4-23-2021 2:00 PM

A class of phase-type ageing models and their lifetime distributions

Boquan Cheng, *The University of Western Ontario*

Supervisor: Mamon, Rogemar, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

Abstract

Ageing is a universal and ever-present biological phenomenon. Yet, describing the ageing mechanism in formal mathematical terms — in particular, capturing the ageing pattern and quantifying the ageing rate — has remained a challenging actuarial modelling endeavour. In this thesis, we propose a class of Coxian-type Markovian models. This class enables a quantitative description of the well-known characteristics of ageing, which is a genetically determined, progressive, and essentially irreversible process. The unique structure of our model features the transition rate for the ageing process and a functional form for the relationship between ageing and death with a shape parameter that captures the biologically deteriorating effect of ageing. The force of moving from one state to another in the Markovian process indicates the intrinsic biological ageing force. The associated increasing exit rate captures the external force of stress due to mortality risk on a living organism.

We define an index, called physiological age, to quantify the heterogeneity between individuals. The physiological age can be used to compare the death rate between individuals in which an individual with a higher physiological age has higher mortality rate. The probability in each state at any time is calculated, and the distribution of the physiological age at any chronological age is obtained. We also prove that the distribution of the physiological age at a given time can be approximated by a normal distribution as the Phase-Type Ageing Model (PTAM) allows for a large number of states. The approximation can be used to quickly compute the probability in each state at any given time. The lifetime distribution for each individual readily follows from their physiological ages whose distribution is helpful in quantifying the variability of individual health status in the population.

We develop an efficient method to evaluate the PTAM's likelihood utilising a lifetime data set. Our likelihood calculation uses vectorisation to find simultaneously the density function at observed lifetimes. Furthermore, our method uses uniformisation strategy to stabilise the numerical calculation with an appropriate error tolerance. We demonstrate that our numerical method is more accurate and faster than the traditional method using matrix exponential. Lastly, we investigate the estimability of the PTAM when only the lifetime data is observable along with some conditions that could improve the model's estimability in terms of parameters' identification.

Keywords: Phase-type aging model, physiological age, uniformisation, estimability

Summary for Lay Audience

Ageing is a universal and ever-present biological phenomenon. Yet, describing the ageing mechanism in formal mathematical terms – in particular, capturing the ageing patterns and quantifying the ageing rate - has remained a challenging endeavour. The modelling of the human ageing process, under a spectrum of uncertainty, is critical in the accurate valuation and robust risk management of insurance and pension products. We put forward a general model having a small number of parameters but flexible enough to produce a variety of lifetime distributions. Our research contributions widen the available actuarial and survival analysis techniques in the following way: (i) A phase-type ageing model (PTAM) is constructed in which the ageing's deteriorating effect and the associated increasing mortality risk are simultaneously taken into account. (ii) A physiological age index is introduced in order to quantify the heterogeneity between individuals. The physiological age index can be used to classify various mortality risk levels. (iii) Some pertinent statistical properties of the PTAM and the physiological age distribution are established. (iv) An efficient method is developed to evaluate the PTAM's likelihood. Some numerical examples, utilising simulated and actual lifetime data sets, are provided to demonstrate that our proposed calculation technique is faster and more accurate than the traditional method based on matrix exponential. (v) The estimability of the PTAM under different lifetime-information scenarios is examined in the context of improving the parameter estimation. Our modelling of the ageing process through lifetime distributions is also of utmost importance in crafting suitable regulatory requirements and policies that will strengthen further the public confidence in the national or even global financial system.

Co-Authorship Statement

Chapter 3 of this thesis has been published. The remaining Chapters consist of materials that delve into the pertinent topics of Chapter 3, dealing with the various aspects of the proposed model. I declare that the research outputs incorporated in this thesis are the direct results of my main research works and efforts during the course of my PhD program.

The research results in Chapter 3 form the main basis of an article with the following citation details:

'Cheng, B., Jones, B., Liu, X., and Ren, J.(2021), The mathematical mechanism of biological aging, *North American Actuarial Journal*, accepted:DOI:10.1080/10920277.2020.1775654.

As the first author of a research paper incorporated in this thesis, I am responsible for the development of the modelling set-up, the implementation of algorithms, and the completion of the manuscript. The research plan, model formulation, and empirical analysis of the results in the published paper are guided and supervised by Dr. Bruce Jones, Dr.Jiandong Ren and Dr. Xiaoming Liu. Some insights in the analysis of the Channing House data set were provided by Dr. Bruce Jones. Dr. Xiaoming Liu oversaw the goodness-of-fit comparison between the proposed model in this thesis and the Lin and Liu's model.

An integrated-article format is employed in line with Western's thesis guidelines. I certify that this thesis is fully a product of my own work. This was conducted from September 2016 to September 2020 under the supervision of Dr. Bruce Jones and Dr. Xiaoming Liu, and from September 2020 to present under the supervision of Dr. Rogemar Mamon at The University of Western Ontario.

Acknowledgements

I would like to thank my current supervisor Dr. Rogemar Mamon for his generous support and unfailing patience throughout the last year of my PhD programme. His professionalism, excellent supervision, broad knowledge and clear guidance ensured the accomplishment and improvement of this thesis.

I would also like to convey my gratitude to Dr. Bruce Jones for his inspiring mentorship and conscientious supervision and to Dr. Jiandong Ren for his constant encouragement in the early stage of my research pursuits. As well, I am grateful to the helpful faculty and staff at the Department of Statistical and Actuarial Sciences (DSAS). The financial support provided by DSAS is equally appreciated.

I also express my deep appreciation to all of my thesis examiners: Dr. Shu Li, Dr. Pauline Barmby, Dr. Pei Yu and Dr. Taehan Bae, for their valuable time in reading and assessing this thesis.

Lastly, I would like to acknowledge my colleagues and friends for their moral support and my parents' love and contribution.

I would not have reached this milestone without all the people mentioned above.

Contents

List of Figures

List of Tables

List of Appendices

Nomenclature

Y_t	Random variable of state being in at time t, page 9
E	Transient state space, page 11
$\boldsymbol{\alpha}$	$1 \times m$ row vector of initial probability, page 11
$\boldsymbol{\Lambda}$	$m \times m$ degenerated transition matrix of a Phase-Type model, page 11
$\boldsymbol{h}$	$m \times 1$ column vector of the absorption rates of a $m + 1$ states Phase-Type distribution, page 11
$\boldsymbol{e}$	$m \times 1$ column vector of ones, page 11
T	Random variable of the time until absorption, page 11
$\boldsymbol{P}(t)$	$m \times m$ transition probability at time t, page 11
$\boldsymbol{I}$	$m \times m$ identity matrix, page 11
$S(t)$	Survival probability of a phase-type distribution beyond time t, page 12
$f(t)$	Probability density function of a phase-type distribution at time t, page 12
$h(t)$	Hazard function of a phase-type distribution at time t, page 12
$\lambda_{i,j}$	Transition rate from state i to state j, page 12
$\mathbf{p}(t)$	$1 \times m$ probability vector that the individual is in each transient state at time t, page 13
λ_i	Transient rate of a Coxian distribution from state i to state $i + 1$, page 15
h_i	Absorption rate of a Coxian distribution in state i, page 15
$P_k(t)$	Probability in state k at time t for a Coxian distribution, page 16
h_1	One of the parameters of the proposed PTAM, the absorption rate in state 1, page 25
h_m	One of the parameters of the proposed PTAM, the absorption rate in the last transient state, page 25
s	One of the parameters of the proposed PTAM, controls the shape of h_i, page 25

m Total number of transient states in the PTAM, page 28

ψ Life span parameter of the proposed PTAM, page 28

λ Transition rate from one transient state to the next transition state for the proposed PTAM, page 28

$l(\hat{\theta})$ Maximum of log-likelihood, page 32

$\widehat{h_1}$ Maximum likelihood estimate of h_1, page 33

$\widehat{h_m}$ Maximum likelihood estimate of h_m, page 33

$\widehat{s}$ Maximum likelihood estimate of s, page 33

$\widehat{\text{TVaR}}_{1-\alpha}(T)$ Empirical estimate of $\text{TVaR}_{1-\alpha}(T)$, page 40

NLL Negative log-likelihood, page 41

pdf Probability density function, page 41

MLE Maximum likelihood estimator, page 41

d_i Eigenvalues of the degenerated transition matrix $\mathbf{\Lambda}$, page 43

Z_t A transformation of Y_t that has ranges $[0, 1]$, page 53

$o(\epsilon)$ A number c that $\lim_{\epsilon \to 0} \frac{c}{\epsilon} = 0$, page 54

$O(\epsilon)$ A number c that $\lim_{\epsilon \to 0} \frac{c}{\epsilon} = M$, M is a finite number, page 54

$\phi(x)$ Probability density function of the standard normal distribution, page 61

$\Phi(x)$ Cumulative density function of the standard normal distribution, page 61

$M_N(u)$ Moment generating function of the standard normal distribution, page 61

$M(u)$ Moment generating function of the proposed PTAM, page 62

$g(t)$ Limit of the resulting hazard rate of the proposed PTAM as $m \to \infty$, page 73

$L(\boldsymbol{\theta})$ Likelihood function, page 80

$l(\boldsymbol{\theta})$ Log-likelihood function, page 80

$\boldsymbol{P}$ $m \times m$ probability transition matrix with the (i, j) element equal to the probability that a transition from state i to state j occurs given a hypothesised transition occur, page 83

$e^{\mathbf{\Lambda} t}$ $m \times m$ matrix exponential of $\mathbf{\Lambda} t$, page 83

$P_{1,i}^{j}$ First row ith column element of $\boldsymbol{P}^j$, page 84

$S_T(t)$ Theoretical survival function, page 85

$S_N(t)$ Numerical survival function by the proposed algorithm, page 85

$\boldsymbol{\beta}$ $n \times m$ matrix whose ith row is $\mathbf{p}(t_i)$, page 87

$\boldsymbol{\gamma}$ $n \times m$ matrix whose ith row is $\mathbf{p}(\ell_i)$, page 88

.* Element-wise multiplication, page 89

./ Element-wise scalar division, page 89

$f_T(t)$ Theoretical probability density function at time t, page 89

$f_N(t)$ Numerical probability density function at time t by the proposed algorithm, page 89

$l_T(\boldsymbol{\theta})$ Theoretical log-likelihood, page 92

$l_N(\boldsymbol{\theta})$ Numerical log-likelihood by the proposed algorithm, page 92

$M(\boldsymbol{\theta})\epsilon$ An upper bound for the difference between the theoretical log-likelihood and the numerical log-likelihood, page 93

$f_{NP}(t)$ Numerical value of the probability density function by the proposed algorithm, page 95

$f_{NT}(t)$ Numerical value of the probability density function by the traditional method using **expm**, page 95

D Two-sample Kolmogorov-Smirnov test statistic, page 104

$F_{1,n}(t)$ Empirical cdf, page 104

R^2 Coefficient of determination, page 107

RAS Rate of algorithm's success, page 110

K Number of clones for the original data, page 126

σ^2 Variance of measurement error, page 132

$\boldsymbol{P^*}$ $m \times m$ probability matrix $e^{\boldsymbol{\Lambda} k}$, page 133

$\boldsymbol{C_{i,j}}$ $m \times m$ matrix used to calculate the log-likelihood when we can measure the states at some ages with unknown measurement errors, page 133

λ_1^* Smallest eigenvalue of the covariance matrix of the posterior, page 138

λ_K^* Largest eigenvalue of the covariance matrix of the posterior, page 138

λ_K^s Standardized largest eigenvalue of the covariance matrix of the posterior, page 138

$\boldsymbol{e_k}$ The $m \times 1$ column vector with first $k-1$ elements equal to 0 and remaining elements equal to 1, page 160

Chapter 1

Introduction

1.1 Background

As human beings, we start our life journey on the day when we were born. We are weak babies when we are delivered. As time passes, our bodies gradually change – from infant to young child, from young child to teenager, from teenager to adult, from adult to middle-aged, from middle-aged to senior, even septuagenarian (70-79), octogenarian (80-89), nonagenarian (90-99), centenarian (100+), if we are lucky enough! The journey of life is a progression of ageing, becoming "old", accompanied by inside and outside changes to our bodies. Several characteristic ageing symptoms are listed below:

- Teenagers lose the young child's ability to hear high-frequency sounds above 20 kHz (Rodríguez Valiente et al., 2014).
- Wrinkles develop mainly due to photoageing, particularly affecting sun-exposed areas (Thurstan et al., 2012).
- After age 30, the human body mass decreases until 70 years and then shows damping oscillations (Gerasimov and Ignatov, 2004).
- At around age 50, hair turns grey (Pandhi and Khanna, 2013). Pattern hair loss by the age of 50 affects about 30 -– 50% of males (Hamilton, 1951) and a quarter of females (Vary Jr, 2016).
- Almost half of people older than 75 have hearing loss (presbycusis) impeding spoken communication (U.S. Department of Health and Human Services, 2016).
- Recent evidence suggests that age-related risk of death plateaus after age 105 (Thompson, 2018).
- The maximum human lifespan is suggested to be 115 years (Dong et al., 2016; Zimmer, 2016).

These common ageing symptoms reflect that there is a ubiquitous ageing process underneath the life of all human beings. Unfortunately, the ageing process is unobservable. Instead, we can only observe some ageing phenotypes (i.e. the above-mentioned ageing symptoms). Therefore, we resort to the utility of stochastic methods in modelling the ageing process.

1.2 Literature Review on Modeling ageing Process

Despite being aware of its characteristics, ageing is conceptual and its definition is vague. Researchers have their own beliefs, understandings, and interpretations of ageing. Typical research studies on ageing concentrate on determining the relation between biological indicators of ageing and longevity. Cevenini et al. (2008) pointed out that ageing is complex and determined by multiple factors, such as genes (Leroi et al., 2005; Salvioli et al., 2006; Beekman et al., 2006; Capri et al., 2008), immunology (Franceschi et al., 1995a,b; Franceschi and Bonafè, 2003; Sansoni et al., 2008), familial component (Atzmon et al., 2004; Perls et al., 2002; Willcox et al., 2006; Schoenmaker et al., 2006), living location (Deiana et al., 1999; Gueresi et al., 2003), and epigenetics (Fraga et al., 2005). The aforementioned studies provided a good overview of ageing phenotype in humans. They reviewed some models utilised to study human ageing and longevity with a particular focus on families with long-living members, twins and cohorts of unrelated subjects. Those factors include familial gene component, twin classification (monozygotic versus dizygotic twins), body mass index, metabolism, and risk factors for cardiovascular diseases.

Mathematical models linking the ageing process and mortality are reviewed in Yashin et al. (2000). The characteristic of deterioration was incorporated into the assumptions of these mortality models. In Yashin et al. (2000), it is claimed that statistical methods based on an appropriate mathematical model are required to analyse the information collected in the studies of ageing and mortality. One such model is the phase-type model, which can approximate many types of lifetime distributions. Following Lin and Liu (2007), we also consider a specific phase-type model to capture the human ageing process.

As stated in Lin and Liu (2007),

> *ageing, as applied to living organisms, is the genetically determined, progressive, and essentially irreversible diminution with the passage of time of the ability of an organism or of one of its parts to adapt to its environment, manifested as diminution of its capacity to withstand the stresses to which it is subjected (i.e. the increase of susceptibility to certain diseases with age), and culminating in the death of the organism.*

From the definition put forward by Jones (1956), the ageing process is also characterised as genetically determined, progressive, and essentially irreversible. It is our view that people age differently; that is, the ageing process is personalised. This is because ageing is determined by both internal factors (e.g., genes) and external factors (family background, education, accidents). According to Herskind et al. (1996), an additive genetic component explains about 2 percent of the variability in life span, indicating that non-genetic factors contributes to some extent to a person's life span. Yashin et al. (2012) found that each factor generates different average age patterns. There are certainly many factors affecting ageing such that the ageing processes are individually different, but the pattern of ageing for individuals in a cohort population can be similar. From this perspective, the ageing process can be treated as a stochastic process, and the ageing experience of each individual is a trajectory of the stochastic process.

Apparently, ageing impacts health and hence, mortality. As ageing progresses, the biological functionality of organs may change. For example, the ageing system can cause a decline in

the T cells (thymus cells) and B cells (bone marrow- or bursa-derived cells), which are the major cellular components of the adaptive immune response according to Prelog (2006). Indeed, as a person gets older, the functionality of his/her body cells deteriorates, making him/her more likely to be affected by illnesses and contract diseases. Thus, the person's mortality will rise eventually in which case ageing affects mortality indirectly through the ageing's influence on individual's health status. On the other hand, the age patterns of mortality risk result from the interaction between the process of individual ageing and external factors as asserted in Yashin et al. (2012) and the references therein. Therefore, ageing has a direct impact on mortality.

We take the position in this thesis that ageing is **progressive**, **essentially irreversible**, **personalised** and **highly correlated with mortality**. Since the ageing process cannot be observed directly, it requires a mathematical model to describe it. The calibration of such a model with ageing-related data is desirable and beneficial in the context of annuity or insurance product valuation and reserve setting. It aids insurance companies identify weak and strong cohorts, thereby lowering insurance premiums for healthier policy holders and adjusting prices for less healthy policy holders. This provides a fairer pricing mechanism than valuations merely based on mortality at each age. Our modelling approach, therefore, complements the underwriting strategy in premium adjustments. We recognise the difficulty of collecting ageing-related data from birth to death of each individual. Nonetheless, the collection of individual's time of death is a straightforward task and as ageing and mortality are strongly correlated, it is reasonable to use mortality data in calibrating a mathematical model for ageing.

It is worth noting that there is a difference between the ageing model and mortality model. The latter focuses on fitting the mortality data, whilst the former is designed to pin down the underlying process. The death time is treated as the terminal time of this process. So, the ageing-process model contains more information than the mortality-rate model.

Stochastic models were previously utilised to examine the relationship between ageing and mortality via a physiological variable; see for example, Yashin et al. (1985); Woodbury and Manton (1977). The physiological variable provides information about the current status of death risk in the ageing process.

Lin and Liu (2007) constructed a structured phase-type model, called phase-type ageing model, by taking advantage of Markov chains to mimic the ageing process. The states are labelled using integers starting from 1 and are interpreted as physiological ages. An excellent fit to the Swedish cohort mortality data (1811-1911) was obtained except for super old ages. The progression of a Markovian structure imitates the ageing process and the corresponding lifetime distribution.

The model's structural framework in this thesis is based on the phase-type ageing model of Lin and Liu (2007). The structure on the absorption rates in Lin and Liu's model is not flexible enough to achieve a variety of lifetime distributions. We replace a pliable form on the absorption rates so that the PTAM can result numerous shape of distributions. Furthermore, the transition rate between transient states in Lin and Liu's model is a free parameter, while in our proposed model is linked to the total number of states. This can ease the interpretation of the model parameters. It has to be noted that Su and Sherris (2012) used another structured phase-type model for the Australian population data, and the calibrated model achieved a good fit. The merits of phase-type models for modeling the human ageing process are akin to the intuitive model interpretation and promising fitting results for mortality data.

The phase-type models, however, have some drawbacks, which include overparameterisa-

tion and a time-consuming calibration. The general phase-type model has a large number of parameters, and the number of parameters is usually more than required to fit the data well, giving rise to the overparameterisation problem just like in Faddy and Wilson (2000). Calibrating phase-type models is computationally expensive when the total number of states is large; refer for instance to the applications in Lin and Liu (2007) and Su and Sherris (2012). Furthermore, a phase-type model with a large number of parameters may have multiple sets of parameters that can maximise the likelihood of a set of data, which is problematic for model inference. This is related to the model estimability (cf, Chapter 6 of this thesis for more details on model estimability).

The purpose of this thesis is to overcome the challenges of modelling the ageing process using phase-type models under the following considerations.

- What does the Markov chain of the proposed model, including the functional form on the transition rates, look like?
- What are the associated model's statistical properties?
- How is the efficient calibration of the model going to be carried out?
- How could the estimability and identifiability aspects of the proposed model be 'improved' so that the maximum likelihood estimation has a unique result?

The main contributions of this thesis are as follows:

- a relatively flexible structured phase-type ageing model to capture the ageing process;
- establishing the statistical properties of the model;
- an efficient algorithm to estimate the parameters of the proposed model; and
- an examination of the conditions to improve the model's estimability and identifiability.

1.3 Description of methodology

In order to accomplish the objectives enumerated in Subsection 1.2, we execute the following:

- Propose a phase-type model with a small number of parameters yet flexible enough to achieve a variety of lifetime distributions. This modelling structure requirement is important to enable a long Markov chain to mimic the ageing process; without imposing this structure, calibration will be challenging.
- Investigate the statistical properties of the proposed model to provide insights for calibration and further analysis.
- Develop an efficient algorithm for the maximum likelihood estimation of the proposed model considering numerous matrix exponential calculations in the estimation process.
- Understand the estimability and the identifiability issues of the proposed model for the purpose of statistical inference and meaningful interpretation of the parameters of our ageing model. It is our goal as well to get the distribution of the calibrated parameter values.

1.4 Outline of the remaining chapters

The succeeding parts of this thesis are organised as follows.

- Chapter 2 provides the relevant mathematical preliminaries for the phase-type models and the Coxian models;
- The proposed model is introduced in Chapter 3 and some examples showing applications of the proposed model on both the real data and simulated data are presented;
- In Chapter 4, certain statistical properties of the proposed model are discussed;
- Chapter 5 describes the parameter estimation including the analyses of the algorithm's accuracy and efficiency;
- The assessment of the identifiability and estimability of the proposed model, from both the estimation and Bayesian perspectives, is the core subject of Chapter 6;
- Finally, Chapter 7 contains some concluding remarks.

The thesis has four appendices. Appendix A gives the proofs of lemmas and theorems. These include the results on the denseness of the generalised phase-type ageing model, on the stochastic-order of two PTAM random variables, and on the physiological age that increases as time passes in the stochastic-order sense. In Appendix B, the issues concerning a large number of states and an inappropriate ψ, which is a life span parameter in our model, are tackled. The flexibility of the resulting distribution from the proposed model is detailed in Appendix C. Lastly, Appendix D documents the MATLAB code for our proposed algorithm.

Chapter 2

Mathematical preliminaries

In this Chapter, the mathematical background for our proposed model is laid out. We start with the most general model, the Markov-chain process, then move on to a specific model called the phase-type ageing model (PTAM).

2.1 Markov process

A Markov process, named after Markov (1954), is a stochastic model possessing the (Markov) property mainly characterised by memorylessness.

Definition 2.1. *Let* $(\Omega, \mathscr{F}, P)$ *be a probability space with a filtration* $(\mathscr{F}_s, s \in I)$*, for some index set* I*, and let* (E, Σ) *be a measurable space. An* (E, Σ)*-valued stochastic process* $Y = Y_t : \Omega \to E_{t\in I}$ *adapted to the filtration is said to possess the Markov property if, for each* $A \in \Sigma$ *and each* $s, t \in I$ *with* $s < t$

$$P(Y_t \in A|\mathscr{F}_s) = P(Y_t \in A|Y_s)\,. \tag{2.1.1}$$

In the case where Σ is a discrete set with discrete sigma algebra, (2.1.1) is same as,

$$P(Y_t = y_t|\mathscr{F}_s) = P(Y_t = y_t|Y_s = y_s)\,.$$

Furthermore, if Σ is the set of states for a Markov model with countable states and $I = [0, \infty)$, then $P(Y_t = y_t|\mathscr{F}_s)$ is the probability that the process is in state y_t at time t given the state being in at any time up till s, and $P(Y_t = y_t|Y_s = y_s)$ is the probability of the process in state y_t at time t conditional on the process in state y_s at time s. In other words, the probability distribution of future states (Y_t), conditional on both past and present states $(\mathscr{F}_s)$, is the same as that conditional on the current state (Y_s). In our research, we focus on a specific Markov model that has countable states and the index set is the set of non-negative real numbers.

Markov chains are widely used in real-world applications. These include queueing (Neuts, 1978; Latouche and Ramaswami, 1999; Klimenok and Dudin, 2006; Karlin, 2014; Hajek, 2015; Gagniuc, 2017), cruise control systems (Zhang and Vahidi, 2011), classical text translation (Markov, 2006), healthcare (Fackrell, 2009; Garg et al., 2010), actuarial problems (Hoem, 1969, 1977; Jones, 1994; Zhang, 2016), and quantitative finance (Mamon and Elliott, 2007; Elliott and Mamon, 2002).

On the one hand, the "memoryless" property of Markov chains simplifies the probability structure, making the calculations of the corresponding probability distributions easier. On the other hand, the assumption that the future states depend only on the present state satisfies many natural dynamics of real-world processes.

The evolution of a Markov chain intuitively matches the process of a lifetime random variable by considering the lifetime as the time in the process until absorption. When modeling lifetime random variables with non-negative domains, we shall illustrate that a particular class of Markov models is used extensively.

2.2 Phase-type distribution

The particular class of Markov models pointed out in Subsection 2.1 refers to the phase-type models, whose distributions are mixtures of exponential distributions. Erlang (1917) first extended the exponential distribution to what is now called the Erlang distribution. This distribution was developed by defining a non-negative random variable that models the time it takes for a process to move through a fixed number of states, spending an exponential amount of time with a fixed rate in each state. Fifty-eight years later, Neuts (1975) generalised the Erlang distribution via a phase-type random variable, which is defined as the time spent in the transient states of a finite-state continuous-time Markov chain with one absorbing state. It has to be noted that the Erlang distribution was showed by David and Larry (1987) to be the least variable phase-type distribution. In other words, it is the phase-type distribution that has the smallest coefficient of variation (i.e., $\mathrm{Var}(T)/\,\mathbb{E}(T)^2$, where T is the lifetime random variable).

The associated probability distribution of a phase-type model is related to the solution of a set of differential equations studied by Neuts (1982). Given the flexible and dynamic structure of phase-type models together with the explicit solutions of the associated probability distribution, phase-type models have been widely employed in areas as diverse as telecommunications (Sengupta, 1989; Asmussen, 1992; Ausın et al., 2004), finance (Asmussen et al., 2004), teletraffic modelling (Thummler et al., 2006), biostatistics (Olsson, 1996), queueing theory (Faddy and McClean, 1999), drug kinetics (Faddy, 1993), reliability theory (Pérez-Ocón and Castro, 2004), classic illness-death model (Kodell and Nelson, 1980), and survival analysis (Aalen, 1995; Olsson, 1996). A good survey of phase-type models can be found in Chapter 1 of Fackrell (2003).

Definition *Let Y_t be a time-homogeneous Markov process defined on a finite state-space $S = E \cup \Delta = \{1, 2, \ldots, m\} \cup \{\Delta\}$, where Δ is an absorbing state and the states in $E = \{1, 2, \ldots, m\}$ are transient. Let Y_t have initial distribution $(\boldsymbol{\alpha}, 0)$, where α is a $1 \times m$ row vector, and infinitesimal generator*

$$\begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{h} \\ \mathbf{0} & 0 \end{pmatrix}$$

where $\boldsymbol{\Lambda}$ is a $m \times m$ matrix whose ith row jth column element is the transition rate from state i to state j for any $i, j \in S$; $\mathbf{0}$ is the $1 \times m$ row vector of zeros; $\boldsymbol{h} = -\boldsymbol{\Lambda}\boldsymbol{e}$; and $\boldsymbol{e}$ is the $m \times 1$ column vector of ones.

Let T denote the time until absorption or the time until death in the human lifetime context. Then T is said to follow a phase-type (PH) distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{\Lambda})$. The matrix

$\mathbf{\Lambda}$ is called the intensity matrix or the degenerated transition matrix. The row vector $\boldsymbol{\alpha}$ is called the initial probability.

Example 2.1. An example of a phase-type model is illustrated in Figure 2.1.

△

Considering state 1 as being healthy, state 2 as being sick, and the absorbing state as death, then the phase-type model with two states in Example 2.1 is known as the disability-income insurance model.

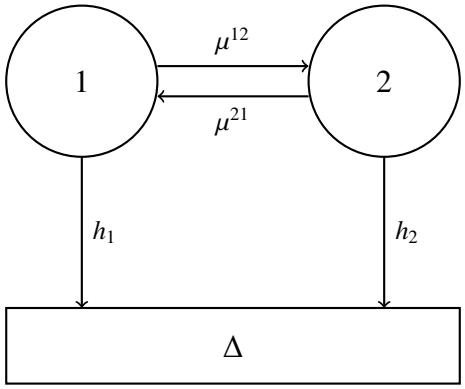

Figure 2.1: A two-state phase-type model.

Phase-type models have intuitive interpretation providing the underlying biological/engineering mechanism of the system that generates the resulting lifetime distribution. In particular, the survival time in such a system is the sum of the sojourn times in all states that the process has ever visited before it is absorbed. The survival probability at any time t is equal to the total probability that the process is in any state, excluding the absorbing state, at time t.

For $s \geq 0, t \geq 0,\ i, j \in E$, let $P_{ij}(s, s+t)$ be the transition probability from state i to state j during the time period $(s, s+t]$:

$$P_{ij}(s, s+t) = \Pr(Y_{s+t} = j | Y_s = i)\,.$$

Because Y_t is a time-homogeneous Markov process, the probability $P_{ij}(s, s+t)$ does not depend on s, so we write $P_{ij}(t) = P_{ij}(s, s+t)$ for all $s \geq 0$.

Suppose $\boldsymbol{P}(t) = \{P_{ij}(t)\}_{i,j\in E}$ is the $m \times m$ transition probability matrix (amongst the transient states) of Y_t in the time interval $(0, t]$. The transition probability matrix satisfies the Kolmogorov forward equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}(t) = \boldsymbol{P}(t)\mathbf{\Lambda}\,, \tag{2.2.2}$$

with the initial condition $\boldsymbol{P}(0) = \boldsymbol{I}$, where $\boldsymbol{I}$ is an $m \times m$ identity matrix.

The Kolmogorov forward equation has the unique solution

$$\boldsymbol{P}(t) = \exp(\mathbf{\Lambda} t), \tag{2.2.3}$$

where

$$\exp(\mathbf{\Lambda} t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{\Lambda}^n.$$

As a result, the survival function of the time-until-absorption random variable T can be expressed as

$$S(t) = \boldsymbol{\alpha} \exp(\mathbf{\Lambda} t)\boldsymbol{e}, \quad t > 0\,,$$

and the probability density function(pdf) of T is

$$f(t) = \boldsymbol{\alpha} \exp(\mathbf{\Lambda} t)\boldsymbol{h}, \quad t > 0\,.$$

Remark 2.1. *We will call the (pdf) and survival function the probability distribution. This terminology also applies to the succeeding Chapters.*

The hazard function of T is the ratio of the pdf to the survival function, or $f(t)/S(t)$. Therefore, the hazard function for the phase-type distribution is

$$h(t) = \frac{\boldsymbol{\alpha} \exp(\mathbf{\Lambda} t)}{\boldsymbol{\alpha} \exp(\mathbf{\Lambda} t)\boldsymbol{e}} \boldsymbol{h}. \tag{2.2.4}$$

It could be verified that $\dfrac{\boldsymbol{\alpha} \exp(\mathbf{\Lambda} t)}{\boldsymbol{\alpha} \exp(\mathbf{\Lambda} t)\boldsymbol{e}}$ is a $1 \times m$ row vector, whose ith element ($i = 1, \ldots m$) is the probability in state i given the process is in transient states at time t. This representation of hazard rate is a weighted average of the absorption rate in each transient state, and the weight is dependent on time t.

The hazard function $h(t)$ converges to the minimal negative diagonal value of the intensity matrix $\mathbf{\Lambda}$ as $t \to \infty$. Specifically, for an $m-$state phase-type model, its intensity matrix $\mathbf{\Lambda}$ is an $m \times m$ matrix. The (i, j) element in $\mathbf{\Lambda}$, $\lambda_{i,j}$, is the transition rate from state i to state j when $i \neq j$, and the (i, i) element is equal to the negative sum of other elements in the ith row so that

$$\lambda_{i,i} = -\sum_{j \neq i} \lambda_{i,j}.$$

We have $h(t) \to \min_{i=1,\ldots,m}(-\lambda_{i,i})$ as $t \to \infty$.

The moment generating function of T is

$$\mathbb{E}\left(e^{sT}\right) = \int_0^{\infty} e^{st} \mathrm{d}F(t) = -\boldsymbol{\alpha}(s\boldsymbol{I} + \mathbf{\Lambda})^{-1}\boldsymbol{h}. \tag{2.2.5}$$

The kth moment of the phase-type distribution is

$$\mathbb{E}\left(T^k\right) = (-1)^k k! \boldsymbol{\alpha} \mathbf{\Lambda}^{-k} \boldsymbol{e},$$

by evaluating the kth derivative of (2.2.5) with respect to s at the point $s = 0$. When $k = 1$, the mean of T is

$$\mathbb{E}(T) = -\boldsymbol{\alpha} \mathbf{\Lambda}^{-1} \boldsymbol{e}. \tag{2.2.6}$$

Similarly, one can derive the variance of T by the formula $\mathbb{E}(T^2) - (\mathbb{E}(T))^2$, and get

$$\text{Var}(T) = 2\boldsymbol{\alpha}\boldsymbol{\Lambda}^{-2}\boldsymbol{e} - \left(\boldsymbol{\alpha}\boldsymbol{\Lambda}^{-1}\boldsymbol{e}\right)^2. \tag{2.2.7}$$

One can use (2.2.6) and (2.2.7) to quickly calculate the life expectancy and its associated variability for the phase-type models.

Notice that the ith element of the $1 \times m$ vector

$$\mathbf{p}(t) = \boldsymbol{\alpha} \exp(\boldsymbol{\Lambda} t) \tag{2.2.8}$$

is just the probability that $Y_t = i$, and $\mathbf{p}(t)\boldsymbol{e}$ gives the probability that $Y_t \in E$.

The class of phase-type distributions is dense in the space of all continuous non-negative distributions. That is, any distribution on $[0, \infty)$ can, at least in principle, be approximated arbitrarily closely by a phase-type distribution (Johnson, 1993). Therefore, phase-type models have been successfully applied in many fields to obtain explicit analytical solutions, i.e., the ruin-related quantities (Feng, 2009; Badescu et al., 2009), and in reliability system (Ruiz-Castro et al., 2008; Asmussen, 2000). However, getting numerical results by solving the differential equations (2.2.2) or calculating the matrix exponential (2.2.3) is not trivial, especially when the dimension of the matrix $\boldsymbol{\Lambda}$ (i.e., the number of states, m) is high.

A general phase-type model allows transitions from any state to any other state. This structure provides the flexibility when approximating any non-negative distribution. A general phase-type distribution has typically many parameters, and some parameterisation methods have the issue of non-unique estimated value; see for example, (Asmussen et al., 1996; Marshall and Zenga, 2009b; Fackrell, 2003). The non-uniqueness is due to the overparameterisation of the phase-type model, i.e., the number of parameters is more than required. For instance, a general phase-type model with m transient states has $m^2 + m$ parameters; that is, m^2 parameters for the degenerated transition matrix and m parameters for the initial probability. When using an m-state phase-type model, with $m \geq 2$ and $m^2 + m$ parameters to approximate an exponential distribution with one parameter, it may be shown that there is more than one set of parameter values resulting from the exponential distribution. Due to its overparameterisation, the parameter estimation of phase-type distributions presents some difficulty. As will be discussed in Chapter 6, the problem of non-uniqueness in parameter specification is usually intermingled with the complication of model estimation, mainly when the model contains a large number of states, and the model parameter values are close to each other.

One generalisation of the phase-type model is obtained by replacing the exponentially distributed assumption on each waiting time to leave the states by a parametric non-exponential distribution (Huzurbazar, 1999), whose structure may depend on the present state or next state to be visited. As a result, the time-until-absorption random variable T is the sum of random variables with specific parametric distributions. For example, the generalised phase-type model has two transient states, and the time spent in each state follows a Weibull distribution. The resulting pdf is the convolution of the pdf's of the Weibull distributions.

Another generalisation of phase-type models is extending the assumption of homogeneous to inhomogeneous transition rates (Liu and Lin, 2012; Sherris and Zhou, 2014; Albrecher and Bladt, 2018). Such a generalisation keeps the denseness in the class of distributions with non-negative domains. The time-inhomogeneous phase-type models are more parsimonious

with fewer parameters in approximating heavy-tailed distributions in comparison to the time-homogeneous phase-type models. This is because the time-homogeneous phase-type models are light-tailed distributions. So, a large number of states is required to fit the tails well when approximating heavy-tailed distributions.

This thesis considers the applications of phase-type models for the human ageing process, similar to Aalen (1995), Lin and Liu (2007), and Su and Sherris (2012). In these applications, the phase-type models involve a large number of states with sparse degenerated transition matrices. A class of phase-type models is more suitable for modeling the human ageing process than the general phase-type models.

2.3 Coxian distributions

A particular class of phase-type models, proposed by Cox (1955) and called Coxian distributions, preserves the denseness in the class of distributions with non-negative domains. The associated Markov chain only allows transitions from each state to either the next transient state or the absorbing state. Such restriction forces the degenerated transition matrix to be bi-diagonal.

Coxian models have numerical advantages in calculating distributional quantities, for example, the pdf of T and its state distribution at a given time. Their applications are extensive in many fields; see Asmussen et al. (1996); Asmussen (1992) and Marshall and Zenga (2009b). Compared with the phase-type models, a Coxian model with m transient states has $2m - 1$ parameters, while a phase-type model with m transient states has $m^2 + m$ parameters. Since the number of parameters is reduced significantly, the Coxian models are preferable in most applications, except for cases having explicit requirements in the underlying Markov chain.

One modification of the Coxian models is to impose a structure on the transition rates. For example, Faddy and McClean (1999) extended the Coxian models by involving regressions with covariates in the transition rates. Recall that the the human ageing process' characteristics are progressive and essentially irreversible. Accordingly, the states in the human ageing process cannot be revisited once the individual has left them. Furthermore, the progressiveness requires that the absorption rates cannot change dramatically when moving to other states. As a result, the transition out of the present state can move to either the absorbing state (death) or the next transient state.

Definition *For $m = 1, 2, \ldots,$ denoting the absorbing state $m + 1$, let T have an associated m-state phase-type distribution with a $1 \times m$ initial probability vector*

$$\boldsymbol{\alpha} = (1, 0, \ldots, 0),$$

and the $m \times m$ degenerated transition matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix} -(\lambda_1 + h_1) & \lambda_1 & & & \\ & -(\lambda_2 + h_2) & \lambda_2 & & \\ & & \ddots & & \\ & & & -(\lambda_{m-1} + h_{m-1}) & \lambda_{m-1} \\ & & & & -h_m \end{bmatrix},$$

where $\lambda_i > 0$ and $h_i > 0$ for all $i \in E$. Then, T follows a Coxian distribution.

The Coxian model is illustrated in Figure 2.2. Since the process only starts from state 1, the survival function of the Coxian model only depends on the first row of the $m \times m$ transition matrix $\boldsymbol{P}(t)$, which is denoted by $\mathbf{p}_1(t)$. Denote the kth element of $\mathbf{p}_1(t)$ by $P_{1k}(t)$ or simply $P_k(t)$.

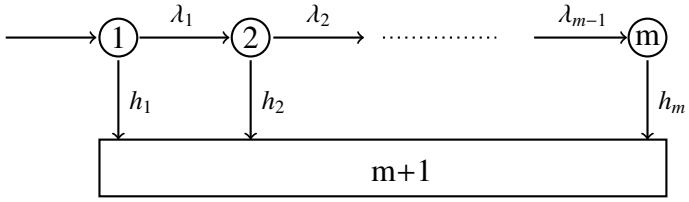

Figure 2.2: Diagram for a Coxian-type Markovian process.

Then, the Kolmogorov forward equation is given by

$$\begin{cases} \frac{\mathrm{d}P_1(t)}{\mathrm{d}t} & = -(\lambda_1 + h_1)P_1(t); \\ \frac{\mathrm{d}P_k(t)}{\mathrm{d}t} & = \lambda_{k-1}P_{k-1}(t) - (\lambda_k + h_k)P_k(t),\ k = 2, 3, \ldots, m-1; \\ \frac{\mathrm{d}P_m(t)}{\mathrm{d}t} & = \lambda_{m-1}P_{m-1}(t) - h_m P_m(t). \end{cases} \tag{2.3.9}$$

The initial conditions are $P_1(0) = 1$ and $P_k(0) = 0,\ k = 2, \ldots, m$. This tells us that

$$P_1(t) = e^{-(\lambda_1 + h_1)t}.$$

In addition, $P_k(t)$ for $k = 2, \ldots, m$ can be obtained iteratively, yielding

$$P_k(t) = \sum_{j=1}^{k} \frac{(-1)^{k-1}\lambda_1 \ldots \lambda_{k-1}}{\prod_{s=1, s\neq j}^{k}(\lambda_j + h_j - \lambda_s - h_s)} e^{-(\lambda_j + h_j)t},\ k = 2, \ldots, m, \tag{2.3.10}$$

by defining $\lambda_m = 0$.

The denominator of (2.3.10) is equal to 0 when $k = 1$, at which the formula is not well-defined. We define $\frac{(-1)^{k-1}\lambda_1 \ldots \lambda_{k-1}}{\prod_{s=1, s\neq j}^{k}(\lambda_j + h_j - \lambda_s - h_s)} = 1$ when $k = 1$ so that the result in (2.3.10) also gives $P_1(t)$. Therefore, the survival function of T is given by

$$S(t) = \sum_{k=1}^{m} P_k(t) = \sum_{k=1}^{m}\sum_{j=1}^{k} \frac{(-1)^{k-1}\lambda_1 \ldots \lambda_{k-1}}{\prod_{s=1, s\neq j}^{k}(\lambda_j + h_j - \lambda_s - h_s)} e^{-(\lambda_j + h_j)t}. \tag{2.3.11}$$

Remark 2.2. *Li and Ng (2008) derived a formula to calculate the probability of being in state k at time t for a Coxian model using the same approach above by solving the Kolmogorov forward equation. However, their formula for $P_k(t)$ is equal to the difference of two sums, specifically,*

$$P_k(t) = \lambda_{k-1}\sum_{j=1}^{k-1} \frac{C_{k-1,j}e^{-(\lambda_j + h_j)t}}{\lambda_k + h_k - \lambda_j - h_j} - \lambda_{k-1}\sum_{j=1}^{k-1} \frac{C_{k-1,j}e^{-(\lambda_k + h_k)t}}{\lambda_k + h_k - \lambda_j - h_j}, \tag{2.3.12}$$

and, $C_{11} = 1$*,*

$$C_{ij} = \begin{cases} \lambda_{i-1} C_{i-1,j}/(\lambda_i + h_i - \lambda_j - h_j), \text{ when } j < i \\ -\lambda_{i-1} \sum_{j=1}^{i-1} C_{i-1,j}/(\lambda_i + h_i - \lambda_j - h_j), \text{ when } j = i \\ 0, \text{ when } j > i \end{cases}$$

Since both (2.3.12) and (2.3.10) are the solutions of the Kolmogorov forward equation, the two formulas are equivalent; both formulas can be expressed as $P_k(t) = \sum_{j=1}^{k} a_j e^{-(\lambda_j + h_j)t}$*. However, there are two subtle differences: firstly, (2.3.10) is more compact than (2.3.12); secondly, (2.3.10) may be more numerically stable to evaluate when the two sums in (2.3.12) are relatively large, but then* $P_k(t)$ *is only relatively small.*

When the dimension of the Coxian model is small, and the parameter values are far enough apart so that the denominator of (2.3.10) is not too small, equation (2.3.10) may be used to evaluate the distribution quickly. However, in lifetime modeling, the model to be used may contain a large number of states with the λ_is and h_is being very close to each other. Using (2.3.10) causes numerical problems because the values of each term in summation are large and have alternating signs.

Coxian models require a substantially smaller number of parameters than the general phase-type models when the total number of states is fixed. Hence, most modelers prefer Coxian models for modeling lifetime random variables. Interestingly, due to the restrictions on the transitions, it usually requires more states for a Coxian model to achieve a good fit to life data than it does for a general phase-type model (Fackrell, 2009).

2.4 Phase-type aging model

Lin and Liu (2007) proposed a specific Coxian model, called the Phase-Type ageing Model (PTAM), to fit the observed mortality data and link the parameters to the physiological mechanism of ageing. In their model, they incorporated the development of ageing periods in the human ageing process. In this thesis, we focus on the ageing period.

The assumptions of the PTAM in Lin and Liu (2007) are:

- the initial probability $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$, assuming each individual starts at state 1;
- the transition rates between states are constant, $\lambda_i = \lambda$, assuming the physiological variables have linear declines agewise on average;
- the absorption rates, ignoring the behaviour-related accident rate, follow a power function $h_i = b + i^p q$;
- the total number of states is fixed.

As a result, they simplified the Coxian model to a four-parameter PTAM with a parameter set (b, p, q, λ). They utilised the progressiveness and irreversibleness of the Coxian-type Markov chains to mimic the ageing process. They labeled the states with integer numbers starting from 1 and interpreted the states as physiological ages. Their model was calibrated using the

Swedish cohort data from the year 1811 to 1911, and an excellent fit to the mortality data was achieved except for extremely old ages. Also, their calibrated model can capture trends in mortality across birth cohorts. Additionally, using (2.3.11) and (2.3.10) allows one to calculate the associated survival distributions and the probabilities in any states at any time easily.

Govorun et al. (2018) extended the PTAM with health information to narrow the state variability at any given time. Their model has similar structures on the transition rates and absorption rates as Lin and Liu's.

The merits of PTAMs in modeling the human ageing process have been demonstrated considering the intuitive model interpretation and promising fitting results. The Markovian structure is suitable in portraying the progressive and essentially irreversible characteristics of the ageing process. Certainly, the individual's health history has impact on their current health status. However, it could be argued that the current health status already encapsulated the information from the past health status. Thus, on the basis of the current health status, the distribution of the future status can be estimated, and this is where we could see the appropriateness of the Markovian modelling framework. Needless to say, the simplicity of this framework enable a much easier calibration than those entailed by other modelling frameworks. The Markovian structure can be understandably interpreted as an imitation of the ageing process, whilst the resulting distribution can be intelligibly interpreted as the lifetime distribution. A recent study on modeling human mortality using the general phase-type models is elaborated in Asmussen et al. (2019).

Nonetheless, the PTAMs also have some drawbacks. For instance, the modelers have to determine judiciously the total number of states and the structure of the absorption rates before other parameters could be estimated. In Lin and Liu (2007) and Govorun et al. (2018), power functions were used for the absorption rates with a total number of states $m = 200$, whilst Su and Sherris (2012) used exponential functions with a total number of states $m = 100$. Since the applications of PTAMs require a large number of states, typically more than 50, PTAMs' calibration is computationally expensive. Certainly, the value of m and the structure on the absorption rates must be carefully selected.

Chapter 3

Phase-type ageing model

In this Chapter, we investigate special models following the Coxian structure. We propose a family of special Coxian distributions, called the generalised phase-type ageing model (GPTAM), by assuming the transitions from any transient state to the next transient state having the same rate. One of the parameters in the GPTAM is the transition rate, and the remaining parameters are the absorption rates in all transient states. We can show that this family of distribution is flexible to approximate any continuous non-negative-valued distribution. However, it is challenging to estimate the parameters of a GPTAM due to its large number of parameters.

We, therefore, propose a structure on the absorption rates that will reduce the number of parameters. The proposed structure is a reminiscent of the Box-Cox transformation, which provides some degrees of flexibility for the absorption rate to achieve a variety of patterns. The proposed PTAM has 5 parameters, making it manageable for calibration. Additionally, the proposed PTAM inherits the Coxian structure, which requires the state progression can only move forward or exit into the absorbing state. Such a structure matches the characteristics of the ageing process in our modelling context. Two numerical applications are included to demonstrate the application of our proposed model. Furthermore, we use some numerical examples to illustrate that the resulting probability distribution of the proposed model can approximate a variety of lifetime distributions reasonably well.

3.1 Modelling background

This thesis takes the perspective that ageing is progressive and essentially irreversible. Viewing the absorbing state as death and its transient states as intermediate statuses of the ageing process, the PTAM's Markovian structure matches our prior knowledge of ageing – progressive and essentially irreversible. In lieu of validation using a data set, which is a challenge to observe and choose the right proxy for the ageing process, our Markovian assumption is simply made on the basis of a generally accepted principle that an ageing process progresses irreversibly and ending in death eventually. Furthermore, some characteristics of the PTAM are as follows:

- The lifetime of an individual is the sum of the sojourn times in each state before absorption;

- For a given realisation of the process, the probability of being in each state is easily determined using the Markov-chain theory;
- Given the current state, the conditional lifetime distribution is determined by another PTAM, although the current state is usually unobservable.

Let us define the transition rates between the transient states as the ageing rates, and the absorption rates as the dying rates. The labeled number for each state determines the physiological age. The definition of physiological age will be addressed later.

Lin and Liu (2007) assumed the dying rates have a power functional form with an increasing number of states. They defined the states as the physiological ages. They argued that $m = 200$ should be large enough to have an excellent fit to the Swedish mortality data. Govorun et al. (2018) assumed a different functional form for the dying rates, which involves a power function in the later states. They believed $m = 230$ is the sufficient number to achieve a good fit for the Canadian male mortality data. The total number of states is reasonably large in both papers, and the parameter estimation is carried out via the least-squared-error or maximum-likelihood approach.

Both models were calibrated with mortality data, although they are ageing models and not mortality models. Furthermore, the model calibration in Govorun et al. (2018) attempted to refine the estimation by incorporating health-related information.

The structure of the dying rate for each model is usually determined by the modeler's beliefs. Furthermore, the total number of states is also fixed by some prior knowledge. Therefore, it is natural to ask the following questions:

- Can the lifetime data itself determine the monotonic pattern and curvature on the dying rates?
- Is it possible to estimate m from the lifetime data?
- Is there a simple and smooth structure on the dying rate such that the corresponding estimation is relatively simple, yet the resulting lifetime distribution is flexible enough to achieve a variety of lifetime distributions?

3.2 General phase-type ageing model

Before introducing our proposed model, we would like to generalise the phase-type ageing models in Lin and Liu (2007) and Govorun et al. (2018).

