Insurance: Mathematics and Economics 107 (2022) 68–87

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

www.elsevier.com/locate/ime

Mortality modeling and regression with matrix distributions

Hansjörg Albrecher [a], Martin Bladt [a,*], Mogens Bladt [b], Jorge Yslas [c]

[a] Department of Actuarial Science, Faculty of Business and Economics, University of Lausanne, UNIL-Dorigny, CH-1015 Lausanne, Switzerland
[b] Department of Mathematics, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark
[c] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, CH-3012 Bern, Switzerland

ARTICLE INFO

Article history:
Received November 2021
Received in revised form July 2022
Accepted 1 August 2022
Available online 17 August 2022

JEL classification:
C13
G22
J11

Keywords:
Survival analysis
Regression models
Phase-type distributions
Inhomogeneous phase-type distributions
Inhomogeneous Markov processes

ABSTRACT

In this paper we investigate the flexibility of matrix distributions for the modeling of mortality. Starting from a simple Gompertz law, we show how the introduction of matrix-valued parameters via inhomogeneous phase-type distributions can lead to reasonably accurate and relatively parsimonious models for mortality curves across the entire lifespan. A particular feature of the proposed model framework is that it allows for a more direct interpretation of the implied underlying aging process than some previous approaches. Subsequently, towards applications of the approach for multi-population mortality modeling, we introduce regression via the concept of proportional intensities, which are more flexible than proportional hazard models, and we show that the two classes are asymptotically equivalent. We illustrate how the model parameters can be estimated from data by providing an adapted EM algorithm for which the likelihood increases at each iteration. The practical feasibility and competitiveness of the proposed approach, including the right-censored case, are illustrated by several sets of mortality and survival data.

1. Introduction

The statistical modeling of human lifetimes is a central topic in actuarial science, as it has immediate consequences for the pricing and management of a wide range of insurance and pension products (see, e.g., Olivieri and Pitacco (2015), Dickson et al. (2019)). Starting all the way back with Gompertz (1825), who proposed a mortality rate that increases exponentially over the lifetime, the field has seen a tremendous development over the last 200 years, and nowadays, sophisticated models are available that take into account the changing nature of mortality over time as well as common and individual trends across different (sub)populations (see, e.g., Pitacco (2004, 2019) for surveys and Denuit and Trufin (2016), Dowd et al. (2020), Lin et al. (2021), Shapovalov et al. (2021), Zeddouk and Devolder (2020) and Renshaw and Haberman (2021) for some very recent developments in different research directions, and Barigou et al. (2021) for a proposition how to combine various alternative models into one). A detailed understanding of these trends is crucial for prudent management of longevity products (Barrieu et al. (2012), Sherris and Zhou (2014)) and the future of old-age provision on the society level in general (see, e.g., Albrecher et al. (2016)). At the same time, the statistical estimation of mortality models from given (often incomplete and interval-censored) mortality data can be quite challenging; see, e.g., Macdonald et al. (2018) for a recent account. In the presence of the many available model options for a given purpose (e.g., constructing lifetables or making mortality projections), a number of countries nowadays rely on an accorded effort of academia and insurance practice for respective recommendations or guidelines (see, e.g., Antonio et al. (2017) for an excellent survey on the current approach of Belgium and the Netherlands).

Most of the models in practical use are based on a purely statistical fitting of observed mortality patterns and subsequently extrapolating the trends of the past into the future, which historically has sometimes led to an underestimation of the actual mortality improvement. In addition, the resulting models typically rely on an enormous number of parameters, like for instance, 300 in the very popular Lee-Carter model (Lee and Carter (1992)), which employs separate parameters for each year and each age, that then need to be estimated from the given data, cf. Li et al. (2009). This may not be considered an issue, as long as the values of these parameters remain reasonably stable

* Corresponding author.
E-mail addresses: hansjoerg.albrecher@unil.ch (H. Albrecher), martin.bladt@unil.ch (M. Bladt), bladt@math.ku.dk (M. Bladt), jorge.yslas@stat.unibe.ch (J. Yslas).

https://doi.org/10.1016/j.insmatheco.2022.08.001

during updates, and if for instance cohort effects are thought of dominating the stochastic nature of mortality. At the same time, it can be interesting to see whether, through other and possibly parsimonious approaches, one can also capture observed mortality patterns to a satisfactory degree. Motivated by corresponding approaches in medicine and early contributions by, e.g., Gutterman and Vanderhoof (1998) in the actuarial context, Lin and Liu (2007) proposed an interesting alternative approach for the modeling of human lifetimes that links its parameters to the biological and physiological mechanisms of aging. Concretely, one may imagine a lifetime as traversing through a number of stages that can be thought of as markers of the biological age, with the time that it takes an individual to go from one stage to the next being random, and an additional possibility to die 'prematurely' at each stage along the way (e.g., due to accidents or other external causes). If such a model can be calibrated to mortality data in a satisfactory way, then one could possibly attribute particular markers to each stage, linking the parameters of the model more directly to the physiological aging mechanism, and opening up the possibility to include expert opinions on particular mortality trends into the models and the resulting forecasts more directly. In addition, such an approach could also allow to compare the mortality patterns of different (sub)populations on the basis of changes of transition rates between those stages, which is a somewhat appealing alternative to the purely statistical approach that is typically followed today, see Lin and Liu (2007) and also Cheng et al. (2020) for a detailed discussion as well as Hassan Zadeh et al. (2014) for an application in a disability context.

A natural candidate for a model of the aging process of the human body is a finite-state homogeneous continuous-time Markov process with a single absorbing state (death). In the context of modeling stages of a disease, such a Markov approach can, e.g., be traced back to Kay (1986), Longini et al. (1989), Guihenneuc-Jouyaux et al. (2000). For early contributions towards human lifetime modeling along these lines, see Gavrilov and Gavrilova (1991), and Aalen (1995) who particularly focused on the modeling of hazard rates. The resulting class of distributions for the lifetime are known as *phase-type distributions*. Phase-type (PH) distributions are particularly tractable and have been systematically studied as matrix distributions over the years (see Neuts (1975, 1981) and Bladt and Nielsen (2017) for a survey). One of the most attractive features of PH distributions from a statistical perspective is that they are dense in the class of distributions on the positive real line in the sense of weak convergence, so that any given distribution on the positive half-line can be approximated arbitrarily well. However, for a good fit to a given histogram of data, often a large number of states (phases) are required, making the estimation procedure complex. Several estimation methods have been proposed in the literature, including the method of moments in Bobbio et al. (2005); Horváth and Telek (2007) and references therein, Bayesian inference in Bladt et al. (2003), and maximum likelihood estimation via the expectation-maximization (EM) algorithm in Asmussen et al. (1996) (for the right-censored case see Olsson (1996)). The EM method considers the full likelihood arising from the path representation of a PH distribution and has been the most popular approach for many applications in applied probability.

In Lin and Liu (2007), the human aging process is modeled by such a PH distribution (concretely, a Coxian distribution). It turns out that for an adequate fit to data (the eventual lifetimes), about 200-250 phases are needed (see also Asmussen et al. (2019), where about 50 phases are suggested in the context of pricing equity-linked insurance products, when the focus is more on the stability of eventual prices rather than the accuracy of the fit). Cheng et al. (2020) introduce additional restrictions of the parameters for the fitting procedure. The main reason for the need of so many phases (here and in other applications) towards an adequate fit is typically due to the fact that all PH distributions have exponentially decaying tails, whereas the actual data do not have that property. Rather than using many phases to correct for that, Albrecher and Bladt (2019) proposed a framework of inhomogeneous Markov processes for the PH construction, where time evolves according to a deterministic transform of the original time, which can be targeted towards capturing the tail behavior of the resulting distribution via the transform rather than the introduction of additional states. The resulting class of *inhomogeneous phase-type distributions* (IPH) is still dense in the class of all distributions on the positive half-line and turns out to lead to adequate fits with substantially fewer phases (and consequently parameters) in various applications. The statistical estimation of IPH distributions was then subsequently developed in Albrecher et al. (2022b). A number of IPH distributions can be seen as matrix extensions of other base distributions, where the latter determine the tail behavior, and the matrix-valued parameters improve the fit of the formerly scalar counterparts, cf. Albrecher and Bladt (2019).

The purpose of this paper is to study the potential of IPH distributions for the modeling of human mortality. In particular, we will look into matrix versions of the Gompertz distribution, which will give the needed additional flexibility for fitting observed mortality rates (hazard functions), yet keeping the number of phases (and consequently parameters) of the model reasonably low. Importantly, the resulting model will allow for an interpretation in terms of the above aging process through phases, with time being transformed according to a deterministic mechanism. The focus in the modeling will be on an adequate representation of the hazard rate. In that way, we combine the original approach of hazard rate modeling with a generalized version of Lin and Liu (2007). The advantage of this combined approach is that the number of necessary phases for a visually acceptable fit goes down from 200-250 to 10-12, and the approach does not have to be restricted to the sub-class of Coxian PH distributions.

In a second step, we will then investigate (for the first time) regression of IPH distributions, which will lead a way to multi-population mortality modeling within the above framework of modeling the aging mechanism. In the life sciences, particular instances of PH distributions (namely Coxian distributions) were considered as survival regression models, within the accelerated failure time (AFT) specification, cf. McGrory et al. (2009); Rizk et al. (2021); Tang et al. (2012) and references therein. The popularity of these models for several decades is the product of their natural interpretation in terms of states and their statistical efficacy. Markov Chain Monte Carlo (MCMC) methods seem to be particularly popular for this type of application. In this paper, we propose survival regression models based on general IPH distributions. Concretely, we propose a proportional intensities model for the regression, which contains a number of classical methods as special cases (such as the proportional hazards model for any parametric family, the AFT model and Coxian regression). The combination of inhomogeneous time transformation together with general sub-intensity matrices makes our model significantly more flexible both in terms of tails and hazard function shapes. Regressing on the inhomogeneity parameters also leads to matrix generalizations of models such as the one of Hsieh and Lavori (2000). A particular case of our specification was studied in Bladt (2021) for severity modeling, where right-censoring, aggregation, and regressing on the inhomogeneity parameters were not considered.

To make the models practically useful, we adapt techniques developed in Asmussen et al. (1996); Olsson (1996) and Albrecher et al. (2022b) to obtain effective estimation procedures based on the EM algorithm. While the focus in this paper is on mortality modeling, the generality of the approach could, with corresponding adaptations of the techniques, also be of interest for other application areas. We would also like to stress that with its great flexibility comes a drawback, which is that our model does not and is not expected to

outperform any individual survival regression model specification in the literature. In particular, the message that we aim to convey is that inhomogeneous Markov aging models are reaching a mature state where they actually are able to resemble many of the classical models when it comes to explaining real data. Prediction and other more delicate aspects of mortality modeling (in particular in connection with time effects and cohort effects) are supported but do not outperform the state-of-the-art machine learning methods. Nonetheless, our models enjoy favorable estimation and closed-form mathematical formulas for the resulting lifetime distribution. The latter is important to build realistic insurance products that can be analyzed explicitly, and thus the contribution is also targeted to researchers who want to use our model as a building block in their own research (with the advantage that they may calibrate them on the entire lifetime distribution).

The structure of the remainder of the paper is as follows. In Section 2, we revisit the necessary existing background on IPH distributions with an emphasis on the intensity function, which plays a central role in the paper, and establish a number of explicit expressions in the context of mortality modeling with matrix distributions, addressing the main PH classes under consideration in the paper in more detail. Section 3 is devoted to regression and proposes a proportional intensities model. Main properties are derived and it is shown that it is asymptotically equivalent to the proportional hazards model, but more flexible for our purposes. Section 4 then develops the estimation methodology for the regression models via the EM algorithm. The presented methods and models are illustrated in Section 5 for mortality data from Denmark, Japan, and the USA, both for the univariate case and the regression. In particular, it is shown that for these data dimensions, around 10-12 in the matrix framework are already reasonably competitive when for instance, compared to the classical Lee-Carter model. For survival data, Section 6 implements the right-censored version of the main algorithm. Finally, Section 7 concludes.

2. Mortality rate modeling using matrix distributions

In this section, we recall some basic properties of inhomogeneous phase-type distributions that will be relevant later on. In the sequel, for a random variable X, we write $X \sim F$ with F a distribution function, density, or acronym, when X follows the distribution uniquely associated with F. For two real-valued functions, g, h the terminology $g(t) \sim h(t)$, as $t \to \infty$ is defined as $\lim_{t\to\infty} g(t)/h(t) = 1$. Unless stated otherwise, equalities between random objects hold almost surely.

We will denote by $\mu_f(s)$ the hazard (or mortality) rate of a lifetime random variable X with density f, so that

$$\overline{F}(x) = \exp\left(-\int_0^x \mu_f(s)ds\right) \tag{2.1}$$

is the corresponding survival function, and $\mu_f(s)ds$ can be interpreted as the probability of dying in the time interval $[s, s+ds)$ given that the individual has survived up until time s. The variation of $\mu_f(s)$ with s is linked to the aging process.

In this paper, we make the assumption that

$$\mu_f(x) = e^{\beta x},$$

for some $\beta > 0$, that is, we have an underlying Gompertz hazard rate. For presentation purposes, and also because many of the formulas also work for other hazard rates, we will still often work abstractly with the symbol $\mu_f(x)$ in the sequel.

