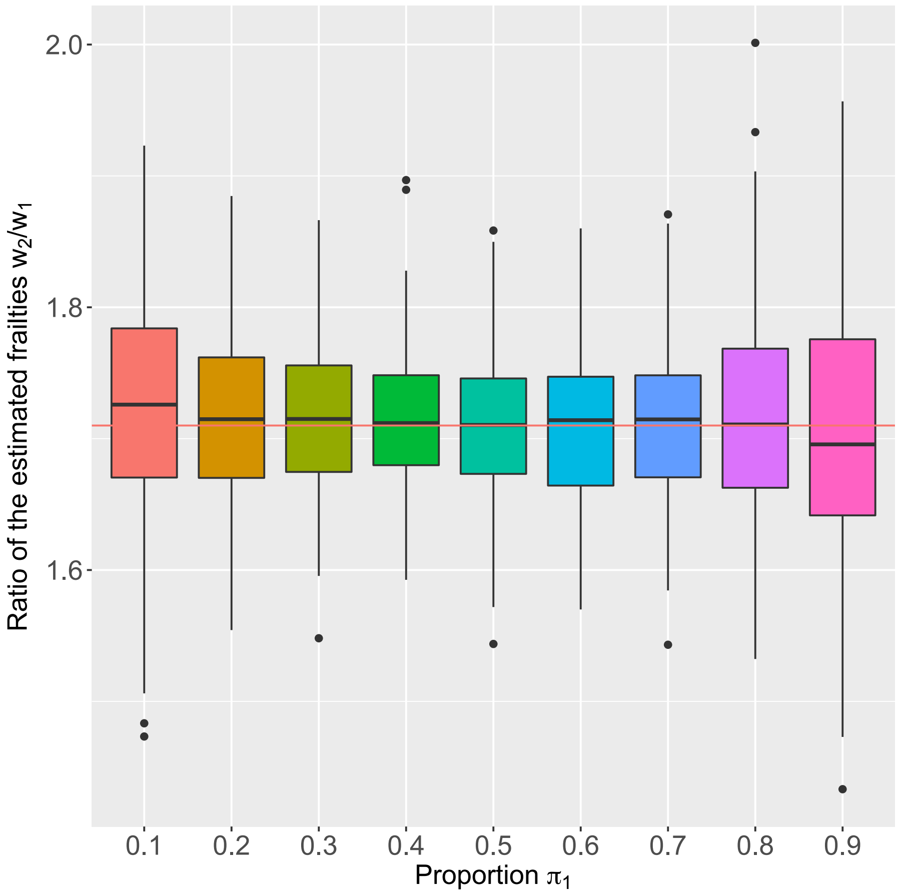


Figure 7: Estimates of the ratio $w_2/w_1$, fixed frailty ratio and nine alternative values of $\pi_1$. We represent the nine boxplot (median and quantiles) of the ratio of the maximum likelihood estimators for $w_1$ and $w_2$ over all 100 simulations, for each case. The red lines represent the real values.

| | $\pi_1$ | $w_1$ | $w_2$ | ratio |
|---|---|---|---|---|
| 1 | 0.30 | 0.70 | 0.80 | 1.14 |
| 2 | 0.30 | 0.70 | 0.90 | 1.29 |
| 3 | 0.30 | 0.70 | 1.00 | 1.43 |
| 4 | 0.30 | 0.70 | 1.10 | 1.57 |
| 5 | 0.30 | 0.70 | 1.20 | 1.71 |
| 6 | 0.30 | 0.70 | 1.40 | 2.00 |
| 7 | 0.30 | 0.70 | 2.10 | 3.00 |

Table 4: Second simulation study

As before, in Fig. 8, 9 and 10, we present the estimates of $K$, the number of latent populations. Broadly for all three model selection methods, the larger the contrast in frailties $w_2/w_1$ between the populations, the more frequently the true $K = 2$ is obtained.

The estimates of the mixing proportion are represented in Fig. 11, and the estimates of the frailty ratio in Fig. 12, showing that more accurate estimates are obtained as the frailty ratio increases, thus as the contrast between the two populations becomes greater.