Definition 3.1. *The general phase-type ageing model (GPTAM) is a phase-type model satisfying three conditions:*

- *Every individual starts in state* 1 *at time* 0*;*
- *The transition of going out from each state must be either to the next transient state or to the absorbing state;*
- *The rates of transition, which are ageing rates, from one transient state to the next transient state are the same.*

The GPTAM is a subclass of Coxian models with the restriction $\lambda_i = \lambda$, for all $i = 1, 2, \ldots, m-1$. With the above-mentioned restrictions, the GPTAM assumes that the ageing rate is uniform over time, so that the rate of increase in physiological age is uniform. This makes some sense because calendar age increases uniformly. Although the assumed ageing rate is constant, there is variability in transition times. Therefore, individuals at the same calendar age may be in different states, representing different physiological ages. This is the same idea used by Lin and Liu (2007) in which the ageing-related transition intensity parameter is specified as a constant. To capture the mortality pattern at infant ages, Lin and Liu (2007) allowed λ_i to vary so that the non-ageing-related mortality causes can be accommodated. However, in our models, we focus on modeling ageing rather than mortality fitting. Given this focus, our models are most applicable at human ages beyond the attainment of adulthood.

There is an intuitive way to understand the concept of GPTAM. Let us consider each state in the GPTAM as the hypothetical year similar to the chronological year. The time spent in each state is similar to the days spent in each chronological year. It is clear that there are 365(366) days in each chronological(leap) year, whilst the time spent in each hypothetical year is random but expected to be the same on average. The force of mortality is different at various chronological times, while the force of dying in each hypothetical year(state) is a constant, but vary in different hypothetical years. Therefore, in the GPTAM the randomness in the time of death is decomposed into two parts: the randomness of time in each hypothetical year (dependent on the force of ageing) and the randomness of death in each hypothetical year (dependent on the force of dying). The GPTAM differentiates two forces, the force of ageing and the force of dying, which mutually determine the force of mortality.

We call a GPTAM with m transient states an $m-$state GPTAM, labeling the absorbing state $m+1$. The Markov chain for an $m-$state GPTAM is graphically exhibited in Figure 2.2. The transition rates $h_i,\ i = 1, \ldots, m$, are the dying rates and the transition rate λ is the ageing rate.

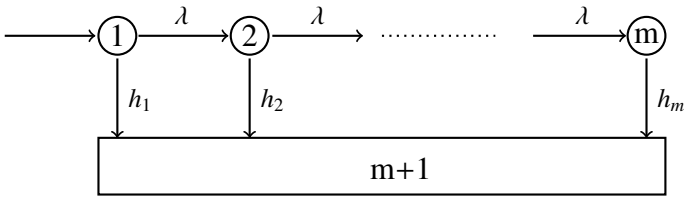

Figure 3.1: State transition diagram for GPTAM.

The resulting pdf and survival function at time t are

$$f(t) = \sum_{k=1}^{m} P_k(t)h_k, \text{ and } S(t) = \sum_{k=1}^{m} P_k(t),$$

respectively, where $P_k(t)$ is the probability in state k at time t. The survival probability at time t is the total probability in transient states at time t, and the hazard rate, $f(t)/S(t)$, at time t is a weighted average of h_k's. The weight in state k is equal to $P_k(t)/S(t)$. Solving the Kolmogorov

forward equation (2.3.9), we get

$$P_k(t) = \sum_{j=1}^{k} \frac{(-\lambda)^{k-1}}{\prod_{s=1,s\neq j}^{k}(\lambda_j + h_j - \lambda_s - h_s)} e^{-(\lambda_j+h_j)t},\ k = 1,\ldots,m, \tag{3.2.1}$$

where $\lambda_i = \lambda$ for $i \neq m$, $\lambda_m = 0$, and $\prod_{s=1,s\neq j}^{k}(\lambda_j + h_j - \lambda_s - h_s) = 1$ for $k = 1$.

Remark 3.1. *The survival function $S(t)$ can be viewed as a weighted average of the survival function for exponential distributions with rate $\lambda_i + h_i$ for $i = 1,\ldots,m$ if negative weights are allowed. This is easy to verify by the fact that $S(0) = 1$, yielding the sum of coefficients equal to* 1.

Equation (3.2.1) is useful in the evaluation of the state distribution at any time $t \geq 0$, when the total number of states is small and $h_1,\ldots,h_m$ are sufficiently far apart that the denominator of (3.2.1) is not too small. It is worth noting that (3.2.1) may not be robust in the numerical calculation when m is large or the denominator is too small. The discussion of a more robust method to numerically calculate the probability distribution is deferred until Chapter 5. Consider the set of the resulting lifetime distribution of all GPTAMs. Then, any non-negative-valued distribution can be approximated well by a GPTAM.

Theorem 3.1. *Let $\mathcal{F}$ be the distribution family of all GPTAMs whose absorbing rate in state i is equal to $h_i \geq 0$ for $i = 1,\ldots,m$; transition rate from state j to $j+1$ is equal to λ for $j = 1,\ldots,m-1$; and the total number of states m is a positive integer. Given a non-negative-valued distribution with domain $(0,T)$, assume its survival function $S(t)$ is continuous. For any $\epsilon > 0$, there is a GPTAM in $\mathcal{F}$ such that its resulting survival function $S^*(t)$ satisfies $|S^*(t) - S(t)| < \epsilon$ for any $t \in (0,T)$. Therefore, $\mathcal{F}$ is dense in the field of all continuous non-negative-valued distributions.*

Proof. See Appendix A.1. ∎

The distribution family of GPTAMs is flexible enough to represent any lifetime distribution. The goodness of fit to any lifetime data is guaranteed almost surely.

If we assume the lifetime distribution is the observed information from an ageing process, then one can find an equivalent GPTAM, whose lifetime distribution is close to the observed lifetime distribution. Meanwhile, we can interpret the lifetime distribution from the ageing perspective, e.g. the associated ageing/dying rate pattern. By doing so, it provides a more intuitive perspective to explain the lifetime distribution – interpreting the lifetime as the terminal time of an ageing process.

Let Y_t be the state variable at time t for any $t > 0$. The domain of Y_t is the set $\{1,\ldots,m+1\}$, where states $1,\ldots,m$ are the transient (alive) states and state $m+1$ is the absorption (death) state. The random variable $Y_t|Y_t \in E$ represents the state occupied at time t given the individual is in the transient states. We can prove that $Y_t|Y_t \in E$ is increasing with respect to t in the stochastic order by Lemma 3.1.

Definition 3.2. *A random variable Y_t is increasing with respect to t in stochastic order if Y_{t_1} is less than Y_{t_2} for any $t_1 < t_2$, or $P(Y_{t_1} > y) \leq P(Y_{t_2} > y)$ for any y in the domain.*

Lemma 3.1. *Consider an $m-$state GPTAM with transition rate from one transient state to the next transient state equal to λ, and the absorption rate in state i is h_i for $i = 1, \ldots, m$. For any $t > 0$, let Y_t be the state variable at time t, then $P(Y_t \geq k | Y_t \in E)$ is an increasing function with respect to t for any $k = 1, \ldots, m$.*

Proof. The proof is in Appendix A.2.

■

By Lemma 3.1, $P(Y_{t_1} \geq k | Y_{t_1} \in E) < P(Y_{t_2} \geq k | Y_{t_2} \in E)$ for any $0 \leq t_1 < t_2$ and any $k = 1, \ldots, m$. Equivalently, $(Y_{t_1} | Y_{t_1} \in E)$ is less than $(Y_{t_2} | Y_{t_2} \in E)$ in stochastic order. In other words, the alive individual is more likely in the later states as time passes.

Example 3.1. Consider a GPTAM with $m = 100$, $\lambda = 3$, and $h_i = 0.005i$ for $i = 1, \ldots, 100$, $i \neq 50$ and $h_{50} = 0.1$. The state distributions, given that the individual being alive at various times, are plotted in Figure 3.2.

The top graph shows that the mode of the state distribution shifts to a higher state as time passes by, and the bottom graph shows that the probability $P(Y_t \geq k | Y_t \in E)$ is greater for a larger t for any k. The intuitive explanation for Lemma 3.1 is that each individual in the ageing process cannot reverse the process. This property matches our belief on the progressiveness

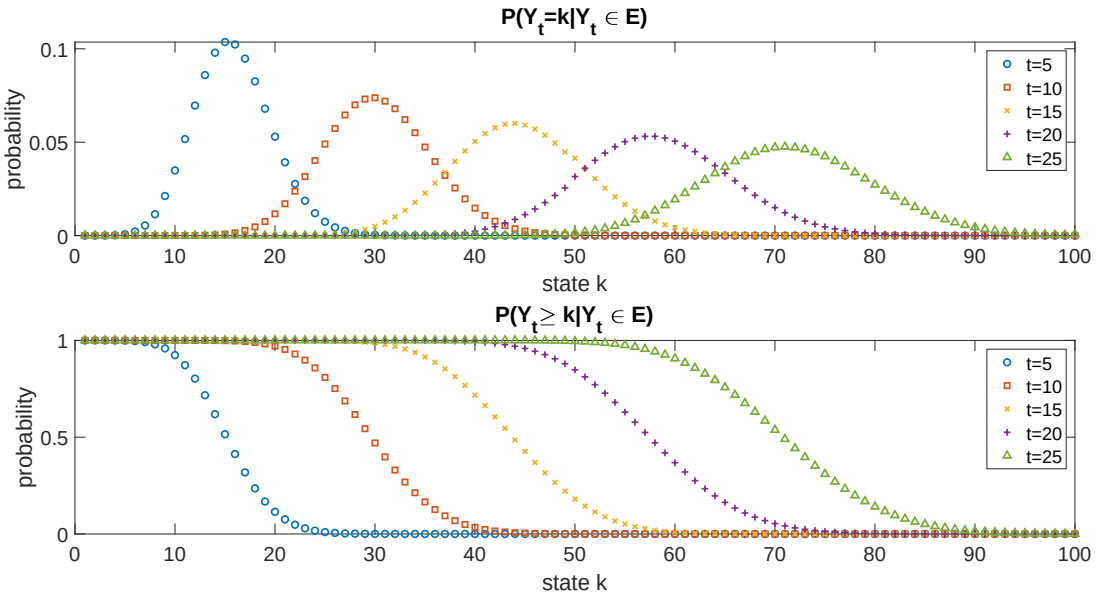

Figure 3.2: The state distribution at various times t for a GPTAM.

and irreversibleness of the ageing process – the process has to progress to the later states as time passes, and it cannot reverse the process back to the early states.

△

Since the dying rate h_i for the GPTAM is flexible, let us assume h_i is monotone. Then, we can establish the relation between the pattern of the dying rate h_i and the resulting pattern of the hazard rate $h(t)$. But, first we need the following result.

Theorem 3.2. *Suppose the absorption rate of a GPTAM is monotone with respect to state label. The resulting hazard rate $h(t)$ of the GPTAM is increasing with respect to age t if and only if the dying rate h_i is increasing with respect to i.*

Proof. For any $t \geq 0$, the resulting hazard rate is a weighted average of the dying rates in each transient state by (2.2.4). Note that

$$h(t) = \frac{\boldsymbol{\alpha} \exp(\boldsymbol{\Lambda} t)}{\boldsymbol{\alpha} \exp(\boldsymbol{\Lambda} t)\boldsymbol{e}} \boldsymbol{h} = \sum_{i=1}^{m} h_i P(Y_t = i | Y_t \in E) = \mathbb{E}\left(h_{Y_t} | Y_t \in E\right),$$

which shows $h(t)$ is the expected value of $(h_{Y_t}|Y_t \in E)$. The ith element of $\boldsymbol{\alpha} \exp(\boldsymbol{\Lambda} t)$ is the probability in state i at time t.

By Lemma 3.1, for any $t_1 < t_2$, we have that Y_{t_1} is less than Y_{t_2} in stochastic order. It is well known that A is less than B in stochastic order, then for any non-decreasing functions u, $\mathbb{E}(u(A)) \leq \mathbb{E}(u(B))$. Hence, when h_i is increasing with respect to i, for any $0 \leq t_1 < t_2$,

$$h(t_1) = \mathbb{E}\left(h_{Y_{t_1}} \middle| Y_{t,1} \in E\right) \leq \mathbb{E}\left(h_{Y_{t_2}} \middle| Y_{t_2} \in E\right) = h(t_2),$$

which yields that $h(t)$ is increasing with respect to age t.

On the other hand, when $h(t)$ is increasing with respect to t, for any $0 \leq t_1 < t_2$,

$$\mathbb{E}\left(h_{Y_{t_1}} \middle| Y_{t_1} \in E\right) = h(t_1) < h(t_2) = \mathbb{E}\left(h_{Y_{t_2}} \middle| Y_{t_2} \in E\right). \tag{3.2.2}$$

Suppose h_i is decreasing with respect to i, then $-h_i$ is an increasing function. We have shown that Y_{t_1} is less than Y_{t_2} in stochastic order. Therefore,

$$\mathbb{E}\left(-h_{Y_{t_1}} \middle| Y_{t_1} \in E\right) \leq \mathbb{E}\left(-h_{Y_{t_2}} \middle| Y_{t_2} \in E\right),$$

yielding $\mathbb{E}\left(h_{Y_{t_1}} \middle| Y_{t_1} \in E\right) \geq \mathbb{E}\left(h_{Y_{t_2}} \middle| Y_{t_2} \in E\right)$, or $h(t_1) \geq h(t_2)$, which contradicts with $h(t)$ is increasing. Therefore, the dying rate h_i is increasing with respect to i.

■

Remark 3.2. *Suppose the absorption rate of a GPTAM is monotone with respect to state label. Then, the resulting hazard rate $h(t)$ is decreasing with respect to age t if and only if the dying rate h_i is decreasing with respect to i.*

Remark 3.3. *Without the monotone assumption, it is still true that $h(t)$ is increasing given h_i is increasing. However, the inverse is not true, i.e., h_i may not be necessarily increasing given $h(t)$ is increasing. In Example 3.1, the dying rate and resulting hazard rate are displayed in Figure 3.3. This is a counterexample showing that the dying rate is not monotone, yet the hazard rate is increasing as time increases. The plateau of $h(t)$ is the hazard rate's limit,* $\lim_{t\to\infty} h(t)$.

Theorem 3.2 provides a guideline on how to select the dying rate pattern for the GPTAM in order to achieve a monotonic hazard rate. For example, it is reasonable to choose an increasing dying rate pattern for the GPTAM to generate a lifetime distribution whose hazard rate is increasing.

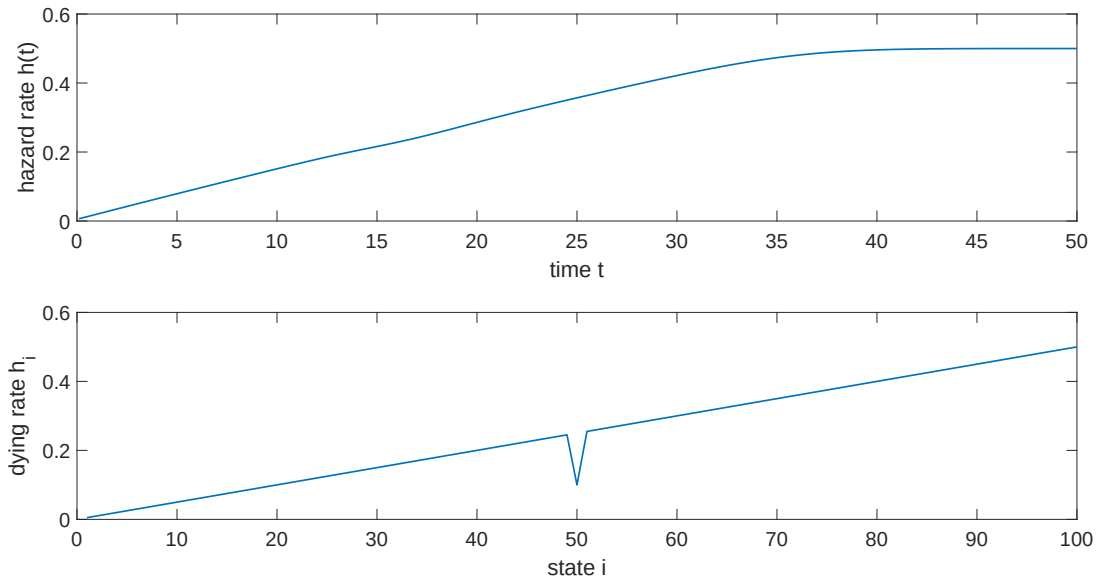

Figure 3.3: Resulting hazard rate and the dying rate for the GPTAM with $m = 100$, $\lambda = 3$, and $h_i = 0.005i$ for $\forall i,\ i \neq 50$ and $h_{50} = 0.1$.

The GPTAM is flexible, and it has a lower number of parameters compared with the Coxian models. However, the number of parameters for a GPTAM is still large when the model has a large number of states. Recall that an $m-$state GPTAM has $m+1$ parameters. Hence, when m is big, the number of parameters is large, and the model calibration is an extensive endeavour. It is impractical to find the global maximum likelihood estimate(s) because the parameter space is too huge to explore; see the curse of dimensionality as discussed in Bellman and Kalaba (1959). To make the parameter estimation feasible and stable, we need to impose a structure on the dying rate to reduce the number of parameters. Meanwhile, the proposed structure should achieve a variety of patterns to reflect some plausible ageing process patterns.

3.3 Our proposed PTAM

We choose an appropriate pattern for the dying rate for the GPTAM to achieve the observed hazard rate pattern. Let us first review some studies on hazard rate. Aalen (1994) claimed that hazard rate as a function of time is the result of selection effects due to both variation between individuals and variation within each individual over time. The selection effects are due to the fact that individuals with higher risks tend to die earlier, and those survivors will tend to be a selected group with lower risks. The Markov chain under the GPTAM can model well the variation between individuals (different individuals may be in various states) and the variation within each individual over time (the individual may progress to different states over time).

Aalen and Gjessing (2001) pointed out that progressive models for lifetime random variables would tend to have increasing hazard rates. Empirical studies of large populations of flies and humans have shown mortality plateaus at extremely old ages (Ricklefs, 1998; Liedo et al.,

1992; Fukui et al., 1993, 1996). Some lifetime variables may have decreasing hazard rate, e.g. infant lifetime (see mortality rate before age 10 in Heligman and Pollard (1980)). As a result, we only consider the lifetime distributions with monotonic hazard rate.

Let us assume that the proposed PTAM has a monotonic dying rate h_i as a function of i. Then, the associated hazard rate is monotone by Theorem 3.2. Furthermore, let us that assume $m > 3$ for two reasons. Firstly, there is no need to propose a structure when $m \leq 3$. Secondly, it requires a large number of states to reproduce the ageing process as in the case of the human ageing process.

We specify the form of the dying rates h_i. Suppose that h_1 and h_m are fixed. These are parameters to be estimated. We then require a smooth pattern of h_i values for i between 1 and m. We achieve this by letting

$$h_i^s = \frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s,$$

where s is a shape parameter that can control the convexity of the h_i pattern. In other words, the powers of h_i are obtained as a linear interpolation between the corresponding powers of h_1 and h_m. We then have

$$h_i = \left(\frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s \right)^{1/s}.$$

We can use this for any real number s except $s = 0$, when the expression is undefined. However, the limiting case as $s \to 0$ gives

$$h_i = h_1^{\frac{m-i}{m-1}} h_m^{\frac{i-1}{m-1}}.$$

In this case, $\log h_i$ is obtained as a linear interpolation between $\log h_1$ and $\log h_m$. Thus, the parameter s can take any real value. For $i = 1, 2, \ldots, m$, let

$$h_i = \begin{cases} \left(\frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s \right)^{1/s} & s \neq 0, \\ h_1^{\frac{m-i}{m-1}} h_m^{\frac{i-1}{m-1}} & s = 0. \end{cases} \tag{3.3.3}$$

Also, let $\beta = \frac{i-1}{m-1}$. Then we have $\beta \in [0, 1]$ and

$$h_i = \begin{cases} \left((1-\beta) h_1^s + \beta h_m^s \right)^{1/s} & s \neq 0 \\ h_1^{1-\beta} h_m^{\beta} & s = 0. \end{cases}$$

By differentiating $\log h_i$ with respect to s,

$$\frac{1}{h_i} \frac{\partial h_i}{\partial s} = \frac{\frac{(1-\beta) h_1^s \log h_1 + \beta h_m^s \log h_m}{(1-\beta) h_1^s + \beta h_m^s} s - \log \left((1-\beta) h_1^s + \beta h_m^s \right)}{s^2}.$$

Taking the limit as $s \to 0$ in the right hand-side of the previous equation,

$$\begin{aligned}
&\lim_{s\to 0} \frac{\frac{(1-\beta)h_1^s \log h_1 + \beta h_m^s \log h_m}{(1-\beta)h_1^s + \beta h_m^s} s - \log\left((1-\beta)h_1^s + \beta h_m^s\right)}{s^2} \\
=&\lim_{s\to 0} \frac{\left((1-\beta)h_1^s (\log h_1)^2 + \beta h_m^s (\log h_m)^2\right)\left((1-\beta)h_1^s + \beta h_m^s\right) - \left((1-\beta)h_1^s \log h_1 + \beta h_m^s \log h_m\right)^2}{2\left((1-\beta)h_1^s + \beta h_m^s\right)^2} \\
=&\frac{\left((1-\beta)(\log h_1)^2 + \beta(\log h_m)^2\right) - \left((1-\beta)\log h_1 + \beta \log h_m\right)^2}{2},
\end{aligned}$$

where the first equation holds by L'Hopital's Rule. This indicates the derivative of h_i with respect to s exists at $s = 0$ when h_1 and h_m are not equal to 0. This representation shows that h_i^s is a weighted average of h_1^s and h_m^s. This structure is reminiscent of the well-known Box-Cox transformation introduced by Box and Cox (1964). The Box-Cox transformation is a monotonic data transformation skewing the curvature by a smooth functional structure. Therefore, the proposed structure provides some curvature flexibility for the dying rate. Typically, the lifetime data have a linear or exponential patterns. Either pattern is a special case of the proposed structure. The structure of (3.3.3) can achieve both increasing and decreasing patterns of h_i. This is formally stated in the next theorem.

Theorem 3.3. *The dying rate h_i is an increasing function of i if and only if $h_1 < h_m$; h_i is a decreasing function if and only if $h_1 > h_m$; h_i's are identical if and only if $h_1 = h_m$.*

Proof. Suppose $h_1 = h_m = h$. It is trivial that $h_i = h$ for $i = 1, \ldots, m$.

Suppose $h_1 < h_m$. When $s > 0$, $h_1^s < h_m^s$. For any $1 \le i < j \le m$,

$$h_i^s = h_1^s p_1 + h_m^s (1 - p_1) < h_1^s p_2 + h_m^s (1 - p_2) = h_j^s,$$

where $p_1 = \frac{m-i}{m-1} > \frac{m-j}{m-1} = p_2$ and the inequality holds using Lemma 3.2. Therefore, $h_i < h_j$ using the fact that x^s is an increasing function of x when $x, s > 0$.

Similarly, when $s < 0$, $h_1^s > h_m^s$. For any $1 \le i < j \le m$,

$$h_i^s = h_1^s p_1 + h_m^s (1 - p_1) > h_1^s p_2 + h_m^s (1 - p_2) = h_j^s.$$

Hence, $h_i < h_j$ since x^s is a decreasing function of x when $x > 0$ and $s < 0$.

When $s = 0$,

$$\log(h_i) = \frac{\log(h_m) - \log(h_1)}{m-1} i + \frac{m \log(h_1) - \log(h_m)}{m-1},$$

which is an increasing function of i if and only if $h_1 < h_m$.

In summary, for any value of s, h_i is an increasing function of i if and only if $h_1 < h_m$. Likewise, one may prove that h_i is a decreasing function with respect to i if and only if $h_1 > h_m$. ∎

Lemma 3.2. *For any* $0 < a < b$ *and* $0 \leq p_2 < p_1 \leq 1$, *the inequality* $ap_1 + b(1 - p_1) < ap_2 + b(1 - p_2)$ *holds.*

Proof.

$$ap_2 + b(1 - p_2) - ap_1 - b(1 - p_1) = (p_2 - p_1)(a - b) > 0.$$

■

The structure given in (3.3.3) can achieve not only increasing and decreasing patterns but also a variety of curvature patterns. The parameter s is the shape parameter controlling the curvature. Note that, when $s = 1$, we have a linear pattern of h_i values with increasing i. When $s = 0$ we have an exponential pattern of h_i values. Furthermore, when $s < 1$, the pattern will be convex, and when $s > 1$, the pattern will be concave. Figure 3.4 shows the patterns of h_i values for $s = 0$, 1 and 2. For human ageing, we would expect that $s = 0$ is the most appropriate of the three values. The actual value is ideally estimated from ageing-related data. However, since such ageing-related information is hard to collect, the value may be estimated from lifetime data.

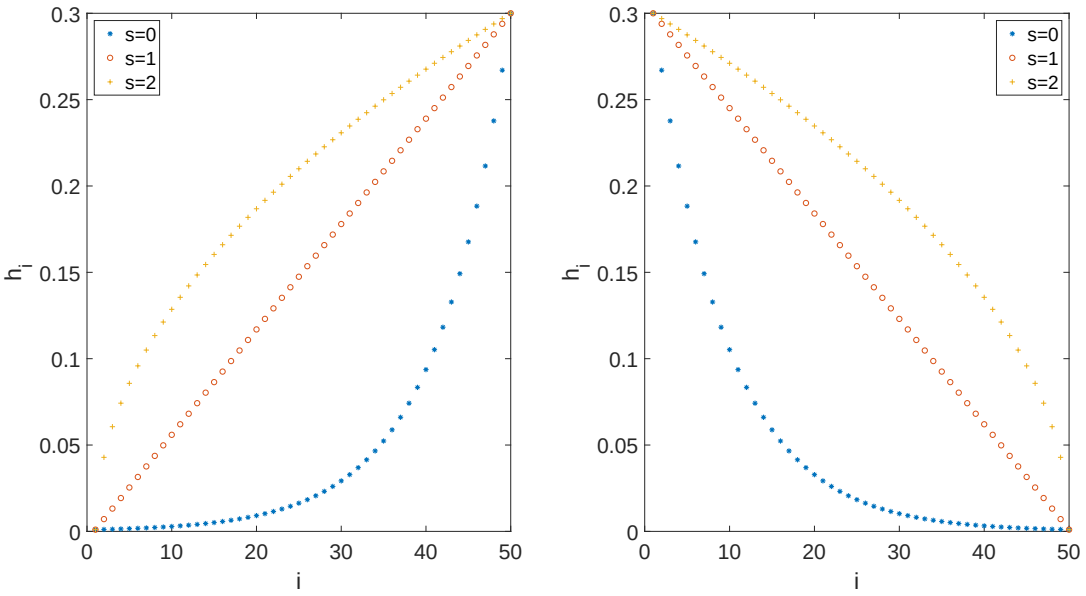

Figure 3.4: Left: Values of h_i determined using $h_1 = 0.001$, $h_m = 0.3$ and $s = 0$, 1 and 2. Right: Values of h_i determined using $h_1 = 0.3$, $h_m = 0.001$ and $s = 0$, 1 and 2.

According to the definition of h_i, it is a discrete function of i. To investigate how flexible the structure is, suppose h_i is a continuous function of i. Then, when $s \neq 0$, the first derivative of h_i^s with respect to i is

$$\frac{\partial h_i^s}{\partial i} = s h_i^{s-1} \frac{\partial h_i}{\partial i} = \frac{h_m^s - h_1^s}{m - 1},$$

yielding

$$\frac{\partial h_i}{\partial i} = \frac{h_m^s - h_1^s}{s(m-1)} h_i^{1-s}.$$

When $s = 1$, $\frac{\partial h_i}{\partial i} = \frac{h_m - h_1}{m-1}$; When $s \neq 1$, the second derivative of h_i with respect to i is

$$\begin{aligned}\frac{\partial^2 h_i^s}{\partial i^2} &= \frac{h_m^s - h_1^s}{s(m-1)}(1-s)h_i^{-s}\frac{\partial h_i}{\partial i} \\ &= \left(\frac{h_m^s - h_1^s}{s(m-1)}\right)^2 (1-s)h_i^{1-2s}.\end{aligned}$$

Hence, the second derivative is positive (h_i is convex) when $s < 1$; and the second derivative is positive (h_i is concave) when $s > 1$.

Recall from Chapter 2 that the ageing rate λ is the transition rate between transient states. The value of λ needs to be compatible with the value of m. We need the rate of progression through the states to be large enough that some individuals will reach state m, or there is no need to have as many as m states. However, we want only a small proportion of individuals to survive to state m, or the model hazard rate will flatten out at older ages. On the other hand, the mean spending time in each transient state is around $1/\lambda$ when $\lambda >> h_i$. Hence, the expected time being in the system for the individual who can reach state m is about m/λ. We therefore let

$$\lambda = m/\psi, \tag{3.3.4}$$

where ψ can be thought of as the life span of individuals in the population of interest. The parameter ψ need not be a limiting age, as there is no limiting age in our model. However, ψ should be a high age at which only a very small proportion of individuals will survive. This parameter can be estimated from the data or chosen based on some prior opinion. For human lifetimes, a value of ψ between 100 and 120 may be reasonable.

Remark 3.4. *It is plausible to extend the proposed PTAM with a λ that has a more generalised setting. This allows the PTAM to gain greater flexibility, although it could complicate the model calibration. Hence, there is a need to balance the trade off between flexibility and inferential power; see Section 6.2.*

In summary, we propose a Coxian distribution with five parameters to be specified: m, ψ, h_1, h_m and s. The parameters m and ψ may be selected based on some prior information of the modeller, or a combination of lifetime/ageing-related data and prior knowledge. Having established the values of m and ψ, λ is determined by (3.3.4), and h_1, h_m and s are easily estimated from lifetime data. We will show two applications before providing the details of the parameter estimation. A deeper understanding of the PTAM is facilitated by numerical examples, which are featured prior to delving into the estimation details. More discussion of the estimation of m and ψ will be provided in Chapter 5.

Ideally, our ageing model would be calibrated using ageing-related data, rather than lifetime data. This means that if the observations for one or more key ageing variables are taken at one

or more times whilst each individual is alive in addition to ages at death, we could estimate the model parameters with less uncertainty. The estimation incorporating with ageing-related data will be addressed in Chapter 6.

It is worth noting that the proposed PTAM can reproduce some special distributions by setting the parameter values. The special cases are summarised in the succeeding remarks.

Remark 3.5. *When $h_1 = h_m = \mu$, formula (3.3.3) yields $h_i = \mu$ for $i = 1, \ldots, m$. The resulting survival function is $S(t) = e^{-\mu t}$, in which the proposed PTAM degenerates to an exponential distribution with rate parameter μ. In this case, there is no need to have more than 1 state in the Markov chain.*

Remark 3.6. *For any $m \geq 2$, when $h_1 \neq h$ and $h_2 = \ldots = h_m = h$, the resulting probability distribution of the GPTAM is the same as the resulting probability distribution of a 2-state Markov model with the 2×2 transition matrix*

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda_1 + h_1) & \lambda_1 \\ & -h \end{bmatrix}.$$

This can be verified by the fact that

$$f(t) = \sum_{k=1}^{m} P_k(t)h_k = P_1(t)h_1 + \sum_{k=2}^{m} P_k(t)h = P_1(t)h_1 + h\sum_{k=2}^{m} P_k(t).$$

By treating states 2 to m as simply a hypothetical state 2^*, the probability in state 2^* at time t is $\sum_{k=2}^{m} P_k(t)$. Therefore, the probability distribution $f(t)$ is equal to that of the 2-states Markov model. As a result, the survival function and hazard function are also the same between the two Markov models.

Remark 3.7. *As $s \to +\infty$ and $h_1, h_m \neq 0$, formula (3.3.3) gives that $h_i = \max(h_1, h_m)$ for $i = 2, \ldots, m-1$, where*

$$\begin{aligned}
\lim_{s\to+\infty} h_i &= \lim_{s\to+\infty} \left(h_1^s \frac{m-i}{m-1} + h_m^s \frac{i-1}{m-1} \right)^{\frac{1}{s}} \\
&= \max(h_1, h_m) \lim_{s\to+\infty} \left(\left(\frac{h_1}{\max(h_1, h_m)} \right)^s \frac{m-i}{m-1} + \left(\frac{h_m}{\max(h_1, h_m)} \right)^s \frac{i-1}{m-1} \right)^{\frac{1}{s}} \\
&= \max(h_1, h_m).
\end{aligned}$$

Furthermore, when $h_1 < h_m$, by Remark 3.6, the resulting probability distribution is equal to the probability distribution of a 2-state Markov model with the 2×2 transition matrix

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ & -\max(h_1, h_m) \end{bmatrix} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ & -h_m \end{bmatrix},$$

and when $h_1 > h_m$, the resulting probability distribution is equal to the probability distribution of a 2-state Markov model with the 2×2 transition matrix

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda + \max(h_1, h_m)) & \lambda \\ & -h_m \end{bmatrix} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ & -h_m \end{bmatrix}.$$

So, the resulting probability distribution is equal to the probability distribution of a 2-states Markov model with the 2×2 *transition matrix:*

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ & -h_m \end{bmatrix}.$$

Remark 3.8. *As* $s \to -\infty$ *and* $h_1, h_m \neq 0$*, formula (3.3.3) provides that* $h_i = \min(h_1, h_m)$ *for* $i = 2, \ldots, m-1$*, where*

$$\begin{aligned} \lim_{s\to-\infty} h_i &= \lim_{s\to-\infty} \left(h_1^s \frac{m-i}{m-1} + h_m^s \frac{i-1}{m-1} \right)^{\frac{1}{s}} \\ &= \min(h_1, h_m) \lim_{s\to-\infty} \left(\left(\frac{h_1}{\min(h_1, h_m)} \right)^s \frac{m-i}{m-1} + \left(\frac{h_m}{\min(h_1, h_m)} \right)^s \frac{i-1}{m-1} \right)^{\frac{1}{s}} \\ &= \min(h_1, h_m). \end{aligned}$$

Moreover, by Remark 3.6, the resulting probability distribution is equal to the probability distribution of a 2-state Markov model with the 2×2 *transition matrix*

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ & -h_m \end{bmatrix}.$$

Remark 3.9. *When* $h_1 = 0$*,* $h_m = \lambda$ *and* $s = 0$*, we have* $h_i = h_1^{\frac{m-i}{m-1}} h_m^{\frac{i-1}{m-1}} = 0$ *for* $i = 1, \ldots, m-1$*. There is no early exit until the last state and the transition rate in each state is equal to* λ*, in which the proposed PTAM is equivalent to a Gamma distribution with a shape parameter* m *and a rate parameter* λ*, which is also called an Erlang distribution (Erlang, 1917). The corresponding pdf is*

$$f(t) = \frac{\lambda^m}{(m-1)!} t^{m-1} e^{-\lambda t}.$$

3.4 Empirical implementation of the proposed PTAM

The proposed PTAM is used for modelling the ageing processes. We will demonstrate two applications of the proposed PTAM in this Section. For simplicity, we call the proposed PTAM "the PTAM" or "our model". We will address the estimation details in Chapter 5.

3.4.1 Application using data from a retirement community

To illustrate how our ageing model can be calibrated using lifetime data, we consider a data set from the Channing House, a retirement community in Palo Alto, California. The data set includes the entry age and age at death (or the date the study ended) for 462 people (97 males and 365 females) who resided in the facility between January 1964 and July 1975. These residents were covered by a health care program, which provided easy access to care at no cost. This may have resulted in lower than average mortality. We fit our ageing model to

the female data only. Thus, we consider data on a relatively small group of homogeneous individuals - all females living in the same community with the same access to health care and, very likely, similar lifestyles. This is not ideal, but this is the best we can do to ensure all other variables that are likely to affect death are as similar as possible among individuals, and the only variable that we cannot control is the underlying ageing process.

Only 362 of the 365 female records were kept, because three records had equal entry and exit ages. Of the 362 females, 130 died whilst in the community, and the other 232 survived until the end of the observation period. The youngest entry age was just over age 61, and the oldest exit age was just under age 101. So the data pertain to ageing over this age range.

Since our model assumes that all individuals start at state 1, but we expect some variability in states by age 61, we assume for this example that the ageing process starts at age 50. This allows us to achieve some variability in the state distribution by age 61. The variability in the state distribution is similar by assuming that the ageing process starts at age 50 or some earlier ages. In particular, a starting age is chosen such that it is close to the youngest entry age in the data set. It is more fitting to assume that the ageing process begins at young ages if some individuals enter the observation study in their youth. Also, we assume, somewhat arbitrarily, that age 105 is the end of the life span. This meets our condition that it should be a high age to which only a very small proportion of individual will survive. In fact, none in our small sample of females were observed to live older than 105. All members of the community were dead or exited by this age. Therefore, $\psi = 105 - 50 = 55$. Also, based on experience, we believe that having $m = 100$ ageing states is appropriate for modelling ageing above age 50. This results in a reasonable amount of variability in the physiological age at various calendar ages, as shown later in Figure 3.10. We then have $\lambda = m/\psi = 100/55 = 1.8182$. Using the maximum likelihood estimation, we obtain the results in Table 3.1.

Table 3.1: Maximum likelihood estimates of parameters along with the approximate 95 percent confidence intervals (based on the profile log-likelihood) for the Channing House data set.

		Approximate 95% CI	
Parameter	Estimate	Lower	Upper
h_1	0.0017	0.0000	0.0099
h_m	1.2750	0.2770	∞
s	-0.0735	-0.5270	0.3550

In Table 3.1, we present approximate confidence intervals for the parameter estimates rather than standard error estimates. This is because asymmetries of the log-likelihood function indicate that the estimates are not approximately normal (Figure 3.5). The asymmetries may be due to the strong correlation amongst h_1, h_m and s; moreover, 100 estimates are not enough to approximate the estimator distributions. Therefore, standard errors would not be easy to interpret. Furthermore, our confidence intervals are based on the profile log-likelihood functions; these particular confidence intervals are known to perform well when the log-likelihood function is asymmetric; see Figure 3.6.

The profile log-likelihood is constructed by fixing one parameter and then maximising the likelihood with respect to the other parameters. In our model, we maximise the likelihood for

each fixed value of h_1, fixed value of h_m and fixed value of s. The 95% confidence interval is determined by the likelihood-ratio test. An approximate 95% confidence interval for θ is the set of values satisfying:

$$l(\hat{\theta}) - l(\theta) < 1.92,$$

where $l(\hat{\theta})$ is the maximum log-likelihood and 1.92 is half of the critical value of the χ^2 statistic with 1 degree of freedom. The upper bound of 95% confidence interval for h_m is ∞. This is because the profile log-likelihood does not change too much for a large value of h_m; see bottom-left graph in Figure 3.6.

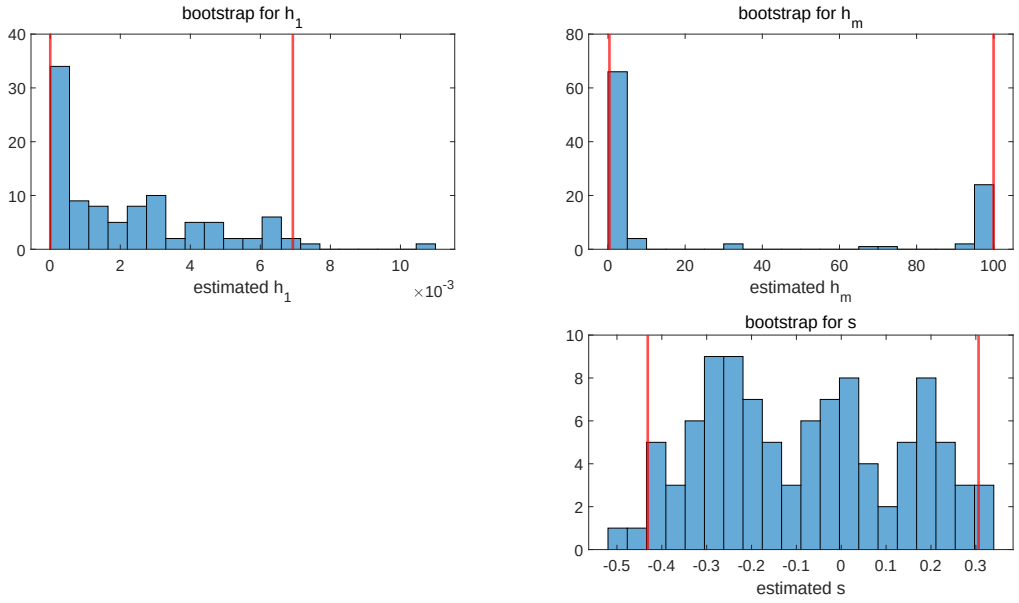

Figure 3.5: Histogram of 100 estimates for h_1, h_m and s using bootstrap for the Channing house data. The red lines locate the lower and upper bounds for the empirical 95 % confidence interval. The upper bound in the numerical optimisation for h_m is set to 100, which has been a very high mortality rate for humans.

The confidence intervals in Table 3.1 suggest considerable variability of the parameter estimates. This is not surprising as the data set includes just 362 individuals, only 130 of which are observed to die, and individuals are observed for just 7 years on average. The wide confidence intervals reflect the fact that our functional form for h_i involves great flexibility in its behaviour for large values of i (near $m = 100$). However, very few individuals are old enough in the data set in order to have a usable information concerning h_i for large i. Our confidence interval for h_m extends to infinity, suggesting that any value of h_m greater than 0.2770 is quite plausible. Depending on which value is used, h_1 and s must be adjusted so that appropriate estimates of h_i are obtained for less extreme values of i.

To gain insight into how much our parameter uncertainty would decrease by increasing the size of the data set, we determine the confidence intervals using data cloning. This is because we want to have the same estimate of the MLE for various sample sizes. The same estimate

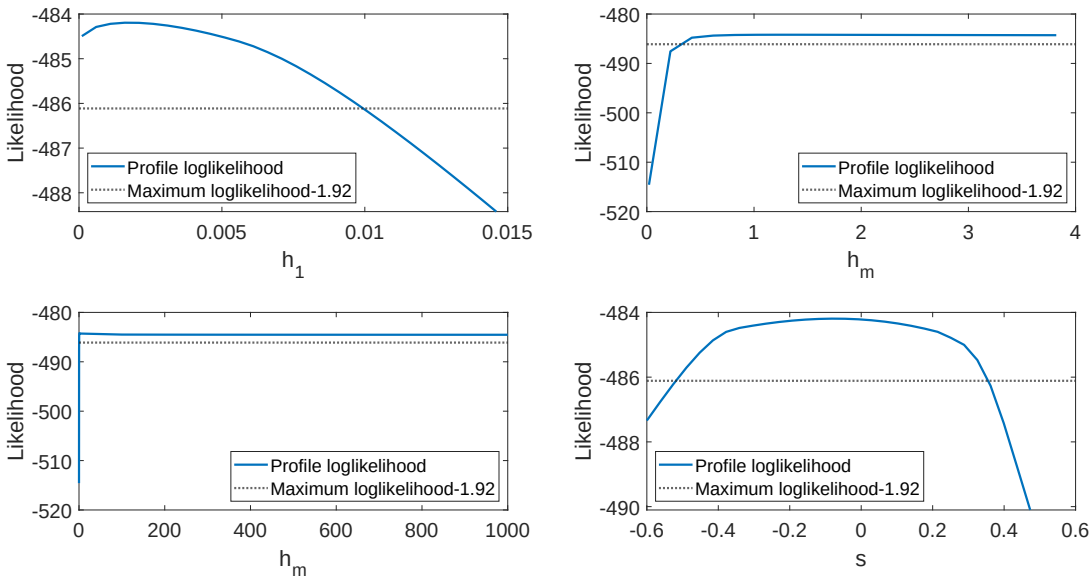

Figure 3.6: Top-left: Profile log-likelihood for h_1;Top-right: Profile log-likelihood for h_m in $[0, 4]$; Bottom-left: Profile log-likelihood for h_m in $[0, 1000]$; Bottom-right: Profile log-likelihood for s.

of the MLE makes it straightforward to compare the 95% confidence interval with different sample sizes. For more details on data cloning, see Chapter 6. We clone the data 10 times and 100 times, and determine the corresponding confidence intervals as before. Note that this does not change the estimates of h_1, h_m and s, as the log-likelihood function is simply multiplied by 10 and 100, respectively. However, this stretch produces narrower confidence intervals. The results are shown in Table 3.2, which illustrates that with 100 repetitions, there are 36,200 observations involving 13,000 deaths and the uncertainty about the parameter values is greatly reduced.

Table 3.2: Approximate 95 percent CIs (based on the profile log-likelihood) for the Channing House data cloned 10 and 100 times.

		Approximate 95% CIs					
		1 clone		10 clones		100 clones	
Parameter	Estimate	Lower	Upper	Lower	Upper	Lower	Upper
h_1	0.0017	0	0.0099	0.0003	0.0042	0.0012	0.0024
h_m	1.2750	0.2770	∞	0.6080	11.8032	0.9610	1.9106
s	-0.0735	-0.5270	0.3550	-0.2890	0.1230	-0.1410	-0.0112

The estimates of h_i that result from $\widehat{h_1}$, $\widehat{h_m}$ and $\widehat{s}$ are shown in Figure 3.7. The pattern of values look reasonable given that they represent instantaneous mortality rates that apply between ages 50 and 105. Figure 3.8 shows the force of mortality and the log (base 10) of the force of mortality based on the fitted model. Once again, the behaviour is quite reasonable.

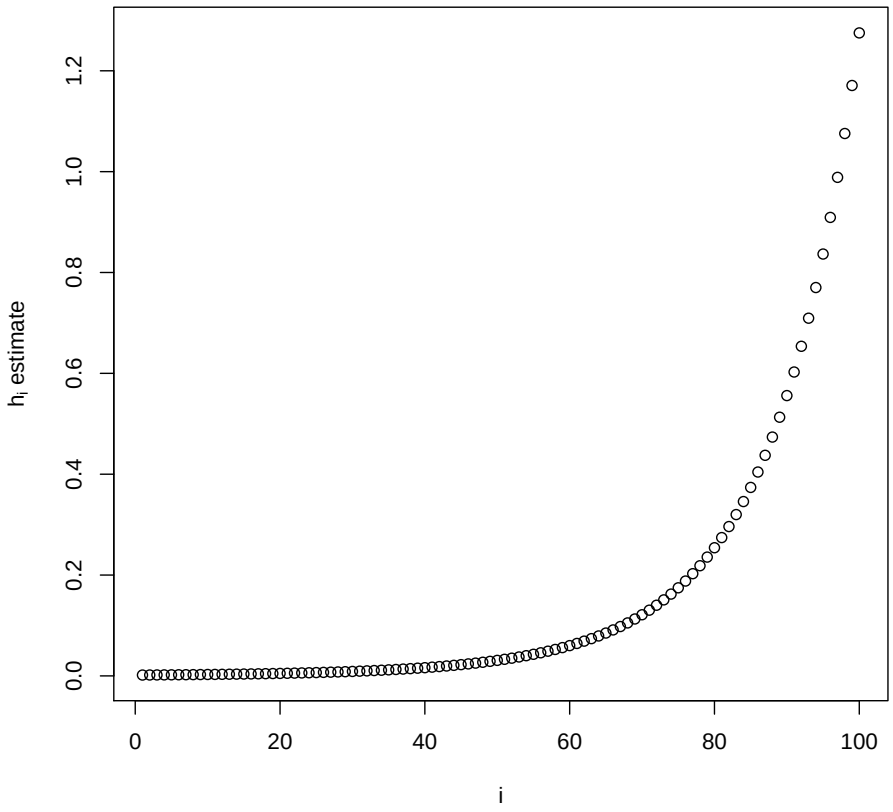

Figure 3.7: Estimates of h_i obtained by calibrating the PTAM using the Channing House female data.

In Figure 3.9, we illustrate the goodness of fit of our model to the Channing House data by plotting the fitted survival function along with the Kaplan-Meier nonparametric survival function estimates. We observe that our model fits quite well, and our fitted model estimates stay within the 95 percent confidence limits based on the Kaplan-Meier estimator.

To summarise the estimation process for the Channing House data, we firstly determine the starting age for the process and the lifespan parameter based on the observations. Then, the estimates of m is chosen by our prior knowledge. Hence, the estimate of m and λ are obtained. The last step is to estimate the remaining parameters h_1, h_m and s by the MLE technique.