Now let $(J_t)_{t\geq 0}$ denote a time-inhomogeneous Markov jump process on a state-space $\{1, \ldots, p, p+1\}$, where states $1, \ldots, p$ are transient and state $p+1$ is absorbing. In other words, $(J_t)_{t\geq 0}$ has an intensity matrix of the form

$$\boldsymbol{\Lambda}(t) = \begin{pmatrix} \boldsymbol{T}(t) & \boldsymbol{t}(t) \\ \boldsymbol{0} & 0 \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}, \quad t \geq 0,$$

where $\boldsymbol{T}(t)$ is a $p \times p$ matrix function and $\boldsymbol{t}(t)$ is a p-dimensional column vector function. Here, for any time $t \geq 0$, $\boldsymbol{t}(t) = -\boldsymbol{T}(t)\boldsymbol{e}$, where $\boldsymbol{e}$ is the p-dimensional column vector of ones. Let $\pi_k = \mathbb{P}(J_0 = k)$, $k = 1, \ldots, p$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ be a probability vector corresponding to the initial distribution of the jump process. In particular, we have $\mathbb{P}(J_0 = p+1) = 0$. Then we say that the time until absorption

$$\tau = \inf\{t \geq 0 \mid J_t = p+1\},$$

has an inhomogeneous phase-type distribution with representation $(\boldsymbol{\pi}, \boldsymbol{T}(t))$. For the purpose of this paper, the following particular case is of interest: $\boldsymbol{T}(t) = \lambda(t)\boldsymbol{T}$, where $\lambda(t)$ is some known non-negative real function, named the *intensity function*, and $\boldsymbol{T} = \{t_{kl}\}_{1\leq k,l\leq p}$ is a sub-intensity matrix. Similarly, we let $-\boldsymbol{T}\boldsymbol{e} = \boldsymbol{t} = \{t_k\}_{1\leq k\leq p}$. In practice, we will usually obtain λ as a hazard rate from a known distribution, for instance $\lambda = \mu_f$. We write

$$\tau \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda).$$

Note that for $\lambda(t) \equiv 1$ one returns to the time-homogeneous case, which corresponds to the conventional phase-type distributions with notation $\mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$; a comprehensive account of PH distributions can be found in Bladt and Nielsen (2017). It is not difficult to show that if $Y \sim \mathrm{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$, then there exists a function g such that

$$Y \sim g(Z), \tag{2.2}$$

where $Z \sim \mathrm{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. Specifically, g is defined in terms of λ by

$$g^{-1}(y) = \int_0^y \lambda(s)ds, \quad y \geq 0,$$

or, equivalently,

$$\lambda(y) = \frac{d}{dy} g^{-1}(y)\,.$$

To avoid degeneracies, we assume that $g^{-1}(y) < \infty$, $\forall y \geq 0$, and $\lim_{y\to\infty} g^{-1}(y) = \infty$. The density f_Y and survival function $\overline{F}_Y$ of $Y \sim \text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$ are given by

$$f_Y(y) = \lambda(y)\boldsymbol{\pi} \exp\left(\int_0^y \lambda(s)ds\, \boldsymbol{T}\right)\boldsymbol{t}\,, \quad y \geq 0\,,$$

$$\overline{F}_Y(y) = \boldsymbol{\pi} \exp\left(\int_0^y \lambda(s)ds\, \boldsymbol{T}\right)\boldsymbol{e}\,, \quad y \geq 0\,. \tag{2.3}$$

For further properties on IPH distributions and motivation for their use in statistical modeling, we refer to Albrecher and Bladt (2019). See also Albrecher et al. (2022b) for extensions to the multivariate case.

The tail behavior of IPH distributions can be derived from the corresponding tail behavior of a PH distribution. Recall that the survival function of a PH distribution is given by

$$\overline{F}_Z(y) = \sum_{j=1}^{m}\sum_{l=0}^{\kappa_j-1} y^l \exp(\Re\left(-\eta_j\right)y)\left[a_{jl}\sin\left(\Im\left(-\eta_j\right)y\right) + b_{jl}\cos\left(\Im\left(-\eta_j\right)y\right)\right]\,, \quad y \geq 0\,, \tag{2.4}$$

where $-\eta_j < 0$ are the eigenvalues of the Jordan blocks $\boldsymbol{J}_j$ of $\boldsymbol{T}$, with corresponding dimensions κ_j, $j = 1, \ldots, m$, and a_{jl} and b_{jl} are constants depending on $\boldsymbol{\pi}$ and $\boldsymbol{T}$. If $-\eta$ is the largest real eigenvalue of $\boldsymbol{T}$ and n is the dimension of the Jordan block of η, then it is easy to see from (2.4) that

$$\overline{F}_Z(y) \sim c y^{n-1}\exp(-\eta y)\,, \quad y \to \infty\,, \tag{2.5}$$

where c is a positive constant. That is, all PH distributions have exponential tails with Erlang-like second-order bias terms. The tail behavior of Y then follows from (2.2).

A number of IPH distributions can be expressed as classical distributions with matrix-valued parameters. For the representation of such distributions, we make use of functional calculus. If ν is an analytic function and $\boldsymbol{A}$ is a matrix, define

$$\nu(\boldsymbol{A}) = \frac{1}{2\pi i}\oint_\Gamma \nu(w)(w\boldsymbol{I} - \boldsymbol{A})^{-1}dw\,,$$

where Γ is a simple path enclosing the eigenvalues of $\boldsymbol{A}$; cf. (Bladt and Nielsen, 2017, Sec. 3.4) for details. In particular, the matrix exponential can be defined in this way.

The matrix distribution of particular interest for the mortality modeling in this paper is defined through the following survival function:

$$\overline{H}(x) = \boldsymbol{\pi} \exp\left(\boldsymbol{T}\int_0^x \mu_f(s)ds\right)\boldsymbol{e} = \boldsymbol{\pi} e^{\boldsymbol{T}\left(e^{\beta x}-1\right)/\beta}\boldsymbol{e}\,.$$

Here, $(\boldsymbol{\pi}, \boldsymbol{T})$ is a p-dimensional PH representation (so that the underlying Markov process has p transient states). In view of (2.3), this is the survival function of the time until absorption of a time-inhomogeneous Markov process initiating according to $\boldsymbol{\pi}$ and with intensity matrix

$$\boldsymbol{\Lambda}(t) = \mu_f(t)\begin{pmatrix}\boldsymbol{T} & \boldsymbol{t}\\ \boldsymbol{0} & 0\end{pmatrix}\,. \tag{2.6}$$

Hence $H \sim \text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \mu_f)$. If $\boldsymbol{T}$ is one-dimensional taking the value -1, then H coincides with F in (2.1).

This *matrix-Gompertz* distribution $H \sim \text{mGompertz}(\beta, \boldsymbol{\pi}, \boldsymbol{T})$ has tails much lighter than exponential. Observe that from an inhomogeneous Markov point of view, the sub-intensity matrix of the underlying jump process is given by $\boldsymbol{T}\exp(\beta x)$, which for scalar $\boldsymbol{T}$ leads back to the exponential hazard function of the Gompertz distribution, the latter being the classical model for human mortality over long age ranges (see, e.g., Macdonald et al. (2018)).

The density function $h(x)$ corresponding to H is given by

$$h(x) = \boldsymbol{\pi} \exp\left(\boldsymbol{T}\int_0^x \mu_f(s)ds\right)\boldsymbol{t}\mu_f(x) = e^{\beta x}\boldsymbol{\pi} e^{\boldsymbol{T}\left(e^{\beta x}-1\right)/\beta}\boldsymbol{t}\,, \tag{2.7}$$

with corresponding hazard function

$$\mu_h(x) = \mu_f(x) \frac{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(s) ds \right) \boldsymbol{t}}{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(s) ds \right) \boldsymbol{e}}. \tag{2.8}$$

Remark 2.1. One way to justify this choice of model is to argue that $\overline{H}(x)$ allows to improve a base model $\overline{F}(x)$ in a parsimonious way. Indeed, since IPH distributions are dense in the class of distributions on the positive real line (in the sense of weak convergence, cf. Albrecher and Bladt (2019)), one can obtain an arbitrarily close approximation to a given lifetime distribution by adding sufficiently many phases to the state space, and in contrast to the homogeneous case, this can typically be achieved with very few states when the base model is already a good first approximation. In addition, this modeling approach also gives rise to the additional interpretation that life (aging) goes through various phases, prone to an overall time-inhomogeneous variation represented by $\mu_f(t)$.

Let $(J_t)_{t\geq 0}$ be the underlying Markov jump process with intensity matrix (2.6). Its transition probabilities are given by

$$p_{kl}(s,t) = \mathbb{P}(J_t = l | J_s = k) = \left[e^{\boldsymbol{T} \int_s^t \mu_f(u) du} \right]_{kl},$$

for $k, l = 1, \ldots, p$ being transient (non-absorbing) states. Now let V_l denote the time spent in state l during a person's life prior to death and let

$$u_{kl} = \mathbb{E}_k(V_l) = \mathbb{E}(V_l | J_0 = k), \quad k, l = 1, \ldots, p,$$

with $\boldsymbol{U} = \{u_{kl}\}_{k,l=1,\ldots,p}$ denoting the corresponding matrix, which is referred to as the Green matrix. Then with $\tau \sim H$,

$$\begin{aligned} u_{kl} &= \mathbb{E}_k \left(\int_0^\tau 1\{J_s = l\} ds \right) \\ &= \int_0^\infty \mathbb{P}_k(J_s = l, \tau \geq s) ds \\ &= \int_0^\infty \left[e^{\boldsymbol{T} \int_0^s \mu_f(u) du} \right]_{kl} ds, \end{aligned}$$

and

$$\boldsymbol{U} = \int_0^\infty e^{\boldsymbol{T} \int_0^s \mu_f(u) du} ds. \tag{2.9}$$

Again, for the one-dimensional case with $\boldsymbol{T} = -1$, we have that

$$\boldsymbol{U} = \int_0^\infty e^{-\int_0^s \mu_f(u) du} ds = \int_0^\infty \overline{F}(s) ds,$$

the mean of F. If $\mu_f(s) = 1$, then $\boldsymbol{U} = (-\boldsymbol{T})^{-1}$.

For a Gompertz hazard rate function $\mu_f(s)$, define a function g in terms of its inverse by

$$g^{-1}(s) = \int_0^s \mu_f(u) du = \frac{1}{\beta} \left(e^{\beta s} - 1 \right), \tag{2.10}$$

or, equivalently,

$$g(s) = \frac{\log(\beta s + 1)}{\beta}.$$

Note that $g^{-1}(s) \to +\infty$ as $s \to +\infty$. Then for a $\eta > 0$, using both substitution and partial integration, we get

$$\begin{aligned} \int_0^\infty e^{-\eta \int_0^s \mu_f(u) du} ds &= \int_0^\infty e^{-\eta g^{-1}(s)} ds \\ &= \int_0^\infty e^{-\eta s} g'(s) ds \end{aligned}$$

$$= \eta \int_0^\infty e^{-\eta s} g(s) ds$$
$$= \eta L_g(\eta) = \eta \frac{e^{\frac{\eta}{\beta}} \mathrm{Ei}_1\left(\frac{\eta}{\beta}\right)}{\beta \eta},$$

where L_g denotes the Laplace transform of g, and where

$$\mathrm{Ei}_1(z) = \int_1^\infty u^{-1} e^{-uz} du = \int_z^\infty u^{-1} e^{-u} du.$$

Correspondingly, (2.9) reduces to

$$\boldsymbol{U} = \int_0^\infty e^{\boldsymbol{T} \int_0^s \mu_f(u) du} ds = -\boldsymbol{T} L_g(-\boldsymbol{T}) = L_g(-\boldsymbol{T})(-\boldsymbol{T}), \tag{2.11}$$

which can then be written as

$$\boldsymbol{U} = -\boldsymbol{T} \beta^{-1} (-\boldsymbol{T})^{-1} e^{-\beta^{-1} \boldsymbol{T}} \mathrm{Ei}_1\left(-\beta^{-1} \boldsymbol{T}\right)$$
$$= \beta^{-1} e^{-\beta^{-1} \boldsymbol{T}} \mathrm{Ei}_1\left(-\beta^{-1} \boldsymbol{T}\right),$$

since $-\boldsymbol{T}$ and $L_g(-\boldsymbol{T})$ commute. More generally, $\boldsymbol{U}$ can be expressed in an explicit form whenever L_g can (the matrix function $L_g(-\boldsymbol{T})$ may be computed in different ways in practice, cf. Bladt and Nielsen (2017)). We then have the following consequence of the above formulas:

Lemma 2.1. *For $X \sim H$ we have*

$$\mathbb{E}(X) = \boldsymbol{\pi} L_g(-\boldsymbol{T}) \boldsymbol{t} = \beta^{-1} \boldsymbol{\pi} e^{-\beta^{-1} \boldsymbol{T}} Ei_1\left(-\beta^{-1} \boldsymbol{T}\right) \boldsymbol{e}.$$

Proof. The expected value of the distribution H can be obtained as the weighted sum of conditional expected times in the different states, where the weights correspond to the initial probabilities. Hence

$$\mathbb{E}(X) = \boldsymbol{\pi} \boldsymbol{U} \boldsymbol{e} = \boldsymbol{\pi} L_g(-\boldsymbol{T})(-\boldsymbol{T}\boldsymbol{e}) = \boldsymbol{\pi} L_g(-\boldsymbol{T}) \boldsymbol{t},$$

and we may use the other expression for $\boldsymbol{U}$ to complete the proof. □

Note that for $X \sim H$, residual lifetimes given survival up to time x, say, are readily obtained by the survival function

$$\mathbb{P}(X > x + y | X > x) = \frac{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(u) du\right)}{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(u) du\right) \boldsymbol{e}} \exp\left(\boldsymbol{T} \int_x^{x+y} \mu_f(u) du\right) \boldsymbol{e}.$$

Hence the residual lifetime distribution given survival until time x is again IPH with initial vector

$$\boldsymbol{\alpha}^x = \frac{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(u) du\right)}{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(u) du\right) \boldsymbol{e}},$$

intensity matrix $\boldsymbol{T}$, and hazard function $\mu_f^x(u) = \mu_f(x+u)$.

We now establish a relationship between the intensity function of an IPH distribution and its hazard function. Note that the intensity function λ of a IPH$(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$ distribution is not the same as its hazard function, μ. The latter is given by

$$\mu(t) = \lambda(t) \frac{\boldsymbol{\pi} \exp\left(\int_0^t \lambda(s) ds\, \boldsymbol{T}\right) \boldsymbol{t}}{\boldsymbol{\pi} \exp\left(\int_0^t \lambda(s) ds\, \boldsymbol{T}\right) \boldsymbol{e}}.$$

This distinction is subtle, but leads to increased flexibility in the interplay of hazards between subgroups, as we will discuss in Section 3. The cumulative hazard function of an IPH$(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$ distribution reads

$$M(t) = \int_0^t \mu(s) ds = -\log\left(\boldsymbol{\pi} \exp\left(\int_0^t \lambda(s) ds\, \boldsymbol{T}\right) \boldsymbol{e}\right).$$

The next result shows that λ is asymptotically equivalent to the hazard function in the upper tail.