# Appendix D

In Appendix D, we show the full results of the third simulation study, where the number of latent populations is varied in three scenarios with $K = 2, 3, 4$ in turn, while the mixing proportions and frailty ratios are fixed at values listed in Table 5.

| K | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\pi_4$ | $w_1$ | $w_2$ | $w_3$ | $w_4$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.40 | 0.60 | - | - | 2 | 3 | - | - |
| 3 | 0.20 | 0.30 | 0.50 | - | 2 | 3 | 5 | - |
| 4 | 0.15 | 0.25 | 0.30 | 0.30 | 2 | 3 | 5 | 8 |

Table 5: Third simulation study

In Fig. 13, 14 and 15, we present the estimates of $K$, the number of latent populations, under three methods of model selection. We note that AIC estimates the true value of $K$ in the majority of simulations in the three considered frameworks, Fig. 13. Both AIC and BIC show the best performances in the case of real $K = 2$ and the worst performances in the case of real $K = 3$. BIC estimates one latent population less than the true value in the majority of cases for both real $K = 3$ and $K = 4$, Fig. 14. Finally, Laird (1978) method tends to estimate the true $K$ in about half of the simulations, but in the other half it tends to estimate one or two latent populations more than the true value, Fig. 15.

The mixing proportions (Fig. 16) and the frailty ratios (Fig. 17) are estimated accurately.

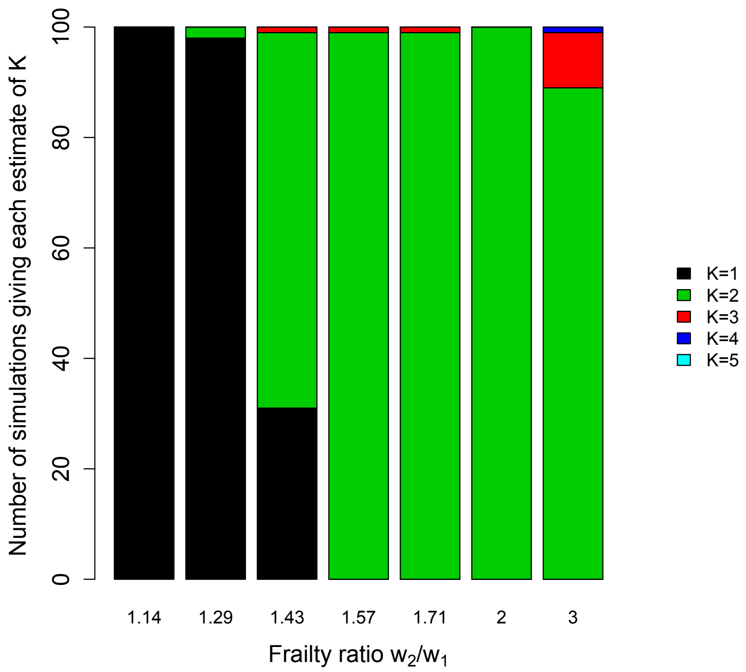


Figure 8: Estimates of number of latent populations, $K$, from model with minimum AIC, fixed $\pi_1$ and seven alternative values of $w_2/w_1$. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the seven scenarios (from the lowest $w_2/w_1$ on the left to the highest $w_2/w_1$ on the right). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$.