With our fitted model, we can perform a variety of analyses of the ageing process. For example, it is relevant to examine the distribution of the state of an individual at different ages, given that the individual is alive at these ages. The probabilities associated with this distribution, that is $P(Y_t = i|Y_t \in E)$, are given by the $1 \times m$ vector

$$\mathbf{p}(t)/\mathbf{p}(t)\mathbf{e},$$

where $\mathbf{p}(t)$ can be calculated using equation (2.2.8). Since these probabilities correspond to age $50 + t$ in our example, it is intuitively appealing to transform the state Y_t to a comparable value

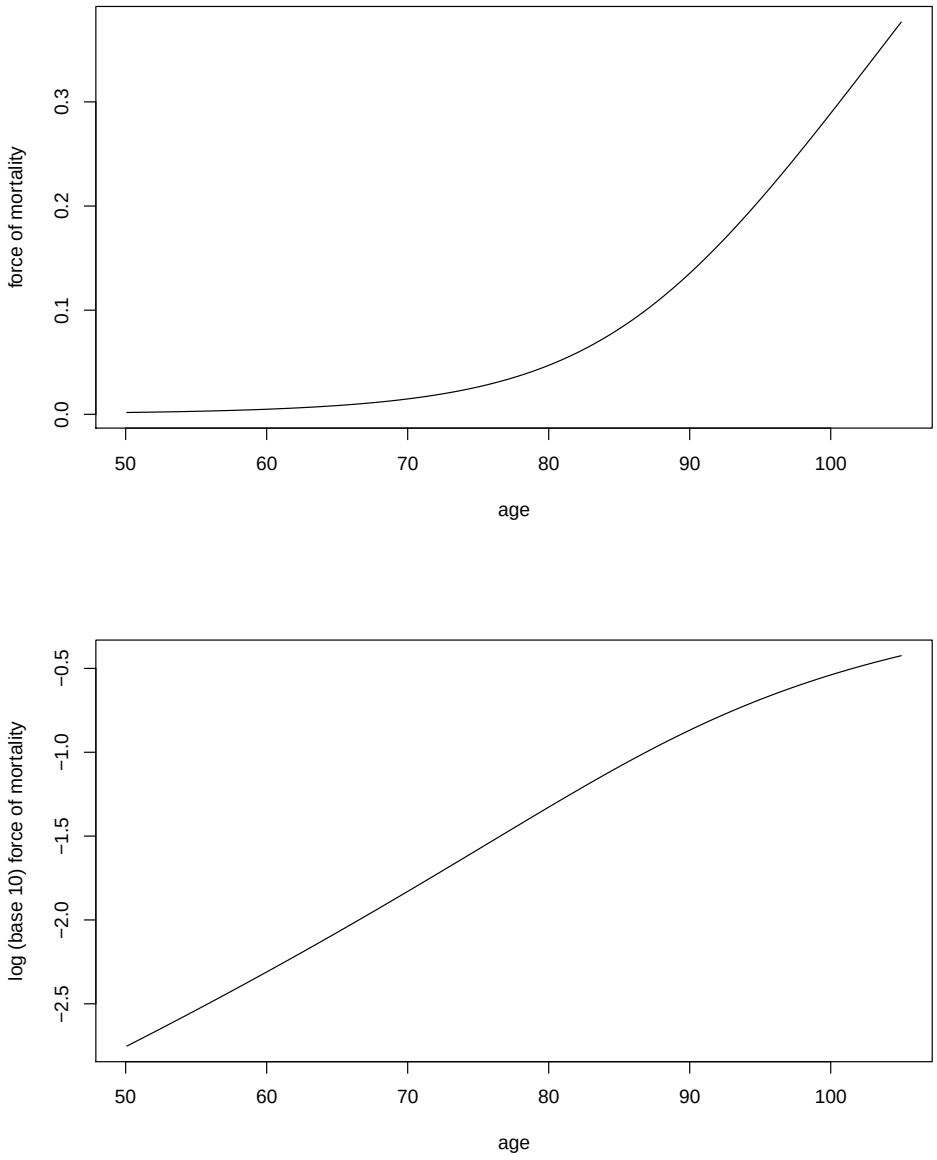

Figure 3.8: Force of mortality and log (base 10) force of mortality based on the PTAM calibrated using the Channing House female data.

that we can interpret as the individual's physiological age. Let

$$\text{Physiological age at calendar age } t = 50 + \frac{Y_t - 1}{m - 1}\,\psi. \tag{3.4.5}$$

Then an individual in state 1 has physiological age 50, and an individual in state 100 has physiological age $50 + \frac{100-1}{99}(55) = 105$. We can now determine the distribution of physiological age at any calendar age.

Figure 3.10 shows the distribution of physiological age at ages 60, 70, 80 and 90 for the

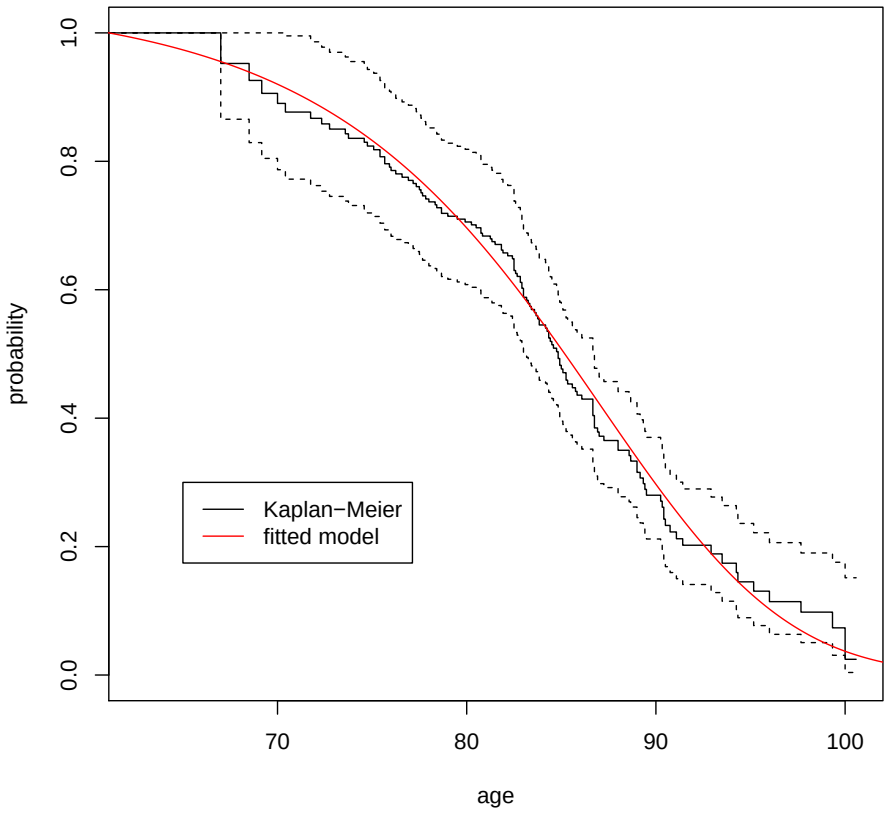

Figure 3.9: Survival function based on the PTAM calibrated using the Channing House female data, along with the Kaplan-Meier estimates of the survival function and corresponding 95 percent confidence limits (dashed).

PTAM calibrated using the Channing House female data. We observe that the physiological age distribution has less variability at younger ages. Small physiological age variation for younger people should be expected because we started the process at age 50, so that all individuals have a physiological age 50 at age 50. At older ages, the variability stabilises. We also observe that the mean physiological age, indicated by the vertical lines, is very close to 60 at age 60. However, at older ages, the mean physiological age becomes noticeably smaller than the age. This is due to mortality selection - individuals with a higher physiological age have a higher mortality rate. Therefore, old-age survivors tend to be those with a lower physiological age.

We can gain further insight into our ageing model by observing how the paths of the process behave. We can do this by simulating several paths and plotting them. To simulate a path, we assume that the individual is in state 1 (physiological age 50) at age 50. The individual will stay in state 1 for a length of time that is exponentially distributed with rate $\lambda + h_1$. We can generate this exponential random variable. At the end of the stay in state 1, the individual will die with probability $h_1/(\lambda + h_1)$ and move to state 2 with probability $\lambda/(\lambda + h_1)$. We can generate a sample from a Uniform (0,1) distribution to determine whether or not the individual

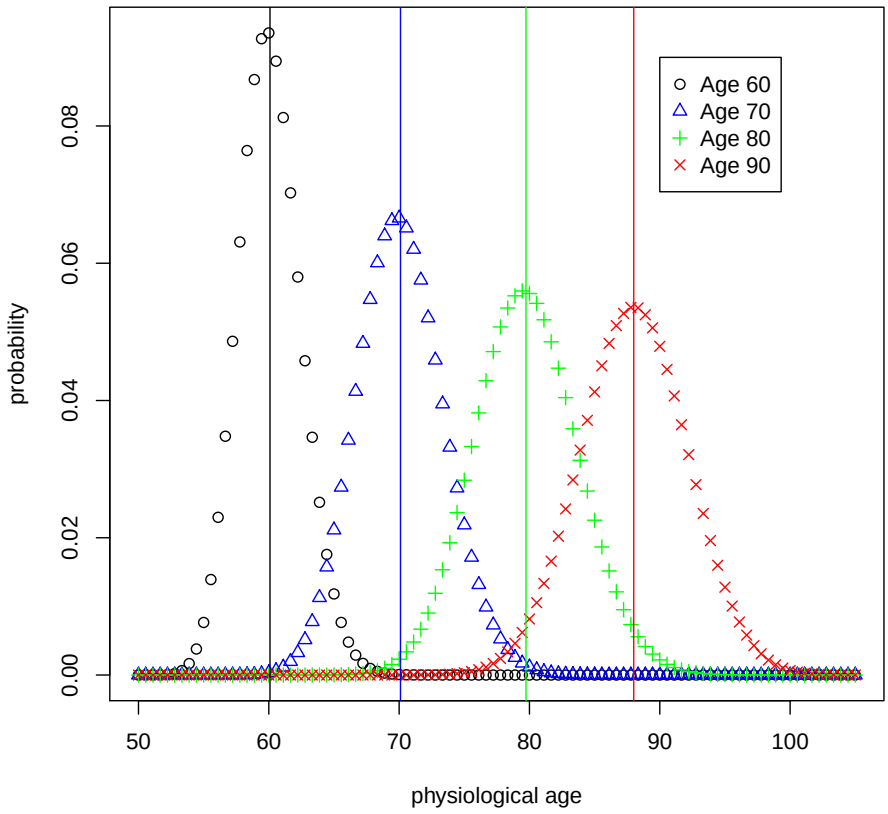

Figure 3.10: Physiological age distribution at ages 60, 70, 80 and 90 based on the PTAM calibrated using the Channing House female data. Vertical lines indicate the means of the distributions.

dies at this time. If not, we generate the time spent in state 2 (exponential with rate $\lambda + h_2$) and continue until the individual dies.

We simulated ten paths of the process and plotted them in Figure 3.11. Once again, the state of the process has been transformed to physiological age as described by (3.4.5). Figure 3.11 illustrates once again the variability in physiological age at different ages, but also shows the extent to which individual paths depart from uniform ageing over time. That is, we observe the "wigglyness" of individual paths.

3.4.2 Further analysis of the PTAM

Our basic modelling principle is that the (length of) life is the result of two important forces: the force of ageing (i.e., the fundamental biological process) and the force of dying (i.e., the external forces from environmental stress including accidents, accessibility of nutrition and medical care, etc). Besides, we want to model the two forces using a **stochastic** approach; specifically, we adopt the Markovian modelling framework of the Coxian distribution. That

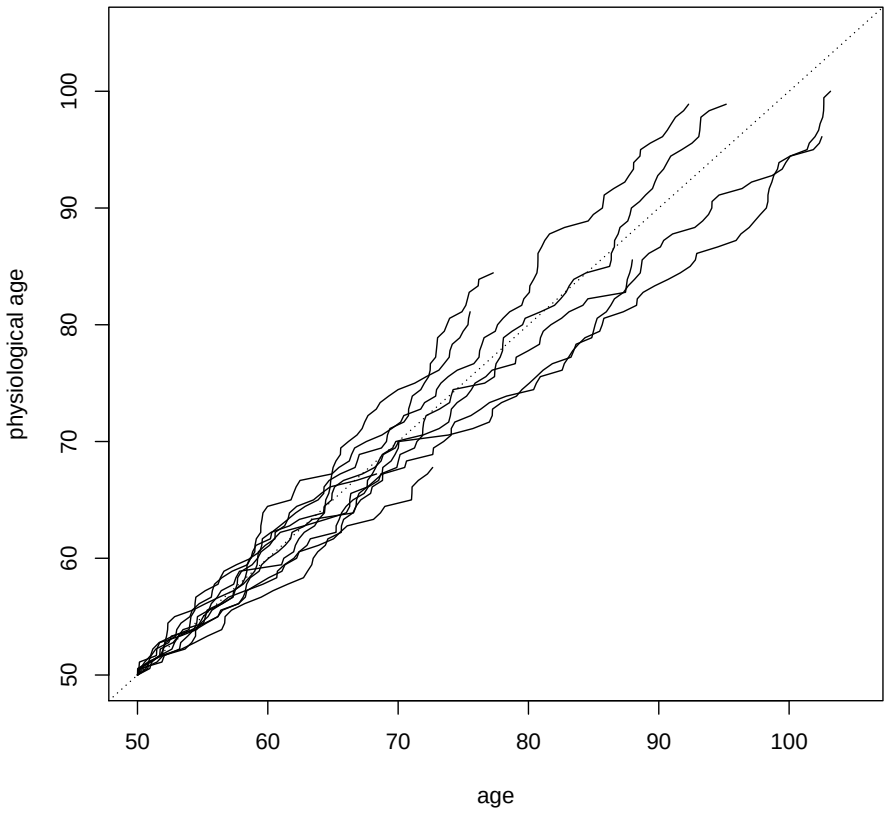

Figure 3.11: Ten simulated paths from age 50 until death based on the PTAM model calibrated using the Channing House female data.

is, we define a finite number of states representing different physiological capacity levels and transition intensities determining the rate of progression through the states. These states (with labelling suitably transformed) are interpreted as an individual's physiological age, since they are used to classify an individual's physiological capacity at the moment. As a result, we have naturally incorporated the concept of heterogeneity into the lifetime dynamic process by introducing the so-called "*physiological age*".

We mentioned in Chapter 1 that, since the middle of the 20th century, many researchers have begun collecting longitudinal data on various physiological variables. The most striking finding may be the fact that the decline of these physiological functions as a result of the underlying ageing process follows a slow, uniform and roughly linear pattern with age. This suggests that, according to our assumption of a uniform rate of increase in physiological age, declines in physiological function are decreasing, approximately linear, functions of physiological age.

Notwithstanding the appeal of our approach, other strategies can be used to construct a Coxian distribution. We discuss one of these below and illustrate that our model is nearly equivalent in terms of the resulting lifetime distribution.

3.4.3 Benchmarking with the Markovian Le Bras model

We now describe a model developed by Szilard (1959) and Le Bras (1976). Szilard (1959) proposed an ageing process theory assuming that chromosomes mutate with a constant rate in cells. If a cell accumulates too many mutations, it will cease functioning. Once a certain percentage of cells stop functioning in a human body, the body will die. Furthermore, Le Bras (1976) added an assumption that the inherited chromosomal mutation in the human body follows a Markov process. For a newborn, each cell mutates with rate λ_0 initially. Then new mutations occur with additional rate $i\lambda$, proportional to the total number of mutations i that have happened in this cell, and each cell dies with rate $i\mu$, proportional to i as well. Hence, let state i of a Markov process represent the state in which cells have accumulated i mutations in total. Then the transition rate from state i to $i+1$ is $\lambda_0 + i\lambda$ and the transition rate from state i to the absorbing state, i.e. to death, is $\mu_0 + i\mu$.

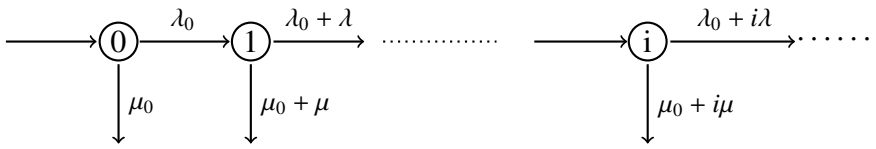

Figure 3.12: Diagram for the Le Bras dual linear model.

The Le Bras model was discussed by Yashin et al. (1994) in the following form: $\lambda_i = \lambda_0 + i\lambda$ and $\mu_i = \mu_0 + i\mu$, where μ_0 is the initial exiting rate. See Figure 3.12 for the diagram of the model. Here, the Le Bras model has an infinite number of states. Following the steps in solving the Kolmogorov forward equation in Chapter 2, one can derive the transition probability

$$P_i(t) = \frac{e^{-(\mu_0+\lambda_0)t}}{i!}\left(\frac{\lambda(1-e^{-(\lambda+\mu)t})}{\lambda+\mu}\right)^i \prod_{k=1}^{i}\left(\frac{\lambda_0}{\lambda}+k-1\right).$$

We recognise that $P_i(t)$ follows the term of a binomial series by considering $\frac{1}{(1-x)^\alpha} = \sum_{i=0}^{\infty}\left(\prod_{k=1}^{i}(\alpha+k-1)/i!\right)x^i$, where $x = \frac{\lambda(1-e^{-(\lambda+\mu)t})}{\lambda+\mu}$, $\alpha = \frac{\lambda_0}{\lambda}$, and $\prod_{k=1}^{i}\left(\frac{\lambda_0}{\lambda}+k-1\right)/i!$ is the binomial coefficient. It may be verified that

$$S(t) = \sum_{i=0}^{\infty} P_i(t) = e^{-(\lambda_0+\mu_0)t}\left(\frac{\lambda+\mu}{\mu+\lambda e^{-(\lambda+\mu)t}}\right)^{\frac{\lambda_0}{\lambda}}.$$

The Le Bras model is one of the first few attempts that have successfully incorporated physical and biological assumptions in a mathematical framework. Hence, it is important to review their approach, as this can help us gain insight on how our proposed model can be useful.

Remark 3.10. *Under the further assumption that $\mu \ll \lambda$, the hazard function of the Le Bras model can be approximated by*

$$\mu(t) = \left(\mu_0 - \frac{\mu\lambda_0}{\lambda}\right) + \frac{\mu\lambda_0}{\lambda}\, e^{(\lambda+\mu)t},$$

which is equivalent to the 3-parameter Gompertz-Makeham mortality model, $\mu(t) = a + be^{ct}$, with

$$a = \mu_0 - \frac{\mu\lambda_0}{\lambda}, \quad b = \frac{\mu\lambda_0}{\lambda}, \quad c = \lambda + \mu\,.$$

See Yashin et al. (1994) for further details. The important contribution of the Le Bras model is the dual linear structure in describing the ageing process (the linear pattern for λ_i) and the deteriorating effect (the linear pattern for μ_i) that could result in a Gompertz form of an exponentially increasing mortality pattern.

Remark 3.11. *In Yashin et al. (1994), the authors discuss the Le Bras model. It was found that, starting from a fixed frailty assumption, it is possible to derive the same mortality model. Hence, it was argued that, in the statistical analysis of data, results and conclusions depend not only on the data but also on underlying assumptions about the mechanism which generated the data. In other words, in reality, the use of lifetime data alone is not sufficient to distinguish between different assumptions on the mechanism that can generate the observed mortality patterns. More sophisticated data need to be used to validate the assumption.*

Remark 3.12. *There are some similarities between the Le Bras model and our proposed model. They are:*

- *both models differentiate the ageing effect from the ageing process;*
- *both models use a Coxian structure to describe the interaction between the intrinsic force of ageing and the external force of dying.*

However, from a practical perspective, the Le Bras model seems too restrictive in the sense that it requires a dual-linear pattern in its parametric form, which cannot be tested or modified.

It is worth noting that both models are ageing models rather than mortality models.

3.4.4 A simulation study

The Le Bras model describes a plausible ageing mechanism. Such model is a good candidate for modeling the human ageing process. However, the associated Markov chain has infinite states, which makes it hard to analyse the physiological age distribution at any chronological age. Since the heterogeneity of lifetimes in a population can be quantified by the physiological age, and the physiological age for our PTAM is easy to analyse, we shall examine if our PTAM can capture the probabilistic features of the Le Bras model.

To explore this, we simulated 5,000 lifetime observations from the Le Bras model with the parameters given in Table 3.3. We then fit our model to the simulated data. Before estimating the other parameters using the MLE method, we set the life span parameter ψ as

$$\psi = \widehat{\mathrm{TVaR}}_{0.999}(T)\,.$$

where $\widehat{\mathrm{TVaR}}_{1-\alpha}(T)$ is an empirical estimate of $\mathrm{TVaR}_{1-\alpha}(T)$, obtained from the simulated data. In particular, $\mathrm{TVaR}_{0.999}(T)$ is the conditional tail expectation of the lifetime that is in the 0.1% of old ages. We use $\widehat{\mathrm{TVaR}}_{1-\alpha}(T)$ to estimate the lifespan ψ for two reasons.

- The lifespan ψ is a high age to which only a very small proportion of individuals will survive and we believe $\text{TVaR}_{0.999}(T)$ is such a high age;
- The sample size is large enough that $\widehat{\text{TVaR}}_{0.999}(T)$ is a robust estimate of $\text{TVaR}_{0.999}(T)$.

Since our simulated sample size is 5,000, $\widehat{\text{TVaR}}_{0.999}(T)$ is the average of the five largest observations. We obtain $\psi = \widehat{\text{TVaR}}_{0.999}(T) = 112.55$. We estimate the remaining parameters using the MLE method. Figure 3.13 shows a histogram of the simulated data, the pdf of the Le Bras model and the pdf of the fitted PTAM. The MLE of m is determined by fitting the model for different but fixed m using the MLE method, and comparing their respective Negative Log-Likelihood (NLL) values. As one can see in Table 3.4, when $m = 225$, we obtain the lowest NLL value. It is worth noting that the log-likelihoods are similar for different fixed m's. Therefore, the estimation of m using lifetime data only has huge variability. This is due to the fact that lifetime data carry very little information about m.

Table 3.3: Parameter values

	λ_0	λ	μ_0	μ
The Le Bras model	0.6	0.07	0.001	0.4×10^{-4}
Our fitted PTAM model	h_1	h_m	λ	s
with $m = 225$	0.0008	1.6535	1.9991	-0.1112

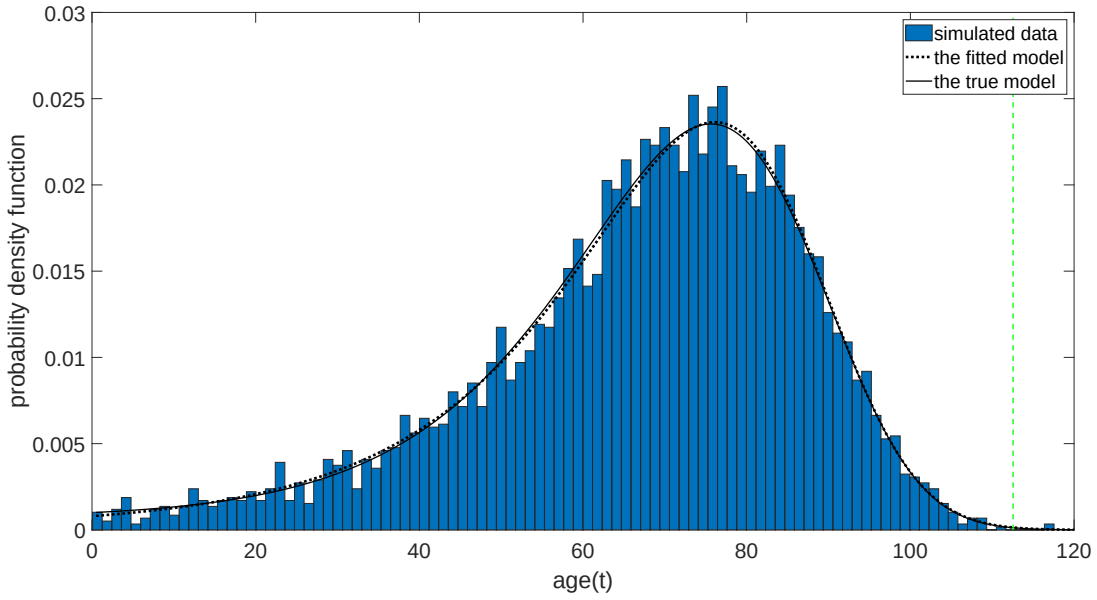

Figure 3.13: Histogram of 5,000 lifetimes simulated from the Le Bras Model. The fitted model with m=225 is plotted vis-à-vis the true model. The dotted vertical line indicates the location of $\psi = 112.55$.

Figure 3.13 shows that if the observed mortality rates are truly generated from the Le Bras model, our model can provide a nearly equivalent representation. Furthermore, we apply the

Pearson's chi-squared test for the estimated results. The frequency table has 120 categories from age 1 to age 120. The limiting age is set to 120 because the survival probability to such old age is tiny. Let O_i and E_i be the observed number of deaths, and the expected number of deaths between age $i-1$ and age i, respectively. The calculated chi-squared statistic is $\chi^2 = 2.54$ with degrees of freedom equal to $120 - 4 = 116$, where

$$\chi^2 = \sum_{i=1}^{120} \frac{(O_i - E_i)^2}{E_i},$$

and the corresponding $p-$value is 1. Therefore, there is not enough evidence to reject the null hypothesis of no difference between the fitted distribution and the distribution from the Le Bras model. We are able to achieve the same goodness of fit as the Le Bras model by allowing the exit rate h_i to increase faster than a linear rate. The fact that s can be flexible to take any value in order to accommodate the data (in this example $s = -0.1112$) is a specific feature of our model.

Hence, we have found an alternative to the Le Bras dual linear model with a constant transition rate λ, though our model has a slightly different interpretation of the underlying ageing mechanism. While the original Le Bras model contains an infinite number of states, our PTAM model re-labels them into m states. The ageing process is still modelled as marching forward from one state to the next, and the impact of ageing is still described as increased frailty to hazard with higher indexed state. The only difference is the way of labeling the states. Transforming from the Le Bras model to our PTAM model, the earlier states may need to be split into more states, while the later states may need to be grouped, but not in a linear fashion. The overall effect is that, in our modelling framework, the transitions from one to another are required to occur uniformly due to the use of the constant transition rate λ. In particular, the pattern for h_i changes from a linear increase to a pattern that has to climb slightly faster than the exponential rate, as indicated by the parameter s taking a small negative value.

In Table 3.4, we show additional results from fitting our model to the simulated data. As required, the estimate of the parameter λ increases with m. We have also plotted the survival functions, pdfs and hazard functions of these fitted models with m ranging from 200 to 250 in Figure 3.14, the dotted vertical line indicates the location of $\psi = 112.55$. For ages up to this value, the distributions are very close to each other. However, for the hazard functions, there are small but noticeable differences beginning near age 100.

Examining the hazard functions in the bottom left panel of Figure 3.14, we find that the fit changes asymmetrically as m moves away from its optimal value. The model with $m = 250$ fits much better than the one with $m = 200$. Also, the model with $m = 250$ gives the highest hazard rate at the right end of the life span. For phase-type distributions, it is a well-known property that

$$\lim_{t\to\infty} h(t) = \min_{i=1,\cdots,m} \{d_1, d_2, \ldots, d_m\},$$

where the values of d_i are the eigenvalues of the transition intensity matrix $\mathbf{\Lambda}$. In our fitted PTAM, we have

$$\lim_{t\to\infty} h(t) = \min\{\lambda + h_1, h_m\},$$

since the eigenvalues $\lambda + h_i$ increase except for the last element h_m. The limit of $h(t)$ for each fitted model is provided in the last column of Table 3.4, and the model with $m = 250$ has the

Table 3.4: Estimation results using different m based on 5,000 lifetimes simulated from the Le Bras limiting distribution. The first column gives the negative log-likelihood - NLL. The last column is the limit of the resulting hazard function $h(t)$ as $t \to \infty$.

NLL	h_1	h_m	λ	s	m	$\min(\lambda + h_1, h_m)$
21631.884	0.00081	2.0033	1.7764	-0.1237	200	1.7772
21631.826	0.00080	1.8392	1.8652	-0.1183	210	1.8392
21631.806	0.00080	1.7079	1.9540	-0.1134	220	1.7079
21631.713	0.00080	1.6535	1.9991	-0.1112	225	1.6535
21631.813	0.00079	1.6006	2.0428	-0.1089	230	1.6006
21631.843	0.00078	1.5108	2.1316	-0.1047	240	1.5108
21631.889	0.00078	1.4354	2.2205	-0.1009	250	1.4354

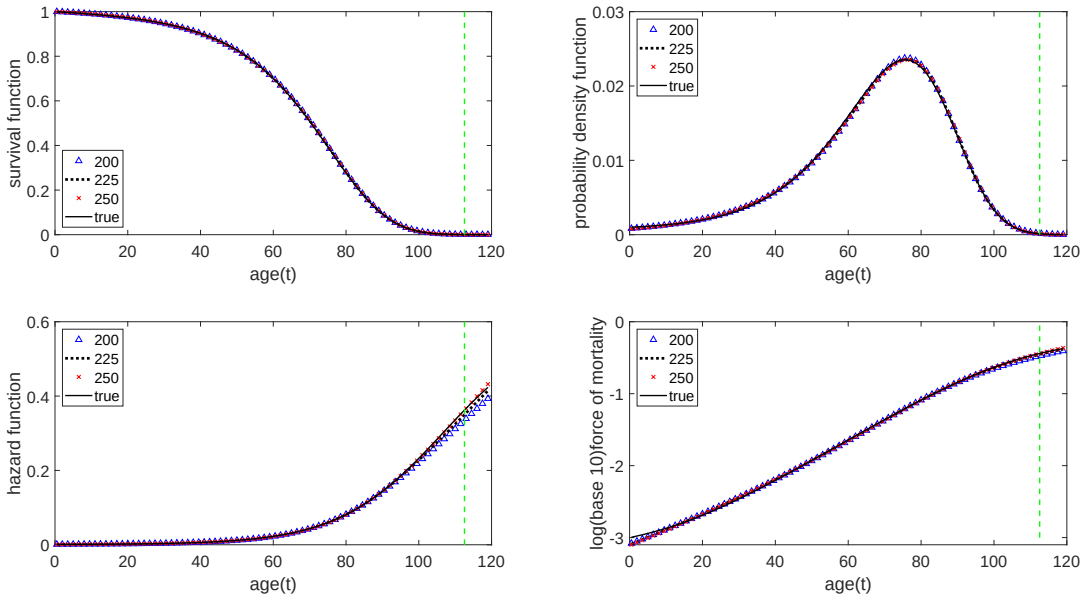

Figure 3.14: Fitted survival function $S(t)$, probability density function $f(t)$, hazard function $h(t)$, and log (base 10) hazard function. Each graph includes four curves corresponding to the fitted model with m=200, 225 and 250, as well as the true model. The dotted vertical line indicates the location of $\psi = 112.55$.

lowest value of this limit.

In Figure 3.15, we show the logarithm of the hazard function of the fitted models from age 80 to 500. We extrapolate the hazard function to 500, an unrealistically old age, to demonstrate the differences in the limiting hazard rate for the fitted results. Although the graph shows differences in tail behaviour, these differences occur well beyond the age range that is relevant to human ages from a practical perspective. When considering the ages that human beings can survive to, the fitted models are equivalent in the sense of their resulting lifetime distributions.

The statistical equivalence of the outcome between two different model structures is of importance. By all means, what we propose here is a statistical model which focuses on de-

scribing the progressive and irreversible feature of the ageing process. The interaction between ageing and mortality is not the ageing process itself; rather, the ageing effect is subject to what we observe and how we observe it. In other words, the model aims at capturing the related increasing mortality phenomenon due to the ageing process. It might be impossible to completely rule out some subjective element in how we perceive the process. The equivalence means that the interpretation of the internal ageing process could be flexible depending on what and how we observe the process, but the model should be "true" to the ultimate observable facts – which are the observed death rates by age in this situation. This is the basic principle to validate a statistical model.

To understand the variability of the estimates of parameters associated with h_i, the simulation study was repeated 200 times. Then, holding ψ fixed at 112.55 and m fixed at 225, h_1, h_m and s were estimated for each of the 200 simulations. Two hundred simulations are large enough when using the estimated values to approximate the estimators' distributions. Confidence intervals based on the 2.5th and 97.5th percentiles of the sample of estimates for each parameter were then constructed; they are shown in Table 3.5.

Table 3.5: Approximate Confidence Intervals based on Simulated Data.

		Approximate 95% CI	
Parameter	Estimate	Lower	Upper
h_1	0.0008	0.0007	0.0009
h_m	1.6530	1.1001	2.7403
s	-0.1110	-0.1442	-0.0671

As anticipated, with 5,000 complete lifetimes, the variability of the parameter estimates is quite reasonable.

Now we move to another important concept that is introduced by our modelling framework. Similar to (3.4.5), we can define a physiological age index based on our calibrated model:

$$\text{Physiological age index } X_t \text{ at calendar age t} = \frac{Y_t - 1}{m - 1}\psi. \tag{3.4.6}$$

So X_t is a random variable transformed from Y_t, and X_t can be interpreted as a physiological age index since it is associated with where an individual is at in the ageing process. Note that X_t takes values from $[0, \psi]$. That is, X_t is not affected by the state number parameter m and has the same scale as the calendar age's upper limit ψ. This makes X_t a good counterpart for age t: for each individual which may follow a personalized ageing process Y_t, there is this corresponding physiological age index X_t telling the health status of the individual at his/her calendar age t.

In Figure 3.16, we plot h_i versus physiological age $\frac{i-1}{m-1}\psi$. The pattern of h_i on $[0, \psi]$ turns out to be quite stable with different m. This is another good feature of the proposed model, indicating that the underlying mechanism does not vary too much with m.

In our analysis so far, we have demonstrated that some critical aspects of our ageing model are quite stable with changing m. Our estimated value of m was determined by maximising the likelihood. However, the likelihood is nearly flat for values of m near the maximum. So there are only small differences in the lifetime distribution that result from different m. This leads to the question: How is the model affected by the value of m and how should m be determined?

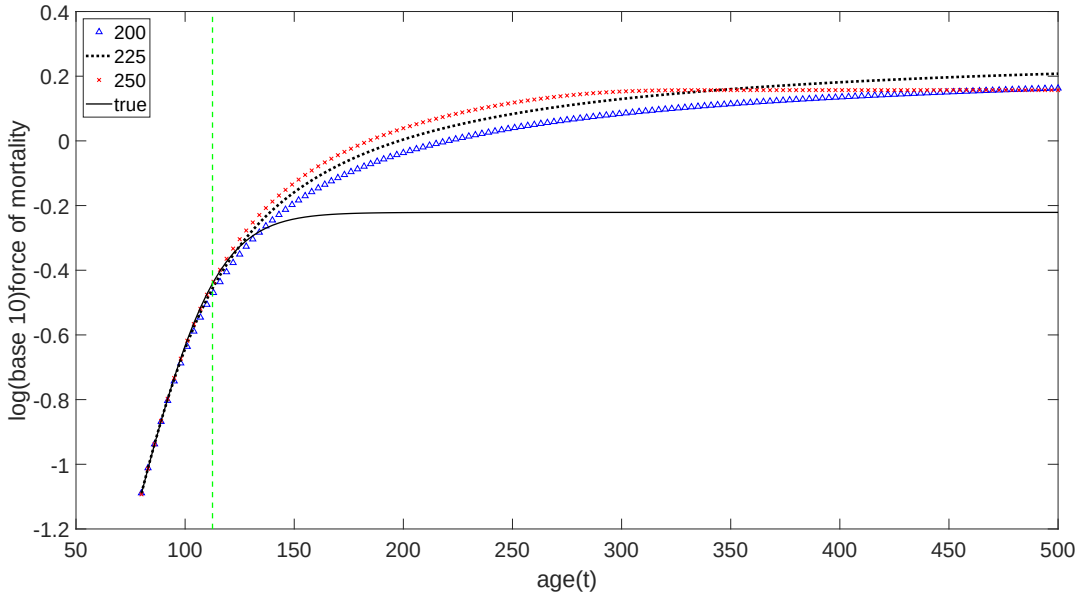

Figure 3.15: Fitted log (base 10) hazard function extended to age 500. Each curve corresponds to the fitted model with m=200, 225 and 250, as well as the true model. The dotted vertical line indicates the location of $\psi = 112.55$.

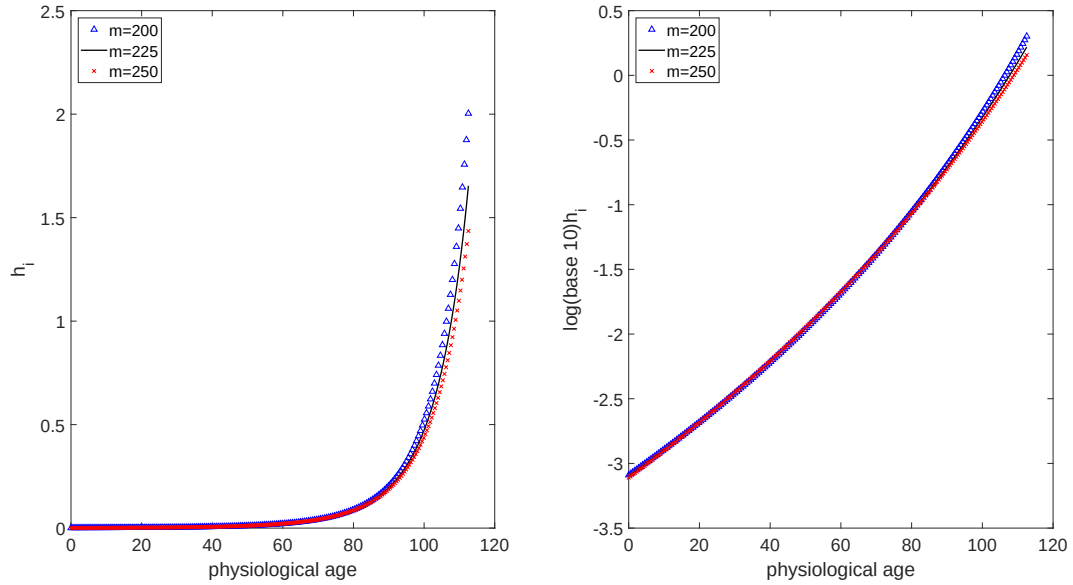

Figure 3.16: The left graph shows the exit rate h_i with m=200, 225 and 250. The right graph shows the log-exit rate with m=200, 225 and 250. Both are plotted against the physiological age $\frac{i-1}{m-1}\psi$.

While the lifetime distribution is relatively insensitive to m, the progression of physiological

age with advancing calendar age is affected by m. In particular, the variability in physiological age is significantly affected by m. In fact, as $m \to \infty$, this variability disappears, and physiological age converges to calendar age. In particular, when m increases, the expected physiological age of an alive individual at calendar age $t \in (0, \psi)$ approaches $\frac{t}{\psi}m$ and the variability decreases. This phenomenon is illustrated in Figure 3.17, where ten simulated sample paths for each of four different values of m: 25, 100, 225, and 1,000 are plotted. In fact, we can prove that as $m \to \infty$, $\mathbb{E}(X_t|Y_t \in E) \to t$ and $Var(X_t|Y_t \in E) \to 0$ in Chapter 4.

Furthermore, the resulting hazard function takes on the behaviour of our h_i values. That is, as m gets large, we observe that the hazard function, which is the expected exit rate, given by $h^m(t) = \mathbb{E}(h_{Y_t}|Y_t \in E)$ approaches

$$h(t) = \begin{cases} \left(\left(1 - \frac{t}{\psi}\right) h_1^s + \frac{t}{\psi} h_m^s\right)^{1/s} & s \neq 0, \\ h_1^{1-\frac{t}{\psi}} \, h_m^{\frac{t}{\psi}} & s = 0. \end{cases}$$

In addition, the variability of h_{Y_t} becomes smaller and smaller when m increases. This is not surprising because of our observation in the last paragraph that Y_t approaches $\frac{t}{\psi}m$ as m goes to infinity.

Figure 3.17 shows ten simulated sample paths for each of four different values of m: 25, 100, 225, and 1,000. We observe that, as m increases, there is less variability in physiological age, and physiological age is converging to calendar age. This suggests that m should not be too large, or the model will not appropriately reflect the variability in physiological age. Obviously, m cannot be too small either, or there will be too much variability in physiological age.

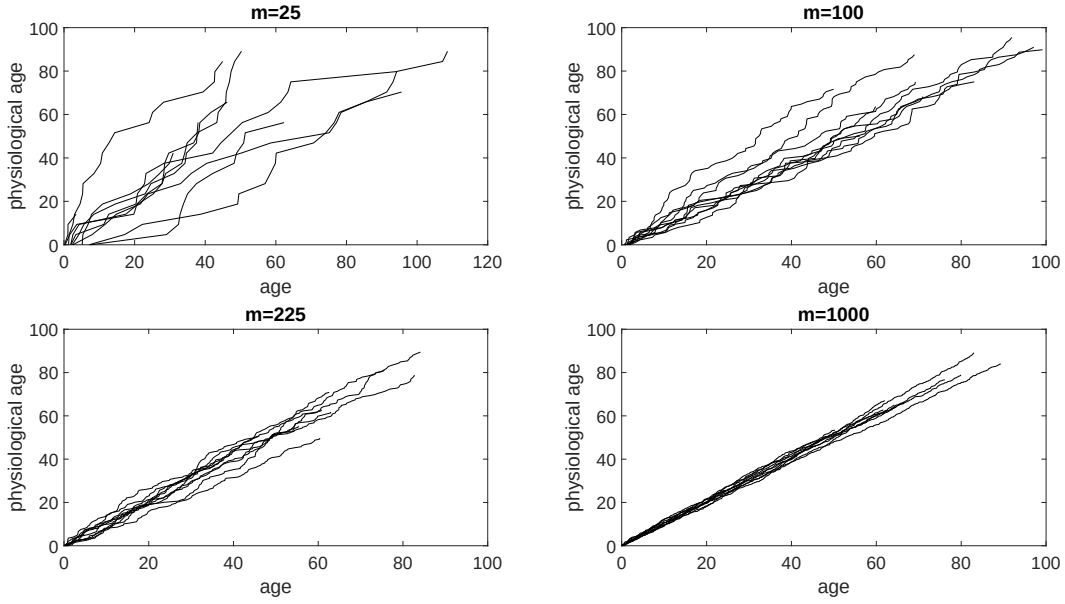

Figure 3.17: Ten simulated sample paths of the fitted PTAM model for each of four different values of m, holding the parameters h_1, h_m and s fixed.

In summary, we intend to choose m and ψ in such a way that the calibrated model can reflect one's opinion about the variability in physiological age. Also since ψ tends to take a value close to the lifespan measured in years, the physiological age has a range between 0 and ψ, which seems to be a reasonable scale conversion from calendar age. However, if one has data that includes information about one or more health variables related to physiological age in addition to the lifetimes of individuals, then additional consideration can be given to estimate m and ψ along with the other parameters. This is the preferred approach if suitable data are available.

Remark 3.13. *The proposed PTAM gains some flexibility when m increases from a small value. This is because, intuitively, a Markov chain with more states can represent more health statuses. However, this flexibility may lose as m keeps increasing to a large value due to the restricted structure of* h_i.

3.4.5 Comparison with the model in Lin and Liu (2007)

Our proposed PTAM is compared with the PTAM of Lin and Liu (2007), who demonstrated how the ageing component, combined with other causes of death, can explain well the age pattern of mortality rates for observed cohorts. However, the main goal of the proposed PTAM is not to reproduce mortality patterns; instead, it aims at finding a way to describe the ageing process, in which ageing-related mortality rates are used to determine a quantitative measurement of the ageing rate and the associated ageing effect under our pre-defined model framework. Hence, the resulting lifetime distribution from our model cannot be treated as a mortality model as in Lin and Liu (2007).

For our proposed PTAM, the structure on dying rate is flexible to achieve a variety patterns. In other words, the proposed functional form (3.3.3) can be used to reproduce the ageing-related mortality rates in equation (3.3) in Lin and Liu (2007), which, for the convenience of readers, is provided below:

$$h_2(i) = i^p q,$$

with parameter p and q estimated from three different cohorts.

We match our proposed functional form (3.3.3) to the three estimated patterns shown in Figure 5 of Lin and Liu (2007). The parameters of h_1, h_m and s for the proposed PTAM are given in Table 3.6 along with the values for q and p estimated from Lin and Liu (2007), respectively.

Table 3.6: Parameters q and p adopted from from Lin and Liu (2007), and their corresponding calibrated values in terms of the proposed structure (3.3.3) for the Swedish cohort data in years 1811, 1861, and 1911.

Year	q	p	h_1	h_m	s
1811	9.3157×10^{-9}	3	1.7103×10^{-8}	0.0745	0.3328
1861	2.6351×10^{-13}	5	8.5212×10^{-13}	0.0844	0.1994
1911	1.8872×10^{-15}	6	5.1411×10^{-14}	0.1206	0.1662

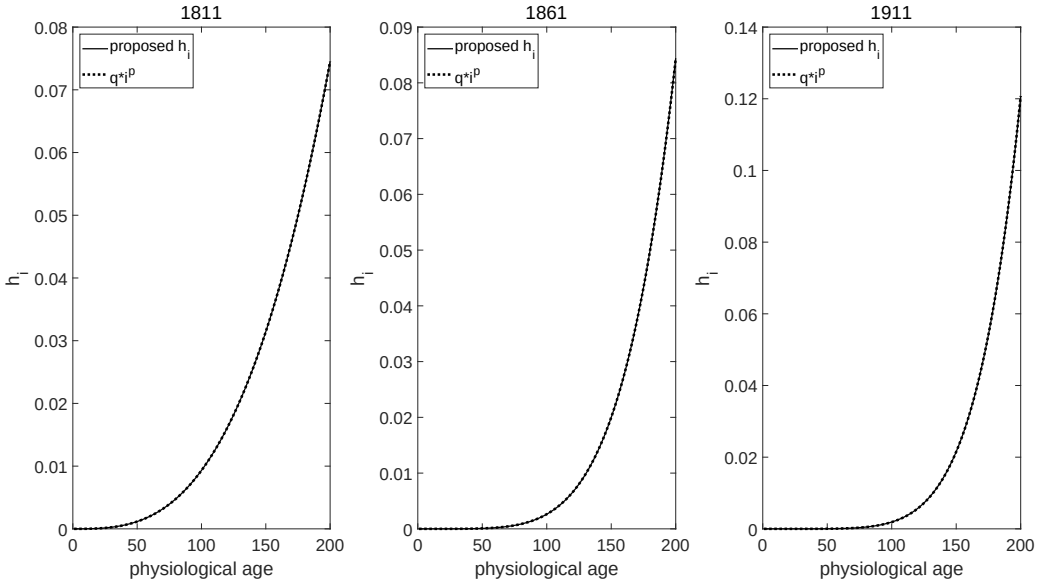

Figure 3.18: Calibrated h_i using the form (3.3.3) versus $h_i = i^p q$ in Lin and Liu (2007) for three cohorts.

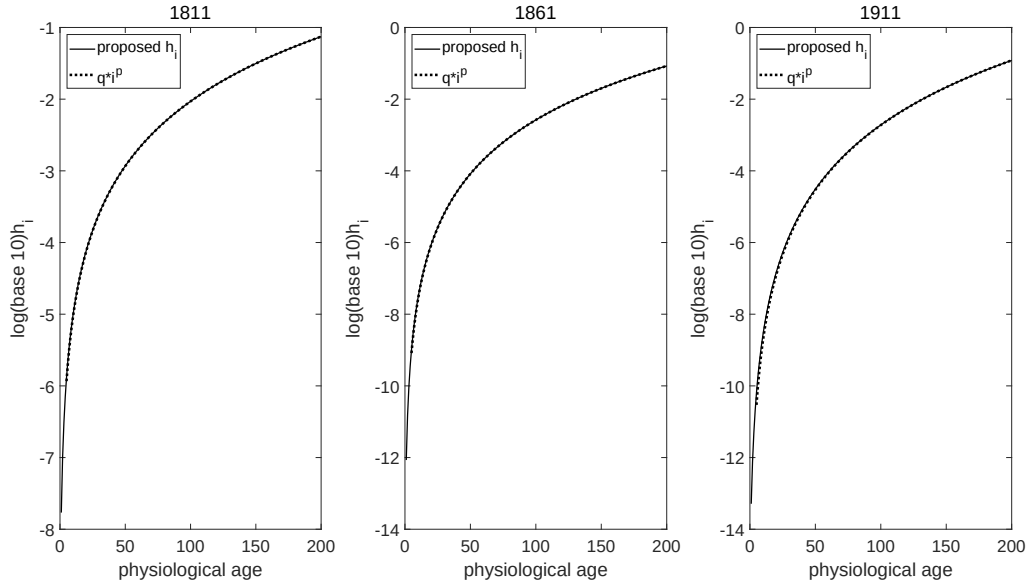

Figure 3.19: Calibrated $\log_{10}(h_i)$ using the form (3.3.3) of versus $h_i = i^p q$ in Lin and Liu (2007) for three cohorts.

As shown in Figures 3.18–3.19, the functional form (3.3.3) can reproduce the three different scenarios in Lin and Liu (2007) well. The parameter s, which is supposed to capture the curvature of the death rates associated with the ageing process, ranges from 0.3328 to 0.1662 for cohorts from 1811 to 1911. This appears to be a good feature of the proposed PTAM. The combinations of p and q are used to capture the ageing-related mortality pattern in Figure 16 of

Cheng et al. (2021). In comparison, the parameter s here is the analogue of the two parameters in Lin and Liu (2007).

3.5 Flexibility of the resulting distribution

The quantitative study of ageing is still in its infancy due to the lack of directly observable ageing-related data. The ageing process is not directly observable even using the most advanced technologies. For this reason, earlier theories of ageing often use the impact of ageing (e.g. the observed increasing death rates with increasing age to validate their underlying hypothesis about ageing). We know two facts about the relation between mortality rate and the rate of ageing:

- The mortality rate is strongly correlated with the rate of ageing;
- It is hard to collect ageing-related data, but it is much easier to collect mortality data.

We used two sets of lifetime data to quantify the ageing effect (Channing House data and simulated data from the Le Bras model). Reasonably good fits on both sets of data are achieved. The promising approximation motivates us to explore the flexibility of the proposed PTAM, so that we can answer the following question: *Can the proposed PTAM achieve a good fit on any set of lifetime data?* The assessment of the PTAM's flexibility in our exploration is based on how well the resulting pdf and hazard function can approximate the true ones, since both functions are easily obtained from the data and are commonly used to validate the calibrated result.

For each living being, we assume that its embedded ageing process still meets our definition of ageing process – " the genetically determined, progressive and essentially irreversible process". However, the dying rate pattern may be distinct for different living beings, and their hazard rate patterns may be dissimilar. For example, the mortality rate for human being has an increasing pattern beyond the attainment of adulthood; while the mortality rate for fruit flies has a flipped U-shaped pattern.

Since the dying rate is monotone for the proposed PTAM, the resulting hazard rate is monotone by Theorem 3.2. It is impossible for the proposed PTAM to produce a non-monotonic hazard rate. Therefore, we only investigate the goodness of fit for some popular lifetime distributions with monotonic hazard rate.

The following list includes the tested lifetime distributions we are interested in:

- Gamma distribution with increasing hazard rate;
- Gamma distribution with decreasing hazard rate;
- Weibull distribution with increasing hazard rate;
- Weibull distribution with decreasing hazard rate;
- Pareto distribution with decreasing hazard rate;
- Convolution of two exponential distributions;

- Convolution of two Weibull distributions;
- Gompertz-Makeham distribution;
- Makeham's second extension of the Gompertz distribution.

According to our experiments, the fitted results are promising for all tested distributions, except for the tail part of heavy-tailed distributions. As expected, it is problematic to use light-tailed distributions, which includes the proposed PTAM, to approximate heavy-tailed distributions, because there is always a distinct tail, which requires a large number of mixtures (states in the PTAM) to approximate the tail well. For example, Figure 3.20 is the typical fitted result for light-tailed distributions, and Figure 3.21 is the typical fitted result for heavy-tailed distributions. More discussion and fitted results are presented in Appendix C.