Theorem 2.2. *Let $Y \sim IPH(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$. Then the hazard function μ and cumulative hazard function M of Y satisfy, respectively,*

$$\mu(t) \sim c\lambda(t), \quad t \to \infty,$$
$$M(t) \sim kg^{-1}(t), \quad t \to \infty,$$

where $g^{-1}(t) = \int_0^t \lambda(s)ds$, and c, k are positive constants.

Proof. We have that if $Z \sim \text{PH}(\boldsymbol{\pi}, \boldsymbol{T})$, then its density f_Z has a representation of the same form as (2.4) (cf. (Bladt and Nielsen, 2017, Sec.4.1.1) for details), which only differs on the multiplicative constants. Thus,

$$\overline{F}_Y(y) \sim c_1[g^{-1}(y)]^{n-1} \exp\left(-\eta[g^{-1}(y)]\right),$$
$$f_Y(y) \sim c_2[g^{-1}(y)]^{n-1} \exp\left(-\eta[g^{-1}(y)]\right) \frac{d}{dy}[g^{-1}(y)],$$

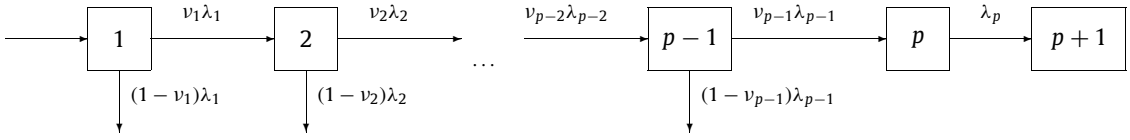
as $y \to \infty$, where $-\eta$ is the largest real eigenvalue of $\boldsymbol{T}$, n is the dimension of the Jordan block of η, and c_1, c_2 are positive constants. The first expression then follows by recalling that $\frac{d}{dy}[g^{-1}(y)] = \lambda(y)$. The second one can be shown in an analogous manner. □

2.1. Special Coxian sub-structures

The analysis becomes particularly tractable when the underlying PH distribution $(\boldsymbol{\pi}, \boldsymbol{T})$ has a simple form. For example, a Coxian PH distribution can be written in the form

$$\boldsymbol{\pi} = (1, 0, \ldots, 0), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda_1 & \nu_1\lambda_1 & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \nu_2\lambda_2 & \ldots & 0 \\ 0 & 0 & -\lambda_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_p \end{pmatrix},$$

where all λ_k, $k = 1, \ldots, p$, are distinct. The phase diagram for such a distribution is given by



That is, the states are traversed sequentially, and at each state, there is a probability to exit to the absorption state $p+1$ (i.e., die) directly. In this case

$$(z\boldsymbol{I} - \boldsymbol{T})^{-1} = \begin{pmatrix} \frac{1}{z+\lambda_1} & \frac{\nu_1\lambda_1}{(z+\lambda_1)(z+\lambda_2)} & \frac{\nu_1\lambda_1\nu_2\lambda_2}{(z+\lambda_1)(z+\lambda_2)(z+\lambda_3)} & \cdots & \frac{\nu_1\lambda_1\cdots\nu_{p-1}\lambda_{p-1}}{(z+\lambda_1)(z+\lambda_2)\cdots(z+\lambda_p)} \\ 0 & \frac{1}{z+\lambda_2} & \frac{\nu_2\lambda_2}{(z+\lambda_2)(z+\lambda_3)} & \cdots & \frac{\nu_2\lambda_2\cdots\nu_{p-1}\lambda_{p-1}}{(z+\lambda_2)(z+\lambda_3)\cdots(z+\lambda_p)} \\ 0 & 0 & \frac{1}{z+\lambda_3} & \cdots & \frac{\nu_3\lambda_3\cdots\nu_{p-1}\lambda_{p-1}}{(z+\lambda_3)(z+\lambda_4)\cdots(z+\lambda_p)} \\ \vdots & \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{z+\lambda_p} \end{pmatrix},$$

so with $\nu_p = 0$ we get the density

$$\begin{aligned} \boldsymbol{\pi}\exp(\boldsymbol{T}x)\boldsymbol{t} &= \frac{1}{2\pi i}\int_\gamma \exp(s)\boldsymbol{\pi}(s\boldsymbol{I} - x\boldsymbol{T})^{-1}\boldsymbol{t}ds \\ &= \frac{1}{2\pi i}\int_\gamma \exp(s)\left[\sum_{k=1}^{p}\frac{(\lambda_1\nu_1x)(\lambda_2\nu_2x)\cdots(\lambda_{k-1}\nu_{k-1}x)\lambda_k(1-\nu_k)}{(s+\lambda_1x)(s+\lambda_2x)\cdots(s+\lambda_kx)}\right]ds \\ &= \sum_{k=1}^{p}\left(\lambda_k(1-\nu_k)\prod_{m=1}^{k-1}\lambda_m\nu_m\right)x^{k-1}\frac{1}{2\pi i}\int_\gamma \frac{\exp(s)}{(s+x\lambda_1)\cdots(s+x\lambda_k)}ds \\ &= \sum_{k=1}^{p}\left(\lambda_k(1-\nu_k)\prod_{m=1}^{k-1}\lambda_m\nu_m\right)x^{k-j}\sum_{m=1}^{k}\frac{\exp(-\lambda_mx)}{\prod_{\substack{n=1\\n\neq m}}^{k}(-x\lambda_m + x\lambda_n)} \end{aligned}$$

$$= \sum_{k=1}^{p} \left(\lambda_k (1-\nu_k) \prod_{m=1}^{k-1} \lambda_m \nu_m \right) \sum_{m=1}^{k} \frac{\exp(-\lambda_m x)}{\prod_{\substack{n=1 \\ n\neq m}}^{k} (\lambda_n - \lambda_m)}.$$

Remark 2.2. If $\boldsymbol{\pi}$ is relaxed to be any initial probability vector (i.e., the Markov jump process can also start in other states than the first one), then one speaks of a *generalized Coxian distribution*. While for the modeling of the aging process as discussed in Section 1 the classical Coxian construction is particularly appealing as it has the direct interpretation of traversing through the p stages of aging, but with a positive probability to die 'prematurely' during one of these states, the generalized Coxian construction is also of interest. In particular, if it leads to a much better fit than the classical Cox construction, it may suggest a significant heterogeneity in the population under consideration.

3. Regression with proportional intensities

In a next step, we now introduce regression in our PH modeling framework. Regression models for life tables were cast into the scene by the introduction of the proportional hazards models, cf. Cox (1972), and these models have remained popular, both in their parametric and non-parametric forms up to today. In the logarithmic scale, the specification implies, however, parallel log-mortality curves, which are typically not observed in datasets for which a joint mortality modeling is of interest. We, therefore, look for introducing regression in a less restrictive way, while still allowing for a direct interpretation of the regression coefficients. This is achieved by letting the intensity function of our IPH distribution (rather than the hazards) vary proportionally with the covariates. As we will see later, this leads to a more flexible model, which is asymptotically equivalent to the proportional hazards construction. A physical interpretation of this construction is that each population subgroup has a natural clock running at a proportional speed relative to other subgroups.

We, therefore, propose a regression model for IPH distributions with representation $\text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda)$, where the intensity $\lambda(\cdot;\boldsymbol{\theta})$ is a non-negative parametric function depending on the vector $\boldsymbol{\theta}$ and incorporate the predictor variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ by specifying

$$\lambda(t \mid \boldsymbol{X}, \boldsymbol{\theta}) = \lambda(t;\boldsymbol{\theta}) m(\boldsymbol{X}\boldsymbol{\beta}), \quad t \geq 0. \tag{3.1}$$

Here $\boldsymbol{\beta}$ is a d-dimensional column vector and m is any positive-valued and measurable function. We call this model the *proportional intensities* (PI) model. The interpretation in terms of the underlying inhomogeneous Markov structure is that time is changed proportionally for a given subgroup when time-transforming the associated PH distribution into an IPH one. In the sequel, we will indistinctly denote by $\boldsymbol{X}$ a vector of covariates at the population level, or a matrix of covariates for finite samples. In the latter case, $\boldsymbol{X}_i$ denotes the i-th row of the matrix, corresponding to the i-th individual in the sample.

Note that the density f and survival function $\overline{F}$ of a random variable with distribution

$$\text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda(\cdot \mid \boldsymbol{X}, \boldsymbol{\theta}))$$

are given by

$$f(y) = m(\boldsymbol{X}\boldsymbol{\beta}) \lambda(y;\boldsymbol{\theta}) \boldsymbol{\pi} \exp\left(m(\boldsymbol{X}\boldsymbol{\beta}) \int_0^y \lambda(s;\boldsymbol{\theta}) ds \, \boldsymbol{T} \right) \boldsymbol{t},$$

$$\overline{F}(y) = \boldsymbol{\pi} \exp\left(m(\boldsymbol{X}\boldsymbol{\beta}) \int_0^y \lambda(s;\boldsymbol{\theta}) ds \, \boldsymbol{T} \right) \boldsymbol{e}.$$

Remark 3.1. Since the exponential function $m(x) = \exp(x)$ in the above specification is the most common choice (for instance, the one used in the original Cox proportional hazards model, Cox (1972)), the PI model draws its name from that particular example. Other choices also go under the proportional name, see for instance Royston and Parmar (2002). In terms of coefficients, the interpretation of $\boldsymbol{\beta}$ is akin to the parameters in a regular Cox regression, except that the effects here are on the intensity of a Markov process, which is only asymptotically equivalent (but not equal) to the hazard rate.

Remark 3.2. Within the above setup, it is possible to add dependence of g on $\boldsymbol{X}$ also. That is, we may consider the model

$$\lambda(t \mid \boldsymbol{X}, \boldsymbol{\gamma}) = \lambda(t;\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma})) m(\boldsymbol{X}\boldsymbol{\beta}), \quad t \geq 0, \tag{3.2}$$

where $\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma})$ is a vector-valued function, mapping the score $\boldsymbol{X}\boldsymbol{\gamma}$ to the parameter space of λ. For instance, in the one-dimensional case,

$$\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma}) = \exp(\gamma_0 + \boldsymbol{X}\boldsymbol{\gamma})$$

is a natural choice. In the sequel, any $\boldsymbol{\theta}$ specification may be of the form $\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma})$, and the notation may be used interchangeably, when there is no risk of confusion. In this specification, the role of $\boldsymbol{\theta}$ is to regulate the Gompertz time-transformation of the underlying Markov dynamics. The incorporation of covariates thus means that shortening or elongating virtual times is necessary to describe the differences observed in the data for individuals of differing characteristics.

Remark 3.3. In the proportional hazards model, one assumes that the hazard function satisfies

$$\mu(t \mid \boldsymbol{X}) = \mu(t) \exp(\boldsymbol{X}\boldsymbol{\beta}),$$

with $\mu(t)$ referred to as the *baseline hazard function* or the *hazard function for a standard subgroup* (with $\boldsymbol{X}\boldsymbol{\beta} = 0$). Common parametric models for $\mu(t)$ in the context of mortality modeling are the exponential function (referring to the Gompertz assumption) and other more refined approximations employed for the fitting of empirical mortality curves (cf. Kostaki (1992), Macdonald et al. (2018)).

One of the advantages of the PI model is that the implied hazard functions between different subgroups can deviate from proportionality in the body of the distribution (for $\dim(\boldsymbol{T}) > 1$), as can be observed in Section 5. This addresses one of the common practical drawbacks of the proportional hazards model. If $p = 1$, we have that $\mu(t) = c\lambda(t)$ for a constant $c > 0$, so that the PI specification reduces to the proportional hazards model for this case.

Another important property of the PI model is that a random variable following this specification can be expressed as a functional of a PH random variable. More specifically, consider $Y \sim \text{IPH}(\boldsymbol{\pi}, \boldsymbol{T}, \lambda(\cdot \mid \boldsymbol{X}, \boldsymbol{\gamma}))$ and let $g(\cdot \mid \boldsymbol{X}, \boldsymbol{\gamma})$ be defined in terms of its inverse function

$$g^{-1}(y \mid \boldsymbol{X}, \boldsymbol{\gamma}) = \int_0^y \lambda(s \mid \boldsymbol{X}, \boldsymbol{\gamma}) ds = m(\boldsymbol{X}\boldsymbol{\beta}) \int_0^y \lambda(s; \boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\gamma})) ds .$$

Then it is not hard to see that

$$Y \overset{d}{=} g(Z \mid \boldsymbol{X}, \boldsymbol{\gamma}),$$

for any $Z \sim \text{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. Conversely,

$$Z = g^{-1}(Y \mid \boldsymbol{X}, \boldsymbol{\gamma}) \sim \text{PH}(\boldsymbol{\pi}, \boldsymbol{T}). \tag{3.3}$$

This distributional property is a key observation for the estimation procedure proposed in Section 4.

Next, we concentrate on two important distributions which, in the presence of covariates, generalize the commonly employed models for parametric proportional hazards modeling: the exponential and the Weibull baseline hazard models.