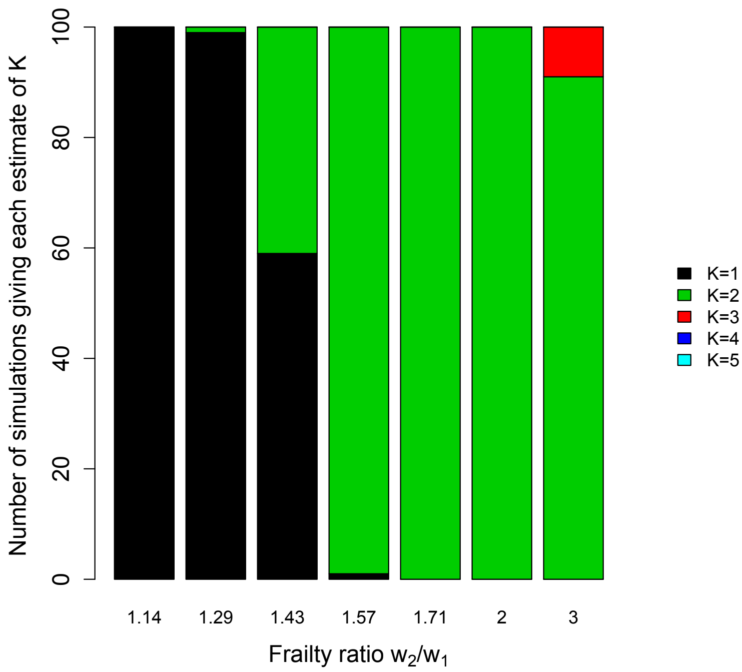


Figure 9: Estimates of number of latent populations, $K$, from model with minimum BIC, fixed $\pi_1$ and seven alternative values of $w_2/w_1$. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the seven scenarios (from the lowest $w_2/w_1$ on the left to the highest $w_2/w_1$ on the right). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$.

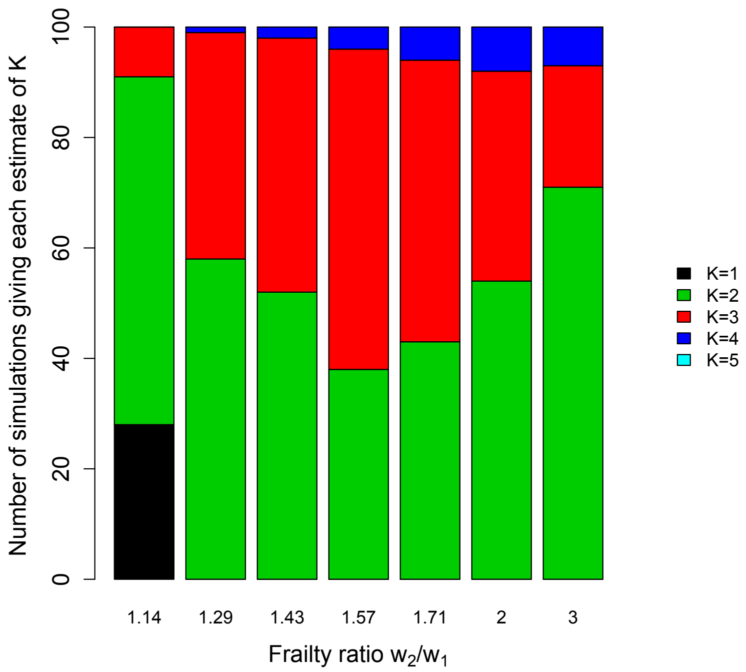


Figure 10: Estimates of number of latent populations, $K$, from model up to Laird (1978) criterium, fixed $\pi_1$ and 7 alternative values of $w_2/w_1$. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the 7 scenarios (from the lowest $w_2/w_1$ on the left to the highest $w_2/w_1$ on the right). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$.