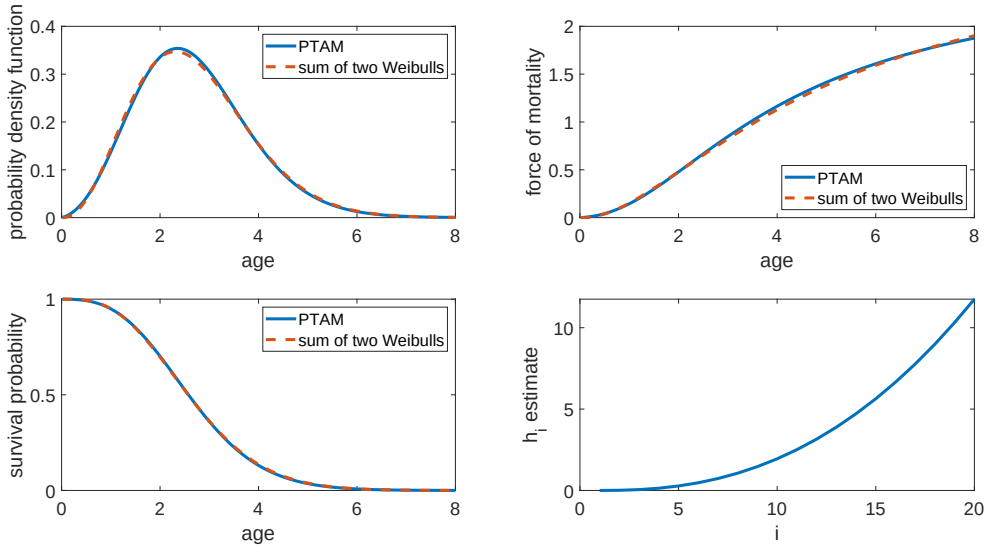

Figure 3.20: Proposed PTAM approximating a convolution of two Weibull distributions with $\lambda_1 = 2, k_1 = 1, \lambda_2 = 1,$ and $k_2 = 1.3$.

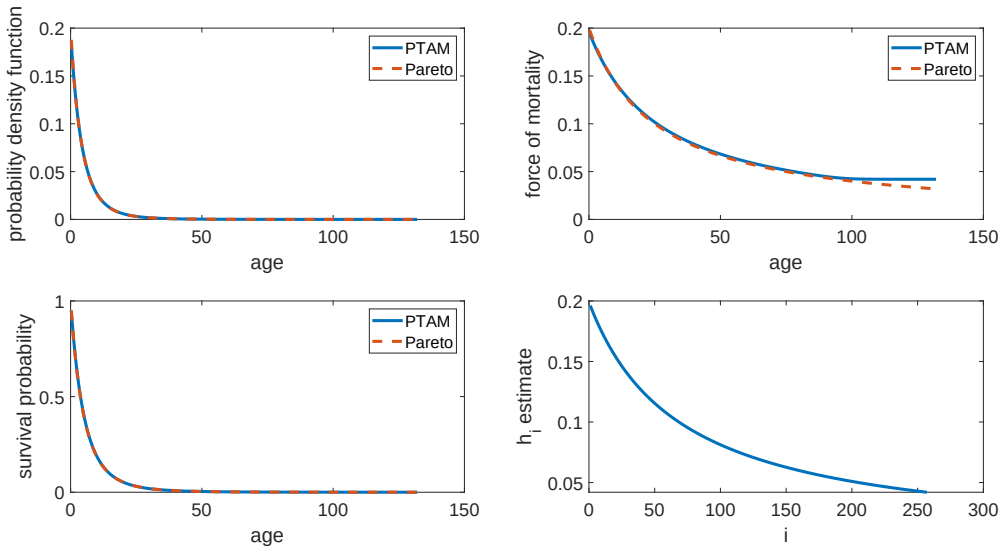

Figure 3.21: Proposed PTAM approximating a Pareto distribution with $k = 0.2$ and $\sigma = 5$.

Chapter 4

State and probability distributions

In this Chapter, we investigate the state distribution at any time and the resulting probability distribution for the proposed PTAM. The investigation is heavily based on the underpinnings of mathematical statistics. One numerical example is provided after each result to gain certain intuitions.

We establish the convergence in probability at time t of a transformation of the state random variable conditional on being in transient states to a function of t as the number of states approaches infinity. This property shows that the physiological age index converges to the chronological age if the PTAM has a large number of states. Furthermore, the distribution of the transformed state random variable asymptotically converges to a normal distribution.

On the other hand, conditional on the current state, the lifetime random variable still follows the proposed PTAM. Based on the resulting probability distribution, we can group the parameters of the proposed PTAM into two parameter classifications: shape and scale. The limit of the resulting hazard function with respect to the number of states in the PTAM is provided at the end of this Chapter.

4.1 The associated state distribution

As we demonstrated using the PTAM to fit the Channing House data, both the resulting lifetime distribution and the state distribution at any time were important.

Given any lifetime t, the $1 \times m$ probability vector of being in each state at time t is $\mathbf{p}(t)$, whose value can be calculated by (2.2.8). Conditional on being alive at time t, the probability vector is updated by

$$\mathbf{p}(t|Y_t \in E) = \frac{\mathbf{p}(t)}{\mathbf{p}(t)\mathbf{e}},$$

where $\mathbf{e}$ is a $m \times 1$ column vector of ones and $\mathbf{p}(t)\mathbf{e}$ is the sum of all elements in $\mathbf{p}(t)$. The kth element of $\mathbf{p}(t|Y_t \in E)$ is the probability in state k conditional on being alive at time t, or $P(Y_t = k|Y_t \in E)$.

The state distribution contains information on the variability of the state at any age. Using (3.4.6), the state label could be transformed to a physiological age index with range $[0, \psi]$.

Using the transformation

$$Z_t = \frac{Y_t - 1}{m - 1},$$

the physiological age index takes values between 0 and 1.

Recall that Figure 3.17 showed $\mathbb{E}\left(\frac{Y_t}{m}\middle|Y_t \in E\right)$ gets closer to $\frac{t}{\psi}$ and $\mathrm{Var}\left(\frac{Y_t}{m}\middle|Y_t \in E\right)$ gets closer to 0 as m increases from 25 to 1000. It is reasonable to guess that the random variable $\left(\frac{Y_t}{m}\middle|Y_t \in E\right)$ converges to $\frac{t}{\psi}$ in probability as $m \to \infty$. The convergence is indeed true as will be shown below and owing to the fact that $\lim_{m\to\infty} \frac{Y_t}{m} = \lim_{m\to\infty} Z_t$. In order to emphasise that Y_t and Z_t are related to the value of m, we use $Y_{t,m}$ to represent Y_t and use $Z_{t,m}$ to represent Z_t in Theorem 4.1. The key idea of the proof is to prove $P_k(t)$ is close to $\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}$ and $P_m(t)$ is close to 0 as $m \to \infty$.

Theorem 4.1. *Suppose the proposed PTAM has m ($m = 1, 2, \ldots$) transient states with labels $1, \ldots, m$ and one absorbing state with label $m+1$. The ageing rate, which is the transition rate from one transient state to the next transient state, is λ. For $i = 1, \ldots, m$, the dying rate in state i is h_i with $0 \le h_1 < h_m < \infty$, and h_i follows (3.3.3). Let ψ be the lifespan parameter. For any time $0 \le t < \psi$, let $Y_{t,m}$ be the state variable at time t, and $Z_{t,m} = \frac{Y_{t,m}-1}{m-1}$, then the sequence of random variables $(Z_{t,m}|Y_{t,m} \in E)$ converges to $\frac{t}{\psi}$ in probability as $m \to \infty$.*

Proof. By Theorem 3.3, the rate h_i is increasing if and only if $h_1 < h_m$. According to (3.3.3), when $s \neq 0$, for any $i = 1, \ldots, m-1$,

$$\begin{aligned}
\lim_{m\to\infty} \frac{h_{i+1}}{h_i} &= \lim_{m\to\infty} \left(\frac{h_1^s \frac{m-(i+1)}{m-1} + h_m^s \frac{(i+1)-1}{m-1}}{h_1^s \frac{m-i}{m-1} + h_m^s \frac{i-1}{m-1}} \right)^{1/s} \\
&= \lim_{m\to\infty} \left(\frac{h_1^s(m-i-1) + h_m^s i}{h_1^s(m-i) + h_m^s(i-1)} \right)^{1/s} \\
&= \lim_{m\to\infty} \left(1 + \frac{h_m^s - h_1^s}{h_1^s(m-i) + h_m^s(i-1)} \right)^{1/s} \\
&= 1.
\end{aligned}$$

Similarly, $\lim_{m\to\infty} \frac{h_{i+1}}{h_i} = 1$ when $s = 0$ because h_i is a continuous function of s. Therefore, for any $\epsilon > 0$, there is an M_1 such that for any $m > M_1$,

$$0 < h_{i+1} - h_i < \epsilon,$$

for any $i = 1, \ldots, m-1$. Letting $\epsilon \to 0$, $M_1 \to \infty$ because $h_{i+1} \neq h_i$ for any fixed m.

For any fixed $m > M_1$ and any $0 \le t < \psi$, the probability in state k at time t is (2.3.9) with initial conditions $P_1(0) = 1$ and $P_k(0) = 0$ for $k = 2, \ldots, m$. The solution is

$$\begin{cases}
P_1(t) &= e^{-(\lambda+h_1)t}, \\
P_k(t) &= \lambda e^{-(\lambda+h_k)t} \int_0^t e^{(\lambda+h_k)u} P_{k-1}(u)\mathrm{d}u, \text{ for } k = 2, \ldots, m-1, \\
P_m(t) &= \lambda e^{-h_m t} \int_0^t e^{h_m u} P_{m-1}(u)\mathrm{d}u.
\end{cases}$$

The probability of being in state 2 at time t is

$$\begin{aligned}
P_2(t) &= \lambda e^{-(\lambda+h_2)t} \int_0^t e^{(\lambda+h_2)u} e^{-(\lambda+h_1)u} \mathrm{d}u \\
&= \lambda e^{-(\lambda+h_2)t} \int_0^t e^{(h_2-h_1)u} \mathrm{d}u \\
&= \lambda e^{-(\lambda+h_2)t} \left(\int_0^t 1 + (h_2 - h_1)u + o(\epsilon) \right) \mathrm{d}u \\
&= \lambda e^{-(\lambda+h_2)t} (t + c_2'(t)) \\
&= \lambda t e^{-(\lambda+h_2)t} (1 + c_2(t)),
\end{aligned}$$

where the third equality uses both the Taylor series on $e^{(h_2-h_1)u}$ and the fact that $0 < h_2 - h_1 < \epsilon$. Furthermore, $c_2(t) = \frac{c_2'(t)}{t}$ and $c_2'(t) = \int_0^t (h_2 - h_1)u + o(\epsilon)\mathrm{d}u \geq 0$. Since $h_2 - h_1 < \epsilon$,

$$\begin{aligned}
\int_0^t (h_2 - h_1)u + o(\epsilon)\mathrm{d}u &< \epsilon \int_0^t u\mathrm{d}u + \int_0^t o(\epsilon)\mathrm{d}u \\
&= \epsilon \frac{t^2}{2} + o(\epsilon)t \\
&= \epsilon t \left(\frac{t}{2} + o(\epsilon) \right),
\end{aligned}$$

where $\lim_{\epsilon \to 0} \epsilon t \left(\frac{t}{2} + o(\epsilon) \right) = 0$ and $\lim_{\epsilon \to 0} \frac{\epsilon t \left(\frac{t}{2} + o(\epsilon) \right)}{\epsilon} = \frac{t^2}{2}$, yielding $c_2(t) = O(\epsilon)$.

Suppose, for any $0 \leq t < \psi$,

$$P_k(t) = \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k)t} (1 + c_k(t)), \tag{4.1.1}$$

where $c_k(t) \geq 0$ and $c_k(t) = O(\epsilon)$. Specifically, $\lim_{\epsilon \to 0} c_k(t) = 0$ and there is a positive number C_k such that $c_k(t) \leq C_k \epsilon$. Then, the probability of being in state $k+1$ at time t is

$$\begin{aligned}
P_{k+1}(t) &= \lambda e^{-(\lambda+h_{k+1})t} \int_0^t e^{(\lambda+h_{k+1})u} \frac{(\lambda u)^{k-1}}{(k-1)!} e^{-(\lambda+h_k)u} (1 + c_k(u))\mathrm{d}u \\
&= \frac{\lambda^k}{(k-1)!} e^{-(\lambda+h_{k+1})t} \int_0^t u^{k-1} e^{(h_{k+1}-h_k)u} (1 + c_k(u))\mathrm{d}u.
\end{aligned}$$

Recall that $0 < h_{k+1} - h_k < \epsilon$. Then, applying the Taylor series expansion on $e^{(h_{k+1}-h_k)u}$, we get

$$\begin{aligned}
\int_0^t u^{k-1} e^{(h_{k+1}-h_k)u} \mathrm{d}u &= \int_0^t u^{k-1} (1 + (h_{k+1} - h_k)u + o(\epsilon))\mathrm{d}u \\
&= \int_0^t u^{k-1} + (h_{k+1} - h_k)u^k + o(\epsilon)u^{k-1} \mathrm{d}u \\
&= \frac{t^k}{k} + c_k''(t),
\end{aligned}$$

or

$$\int_0^t u^{k-1} e^{(h_{k+1}-h_k)u} \mathrm{d}u = \frac{t^k}{k} (1 + c_k'(t)), \tag{4.1.2}$$

where $c_k'(t) = \frac{kc_k''(t)}{t^k}$ and $c_k''(t) = \int_0^t (h_{k+1} - h_k)u^k + o(\epsilon)u^{k-1}\mathrm{d}u \geq 0$. Since $h_{k+1} - h_k < \epsilon$,

$$\begin{aligned}\int_0^t (h_{k+1} - h_k)u^k + o(\epsilon)u^{k-1}\mathrm{d}u &\leq \epsilon \int_0^t u^k \mathrm{d}u + o(\epsilon)\int_0^t u^{k-1}\mathrm{d}u \\ &= t^k\left(\frac{t\epsilon}{k+1} + \frac{o(\epsilon)}{k}\right).\end{aligned}$$

Therefore, $c_k'(t) \leq \frac{kt\epsilon}{k+1} + o(\epsilon)$ and $c_k'(t) = O(\epsilon)$. And

$$\begin{aligned}\int_0^t u^{k-1}e^{(h_{k+1}-h_k)u}c_k(u)\mathrm{d}u &\leq \max_{u\in[0,t]}(c_k(u))\int_0^t u^{k-1}e^{(h_{k+1}-h_k)u}\mathrm{d}u \\ &= C_k'\frac{t^k}{k}\left(1 + c_k'(t)\right),\end{aligned}$$

by letting $C_k' = \max_{u\in[0,t]}(c_k(u))$. Then, $0 \leq C_k' = O(\epsilon)$ because $c_k(t) = O(\epsilon)$ for any t in $[0, \psi)$. There exists a $0 \leq C_k'(t) = O(\epsilon)$ such that

$$\int_0^t u^{k-1}e^{(h_{k+1}-h_k)u}c_k(u)\mathrm{d}u = C_k'(t)\frac{t^k}{k}(1 + c_k'(t)). \tag{4.1.3}$$

Therefore, by (4.1.2) and (4.1.3),

$$\int_0^t u^{k-1}e^{(h_{k+1}-h_k)u}(1 + c_k(u))\mathrm{d}u = \frac{t^k}{k}(1 + c_{k+1}(t)),$$

where $c_{k+1}(t) = (1 + c_k'(t))(1 + C_k'(t)) - 1$.

$$(1 + c_k'(t))(1 + C_k'(t)) - 1 = c_k'(t) + C_k'(t) + c_k'(t)C_k'(t),$$

yielding $0 \leq c_{k+1}(t) = O(\epsilon)$ because $c_k'(t) = O(\epsilon)$ and $C_k'(t) = O(\epsilon)$ are non-negative. As a result,

$$P_{k+1}(t) = \frac{(\lambda t)^k}{k!}e^{-(\lambda+h_{k+1})t}(1 + c_{k+1}(t)).$$

By induction, (4.1.1) holds for $k = 1, \ldots, m-1$. When $k = m$, we have

$$\begin{aligned}P_m(t) &= \lambda e^{-h_m t}\int_0^t e^{h_m u}\frac{(\lambda u)^{m-2}}{(m-2)!}e^{-(\lambda+h_{m-1})u}(1 + c_{m-1}(u))\mathrm{d}u \\ &= \lambda e^{-h_m t}\int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!}e^{-\lambda u}e^{(h_m - h_{m-1})u}(1 + c_{m-1}(u))\mathrm{d}u.\end{aligned}$$

By the fact that $e^{(h_m - h_{m-1})u} \geq 1$ for any $u \geq 0$, we have

$$\begin{aligned}\lambda\int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!}e^{-\lambda u}e^{(h_m - h_{m-1})u}\mathrm{d}u &\geq \lambda\int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!}e^{-\lambda u}\mathrm{d}u \\ &= \int_0^{\lambda t}\frac{w^{m-2}}{(m-2)!}e^{-w}\mathrm{d}w \\ &= 1 - \sum_{j=0}^{m-2}\frac{(\lambda t)^j}{j!}e^{-\lambda t},\end{aligned}$$

or

$$\lambda \int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!} e^{-\lambda u} e^{(h_m - h_{m-1})u} \mathrm{d}u \geq \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t}, \tag{4.1.4}$$

where $w = \lambda u$. The integration $\int_0^{\lambda t} \frac{w^{m-2}}{(m-2)!} e^{-w} \mathrm{d}w$ is calculated using integration by parts recursively. The last equation holds because $\sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t} = 1$.

On the other hand, by the fact that $e^{(h_m - h_{m-1})s_1} < e^{(h_m - h_{m-1})s_2}$ for any $0 \leq s_1 < s_2$, we have

$$\lambda \int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!} e^{-\lambda u} e^{(h_m - h_{m-1})u} \mathrm{d}u \leq e^{(h_m - h_{m-1})t} \lambda \int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!} e^{-\lambda u} \mathrm{d}u,$$

or

$$\lambda \int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!} e^{-\lambda u} e^{(h_m - h_{m-1})u} \mathrm{d}u \leq e^{(h_m - h_{m-1})t} \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t}. \tag{4.1.5}$$

Since $0 < h_m - h_{m-1} < \epsilon$, $1 \leq e^{(h_m - h_{m-1})t} = 1 + O(\epsilon)$ by the Taylor series. Hence, by (4.1.4) and (4.1.5),

$$\lambda \int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!} e^{-\lambda u} e^{(h_m - h_{m-1})u} \mathrm{d}u = (1 + c'_m(t)) \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t},$$

where $c'_m(t) \geq 0$ and $c'_m(t) = O(\epsilon)$. Similar to the arguments in the previous proof of (4.1.3), we have

$$\lambda \int_0^t \frac{(\lambda u)^{m-2}}{(m-2)!} e^{-\lambda u} e^{(h_m - h_{m-1})u} c_{m-1}(u) \mathrm{d}u = c''_m(t)(1 + c'_m(t)) \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t},$$

where $c''_m(t) \geq 0$ and $c''_m(t) = O(\epsilon)$. As a result,

$$P_m(t) = (1 + c_m(t)) e^{-h_m t} \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t}, \tag{4.1.6}$$

where $c_m(t) = (1 + c'_m(t))(1 + c''_m(t)) - 1 = c'_m(t) + c''_m(t) + c'_m(t)c''_m(t)$. Meanwhile, $c_m(t) \geq 0$ and $c_m(t) = O(\epsilon)$ by checking

$$c'_m(t) \geq 0,$$
$$c''_m(t) \geq 0,$$

and

$$\lim_{\epsilon \to 0} c_m(t) = \lim_{\epsilon \to 0} c'_m(t) + c''_m(t) + c'_m(t)c''_m(t) = 0,$$
$$\lim_{\epsilon \to 0} \frac{c_m(t)}{\epsilon} = \lim_{\epsilon \to 0} \frac{c'_m(t)}{\epsilon} + \frac{c''_m(t)}{\epsilon} + \frac{c'_m(t)c''_m(t)}{\epsilon} = c''(t),$$

where $c''(t) = \lim_{\epsilon\to 0} \frac{c'_m(t)}{\epsilon} + \frac{c''_m(t)}{\epsilon}$ is a positive number.

Now, we show that $P_m(t) \to 0$ as $m \to \infty$. Let a random variable X follow a Poisson distribution with rate λt. Then, $P(X \geq m-1) = \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t}$. On the other hand, $\lambda t = m\frac{t}{\psi} < m$ by recalling $\lambda = m/\psi$ and $0 \leq t < \psi$. Furthermore, when $0 \leq t < \psi\left(1 - \frac{1}{m}\right)$, we have $\lambda t = \frac{mt}{\psi} < m - 1$. According to a Chernoff bound argument (Upfal, 2005), which is $P(X \geq x) \leq \frac{(e\lambda)^x e^{-\lambda}}{x^x}$ for $x > \lambda$ and $X \sim \text{Poisson}(\lambda)$, when $m - 1 > \lambda t$,

$$\begin{aligned} P(X \geq m-1) &\leq \frac{(e\lambda t)^{m-1} e^{-\lambda t}}{(m-1)^{m-1}} \\ &= \left(\frac{m}{m-1}\right)^{m-1} \left(\frac{t}{\psi}\right)^{m-1} e^{m-1-mt/\psi} \\ &= \left(1 + \frac{1}{m-1}\right)^{m-1} e^{(m-1)(1-t/\psi+\log(t/\psi))} e^{-t/\psi}. \end{aligned}$$

Suppose $g(x) = 1 - x + \log(x)$ where $x \in [0, 1]$. Then, the first derivative of $g(x)$ is

$$\frac{\mathrm{d}g(x)}{\mathrm{d}x} = -1 + \frac{1}{x} \geq 0.$$

Therefore, $g(x)$ is increasing in $[0, 1]$. Combining with $g(1) = 0$, we have $g(x) \leq 0$ for any $x \in [0, 1]$.

Hence, $1 - t/\psi + \log(t/\psi) < 0$ for any $0 \leq t < \psi$. On the other hand, $\lim_{m\to\infty} \left(1 + \frac{1}{m-1}\right)^{m-1} = e$. As a result,

$$\lim_{m\to\infty} \left(1 + \frac{1}{m-1}\right)^{m-1} e^{(m-1)(1-t/\psi+\log(t/\psi))} e^{-t/\psi} = 0,$$

yielding the upper bound of $\lim_{m\to\infty} P(X \geq m-1)$ is 0. Since $\lim_{m\to\infty} P(X \geq m-1)$ is non-negative, we have $\lim_{m\to\infty} P(X \geq m-1) = 0$, or $\sum_{j=m}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} e^{-\lambda t} \to 0$ as $m \to \infty$. According to (4.1.6), this yields $P_m(t) \to 0$ as $m \to \infty$.

Now we are going to prove $\frac{P_m(t)}{\sum_{k=1}^{m-1} P_k(t)} \to 0$ as $m \to \infty$. We have shown that $\sum_{j=m}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} e^{-\lambda t} \leq \left(1 + \frac{1}{m-1}\right)^{m-1} e^{(m-1)(1-t/\psi+\log(t/\psi))} e^{-t/\psi}$. Therefore,

$$\begin{aligned} \sum_{j=1}^{m-1} \frac{(\lambda t)^{j-1}}{(j-1)!} e^{-\lambda t} &= 1 - \sum_{j=m}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} e^{-\lambda t} \\ &\geq 1 - \left(1 + \frac{1}{m-1}\right)^{m-1} e^{(m-1)(1-t/\psi+\log(t/\psi))} e^{-t/\psi}. \end{aligned}$$

By (4.1.1) and (4.1.6), we have

$$\begin{aligned}\frac{P_m(t)}{\sum_{k=1}^{m-1} P_k(t)} &= \frac{(1+c_m(t))e^{-h_m t}\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}{\sum_{k=1}^{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}(1+c_k(t))}\\ &= \frac{(1+c_m(t))\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}{\sum_{k=1}^{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k-h_m)t}(1+c_k(t))}\\ &\le \frac{(1+c_m(t))\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}{\sum_{k=1}^{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t}}\\ &\le \frac{(1+c_m(t))\left(1+\frac{1}{m-1}\right)^{m-1}e^{(m-1)(1-t/\psi+\log(t/\psi))}e^{-t/\psi}}{1-\left(1+\frac{1}{m-1}\right)^{m-1}e^{(m-1)(1-t/\psi+\log(t/\psi))}e^{-t/\psi}},\end{aligned}$$

which converges to 0 as $m\to\infty$ by recalling $\lim_{m\to\infty}\left(1+\frac{1}{m-1}\right)^{m-1}e^{(m-1)(1-t/\psi+\log(t/\psi))}e^{-t/\psi}=0$. The first inequality above holds because $e^{(h_m-h_k)t}\ge 1$ and $c_k(t)\ge 0$. Since $\frac{P_m(t)}{\sum_{k=1}^{m-1}P_k(t)}\ge 0$, $\frac{P_m(t)}{\sum_{k=1}^{m-1}P_k(t)}\to 0$ as $m\to\infty$. Meanwhile,

$$\frac{P_m(t)}{\sum_{k=1}^{m}P_k(t)} = \frac{\frac{P_m(t)}{\sum_{k=1}^{m-1}P_k(t)}}{1+\frac{P_m(t)}{\sum_{k=1}^{m-1}P_k(t)}} = \frac{\sum_{k=m}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}(1+c_k(t))+(1+c_m(t))e^{-h_m t}\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}},$$

which converges to 0 as $m\to\infty$.

Now we are ready to prove the limit of $\mathbb{E}(Z_{t,m}|Y_{t,m}\in E)$ as m goes to infinity. For any fixed m, according to (4.1.1) and (4.1.6),

$$\begin{aligned}\mathbb{E}(Z_{t,m}|Y_{t,m}\in E) &= \frac{\sum_{k=1}^{m}\frac{k-1}{m-1}P_k(t)}{\sum_{k=1}^{m}P_k(t)}\\ &= \frac{\sum_{k=1}^{m-1}\frac{k-1}{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}(1+c_k(t))+(1+c_m(t))e^{-h_m t}\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}{\sum_{k=1}^{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}(1+c_k(t))+(1+c_m(t))e^{-h_m t}\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}\\ &= \frac{\sum_{k=1}^{\infty}\min\left(\frac{k-1}{m-1},1\right)\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))},\end{aligned}$$

where $h_k^*=h_k$, $c_k^*(t)=c_k(t)$ for $k=1,\dots,m$ and $h_k^*=h_m$, $c_k^*(t)=c_m(t)$ for $k>m$.

Consider two sequences $a_m=\frac{\sum_{k=1}^{\infty}\min\left(\frac{k-1}{m-1},1\right)\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$ and $b_m=\frac{\sum_{k=1}^{\infty}\frac{k-1}{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$, then $a_m\le b_m$, or $b_m-a_m\ge 0$ for any m. On the other hand,

$$b_m-a_m = \frac{\sum_{k=m+1}^{\infty}\left(\frac{k-1}{m-1}-1\right)\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \tag{4.1.7}$$

$$= \frac{\frac{\lambda t}{m-1}\sum_{k=m}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_m)t}(1+c_m(t))-\sum_{k=m+1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}, \tag{4.1.8}$$

where $\frac{\lambda t}{m-1}$ converges to $\frac{t}{\psi}$, $\frac{\sum_{k=m}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$ and $\frac{\sum_{k=m+1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$ converge to 0 as $m\to\infty$. Therefore, $\lim_{m\to\infty} b_m = \lim_{m\to\infty} a_m$.

Now,

$$\begin{aligned}\lim_{m\to\infty}\frac{\sum_{k=1}^{\infty}\frac{k-1}{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} &= \lim_{m\to\infty}\frac{\lambda t}{m-1}\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}\\ &= \frac{t}{\psi},\end{aligned}$$

yielding $\lim_{m\to\infty}\mathbb{E}(Z_{t,m}|Y_{t,m}\in E) = \frac{t}{\psi}$. The last equation holds because

$$\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} = \frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+d_k^*(t))(1+c_k^*(t))},$$

by letting $e^{(h_{k+1}^*-h_k^*)t} = 1+d_k^*(t)$, where $0\le d_k^*(t) = O(\epsilon)$ when $k<m$, and $d_k^*(t)=0$ when $k\ge m$.

On the other hand, as $m\to\infty$, we can let $\epsilon\to 0$. Therefore,

$$\lim_{m\to\infty}\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+d_k^*(t))(1+c_k^*(t))} = 1.$$

Similarly, we can prove the limit of $\mathbb{E}(Z_{t,m}^2|Y_{t,m}\in E)$ as m approaches infinity. For any fixed m,

$$\begin{aligned}\mathbb{E}(Z_{t,m}^2|Y_{t,m}\in E) &= \frac{\sum_{k=1}^{m}\left(\frac{k-1}{m-1}\right)^2 P_k(t)}{\sum_{k=1}^{m}P_k(t)}\\ &= \frac{\sum_{k=1}^{m-1}\left(\frac{k-1}{m-1}\right)^2\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}(1+c_k(t)) + (1+c_m(t))e^{-h_m t}\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}{\sum_{k=1}^{m-1}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k)t}(1+c_k(t)) + (1+c_m(t))e^{-h_m t}\sum_{j=m}^{\infty}\frac{(\lambda t)^{j-1}}{(j-1)!}e^{-\lambda t}}\\ &= \frac{\sum_{k=1}^{\infty}\min\left(\left(\frac{k-1}{m-1}\right)^2,1\right)\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}.\end{aligned}$$

Similar to the arguments in (4.1.7), $\mathbb{E}(Z_{t,m}^2|Y_{t,m}\in E)$ converges to the following value as m

approaches infinity. Once again, recall that $\lambda = m/\psi$, and so

$$\begin{aligned}
&\lim_{m\to\infty} \frac{\sum_{k=1}^{\infty}\left(\frac{k-1}{m-1}\right)^2 \frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&= \lim_{m\to\infty} \frac{\lambda t}{(m-1)^2}\frac{\sum_{k=1}^{\infty} k\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&= \lim_{m\to\infty} \frac{\lambda t}{(m-1)^2}\frac{\sum_{k=1}^{\infty}(k-1)\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t)) + \sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&= \lim_{m\to\infty} \frac{(\lambda t)^2}{(m-1)^2}\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+2}^*)t}(1+c_{k+2}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} + \frac{\lambda t}{(m-1)^2}\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&= \lim_{m\to\infty} \frac{(mt)^2}{((m-1)\psi)^2}\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+2}^*)t}(1+c_{k+2}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} + \\
&\qquad \lim_{m\to\infty} \frac{mt}{(m-1)^2\psi}\frac{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_{k+1}^*)t}(1+c_{k+1}^*(t))}{\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&= \frac{t^2}{\psi^2},
\end{aligned}$$

giving $\lim_{m\to\infty}\mathbb{E}(Z_{t,m}^2|Y_{t,m}\in E) = \frac{t^2}{\psi^2}$. Hence, we have

$$\begin{aligned}
\lim_{m\to\infty}\mathrm{Var}(Z_{t,m}|Y_{t,m}\in E) &= \lim_{m\to\infty}\mathbb{E}\left((Z_{t,m})^2|Y_{t,m}\in E\right) - \left(\lim_{m\to\infty}\mathbb{E}(Z_{t,m}|Y_{t,m}\in E)\right)^2 \\
&= \frac{t^2}{\psi^2} - \left(\frac{t}{\psi}\right)^2 \\
&= 0.
\end{aligned}$$

Write $Z'_{t,m} := Z_{t,m}|Y_{t,m}\in E$. For any $\epsilon_1 > 0$, we have

$$P\left(|Z'_{t,m} - \mathbb{E}(Z'_{t,m})| \geq \epsilon_1\right) \leq \frac{\mathrm{Var}(Z'_{t,m})}{\epsilon_1^2},$$

by Chebyshev's inequality.

Since $\lim_{m\to\infty}\mathbb{E}(Z'_{t,m}) = \frac{t}{\psi}$, for any $\xi > 0$, there is an M_2, such that for any $m > M_2$,

$$\left|\mathbb{E}(Z'_{t,m}) - \frac{t}{\psi}\right| < \xi. \tag{4.1.9}$$

Rearranging (4.1.9), we have $\frac{t}{\psi} - \xi < \mathbb{E}(Z'_{t,m}) < \frac{t}{\psi} + \xi$, which implies $\left|Z'_{t,m} - \mathbb{E}(Z'_{t,m})\right| > \left|Z'_{t,m} - \frac{t}{\psi}\right| - \xi$. Therefore, if $\left|Z'_{t,m} - \frac{t}{\psi}\right| - \xi > \epsilon_1$ then $\left|Z'_{t,m} - \mathbb{E}(Z'_{t,m})\right| > \epsilon_1$, or the set of $\left|Z'_{t,m} - \frac{t}{\psi}\right| > \epsilon_1 + \xi$

is a subset of $\left|Z'_{t,m} - \mathbb{E}(Z'_{t,m})\right| > \epsilon_1$, yielding $P\left(\left|Z'_{t,m} - \mathbb{E}(Z'_{t,m})\right| > \epsilon_1\right) \geq P\left(\left|Z'_{t,m} - \frac{t}{\psi}\right| > \epsilon_1 + \xi\right)$. Combine with the Chebyshev's inequality,

$$\begin{aligned} P\left(\left|Z'_{t,m} - \frac{t}{\psi}\right| > \epsilon_1 + \xi\right) &\leq P\left(\left|Z'_{t,m} - \mathbb{E}(Z'_{t,m})\right| > \epsilon_1\right) \\ &\leq \frac{\mathrm{Var}(Z'_{t,m})}{\epsilon_1^2}. \end{aligned}$$

By letting $m \to \infty$,

$$\lim_{m\to\infty} P\left(\left|Z'_{t,m} - \frac{t}{\psi}\right| \geq \epsilon_1 + \xi\right) \leq \lim_{m\to\infty} \frac{\mathrm{Var}(Z'_{t,m})}{\epsilon_1^2} = 0,$$

so that

$$\lim_{m\to\infty} P\left(\left|Z'_{t,m} - \frac{t}{\psi}\right| \geq \epsilon_1 + \xi\right) = 0.$$

Since ϵ_1, ξ can be arbitrary small, $Z'_{t,m}$ converges in probability to $\frac{t}{\psi}$ as $m \to \infty$.

■

Invoking Theorem 4.1 the physiological age of an alive individual at calendar age t $(Z_t|Y_t \in E)$ converges in probability to the chronological age (t) as $m \to \infty$. Therefore, m cannot be too large if we require some variability on the state distribution of an alive individual at any chronological age.

The state distribution conditional on being alive at any age resembles a bell-shaped curve in the top graph of Figures 3.2 and 3.10 . Theorem 4.2 describes the shape of state distribution of an alive individual when m is large enough. The main idea of the proof is to show that the moment generating function of the state variable converges to the moment generating function of the standard normal distribution as m goes to infinity.

Recall the standard normal distribution, which is a normal distribution with mean 0 and standard deviation of 1. The respective pdf, cumulative distribution function, and moment-generating function of a random variable are denoted by $\phi(x)$, $\Phi(x)$, and $M_N(u)$, where

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ \Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \mathrm{d}u \\ M_N(u) &= e^{\frac{u^2}{2}}. \end{aligned}$$

Theorem 4.2. *A proposed PTAM has m transient states with labels* $1, \ldots, m$ *and one absorbing state with label* $m + 1$. *The ageing rate from one transient state to the next transient state is* λ. *For* $i = 1, \ldots, m$, *the dying rate in state i is* h_i, *with* $0 \leq h_1 < h_m < \infty$, *and* h_i *follows (3.3.3). Let* ψ *be the lifespan parameter. For any* $0 \leq t < \psi$, *let* $Y_{t,m}$ *be the state variable at time t.* $Y'_{t,m} = \frac{Y_{t,m} - 1 - \lambda t}{\sqrt{\lambda t}}$ *is a transformation of* $Y_{t,m}$. *Then, the distribution of* $(Y'_{t,m}|Y_{t,m} \in E)$ *converges to the standard normal distribution as* $m \to \infty$.

Proof. In Theorem 4.1's proof, for any $\epsilon > 0$ and $0 \le t < \psi$, there is an M_1 such that for any $m > M_1$, we have (4.1.1) and (4.1.6) stated as

$$P_k(t) = \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k)t}(1 + c_k(t)), \text{ for } k = 1, \ldots, m-1;$$
$$P_m(t) = (1 + c_m(t))e^{-h_m t} \sum_{j=m-1}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t},$$

where $P_k(t) = P(Y_{t,m} = k)$ and $c_k(t) = O(\epsilon)$ for $k = 1, \ldots, m$. Equivalently, there is a positive number $C(t)$ such that $c_k(t) \le C(t)\epsilon$ for $k = 1, \ldots, m$.

For any $0 \le t < \psi$, the probability of being in state k at time t conditional on being in transient states at time t is

$$P(Y_{t,m} = k | Y_{t,m} \in E) = \frac{P_k(t)}{\sum_{k=1}^{m} P_k(t)} = \frac{\frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))},$$

for $k = 1, \ldots, m-1$, and

$$P(Y_{t,m} = m | Y_{t,m} \in E) = \frac{P_m(t)}{\sum_{k=1}^{m} P_k(t)} = \frac{\sum_{k=m}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))},$$

where $h_k^* = h_k$, $c_k^*(t) = c_k(t)$ for $k = 1, \ldots, m$ and $h_k^* = h_m$, $c_k^*(t) = c_m(t)$ for $k > m$.

Let $Y_{t,m}^* = (Y_{t,m} - 1 | Y_{t,m} \in E)$ for the sake of simplicity. The domain of $Y_{t,m}$ is $\{1, \ldots, m\}$. The moment-generating function of $Y_{t,m}^*$, denoted by $M(u)$, is

$$\begin{aligned} M(u) &= \mathbb{E}(e^{uY_{t,m}^*}) \\ &= \sum_{k=1}^{m} e^{u(k-1)} P(Y_{t,m} = k | Y_{t,m} \in E) \\ &= \frac{\sum_{k=1}^{m-1} e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t)) + \sum_{k=m}^{\infty} e^{u(m-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}, \end{aligned}$$

or

$$M(u) = \frac{\sum_{k=1}^{\infty} e^{u \min(k-1, m-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}. \tag{4.1.10}$$

Let $a_m = \frac{\sum_{k=1}^{\infty} e^{u \min(k-1, m-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$ and $b_m = \frac{\sum_{k=1}^{\infty} e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$. Then, $b_m - a_m \ge 0$ for any m and $u \ge 0$. We are going to prove that as far as there is a non-negative u near 0 such that $M(u) < \infty$ and $b_m - a_m$ converges to 0 as $m \to \infty$. The proof follows similar steps in (4.1.7).

$$b_m - a_m = \frac{\sum_{k=m+1}^{\infty} \left(e^{u(k-1)} - e^{u(m-1)}\right) \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1 + c_m(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t))}$$

$$b_m - a_m = \frac{\sum_{k=m+1}^{\infty} \frac{(e^u\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} - \frac{\sum_{k=m+1}^{\infty} e^{u(m-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \quad (4.1.11)$$

Suppose $u > 0$ is near 0 such that $1 - t/\psi + \log(t/\psi) + u < 0$, then

$$\begin{aligned}
&\frac{\sum_{k=m+1}^{\infty} e^{u(m-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&\leq e^{u(m-1)} \frac{\sum_{k=m+1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{m} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \\
&= e^{u(m-1)} \frac{\sum_{k=m+1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}(1+c_m(t))}{\sum_{k=1}^{m} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*-h_m)t}(1+c_k^*(t))} \\
&\leq (1+c_m(t)) e^{u(m-1)} \frac{\sum_{k=m+1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}}{\sum_{k=1}^{m} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}}.
\end{aligned}$$

The first inequality holds because $\sum_{k=m+1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t)) \geq 0$ in the denominator. The second inequality holds because $c_k^*(t) \geq 0$ and $h_m - h_k^* \geq 0$. Recall that h_i is increasing when $h_1 < h_m$ by Theorem 3.3.

According to a Chernoff-bound argument, when $m + 1 > \lambda t$ (equivalent to $m + 1 > mt/\psi$ by recalling $\lambda = t/\psi$ and $0 \leq t < \psi$),

$$\begin{aligned}
\sum_{k=m+1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} &\leq \frac{(e\lambda t)^{m+1} e^{-\lambda t}}{(m+1)^{m+1}} \\
&= \left(\frac{m}{m+1}\right)^{m+1} \left(\frac{t}{\psi}\right)^{m+1} e^{m+1-mt/\psi} \\
&= \left(1 - \frac{1}{m+1}\right)^{m+1} e^{(m+1)(1-t/\psi+\log(t/\psi))} e^{t/\psi},
\end{aligned}$$

and

$$\sum_{k=1}^{m} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} \geq 1 - \left(1 - \frac{1}{m+1}\right)^{m+1} e^{(m+1)(1-t/\psi+\log(t/\psi))} e^{t/\psi}.$$

Therefore,

$$\frac{\sum_{k=m+1}^{\infty} e^{u(m-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} \tag{4.1.12}$$

$$\leq (1+c_m(t))e^{u(m-1)} \frac{\sum_{k=m+1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}}{\sum_{k=1}^{m} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}} \tag{4.1.13}$$

$$\leq (1+c_m(t))e^{u(m-1)} \frac{\left(1-\frac{1}{m+1}\right)^{m+1} e^{(m+1)(1-t/\psi+\log(t/\psi))} e^{t/\psi}}{1-\left(1-\frac{1}{m+1}\right)^{m+1} e^{(m+1)(1-t/\psi+\log(t/\psi))} e^{t/\psi}} \tag{4.1.14}$$

$$= (1+c_m(t))e^{-2u} \frac{\left(1-\frac{1}{m+1}\right)^{m+1} e^{(m+1)(1-t/\psi+\log(t/\psi)+u)} e^{t/\psi}}{1-\left(1-\frac{1}{m+1}\right)^{m+1} e^{(m+1)(1-t/\psi+\log(t/\psi))} e^{t/\psi}}, \tag{4.1.15}$$

where in the last equation $e^{u(m-1)}$ is moved to the numerator. Since $1 - t/\psi + \log(t/\psi) + u < 0$, $\lim_{m\to\infty} e^{(m+1)(1-t/\psi+\log(t/\psi)+u)} = 0$, and (4.1.15) converges to 0 as $m \to \infty$.

For the first term of $b_m - a_m$,

$$\frac{\sum_{k=m+1}^{\infty} \frac{(e^u\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_m)t}(1+c_m(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))} = \frac{(1+c_m(t))e^{e^u\lambda t-(\lambda+h_m)t} \sum_{k=m+1}^{\infty} \frac{(e^u\lambda t)^{k-1}}{(k-1)!} e^{-e^u\lambda t}}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}. \tag{4.1.16}$$

Using the Chernoff-bound argument again, when $m + 1 > e^u\lambda t$ (this condition is equivalent to $-\log(1 + 1/m) + \log(t/\psi) + u < 0$ which naturally holds if $1 - t/\psi + \log(t/\psi) + u < 0$),

$$\begin{aligned}
\sum_{k=m+1}^{\infty} \frac{(e^u\lambda t)^{k-1}}{(k-1)!} e^{-e^u\lambda t} &\leq \frac{(ee^u\lambda t)^{m+1}}{(m+1)^{m+1}} e^{-e^u\lambda t} \\
&= \frac{(e^{u+1}mt)^{m+1}}{\psi^{m+1}(m+1)^{m+1}} e^{-e^u\lambda t} \\
&= \left(1-\frac{1}{m+1}\right)^{m+1} e^{(m+1)(u+1+\log(t/\psi))} e^{-e^u\lambda t}.
\end{aligned}$$

The numerator of (4.1.16) is

$$\begin{aligned}
&(1+c_m(t))e^{e^u\lambda t-(\lambda+h_m)t} \sum_{k=m+1}^{\infty} \frac{(e^u\lambda t)^{k-1}}{(k-1)!} e^{-e^u\lambda t} \\
&\leq (1+c_m(t))e^{-(m/\psi+h_m)t} \left(1-\frac{1}{m+1}\right)^{m+1} e^{(m+1)(u+1+\log(t/\psi))} \\
&= (1+c_m(t)) \left(1-\frac{1}{m+1}\right)^{m+1} e^{t/\psi-h_m t} e^{(m+1)(u+1-t/\psi+\log(t/\psi))},
\end{aligned}$$

which converges to 0 as $m \to \infty$ by the fact that $u + 1 - t/\psi + \log(t/\psi) < 0$. Meanwhile, the denominator of (4.1.16) converges to 1 as $m \to \infty$. As a result, (4.1.16) converges to 0 as $m \to \infty$.

Since both (4.1.12) and (4.1.16) converge to 0 as $m \to \infty$, $b_m - a_m$ converges to 0, or $M(u) \to \frac{\sum_{k=1}^{\infty} e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1+c_k^*(t))}$, as $m \to \infty$ by (4.1.11).

On the other hand, we can check that

$$\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}(1 + c_k^*(t)) \le (1 + \max_{k=1,\ldots,m} c_k^*(t)) \sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t},$$

which converges to $\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}$ as $m \to \infty$, because $c_k^*(t) = O(\epsilon)$ and we can let $\epsilon \to 0$ as $m \to \infty$. Therefore, for any u such that $1 - t/\psi + \log(t/\psi) + u < 0$, the limit of $M(u)$ from (4.1.10) is the same as the limit of $M^*(u)$ as m goes to infinity ($\lim_{m\to\infty}(M(u) - M^*(u)) = 0$), where $M^*(u)$ is

$$M^*(u) = \frac{\sum_{k=1}^{\infty} e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}}.$$

The first derivative of $M^*(u)$ with respect to u is

$$\begin{aligned}\frac{\mathrm{d}M^*(u)}{\mathrm{d}u} &= \frac{\sum_{k=1}^{\infty}(k-1)e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}} \\ &= e^u \lambda t \frac{\sum_{k=1}^{\infty} e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_{k+1}^*)t}}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}},\end{aligned}$$

or

$$\frac{\mathrm{d}M^*(u)}{\mathrm{d}u} = e^u \lambda t \frac{\sum_{k=1}^{\infty} e^{u(k-1)} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t} e^{(h_{k+1}^* - h_k^*)t}}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda+h_k^*)t}} \tag{4.1.17}$$

Recall that for any m such that $m > M_1$, $0 < h_{k+1}^* - h_k^* < \epsilon$. Then, $1 < e^{(h_{k+1}^* - h_k^*)t} = 1 + O(\epsilon)$ by the Taylor series representation. As a result, there is a $C(t) \ge 0$ such that $e^{(h_{k+1}^* - h_k^*)t} < 1 + C(t)\epsilon$ for any k. Taking this result back to (4.1.17),

$$e^u \lambda t M^*(u) \le \frac{\mathrm{d}M^*(u)}{\mathrm{d}u} \le (1 + C(t)\epsilon) e^u \lambda t M^*(u) \tag{4.1.18}$$

As ϵ approaches 0, $\frac{\mathrm{d}M^*(u)}{\mathrm{d}u}$ approaches $e^u \lambda t M^*(u)$.

Consider the differential equation

$$\frac{\mathrm{d}X(u)}{\mathrm{d}u} = e^u \lambda t \, X(u) \tag{4.1.19}$$

with the boundary condition $X(0) = 1$. The solution of (4.1.19) is

$$X(u) = e^{\lambda t(e^u - 1)}.$$

It may be verified that

$$M^*(0) = \frac{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda + h_k^*)t}}{\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-(\lambda + h_k^*)t}} = 1.$$

From (4.1.18), for any $\epsilon > 0$ and any $m > M_1$,

$$e^{\lambda t(e^u - 1)} \leq M^*(u) \leq (1 + C(t)\epsilon) e^{\lambda t(e^u - 1)}. \tag{4.1.20}$$

On the other hand, we showed $\lim_{m\to\infty}(M(u) - M^*(u)) = 0$. For any $\xi > 0$, there is an M_2 such that for any $m > M_2$,

$$|M(u) - M^*(u)| < \xi.$$

By (4.1.20), for any $m > \max(M_1, M_2)$,

$$e^{\lambda t(e^u - 1)} - \xi \leq M(u) \leq (1 + C(t)\epsilon) e^{\lambda t(e^u - 1)} + \xi \tag{4.1.21}$$

For any m, the moment-generating function of $(Y'_{t,m} | Y_{t,m} \in E)$ is

$$\begin{aligned}
\mathbb{E}\left(e^{uY'_{t,m}} \middle| Y_{t,m} \in E\right) &= \mathbb{E}\left(e^{u\frac{Y_{t,m}-1-\lambda t}{\sqrt{\lambda t}}} \middle| Y_{t,m} \in E\right) \\
&= e^{-u\sqrt{\lambda t}}\, \mathbb{E}\left(e^{\frac{u}{\sqrt{\lambda t}}(Y_{t,m}-1)} \middle| Y_{t,m} \in E\right) \\
&= e^{-u\sqrt{\lambda t}} M\left(u/\sqrt{\lambda t}\right).
\end{aligned}$$

Based on (4.1.21), for any $m > \max(M_1, M_2)$,

$$e^{-u\sqrt{\lambda t}}\left(e^{\lambda t(e^{u/\sqrt{\lambda t}}-1)} - \xi\right) \leq e^{-u\sqrt{\lambda t}} M\left(u/\sqrt{\lambda t}\right) \leq e^{-u\sqrt{\lambda t}}\left((1 + C(t)\epsilon) e^{\lambda t(e^{u/\sqrt{\lambda t}}-1)} + \xi\right) \tag{4.1.22}$$

Recall that $\lambda = m/\psi$. By the Taylor series representation of $e^{u/\sqrt{\lambda t}} - 1$, we have

$$\begin{aligned}
e^{u/\sqrt{\lambda t}} - 1 &= \frac{u}{\sqrt{\lambda t}} + \frac{1}{2}\frac{u^2}{\lambda t} + o\left(\frac{1}{\lambda}\right) \\
&= \frac{u}{\sqrt{\lambda t}} + \frac{1}{2}\frac{u^2}{\lambda t} + o\left(\frac{1}{m}\right).
\end{aligned}$$

Therefore, by applying this result on both sides of (4.1.22), the left hand side is

$$\begin{aligned}
e^{-u\sqrt{\lambda t} + \lambda t(e^{u/\sqrt{\lambda t}}-1)} - \xi e^{-u\sqrt{\lambda t}} &= e^{-u\sqrt{\lambda t} + \lambda t\left(\frac{u}{\sqrt{\lambda t}} + \frac{1}{2}\frac{u^2}{\lambda t} + o\left(\frac{1}{m}\right)\right)} - \xi e^{-u\sqrt{mt/\psi}} \\
&= e^{\frac{1}{2}u^2 + \frac{tm}{\psi} o\left(\frac{1}{m}\right)} - \xi e^{-u\sqrt{mt/\psi}},
\end{aligned}$$

and the right hand side is

$$\begin{aligned}
(1 + C(t)\epsilon) e^{-u\sqrt{\lambda t} + \lambda t(e^{u/\sqrt{\lambda t}}-1)} + \xi e^{-u\sqrt{\lambda t}} &= (1 + C(t)\epsilon) e^{-u\sqrt{\lambda t} + \lambda t\left(\frac{u}{\sqrt{\lambda t}} + \frac{1}{2}\frac{u^2}{\lambda t} + o\left(\frac{1}{m}\right)\right)} + \xi e^{-u\sqrt{mt/\psi}} \\
&= (1 + C(t)\epsilon) e^{\frac{1}{2}u^2 + \frac{tm}{\psi} o\left(\frac{1}{m}\right)} + \xi e^{-u\sqrt{mt/\psi}},
\end{aligned}$$

where $\lim_{m\to\infty} \frac{tm}{\psi} o\left(\frac{1}{m}\right) = 0$ and $\lim_{m\to\infty} e^{-u\sqrt{mt/\psi}} = 0$. We can let $\epsilon, \xi \to 0$ as $m \to \infty$ so that $e^{-u\sqrt{\lambda t}} M\left(u/\sqrt{\lambda t}\right)$ converges to $e^{\frac{1}{2}u^2}$ by (4.1.22). Hence,

$$\lim_{m\to\infty} \mathbb{E}\left(e^{uY'_{t,m}} \middle| Y_{t,m} \in E\right) = e^{\frac{1}{2}u^2},$$

which is equal to the moment-generating function of the standard normal distribution. Since the moment-generating function and the distribution have a one-to-one mapping, the distribution of $(Y'_{t,m}|Y_{t,m} \in E)$ converges to the standard normal distribution as $m \to \infty$.