Example 3.1 *(PH regression).* Consider $\lambda \equiv 1$, thus $\lambda(t \mid \boldsymbol{X}) = m(\boldsymbol{X}\boldsymbol{\beta})$ for all $t > 0$. The survival function takes the form

$$\overline{F}(y) = \boldsymbol{\pi} \exp\left(m(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{T} y\right)\boldsymbol{e}.$$

Note also that $Y \overset{d}{=} Z/m(\boldsymbol{X}\boldsymbol{\beta})$, where $Z \sim \text{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. In particular, this implies that

$$\mathbb{E}(Y) = \frac{\mathbb{E}(Z)}{m(\boldsymbol{X}\boldsymbol{\beta})} = \frac{\boldsymbol{\pi}(-\boldsymbol{T})^{-1}\boldsymbol{e}}{m(\boldsymbol{X}\boldsymbol{\beta})}.$$

Example 3.2 *(Matrix-Weibull regression).* Consider $\lambda(t) = \theta t^{\theta-1}$, thus $\lambda(t \mid \boldsymbol{X}) = m(\boldsymbol{X}\boldsymbol{\beta})\theta t^{\theta-1}$ for all $t > 0$. The survival function takes the form

$$\overline{F}(y) = \boldsymbol{\pi} \exp\left(m(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{T} y^{\theta}\right)\boldsymbol{e}.$$

Note also that $Y \overset{d}{=} (Z/m(\boldsymbol{X}\boldsymbol{\beta}))^{1/\theta} = m(\boldsymbol{X}\boldsymbol{\beta})^{-1/\theta} Z^{1/\theta}$, where $Z \sim \text{PH}(\boldsymbol{\pi}, \boldsymbol{T})$. The expression for the expected value is then given by

$$\mathbb{E}(Y) = m(\boldsymbol{X}\boldsymbol{\beta})^{-1/\theta}\mathbb{E}(Z^{1/\theta}) = m(\boldsymbol{X}\boldsymbol{\beta})^{-1/\theta}\Gamma(1 + 1/\theta)\boldsymbol{\pi}(-\boldsymbol{T})^{-1/\theta}\boldsymbol{e}.$$

Remark 3.4. For $p = 1$, we recover the conventional proportional hazards models with exponential and Weibull baseline hazards. Moreover, both cases have the interpretation that the expected values are proportional among subgroups. That is, for arbitrary matrix dimension, when the baseline hazard is PH or matrix-Weibull, the PI model is equivalent to the *accelerated failure time* (AFT) model (cf. Pike (1966) for the Weibull case, and more generally, the log-location scale model in Lawless (2011)). It is also not hard to see that homogeneous functions g simultaneously satisfy both models. However, for our purposes the AFT model is not of primary interest, since it regresses proportionally on the means instead of on the underlying intensity, and the physical interpretation in terms of hidden Markov models would typically be lost.

4. Parameter estimation

To estimate the hazard rate of a distribution, there are several approaches one may take. It is common in mortality modeling to consider the problem as a regression on the mortality or log-mortality curve, as a function of age, and with an appropriate loss function. However, for PH distributions, such an approach can only be used if very good (or near-optimal) initial parameters are supplied. Otherwise, local maxima and boundaries of the parameter space are almost unavoidable, and fitting times may be very long.

With that in mind, we present in this section two estimation approaches for the PI model. The first one is the target approach, which deals with the mortality curve directly and is a general procedure, here developed without covariates. It can also be seen as an alternative to the parametric mortality curve fitting in the spirit of Heligman and Pollard (1980), Kostaki (1992), now for the hazard rate of a matrix-Gompertz distribution. The second is a maximum-likelihood approach via the EM algorithm which provides a fast way of obtaining good

initial parameters for the first method[1,2] The main conceptual difference between the two approaches is that the first one gives equal weight to all ages, while the second gives weights proportional to their abundance in the dataset.

4.1. Direct mortality modeling

We define a global loss function ℓ as the aggregated age-specific losses L, which is a function of the observed and fitted mortality curves, as follows:

$$\ell(\boldsymbol{\pi}, \boldsymbol{T}, f|\boldsymbol{\mu}) = \sum_{x=0}^{N} L\left(\mu_h(x), \mu_x\right),$$

where μ_x is the observed mortality at age x and μ_h is given by (2.8). Several choices for L have been proposed, arising either from graphical considerations (for instance, Euclidean or L_p norms, absolute or relative differences) or from probabilistic considerations on the counts of the life table itself (for instance, Poisson, cf. Broström (1985), or other count distributions such as Binomial or Negative Binomial). The empirical work for the present manuscript yielded that the sturdiest choice for the PI model is given by $L(\mu, \nu) = (\log(\mu) - \log(\nu))^2$. In that case, we may apply the following decomposition:

$$\ell(\boldsymbol{\pi}, \boldsymbol{T}, f|\boldsymbol{\mu}) = \sum_{x=0}^{N} (\log(\mathcal{C}(x|\boldsymbol{\pi}, \boldsymbol{T}, f)) + \log(\mu_f(x)) - \log(\mu_x))^2,$$

with the correction factor

$$\mathcal{C}(x|\boldsymbol{\pi}, \boldsymbol{T}, f) = \frac{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(s)ds\right)\boldsymbol{t}}{\boldsymbol{\pi} \exp\left(\boldsymbol{T} \int_0^x \mu_f(s)ds\right)\boldsymbol{e}}.$$

It is noteworthy to remark that although this method will yield visually better estimation of the mortality (or log-mortality) curves, the resulting PI model will have a lower likelihood than the one arising from the EM algorithm.

4.2. Maximum likelihood estimation

We now present estimation via maximum-likelihood (ML) for the PI model. We consider the estimation procedure for the general setting where the parameters of the inhomogeneity transform are also regressed, incurring in no additional mathematical complexity.

In many applications, a large proportion of the data is not entirely observed, or censored. In this paper, we consider the case of right-censoring, since it arises naturally in the context of its applications in survival analysis[3] The main difference is that we no longer observe exact data points $Y = y$, but only $Y \in (v, \infty)$, conditionally on the corresponding covariate information. In particular, by monotonicity of g we have that $g^{-1}(Y; \boldsymbol{\theta}) \in (g^{-1}(v; \boldsymbol{\theta}), \infty)$, which can be interpreted as a censored observation of a random variable with conventional PH distribution. Some of the formulas from Olsson (1996) are useful to adapt an expectation-maximization (EM) algorithm for IPH distributions in the presence of covariates and censored data.

Let $(z_1, \delta_1), \ldots, (z_N, \delta_N)$ be a possibly right-censored i.i.d. sample from of a PH distributed random variable Z with parameters $(\boldsymbol{\pi}, \boldsymbol{T})$, where the δ_i, $i = 1, \ldots, N$ are binary indicators taking the value 1 if case of observation and 0 in case of right-censoring. Now, for each $k, l \in \{1, \ldots, p\}$, let B_k be the number of times that the underlying jump-process $(J_t)_{t\geq 0}$ initiates in state k, N_{kl} the total number of jumps from state k to l, N_k the number of times that we reach the absorbing state $p+1$ from state k and let V_k be the total time that the underlying Markov jump process spends in state k prior to absorption. These statistics are hidden, in the sense that Z alone cannot account for them. If they were not hidden, and given a sample of absorption times $\boldsymbol{z}$, the completely observed likelihood $\mathcal{L}_c$ can be written in terms of these sufficient statistics as follows:

$$\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T}; \boldsymbol{z}) = \prod_{k=1}^{p} \pi_k{}^{B_k} \prod_{k=1}^{p} \prod_{l\neq k} t_{kl}{}^{N_{kl}} e^{-t_{kl}V_k} \prod_{k=1}^{p} t_k{}^{N_k} e^{-t_k V_k}, \tag{4.1}$$

which can be brought into the canonical form of the exponential dispersion family of distributions, which possess explicit maximum likelihood estimators.

However, since the full data is not observed, we employ the EM algorithm to obtain the ML estimators. At each iteration the E-step consists in computing the conditional expectations of the sufficient statistics B_k, N_{kl}, N_k and V_k given that $Z = z$, for the fully observed

[1] However, also note that if the modeler is interested in tail probabilities instead of mortality estimates, it is much preferable to exclusively use the EM algorithm to directly estimate the density function. The alternative, that is, obtaining survival estimates through the integrated hazard, will incur model errors that propagate from younger into older ages.

[2] The initial parameters of the EM algorithm itself are of lesser concern since the routine is much faster and more robust. In all our examples, for the EM algorithm, we randomly simulated a valid PH structure (initial vector summing to one, and sub-intensity matrix having negative diagonal and rows summing to less than zero) as initial parameters, and the number of EM iterations was 2000, which was sufficient for the final changes in likelihood to be less than 0.01% in all cases.

[3] We would like to remark that for the mortality applications considered in the sequel, the EM algorithm for right-censored data will not be used in its full strength, since our data will come from 1x1 mortality tables. However, Section 6 illustrates the implementation of the right-censored algorithm for survival data.

case, and given that $Z > z$ for the right-censored case. The M-step maximizes $\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{z})$ using the estimates of the sufficient statistics from the previous step, obtaining updated parameters $(\boldsymbol{\pi}, \boldsymbol{T})$.

Below we write the formulas needed for the E- and M-steps for a sample of size N. We denote by $\boldsymbol{e}_k$ the column vector with all elements equal to zero besides the k-th entry which is equal to one, that is, the k-th element of the canonical basis of $\mathbb{R}^d$.

We now consider a sample of size N. For any two sets, A, B define their product $A \times B = \{(a, b) | a \in A, b \in B\}$, with the definition extending inductively to the case of more than two sets. Define the following product set $\mathcal{Z} = \mathcal{Z}(\boldsymbol{z}) = \prod_{i=1}^{N} A_i$, where $A_i = \{z_i\}$ if $\delta_i = 1$, and $A_i = (z_i, \infty)$ otherwise. Then it can be shown that the conditional expectations can be written as follows:

1) *E-step, conditional expectations:*

$$\mathbb{E}(B_k \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^{N} \left\{ \delta_i \frac{\pi_k \boldsymbol{e}_k^{\top} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}} + (1-\delta_i) \frac{\pi_k \boldsymbol{e}_k^{\top} \exp(\boldsymbol{T} z_i)\boldsymbol{e}}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}} \right\},$$

$$\mathbb{E}(V_k \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^{N} \left\{ \delta_i \frac{\int_0^{z_i} \boldsymbol{e}_k^{\top} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{t}\boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}} + (1-\delta_i) \frac{\int_0^{z_i} \boldsymbol{e}_k^{\top} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{e}\boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}} \right\},$$

$$\mathbb{E}(N_{kl} \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^{N} t_{kl} \left\{ \delta_i \frac{\int_0^{z_i} \boldsymbol{e}_l^{\top} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{t}\boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}} + (1-\delta_i) \frac{\int_0^{z_i} \boldsymbol{e}_l^{\top} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{e}\boldsymbol{\pi} \exp(\boldsymbol{T} u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}} \right\},$$

$$\mathbb{E}(N_k \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^{N} \delta_i t_k \frac{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}_k}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}.$$

2) *M-step, explicit maximum likelihood estimators:*

$$\hat{\pi}_k = \frac{\mathbb{E}(B_k \mid \boldsymbol{Z} \in \mathcal{Z})}{N}, \quad \hat{t}_{kl} = \frac{\mathbb{E}(N_{kl} \mid \boldsymbol{Z} \in \mathcal{Z})}{\mathbb{E}(V_k \mid \boldsymbol{Z} \in \mathcal{Z})}$$

$$\hat{t}_k = \frac{\mathbb{E}(N_k \mid \boldsymbol{Z} \in \mathcal{Z})}{\mathbb{E}(V_k \mid \boldsymbol{Z} \in \mathcal{Z})}, \quad \hat{t}_{kk} = -\sum_{s \neq k} \hat{t}_{ks} - \hat{t}_k.$$

We set

$$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_p), \quad \hat{\boldsymbol{T}} = \{\hat{t}_{kl}\}_{k,l=1,2,\ldots,p}, \quad \hat{\boldsymbol{t}} = (\hat{t}_1, \ldots, \hat{t}_p)^{\top}.$$

Note that the contributions from the conditional expectation of N_k given that $Z > z_i$ are zero, since absorption has not yet taken place.

For convenience, a summary of the entire procedure is provided in Algorithm 1.

Remark 4.1. Algorithm 1 generalizes the EM algorithm for IPH distributions in Albrecher et al. (2022b) in the following two settings: i) presence of censored observations; and ii) incorporation of covariate information.

Remark 4.2. The E-steps require the computation of matrix exponentials which can be performed in multiple ways (Moler and Van Loan, 1978). In Asmussen et al. (1996), this is done by converting the problem into a system of ODEs, which are then solved via a Runge-Kutta method of fourth-order (a C implementation, called EMpht, is available online Olsson (1998)). We employed a new implementation in Rcpp of such a method for the illustration here and everywhere else in the paper. An important consideration is that the Runge-Kutta method requires the sample to be ordered. This is relevant since the transformed data of the above algorithm, $z_i = g^{-1}(y_i \,|\, \boldsymbol{x}_i, \boldsymbol{\gamma})$ will in general not be ordered anymore, given that covariate information varies across subjects. Thus, intermediate ordering steps have to be introduced before proceeding to the numerical implementation of the EM algorithm of Olsson (1996) by the Runge-Kutta method. The re-ordering also modifies the ties in the data, which must be accounted for. Additionally, the evaluation of the likelihood in Step 2 may be done using the by-products of Step 1, which implies that re-ordering and re-weighting will also be necessary at each iteration within the optimization inside Step 2. These additional steps pose a significant computational burden relative to the non-covariate algorithm. The uniformization method (Albrecher et al., 2022b) is an alternative, but the comparison of the two methods' performance is beyond the scope of this paper.

The following assertion is a general property of any EM algorithm.

Proposition 4.1. *Employing Algorithm 1, the likelihood function increases at each iteration. Since the likelihood is bounded for fixed p, it is guaranteed to converge to a (possibly local) maximum.*

5. Examples

All examples in this section were preliminarily fitted with the `c++` wrapper functions `fit` (for the no-covariates case) and `reg` (for the PI model), available in the `matrixdist` package in `R`, cf. Bladt and Yslas (2021a,b). Subsequently, the direct mortality modeling approach with $L(\mu, \nu) = (\log(\mu) - \log(\nu))s^2$ was implemented in the same language, and optimized with the `fortran` wrapper function `optim`.

Algorithm 1 Full EM algorithm for the Proportional Intensities (PI) model.