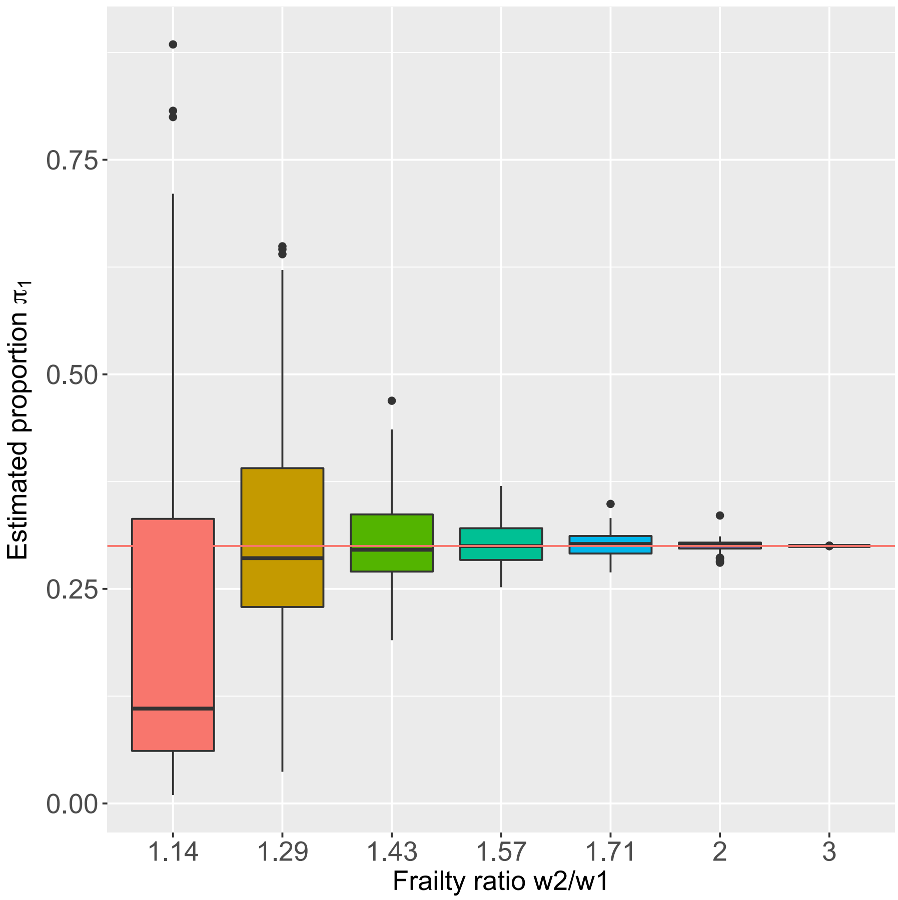


Figure 11: Estimates of $\pi_1$, fixed $\pi_1$ and seven alternative values of $w_2/w_1$. We represent the seven boxplot (median and quantiles) of the maximum likelihood estimators for $\pi_1$ over all 100 simulations, for each case. The red lines represent the real values.

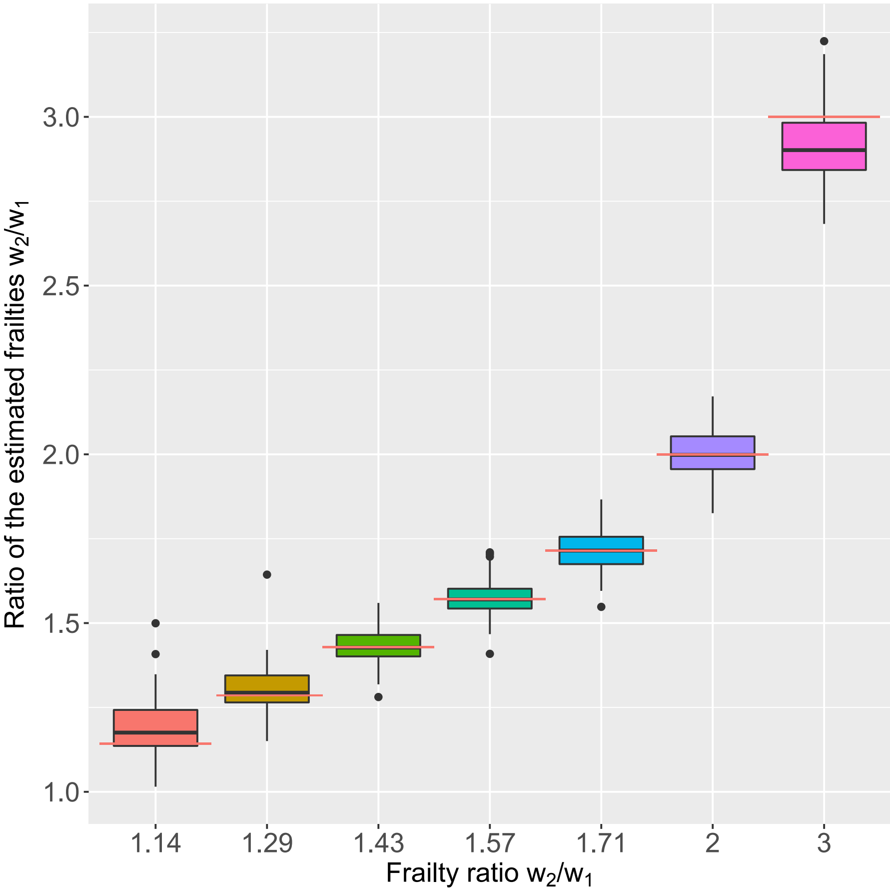


Figure 12: Estimates of $w_2/w_1$, fixed $\pi_1$ and seven alternative values of $w_2/w_1$. We represent the seven boxplot (median and quantiles) of the ratio of the maximum likelihood estimators for $w_2$ and $w_1$ over all 100 simulations, for each case. The red lines represent the real values.