■

Recall that we defined $Z_t = \frac{Y_t - 1}{m-1}$ as the physiological age index. From Theorem 4.2, when m is large enough, we have

$$P\left(Z_t \le \frac{k-1}{m-1} \middle| Y_t \in E\right) = P\left(Y_t \le k | Y_t \in E\right) \approx \Phi\left(\frac{k-1-\lambda t}{\sqrt{\lambda t}}\right), \text{ for } k = 1, \ldots, m.$$

The approximation can be used as a quick assessment of the state distribution when m is large.

Example 4.1. We fix all parameters, except m, to the estimated values based on the Channing house data ($h_1 = 0.0017, h_m = 1.2750, s = -0.0735, \psi = 55$), and plot the state distribution $P(Z_t \le \frac{k-1}{m-1} | Y_t \in E)$ at age 80 ($t = 30$ because we assumed ageing starts at age 50 in the analysis) and the normal distribution with mean $1 + \lambda t$ standard deviation $\sqrt{\lambda t}$ for different m's in Figure 4.1. Two lines are distinguishable in Figure 4.1 when $m = 100$, whilst two lines overlap gradually as m increases. In particular, two lines overlap with each other when $m = 10,000$. Therefore, the result shows the transformation of the state variable converges to the normal distribution.

△

4.2 Lifetime distribution conditional on the current state

Suppose we can observe the current state, then what is the updated distribution? Theorem 4.3 tells us that the updated distribution still follows the PTAM structure but with new parameter values. The intuition is that the current state can be treated as state 1 for a new Markov chain.

Theorem 4.3. *Let an individual X follow the process of the proposed PTAM with parameter values $(h_1 = h_1^*, h_m = h_m^*, s = s^*, \psi = \psi^*, m = m^*)$. At age $t > 0$, the individual is observed in state i, where i can be any integer between 1 and m^*. Conditional on the observed state, the individual follows the process of the proposed PTAM with updated parameter values $(h_1 = h_i^*, h_m = h_m^*, s = s^*, \psi = \psi^* \frac{m^*-i+1}{m^*}, m = m^* - i + 1)$, where h_i^* is calculated through (3.3.3).*

Proof. For any age $t > 0$, the individual is observed in state i. Consider another individual X^* that follows the process of the proposed PTAM with parameter values $(h_1 = h_i^*, h_m = h_m^*, s = s^*, \psi = \psi^* \frac{m^*-i+1}{m^*}, m = m^* - i + 1)$. For $j = i, \ldots, m^*$, the dying rate in state j for X is

$$h_j = \begin{cases} \left(\frac{m^*-j}{m^*-1}(h_1^*)^{s^*} + \frac{j-1}{m^*-1}(h_m^*)^{s^*}\right)^{1/s^*} & s^* \neq 0, \\ (h_1^*)^{\frac{m^*-j}{m^*-1}} (h_m^*)^{\frac{j-1}{m^*-1}} & s^* = 0. \end{cases}$$

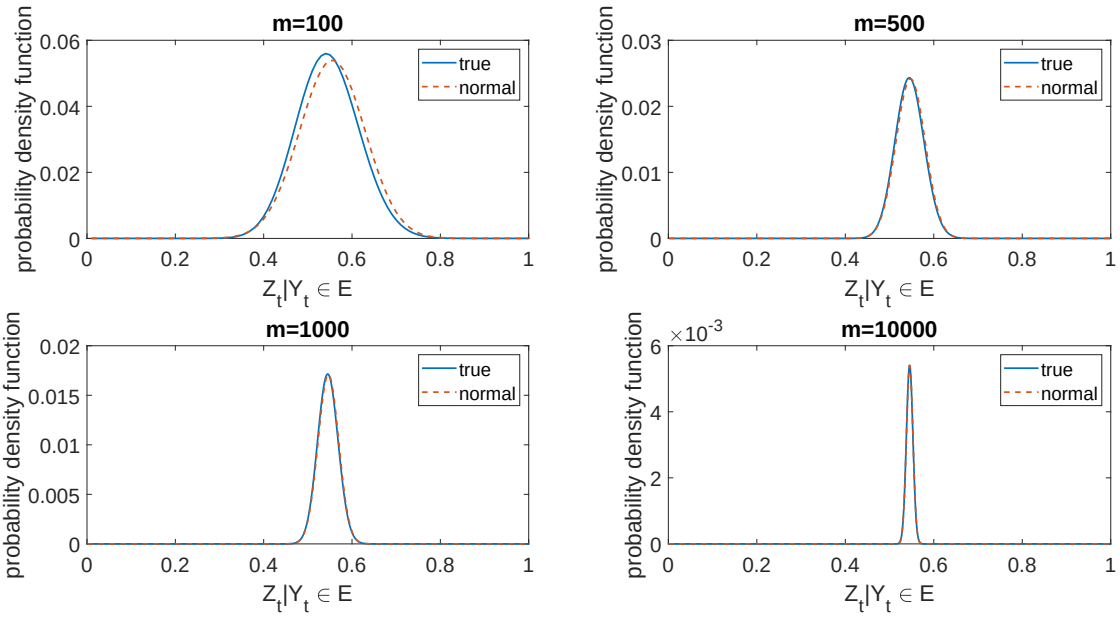

Figure 4.1: State distribution conditional on being alive is approximated by a normal distribution.

Also, the dying rate in state $j-i+1$ for X^* is

$$h_{j-i+1} = \begin{cases} \left(\frac{(m^*-i+1)-(j-i+1)}{(m^*-i+1)-1}(h_i^*)^{s^*} + \frac{(j-i+1)-1}{(m^*-i+1)-1}(h_m^*)^{s^*}\right)^{1/s^*} & s^* \neq 0, \\ (h_i^*)^{\frac{(m^*-i+1)-(j-i+1)}{(m^*-i+1)-1}}\,(h_m^*)^{\frac{(j-i+1)-1}{(m^*-i+1)-1}} & s^* = 0, \end{cases}$$

which is the same as the dying rate in state j for X because of the following facts:

When $s^* \neq 0$,

$$(h_i^*)^{s^*} = \frac{m^*-i}{m^*-1}(h_1^*)^{s^*} + \frac{i-1}{m^*-1}(h_m^*)^{s^*};$$

then,

$$\begin{aligned} &\frac{(m^*-i+1)-(j-i+1)}{(m^*-i+1)-1}(h_i^*)^{s^*} + \frac{(j-i+1)-1}{(m^*-i+1)-1}(h_m^*)^{s^*} \\ &= \frac{m^*-j}{m^*-i}(h_i^*)^{s^*} + \frac{j-i}{m^*-i}(h_m^*)^{s^*} \\ &= \frac{m^*-j}{m^*-i}\left(\frac{m^*-i}{m^*-1}(h_1^*)^{s^*} + \frac{i-1}{m^*-1}(h_m^*)^{s^*}\right) + \frac{j-i}{m^*-i}(h_m^*)^{s^*} \\ &= \frac{m^*-j}{m^*-1}(h_1^*)^{s^*} + \frac{j-1}{m^*-1}(h_m^*)^{s^*}, \end{aligned}$$

When $s^* = 0$,

$$h_i^* = (h_1^*)^{\frac{m^*-i}{m^*-1}}\,(h_m^*)^{\frac{i-1}{m^*-1}},$$

then

$$\begin{aligned}(h_i^*)^{\frac{(m^*-i+1)-(j-i+1)}{(m^*-i+1)-1}}\,(h_m^*)^{\frac{(j-i+1)-1}{(m^*-i+1)-1}} &= (h_i^*)^{\frac{m^*-j}{m^*-i}}\,(h_m^*)^{\frac{j-i}{m^*-i}}\\ &= \left((h_1^*)^{\frac{m^*-i}{m^*-1}}\,(h_m^*)^{\frac{i-1}{m^*-1}}\right)^{\frac{m^*-j}{m^*-i}}\,(h_m^*)^{\frac{j-i}{m^*-i}}\\ &= (h_1^*)^{\frac{m^*-j}{m^*-1}}\,(h_m^*)^{\frac{j-1}{m^*-1}},\end{aligned}$$

where

$$\begin{aligned}\frac{i-1}{m^*-1}\frac{m^*-j}{m^*-i}+\frac{j-i}{m^*-i} &= \frac{(i-1)(m^*-j)+(j-i)(m^*-1)}{(m^*-1)(m^*-i)}\\ &= \frac{(j-1)(m^*-i)}{(m^*-1)(m^*-i)}.\end{aligned}$$

On the other hand, the ageing rate for X is

$$\lambda = \frac{m^*}{\psi^*},$$

and the ageing rate for X^* is

$$\lambda = \frac{m^*-i+1}{\psi^*\frac{m^*-i+1}{m^*}} = \frac{m^*}{\psi^*}.$$

Therefore, the ageing rates are the same for X and X^*. As a result, the individual X conditional on being in state i at time t follows the same process as individual X^* when treating time t for X as time 0 for X^*. Hence, the updated process for X is the same as that for X^*.

■

By Theorem 4.3, the process and the resulting lifetime distribution could be updated for an individual if its current state is observed. For example, suppose an individual follows the PTAM with parameter values estimated from the Channing house data, and the individual is at age 80. If the individual is at physiological age 70, 80, or 90 from (3.4.5), then the corresponding updated parameter values are in Table 4.1 and the distributions are in Figure 4.2. It is clear that the hazard rate is higher if the individual is in a higher physiological age as expected.

Table 4.1: Updated PTAM given the current physiological age.

current physiological age	h_1	h_m	s	ψ	m
70	0.1213	1.2750	-0.0735	17.0500	31
80	0.2542	1.2750	-0.0735	11.5500	21
90	0.5559	1.2750	-0.0735	6.0500	11

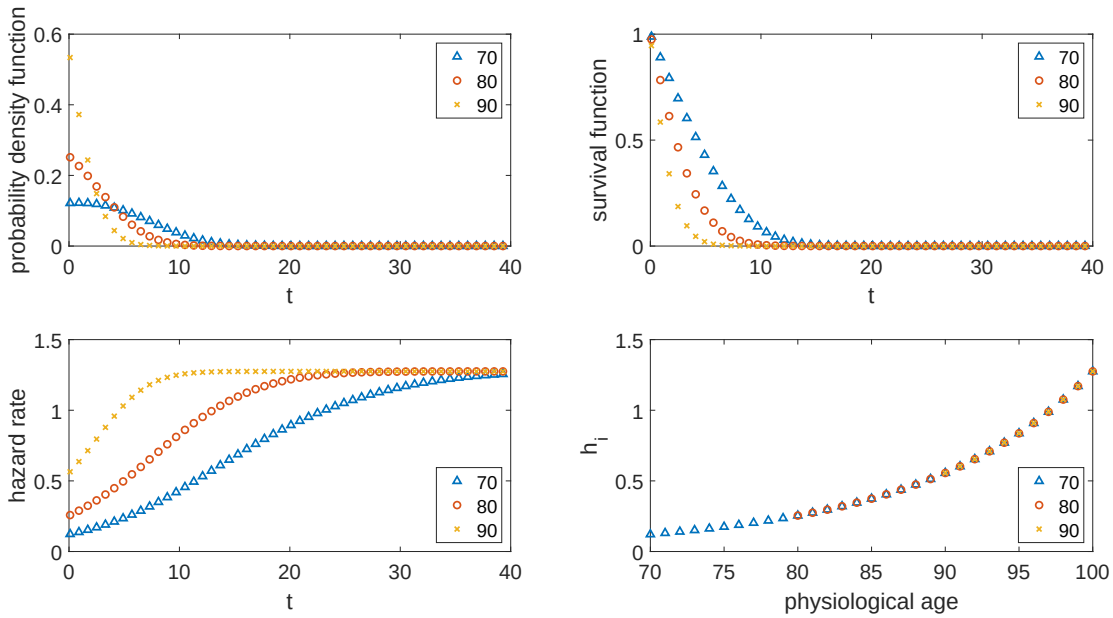

Figure 4.2: Distributions at physiological age 70, 80, and 90.

4.3 Distribution of the PTAM

In this Section, we explore some mathematical properties related to the resulting distribution of the proposed PTAM.

4.3.1 Shape and scale parameters

We refer to the parameter s as the shape parameter. A result via Theorem 4.4 is provided to justify that indeed s is the parameter that controls the distribution's shape. A preliminary result (Lemma 4.1) is needed for the proof of Theorem 4.4.

Lemma 4.1. *Let* $f(s) = \frac{\log(ab^s+(1-a)c^s)}{s}$ *when* $s \neq 0$ *and* $f(s) = a\log(b) + (1-a)\log(c)$ *when* $s = 0$*, where* $0 \leq a \leq 1$*,* $b \geq 0$*, and* $c \geq 0$ *but* b, c *cannot be* 0 *at the same time. Then,* $f(s)$ *is an increasing function of* s*.*

Proof. The proof is detailed in Appendix A.3. ∎

Theorem 4.4. *Let two random variables* X_1 *and* X_2 *follow the proposed PTAM with the same parameter values for* h_1*,* h_m*,* m*, and* ψ*. Also, let* s_1 *be the value of* s *for* X_1 *and* s_2 *the corresponding value for* X_2 *with* $s_1 < s_2$*. Then,* X_1 *is greater than* X_2 *in a stochastic-order sense, or* $S_{X_1}(t) \geq S_{X_2}(t)$ *for any* $t \geq 0$*.*

Proof. Let $h_i^{X_k}$ be the absorption rate for X_k at state i, where $i = 1, \ldots, m$ and $k = 1, 2$.

The structure on the PTAM's absorption rates is

$$h_i = \begin{cases} \left(\frac{m-i}{m-1}h_1^s + \frac{i-1}{m-1}h_m^s\right)^{1/s} & s \neq 0, \\ h_1^{\frac{m-i}{m-1}}\, h_m^{\frac{i-1}{m-1}} & s = 0, \end{cases}$$

The monotonic pattern of h_i is the same as the pattern of $\log h_i$,

$$\log h_i = \begin{cases} \frac{\log(ab^s+(1-a)c^s)}{s} & s \neq 0, \\ a\log b + (1-a)\log c & s = 0. \end{cases}$$

where $a = \frac{m-i}{m-1}, b = h_1, c = h_m$. When $h_1 = h_m = 0$, $h_i = 0$ for any $i = 1, \ldots, m$, in which the process never enters the absorbing state. Therefore, at least one of h_1 and h_m is positive.

By Lemma 4.1, $\log h_i$ is an increasing function with respect to s, which shows h_i is an increasing function with respect to s as well. Since $s_1 < s_2$, we have $h_i^{X_1} < h_i^{X_2}$.

On the other hand, since ψ and m are the same for both variable, the ageing rates for X_1 and X_2 are the same. Define $\mathbf{\Lambda_k}$ as the degenerate transition matrix for X_k, $k = 1, 2$, i.e.,

$$\mathbf{\Lambda_k} = \begin{bmatrix} -\left(\lambda + h_1^{X_k}\right) & \lambda & & & \\ & -\left(\lambda + h_2^{X_k}\right) & \lambda & & \\ & & \ddots & & \\ & & & -\left(\lambda + h_{m-1}^{X_k}\right) & \lambda \\ & & & & -h_m^{X_k} \end{bmatrix}.$$

The (i, j) entry of $\mathbf{\Lambda_k}$, denoted $\mathbf{\Lambda_k}(i, j)$, satisfies the inequality $\mathbf{\Lambda_1}(i, j) \geq \mathbf{\Lambda_2}(i, j)$ for $i, j = 1, \ldots, m$. The survival function for X_k at time $t \geq 0$ is $S_{X_k}(t) = \boldsymbol{\alpha} e^{\mathbf{\Lambda_k} t}\boldsymbol{e}$, where $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$ and $\boldsymbol{e} = (1, 1, \ldots, 1)^\intercal$. Thus,

$$\begin{aligned} S_{X_1}(t) - S_{X_2}(t) &= \boldsymbol{\alpha} e^{\mathbf{\Lambda_1} t}\boldsymbol{e} - \boldsymbol{\alpha} e^{\mathbf{\Lambda_2} t}\boldsymbol{e} \\ &= \boldsymbol{\alpha}\left(e^{\mathbf{\Lambda_1} t} - e^{\mathbf{\Lambda_2} t}\right)\boldsymbol{e} \\ &\geq 0, \end{aligned}$$

where the last inequality holds because $e^{\mathbf{\Lambda_1} t}$ is greater than or equal to $e^{\mathbf{\Lambda_2} t}$ elementwise by considering $\mathbf{\Lambda_1} = \mathbf{\Lambda_2} + \boldsymbol{A}$, where $\boldsymbol{A}$ is non-negative diagonal matrix. Therefore, X_1 is greater than X_2 in stochastic order for any $t \geq 0$.

■

The parameter s controls not only the curvature on the dying rate (more convexity with a smaller value of s), but also the stochastic order of the lifetime random variable by Theorem 4.4. Recall that a random variable X is higher than another random variable Y in stochastic order means X is greater than Y in the probability sense. Therefore, the lifetime random variable following the proposed PTAM with a smaller value of s has a longer life in the probability sense when fixing h_1, h_m, m, ψ.

Example 4.2. We use the PTAM learnt from the Channing house data to compare with the PTAM with the same h_1, h_m, m, ψ but $s = 0$ in Figure 4.3. The dying rate pattern has more convexity for $s = 0$ and the survival probability (hazard rate) is smaller (higher) for $s = 0$ at any age.

△

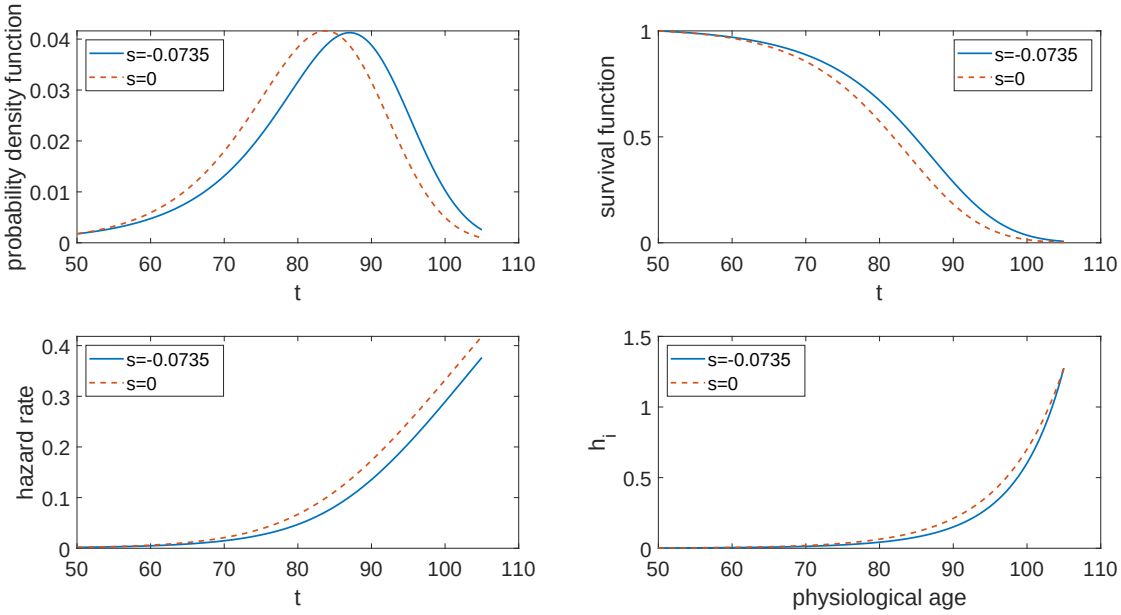

Figure 4.3: Some distributions with different s's but fixed h_1, h_m, m, ψ.

Theorem 4.5. *Consider a proposed PTAM with parameter $\theta = (h_1, h_m, s, \psi, m)$. For any fixed s and m, the parameter $\theta^* = \left(h_1, h_m, \frac{1}{\psi}\right)$ is a scale parameter.*

Proof. For any fixed s and m, let $F(t; \theta^*)$ be the resulting cumulative distribution function of the PTAM,

$$F(t; \theta^*) = 1 - \boldsymbol{\alpha} e^{\boldsymbol{\Lambda}(\boldsymbol{\theta^*})t} \boldsymbol{e},$$

where $\boldsymbol{\Lambda}(\boldsymbol{\theta^*})$ is an $m \times m$ degenerate transition matrix and its entries are determined by θ^*.

For any $a > 0$, the diagonal elements of $\boldsymbol{\Lambda}\left(\frac{\boldsymbol{\theta^*}}{\boldsymbol{a}}\right)$ are

$$-\frac{\lambda + h_i}{a} = -\frac{\lambda}{a} - \frac{h_i}{a} = -\frac{m}{a\psi} - \frac{h_i}{a},$$

for $i = 1, \ldots, m-1$ and $-\frac{h_m}{a}$ for $i = m$. The super-diagonal elements of $\boldsymbol{\Lambda}\left(\frac{\boldsymbol{\theta^*}}{\boldsymbol{a}}\right)$ are $\frac{\lambda}{a} = \frac{m}{a\psi}$. In addition, by the h_i structure, when $s \neq 0$,

$$\frac{h_i}{a} = \frac{\left(\frac{m-i}{m-1} h_1^s + \frac{i-1}{m-1} h_m^s\right)^{1/s}}{a} = \left(\frac{m-i}{m-1}\left(\frac{h_1}{a}\right)^s + \frac{i-1}{m-1}\left(\frac{h_m}{a}\right)^s\right)^{1/s},$$

and when $s = 0$,

$$\frac{h_i}{a} = \frac{h_1^{\frac{m-i}{m-1}} h_m^{\frac{i-1}{m-1}}}{a} = \left(\frac{h_1}{a}\right)^{\frac{m-i}{m-1}} \left(\frac{h_m}{a}\right)^{\frac{i-1}{m-1}}.$$

Therefore, we have

$$\frac{\boldsymbol{\Lambda}(\boldsymbol{\theta}^*)}{a} = \boldsymbol{\Lambda}\left(\frac{\boldsymbol{\theta}^*}{\boldsymbol{a}}\right). \tag{4.3.23}$$

The cumulative distribution function evaluated at $\frac{t}{a}$ with parameter θ^* is

$$F\left(\frac{t}{a}; \theta^*\right) = 1 - \boldsymbol{\alpha} e^{\boldsymbol{\Lambda}(\boldsymbol{\theta}^*)\frac{t}{a}} \boldsymbol{e} = 1 - \boldsymbol{\alpha} e^{\frac{\boldsymbol{\Lambda}(\boldsymbol{\theta}^*)}{a} t} \boldsymbol{e} = F\left(t; \frac{\theta^*}{a}\right),$$

where the last equation holds by (4.3.23). Hence, the parameter θ^* is a scale parameter.

■

With Theorem 4.5, the modellers do not need to worry about the lifetime unit (i.e. day, month, year) when applying the proposed PTAM. This is because the scale parameter can be adjusted to accommodate the unit. On the other hand, it is relatively easy to re-calibrate the model by scaling the scale parameter when altering the lifetime unit, e.g., converting the data in months to data in years.

4.3.2 The hazard rate as m goes to infinity

The resulting hazard rate of the proposed PTAM can be calculated by (2.2.4). For any fixed $h_1, h_m, s,$ and ψ, such a hazard rate is a function of m. We also give the limit of the resulting hazard rate as $m \to \infty$.

Theorem 4.6. *For any time t in $[0, \psi)$, let $h(t; m)$ be the resulting hazard rate of an m-states proposed PTAM, then $h(t; m) \to g(t)$ as $m \to \infty$, where*

$$g(t) = \begin{cases} \left(\left(h_m^s - h_1^s\right)\frac{t}{\psi} + h_1^s\right)^{1/s} & \text{when } s \neq 0; \\ h_1^{(1-t/\psi)} h_m^{t/\psi} & \text{when } s = 0. \end{cases}$$

Proof. Recall that

$$h_i = \begin{cases} \left(\frac{m-i}{m-1}h_1^s + \frac{i-1}{m-1}h_m^s\right)^{1/s} = \left(\frac{i-1}{m-1}(h_m^s - h_1^s) + h_1^s\right)^{1/s} & s \neq 0, \\ h_1^{\frac{m-i}{m-1}} h_m^{\frac{i-1}{m-1}} = h_1^{1-\frac{i-1}{m-1}} h_m^{\frac{i-1}{m-1}} & s = 0. \end{cases}$$

The resulting hazard rate is

$$\begin{aligned} h(t; m) &= \sum_{i=1}^{m} h_i P(Y_t = i | Y_t \in E) \\ &= \sum_{i=1}^{m} h_i P\left(Z_t = \frac{i-1}{m-1} \middle| Y_t \in E\right) \\ &= \mathbb{E}(g_1(Z_t)), \end{aligned}$$

where $g_1(x)$ is a monotone and continuous function of x,

$$g_1(x) = \begin{cases} \left(\left(h_m^s - h_1^s\right) x + h_1^s\right)^{1/s}, & \text{when } s \neq 0; \\ h_1^{1-x} h_m^x, & \text{when } s = 0. \end{cases}$$

By Theorem 4.1, $Z_t \to \frac{t}{\psi}$ in probability as $m \to \infty$. Therefore, for any $\epsilon > 0$,

$$\lim_{m\to\infty} P\left(\left|Z_t - \frac{t}{\psi}\right| > \epsilon\right) = 0.$$

Considering that $g_1(x)$ is a monotonic function, we have

$$\min\left(g_1\left(\frac{t}{\psi} - \epsilon\right), g_1\left(\frac{t}{\psi} + \epsilon\right)\right) \leq g_1(z) \leq \max\left(g_1\left(\frac{t}{\psi} - \epsilon\right), g_1\left(\frac{t}{\psi} + \epsilon\right)\right),$$

for any z that $\left|z - \frac{t}{\psi}\right| \leq \epsilon$.

Since the domain of Z_t is $[0, 1]$, $|Z_t - t/\psi| \leq \max(t/\psi, 1 - t/\psi)$. The expected value of $g_1(Z_t)$ can be calculated as

$$\mathbb{E}(g_1(Z_t)) = \sum_{|z-t/\psi|\leq\epsilon} g_1(z) P\left(Z_t = z\right) + \sum_{|z-t/\psi|>\epsilon} g_1(z) P\left(Z_t = z\right),$$

where the total probability in the first term converges to 1 and the total probability in the second converges to 0 as $m \to \infty$. Meanwhile, $g_1(z)$ is bounded when $0 \leq z \leq 1$. Let U be an upper bound of $g_1(z)$, $g_1(z) \leq U$ for any z, when $\epsilon < |z - t/\psi| \leq \max(t/\psi, 1 - t/\psi)$. Then, we have

$$\begin{aligned} \lim_{m\to\infty} \mathbb{E}(g_1(Z_t)) &= \lim_{m\to\infty} \sum_{|z-t/\psi|\leq\epsilon} g_1(z) P\left(Z_t = z\right) + \lim_{m\to\infty} \sum_{\epsilon<|z-t/\psi|} g_1(z) P\left(Z_t = z\right) \\ &\leq \max\left(g_1\left(\frac{t}{\psi} - \epsilon\right), g_1\left(\frac{t}{\psi} + \epsilon\right)\right) \lim_{m\to\infty} P\left(\left|z - \frac{t}{\psi}\right| \leq \epsilon\right) + U \lim_{m\to\infty} P\left(\left|z - \frac{t}{\psi}\right| > \epsilon\right) \\ &= \max\left(g_1\left(\frac{t}{\psi} - \epsilon\right), g_1\left(\frac{t}{\psi} + \epsilon\right)\right), \end{aligned}$$

and

$$\begin{aligned} \lim_{m\to\infty} \mathbb{E}(g_1(Z_t)) &= \lim_{m\to\infty} \sum_{|z-t/\psi|\leq\epsilon} g_1(z) P\left(Z_t = z\right) + \lim_{m\to\infty} \sum_{\epsilon<|z-t/\psi|} g_1(z) P\left(Z_t = z\right) \\ &\geq \min\left(g_1\left(\frac{t}{\psi} - \epsilon\right), g_1\left(\frac{t}{\psi} + \epsilon\right)\right) \lim_{m\to\infty} P\left(\left|z - \frac{t}{\psi}\right| \leq \epsilon\right) \\ &= \min\left(g_1\left(\frac{t}{\psi} - \epsilon\right), g_1\left(\frac{t}{\psi} + \epsilon\right)\right). \end{aligned}$$

By letting ϵ approach 0, we have

$$g_1\left(\frac{t}{\psi}\right) \leq \lim_{m\to\infty} \mathbb{E}(g_1(Z_t)) \leq g_1\left(\frac{t}{\psi}\right);$$

henceforth,

$$\lim_{m\to\infty} \mathbb{E}(g_1(Z_t)) = g_1\left(\frac{t}{\psi}\right).$$

Therefore, $\lim_{m\to\infty} h(t;m) = g(t)$.

■

Example 4.3. In Figure 4.4, the hazard rate of the PTAM is displayed using $h_1 = 0.0017, h_m = 1.2750, s = -0.0735, \psi = 55$, under different m's ($m = 100, 1000, 10000$). The limit of the resulting hazard function $g(t)$, as described in Theorem 4.6, is also illustrated. The resulting hazard rate gets closer to $g(t)$ as m increases, and it is very close to $g(t)$ when $m = 10000$.

△

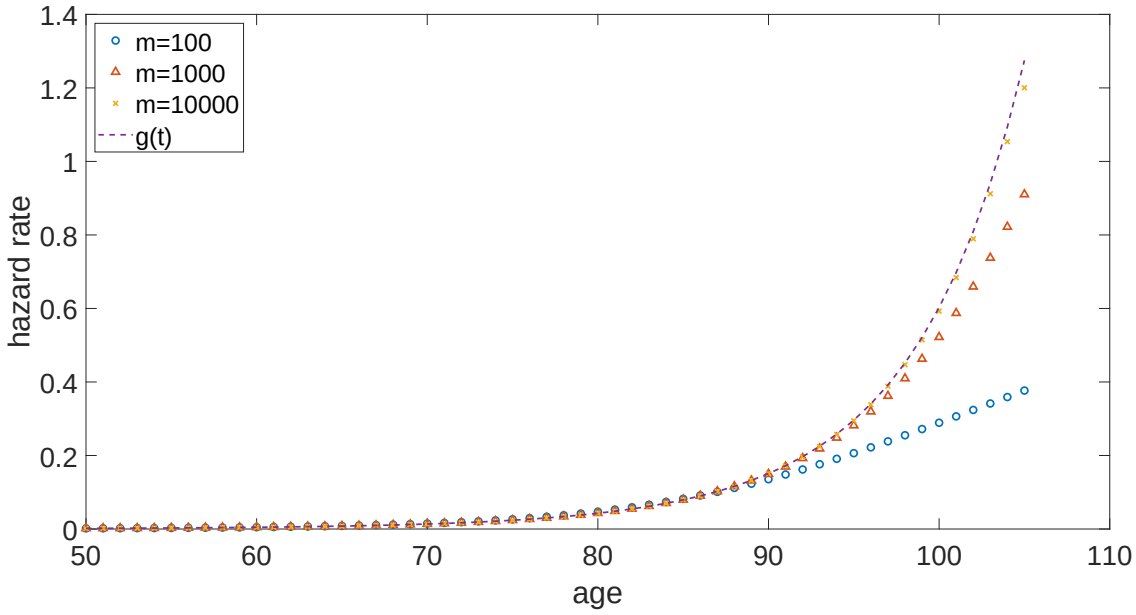

Figure 4.4: Hazard rate for the proposed PTAM with different m's and the limit $g(t)$ of the resulting hazard rate.

4.4 Conclusion

In this Chapter, we proved that a transformation of the state variable conditional on the individual being alive converges to the standard normal distribution, and the physiological index converges to the chronological age in probability as m goes to infinity. Meanwhile, the resulting hazard rate converges to a special function $g(t)$ as m goes to infinity. These results describe the properties of the PTAM with infinite states. The properties can be useful for the analysis when m is relatively large (e.g. $m = 1000, m = 10000$). Given the current state, the updated

distribution follows the PTAM with updated parameter values. The PTAM's parameters can be divided into 3 groups: the s parameter that controls the convexity of the dying rate and the convexity of the hazard rate; the $(h_1, h_m, 1/\psi)$ parameters that govern the scale; and the parameter m that controls the state variability at any age.

Chapter 5

Model calibration

We introduce in this Chapter an efficient algorithm to calculate the likelihood of the PTAM given a set of parameter values. The proposed algorithm uses the uniformization method to stabilize the numerical calculation. It uses a vectorised formula to only calculate the necessary elements for the probability distribution. An error upper bound is provided for the proposed algorithm. It can be easily adjusted to calculate the likelihood of the Coxian models.

Furthermore, we compare the speed and the accuracy of the proposed algorithm with the traditional method using the matrix exponential. The proposed algorithm is faster and more accurate than the traditional method in calculating the likelihood.

Based on our experiments, we recommend using 20 sets of randomly-generated initial values for the optimisation to get an estimate, at which the evaluated likelihood is close to the maximum likelihood.

5.1 Background on the estimation of Coxian model

Model calibration is a process of finding a set of parameter values that optimise a selected objective function; for example, maximising the likelihood, minimising the Akaike Information Criterion (AIC) or the the Bayesian Information Criterion (BIC), and minimising the mean-squared error. Since the analytical solution for the optimisation problem under the Coxian model is not available in general, the calibration procedure becomes a numerical search optimisation process, which has two parts: (i) construction of a function whose inputs are the model parameter values and the output is the selected objective function; and (ii) optimising numerically the objective function.

The Coxian distribution is a particular case of a phase-type distribution. It is then natural to use the parameter estimation procedure for phase-type models in recovering the parameters of a Coxian model. There are different ways to calibrate a Coxian model based on the formulation of the objective function and the optimisation strategy. For example, Bobbio and Cumani (1992) suggested maximising the log-likelihood by solving an iterative linearisation method of estimation. Asmussen et al. (1996) put forward the maximisation of the likelihood by a fitting procedure based on the Expectation-Maximisation (EM) algorithm. Faddy (1994, 1998) utilised the optimisation algorithm proposed by Nelder and Mead (1965) to maximise the log-likelihood function. By way of the penalised likelihood, Faddy (2002) facilitated the

convergence of the algorithm. Lin and Liu (2007) as well as Govorun et al. (2018) numerically searched for the estimates in the least-squares sense. The use of the maximum-likelihood approach was carried out by Rizk et al. (2019) approach and the optimal number of states is obtained by minimising the AIC and BIC.

Each above-mentioned calibration procedure searched for the estimate in a recursive fashion, i.e., the estimation starts with an initial value and proceeds with recursive updates until certain conditions are satisfied. In each recursion, the evaluation of the probability distribution at the observations is required. The probability distribution includes the pdf and survival function, which can be calculated by the matrix exponential or by (2.3.10). For a Coxian model with a large number of states, it is numerically unstable to calculate the probability distribution using (2.3.10). The traditional method to calculate the probability distribution is to calculate the matrix exponential. The matrix exponential calculation employs the built-in matrix exponential function (e.g. **expm** in MATLAB, **expm** in R). From here onwards, we shall refer to the method involving the matrix exponential as the traditional method. However, the calculation is relatively slow when the total number of states is large. Therefore, both methods (i.e., (2.3.10) and traditional method) have the disadvantage in the calculation of the probability distribution of a Coxian model when m is large. How then do we improve available techniques in the probability-distribution computation?

In terms of the optimisation process, the numerical algorithm requires an initial value. We found that the numerical optimisation could produce different results for different initial values. It is noted that the Coxian model is sensitive to the initial value; see Marshall and Zenga (2009a). The sensitivity of the optimisation results to the initial values will be termed in this thesis as the sensitivity issue. We suspect that the numerical optimisation outputs for some initial values may not be optimum, and these outputs cannot be treated as providing the estimated values. When using the maximum likelihood approach, the total number of states is fixed. We found that the log-likelihood increases only a little when changing the total number of states after a large value, for example $m > 200$ in section 3.4.3. There are two questions that ensue: (i) How could the sensitivity problem be rectified, even partially? and (ii) What could be done to test if if the small log-likelihood increases are due to numerical error or actuarial difference?

As demonstrated in Chapter 3, a large number of states is required to model the human ageing process. An efficient method is then necessary to calculate the probability distribution of a Coxian model, especially when there is a large number of states, and to propose a calibration procedure that reduces the sensitivity issue. Our objective in this Chapter is to develop an algorithm that is fast and accurate in the evaluation of the probability distribution of a Coxian model. The algorithm shall speed up the estimation by obtaining the probability distribution at the observations simultaneously. The algorithm's speed and accuracy will be quantified from the theoretical and numerical comparison perspectives with the traditional method.

5.2 Traditional method for the probability distribution calculation

There are different methods to compute the probability distribution; see Duan and Liu (2015). One of the methods is to compute a matrix exponential $e^{\mathbf{\Lambda}t}$, where $\mathbf{\Lambda}$ is the degenerated transition matrix. The matrix exponential computation relies on the software's built-in function. Let $\boldsymbol{\theta}$ be the parameter vector for the Coxian model. Then, the probability distribution at time t is given by

$$S(t;\boldsymbol{\theta}) = \boldsymbol{\alpha} e^{\mathbf{\Lambda}t}\, \boldsymbol{e}, \tag{5.2.1}$$

and

$$f(t;\boldsymbol{\theta}) = \boldsymbol{\alpha} e^{\mathbf{\Lambda}t}\, \boldsymbol{h}. \tag{5.2.2}$$

Both $\mathbf{\Lambda}$ and $\boldsymbol{h}$ are functions of $\boldsymbol{\theta}$. For a given $\boldsymbol{\theta}$, it is common to calculate $e^{\mathbf{\Lambda}t}$ first, and then find $f(t;\boldsymbol{\theta})$ or $S(t;\boldsymbol{\theta})$ by matrix multiplication. The calculation of $e^{\mathbf{\Lambda}t}$ is the key to determine the probability distribution.

5.2.1 Matrix exponential: Scaling and squaring method

We consider the problem on how to calculate the matrix exponential $e^{\mathbf{A}}$ efficiently for a given matrix $\mathbf{A}$. In particular,

$$e^{\mathbf{A}} = \sum_{n=0}^{\infty} \frac{1}{n!} \mathbf{A}^n.$$

Ward (1977) summarised several algorithms to calculate a matrix exponential based on various techniques such as eigenvalue and eigenvector representations (Gantmacher, 1959; Putzer, 1966; Kirchner, 1967), numerical integration (Choudhury et al., 1968; Healey, 1973), truncated power series approximations (Liou, 1966; Bickart, 1968; Healey, 1973; Källström, 1973), and rational approximations (Cody et al., 1969; Saff, 1971). An influential paper featured 19 practical ways concerning matrix exponential calculation (Moler and Van Loan, 2003). Moler and Van Loan claimed 19 dubious methods since they do not know enough detailed performance or careful implementations of various methods. It was found that, for a given matrix, some methods may be better than the others, but the performance of each method highly depends on the matrix structure; in essence, coming to the conclusion that no method outperforms the others under all situations.

In general, some computational approaches may be better than others when the matrix structure is unknown. The only generally competitive series method is the scaling and squaring method, and Ward's program (Ward, 1977) implementing this method is certainly amongst the best available approaches (Moler and Van Loan, 2003). The scaling and squaring method exploits the following facts:

- For any matrix $\mathbf{A}$, $e^{\mathbf{A}} = \left(e^{\mathbf{A}/u}\right)^u$ for any value of u.

- A Padé approximant can approximate $e^{\mathbf{A}}$ well when the norm of $\mathbf{A}$ is small.

A major drawback of the Padé approximation is its poor approximation when the norm of the matrix $\mathbf{A}$ is big. Different strategies to select the optimal value of u are proposed so that it can improve the accuracy of calculating $e^{\mathbf{A}}$ (Ward, 1977; Sidje, 1998; Higham, 2005; Al-Mohy and Higham, 2009).

For any matrix $\mathbf{A}$, there is a value u such that $\mathbf{A}/u$ is near the origin. Thus, $e^{\mathbf{A}/u}$ can be approximated by a Padé approximant by the second fact, and $e^{\mathbf{A}}$ can be calculated by the first fact. This is the main idea of the scaling and squaring method.

Currently, the default algorithm of **expm** in MATLAB is the algorithm proposed by Al-Mohy and Higham (2009). We can treat their algorithm as the most efficient and one that applies the scaling and squaring method to calculate a matrix exponential. It will be our benchmark to compare with our proposed algorithm.

We now investigate the objective function calculated by the traditional method.

5.2.2 Objective function calculation

The calculation of the objective function entails the evaluation of the probability distribution at each observation. We will use log-likelihood as an example to demonstrate how to calculate the objective function by the traditional method.

Suppose we observe the lifetime of n individuals. Specifically, for each observed individual $j = 1, \ldots, n$ we observe (ℓ_j, t_j, δ_j), where ℓ_j is the age at which individual j entered observation, t_j is the age at which individual j left observation, and $\delta_j = 1$ if individual j died at time t_j and $\delta_j = 0$ otherwise.

The likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} \frac{f(t_j;\boldsymbol{\theta})^{\delta_j} S(t_j;\boldsymbol{\theta})^{1-\delta_j}}{S(\ell_j;\boldsymbol{\theta})},$$

where $S(t;\boldsymbol{\theta})$ is the survival function, and $f(t;\boldsymbol{\theta})$ is the pdf. So, the corresponding log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \delta_j \log\left(\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t_j} \boldsymbol{h}\right) + (1-\delta_j) \log\left(\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t_j} \boldsymbol{e}\right) - \log\left(\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} \ell_j} \boldsymbol{e}\right), \tag{5.2.3}$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} -(\lambda_1 + h_1) & \lambda_1 & & & \\ & -(\lambda_2 + h_2) & \lambda_2 & & \\ & & \ddots & & \\ & & & -(\lambda_{m-1} + h_{m-1}) & \lambda_{m-1} \\ & & & & -h_m \end{bmatrix}$$

$\boldsymbol{\Lambda}$ is an $m \times m$ matrix. Since δ_j can be either 0 or 1, the log-likelihood has $2n$ matrix-exponential calculations for a given $\boldsymbol{\theta}$. Furthermore, if we can observe all individuals at age

0, then $S(0;\boldsymbol{\theta}) = 1$ in the denominator of $L(\boldsymbol{\theta})$, and the log-likelihood involves n matrix-exponential calculations. In either case, the log-likelihood calculation is relatively slow when the sample size and the total number of states is relatively large. Some numerical examples will be provided to show how slow the traditional method is. But, why is this so?

According to the literature, the scaling and squaring method requires the calculation of every element in $e^{A/u}$ so that e^A can be calculated by $e^{A/u}$ to the power of u. Therefore, the traditional method has to calculate every element in $e^{\boldsymbol{\Lambda}t}$ before calculating the probability distribution. For each observation (ℓ_j, t_j, δ_j), the scaling and squaring method calculates $m \times m$ elements in the scaling step. In the squaring step, the algorithm performs k (the number of scaling and $u = 2^k$) matrix multiplications to obtain $e^{\boldsymbol{\Lambda}t_j}$ and $e^{\boldsymbol{\Lambda}\ell_j}$. Lastly, these two steps need to repeat n times to get the log-likelihood. Therefore, using the traditional method to compute the log-likelihood performs a large number of matrix multiplications, which makes the calculation very time-consuming especially when the matrix dimension is large.

If we examine the probability distributions (5.2.1) and (5.2.2), it is sufficient to calculate the first row of each $e^{\boldsymbol{\Lambda}t_j}$ and $e^{\boldsymbol{\Lambda}\ell_j}$ for the log-likelihood (5.2.3), due to the fact that $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$. Hence, it is redundant to calculate the elements from the second row to the last row, and calculating these redundant elements slows down the log-likelihood evaluation significantly. For example, suppose we only have one observation who entered at age 0 and died at time t. Then the log-likelihood is $l(\boldsymbol{\theta}) = \log f(t)$. To calculate the log-likelihood, it is sufficient to calculate the first row m elements of $e^{\boldsymbol{\Lambda}t}$ and then multiply the row vector by the column vector $\boldsymbol{h}$, whilst the traditional method calculates m^2 elements of $e^{\boldsymbol{\Lambda}t}$ and then performs the matrix multiplication. Only $1/m$ of the elements in $e^{\boldsymbol{\Lambda}t}$ are useful for (5.2.3). The proportion of useful elements in $e^{\boldsymbol{\Lambda}t}$ decreases dramatically as m increases; in particular, only 1% of the results are useful when $m = 100$. Therefore, under the traditional method, considerable time is wasted in the evaluation of unnecessary elements.

On the other hand, the numerical accuracy of the traditional method is a concern to us because using **expm** involves a Padé approximation, which incurs an approximation bias. Although Higham (2005) and Al-Mohy and Higham (2009) showed that the numerical error can be small for the scaling and squaring method, the numerical error of each $e^{\boldsymbol{\Lambda}t_j}$ or $e^{\boldsymbol{\Lambda}\ell_j}$ may accumulate to a significant error in (5.2.3), which could cause numerical bias on the log-likelihood calculation. The numerical errors from the matrix-exponential computation are unlikely to cancel out each other in the log-likelihood calculation. Therefore, we are interested in the accuracy of the log-likelihood calculation.

In summary, although **expm** is "optimal" to calculate a general matrix exponential, it seems not the best way to compute the probability distribution in terms of algorithm speed and algorithm accuracy.

5.3 Uniformisation method and the proposed algorithm

To compute the probability distribution, it is straightforward to use (2.3.10). This method may work well for a Coxian model with a small number of states. However, the numerical result will be unstable for a Coxian model with a large number of states. For example, the probability

in state 2 at time t by (2.3.10) is given

$$\begin{aligned} P_2(t) &= -\frac{\lambda_1}{\lambda_1+h_1-\lambda_2-h_2}e^{-(\lambda_1+h_1)t} - \frac{\lambda_1}{\lambda_2+h_2-\lambda_1-h_1}e^{-(\lambda_2+h_2)t} \\ &= -\frac{\lambda_1}{\lambda_1+h_1-\lambda_2-h_2}e^{-(\lambda_1+h_1)t} + \frac{\lambda_1}{\lambda_1+h_1-\lambda_2-h_2}e^{-(\lambda_2+h_2)t}, \end{aligned}$$

where both coefficients have opposite signs. The formula for $P_k(t)$ requires multiple subtraction operations and division operations. Numerical minus operation may cause loss of significance. This is problematic when two numbers in the minus operation are relatively large but the result is relatively small, and numerical division is problematic when the denominator is close to 0. Round-off error affects the result significantly in both cases. When the total number of state m is large, it is typical to have the fitted $\lambda_i + h_i$ and $\lambda_j + h_j$ very close that the denominator $\prod_{s=1,s\neq j}^{k}(\lambda_j + h_j - \lambda_s - h_s)$ is small, and the absolute value of each $\frac{(-1)^{k-1}\lambda_1\ldots\lambda_{k-1}}{\prod_{s=1,s\neq j}^{k}(\lambda_j+h_j-\lambda_s-h_s)}e^{-(\lambda_j+h_j)t}$ is relatively large. Under such a scenario, the numerical sum of $P_k(t)$ becomes unstable.

Example 5.1. A numerical example is to calculate $P_2(t)$ at $t = 10^{-6}$ for a proposed PTAM with parameter value $h_1 = 10^{-7}$, $h_m = 2\times10^{-7}$, $s = 1$, $\lambda = 0.016$ and $m = 1000$. The numerical value using the formula (2.3.10) is 2.9802×10^{-8}, while the numerical value using either the proposed algorithm or the traditional method is 1.6×10^{-8}. The numerical value of $-\frac{\lambda}{h_1-h_2}e^{-(\lambda+h_1)t}$ has a magnitude of 10^8, but the numerical value for $P_2(t)$ has a magnitude of 10^{-8}. The round-off error has a significant effect on the result after subtraction. Furthermore, $p_k(t)$ is significantly greater than 1 for $k > 3$. Therefore, (2.3.10) is unstable in the numerical sense when the total number of states is relatively large and the difference between h_i and h_j is small.