Input: *positive data points* $\boldsymbol{y} = (y_1, y_2, \dots, y_N)^\top$, *censoring indicators* $(\delta_1, \delta_2, \dots, \delta_N)^\top$, *covariates* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, *and initial parameters* $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \boldsymbol{\gamma})$

1) *Transformation:* Transform the data into $z_i = g^{-1}(y_i \,|\, \boldsymbol{x}_i, \boldsymbol{\gamma})$, $i = 1, \dots, N$.

2) *E-step:* Define the set $\mathcal{Z} = \mathcal{Z}(\boldsymbol{z}) = \prod_{i=1}^N A_i$, where $A_i = \{z_i\}$ if $\delta_i = 1$, and $A_i = (z_i, \infty)$ otherwise. Then, compute the statistics

$$\mathbb{E}(B_k \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^N \left\{ \delta_i \frac{\pi_k \boldsymbol{e}_k^\top \exp(\boldsymbol{T} z_i)\boldsymbol{t}}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}} + (1-\delta_i) \frac{\pi_k \boldsymbol{e}_k^\top \exp(\boldsymbol{T} z_i)\boldsymbol{e}}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}} \right\},$$

$$\mathbb{E}(V_k \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^N \left\{ \delta_i \frac{\int_0^{z_i} \boldsymbol{e}_k^\top \exp(\boldsymbol{T}(z_i-u))\boldsymbol{t}\boldsymbol{\pi} \exp(\boldsymbol{T}u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}} + (1-\delta_i) \frac{\int_0^{z_i} \boldsymbol{e}_k^\top \exp(\boldsymbol{T}(z_i-u))\boldsymbol{e}\boldsymbol{\pi} \exp(\boldsymbol{T}u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}} \right\},$$

$$\mathbb{E}(N_{kl} \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^N t_{kl} \left\{ \delta_i \frac{\int_0^{z_i} \boldsymbol{e}_l^\top \exp(\boldsymbol{T}(z_i-u))\boldsymbol{t}\boldsymbol{\pi} \exp(\boldsymbol{T}u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}} + (1-\delta_i) \frac{\int_0^{z_i} \boldsymbol{e}_l^\top \exp(\boldsymbol{T}(z_i-u))\boldsymbol{e}\boldsymbol{\pi} \exp(\boldsymbol{T}u)\boldsymbol{e}_k du}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}} \right\},$$

$$\mathbb{E}(N_k \mid \boldsymbol{Z} \in \mathcal{Z}) = \sum_{i=1}^N \delta_i t_k \frac{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{e}_k}{\boldsymbol{\pi} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}.$$

3) *M-step: let*

$$\hat{\pi}_k = \frac{\mathbb{E}(B_k \mid \boldsymbol{Z} \in \mathcal{Z})}{N}, \quad \hat{t}_{kl} = \frac{\mathbb{E}(N_{kl} \mid \boldsymbol{Z} \in \mathcal{Z})}{\mathbb{E}(V_k \mid \boldsymbol{Z} \in \mathcal{Z})}$$

$$\hat{t}_k = \frac{\mathbb{E}(N_k \mid \boldsymbol{Z} \in \mathcal{Z})}{\mathbb{E}(V_k \mid \boldsymbol{Z} \in \mathcal{Z})}, \quad \hat{t}_{kk} = -\sum_{l \neq k} \hat{t}_{kl} - \hat{t}_k.$$

4) Maximize

$$\begin{aligned}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})}{\arg\max} \left\{ \sum_{i=1}^N \delta_i \log(\overline{F}_Y(y_i; \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \boldsymbol{\beta}, \boldsymbol{\gamma})) + (1-\delta_i) \log(f_Y(y_i; \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \boldsymbol{\beta}, \boldsymbol{\gamma})) \right\} \\ &= \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})}{\arg\max} \left\{ \sum_{i=1}^N \delta_i \log \left(m(\boldsymbol{x}_i\boldsymbol{\beta}) \lambda(y; \boldsymbol{\theta}(\boldsymbol{x}_i\boldsymbol{\gamma})) \hat{\boldsymbol{\pi}} \exp\left(m(\boldsymbol{x}_i\boldsymbol{\beta}) \int_0^y \lambda(s; \boldsymbol{\theta}(\boldsymbol{x}_i\boldsymbol{\gamma}))ds\, \hat{\boldsymbol{T}} \right) \hat{\boldsymbol{t}} \right) \right. \\ &\qquad \left. + (1-\delta_i) \log \left(\hat{\boldsymbol{\pi}} \exp\left(m(\boldsymbol{x}_i\boldsymbol{\beta}) \int_0^y \lambda(s; \boldsymbol{\theta}(\boldsymbol{x}_i\boldsymbol{\gamma}))ds\, \hat{\boldsymbol{T}} \right) \boldsymbol{e} \right) \right\}.\end{aligned}$$

5) Update the current parameters to $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$. Return to step 1 unless a stopping rule is satisfied.

Output: *fitted parameters of the proportional intensities model* $(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{\beta}, \boldsymbol{\gamma})$.

The preliminary fitting was done using as log-likelihood weights the implied density from the mortality rates (death to exposure ratio), and using the midpoints between ages as the observed ages (corresponding to the data $\boldsymbol{y}$). An interval-censored approach would be slightly more accurate, but is outside the scope of the current paper, since we are using the EM procedure only to identify good initial parameters. For numerical purposes, age was divided by 100 at the time of fitting, and all given parameters are on that scale. All plots are scaled back to the original size. All data was obtained from https://www.mortality.org/. In terms of running times, all analyses were completed individually within two minutes to 1.5 hours,[4] with the more complex examples being lengthier to optimize.

5.1. Danish female mortality without covariates

We first consider the modeling of Danish females since the year 2000 using a generalized Coxian structure as well as a Coxian structure. The distinction is made since the former seems to provide better fitted models, while the latter has a possibly more intuitive interpretation in terms of state transitions (as all individuals start in the first state).

The number of phases was chosen empirically. The left panel of Fig. 5.1 shows with a dashed line the $p = 1$ case, corresponding to a classical Gompertz distribution. Thereafter, increasing dimensions of $\boldsymbol{T}$ were checked, for $p \leq 20$. At some point, additional parameters do not provide increased performance,[5] and we choose the last p before this occurs. We observe a satisfactory fit, and the parameters for the generalized Coxian case are given by[6]

$$\begin{aligned}\boldsymbol{\pi} &= (0.01, 0, 0.04, 0.07, 0, 0.23, 0.06, 0.25, 0.32, 0, 0, 0, 0, 0.01, 0), \\ \mathrm{Diag}(\boldsymbol{T}) &= -(3.14, 1.81, 95.92, 25.29, 24.27, 2.98, 0.03, 0, 17.19, 0.13, 1.23 \times 10^{12}, 7.28, 3.29, 0.01), \\ \boldsymbol{t} &= (0, 0.42, 1.77, 0, 0.01, 0, 0, 0, 0, 10.16, 0.13, 492.2, 0, 0, 0.01)^\top, \\ \beta &= 10.68.\end{aligned}$$

[4] All computations were done in a standard MacBook Pro (15-inch, 2017) Laptop with a 2.8 GHz Quad-Core Intel Core i7 processor and a 16 GB 2133 MHz LPDDR3 memory.

[5] Increased performance in this context means a substantial decrease in loss function value upon convergence for each model of varying dimensions. Information criteria such as AIC and BIC are not helpful here, since the effective degrees of freedom of PH models are generally unknown (see also Albrecher et al. (2022a)). Thus we chose a relative decrease of less than 0.1% as the stopping rule.

[6] Here, and in subsequently reported parameters, zeros are only exactly zero when the structure implies it (such as a Coxian structure). Else, they represent the rounding of a small number.

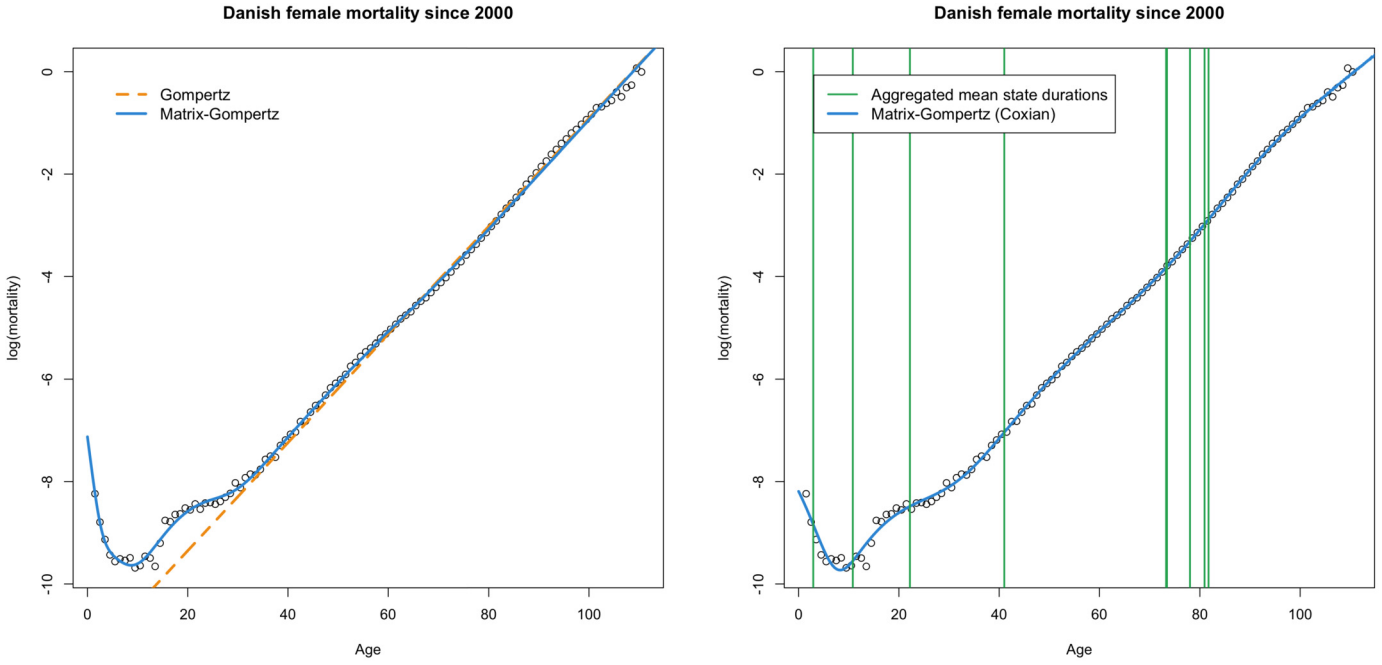


Fig. 5.1. Fitted matrix-Gompertz distributions to Danish female mortality data (generalized Coxian PH on the left, Coxian PH on the right). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Note that for the classical Gompertz fit on the left-hand side one obtains $\beta = 10.55$, but one is not able to take into account the humps at the left end at all. On the other hand, the additional flexibility of the generalized Coxian can accomplish that, maintaining a similar value $\beta = 10.68$ for the transformation.

While the choice of the generalized Coxian is motivated by the tractability of the corresponding estimation procedure (it is much more tractable than fitting a general PH distribution) within the purpose of obtaining a good statistical fit, the restriction to a classical Coxian distribution (starting in state 1 with probability 1) opens up the interpretability in terms of stages of aging as discussed in Section 1. The right panel of Fig. 5.1 depicts the resulting Coxian fit to the same dataset, together with the cumulative aggregated means of the sojourn times for each state of the underlying state space (here 10 states). In the Coxian construction, these are traversed sequentially, and thus the last vertical line is equal to the life expectancy. The parameters and line locations are given by

$$\begin{aligned}
\boldsymbol{\pi} &= \boldsymbol{e}_1,\\
\mathrm{Diag}(\boldsymbol{T}) &= -(27.72, 5.95, 1.84, 0.29, 0.02, 2.32, 2.88 \times 10^{11}, 0.04, 0.03, 0.04),\\
\boldsymbol{t} &= (0.03, 0, 0.01, 0, 0.01, 0, 189.94, 0.01, 0.01, 0.04)^{\top},\\
\mathrm{lines} &= (2.92, 10.81, 22.20, 41.02, 73.30, 73.45, 73.45, 78.02, 80.93, 81.73),\\
\beta &= 7.85.
\end{aligned}$$

One can see that the wavy behavior of the log-mortality curve is also captured quite well with a pure Coxian PH distribution. Recall that the Coxian construction allows for the interpretation that life traverses through different states until death (absorption), staying at each one for a random duration, and with many individuals going through all the states, whereas some individuals die (go to absorption) from an earlier state directly. From the above figures, one readily computes the probabilities of dying while being in state i $(i = 1, \ldots, 10)$:

$$(0.001, 0, 0.003, 0.006, 0.326, 0.001, 0, 0.217, 0.428, 1).$$

The probability of reaching state $p = 10$ before death (i.e., going through all states) here is 0.299. Observe that some states have a very short duration, which can serve as an artifact to generate possible deaths at these concrete points in time. Note also that the Coxian matrix-Gompertz leads to a nice fit, but 'needs' a different value $\beta = 7.85$ for the transformation than the classical Gompertz distribution in order to fit the humps at the left end. The generalized Coxian matrix-Gompertz then is flexible enough again to return to a similar β-value as the classical Gompertz, which is determined by the right end. From the concrete resulting parameters, one sees that there is a high probability to start in state 5, 7 or 8 only, so that the resulting model can be interpreted in terms of more flexible mixtures of Coxian distributions with lower state space.

5.2. Female mortality with country as a covariate

Let us now move on to apply our approach to multi-population fitting with regression. As a first illustration of the PI model, we consider the mortality curves from two countries with substantially different life expectancies, the United States of America and Japan. Thus, the covariate is a binary one: USA $= 1$ and Japan $= 0$. We study the female population in the time interval $2000 - 2010$ (the observations are aggregated). The same approach for selecting the order of the fitted model as above was employed for this dataset.