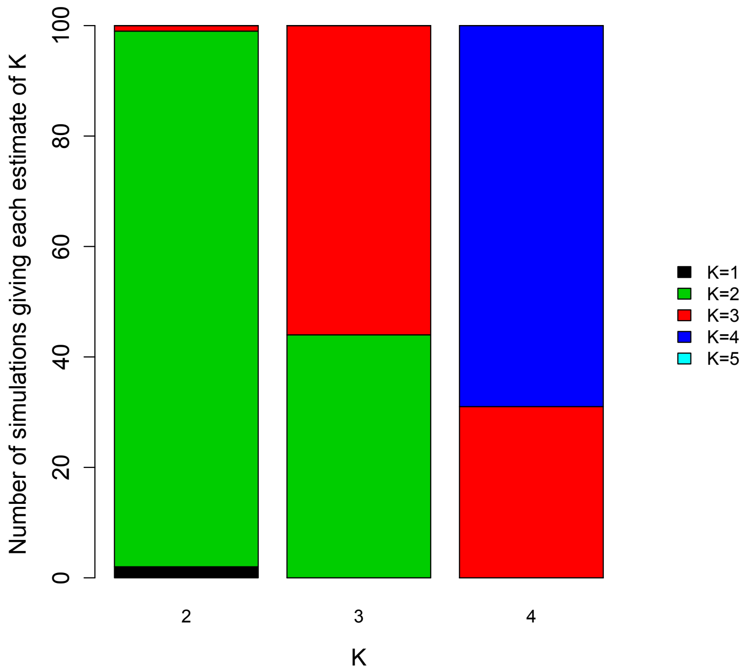


Figure 13: Estimates of number of latent populations, $K$, from model with minimum AIC. In this case, we vary all the parameters up to Table 5. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the three scenarios (from the lowest $K$ on the left to the highest $K$ on the right). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$ in the first bar on the left, $K = 3$ in the central bar and $K = 4$ in the third bar from the left.

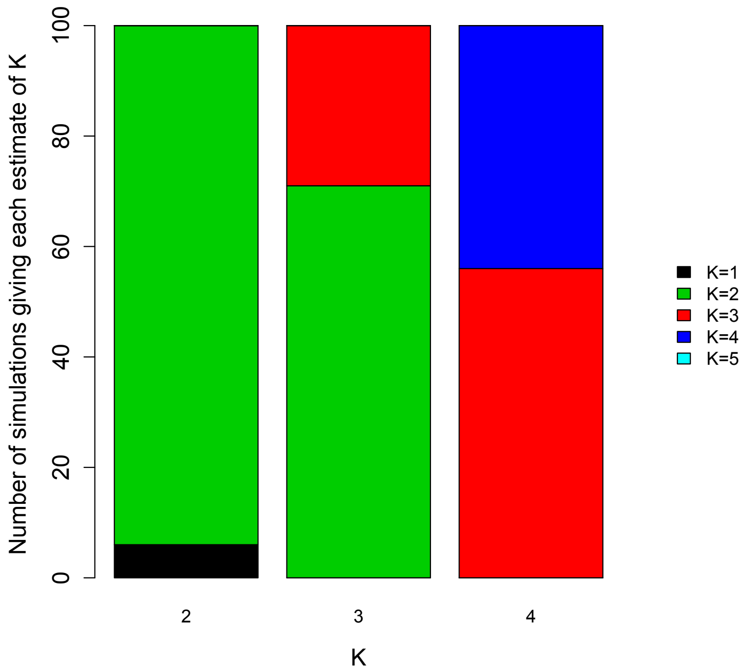


Figure 14: Estimates of number of latent populations, $K$, from model with minimum BIC. In this case, we vary all the parameters up to Table 5. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the three scenarios (from the lowest $K$ on the left to the highest $K$ on the right). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$ in the first bar on the left, $K = 3$ in the central bar and $K = 4$ in the third bar from the left.