△

Motivated by Example 5.1, we develop a numerically robust method or algorithm for the calculation of the probability distribution. It is well-known that the uniformisation or Jensen's method (Jensen, 1953; Stewart, 1994) can stabilise the numerical calculation of the probability distribution of a Coxian model.

5.3.1 Introduction to the uniformisation method

The uniformisation method is a method of computing the transient solutions of continuous-time Markov chains by the process of a discrete-time Markov chain. The basic idea of the uniformisation is assuming all transitions occur at the highest rate out of each state (hypothesised rate) in the original Markov chain, but only a fraction of transitions are real transitions (transitions that are out of the current state) and the remaining ones are fictitious (transitions that remain in the current state). The fraction of real transitions that occur is equal to the ratio of the associated real rate over the hypothesised rate. A detailed introduction of uniformisation can be found in Section 6.7 of Ross (2014).

5.3.2 Probability distribution calculation

Suppose we have a Coxian model whose Markov chain is depicted in Figure 5.1. We shall use

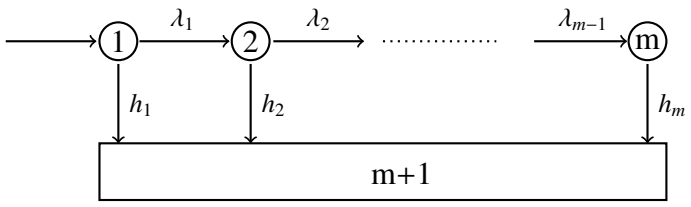

Figure 5.1: State transition diagram for a Coxian model.

the uniformisation method to calculate the probability distribution. Recall that the pdf is

$$f(t) = \boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t} \boldsymbol{h},$$

and the survival function is

$$S(t) = \boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t} \boldsymbol{e}.$$

By the uniformisation method, all transition rates are uniformised to the highest rate, and only a fraction of hypothesised transitions are real transitions. The associated probability transition diagram is displayed in Figure 5.2, where $P_{i,i+1} = \frac{\lambda_i}{\nu}$, $P_{i,m+1} = \frac{h_i}{\nu}$, $P_{i,i} = 1 - \frac{\lambda_i + h_i}{\nu}$ for $i = 1, \ldots, m$, and $\nu = \max(\lambda_i + h_i)$ is the hypothesized rate by letting $\lambda_m = 0$.

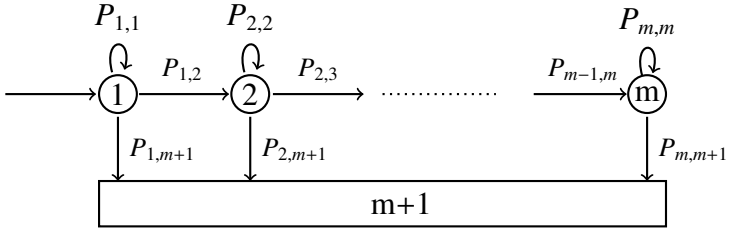

Figure 5.2: Transition probability diagram for the uniformisation method. $P_{i,j}$ is the probability of moving from state i to state j given a hypothesised transition occuring in state i.

Let $\boldsymbol{P}$ be the $m \times m$ probability transition matrix with the (i, j) element equal to the probability that a transition from state i to state j occurs given a hypothesised transition occurs. Then $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = \begin{bmatrix} 1 - \frac{(\lambda_1 + h_1)}{\nu} & \frac{\lambda_1}{\nu} & & & \\ & 1 - \frac{(\lambda_2 + h_2)}{\nu} & \frac{\lambda_2}{\nu} & & \\ & & \ddots & & \\ & & & 1 - \frac{(\lambda_{m-1} + h_{m-1})}{\nu} & \frac{\lambda_{m-1}}{\nu} \\ & & & & 1 - \frac{h_m}{\nu} \end{bmatrix}.$$

The matrix $e^{\boldsymbol{\Lambda} t}$ can be expressed as

$$e^{\boldsymbol{\Lambda} t} = \sum_{j=0}^{\infty} \boldsymbol{P}^j e^{-\nu t} \frac{(\nu t)^j}{j!}. \tag{5.3.4}$$

Since all elements in $\boldsymbol{P}$ are non-negative, every term in (5.3.4) is positive, and the numerical computation is stabilised.

Define $P_{1,i}^{j}$ as the element in the first row ith column in matrix $\boldsymbol{P}^{j}$. According to the definition of matrix power, we have $\boldsymbol{P}^{j+1} = \boldsymbol{P}^{j}\boldsymbol{P}$. Since $\boldsymbol{P}$ is sparse and upper-diagonal, we can derive the following recursive formula to calculate $P_{1,i}^{j}$ more efficiently:

$$\begin{cases} P_{1,1}^{j+1} = \left(1 - \frac{\lambda_1 + h_1}{\nu}\right) P_{1,1}^{j} \\ P_{1,i+1}^{j+1} = \frac{\lambda_i}{\nu} P_{1,i}^{j} + \left(1 - \frac{\lambda_{i+1} + h_{i+1}}{\nu}\right) P_{1,i+1}^{j}, \ i = 1, \ldots, m-2 \\ P_{1,m}^{j+1} = \frac{\lambda_{m-1}}{\nu} P_{1,m-1}^{j} + \left(1 - \frac{h_j}{\nu}\right) P_{1,m}^{j} \end{cases} \tag{5.3.5}$$

By defining $\lambda_m = \lambda_0 = 0$ and $P_{1,0} = 0$, then 3 recursive formulae are summarised as

$$P_{1,i}^{j} = \frac{\lambda_{i-1}}{\nu} P_{1,i-1}^{j-1} + \left(1 - \frac{\lambda_i + h_i}{\nu}\right) P_{1,i}^{j-1}, \ i = 1, \ldots, m, \ j = 1, \ldots \tag{5.3.6}$$

The $1 \times m$ initial probability vector is $\left(P_{1,1}^{0}, \ldots, P_{1,m}^{0}\right) = (1, 0, \ldots, 0)$, because $\boldsymbol{P}^{0}$ is the identity matrix.

An intuitive explanation for (5.3.6) is that an individual moves from state 1 to state i in j transitions, which decomposes to two probability events. The individual moves from state 1 to state $i-1$ in $j-1$ transitions, and then moves from state $i-1$ to state i in the jth transition; or the individual moves from state 1 to state i in $j-1$ transitions, and then the jth transition is a fictitious transition.

The recursive formula (5.3.6) is easy to code. In particular, the algorithm can concurrently update the $1 \times m$ vector $\left(P_{1,1}^{j}, \ldots, P_{1,m}^{j}\right)$, denoting $P_{1,\bullet}^{j}$, for each j. Compared with calculating the $m \times m$ $\boldsymbol{P}^{j}$ by matrix multiplication, it is faster to calculate the vector $P_{1,\bullet}^{j}$ by (5.3.6) because vector calculation is more efficient than matrix calculation in computers [1]. For the proposed PTAM, $\lambda_i = \lambda$ for $i = 1, \ldots, m-1$, and h_i is monotone by (3.3.3). It is easy to check that $\nu = \max(\lambda + h_{m-1}, h_m)$ if $h_1 < h_m$, while $\nu = \max(\lambda + h_1, h_m)$ if $h_1 > h_m$. The recursive formula can be simplified further,

$$P_{1,i}^{j} = \frac{\lambda}{\nu} P_{1,i-1}^{j-1} + \left(1 - \frac{\lambda + h_i}{\nu}\right) P_{1,i}^{j-1}, \ i = 1, \ldots, m, \ j = 1, \ldots$$

The pdf at time t is

$$\begin{aligned} f(t) &= \boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t} \boldsymbol{h} \\ &= \sum_{j=0}^{\infty} \boldsymbol{\alpha} \boldsymbol{P}^{j} \boldsymbol{h} e^{-\nu t} \frac{(\nu t)^{j}}{j!} \\ &= \sum_{j=0}^{\infty} \sum_{i=1}^{m} P_{1,i}^{j} h_i \, e^{-\nu t} \frac{(\nu t)^{j}}{j!}, \end{aligned}$$

[1] The speed advantage can be attributed to the computational operation efficiency involving vectors rather than matrices in software such as R or Matlab.

and the survival function at time t is

$$S(t) = \boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t} \boldsymbol{e} = \sum_{j=0}^{\infty} \boldsymbol{\alpha} \boldsymbol{P}^j \boldsymbol{e} e^{-\nu t} \frac{(\nu t)^j}{j!},$$

or

$$S(t) = \sum_{j=0}^{\infty} \sum_{i=1}^{m} P_{1,i}^j \, e^{-\nu t} \frac{(\nu t)^j}{j!}, \tag{5.3.7}$$

where $\boldsymbol{h} = (h_1, \ldots, h_m)^\top$ is a $m \times 1$ column vector, $\boldsymbol{e}$ is a $m \times 1$ column vector with ones and $P_{1,i}^j$ is obtained by (5.3.5). It is worth noting that $\boldsymbol{\alpha} \boldsymbol{P}^j$ is a $1 \times m$ row vector equal to the first row of $\boldsymbol{P}^j$.

Additionally, the first-row kth ($k = 1, \ldots, m$) column element of $e^{\boldsymbol{\Lambda} t}$ is the probability that the individual in state k at time t, denoting $P_k(t)$, and

$$P_k(t) = \sum_{j=0}^{\infty} P_{1,k}^j e^{-\nu t} \frac{(\nu t)^j}{j!},$$

where $P_{1,k}^j$ is the probability in state k given j hypothesised transitions occur, and $e^{-\nu t} \frac{(\nu t)^j}{j!}$ is the probability j transitions occur in the uniformised process. An intuitive interpretation for the formula is that $P_k(t)$ is the total probability in state k at time t after j transitions. It is worth noting that $e^{-\nu t} \frac{(\nu t)^j}{j!}$ is the pdf of a Poisson distribution with rate νt, and such a probability can be accurately calculated by the built-in function (e.g. **poisspdf** in MATLAB, **ppois** in R).

So far, there is no approximation involved. However, since $f(t)$ or $S(t)$ is a sum with infinite terms, it needs truncation in practice. Specifically, let J be the truncation point, and $S_N(t)$ and $S_T(t)$ be the respective numerical value and the theoretical value of the survival function, where

$$S_T(t) = \sum_{j=0}^{\infty} \boldsymbol{\alpha} \boldsymbol{P}^j \boldsymbol{e} e^{-\nu t} \frac{(\nu t)^j}{j!}$$

$$S_N(t) = \sum_{j=0}^{J} \boldsymbol{\alpha} \boldsymbol{P}^j \boldsymbol{e} e^{-\nu t} \frac{(\nu t)^j}{j!}.$$

Gross and Miller (1984) proved the following result.

Remark 5.1. *For any error tolerance $\epsilon > 0$, if the truncation point J satisfies the following condition*

$$1 - e^{-\nu t} \sum_{j=0}^{J} \frac{(\nu t)^j}{j!} \leqslant \epsilon, \tag{5.3.8}$$

then the difference between the numerical result and the true value of the survival function is less than ϵ, or $S_T(t) - S_N(t) \leq \epsilon$.

Since each term in the difference of the two sums is positive, i.e.,

$$S_T(t) - S_N(t) = \sum_{j=J+1}^{\infty} \left(\boldsymbol{\alpha}\boldsymbol{P}^j\boldsymbol{e}\right) e^{-\nu t}\frac{(\nu t)^j}{j!} > 0,$$

we have $S_N(t) < S_T(t)$. Similarly, let $f_N(t) = \sum_{j=0}^{J} \boldsymbol{\alpha}\boldsymbol{P}^j\boldsymbol{h}e^{-\nu t}\frac{(\nu t)^j}{j!}$ and $f_T(t) = \sum_{j=0}^{\infty} \boldsymbol{\alpha}\boldsymbol{P}^j\boldsymbol{h}e^{-\nu t}\frac{(\nu t)^j}{j!}$ be the numerical value and the theoretical value of the pdf, respectively. We have $f_N(t) < f_T(t)$, which is trivial by the fact that each term of sum is positive.

Theorem 5.1. *When (5.3.8) holds, the difference between the numerical value and the theoretical value of the pdf is less than* $\max_{i=1,\dots,m}(h_i)\epsilon$*, i.e.,* $f_T(t) - f_N(t) \leq \max_{i=1,\dots,m}(h_i)\epsilon$.

Proof.

$$\begin{aligned} f_T(t) - f_N(t) &= \boldsymbol{\alpha}\left(\sum_{j=0}^{\infty} \boldsymbol{P}^j e^{-\nu t}\frac{(\nu t)^j}{j!}\right)\boldsymbol{h} - \boldsymbol{\alpha}\left(\sum_{j=0}^{J} \boldsymbol{P}^j e^{-\nu t}\frac{(\nu t)^j}{j!}\right)\boldsymbol{h} \\ &= \sum_{j=J+1}^{\infty} \left(\boldsymbol{\alpha}\boldsymbol{P}^j\boldsymbol{h}\right) e^{-\nu t}\frac{(\nu t)^j}{j!} \\ &\leq \max_{i=1,\dots,m}(h_i) \sum_{j=J+1}^{\infty} e^{-\nu t}\frac{(\nu t)^j}{j!} \leq \max_{i=1,\dots,m}(h_i)\epsilon, \end{aligned}$$

where the first inequality holds because $\boldsymbol{\alpha}\boldsymbol{P}^j$ is a probability vector, and $\boldsymbol{\alpha}\boldsymbol{P}^j\boldsymbol{h}$ is a weighted average of absorption rates. In particular, $\max_{i=1,\dots,m}(h_i) = \max(h_1, h_m)$ for the proposed PTAM. ■

The numerical error for the uniformisation method is caused by the round-off error and the truncation error, where the truncation error can be controlled by the error tolerance ϵ. In summary, the formulae for $f(t)$ and $S(t)$ by the uniformisation method only calculate the first row of $e^{\boldsymbol{\Lambda}t}$. This method calculates a significantly less number of elements than the traditional method. Meanwhile, the uniformisation method can achieve any required accuracy within a specified error tolerance. The next step is to evaluate the probability distribution at multiple points simultaneously.

5.3.3 The proposed algorithm

We propose an algorithm that can compute the probability distribution at all observations at the same time in a matrix form using the uniformisation method. According to (5.3.4), $\boldsymbol{P}^j$ and $e^{-\nu t}\frac{(\nu t)^j}{j!}$ must be calculated to get $e^{\boldsymbol{\Lambda}t}$. Each $\boldsymbol{P}^j$ is independent of t for any $j = 1, 2, \dots$ As a result, the algorithm can vectorise the calculation of each first row of $e^{\boldsymbol{\Lambda}t_j}$ and $e^{\boldsymbol{\Lambda}\ell_j}$ for (5.2.3).

No censored or truncated data

Suppose all individuals entered the observation at age 0, and all individuals died during the observation study. Then, $\boldsymbol{\alpha} e^{\boldsymbol{\Lambda}\ell_j}\boldsymbol{e} = 1$ and $\delta_j = 1$ for each individual. Equation (5.2.3) may be simplified to

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log\left(\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t_j}\boldsymbol{h}\right).$$

From (5.3.4), the associated matrix exponential for the ith individual is given by

$$e^{\boldsymbol{\Lambda} t_i} = \sum_{j=0}^{\infty} \boldsymbol{P}^j e^{-\nu t_i} \frac{(\nu t_i)^j}{j!},$$

Let $\beta_{\mathbf{i}}$ be the first row vector of $e^{\boldsymbol{\Lambda} t_i}$, then

$$\beta_{\mathbf{i}} = \sum_{j=0}^{\infty} e^{-\nu t_i} \frac{(\nu t_i)^j}{j!} \left(P^j_{1,1}, \ldots, P^j_{1,m}\right),$$

where $e^{-\nu t_i}\frac{(\nu t_i)^j}{j!}$ is a number, and $(P^j_{1,1}, \ldots, P^j_{1,m})$ is a $1 \times m$ vector calculated by (5.3.6). Then, the $n \times m$ matrix $\boldsymbol{\beta}$, whose ith row equals $\beta_{\mathbf{i}}$, is

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\mathbf{1}} \\ \vdots \\ \beta_{\mathbf{n}} \end{bmatrix} = \sum_{j=0}^{\infty} \left\{ \begin{bmatrix} e^{-\nu t_1}\frac{(\nu t_1)^j}{j!} \\ \vdots \\ e^{-\nu t_n}\frac{(\nu t_n)^j}{j!} \end{bmatrix} \left(P^j_{1,1}, \ldots, P^j_{1,m}\right) \right\}.$$

Each term in the sum is equal to an $n \times 1$ matrix multiplied by a $1 \times m$ matrix, and $\boldsymbol{\beta}$ is an $n \times m$ matrix. The (i, j) element of $\boldsymbol{\beta}$ is $P_j(t_i)$, the probability of being in state j at time t_i. The pdf evaluated at each observation can be obtained by computing

$$(f(t_1), \ldots, f(t_n))^\top = \boldsymbol{\beta}\boldsymbol{h}, \tag{5.3.9}$$

Now, the log-likelihood (5.2.3) is calculated as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(t_i).$$

Similarly, the calculation of any objective function is immediate once $(f(t_1), \ldots, f(t_n))^\top$ is obtained. Furthermore, if the pdf in $[T_1, T_2]$ is sought, the domain can be discretised into $T_1 = t_1 < \ldots < t_n = T_2$, and then (5.3.9) can be used to obtain all pdf values by considering the discretised points as new observations at the same time.

Case of censored and truncated data

Suppose we observe n individuals with 3 vectors $(t_1, t_2, \ldots, t_n)$, $(\ell_1, \ell_2, \ldots, \ell_n)$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^\top$. Recall that for each observed individual $j = 1, \ldots, n$ we observe (ℓ_j, t_j, δ_j), where ℓ_j is the age

at which an individual j entered the observation study, t_j is the age at which an individual j left the observation, and $\delta_j = 1$ if an individual j died at time t_j and $\delta_j = 0$ otherwise. We need an efficient algorithm to compute the log-likelihood (5.2.3) for a given $\boldsymbol{\theta}$.

According to (5.3.4) and (5.2.3), the associated matrix exponentials for the ith individual are

$$e^{\boldsymbol{\Lambda}\ell_i} = \sum_{j=0}^{\infty} \boldsymbol{P}^j e^{-\nu\ell_i} \frac{(\nu\ell_i)^j}{j!};$$

and

$$e^{\boldsymbol{\Lambda}t_i} = \sum_{j=0}^{\infty} \boldsymbol{P}^j e^{-\nu t_i} \frac{(\nu t_i)^j}{j!},$$

respectively.

Let $\gamma_{\mathbf{i}}$ be the first row vector of $e^{\boldsymbol{\Lambda}\ell_i}$, and β_i be the first row vector of $e^{\boldsymbol{\Lambda}t_i}$. Then,

$$\beta_{\mathbf{i}} = \sum_{j=0}^{\infty} e^{-\nu t_i} \frac{(\nu t_i)^j}{j!} \left(P_{1,1}^j, \ldots, P_{1,m}^j\right),$$

$$\gamma_{\mathbf{i}} = \sum_{j=0}^{\infty} e^{-\nu \ell_i} \frac{(\nu \ell_i)^j}{j!} \left(P_{1,1}^j, \ldots, P_{1,m}^j\right),$$

where $P_{1,i}^j$ is calculated by (5.3.6). Let $\boldsymbol{\gamma}$ be the $n \times m$ matrix whose ith row is equal to $\gamma_{\mathbf{i}}$, and $\boldsymbol{\beta}$ be the matrix whose ith row is equal to $\beta_{\mathbf{i}}$. Then,

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{\mathbf{1}} \\ \vdots \\ \gamma_{\mathbf{n}} \end{bmatrix} = \sum_{j=0}^{\infty} \left\{ \begin{bmatrix} e^{-\nu\ell_1} \frac{(\nu\ell_1)^j}{j!} \\ \vdots \\ e^{-\nu\ell_n} \frac{(\nu\ell_n)^j}{j!} \end{bmatrix} \left(P_{1,1}^j, \ldots, P_{1,m}^j\right) \right\},$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\mathbf{1}} \\ \vdots \\ \beta_{\mathbf{n}} \end{bmatrix} = \sum_{j=0}^{\infty} \left\{ \begin{bmatrix} e^{-\nu t_1} \frac{(\nu t_1)^j}{j!} \\ \vdots \\ e^{-\nu t_n} \frac{(\nu t_n)^j}{j!} \end{bmatrix} \left(P_{1,1}^j, \ldots, P_{1,m}^j\right) \right\}.$$

Similar to the case when no censored or truncated data, each term in the sum is equal to an $n \times 1$ matrix multiplied by a $1 \times m$ matrix, and both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are $n \times m$ matrices. The $\boldsymbol{\beta}$ here is the same as the one in the previous case, and $\boldsymbol{\gamma}$ has the same meaning as $\boldsymbol{\beta}$, except that the time of evaluation is different. The required probability distributions evaluated at each lifetime can be obtained simultaneously by

$$\begin{aligned} (f(t_1), \ldots, f(t_n))^\top &= \boldsymbol{\beta}\boldsymbol{h}, \\ (S(t_1), \ldots, S(t_n))^\top &= \boldsymbol{\beta}\boldsymbol{e_1}, \\ (S(\ell_1), \ldots, S(\ell_n))^\top &= \boldsymbol{\gamma}\boldsymbol{e_1}, \end{aligned}$$

where $\boldsymbol{e_1}$ is an $m \times 1$ column vector of ones, and $\boldsymbol{h}$ is an $m \times 1$ column vector of h_i.

Let $g(t_i, \ell_i, \delta_i)$ be the contribution of the ith individual to the likelihood, or

$$g(t_i, \ell_i, \delta_i) = \frac{f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}}{S(\ell_i)}. \tag{5.3.10}$$

Then, $g(t_i, \ell_i, \delta_i)$ is computed as

$$(g(t_1, \ell_1, \delta_1), \ldots, g(t_n, \ell_n, \delta_n))^\top = (\boldsymbol{\beta h} .* \boldsymbol{\delta} + \boldsymbol{\beta e_1} .* (\boldsymbol{e_2} - \boldsymbol{\delta})) ./ (\boldsymbol{\gamma e_1}), \tag{5.3.11}$$

where $\boldsymbol{e_2}$ is an $n \times 1$ column vector of ones, and $\boldsymbol{\delta}$ is an $n \times 1$ column vector whose ith element is δ_i. We use the MATLAB operator notation $.*$ and $./$ for the element-wise scalar multiplication and element-wise scalar division, respectively. It is worth noting that δ_i can be either 0 or 1, so $g(t_i, \ell_i, \delta_i) = f(t_i)/S(\ell_i)$ when $\delta_i = 1$, and $g(t_i, \ell_i, \delta_i) = S(t_i)/S(\ell_i)$ when $\delta_i = 0$. On the other hand, $\boldsymbol{\beta h} .* \boldsymbol{\delta} + \boldsymbol{\beta e_1} .* (\boldsymbol{e_2} - \boldsymbol{\delta})$ is a column vector, whose elements are equal to $\boldsymbol{\beta h}$ when the corresponding column of $\boldsymbol{\delta}$ is equal to 1, and whose elements are equal to $\boldsymbol{\beta e_1}$ when the corresponding column of $\boldsymbol{\delta}$ is equal to 0. Therefore, the left-hand side of (5.3.11) is equal to its right-hand side even though the formula is different from (5.3.10). The log-likelihood function is then given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log g(t_i, \ell_i, \delta_i).$$

Similarly, the objective function follows once the value coming from (5.3.11) is obtained.

5.3.4 Truncation in the numerical calculation

Since $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ have infinite terms in the sum, the algorithm needs to be truncated at some point J in the implementation process. Suppose we have an error tolerance ϵ and the infinite-terms sum truncates at J. Let J be the minimum integer such that

$$\max_{i=1,\ldots,n} \left(1 - e^{-\nu t_i} \sum_{j=0}^{J} \frac{(\nu t_i)^j}{j!}, 1 - e^{-\nu \ell_i} \sum_{j=0}^{J} \frac{(\nu \ell_i)^j}{j!} \right) \leqslant \epsilon. \tag{5.3.12}$$

From Remark 5.1, the difference between the theoretical value and the numerical value of the survival function is less than ϵ,

$$S_T(t_i) - S_N(t_i) < \epsilon, \text{ and } S_T(\ell_i) - S_N(\ell_i) < \epsilon.$$

Moreover, the difference between the theoretical value and the numerical value of the probability density function is less than $\max_{i=1,\ldots,m}(h_i)\epsilon$, i.e.,

$$f_T(t_i) - f_N(t_i) < \max_{i=1,\ldots,m} (h_i)\epsilon.$$

Let us investigate (5.3.12). Suppose two random variables (A and B) follow Poisson distributions with different rates. That is, suppose the rate of A is less than that of B. Then,

the random variable with the higher rate is greater than the other (i.e., B is greater than A in stochastic order). The probability that the random variable is greater than any number is higher for the Poisson distribution with the higher rate, i.e., $P(A > k) \leq P(B > k)$ for any k. Therefore, we have

$$\max_{i=1,\ldots,n}\left(1 - e^{-\nu t_i}\sum_{j=0}^{J}\frac{(\nu t_i)^j}{j!}, 1 - e^{-\nu \ell_i}\sum_{j=0}^{J}\frac{(\nu \ell_i)^j}{j!}\right) = 1 - e^{-\nu t_{\max}}\sum_{j=0}^{J}\frac{(\nu t_{\max})^j}{j!},$$

where $t_{\max} = \max_{i=1,\ldots,m}(t_i) = \max_{i=1,\ldots,m}(t_i, \ell_i)$ by the fact that $\ell_i \leq t_i$ for any i. This is because $\nu t_{\max}$ is the highest rate, and $1-e^{-\nu t_{\max}}\sum_{j=0}^{J}\frac{(\nu t_{\max})^j}{j!}$ is greater than or equal to any $1-e^{-\nu t_i}\sum_{j=0}^{J}\frac{(\nu t_i)^j}{j!}$ and $1 - e^{-\nu \ell_i}\sum_{j=0}^{J}\frac{(\nu \ell_i)^j}{j!}$ for any j, where $1 - e^{-a}\sum_{j=0}^{J}\frac{a^j}{j!}$ is the probability that a Poisson variable with rate a is greater than J. Consequently, (5.3.12) is equivalent to

$$1 - e^{-\nu t_{\max}}\sum_{j=0}^{J}\frac{(\nu t_{\max})^j}{j!} \leqslant \epsilon, \tag{5.3.13}$$

or J is the $1 - \epsilon$ quantile of the Poisson distribution with rate $\nu t_{\max}$. The determination of J is straightforward by the built-in quantile function of the Poisson distribution in MATLAB.

The default value for the error tolerance is $\epsilon = 10^{-10}$ in our numerical calculations. To achieve a higher accuracy, one can set a smaller ϵ. We show how to calculate the pdf of a GPTAM by 3 methods: the proposed algorithm, the traditional method, and formula (2.3.10).

Example 5.2. Suppose a GPTAM has a total of 50 states. The absorption rate in state i follows $h_i = 0.1i + 0.15$ and the transition rate to the next state is $\lambda = 1.8$. The numerical results are shown in Figure 5.3, indicating more stability for the proposed algorithm than using (2.3.10). The numerical results have little difference between the traditional method and the proposed method in this example.

Our proposed algorithm only calculates the necessary elements for the probability distribution at all observations in a matrix. The numerical results are stable, compared with (2.3.10). A truncation condition is provided for numerical computation so that the numerical error of the probability distribution at each point can be controlled.

△

In the succeeding discussion, we compare the algorithm's speed and accuracy with those of the traditional method on the basis of the log-likelihood calculation.

5.4 Algorithm's accuracy

We are interested in the accuracy of the proposed algorithm because the algorithm is useful only if it can achieve a given accuracy. In this section, we assess the algorithm's accuracy for our proposed PTAM. All individuals entered the observation study at age 0; the times of death of all individuals are observed at $t_1, \ldots, t_n$. The log-likelihood is $l(\boldsymbol{\theta}) = \sum_{j=1}^{n}\log f(t_i)$, where $f(t_i) > 0$.

We derive an upper bound for the numerical error of the computed log-likelihood. Then, we will compare the proposed algorithm with the traditional method in terms of the numerical error

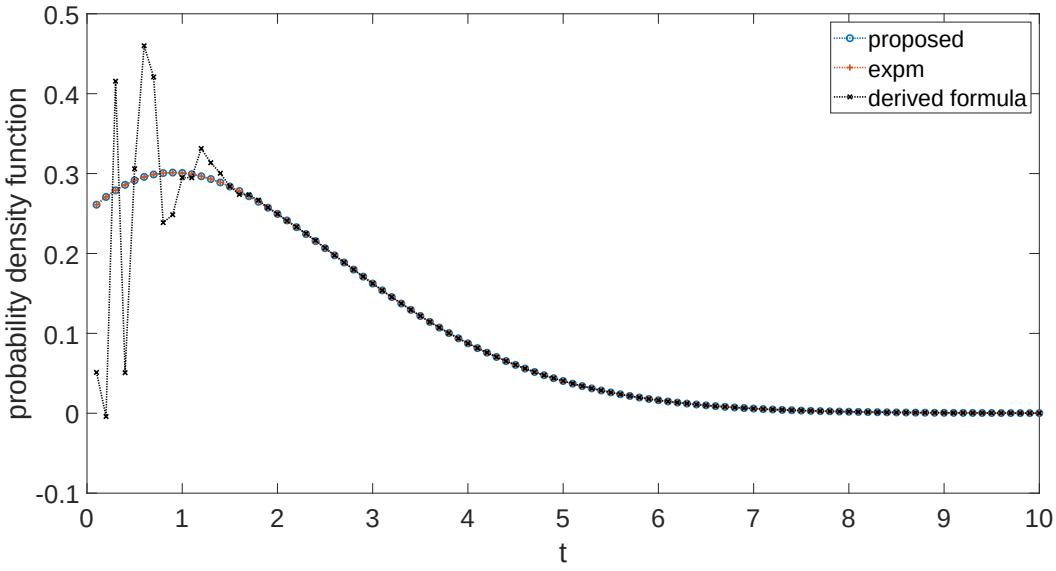

Figure 5.3: The pdf calculated by the proposed algorithm, the traditional method, and the derived formula (2.3.10).

of the probability distribution calculation. The numerical error of the log-likelihood calculation will also be evaluated under a special PTAM, whose probability distribution can be obtained accurately by an analytic solution that is stable in numerical computing. At the end of this section, we establish a condition that tests if the true log-likelihoods are different for various parameter values.

5.4.1 An upper bound for the numerical error in the log-likelihood

For a given parameter value $\boldsymbol{\theta}$, let $l_N(\boldsymbol{\theta})$ and $l_T(\boldsymbol{\theta})$ be the respective numerical value and the theoretical value of the log-likelihood. Specifically,

$$l_T(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_T(t_i),$$

$$l_N(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_N(t_i),$$

where the pdf is calculated by the uniformisation method. The $1 \times m$ row vectors of the pdf evaluated at the observations are:

$$(f_T(t_1), \ldots, f_T(t_n))^\top = \left\{ \sum_{j=0}^{\infty} \begin{bmatrix} e^{-\nu t_1} \frac{(\nu t_1)^j}{j!} \\ \vdots \\ e^{-\nu t_n} \frac{(\nu t_n)^j}{j!} \end{bmatrix} \left(P^j_{1,1}, \ldots, P^j_{1,m} \right) \right\} \boldsymbol{h},$$

$$(f_N(t_1), \ldots, f_N(t_n))^\top = \left\{ \sum_{j=0}^{J} \begin{bmatrix} e^{-\nu t_1} \frac{(\nu t_1)^j}{j!} \\ \vdots \\ e^{-\nu t_n} \frac{(\nu t_n)^j}{j!} \end{bmatrix} \left(P^j_{1,1}, \ldots, P^j_{1,m} \right) \right\} \boldsymbol{h}.$$

Recall that J is the minimum integer such that

$$1 - e^{-\nu t_{\max}} \sum_{j=0}^{J} \frac{(\nu t_{\max})^j}{j!} \leq \epsilon,$$

where $t_{\max} = \max_{i=1,\ldots,n}(t_i)$ and ϵ is the error tolerance.

Since $f_N(t_i) < f_T(t_i)$ for each t_i, we have $l_N(\boldsymbol{\theta}) < l_T(\boldsymbol{\theta})$. Therefore, the numerical error for the log-likelihood is positive, i.e., $l_T(\boldsymbol{\theta}) - l_N(\boldsymbol{\theta}) > 0$. An upper bound for such a numerical error is given in the next theorem.

Theorem 5.2. *The numerical error of the proposed algorithm to compute the log-likelihood has an upper bound of* $\sum_{i=1}^{n} \frac{\max_{i=k,\ldots,m}(h_k)}{f_N(t_i)} \epsilon$.

Proof.

$$\begin{aligned} l_T(\boldsymbol{\theta}) - l_N(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log(f_T(t_i)) - \sum_{i=1}^{n} \log(f_N(t_i)) \\ &= \sum_{i=1}^{n} \log\left(\frac{f_T(t_i)}{f_N(t_i)} \right) \\ &= \sum_{i=1}^{n} \log\left(1 + \frac{f_T(t_i) - f_N(t_i)}{f_N(t_i)} \right) \\ &\leq \sum_{i=1}^{n} \frac{1}{f_N(t_i)} (f_T(t_i) - f_N(t_i)) \leq M(\boldsymbol{\theta}) \epsilon, \end{aligned}$$

where

$$M(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\max_{i=k,\ldots,m}(h_k)}{f_N(t_i)}.$$

The first inequality holds because $\log(1 + x) \leq x$ for any $x \geq 0$, and the second inequality is justified due to $f_T(t_i) - f_N(t_i) \leq \max_{k=1,\ldots,m}(h_k)\epsilon$ as per Theorem 5.1.

■

We turn our attention to the upper bound $M(\boldsymbol{\theta})\epsilon$, where the value of $M(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. As ϵ and $f_T(t_i) - f_N(t_i)$ become smaller, leading to a bigger $f_N(t_i)$ that is closer to $f_T(t_i)$, and resulting to a smaller $M(\boldsymbol{\theta})$ that is closer to $\sum_{i=1}^{n} \frac{\max_{i=k,\ldots,m}(h_k)}{f_T(t_i)}$. Therefore, the numerical error of the log-likelihood calculation asymptotically converges to 0 as ϵ approaches 0. In particular,

$$\begin{aligned}\lim_{\epsilon\to 0} l_T(\boldsymbol{\theta}) - l_N(\boldsymbol{\theta}) &\leq \lim_{\epsilon\to 0} M(\boldsymbol{\theta})\epsilon \\ &= \sum_{i=1}^{n} \frac{\max_{i=k,\ldots,m}(h_k)}{f_T(t_i)}(0) \\ &= 0.\end{aligned}$$

Therefore, the numerical log-likelihood of the proposed algorithm can be arbitrarily close to the actual value in theory. In practice, the round-off error may limit the algorithm's accuracy, but this type of analysis is out of the scope in this thesis. From a practical point of view, we shall conduct the following experiments to compare the accuracy of the proposed algorithm with that of the traditional method by computing the probability distribution and log-likelihoods using randomly simulated data.

5.4.2 Comparison with the traditional method

Definition 5.1. *The measure of the algorithm's accuracy refers to the absolute difference between the theoretical value and the numerical result from the algorithm.*

It has to be noted that the smaller the difference, the higher the algorithm's accuracy.

To compare the accuracy between algorithms, we need the theoretical value accurately. It was demonstrated that (2.3.10) is not stable for numerical calculation, creating a significant bias against the theoretical value. It turns out that the probability distribution of a special GPTAM has an analytic form that is stable in performing numerical calculation. The special GPTAM has a restriction that h_i is a linear function of i. Then, the analytic solution to (2.3.9) can be obtained by the following theorem.

Theorem 5.3. *Suppose h_i is given by*

$$h_i = h_1 + \mu(i-1),$$

and $\lambda_i = \lambda$ for $i = 1, \ldots, m-1$. The solution for (2.3.9) is

$$P_k(t) = e^{-(\lambda+h_1)t}\frac{(c(t))^{k-1}}{(k-1)!},$$

for $k = 1, \ldots, m-1$, and

$$P_m(t) = \lambda e^{-(h_1+\mu(m-1))t}\int_0^t e^{((m-1)\mu-\lambda)u}\frac{(c(u))^{m-2}}{(m-2)!}\mathrm{d}u$$

where $c(t) = \frac{\lambda}{\mu}(1 - e^{-\mu t})$.

Proof. We obtain $P_1(t) = e^{-(\lambda+h_1)t}$ as a solution to $\frac{\mathrm{d}P_1(t)}{\mathrm{d}t} = -(\lambda + h_1)P_1(t)$. For $k = 2, \ldots, m-1$, the unique solution is

$$P_k(t) = e^{-(\lambda+h_k)t}\lambda \int_0^t e^{(\lambda+h_k)u} P_{k-1}(u)\mathrm{d}u. \tag{5.4.14}$$

Suppose $P_k(t) = \frac{e^{-(\lambda+h_1)t}}{(k-1)!}\left(\frac{\lambda(1-e^{-\mu t})}{\mu}\right)^{k-1}$. Then, according to (5.4.14),

$$\begin{aligned}
P_{k+1}(t) &= e^{-(\lambda+h_{k+1})t}\lambda \int_0^t e^{(\lambda+h_{k+1})u} \frac{e^{-(\lambda+h_1)u}}{(k-1)!}\left(\frac{\lambda(1-e^{-\mu u})}{\mu}\right)^{k-1} \mathrm{d}u \\
&= e^{-(\lambda+h_{k+1})t}\frac{\lambda^k}{\mu^{k-1}(k-1)!}\int_0^t e^{k\mu u}(1-e^{-\mu u})^{k-1}\mathrm{d}u \\
&= e^{-(\lambda+h_{k+1})t}\frac{\lambda^k}{\mu^{k-1}(k-1)!}\int_0^t e^{u\mu}(e^{u\mu}-1)^{k-1}\mathrm{d}u \\
&= e^{-(\lambda+h_{k+1})t}\frac{\lambda^k}{\mu^{k}(k-1)!}\int_0^t (e^{u\mu}-1)^{k-1}\mathrm{d}e^{u\mu} \\
&= e^{-(\lambda+h_1+k\mu)t}\frac{\lambda^k}{\mu^k(k-1)!}\frac{(e^{t\mu}-1)^k}{k} \\
&= \frac{e^{-(\lambda+h_1)t}}{k!}\left(\frac{\lambda(1-e^{-\mu t})}{\mu}\right)^k.
\end{aligned}$$

Therefore, $P_k(t) = \frac{e^{-(\lambda+h_1)t}}{(k-1)!}\left(\frac{\lambda(1-e^{-\mu t})}{\mu}\right)^{k-1}$ holds for $k = 2, \ldots, m-1$ by induction. For $k = m$,

$$\begin{aligned}
P_m(t) &= e^{-h_m t}\lambda \int_0^t e^{h_m u} P_{m-1}(u)\mathrm{d}u \\
&= e^{-h_m t}\lambda \int_0^t e^{h_m u}\frac{e^{-(\lambda+h_1)u}}{(m-2)!}\left(\frac{\lambda(1-e^{-\mu u})}{\mu}\right)^{m-2}\mathrm{d}u \\
&= e^{-(h_1+(m-1)\mu)t}\lambda \int_0^t e^{((m-1)\mu-\lambda)u}\frac{\left(\frac{\lambda(1-e^{-\mu u})}{\mu}\right)^{m-2}}{(m-2)!}\mathrm{d}u
\end{aligned}$$

■

Remark 5.2. *Given the value of $P_k(t)$, $P_{k+1}(t)$ can be calculated in a recursive fashion,*

$$P_{k+1}(t) = P_k(t)\frac{c(t)}{k}, \text{ for } k = 1, \ldots, m-1, \tag{5.4.15}$$

starting with $P_1(t) = e^{-(\lambda+h_1)t}$. In our experience, the recursive formula is accurate and efficient to calculate $P_k(t)$ for any k, except at $k = m$. The probability of being in state m at time t, $P_m(t)$, entails a numerical integration. Such a numerical integration is performed through MATLAB using a long format with the highest precision. The results using the recursive formula (5.4.15)

in conjunction with the numerical integration are treated as the theoretical values, which will serve as benchmarks in our comparison. It is worth noting that the "theoretical" values still have numerical errors attributed to both round-off and numerical-integration errors, but these results are the most accurate results we can obtain.

Example 5.3. Consider a GPTAM with the following parameter values: $h_1 = 0.025$, $\mu = 0.01$, $m = 50$, and $\lambda = 1.6$. The corresponding pdf, survival function, hazard rate, and dying rate are graphically presented in Figure 5.4. The plot of the survival function shows that almost nobody can survive to age 30 by assuming the process starts at age 0. The survival probabilities to ages 18, 25, and 30 are 1.6%, 0.2%, and 0.069%, respectively.

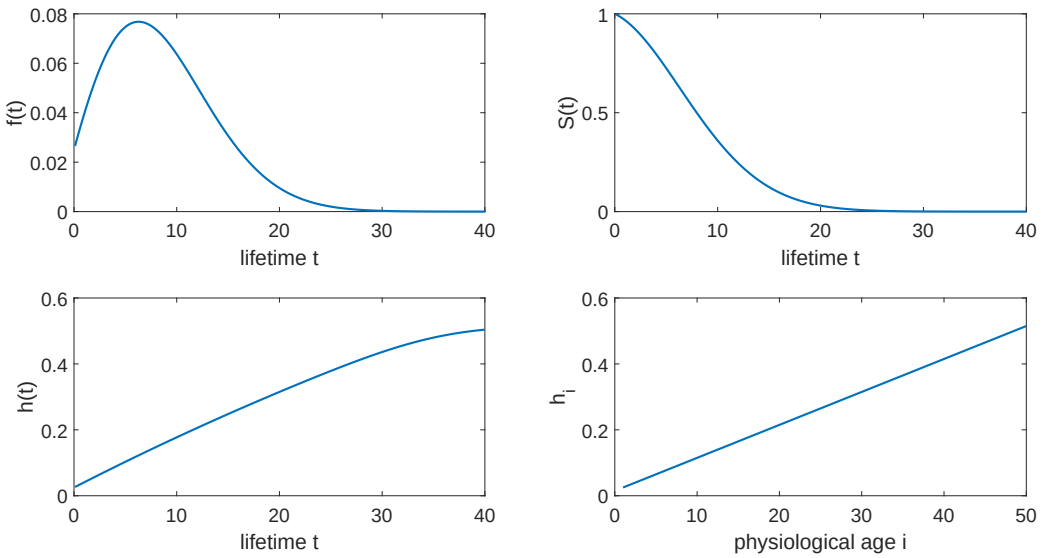

Figure 5.4: Results for the GPTAM with $h_1 = 0.025$, $\mu = 0.01$, $m = 50$, and $\lambda = 1.6$. Top-left: pdf; Top-right: Survival function; Bottom-left: hazard rate; Bottom-right: dying rate at physiological age.

We compare the accuracy of the proposed algorithm and that of the traditional method in terms of the probability distribution calculation. Recall that the numerical error of the pdf at time t is $|f_T(t) - f_N(t)|$, where $f_T(t)$ is the calculated theoretical value and $f_N(t)$ is the numerical value from the tested algorithm.

Let $f_{NP}(t)$ be the numerical value from the proposed algorithm and $f_{NT}(t)$ be the numerical value from the traditional method. The ratio $|f_T(t) - f_{NT}(t)|/|f_T(t) - f_{NP}(t)|$ gives the relative error between the two algorithms. The proposed algorithm has better accuracy when the ratio is higher than 1, and the greater the ratio, the more significant the accuracy advantage. Our experiment results (see Figure 5.5 for the plot of $|f_T(t) - f_{NT}(t)|/|f_T(t) - f_{NP}(t)| - 1$) show that the maximum of the ratio has magnitude of 10^{15}. The ratio is significantly greater than 1 after time 25, which implies that the numerical error is relatively large for the traditional method.

Remark 5.3. *The actual values at the first 4 peaks before $t = 25$ in Figure 5.5 are infinite because $|f_T(t) - f_{NP}(t)| = 0$, and we manually set them equal to the maximum of the remaining values.*

By checking the numerical error for each algorithm, based on Figure 5.6, the numerical error for both methods are close to 0 when $t < 18$. The numerical error for the traditional method is significantly greater than 0, whilst the numerical error for the proposed algorithm remains near 0 when $t > 18$. The traditional method computes less accurately than the proposed algorithm.

Figures 5.5 – 5.6 show that the numerical error is small when t is small (e.g. $t < 18$ in the example) for both methods. However, the traditional method has relatively larger numerical error than the proposed method when t is relatively big (e.g. $t > 25$ in the example). This shows both methods are accurate to calculate the distribution in most domains, but the proposed algorithm outperforms the traditional method in the right tail of the distribution. Therefore, the proposed algorithm is better than the traditional method in terms of accuracy of computing probability distribution.

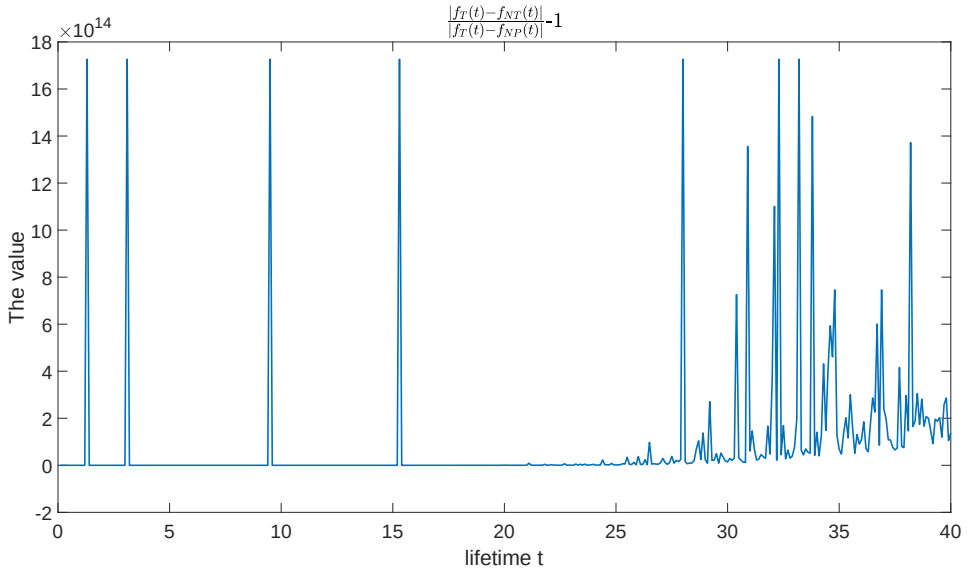

Figure 5.5: The difference between the ratio $|f_T(t) - f_{NT}(t)|/|f_T(t) - f_{NP}(t)|$ and 1 as t increases.

The next comparison is the accuracy of log-likelihood calculation by both methods. We simulate 20 sets of observations with different sample sizes. For each set of data, we randomly simulate n lifetimes, where $n = 50j$, $j = 1, \ldots, 20$, for the jth set of data. In other words, the smallest data set has 50 observations, whilst the largest data-set has 1,000 observations. For each data set, the true log-likelihood is calculated by Theorem 5.3, and the numerical log-likelihood is calculated by the proposed algorithm and the traditional method. The relative numerical error of the log-likelihood is the ratio of the difference between the true log-likelihood and numerical log-likelihood over the true log-likelihood. That is,

$$\text{Relative numerical error of the log-likelihood} = \frac{|l_T(\boldsymbol{\theta}) - l_N(\boldsymbol{\theta})|}{l_T(\boldsymbol{\theta})}.$$

The relative numerical error of the log-likelihoods for different sample sizes by different methods are compared in Figure 5.7. The relative numerical error for the proposed algorithm is close to 0, yielding the numerical error for the proposed algorithm is negligible for the log-likelihood

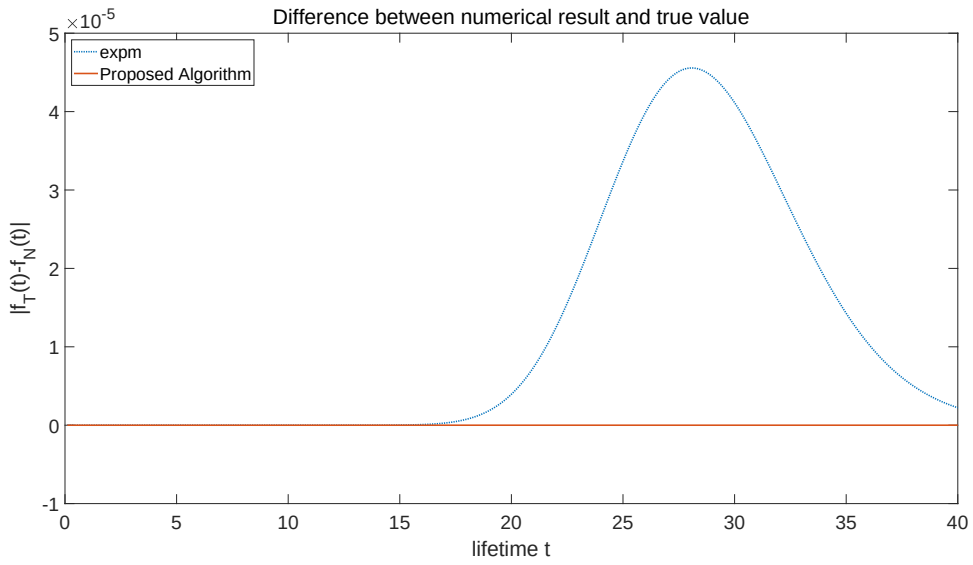

Figure 5.6: Numerical errors in the calculation of the Coxian model's pdf with $h_1 = 0.025$, $\mu = 0.01$, $m = 50$, and $\lambda = 1.6$.

calculation. On the other hand, the relative numerical error for the traditional method is small in the magnitude of 10^{-4}, but the error is larger than that for the proposed algorithm. The result is consistent with the probability distribution calculation. As a result, the proposed algorithm achieves a higher accuracy than the traditional method in log-likelihood calculation.