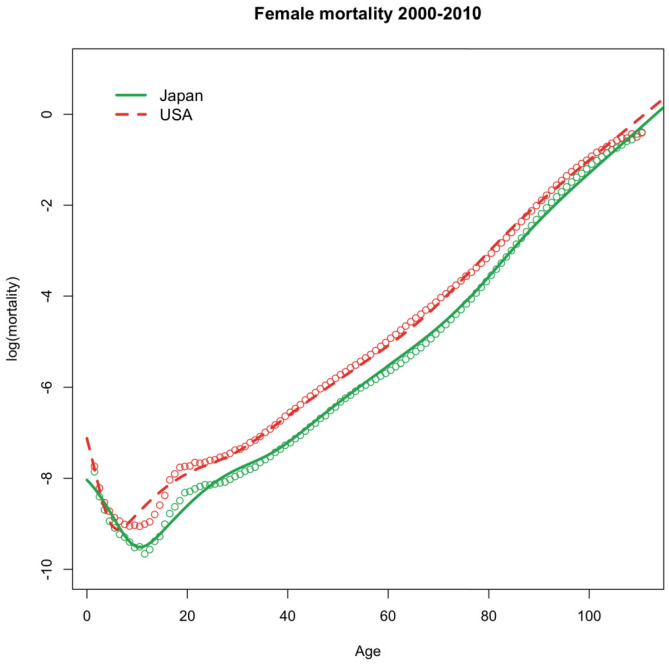


Fig. 5.2. PI model applied to country as a covariate.

Fig. 5.2 shows that although there is a slight misfit at the teenage years, the overall shape of both fitted curves resemble closely the observed mortality. Here we have $p = 12$ states and the parameters are given by

$$\begin{aligned}
\boldsymbol{\pi} &= (0.07, 0.03, 0, 0, 0, 0.15, 0, 0.05, 0.24, 0.45, 0, 0),\\
\mathrm{Diag}(\boldsymbol{T}) &= -(0.02, 0.06, 22.37, 8.78, 0.20, 1.37, 1.22, 44.17, 404.86, 0, 31.44, 0),\\
\boldsymbol{t} &= (0, 0, 13.89, 0.25, 0, 0, 0.02, 0, 0, 0, 0.35, 0)^{\top},\\
\boldsymbol{\beta} &= (0.91),\\
\boldsymbol{\theta} &= (2.28, -0.07).
\end{aligned}$$

The positive value of $\boldsymbol{\beta}$ indicates an increased underlying intensity (and thus shorter sojourn times) for the USA population, which is multiplicative with factor $\exp(0.91) \approx 2.48$. On the other hand, the parameters of $\boldsymbol{\theta}$ indicate that there is a baseline Gompertz factor of size $\exp(2.28) \approx 9.78$, and the USA population has a multiplicative decrease of this factor by $\exp(-0.07)$, or roughly 93%, incurring in an overall Gompertz factor of size 9.16. The latter means in terms of the PI model that there is an "environment" lifetime generating Gompertz process which is unfavorable for the Japanese female population, but their underlying slower traversing through the multi-state process (of their lives) makes up for it, to the degree that can be observed in Fig. 5.2.

Theoretically, by Theorem 2.2, both curves are asymptotically parallel, and although we see a substantial difference at most ages, the gap closes and is quite small at very advanced ages. This observation can be made from the data itself, implying that old-age mortality in both countries seems to be quite similar. However, the takeaway from the hidden Markov point of view is that the flexibility gained from considering intensities instead of hazard rates is advantageous for modeling more delicate inter-curve features in regression analyses.

5.3. Danish female mortality with time as a covariate

As a second example of the PI model, let us consider the Danish female population during the interval $1950 - 2000$, and examine the effect of time when considering it as a numerical covariate. For numerical convenience, we actually use the transformed covariate $t^* = (\text{year} - 1950)/1000$. The same approach as in the two previous examples is used to select the size p of the underlying state space.

The proportional hazards model is not well suited for this task (the resulting log-mortality curves for different time periods would be parallel!), and so we use the Lee-Carter (LC, cf. Lee and Carter (1992)) model as a benchmark. The LC model specifies the log-mortality by

$$\log(\mu_{x,t}) = a_x + b_x k_t + \epsilon_{x,t},$$

where $\epsilon_{x,t}$ are Gaussian random variables. The a_x term is estimated as the average log-mortality over time at each age x, and then b_x and k_t are computed from a singular value decomposition of $\log(\mu_{x,t}) - a_x$. The procedure is non-parametric and thus can capture very minute changes in individual mortality curves across different ages.

On the other hand, by Theorem 2.2 the PI model asymptotically satisfies, as x grows,

$$\log(\mu_{x,t}) \to a + b_{x,t} + k_t,$$

where $a = \log(c)$ is a constant which depends on the parameters $\boldsymbol{\pi}$ and $\boldsymbol{T}$, $b_{x,t} = \log(\lambda(x; \exp(\theta_0 + \theta_1 t^*))) = \exp(\theta_0 + \theta_1 t^*)\log(x)$ (in our case) is the logarithm of the time-dependent Gompertz intensity, and $k_t = \beta_1 t^*$ (also in our case) is the logarithm of the proportionality

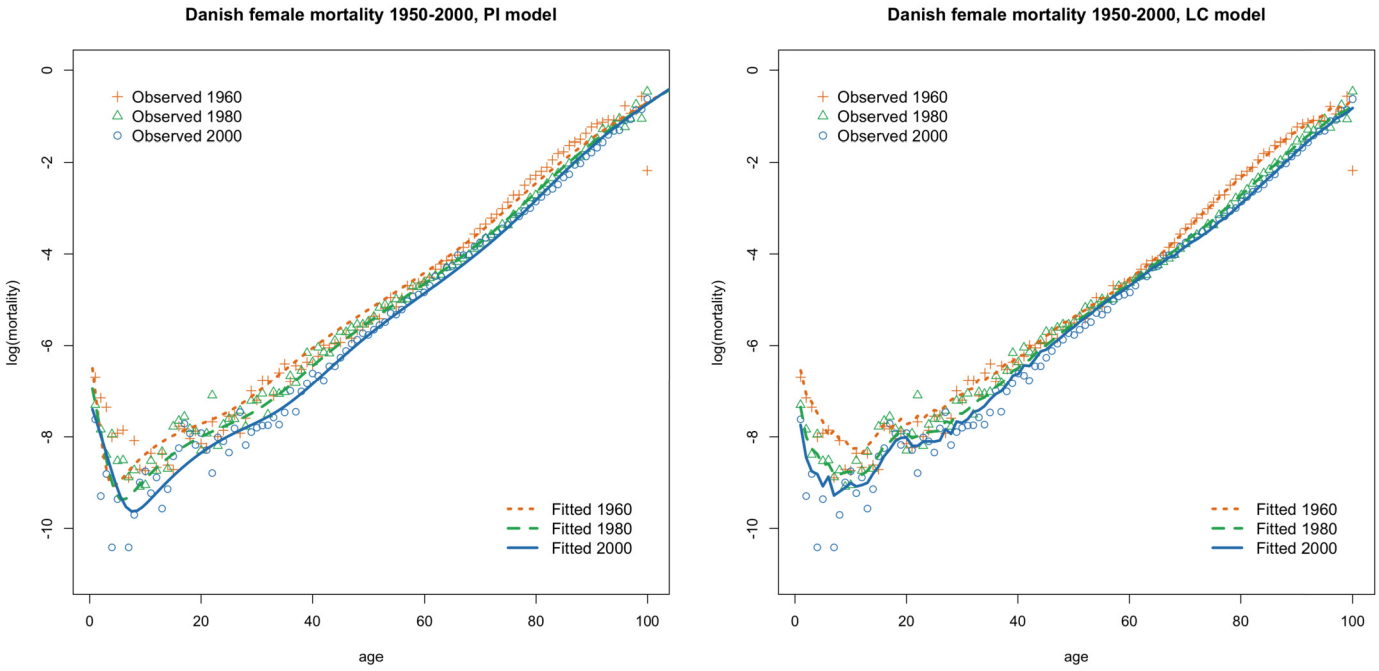


Fig. 5.3. PI model using time as a covariate, plotted for the years 1960, 1980 and 2000, and for the Danish females dataset.

factor between the underlying intensities. For smaller x (before convergence to the limit) the log-mortality does not have such a simple decomposition, and is instead given by

$$\log(\mu_{x,t}) = b_{x,t} + \log\left(\boldsymbol{\pi}\exp\left(\int_0^x \exp(b_{s,t}+k_t)ds\,\boldsymbol{T}\right)\boldsymbol{t}\right) - \log\left(\boldsymbol{\pi}\exp\left(\int_0^x \exp(b_{s,t}+k_t)ds\,\boldsymbol{T}\right)\boldsymbol{e}\right). \tag{5.1}$$

Fig. 5.3 shows the resulting fitted PI and LC models for the selected years 1960, 1980, and 2000, where in both cases we observe a downward trend in mortality every 20 years. Both models have room for improvement, but overall show accurate descriptions. The PI model seems to not capture a surge in mortality for old ages in 1960, while the LC model does not seem to capture the decreases of mortality from 1980 to 2000 for young ages. The non-parametric nature of the LC model is sturdier to outliers as is the case for an old-age point of 1960, while the PI seeks to compromise large parts of the curve for this one observation. However, observe that the LC model has some wiggling occurring for ages 0 to 40, which may indicate a slight degree of overfitting, a known issue for smaller populations and one which the PI model seems to handle better in this case.

The parameters of the PI model are as follows:

$$\begin{aligned}
\boldsymbol{\pi} &= (0.31, 0.15, 0.17, 0.11, 0.12, 0.01, 0, 0.01, 0.04, 0.03, 0.04, 0),\\
\mathrm{Diag}(\boldsymbol{T}) &= -(36.47, 0.09, 1.3, 0.06, 9.49, 5.34, 0.19, 2.25, 139.68, 0.22, 0.89, 0.66),\\
\boldsymbol{t} &= (0, 0, 0, 0, 0, 0.17, 0, 0, 4.16, 0, 0, 0.66)^\top,\\
\boldsymbol{\beta} &= (-22.54),\\
\boldsymbol{\theta} &= (1.92, 2.94).
\end{aligned}$$

In contrast to the previous illustration, we have that the negative value of $\boldsymbol{\beta}$ indicates a decreased underlying intensity (and thus longer sojourn times) as years progress, which is multiplicative with factor $\exp(-22.54\cdot t^*)$. In turn, the parameters of $\boldsymbol{\theta}$ imply a baseline Gompertz factor given by $\exp(1.92)\approx 6.82$ at $t^*=0$ (which amounts to the year 1950), thereafter it increases with the factor $\exp(2.94\cdot t^*)$, which itself increases the mortality as time progresses. Fig. 5.3 shows that the overall tradeoff from increased "environmental" mortality and decreased state-traversing speed leads to a net decrease in mortality through the years at young ages, but virtually no changes at older ages.

In order to better understand the issue of potential overfitting of the LC model for smaller populations (like the case of Denmark above), we did an analogous analysis for USA females, and the resulting parameters of the PI model in this case are

$$\begin{aligned}
\boldsymbol{\pi} &= (0.3, 0.15, 0.13, 0.13, 0.15, 0.01, 0, 0.01, 0.05, 0.03, 0.04, 0),\\
\mathrm{Diag}(\boldsymbol{T}) &= -(1.15, 0.11, 0.65, 0.05, 7.82, 4.02, 0.19, 2.81, 56.14, 0.2, 0.77, 0.70),\\
\boldsymbol{t} &= (0, 0, 0, 0, 0, 0.22, 0, 0, 2.72, 0, 0, 0.7)^\top,\\
\boldsymbol{\beta} &= (-17.98),\\
\boldsymbol{\theta} &= (1.93, 1.74).
\end{aligned}$$

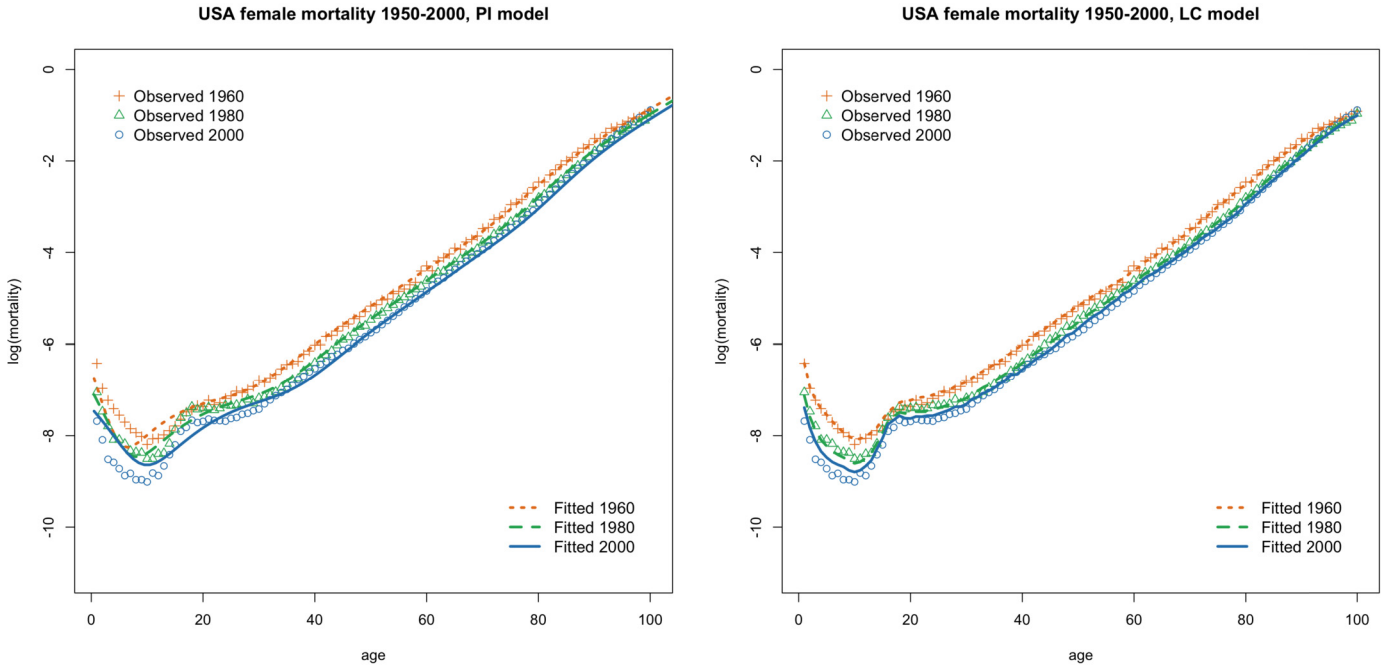


Fig. 5.4. PI model using time as a covariate, plotted for the years 1960, 1980 and 2000, and for the USA females dataset.