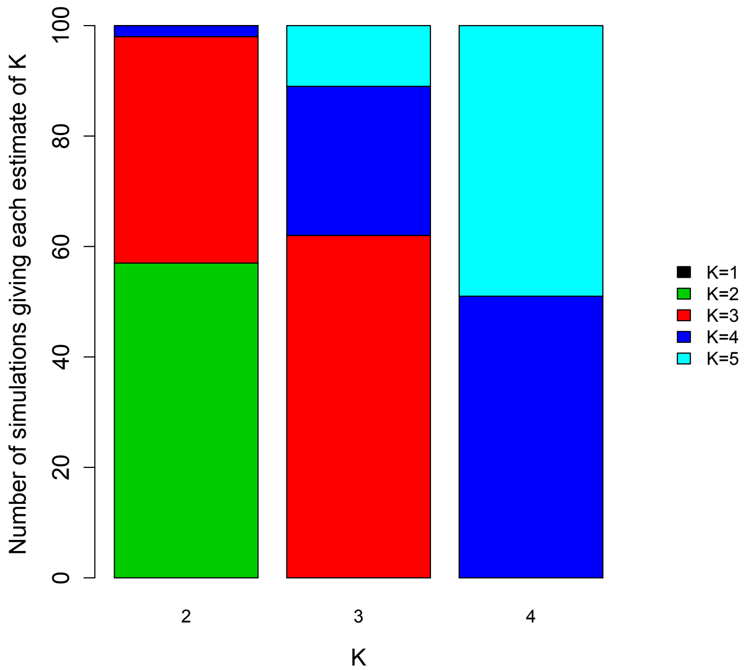


Figure 15: Estimates of number of latent populations, $K$, from model up to Laird (1978). In this case, we vary all the parameters up to Table 5. In the y-axis we count the total number of simulations giving each estimate of K, while on x-axis we show the three scenarios (from the lowest $K$ on the left to the highest $K$ on the right). The part of the bars in black is the total number of simulations in which the best model estimate $K = 1$, the green part of the bars represents the total number of simulations in which the best model estimate $K = 2$ and so o and so forth. The true model has $K = 2$ in the first bar on the left, $K = 3$ in the central bar and $K = 4$ in the third bar from the left.

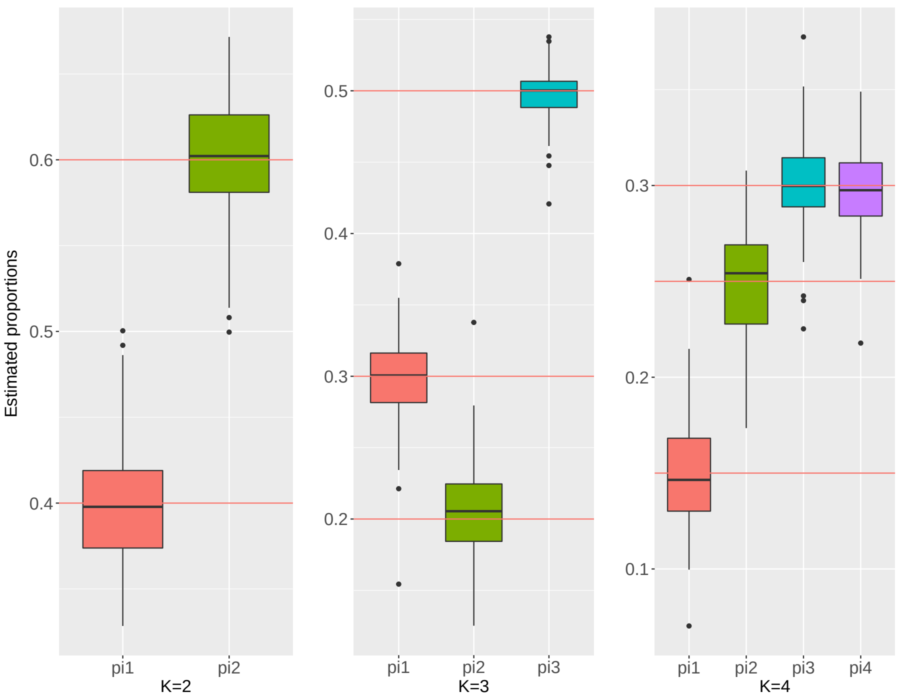


Figure 16: Estimates of the mixing proportions. In the first boxplot from the left we have $K = 2$ and the real values are $\boldsymbol{w} = [0.4, 0.6]$, in the second we have $K = 3$ and the real values are $\boldsymbol{\pi} = [0.2, 0.3, 0.5]$, in the third we have $K = 4$ and the real values are $\boldsymbol{\pi} = [0.15, 0.25, 0.30, 0.30]$. The red lines represent the real values.