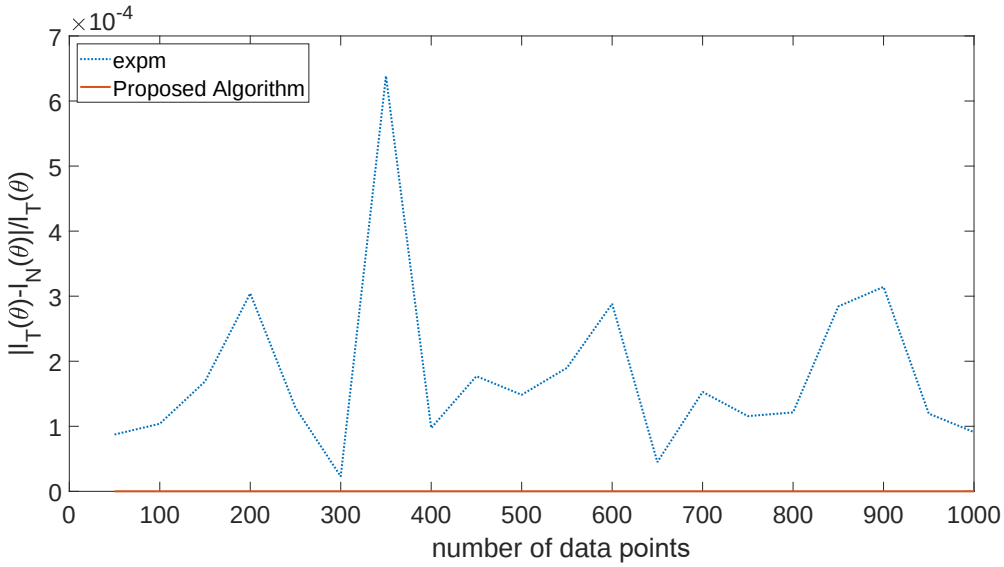

Figure 5.7: Numerical error for the log-likelihood with increasing sample size using the proposed algorithm and the traditional method.

The numerical comparisons demonstrate that the proposed algorithm achieves higher accuracy than the traditional method.

△

5.4.3 Test condition on the log-likelihood difference

As demonstrated in the Le Bras simulation, the maximum log-likelihoods are similar for different values of m. We explore if the true log-likelihoods are different by comparing the numerical results. If the true log-likelihoods are different for two sets of parameter values, then at least one of them cannot be the MLE.

Suppose we have two numerically computed log-likelihoods obtained by the proposed algorithm for two sets of parameter values $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ corresponding to the log-likelihood $l_N(\boldsymbol{\theta_1})$ and $l_N(\boldsymbol{\theta_2})$, respectively. The respective error upper bounds in Theorem 5.2 are $M(\boldsymbol{\theta_1})\epsilon$ and $M(\boldsymbol{\theta_2})\epsilon$; we also assume that $l_N(\boldsymbol{\theta_1}) < l_N(\boldsymbol{\theta_2})$. The following theorem provides a sufficient condition that the true log-likelihood with parameter equal to $\boldsymbol{\theta_2}$ is the larger one, i.e., $l_T(\boldsymbol{\theta_1}) < l_T(\boldsymbol{\theta_2})$.

Theorem 5.4. *If*

$$l_N(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_1}) > M(\boldsymbol{\theta_1})\epsilon \tag{5.4.16}$$

then the true log-likelihood is greater at $\boldsymbol{\theta_2}$, or $l_T(\boldsymbol{\theta_2}) - l_T(\boldsymbol{\theta_1}) > 0$.

Proof.

$$\begin{aligned} l_T(\boldsymbol{\theta_2}) - l_T(\boldsymbol{\theta_1}) &= (l_T(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_2})) + l_N(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_1}) - (l_T(\boldsymbol{\theta_1}) - l_N(\boldsymbol{\theta_1})) \\ &> (l_T(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_2})) + M(\boldsymbol{\theta_1})\epsilon - (l_T(\boldsymbol{\theta_1}) - l_N(\boldsymbol{\theta_1})) \\ &\geq 0, \end{aligned}$$

where the last inequality holds because $l_T(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_2}) \geq 0$ and $l_T(\boldsymbol{\theta_1}) - l_N(\boldsymbol{\theta_1}) \leq M(\theta_1)\epsilon$. Therefore, the theoretical log-likelihood is greater at $\boldsymbol{\theta_2}$.

■

Since $M(\boldsymbol{\theta_1})\epsilon$ asymptotically converges to 0 as ϵ approaches 0, there exists a small ϵ such that $l_N(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_1}) > M(\boldsymbol{\theta_1})\epsilon$, whenever the theoretical log-likelihoods are different at $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$. As a result, $l_N(\boldsymbol{\theta_2}) - l_N(\boldsymbol{\theta_1}) > M(\boldsymbol{\theta_1})\epsilon$ is a useful condition to diagnose if the actual log-likelihood increases by changing parameter values. The following example is a demonstration of this condition.

Example 5.4. Five thousand lifetimes are randomly generated from the proposed PTAM with $\lambda = 1$, $h_1 = 0.05$, $h_m = 1$, $s = 1$ and $m = 20$. Specifically, the absorption rate in state i is $h_i = 0.05i$ for $i = 1, \ldots, 20$. We estimate m using the maximum likelihood approach, and m is restricted to the values in the set $\{5, 10, 15, 20, 25, 30, 35, 40\}$. For each fixed m, the other parameters are estimated by MLE method.

The estimation results are summarised in Table 5.1. From the $l_N(\theta)$ column, the log-likelihood increases as m increases from 5 to 40, indicating that the maximum likelihood is achieved when $m = 40$. The values in $l_N(\theta_2) - l_N(\theta_1)$ column from $m = 5$ to $m = 25$ are greater than the corresponding values in the $M\epsilon$ column; thus, the theoretical log-likelihood increases

indeed by changing from $m = 5$ to $m = 30$. However, the values in $l_N(\theta_2) - l_N(\theta_1)$ column for $m = 30$ and $m = 35$ are less than the corresponding values in the column named $M\epsilon$ in Table 5.1. Therefore, we cannot conclude if the log-likelihood actually increases by changing from $m = 30$ to $m = 35$, or from $m = 35$ to $m = 40$. As a result, we can confirm that the estimate of m is greater than 25 by the MLE, but we cannot tell if the theoretical log-likelihood increases by changing from $m = 30$ to $m = 35$ and from $m = 35$ to $m = 40$. So, Theorem 5.4 can help identify some cases that the true log-likelihood can increase by changing the parameter value. Under the special case that the actual log-likelihoods are the same for two different sets of parameter values, the model is not estimable. Model estimability will be dealt with in Chapter 6.

Table 5.1: Estimation results based on 5,000 lifetimes simulated from the proposed PTAM for different m's. The $l_N(\theta)$ column is the log-likelihood. The $M\epsilon$ column is the right-hand side of (5.4.16). The ith value in the $l_N(\theta_2) - l_N(\theta_1)$ column is the left-hand side of (5.4.16) between the ith and $(i + 1)$th values.

$l_N(\theta)$	h_1	h_m	λ	s	m	$M\epsilon$	$l_N(\theta_2) - l_N(\theta_1)$
−12285.1545	0.0610	0.6750	0.5770	0.0090	5	6.584×10^{-6}	3.6200
−12281.5342	0.0550	0.7430	0.7880	0.7600	10	7.736×10^{-6}	0.2380
−12281.2960	0.0540	0.8310	0.9560	0.9190	15	8.860×10^{-6}	0.0140
−12281.2817	0.0530	1.0640	0.9880	0.9410	20	1.140×10^{-5}	0.0010
−12281.2800	0.0530	1.3160	1.0000	0.9470	25	1.412×10^{-5}	6.816×10^{-5}
−12281.2799	0.0530	1.5900	1.0010	0.948	30	1.707×10^{-5}	7.476×10^{-7}
−12281.2799	0.0530	1.8710	1.0000	0.9470	35	2.008×10^{-5}	5.527×10^{-8}
−12281.2799	0.0530	2.1520	1.0000	0.9470	40	2.310×10^{-5}	

△

5.5 Algorithm's efficiency

In this Section, we compare the speed of the proposed algorithm and the traditional method in the computation of the pdf (log-likelihood). When the total number of states (m) is large, it is time-consuming for the traditional method to calculate the log-likelihood. The first comparison is based on the theoretical assessment whilst the second comparison is based on numerical experiments.

5.5.1 Required flops

The first algorithm efficiency comparison is based on the concept of flop, which is proposed by Moler and Van Loan (2003) to quantify the magnitude of the time required for computations. More precisely, they defined a flop to be the time required for a particular computer system to

execute the FORTRAN statement

$$A(I, J) = A(I, J) + T * A(I, K),$$

where $A(I, J)$ represents the (I, J) element of matrix $\mathbf{A}$, T is a number, and $*$ is multiplication operation. This involves one floating-point multiplication, one floating-point addition, a few subscripts, index calculations, and a few storage references. The smaller the required flops, the faster the algorithm.

We start with the required flops for the traditional method. Moler and Van Loan (2003) derived the required flops for the scaling and squaring method (the algorithm used in **expm**) to calculate the matrix exponential $e^{\Lambda t}$, where Λ is a $m \times m$ matrix, are $O(m^3)$. Therefore, the required flops for the traditional method to calculate the probability distribution are $O(m^3)$. For a set of observation with n individuals, the required total flops for the traditional method to calculate the log-likelihood is $O(nm^3)$.

Let us assess the required flops for the proposed algorithm. Recall that the proposed algorithm evaluates the probability distribution at multiple points simultaneously by (5.3.9), and the calculation is truncated at J and satisfies (5.3.13). When evaluating the probability distribution at one point $\boldsymbol{\beta}$, which is a $1 \times m$ vector, the required flops are $Jm + J$. The calculation requires Jm flops for J $1 \times m$ vectors $\left(P^j_{1,1}, \ldots, P^j_{1,m}\right)$ $j = 1, \ldots, J$, and additional J flops for $\boldsymbol{\beta h}$. For a set of observations with n individuals, the required flops for $\boldsymbol{\beta}$ are $Jmn + J$, because the calculation involves J terms of $n \times 1$ matrices multiplying by $1 \times m$ matrices. The required flops are nm for each $n \times 1$ matrix multiplied by $1 \times m$ matrix. For the sum to obtain $\boldsymbol{\beta}$, the required flops are J. Therefore, the required flops for the proposed algorithm to calculate the log-likelihood are $O(Jmn)$.

Focusing on J for the proposed PTAM, its value depends on both the error tolerance ϵ and the rate $\nu t_{\max}$ by (5.3.13). The higher ϵ or $\nu t_{\max}$, the bigger the J. That a Poisson distribution with rate λ can be approximated by a normal distribution with mean equal to λ and standard deviation equal to $\sqrt{\lambda}$, when λ is large enough, is a well-known fact. For a fixed set (h_1, h_m, s, ψ) in the proposed PTAM, $\nu = m/\psi + \max_{i=1,\ldots,m-1}(h_i)$ when $m > \psi(h_m - \max_{i=1,\ldots,m-1}(h_i))$. This is because

$$m/\psi + \max_{i=1,\ldots,m-1}(h_i) > h_m,$$

and

$$\nu = \max_{i=1,\ldots,m-1}(m/\psi + h_i, h_m).$$

Since J is the $1 - \epsilon$ quantile of a Poisson random variable with rate $\nu t_{\max}$, when $\nu t_{\max}$ is large,

$$\Phi\left(\frac{J - \nu t_{\max}}{\sqrt{\nu t_{\max}}}\right) \approx 1 - \epsilon,$$

or

$$J \approx \nu t_{\max} + \Phi^{-1}(1 - \epsilon)\sqrt{\nu t_{\max}},$$

where Φ is the cumulative density function for the standard normal distribution and Φ^{-1} is the inverse function of Φ, $\Phi^{-1}(\Phi(x)) = x$. Hence,

$$J \approx m\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max} + \Phi^{-1}(1-\epsilon)\sqrt{m\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max}}, \tag{5.5.17}$$

when $\nu t_{\max}$ is large.

Remark 5.4. *Recall that* $\Phi^{-1}(1-10^{-10}) = 6.36$. *The approximation in (5.5.17) is dominated by* $m\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max}$, *when* $m\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max}$ *is large. For instance, when* $m\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max} = 10000$, $J \approx 10000 + 6.36 \times \sqrt{10000} = 10636$.

The required flops for the proposed algorithm to calculate the log-likelihood are

$$O(Jmn) \approx O\left(m^2 n\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max}mn + \Phi^{-1}(1-\epsilon)mn\sqrt{m\frac{t_{\max}}{\psi} + \max_{i=1,\ldots,m-1}(h_i)t_{\max}}\right),$$

whose dominant term is $m^2 n\frac{t_{\max}}{\psi}$ when m is large. As a result, the total required flops have the magnitude of $O(nm^2)$ when m is large.

The required flops for the proposed algorithm to calculate the log-likelihood of the proposed PTAM are about $1/m$ of that for the traditional method. For instance, the required flops for the proposed algorithm are about 1% of that for the traditional method when $m = 100$.

5.5.2 Comparison with the traditional method

It was shown that the proposed algorithm is theoretically faster than the traditional method. A speed comparison in so far as computing the log-likelihoods of simulated data under the proposed PTAM is concerned, will be conducted in this Subsection. All calculations are performed in MATLAB on the same computer equipped with an i7-6700k @4.0GHz CPU and 16GB RAM.

Required time versus m

Example 5.5. We compare the required time to calculate the log-likelihood by each method under different values of m. The assigned values of m are 10, 25, 50 and 100. For a given value of m, 1,000 sets of parameter values are randomly generated. For each set of parameter values, 5,000 observations are randomly simulated from the PTAM and the corresponding log-likelihood is calculated using both methods. The 1,000 required times to calculate the log-likelihoods in each method are obtained.

The empirical distribution of the required times for each m is presented in Figures 5.8-5.11; the red dotted lines in each histogram are the mean of the required times for each method.

Remark 5.5. *Figures 5.8-5.11 only reflect the speed of obtaining the log-likelihood for a set of parameter values. In the calculations entailed in the search for the MLE, thousands of log-likelihood evaluations are required. Certainly, the cumulative time in completing the entire MLE-search process would be significantly different in each method.*

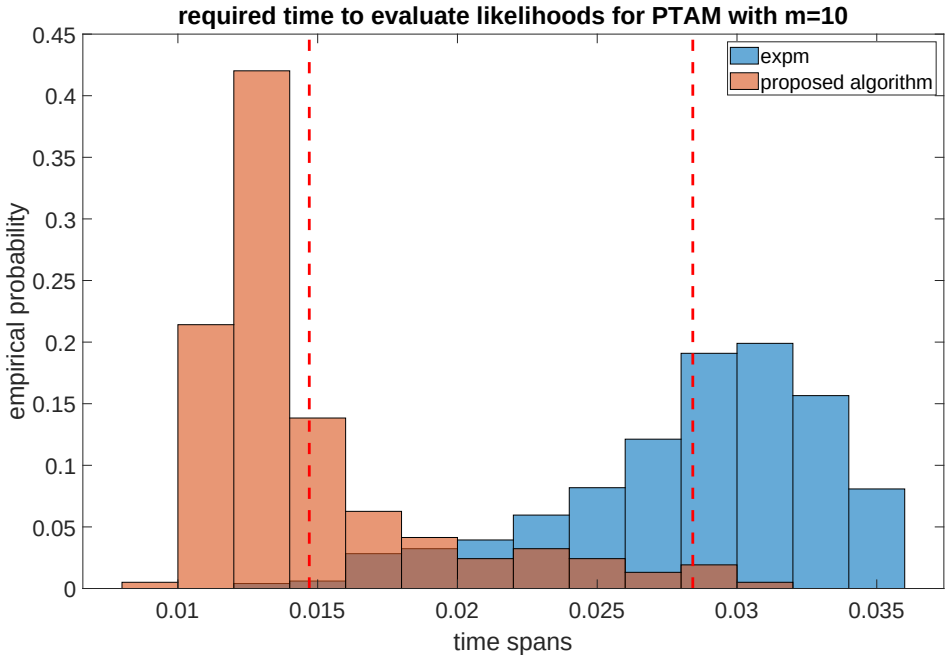

Figure 5.8: Required time (in seconds) to calculate the log-likelihood when $m = 10$.

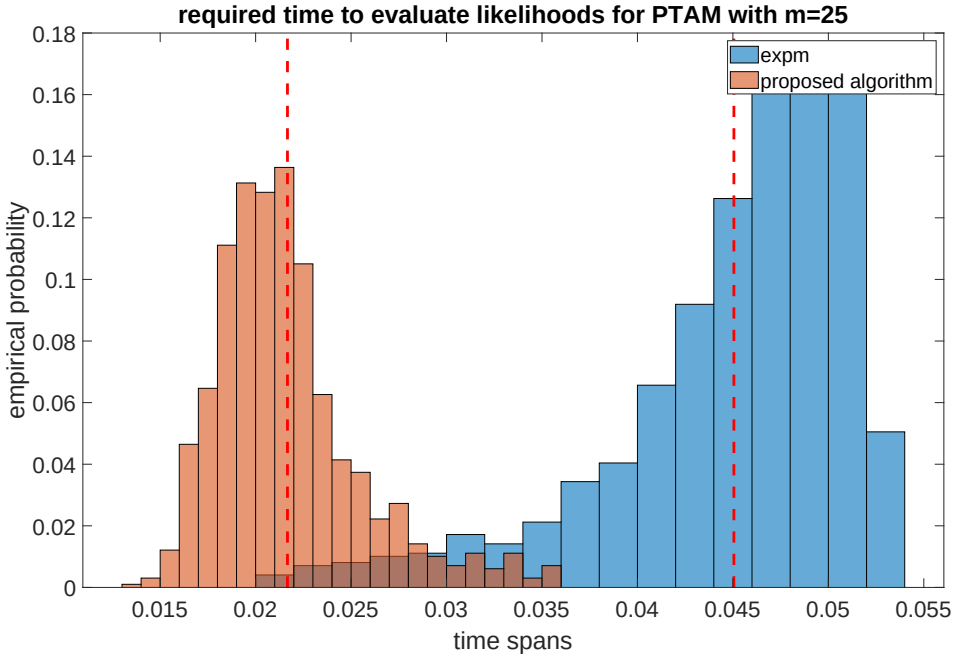

Figure 5.9: Required time (in seconds) to calculate the log-likelihood when $m = 25$.

In Figures 5.8-5.9, the right tail of the histogram for the proposed algorithm overlaps with the left tail of the histogram for the traditional method method when m is relatively small (e.g., $m = 10$ and $m = 25$). Meanwhile, the mean of the required times is substantially less for

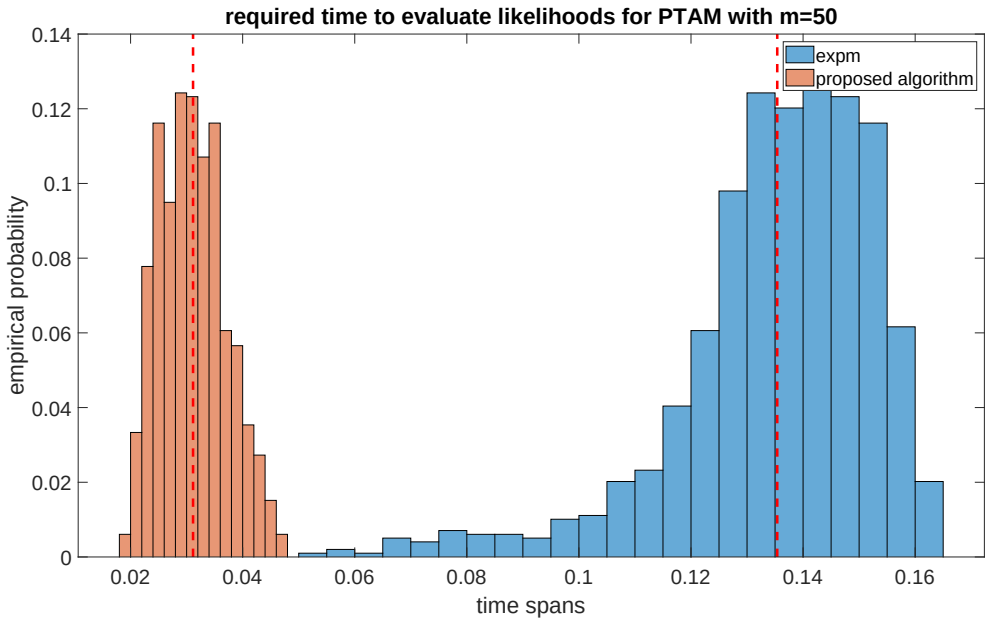

Figure 5.10: Required time (in seconds) to calculate the log-likelihood when $m = 50$.

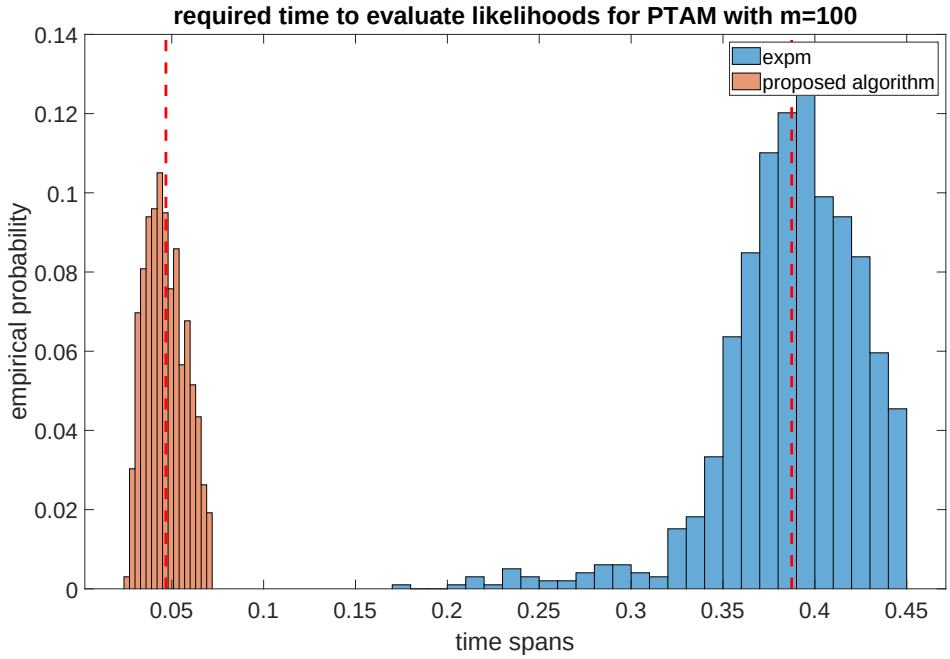

Figure 5.11: Required time (in seconds) to calculate the log-likelihood when $m = 100$.

the proposed algorithm. This shows that the proposed algorithm is faster than the traditional method on average, but the speed advantage is not significant. It is worth noting that more than 50% of the recorded times are smaller for the proposed algorithm in Figures 5.8 – 5.9. On the other hand, when m is relatively big (e.g., $m = 50$ and $m = 100$), all the recorded times are

smaller than those of the proposed algorithm; see Figures 5.10- 5.11. Additionally, the distance between the required time distributions for each method go further away as m increases from $m = 50$ to $m = 100$. This indicates that the fast speed advantage of the proposed algorithm is more significant for a larger m. One can also confirm the speed advantage of the proposed algorithm by comparing the mean required times. For example, in Figure 5.10, the mean of the required times for the traditional method is 0.135 seconds, whilst for the proposed algorithm is 0.03 seconds or 22.2% of the mean required time for the traditional method when $m = 50$. In Figure 5.11, the mean of the required times for the traditional method is 0.40 seconds, whilst for the proposed algorithm it is 0.05 seconds, or 12.5% of the time for the traditional method when $m = 100$. On average, the ratio of the required time for the proposed algorithm over that for the traditional method becomes smaller as m increases, attesting to the speed advantage of the proposed algorithm for a large m (e.g., $m \geq 50$).

Let us apply the two-sample Kolmogorov-Smirnov test on the empirical cumulative density distribution (cdf) of required time(s) for each method, under the null hypothesis that two required time distributions are the same. The test statistic is

$$D = \sup_t |F_{1,n}(t) - F_{2,n}(t)|,$$

where $F_{1,n}(t)$ and $F_{2,n}(t)$ are the empirical cdf's of the required time(s) for methods 1 and 2, respectively. The corresponding statistics under each situation is displayed in Table 5.2. Very low p-values (close to 0) indicate that the distributions of the required times to evaluate the log-likelihood for a set of parameter values are significantly different between two methods. The speed superiority of our method is definitely going to be magnified when taking into account the complete MLE search.

Table 5.2: Two-sample Kolmogorov–Smirnov test.

	D	p-value
$m = 10$	0.9670	0
$m = 25$	0.9720	0
$m = 50$	1.0000	0
$m = 100$	1.0000	0

△

Trend of mean required time versus m

Example 5.6. The second comparison covers the trend of the mean required time to calculate the log-likelihood as m increases from 10 to 200 by each method. The sample size is fixed at 5,000 for each log-likelihood, and we use the same procedure to collect the required times to calculate the log-likelihoods by both methods. For each value of m, we randomly generate

1,000 sets of parameter values. For each set of parameter values, we randomly simulate 5,000 data, and compute the corresponding log-likelihood using both methods.

The mean of 1,000 required times for each method is a statistic that represents the empirical algorithm's speed and the result is shown in Figure 5.12. The mean required time for the traditional method (blue line) increases much faster than that for the proposed algorithm (red line) as m increases.

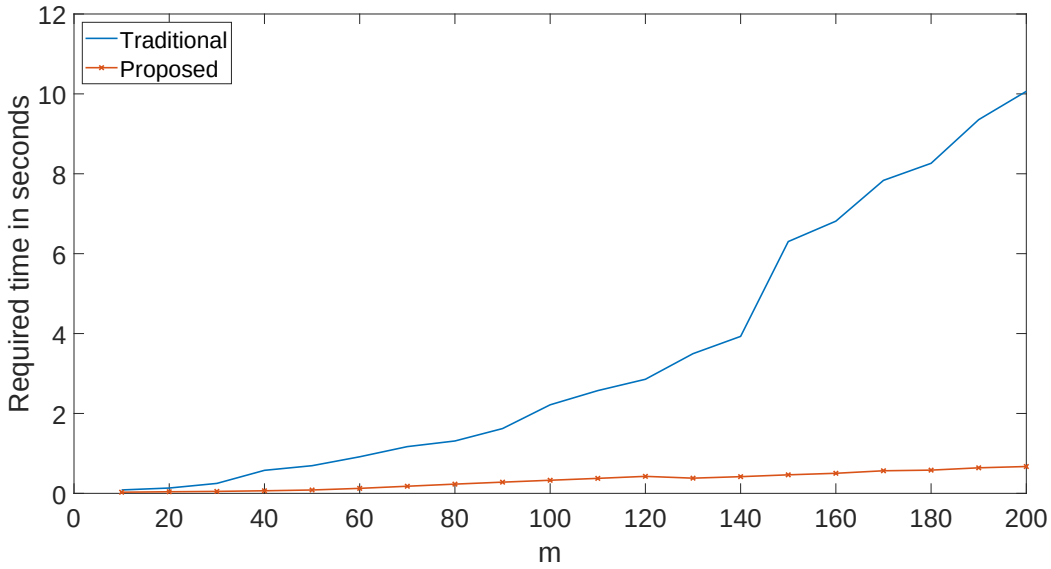

Figure 5.12: Mean of 1,000 required times in the calculation of the log-likelihood with different m for the traditional and proposed methods.

According to our previous analysis, the required flops for the traditional method are $O(nm^3)$, whilst the required flops for the proposed algorithm are $O(nm^2)$. We shall predict the required time to compute the log-likelihood for any given m by each method. To do this, we fit, in the least-squares sense, the required times under the traditional method to a polynomial $t(m)$ of degree 3, where $t(m) = a_3m^3 + a_2m^2 + a_1m + a_0$, and the required times under the proposed algorithm to a polynomial $t(m)$ of degree 2, where $t(m) = a_2m^2 + a_1m + a_0$. The estimates for the coefficients are given in Table 5.3.

Table 5.3: Coefficient estimates of the fitted curve for the required time to calculate the log-likelihood under different m's by each method.

	Traditional	Proposed
$\hat{a_0}$	0.3465	−0.0397
$\hat{a_1}$	−0.0115	0.0032
$\hat{a_2}$	0.0003	1.7448×10^{-6}
$\hat{a_3}$	1.1716×10^{-7}	

The respective fitted results for the traditional and proposed methods with a 95% prediction

interval are diagrammed in Figures 5.13 and 5.14. We extrapolate from both fitted curves up to $m = 300$. It is clear that the increasing trend is less dramatic for the proposed algorithm by checking the magnitude of the vertical axis (0-35 for the traditional method versus 0-1.4 for the proposed algorithm).

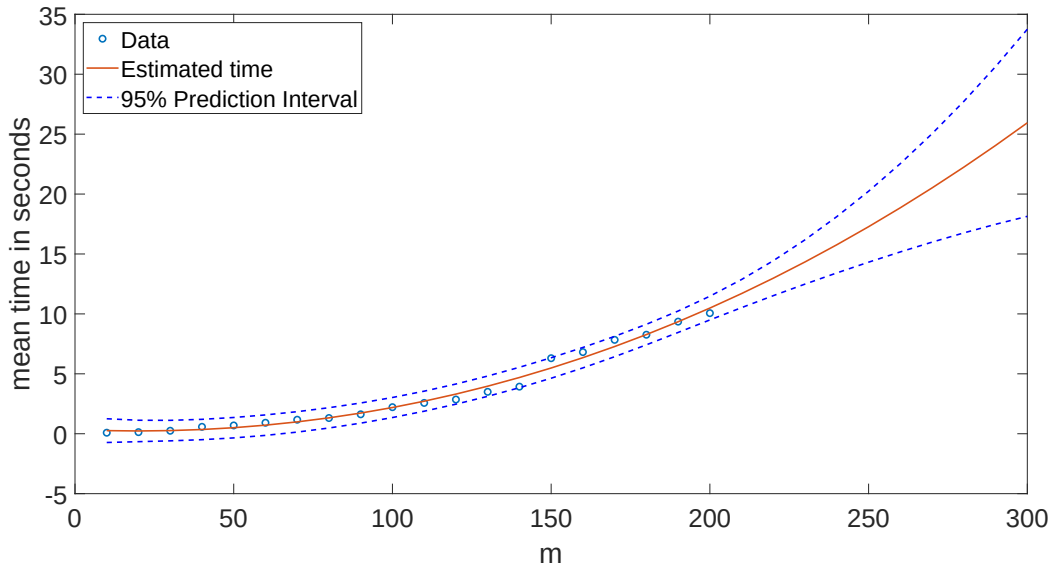

Figure 5.13: Estimated time with a 95% prediction interval in the calculation of the log-likelihood with different m's under the traditional method.

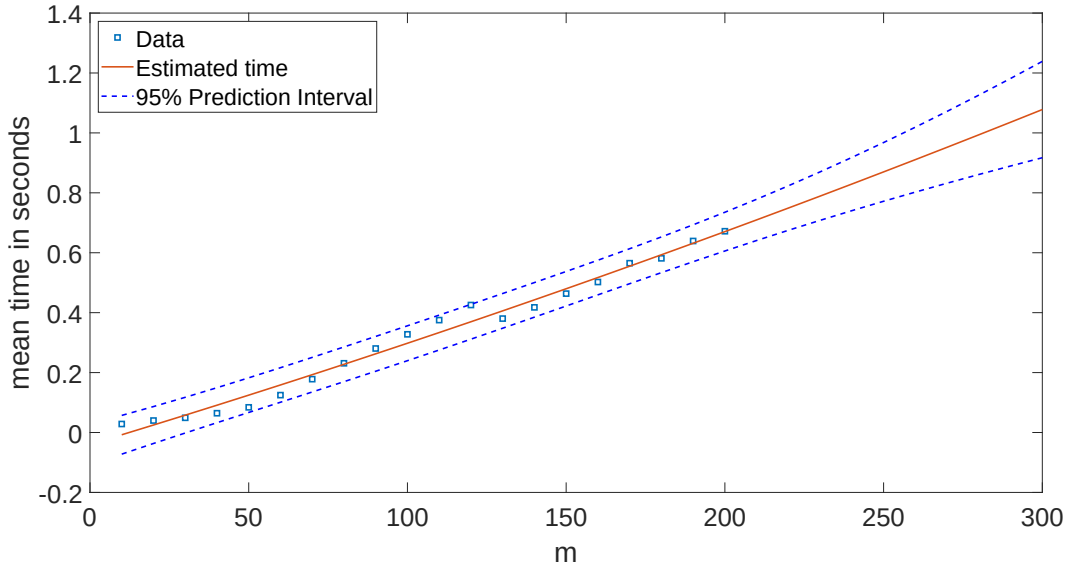

Figure 5.14: Estimated time with a 95% prediction interval to calculate the log-likelihood with different m's under the proposed method.

The estimated required times for the traditional method and for the proposed algorithm in

the calculation of the log-likelihood involving 5,000 observations with an $m-$state PTAM are $t_{trd}(m) = 0.3465-0.0115m+0.0003m^2+1.1716\times10^{-7}m^3$ with coefficient of determination $R^2 = 0.9842$ and $t_{prop}(m) = -0.0397 + 0.0032m + 1.7448 \times 10^{-6}m^2$ with coefficient of determination $R^2 = 0.9857$. The high value of R^2 supports a goodness-of-fit of both polynomial models. Thus, with $m = 300$, the approximated time to calculate the log-likelihood by the traditional method is 26 seconds, whilst the approximated time by the proposed algorithm is only 1 second.

The difference in the estimated times by both methods is $t_{trd}(m) - t_{prop}(m) = 1.1716 \times 10^{-7}m^3 + 2.8597 \times 10^{-4}m^2 - 0.0147m + 0.3863$. The first derivative of $t_{trd}(m) - t_{prop}(m)$ with respect to m is $3.5148 \times 10^{-7}m^2 + 5.7194 \times 10^{-4}m - 0.0147$, which is greater than 0 when $m > 25$. So the difference in the predicted time by each method is an increasing function of m when $m > 25$.

△

Trend of mean required time versus sample size

Example 5.7. The third investigation is concerned with the trend of the mean required time for each method in calculating the log-likelihood as sample size increases. We fix $m = 200$ and try different sample-sizes in this experiment. Similar to previous experiments, we randomly generate 1,000 sets of parameter values. For each set of parameter value, n lifetimes are randomly simulated, where $n = 100, 200, 300, \ldots, 5000$ for each set of observations. The log-likelihood for each set of observations is calculated by both methods. For each n, 1,000 required times to calculate the log-likelihoods are recorded in each method. The mean of the 1,000 empirically required times is used to represent the required time to calculate the log-likelihood for n individuals.

The mean required time for each n by each method is plotted in Figure 5.17. The mean required time for both methods increase linearly as the sample size increases, which is consistent with our theoretical analysis. We fit the required time to a linear model $t(n) = b_1n + b_0$ in the least-squares sense. The estimates of the coefficients are exhibited in Table 5.4. The estimated time $t(n)$ for the proposed algorithm follows $t(n) = 0.0001n + 0.1137$, and the estimated time for the traditional method is $t(n) = 0.0019n + 1.3109$. The corresponding R^2 are 0.9726 and 0.9744. Again, the high value of R^2 indicates both linear models fit reasonably well. The fitted lines with their 95% prediction intervals are displayed in Figures 5.15-5.16.

Table 5.4: Coefficient estimates of the fitted line for the required time in calculating the log-likelihood for different n's in each method.

	Traditional	Proposed
$\hat{b_0}$	1.3109	0.1137
$\hat{b_1}$	0.0019	0.0001

Both slope estimates are greater than 0 implying that the required time increases as the sample size increases. The slope for the proposed algorithm is 0.0001, which is $0.0001/0.0019 = 5.26\%$ of the slope for the traditional method. Therefore, the predicted required time increases

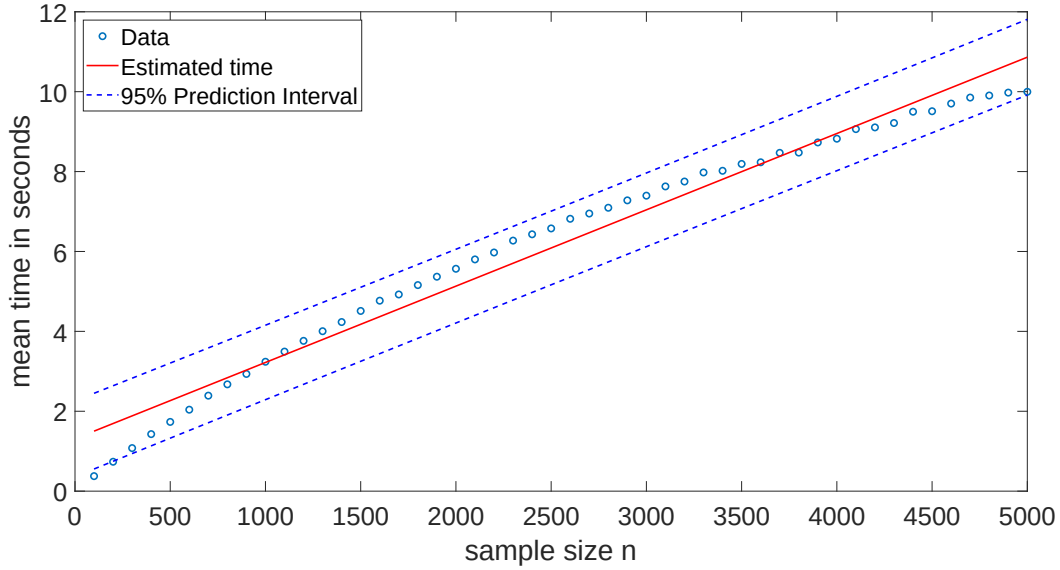

Figure 5.15: Estimated time with 95% prediction interval to calculate the log-likelihood with different n for the traditional method.

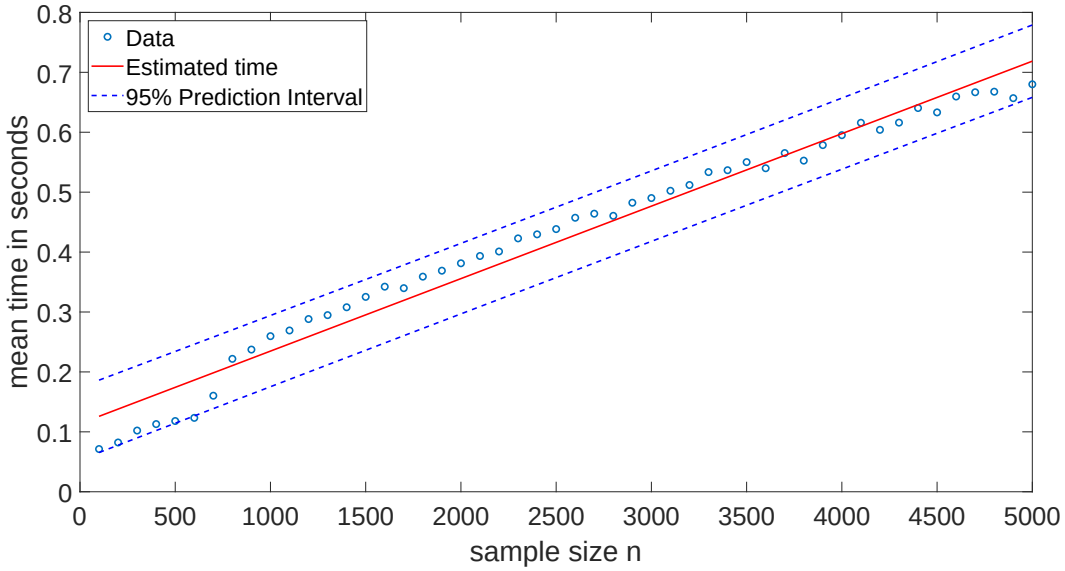

Figure 5.16: Estimated time with 95% prediction interval to calculate the log-likelihood with different n for the proposed method.

much slower for the proposed algorithm as the sample size increases. Thus, the larger the sample size, the more significant the speed advantage of the proposed algorithm.

△

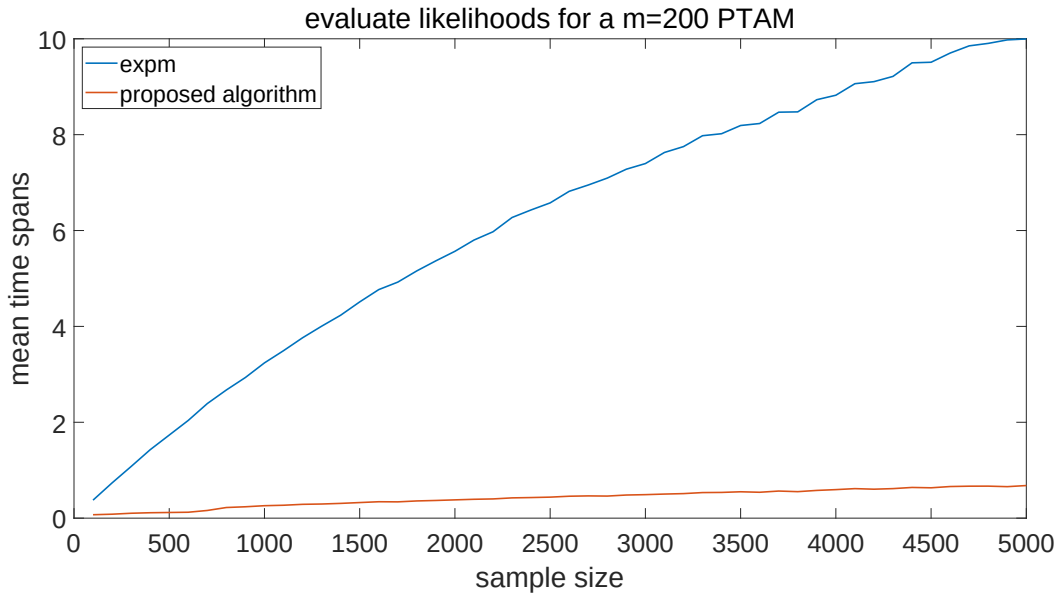

Figure 5.17: Mean required time in the calculation of the log-likelihood for different sample sizes.

5.6 Proposed calibration procedure

One particular issue when optimising the objective function for the Coxian model is that the optimised result may be sensitive to the initial value. Marshall and Zenga (2012) assessed the fitting process for a Coxian model and found that the estimation result was largely dependent on both the initial parameter value and the actual data set. The sensitivity issue may cause the estimation to be unreliable. We probe the sensitivity issue in this section by providing the following example.

Example 5.8. We simulate 4 sets of lifetime observations with n lifetimes ($n = 500, 1000, 1500, 2000$) from the proposed PTAM with parameter values $h_1 = 0.0018$, $h_m = 1.2752$, $s = -0.0734$, $\psi = 55$ and $m = 100$, which are estimated from the Channing house data. Furthermore, it is assumed that $m = 100$ and the parameters to be estimated (the inputs of the objective function) are h_1, h_m, s and ψ. For each set of observations, 100 initial values are randomly simulated to initiate the numerical optimisation in search for the MLE.

The maximised log-likelihoods are plotted in Figure 5.18, in which most optimisation outputs are located at the maximum log-likelihood value. As expected, some are smaller than the maximum log-likelihood value; that is, the maximum log-likelihood value is about -1860 whilst a few outputs are around -1873 for $n = 500$.

△

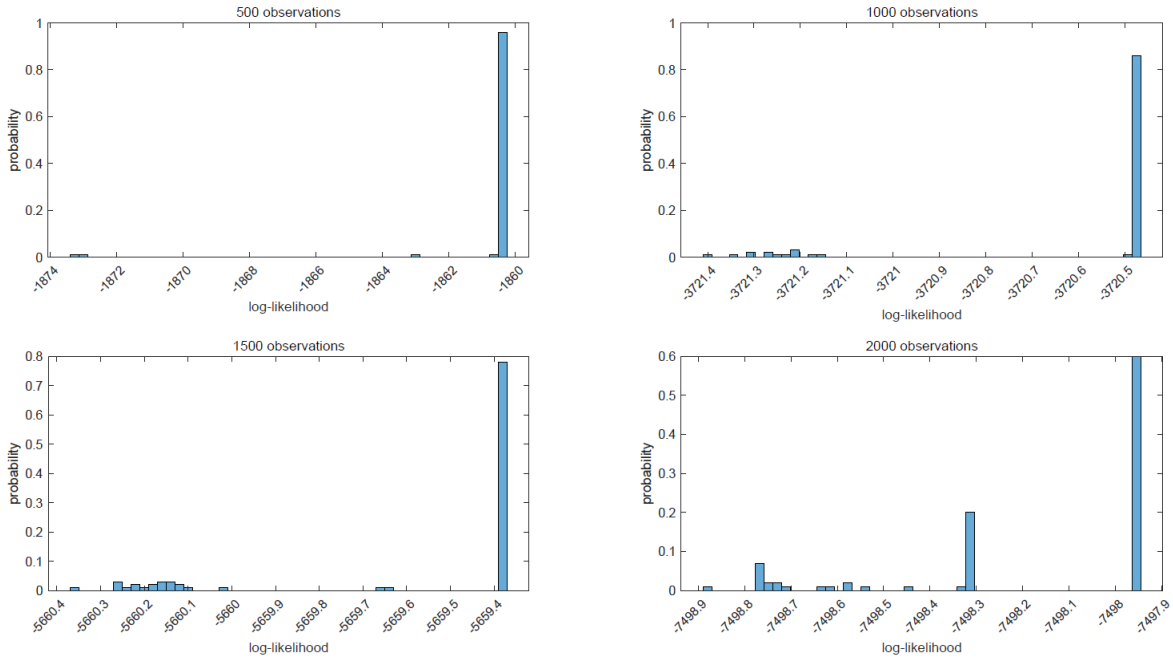

Figure 5.18: Empirical distribution of one hundred estimates using 100 randomly simulated initial values for different sample sizes.

5.6.1 Sensitivity assessment

Marshall and Zenga (2012) proposed a measure of convergence performance called rate of algorithm's success (RAS) defined by

$$\text{RAS} = \frac{\text{number of acceptable results}}{\text{number of successful estimations}}(100\%),$$

where the successful estimations refer to the estimation procedure that complete the entire calculation process. The acceptable results are those that fall in the range of acceptable parameters.

On the basis of a sufficient condition in Theorem 5.4, the range of acceptable parameters are

$$l_N(\hat{\theta}) - l_N(\theta) < M(\theta)\epsilon,$$

where $\hat{\theta}$ is the parameter value that maximises the log-likelihood. Recall the default value $\epsilon = 10^{-10}$. The maximum log-likelihoods in Example 5.6 are -1860.2659, -3720.4667, -5659.3706 and -7497.9491 for $n = 500, 1000, 1500, 2000$, respectively.

We use the RAS to assess how sensitive the optimisation output is to the initial value. The RAS for each sample is reported in in Table 5.5. There are 3 unsuccessful estimations for $n = 500$ and 1 unsuccessful estimation for $n = 1500$. The Hessian matrix evaluated at each unsuccessful estimation is not negative definite (not all eigenvalues of the Hessian matrix are positive), implying that a local maximum is not found. The RAS is poor, i.e., less than 20%, for $n = 500$, whilst the RAS is greater than 50% for $n = 1000, 1500$, and 2000. The RAS

increases when the sample size increases from $n = 500$ to $n = 1000$; then, the RAS decreases as the sample size continues to increase from $n = 1000$ to $n = 2000$. The low RAS suggests that the estimated result based on one set of initial values is unreliable.

Table 5.5: Number of acceptable results, number of successful estimations, and the RAS for 4 sets of simulated lifetimes.

	# of acceptable results	# of successful estimations	RAS (%)
n=500	19	97	19.6
n=1000	75	100	75
n=1500	66	99	66.7
n=2000	56	100	56

Furthermore, by checking the eigenvalues of the Hessian matrix evaluated at the estimated point, the smallest eigenvalue is always close to 0 (around 0.01 for $n = 500$, around 0.05 for $n = 1000$, around 0.1 for $n = 1500$, and around 0.2 for $n = 2000$). The smallest eigenvalue near 0 means the log-likelihood surface is flat, which is problematic for optimisation. As a result, it is not easy to find the global maximum-likelihood estimate for this illustration.

In terms of the estimate for each initial value, the distributions of estimate for each parameter for the successful estimations are in Figure 5.19, Figure 5.20, Figure 5.21, and Figure 5.22. The estimated results for each parameter are stable with a distinguishable tall bar in each graph. Therefore, the estimate for each parameter is stable. The estimates for all parameter are significantly biased when $n = 500$, and, as we expected, the estimates are closer to the true values as sample size increases. The estimates of h_m are far from the true value, but our focus for this example is on the stability of the estimate. Therefore, we conclude that the estimated result is stable in the general sense.

Remark 5.6. *In all Figures in this Section, the red line locates the mean and the black dash lines are the lower and upper limits of the 95% confidence interval of the estimate.*

5.6.2 Strategy to overcome the initial-value sensitivity

We demonstrated that the estimate may be sensitive to one set of initial values, but it is generally stable. A random simulation of a number sets of initial values can mitigate, to some extent, the sensitivity issue. In particular, the calibration process starts with generating a number of initial values, and then runs the numerical optimisation with each initial value. The final estimate is the one that gives the best objective function amongst the optimisation outputs. For example, Marshall and Zenga (2012, 2009a) used 103 initial values to find the MLE for their 4-state Coxian model.