Fig. 5.4 depicts the resulting log-mortality curves for 1960, 1980 and 2000 together with the empirical observations. Indeed, here the wiggling of the LC curves is considerably improved. One sees that in this case from a purely statistical point of view the LC model may be preferable over the PI model, but one should keep in mind that the PI model offers a more direct interpretation of the resulting model, with much fewer parameters, while still providing a visually acceptable statistical fit.

Remark 5.1. Forecasting with such a model is straightforward if we are only interested in non-autoregressive effects,[7] and the incorporation of cohort effects can also be done as with time, and with possible non-linear terms. However, the predictive performance falls short of models which are specifically designed for this task, and incorporating these temporal dependencies is an interesting subject of future research.

6. General survival modeling

This section illustrates the use of inhomogeneous Markov models for survival analysis, showcasing the implementation of Algorithm 1 in the right-censored case. The first is a simulated example, while the second is a Veterans' lung cancer dataset, both of which are subject to right-censoring and include covariate information, and thus the regression methods from this paper can be applied.

6.1. Simulated data

Consider the following simple two-group setup. We study two groups of equal size, 500 each. For the first group,

$$Z_i \sim \text{IPH}\left((1/4,\, 1/2,\, 1/4),\, \begin{pmatrix} -10 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1/10 \end{pmatrix},\, \frac{3}{2}s^{1/2}\right)$$

while for the second

$$Z_i \sim \text{PH}\left((1/4,\, 1/2,\, 1/4),\, \begin{pmatrix} -20 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2/10 \end{pmatrix}\right).$$

Subsequently, we randomly censor the Z_i by 1000 exponential variables E_i of rate $1/10$ by

$$Y_i = \min\{Z_i,\, E_i\}, \quad \delta_i = 1\{Y_i = Z_i\}, \quad i = 1, \ldots, 1000,$$

leading to roughly 10% right-censored observations. Notice that in the above setting, the inhomogeneity function also changes with the covariates, such that we need to implement the fitting procedure associated with model (3.2). Additionally, for comparison, we fitted two misspecified models: a standard Weibull proportional hazards model and the scalar ($\dim(\boldsymbol{T}) = 1$) version of model (3.2). The resulting

[7] Given a desired time t (and thus t^*) and age x, Equation (5.1) provides the corresponding mortality estimate. Moreover, all the implied distributional properties can also be extracted using the fitted parameters, using Equation (2.7).

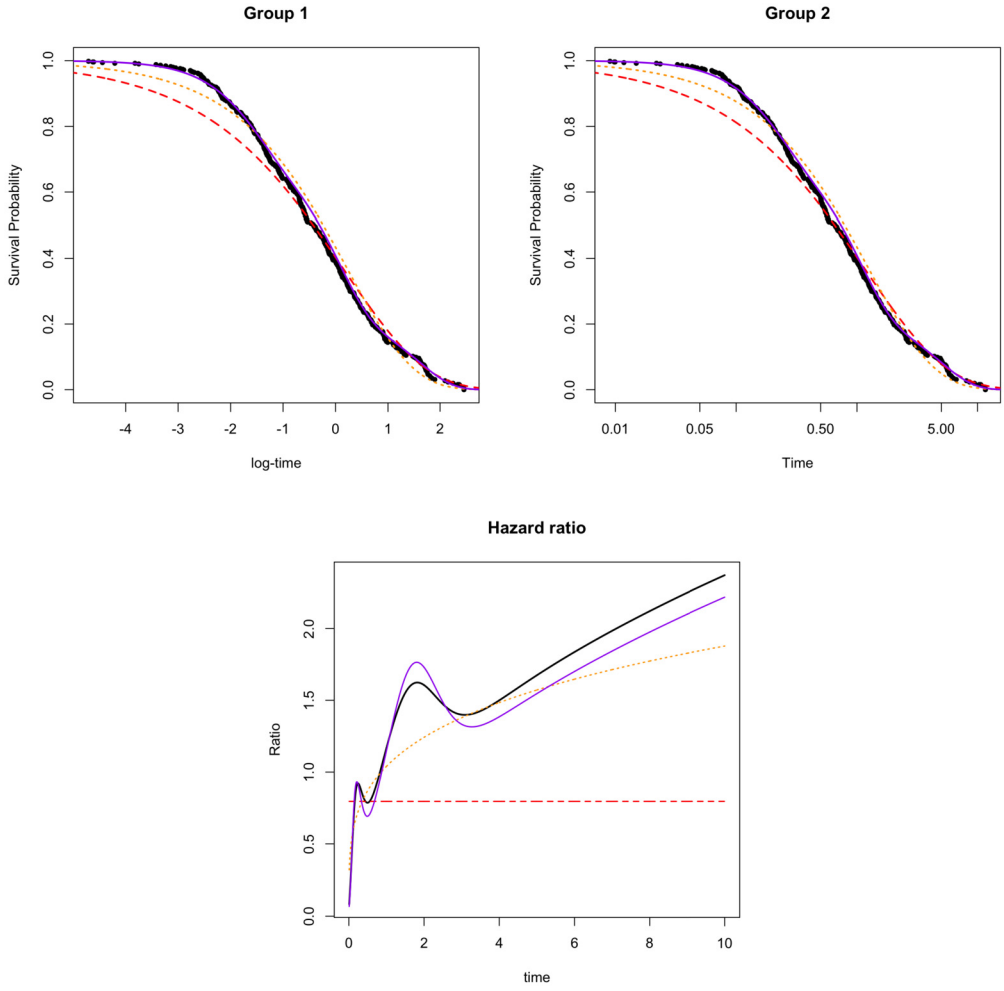


Fig. 6.1. Simulated example. Top: empirical (black, dots) versus fitted survival curves for Weibull proportional hazards model (red, dashed), scalar version of (3.2) (orange, dotted), and correctly specified (3.2) (purple, solid). Bottom: theoretical hazard function ratio between the two groups (black solid) versus the aforementioned models (same line code).

empirical versus fitted survival curves for each group, as well as the hazard ratio between the two groups, are given in Fig. 6.1. We observe that only the correctly specified model recovers the oscillating shape of the theoretical hazard ratio, and within each group, the survival curve is better estimated in the tails of the distribution. Additionally, the goodness of fit can be evaluated by considering a generalization of the Cox-Snell residuals given by

$$r_i = -\log\left(\boldsymbol{\pi}\exp\left(m(\boldsymbol{x}_i\boldsymbol{\beta})\int_0^{y_i}\lambda(s;\boldsymbol{\theta}(\boldsymbol{x}_i\boldsymbol{\gamma}))ds\,\boldsymbol{T}\right)\boldsymbol{e}\right), \quad i=1,\dots,100,$$

which under the null hypothesis of correct specification of the model, should constitute a right-censored unit-exponential dataset. Fig. 6.2 depicts the Kaplan-Meier and Nelson-Aalen non-parametric estimators of the survival curve of the residuals r_i to assess standard exponential behavior visually. We observe that only the correctly specified model provides an adequate fit in the sense of staying within the confidence bounds implied by Greenwood's formula for the first approach, and by aligning itself with the identity for the second one.

6.2. Veterans data

We consider a well-known dataset from the Veterans' Administration Lung Cancer study (publicly available in the R package *survival*).

The data consist of 137 observations and 12 variables coming from a randomized trial where two groups are given different treatment regimens for lung cancer. The variables of interest for our analysis are survival time, censoring status (binary), treatment (binary), prior therapy (binary), and Karnofsky performance score (from 0 to 100). Previous studies (cf. for instance, the vignette https://cran.rstudio.com/web/packages/survival/vignettes/timedep.pdf, pp. 15-22) have shown that only the Karnofsky score violates the proportional hazards assumption. Common ways to address this issue are stratifying the dataset into time groups or considering continuous time-dependent regression coefficients. However, the coefficient for treatment becomes significant when improving the model's fit. Consequently, we presently focus only on the goodness of fit of the model introduced in the text. The latter is measured by its log-likelihood and residuals.

For comparison, we consider a classical model: a) Weibull proportional hazards model. Given the modest size of the dataset, we avoid large matrices $\boldsymbol{T}$ for the IPH-based model. Thus, we present a model based on $\dim(\boldsymbol{T}) = 2$ with a Coxian structure: b) Matrix-Weibull PI model.

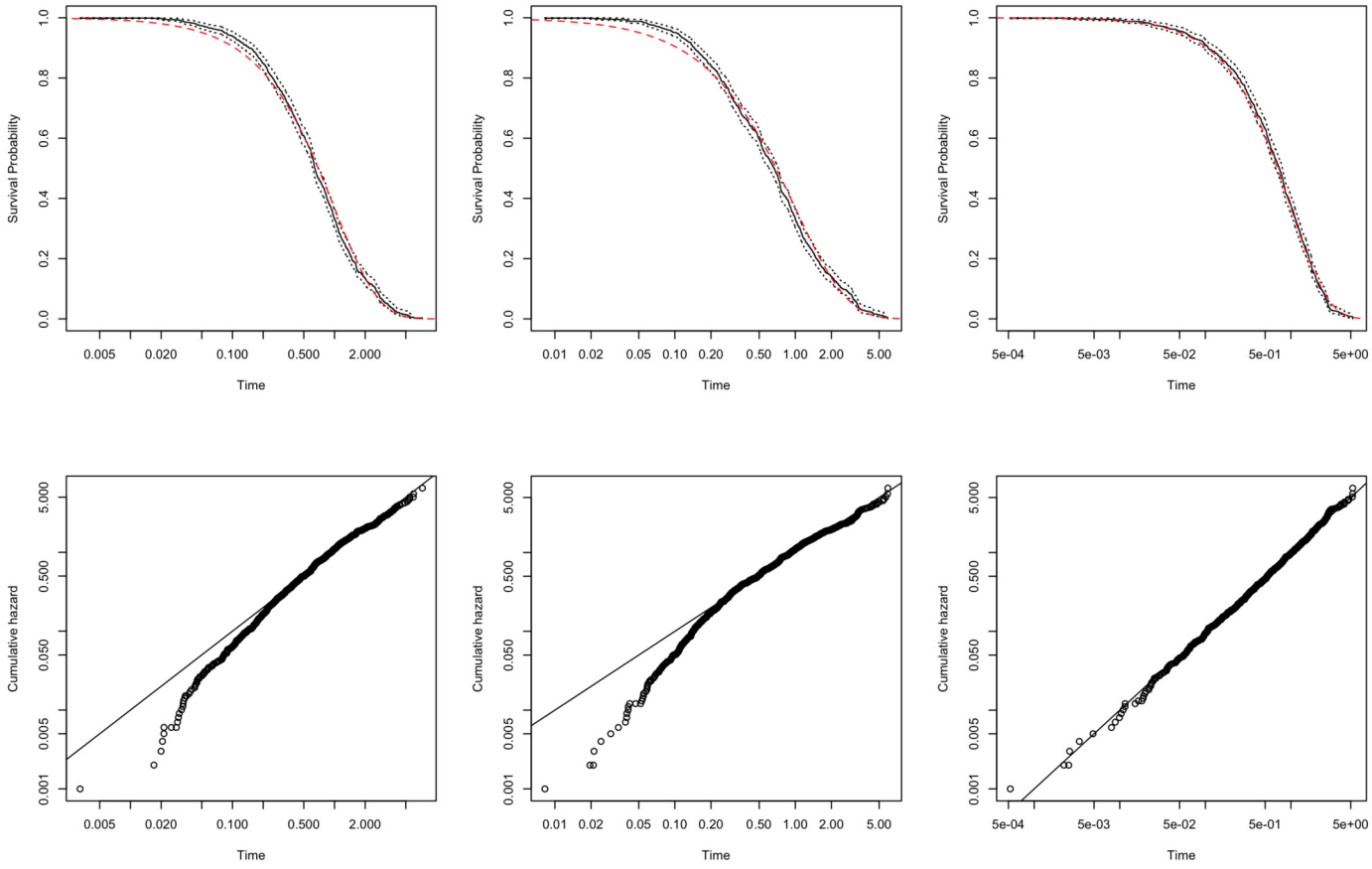


Fig. 6.2. Simulated example goodness of fit based on the three models from Fig. 6.1. Top: survival curves based on the Kaplan-Meier estimator of the residuals r_i. Bottom: plot between r_i and $-\log(\widehat{\widehat{F}}(r_i))$ based on the Nelson-Aalen estimator.

The results are summarized in the table below and in Fig. 6.3.

Model	Log-likelihood	Nb. Parameters	AIC	BIC
a	-136.21	5	282.42	297.02
b	-127.74	7	269.48	289.92

We observe that model b) improves the fit to where the residuals are within bounds, and the log-likelihood, AIC and BIC reflect this.

The AIC and BIC computed here for the matrix-distributions works well for comparison purposes, given the low dimensions of the distributions. However, in general, one should proceed with caution (especially in high dimensions) due to the well-known identifiability issue for PH distributions, which carries over to IPH survival models. Namely, different dimensions and parameter configurations may lead to very similar density shapes.