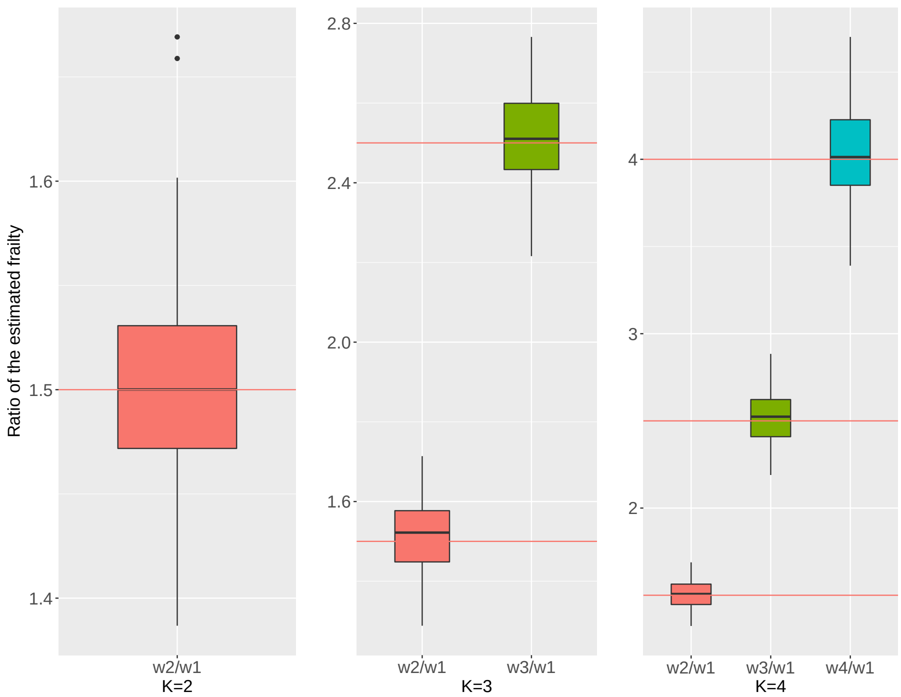


Figure 17: Ratios of the estimated ratios. In the first boxplot from the left we have $K = 2$ and the real values are $w_2/w_1 = 1.5$, in the second we have $K = 3$ and the real values are $\mathbf{w}/w_1 = [1.5, 2.5]$, in the third we have $K = 4$ and the real values are $\mathbf{w}/w_1 = [1.5, 2.5, 4]$. The red lines represent the real values.

# MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

**43/2017** Bottle, A.; Ventura, C.M.; Dharmarajan, K.; Aylin, P.; Ieva, F.; Paganoni, A.M.
*Regional variation in hospitalisation and mortality in heart failure: comparison of England and Lombardy using multistate modelling*

**44/2017** Martino, A.; Ghiglietti, A.; Ieva, F.; Paganoni, A.M.
*A k-means procedure based on a Mahalanobis type distance for clustering multivariate functional data*

**42/2017** Gower, AL, Shearer, T, Ciarletta P
*A new restriction for initially stressed elastic solids*

**41/2017** Beretta, E.; Micheletti, S.; Perotto, S.; Santacesaria, M.
*Reconstruction of a piecewise constant conductivity on a polygonal partition via shape optimization in EIT*

**39/2017** Ciarletta, P.
*Matched asymptotic solution for crease nucleation in soft solids*

**40/2017** Ciarletta, P.
*Matched asymptotic solution for crease nucleation in soft solids*

**38/2017** Bonaventura, L.; Fernandez Nieto, E.; Garres Diaz, J.; Narbona Reina, G.;
*Multilayer shallow water models with locally variable number of layers and semi-implicit time discretization*

**37/2017** Formaggia, L.; Vergara, C.; Zonca, S.
*Unfitted Extended Finite Elements for composite grids*

**35/2017** Piercesare Secchi
*On the role of statistics in the era of big data: a call for a debate*

**36/2017** Koeppl, T.; Vodotto, E.; Wohlmuth, B.; Zunino, P.
*Mathematical modelling, analysis and numerical approximation of second order elliptic problems with inclusions*