7. Conclusion

In this paper we investigated the potential of matrix-Gompertz distributions as special cases of inhomogeneous phase-type distributions for the modeling of mortality across the entire lifespan as a parsimonious alternative to popular modeling approaches. The goal is not to replace the latter but rather to see how close in performance one can get with this type of models, as they also allow a causal interpretation in terms of biological aging, with lifetime traversing through different stages and a possible early death in each of them. In that latter context, the introduction of the time-transform inherent in the construction of inhomogeneous phase-type distributions allowed to reduce the necessary dimension for adequate modeling reported in previous studies from 200 - 250 to 10 - 12. In addition, we studied a regression approach that can be used for multi-population mortality modeling based on proportional intensities. As the first contribution on regressing IPH distributions, this part may also be of independent interest. We developed an estimation procedure and illustrated the models for female mortality data from Denmark, Japan and the USA.

Since matrix-Gompertz distributions are dense in the class of all lifetime distributions, from a general perspective we have illustrated that misspecified survival models can sometimes be just one matrix away from providing a good fit. The practical implementation of the methods in this paper relies on the EM algorithm, which for large matrices and datasets can become significantly slow. It is our experience that the use of an inhomogeneous intensity function in the underlying stochastic jump process can alleviate the need for a large number of phases, thus making estimation feasible.

Although we have used the midpoint approach, which is standard and works well for 1×1 resolution life tables, an interval-censored estimation procedure is statistically more accurate and has potential to be applied to data at lower resolutions, which will be the subject

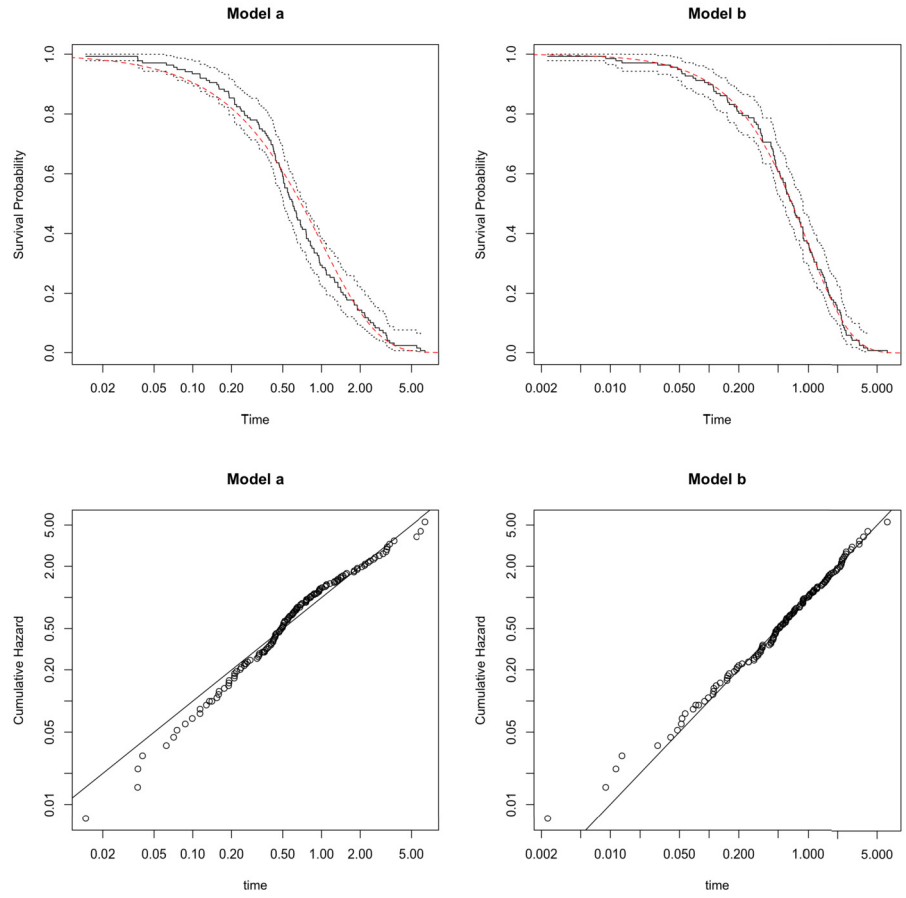


Fig. 6.3. Veterans dataset. Goodness of fit diagnostics based on the Kaplan-Meier estimator of the residuals (top row), and on the residuals versus Nelson-Aalen implied cumulative hazard plot (bottom row). The corresponding models are: a) Weibull proportional hazards model; b) Matrix-Weibull PI model.

of future research. In the present paper we considered regression according to proportional intensities, as a first and workable approach for IPH regression. We also showcased how the right-censored version of our algorithm can be used in survival settings. In general, for multi-population mortality modeling it is of course restrictive to assume that the operational time for each population is scaled with a constant across the entire lifetime. Another promising direction is hence the regression of individual rates of the intensity matrix, as this would allow for interesting conclusions in terms of the aging interpretation underlying the IPH models. However, if applied to all possible parameters, the model will be highly overparametrized, and hence some constraints will have to be imposed, challenging the corresponding estimation methods. It will be an interesting subject of future research to identify workable compromises in this direction, which should increase the versatility of the resulting models.

Declaration of competing interest

The authors declare that there is no competing interest.

Acknowledgement

The authors would like to thank Johannes Thuswaldner for conscientiously implementing some tests for the fitting procedures at an early stage of the project, and Peter Hieber for interesting discussions on the topic. HA and MaB would like to acknowledge financial support from the Swiss National Science Foundation Project 200021_191984. JY would like to acknowledge financial support from the Swiss National Science Foundation Project IZHRZ0_180549.

References

Aalen, O.O., 1995. Phase type distributions in survival analysis. Scandinavian Journal of Statistics 22 (4), 447–463.
Albrecher, H., Bladt, M., 2019. Inhomogeneous phase-type distributions and heavy tails. Journal of Applied Probability 56 (4), 1044–1064.
Albrecher, H., Bladt, M., Müller, A.J., 2022a. Penalised likelihood methods for phase-type dimension selection. Preprint.
Albrecher, H., Bladt, M., Yslas, J., 2022b. Fitting inhomogeneous phase-type distributions to data: the univariate and the multivariate case. Scandinavian Journal of Statistics 49 (1), 44–77.
Albrecher, H., Embrechts, P., Filipović, D., Harrison, G.W., Koch, P., Loisel, S., Vanini, P., Wagner, J., 2016. Old-age provision: past, present, future. European Actuarial Journal 6 (2), 287–306.
Antonio, K., Devriendt, S., de Boer, W., de Vries, R., De Waegenaere, A., Kan, H.-K., Kromme, E., Ouburg, W., Schulteis, T., Slagter, E., Van der Winden, M., Van Iersel, C., Vellekoop, M., 2017. Producing the Dutch and Belgian mortality projections: a stochastic multi-population standard. European Actuarial Journal 7 (2), 297–336.
Asmussen, S., Laub, P.J., Yang, H., 2019. Phase-type models in life insurance: fitting and valuation of equity-linked benefits. Risks 7 (1), 17.

Asmussen, S., Nerman, O., Olsson, M., 1996. Fitting phase-type distributions via the EM algorithm. Scandinavian Journal of Statistics 23 (4), 419–441.
Barigou, K., Goffard, P.-O., Loisel, S., Salhi, Y., 2021. Bayesian model averaging for mortality forecasting using leave-future-out validation. ArXiv preprint. arXiv:2103.15434.
Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., Salhi, Y., 2012. Understanding, modelling and managing longevity risk: key issues and main challenges. Scandinavian Actuarial Journal 2012 (3), 203–231.
Bladt, M., 2021. Phase-type distributions for claim severity regression modeling. ASTIN Bulletin: The Journal of the IAA, 1–32.
Bladt, M., Gonzalez, A., Lauritzen, S.L., 2003. The estimation of phase-type related functionals using Markov chain Monte Carlo methods. Scandinavian Actuarial Journal 2003 (4), 280–300.
Bladt, M., Nielsen, B.F., 2017. Matrix-Exponential Distributions in Applied Probability, vol. 81. Springer.
Bladt, M., Yslas, J., 2021a. matrixdist: an R package for inhomogeneous phase-type distributions. ArXiv preprint. arXiv:2101.07987.
Bladt, M., Yslas, J., 2021b. matrixdist: Statistics for Matrix Distributions. R package version 1.1.3.
Bobbio, A., Horváth, A., Telek, M., 2005. Matching three moments with minimal acyclic phase type distributions. Stochastic Models 21 (2–3), 303–326.
Broström, G., 1985. Practical aspects on the estimation of the parameters in Coale's model for marital fertility. Demography 22 (4), 625–631.
Cheng, B., Jones, B., Liu, X., Ren, J., 2020. The mathematical mechanism of biological aging. North American Actuarial Journal 25 (1), 73–93.
Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society, Series B, Methodological 34 (2), 187–202.
Denuit, M., Trufin, J., 2016. From regulatory life tables to stochastic mortality projections: the exponential decline model. Insurance. Mathematics & Economics 71, 295–303.
Dickson, D., Hardy, M., Waters, H., 2019. Actuarial Mathematics for Life Contingent Risks. Cambridge University Press.
Dowd, K., Cairns, A.J., Blake, D., 2020. CBDX: a workhorse mortality model from the Cairns–Blake–Dowd family. Annals of Actuarial Science 14 (2), 445–460.
Gavrilov, L.A., Gavrilova, N.S., 1991. The Biology of Life Span: A Quantitative Approach. Harwood Academic, New York.
Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London 115, 513–583.
Guihenneuc-Jouyaux, C., Richardson, S., Longini Jr, I.M., 2000. Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline. Biometrics 56 (3), 733–741.
Gutterman, S., Vanderhoof, I.T., 1998. Forecasting changes in mortality: a search for a law of causes and effects. North American Actuarial Journal 2 (4), 135–138.
Hassan Zadeh, A., Jones, B.L., Stanford, D.A., 2014. The use of phase-type models for disability insurance calculations. Scandinavian Actuarial Journal 2014 (8), 714–728.
Heligman, L., Pollard, J.H., 1980. The age pattern of mortality. Journal of the Institute of Actuaries 107 (1), 49–80.
Horváth, A., Telek, M., 2007. Matching more than three moments with acyclic phase type distributions. Stochastic Models 23 (2), 167–194.
Hsieh, F., Lavori, P.W., 2000. Sample-size calculations for the Cox proportional hazards regression model with non-binary covariates. Controlled Clinical Trials 21 (6), 552–560.
Kay, R., 1986. A Markov model for analysing cancer markers and disease states in survival studies. Biometrics, 855–865.
Kostaki, A., 1992. A nine-parameter version of the Heligman-Pollard formula. Mathematical Population Studies 3 (4), 277–288.
Lawless, J.F., 2011. Statistical Models and Methods for Lifetime Data, vol. 362. John Wiley & Sons.
Lee, R.D., Carter, L.R., 1992. Modeling and forecasting US mortality. Journal of the American Statistical Association 87 (419), 659–671.
Li, J.S.-H., Hardy, M.R., Tan, K.S., 2009. Uncertainty in mortality forecasting: an extension to the classical Lee-Carter approach. ASTIN Bulletin: The Journal of the IAA 39 (1), 137–164.
Lin, T., Wang, C.-W., Tsai, C.C.-L., 2021. Correlated age-specific mortality model: an application to annuity portfolio management. European Actuarial Journal 11 (2), 413–440.
Lin, X.S., Liu, X., 2007. Markov aging process and phase-type law of mortality. North American Actuarial Journal 11 (4), 92–109.
Longini Jr, I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F., Hethcote, H.W., 1989. Statistical analysis of the stages of HIV infection using a Markov model. Statistics in Medicine 8 (7), 831–843.
Macdonald, A.S., Richards, S.J., Currie, I.D., 2018. Modelling Mortality with Actuarial Applications. Cambridge University Press.
McGrory, C.A., Pettitt, A.N., Faddy, M.J., 2009. A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. Computational Statistics & Data Analysis 53 (12), 4311–4321.
Moler, C., Van Loan, C., 1978. Nineteen dubious ways to compute the exponential of a matrix. SIAM Review 20 (4), 801–836.
Neuts, M.F., 1975. Probability distributions of phase type. In: Liber Amicorum Professor Emeritus H. Florin. Department of Mathematics, University of Louvian, Belgium, pp. 173–206.
Neuts, M.F., 1981. Matrix-Geometric Solutions in Stochastic Models. Johns Hopkins Series in the Mathematical Sciences, vol. 2.
Olivieri, A., Pitacco, E., 2015. Introduction to Insurance Mathematics: Technical and Financial Features of Risk Transfers. Springer.
Olsson, M., 1996. Estimation of phase-type distributions from censored data. Scandinavian Journal of Statistics 23 (4), 443–460.
Olsson, M., 1998. The EMpht Programme. Manual. Chalmers University of Technology and Götborg University.
Pike, M., 1966. A method of analysis of a certain class of experiments in carcinogenesis. Biometrics 22 (1), 142–161.
Pitacco, E., 2004. Survival models in a dynamic context: a survey. Insurance. Mathematics & Economics 35 (2), 279–298.
Pitacco, E., 2019. Heterogeneity in mortality: a survey with an actuarial focus. European Actuarial Journal 9 (1), 3–30.
Renshaw, A., Haberman, S., 2021. Modelling and forecasting mortality improvement rates with random effects. European Actuarial Journal 11 (2), 381–412.
Rizk, J., Walsh, C., Burke, K., 2021. An alternative formulation of Coxian phase-type distributions with covariates: application to emergency department length of stay. Statistics in Medicine 40 (6), 1574–1592.
Royston, P., Parmar, M.K., 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 21 (15), 2175–2197.
Shapovalov, V., Landsman, Z., Makov, U., 2021. Exchangeable mortality projection. European Actuarial Journal 11 (1), 113–133.
Sherris, M., Zhou, Q., 2014. Model risk, mortality heterogeneity, and implications for solvency and tail risk. In: Recreating Sustainable Retirement: Resilience, Solvency, and Tail Risk, pp. 113–133.
Tang, X., Luo, Z., Gardiner, J.C., 2012. Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. Statistics in Medicine 31 (14), 1502–1516.
Zeddouk, F., Devolder, P., 2020. Mean reversion in stochastic mortality: why and how? European Actuarial Journal 10 (2), 499–525.