On the one hand, the more initial values we generate, the more likely the final estimate can reach the optimum value. On the other hand, the more initial values we generate, the slower the estimation process. To balance both aspects in implementation, we suggest that 20 sets of

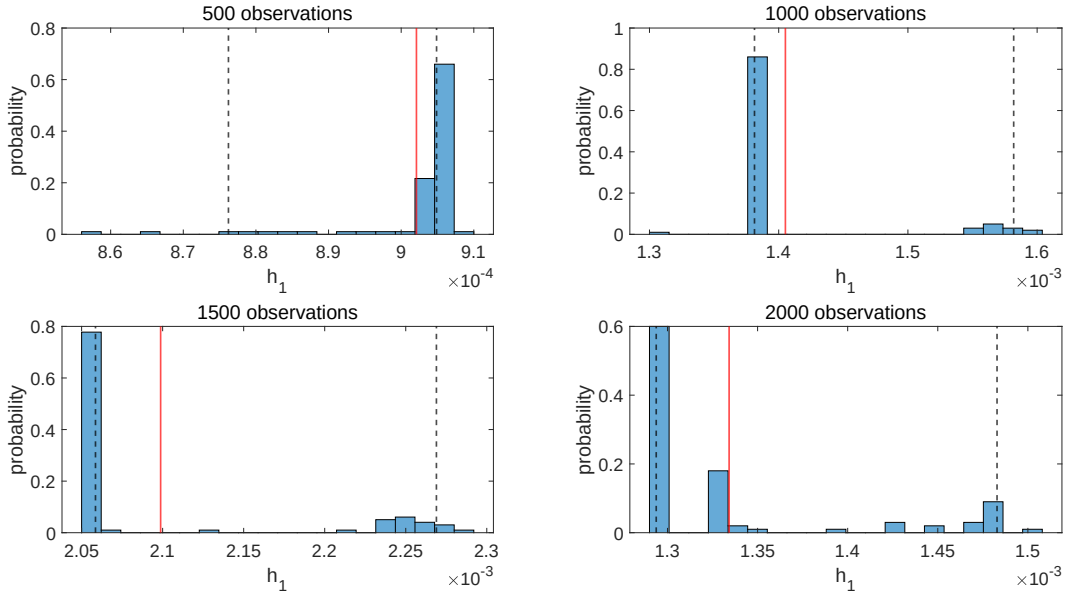

Figure 5.19: Empirical distribution of one hundred estimates of $\hat{h}_1$ under different sample sizes.

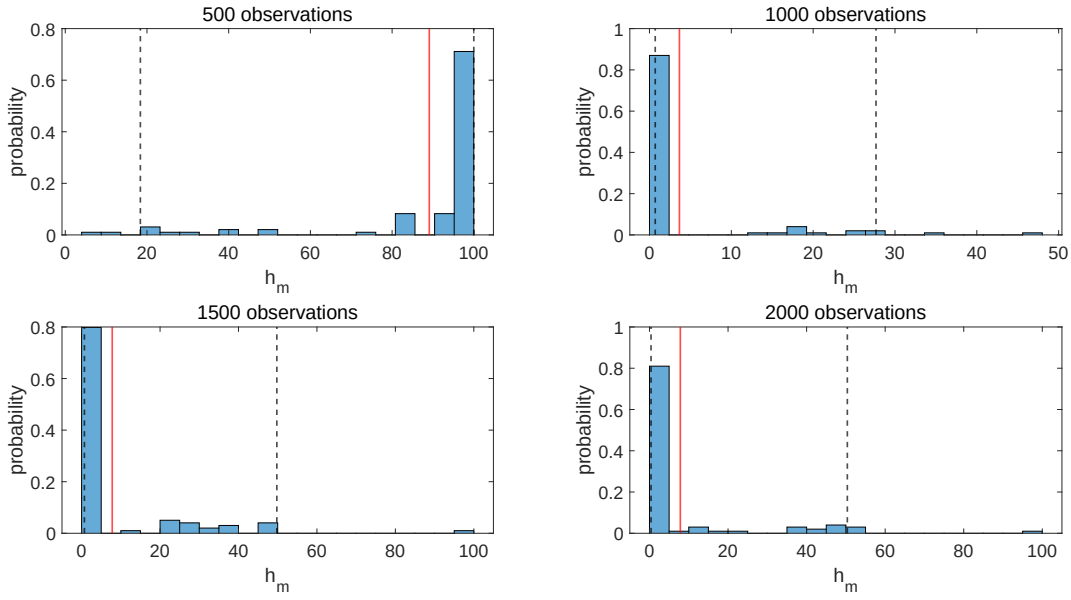

Figure 5.20: Empirical distribution of one hundred estimates of $\hat{h}_m$ under different sample sizes.

initial values will be sufficient for the proposed PTAM. For example, the RAS in the above example is between 20% and 75%. The probability that the final estimate is not in the range of acceptable parameters is between $0.25^{20} = 9\times10^{-13}$ and $0.8^{20} = 1.15\%$. Such probabilities in this range are so small that it is likely the final estimate is in the acceptable range.

In summary, the estimate based on one set of initial values may be sensitive to the initial point, but the estimate based on a number of randomly generated initial values is relatively

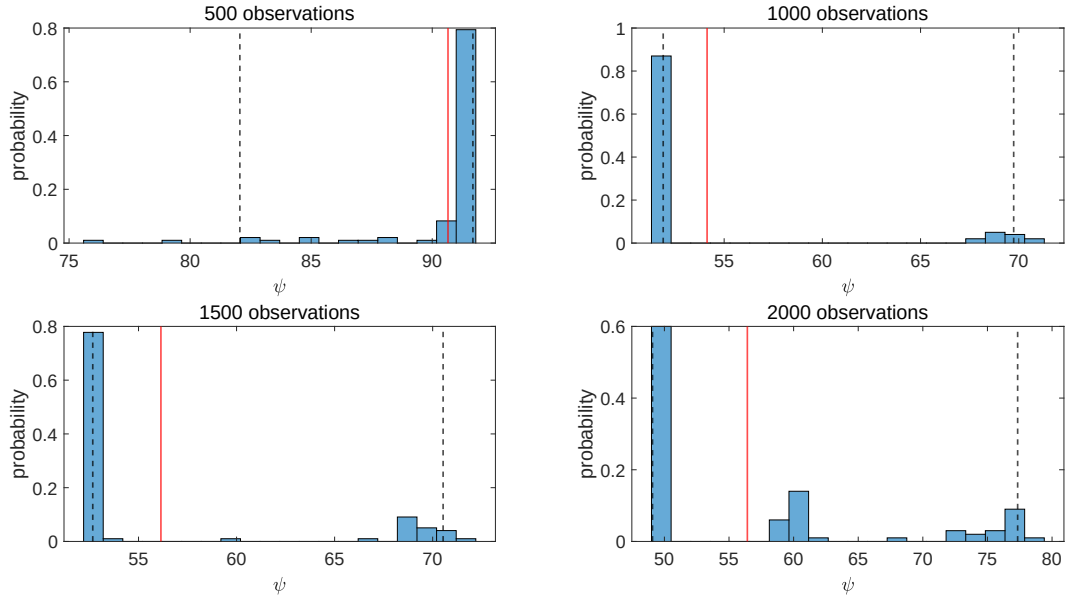

Figure 5.21: Empirical distribution of one hundred estimates of $\hat{\psi}$ under different sample sizs.

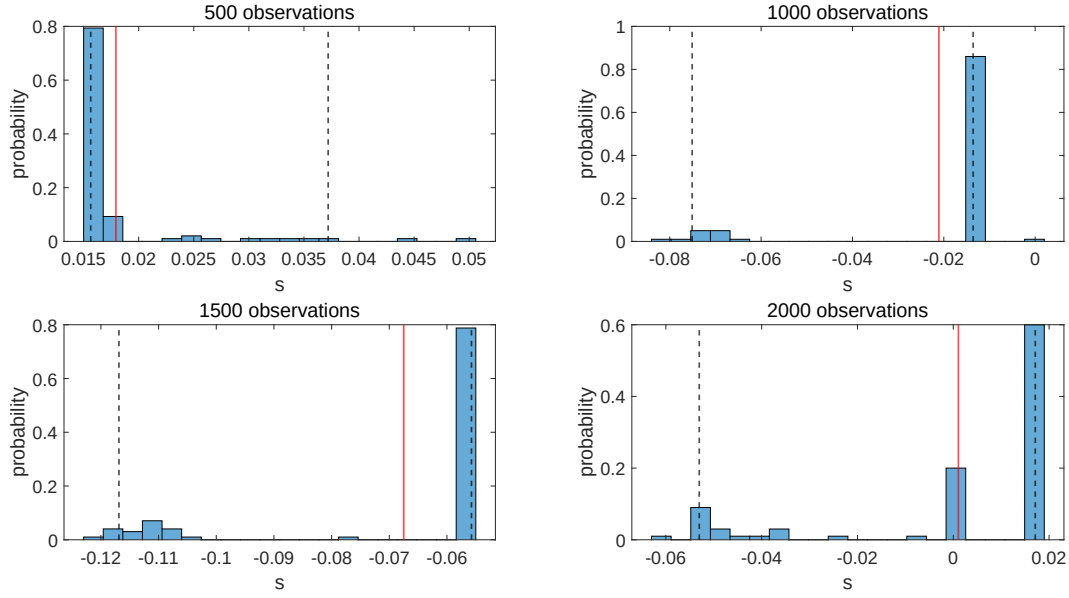

Figure 5.22: Empirical distribution of one hundred estimates of $\hat{s}$ under different sample sizs.

computationally helpful.

Chapter 6

Identifiability and estimability

We investigate the identifiability and the estimability of the proposed PTAM in this Chapter. The proposed PTAM will be shown to be identifiable, but it has poor estimability when the observation is only the time until absorption. The poor estimability can be visualised by the narrow ellipses in the contour plots, and the flat marginal likelihood functions.

We use a data-cloning method to assess model estimability. Through this method, the PTAM's estimability under different scenarios could be compared, from the best scenario that the Markov process is fully observable for each individual to the worst scenario that the only observation is the time until absorption for each individual. Some scenarios in the middle are: (i) the state can be observed every couple years; and (ii) the state can be measured with some error. The additional information about the state improve the PTAM's estimability.

6.1 Some pertinent background on model identifiability and model estimability

A simple version of the relationship between model inference and model identifiability is that model inference can be accurate if and only if the model is identifiable. When the model is identifiable, the true parameter values can be learnt theoretically. Thus, model identifiability is one of the fundamental statistical properties required for both model inference and hypothesis testing. Hsiao (1983) included a survey on the development of the model identifiability conditions since the publication of Fisher's book (Fisher, 1966). Hsiao (1983) attempted to describe the identifiability issues in a mathematical manner. The identifiability criteria for some specific models were provided. Furthermore, a discussion about the identifiability issue from the perspective of the Bayesian approach was addressed as well. In Casella and Berger (2002), and Lehmann and Casella (2006), some examples of models that are non-identifiable were given. The true parameter values in those non-identifiable models can never be learnt from the data.

Model inference plays an important role in recovering the true parameters based on observed data. For example, either a point estimate paired with its estimated standard error or the confidence interval for each parameter is helpful to get some insights on the underlying model. However, the performance of such statistics depends highly on model identifiability and model estimability. In other words, two basic questions need to be answered before starting model inference: *Which models are identifiable?* and *Which models are estimable?* The formal def-

inition of each concept will be provided in the later part of this Chapter. We start with some general ideas.

One of the popular ways to estimate the model parameters is through the MLE method. Given a data set, the MLE of a model is plausibly unique only when the likelihood surface has a distinguishable peak and the likelihood surface has much curvature around the peak. The uniqueness of the MLE is related to model estimability, which has been well studied for linear models. For example, Alalouf et al. (1979) studied the required conditions on $\boldsymbol{X}$ such that a general linear model $g(\boldsymbol{Y}) = \boldsymbol{X\beta}$ is estimable. Bunch (1991) demonstrated some non-identifiable covariance parameters in a model that has a linear-in-parameters multinomial probit framework. Lele et al. (2010) applied data cloning to test the estimability of generalised linear mixed models.

There are some studies on model estimability for general models. For instance, Miettinen (1976) assessed the estimability of the ratio of incidence densities in case-referent studies. Jacquez and Greif (1985) attempted to develop practical approaches to examine local identifiability for particular parameters. They showed that if one has initial estimates of the parameters, then the local identifiability, estimability, and optimal sampling design involve similar considerations. To generalise the required conditions for a model to be identifiable and estimable, Bunke and Bunke (1974) developed a general theory of both parameter identifiability of unbiased decision functions and estimable optimal decision sets. Furthermore, they showed that estimability and identifiability coincide for linear models, and the linear parameters in multivariate linear models can be viewed as estimable and identifiable with a suitable loss function. A good survey of previous studies on identifiability and estimability is McLean and McAuley (2012). In their survey, they compared the questions to answer between identifiability analysis and estimability analysis, which clearly distinguish one analysis from the other. Furthermore, they recommended some studies on identifiability analysis and some studies on estimability analysis in each of the following categories:

- alternative names
- information used
- mathematical techniques
- model type
- model complexity

The relationships between identifiability and estimability are as follows:

- a non-identifiable model must be non-estimable;
- an estimable model must be identifiable.

Nevertheless, it is worth noting that the opposite of previous relationships may not be true. An identifiable model is not necessarily estimable for any set of data, and a non-estimable model is not necessarily non-identifiable. After defining model identifiability and model estimability, we provide a counterexample to show the opposite of the relationships may not hold.

Recall the fact that a model has accurate model inference if the model is identifiable, and another fact that the MLE is unique if the model is estimable. There are some challenges on interpreting the calibrated result if the MLE is not unique. Since the proposed PTAM is a model for ageing process, both the model inference and model interpretation are important.

6.2 Trade-off between model flexibility and inferential power

A model with more parameters is more flexible in reproducing a variety of distributions. Nonetheless, including more parameters in a model increases model complexity. A well-known theorem, proved initially by Cox (1955) and summarised by Bolch et al. (2006), states that any distribution with non-negative support can be approximated well by a phase-type distribution or a Coxian distribution. The models with more parameters have a higher degree of freedom resulting distribution, and their distributions are more flexible. From the standpoint of model flexibility, a model with more parameters is preferable.

A model with more parameters though has less inferential power. As the model includes more parameters, it is less likely to learn the "true" parameter values from the data. This phenomenon is related to the curse of dimensionality – when the dimension increases, the volume of the parameter space increases so fast that the available data becomes sparse, especially in high dimensional data. All objects then appear to be sparse and dissimilar in many ways, which prevent common data-organisation strategies from being efficient. This sparsity is problematic for model inference. Trunk (1979) constructed a mixed Gaussian example such that $P(X \mid \omega_1) \sim N(\boldsymbol{\mu_1}, \boldsymbol{I})$, $P(X \mid \omega_2) \sim N(\boldsymbol{\mu_2}, \boldsymbol{I})$, where $\boldsymbol{\mu_1} = \boldsymbol{\mu} = -\boldsymbol{\mu_2}$ is an $n-$vector mean value whose ith component is $(1/i)^{1/2}$, and $P(\omega_1) = P(\omega_2) = \frac{1}{2}$. They proved that the probability error, $P\left(X^\top \boldsymbol{\mu} \geq 0 \mid \omega_2\right)$, approaches zero as the dimensionality increases when the mean values were known; whilst the probability error approaches 0.5 as the dimensionality increases when the mean values were estimated from a finite number of samples. Friedman (1997) demonstrated an example that the optimal K for the K-nearest neighbor method increased rapidly as the dimension of parameters increased. Both examples demonstrate that the issues occur in high dimensionality. From the perspective of inferential power, a model with fewer parameters is preferable.

To strike a balance between model flexibility and inferential power, the "best" model is typically selected from some model candidates through the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). These criteria deal with the trade-off between the goodness of fit and the model complexity. Cavanaugh and Neath (2019) summarised the background, derivation, properties, application, interpretation, and refinements of the AIC, and similar research endeavours were contributed by Neath and Cavanaugh (2012) for the BIC. Vrieze (2012) compared the AIC and the BIC, and found that:

- The BIC is consistent in selecting the right model if the actual model is amongst the candidates, under which the BIC is more efficient than the AIC.
- The AIC is more efficient if the actual model is not amongst the candidates since the AIC asymptotically chooses the model by minimising the mean squared error of estimation.

Since the number of parameters in the proposed PTAM is fixed for any m, and both the AIC and the BIC are equal to a constant minus two times the log-likelihood, minimising either the

AIC or the BIC is equivalent to maximising the log-likelihood (MLE), which is our criterion to calibrate the PTAM. As a result, when calibrating the PTAM, we do not need to worry about the trade-off between goodness of fit and the model simplicity.

From the perspective of model simplicity, the proposed PTAM has 5 parameters, and this number of parameters is relatively small compared with the models in Lin and Liu (2007) and Govorun et al. (2018). From the perspective of model flexibility, the proposed PTAM allows some flexibility on the dying rates resulting to a variety of lifetime distributions (see section 3.5).

6.3 Identifiability of the proposed PTAM

Which models are identifiable? To answer this question, we need to define what model identifiability is.

Definition 6.1. *Given the model parameter space Θ and the model support T, the model parameter $\theta \in \Theta$ is identifiable if, for any $\theta_1, \theta_2 \in \Theta$ such that the probability density functions $f_{\theta_1}(t) = f_{\theta_2}(t)$ for $\forall\ t \in T$, then it must be that $\theta_1 = \theta_2$. Or equivalently, for any $\theta_1 \neq \theta_2 \in \Theta$, there is a $t_0 \in T$ such that $f_{\theta_1}(t_0) \neq f_{\theta_2}(t_0)$. If the (model) parameter fails to be identifiable, then the (model) parameter is non-identifiable.*

It is well-known that the general phase-type models are non-identifiable. Marshall and Zenga (2009b) indicated that finding efficient numerical procedures in the estimation of the phase-type model parameters remains an open problem. A part of the estimation challenge is due to the phase-type model's non-identifiability.

Remark 6.1. *There may be two different phase-type models whose resulting lifetime distributions are indistinguishable, i.e., their pdf's are the same but their corresponding set of parameter values could be different.*

Example 6.1. A case of a non-identifiable phase-type model is the $m \times m$ degenerated transition matrix

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda_1 + h) & \lambda_1 & & & \\ & -(\lambda_2 + h) & \lambda_2 & & \\ & & \ddots & & \\ & & & -(\lambda_{m-1} + h) & \lambda_{m-1} \\ & & & & -h \end{bmatrix}. \tag{6.3.1}$$

Under this situation, the transition rates from any transient states to an absorbing state are identical. The resulting pdf for any $\lambda_i \geq 0$, $i = 1, ..., m-1$, is equal to the pdf of an exponential distribution with rate h, i.e., $f(t) = he^{-ht}$.

△

Ramírez-Cobo et al. (2010) constructed a simple but non-identifiable two-state hidden Markov model. The non-identifiability of the general phase-type model is due to overparameterisation – the number of model parameters is more than needed to generate the required distribution.

Different restrictions on the phase-type models are imposed to construct identifiable phase-type models. One popular class of phase-type models is the Coxian model, which is proved to be dense in the field of non-negative-valued distributions (Cox, 1955). The identification of the Coxian models is an open problem, and many contributions to the Coxian model calibration have been given in the past few decades (e.g., Dempster et al. (1977) and Asmussen et al. (1996) utilising the EM algorithm to maximise likelihoods). In order to enhance the interpretations, Augustin and Büscher (1982) presented a new representation for the Coxian models by decomposing the Markov chain into several Markov chains with associated entry probabilities. In Faddy (1994), Coxian models are fitted to several example data sets; whilst in Faddy (1998), inferring the number of states for a Coxian model by likelihood ratio testing was shown. Faddy (2002) exploited the penalised maximum likelihood estimation in distinguishing the eigenvalues of the transition intensity matrix. The characterisation of the Coxian model as it relates to its identifiability is contained in the next theorem.

Theorem 6.1. *A Coxian model has a unique minimal representation.*

Proof. See Cumani (1982).

■

Remark 6.2. *Recall that the representation for a Coxian model is* $(\boldsymbol{\alpha}, \boldsymbol{\Lambda})$*, where the dimension of* $\boldsymbol{\Lambda}$ *depends on the total number of states m. The representation with the smallest m such that it can generate the target lifetime distribution is called the minimal representation.*

Chapter 2 of Fackrell (2003) discusses the non-uniqueness of representations for the general phase-type models and the uniqueness of minimal representation for the Coxian models. Cumani's theorem implies that the minimal representation of a Coxian model is identifiable. The minimal representation of Example 6.1 has the following 1×1 degenerated transition matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix} -h \end{bmatrix}.$$

Such a representation is identifiable. However, other representations of the form (6.3.1) are non-identifiable.

Since the proposed PTAM is a type of a Coxian model, its minimal representation is unique and identifiable. When $h_1 \neq h_m$, the minimal representation of the proposed PTAM is $(\boldsymbol{\alpha}, \boldsymbol{\Lambda})$, where $\boldsymbol{\alpha}$ is a $1 \times m$ row vector with the first element equal to 1 and the others are equal to 0, and $\boldsymbol{\Lambda}$ is a $m \times m$ matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix} -(\lambda + h_1) & \lambda & & & \\ & -(\lambda + h_2) & \lambda & & \\ & & \ddots & & \\ & & & -(\lambda + h_{m-1}) & \lambda \\ & & & & -h_m \end{bmatrix},$$

and h_i follows (3.3.3). The minimal representation has such a form because the absorption rates in different states are different. Thus, the dimension of the intensity matrix in the minimal representation cannot be reduced further. As a result, both the absorption rates and the transition rates for the proposed PTAM are identifiable. The parameters for the proposed PTAM are identifiable as well. However, one trivial case occurs when $h_1 = h_m = h$, yielding $h_i = h$, and the degenerated transition matrix $\mathbf{\Lambda}$ is a special case of (6.3.1) with $\lambda_i = \lambda$. This example illustrates the concept of non-identifiablility as per our previous analysis.

From another perspective, the resulting distribution is a mixture of exponential distributions with different rates by (2.3.11). The rate parameters uniquely determine the resulting distribution. Thus, insofar as the rates are different, the rate parameters should be identifiable, which is consistent with the theorem of Cumani (1982). Two proposed PTAMs, except for the trivial example having all absorption rates equal, with different total numbers of states cannot generate the same pdf in the domain. This is trivial when considering the resulting pdf as a mixture of exponential distributions, in which the number of components is equal to the total number of states and the rates for different components are distinct.

In summary, the proposed PTAM is identifiable. Therefore, the true parameter values can be inferred from the lifetime data. As per our investigations, however, it is typical to achieve similar likelihoods for different values of m for the same set of lifetime data. When involving parameter estimation, a desired requirement is to introduce model estimability.

6.4 Estimability of the proposed PTAM

The identifiability of the proposed PTAM provides a solid background for accurate model inference in theory. However, there are challenges when calibrating the PTAM. Given a set of lifetime data, it is typical to obtain similar likelihoods for different values of m. In order to facilitate the interpretation of the calibrated results, only one set of estimated values will be used in the final calibrated model. Therefore, it is important to answer the question: *which set of estimated values is the most suitable for the final calibrated model?*

Let us start with the parameter estimation of the phase-type models. The parameter estimation is challenging, especially when the total number of states is unknown and it has to be estimated from the data. To overcome the estimation challenges, many studies have defined restrictions on the Coxian representation. Bobbio and Cumani (1992) suggested maximising the log-likelihood by solving an iterative linearisation method of estimation. Asmussen et al. (1996) proposed a fitting procedure based on the Expectation-Maximisation (EM) algorithm. Faddy (1994) and Faddy (1998) utilised the optimisation algorithm proposed by Nelder and Mead (1965) to maximise the log-likelihood function. The problem of non-convergence of the algorithm by penalised likelihoods was remedied in Faddy (2002).

It is typical for Coxian models to achieve a flat log-likelihood surface around the maximum. This is troublesome when searching for the MLE numerically, and it is a symptom of poor model estimability, which we will address after defining model estimability. Notably, the proposed PTAM has a flat log-likelihood, and the result from the simulation study in Subsection 3.4.3 is used to demonstrate this particular issue of flatness. Consider the estimated results for other parameters when fixing the value of m in Table 6.1. The log-likelihood differences for various values of m are relatively small.

Table 6.1: Estimation results using different m's based on 5,000 lifetimes simulated from the Le Bras's limiting distribution. The first column gives the negative log-likelihood NLL. The last column is the limit of the resulting hazard function $h(t)$ as $t \to \infty$.

NLL	h_1	h_m	$\lambda = \frac{m}{\psi}$	s	m	$\min(\lambda + h_1, h_m)$
21631.884	0.00081	2.0033	1.7764	-0.1237	200	1.7772
21631.826	0.00080	1.8392	1.8652	-0.1183	210	1.8392
21631.806	0.00080	1.7079	1.9540	-0.1134	220	1.7079
21631.713	0.00080	1.6535	1.9991	-0.1112	225	1.6535
21631.813	0.00079	1.6006	2.0428	-0.1089	230	1.6006
21631.843	0.00078	1.5108	2.1316	-0.1047	240	1.5108
21631.889	0.00078	1.4354	2.2205	-0.1009	250	1.4354

The goodness of fit for some values of m is displayed in Figure 6.1. Graphically, all fitted PTAMs approximate the Le Bras model well for most ages, except for the extremely old ages above 100. The survival probabilities to such extremely old ages are so small that only a few extremely old lifetimes can be observed. There could be hardship in validating the goodness of fit at the tail due to the lack of data. This validation may require an unreasonably large amount of data, which is unrealistic to collect in practice. Since the PTAM's calibrated results under different fixed m's are close, it is challenging to estimate m using lifetime data only. This is consistent with our beliefs that m controls the variability of physiological age at any chronological age, and the lifetime data provide little information about the variability of physiological age. Our experiments show that the PTAM's estimability is relatively poor when m is one of the parameters, whilst the estimability of the other parameters is improved significantly when m is fixed.

The main idea of estimability is whether the MLE is unique. If the MLE is not unique, it is impossible to distinguish those estimated results from each other when comparing their resulting distributions.

Definition 6.2. *Given a set of data* $\mathbf{y} = (y_1, ..., y_n)$. *If* $N(\theta) = \{\theta \in \Theta : l(\theta, \mathbf{y}) = \max l(\theta, \mathbf{y})\}$ *is a single set, where* $l(.)$ *is the log-likelihood, then we say* θ *is estimable based on* $\mathbf{y}$. *Or, we simply call the model estimable.*

According to the definition, the estimability of the model parameter relies on both the model identifiability and the data quality. The parameter is estimable only when both the model is identifiable and the data quality is good enough to generate a distinguishable log-likelihood peak around the MLE. However, an identifiable model is not necessary to be estimable for any set of data. A counterexample is

Example 6.2. Suppose the random variable X follows a PTAM with a 2×2 intensity matrix

$$\mathbf{\Lambda} = \begin{bmatrix} -(\lambda + h_1) & \lambda \\ & -h_2 \end{bmatrix},$$

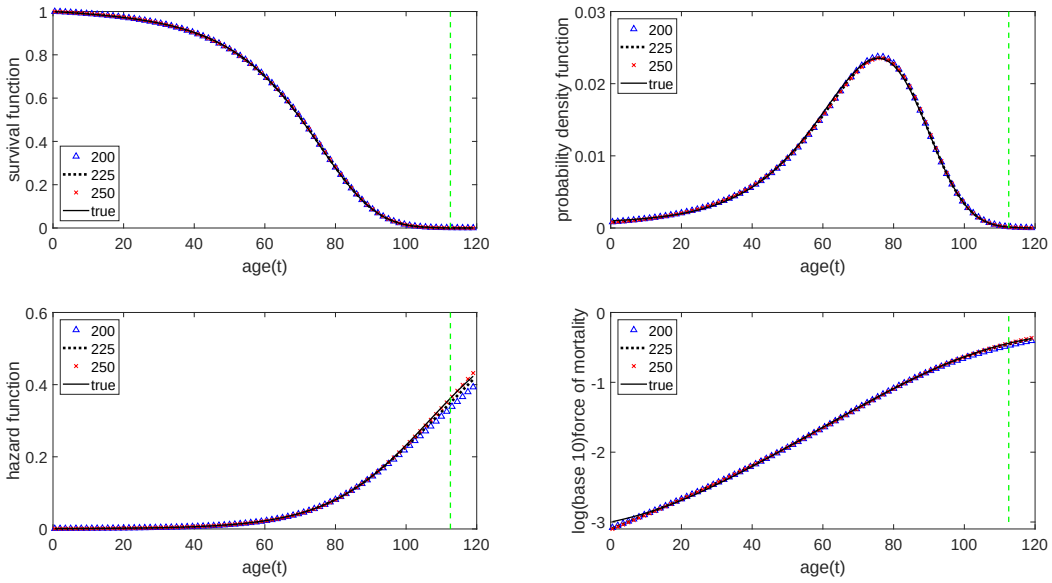

Figure 6.1: Fitted survival function $S(t)$, pdf $f(t)$, hazard function $h(t)$, and log (base 10) hazard function. Each graph includes four curves corresponding to the fitted model with m=200, 225 and 250, as well as the true model. The dotted vertical line indicates the location of $\psi = 112.55$.

where $h_1 \neq h_2$. From Theorem 6.1, the PTAM is identifiable. Suppose only one data $x = \epsilon$, where ϵ is a small value close to 0, is observed. Then the likelihood is

$$L(h_1, h_m, \lambda) = (1, 0)e^{\mathbf{\Lambda}\epsilon}\boldsymbol{h} \approx (1, 0)(\mathbf{I} + \mathbf{\Lambda}\epsilon)\boldsymbol{h} = h_1 - ((\lambda + h_1)h_1 - \lambda h_2)\epsilon,$$

where the approximation is due to the Taylor series $e^{\mathbf{\Lambda}\epsilon} = \mathbf{I} + \mathbf{\Lambda}\epsilon + o(\mathbf{\Lambda}\epsilon)$. The values of λ and h_2 have little impact on the likelihood value. Hence, the likelihood around the neighbourhood of MLE is flat and it is problematic to maximize the likelihood.

△

If more observations can be collected in Example 6.2, we have more information about the process. Thus, it is easier for the numerical optimisation process to find an unique MLE. This maximisation challenge in the numerical process is due to poor data quality. For instance, there is only one observation in Example 6.2. This observation has the information about the first state, but tiny information about the second state. Therefore, the likelihood remains the same level when change the parameter value in the second state.

Recall that the observed lifetimes are the sum of sojourn times spent in each state before absorption. If the complete information for each individual (the sojourn time spent in each state before absorption) is observed, the estimability of the parameters in the PTAM improves significantly. The improvement is attributed to the capacity of being able to estimate the rate parameters in each state through the sample sojourn times; meanwhile, the total number of states m can be estimated by the maximum observed state. However, complete information is unattainable in reality and it is typical to estimate the parameters with lifetime data only. Our experiments show nonetheless that the estimability of the proposed PTAM is relatively poor if

only the lifetime data are relied upon, especially when m needs to be estimated. It is typical to achieve a fairly flat log-likelihood surface for some values of m. On the other hand, if m is fixed, the estimates of other parameters are more stable and are easier to find, which indicates that model estimability could be brought about by fixing m.

We monitor closely the marginal log-likelihood in the neighbourhood of the MLE for the Le Bras model simulation study in Figure 6.2. That is, each marginal log-likelihood is calculated by changing one parameter value near the MLE and fixing the others at the MLE. As we could see from Figure 6.2, the marginal log-likelihood has more curvature when h_1, h_m, s, and λ are changed. The marginal log-likelihood is relatively flat though when changing ψ and m. Furthermore, when fixing both m and ψ, the other parameters are highly correlated, which is shown by three contour plots in Figure 6.3. There is a narrow ridge in each graph, which means the log-likelihood can remain the same even when other parameter values are changed with one parameter value having a drastic change. The correlation between h_m and s is relatively stronger, compared with the other two correlations. This is because a smaller s can achieve a flatter h_i pattern in the early states when h_m is overstated (see Figure 3.4); this somehow mitigates the biased estimate of h_m.

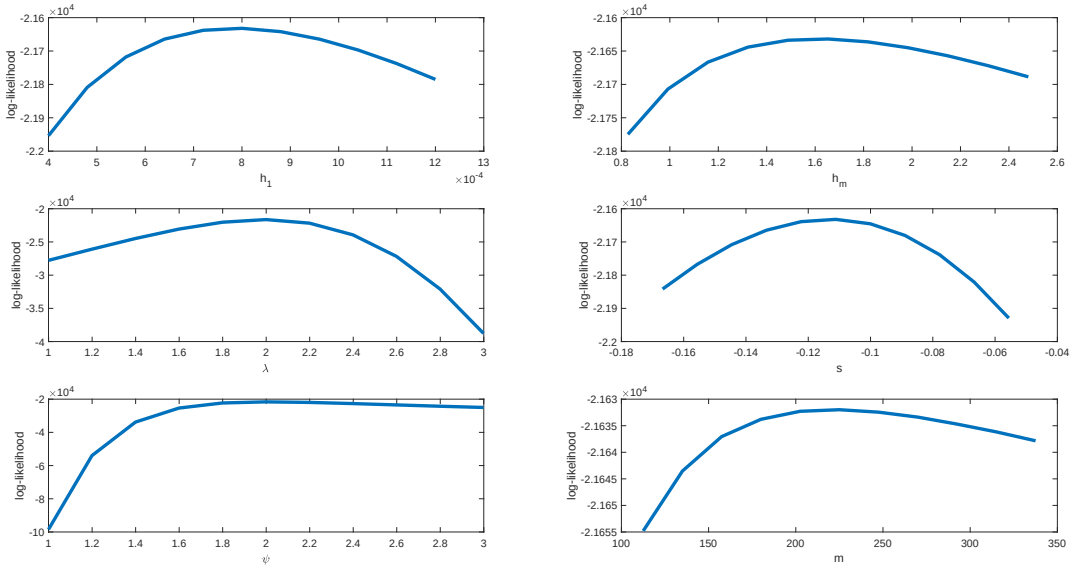

Figure 6.2: Left-Up: Log-likelihood function with h_1 changing and other parameters fixed around the MLE; Right-Up: Lhe log-likelihood function with h_m changing and other parameters fixed around the MLE; Left-Middle: Log-likelihood function with λ changing and other parameters fixed around the MLE; Right-Middle: Log-likelihood function with s changing and other parameters fixed around the MLE; Left-Down: Log-likelihood function with ψ changing and other parameters fixed around the MLE; Right-Down: Log-likelihood function with m changing and other parameters fixed around the MLE.

Example 6.3. To illustrate the strong pairwise correlation amongst h_1, h_m and s, we assign $m = 225$ and $\lambda = 1.99908$. After fixing m and λ, we fix one of the parameters in the set $\{h_1, h_m,$

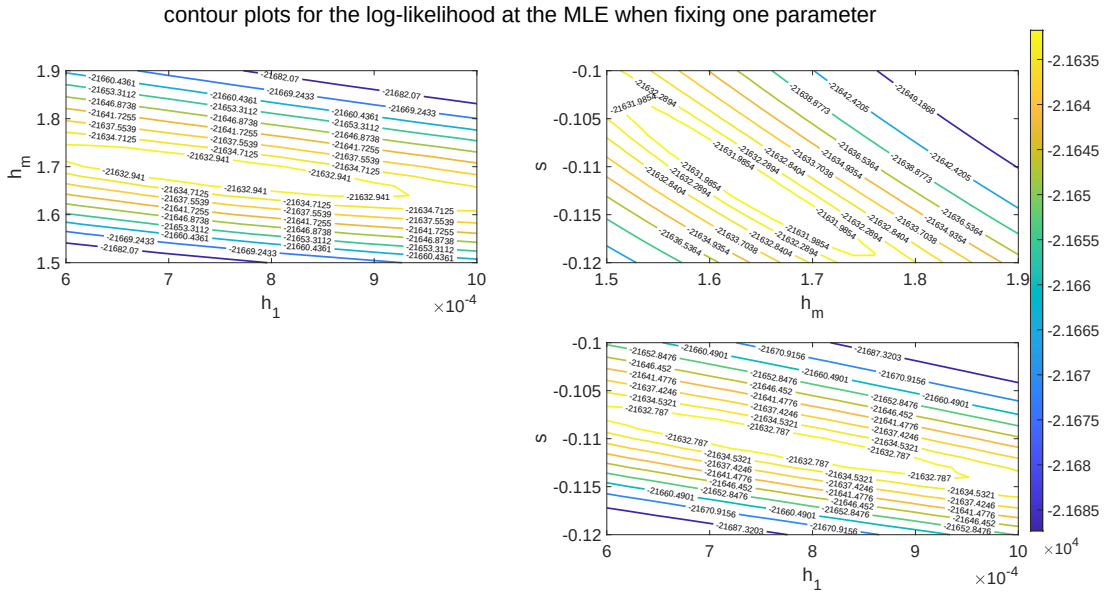

Figure 6.3: Contour plots of the log-likelihood at the MLE for the Le Bras simulation study with one parameter fixed.

$s\}$, and estimate the other two parameters by the MLE method. The fixed values are:

- $h_1 = 10^{-5}$, resulting to relatively low dying rates in the early states.
- $h_m = 100$, resulting to extremely high dying rates in the last few states.
- $s = 0$, resulting to an exponentially increasing dying rate with respect to physiological age.

  The estimated h_i's under each scenario are plotted in Figure 6.4. It is clear that the estimated dying rates under each scenario are almost identical before state 100; they are slightly distinguishable from state 100 to state 150; and they are significantly distinguishable after state 150. Since only few individuals can survive to extremely high physiological ages (states), the maximised likelihoods under each scenario are similar to each other.

△

When calibrating the PTAM with lifetime data only, the estimate of h_1 is determined by the mortality rates at extremely young ages, and the estimate of h_m is determined by the mortality rates at extremely old ages. The logic behind this is intuitive when checking the physiological age distributions at young ages and the physiological age distributions at old ages. Recall another fact that the impact on the resulting lifetime distribution due to the biased estimate of h_m can be mitigated by a biased estimate of h_1. Therefore, as far as the sample size is large enough that there are enough sample individuals surviving to extremely old ages, the estimates of h_1 and h_m should be independent and accurate. However, it is extremely rare that there are enough death time samples at extremely old ages, which causes huge uncertainty on the

estimate of h_m. The strong correlations between h_1 and h_m, between h_1 and s, and between h_m and s make the PTAM calibration quite a hurdle, especially when dealing with model inference.

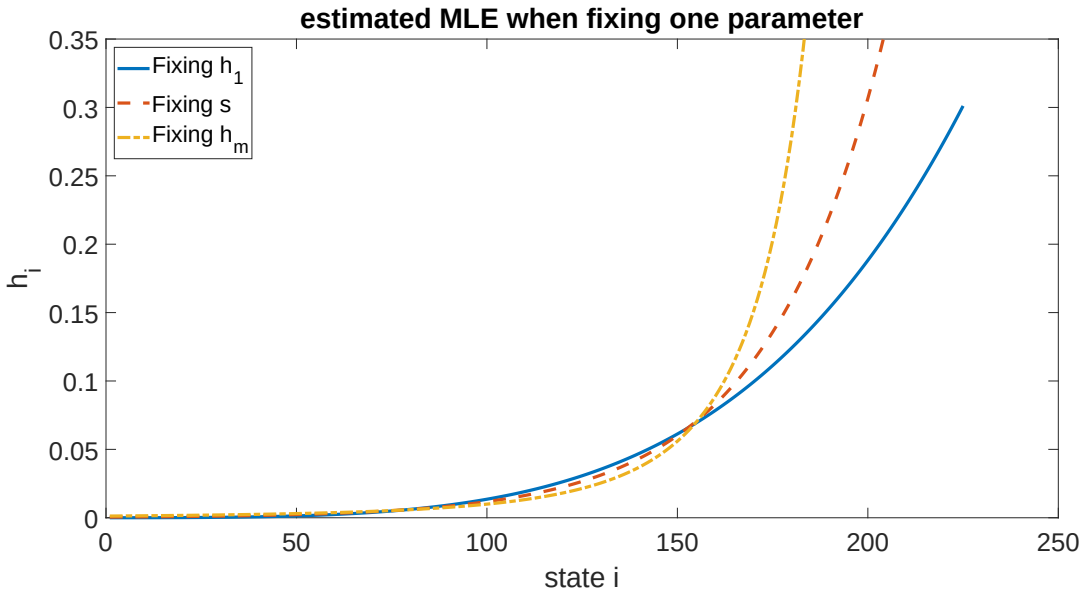

Figure 6.4: Estimated h_i for the Le Bras simulation study with one parameter fixed.

6.4.1 Assessing model estimability

As mentioned before, the model estimability relies on both model identifiability and data quality. For a given set of lifetimes, different parameter values for the proposed PTAM may generate similar log-likelihoods. It is essential to test whether the log-likelihood difference is due to computation errors or the nature of poor estimability. Our objective is to be able to assess model estimability.

Some tools in the exploration of model estimability could be grouped into five categories:

- Interval estimation or confidence interval for the parameter;
- Hessian matrix evaluated at the MLE;
- Constructed testing function;
- Sensitivity testing on priors;
- Data cloning.

Interval estimation

Most papers (e.g., Titman and Sharples (2010)), with the use of a data set, assess model estimability by checking the width of the confidence interval for each parameter. When the confidence interval is "wide", the estimates' variance is large, and there is more uncertainty about the actual parameter values. Therefore, the model estimability is poor when the confidence interval

is "wide". One advantage of this method is that the concept is easy to understand, and another advantage is its power to detect those models that are non-estimable. Note that this method is qualitative in terms of judging "wide" confidence interval. For those cases with not so wide confidence intervals, model estimability may still be poor.

Hessian matrix at the MLE

When numerically searching for the MLE, it is easy to evaluate the Hessian matrix of the log-likelihood at the optimum. Moriguchi and Murota (2012) reported the characteristics of the Hessian matrix, one of which is that the curvature of log-likelihood around the MLE can be quantified by the Hessian matrix evaluated at the MLE. If the evaluated Hessian matrix is negative-definite, i.e., eigenvalues are all negative, then the optimum is a local maximum. Moreover, the bigger the minimum of the absolute value of eigenvalues for the evaluated Hessian matrix is, the more curvature the log-likelihood surface exhibits, the better the model estimability. Unfortunately, when estimating the parameters for the proposed PTAM, it is typical to find that the minimum eigenvalue, in absolute value terms, of the Hessian matrix evaluated at the estimated point is relatively very small; causing poor estimability in the proposed PTAM. The Hessian matrix evaluated at the MLE is equal to the negative of the Fisher information. The observed Fisher information can be numerically obtained as robust numerical methods to calculate the Hessian matrix are available in most software such as MATLAB and R.

Testing function

The third tool to assess estimability is to construct a testing function by utilising the alternating conditional expectation (Hengl et al., 2007). The testing function is the empirical variance of the average ranked transformation, which is designed to detect models with poor estimability. However, the testing function is not powerful enough to detect estimable models.

Sensitivity testing on priors

When using Bayesian approaches, the model parameter is estimable for data set if the posterior is not sensitive to the choice of the prior. Eberly and Carlin (2000) selected three different priors for a non-estimable parameter in a hierarchical Gaussian model, resulting to 3 different posteriors. Similarly, Mu (2019) chose four different priors for a non-estimable parameter in their binomial model, resulting to four distinguishable posteriors. Both examples demonstrated the assessment of estimability from the Bayesian perspective. The embedded mechanism is intuitive. If the model parameter is estimable for a given data set, the data has enough information for learning the parameter value, and the posterior should not be affected too much by the priors, and vice versa. It is inevitable to use the Markov Chain Monte Carlo (MCMC) technique to obtain the posterior for the proposed PTAM. To test whether the posterior is affected by the choice of the prior, we need to select a variety of priors and run the MCMC for each prior, which is a time-consuming process. There is no quantitative way to test whether the posterior is sensitive to the choice of prior. Currently, this approach is mainly based on the modeler's graphical judgement of the posteriors.

Data cloning

Another method to assess estimability using the Bayesian approaches is called data cloning. The basic idea of data cloning is to 'exaggerate' the log-likelihood's curvature by repeating

the data multiple times. Lele et al. (2010) utilised data cloning to explore the estimability of parameters in some hierarchical models, logistic-normal model, and mixed binary-regression model. They proved that the variance of the posterior for the estimable parameter converges to 0 with increasing clones, whilst the variance of the posterior for the non-estimable parameter does not converge to 0 with increasing clones. Some numerical examples to demonstrate the variance of the posterior for both estimable and non-estimable parameters were provided in their paper. In the context of data cloning, Campbell and Lele (2014) assumed a linear model for the posterior mean with k clones and the pth prior. They performed ANOVA testing on two hypotheses: (i) the point estimates are not significantly different when changing the number of clones, and (ii) the point estimates are not significantly different when changing priors. Their conclusion is that the parameter is estimable if the result passes both ANOVA tests. In general, the process of data cloning can be summarised in 3 steps:

- Step 1: Clone the data multiple times and treat them as observations;
- Step 2: Obtain the posterior based on the observations;
- Step 3: Check if the variance of the posterior converges to 0 with the increasing number of clones.

Theorem 6.2. *Suppose K is the number of clones, n is the number of observations, and $N(\theta) = \{\theta \in \Theta : l(\theta, \boldsymbol{y}) = \max l(\theta, \boldsymbol{y})\}$ is a single-point set, that is, the likelihood function is identical over the set $N(\theta)$. As $K \to \infty$, the posterior distribution converges to a distribution with density $\frac{\pi(\theta)}{\int_{N(\theta)} \pi(\theta) d\theta}$ for $\theta \in N(\theta)$, where $\pi(\theta)$ is the prior distribution of θ. If the set $N(\theta)$ is not a single-point set, $\sigma^2_{K,n}$, the largest eigenvalue of the posterior variance matrix, does not converge to 0.*

Proof. The proof is the same as that in Lele et al. (2010). However, some details are supplemented to clarify certain steps in the original proof.

Consider the ratio of the posterior distributions with two sets of parameter values:

$$\frac{\pi_K(\theta|\boldsymbol{y})}{\pi_K(\theta_{(n)}|\boldsymbol{y})} = \frac{\pi(\theta)}{\pi(\theta_{(n)})} \frac{f^K(\boldsymbol{y}|\theta)}{f^K(\boldsymbol{y}|\theta_{(n)})},$$

where $\theta_{(n)}$ is a parameter value that maximises the log-likelihood $l(\theta; \boldsymbol{y})$. For $\theta \notin N(\theta)$,

$$\frac{\pi_K(\theta|\boldsymbol{y})}{\pi_K(\theta_{(n)}|\boldsymbol{y})} = \frac{\pi(\theta)}{\pi(\theta_{(n)})} \left(\frac{f(\boldsymbol{y}|\theta)}{f(\boldsymbol{y}|\theta_{(n)})} \right)^K \to 0,$$

as $K \to \infty$ because $\frac{f(\boldsymbol{y}|\theta)}{f(\boldsymbol{y}|\theta_{(n)})} < 1$. For $\theta_1, \theta_2 \in N(\theta)$, $\frac{\pi_K(\theta_1|\boldsymbol{y})}{\pi_K(\theta_2|\boldsymbol{y})} = \frac{\pi(\theta_1)}{\pi(\theta_2)} \frac{f^K(\boldsymbol{y}|\theta_1)}{f^K(\boldsymbol{y}|\theta_2)} = \frac{\pi(\theta_1)}{\pi(\theta_2)}$. Thus, $N(\theta)$ is a single-point set if and only if the variance of the posterior converges to 0 as $K \to \infty$.

■

The sufficient and necessary condition, indicated in the proof of Theorem 6.2, provides a quantitative method to assess model estimability. It is inevitable to use MCMC in obtaining posteriors under a Bayesian approach as the posteriors have generally intricate structures.

Similarly, we can prove that, given lifetime data set, the MLE is unique if and only if the MLE for such a data set with K clones is unique as $K \to \infty$. Hence, data cloning is a powerful method to test estimability from both the Bayesian and likelihood-based approaches.

Theorem 6.3. *Consider a lifetime data* $\boldsymbol{y} = (y_1, ..., y_n)$ *and any calculation tolerance* ξ *and suppose* $N^K(\theta) = \{\theta \in \Theta : K|l(\theta, \boldsymbol{y}) - \max l(\theta, \boldsymbol{y})| < \xi\}$*, where* $l(.)$ *is the log-likelihood. The parameter* θ *is estimable based on lifetime data* $\boldsymbol{y}$ *if and only if* $N^K(\theta)$ *is a single-point set as* $K \to \infty$*.*

Proof. When $N(\theta)$ is a single-point set, $N^1(\theta)$ may contain more than one point, which means there are at least two different sets of parameter values such that the log-likelihoods are numerically identical. Since $N(\theta)$ is a single-point set, there is a unique θ_0 such that $l(\theta_0, \boldsymbol{y}) = \max l(\theta, \boldsymbol{y})$. Therefore, for any $\xi > 0$ and any $\theta \neq \theta_0 \in N^1(\theta)$ there is a K_0 such that for any $K > K_0$,

$$K|l(\theta, \boldsymbol{y}) - l(\theta_0, \boldsymbol{y})| > \xi,$$

which indicates $\theta \notin N^K(\theta)$. As $K \to \infty$, we have $N^K(\theta) = \{\theta_0\}$.

On the other hand, suppose $N^K(\theta)$ is a single-point set as $K \to \infty$ and there are $\theta_1, \theta_2 \in \Theta$ such that $l(\theta_1, \boldsymbol{y}) = l(\theta_2, \boldsymbol{y}) = \max l(\theta, \boldsymbol{y})$. By definition, $\theta_1, \theta_2 \in N^K(\theta)$ for $K = 1, 2, ...$ Hence, $\theta_1, \theta_2 \in N^K(\theta)$ as $K \to \infty$. Since $N^K(\theta)$ is a single-point set as $K \to \infty$, we have $\theta_1 = \theta_2$, which implies $N(\theta)$ is a single-point set.

■

Compared with other methods, data cloning is the most suitable method to explore estimability for two reasons:

- It is the only method that provides a sufficient and necessary condition to distinguish estimable models from non-estimable models;
- It is a quantitative method providing a robust estimability procedure.

We shall therefore use data cloning in the pursuit of probing PTAM's estimability. In the next Section PTAM's estimability will be investigated and the results will be compared under different scenarios. This investigation is an assessment of the quality of model estimability in a relative sense rather than in an absolute sense. In particular, a model has better estimability compared with another model if the posterior of the model has a smaller variance than that of the other.

6.4.2 PTAM estimability under some scenarios

Govorun et al. (2018) attempted to incorporate health-related information when calibrating their phase-type model. They utilised health information to refine the physiological age distribution at any chronological age by assuming a linear relationship between the physiological age and medical cost. By doing so, their calibrated results showed that the physiological age variability at any chronological age was reduced with the incorporation of medical-cost information. Thus, health-related information of a person apparently provides some information on

the person's physiological age. Refining the physiological age distribution at any chronological age could mitigate the estimability issue. Inclusion of health-related information, in addition to lifetime data, is a possible way to improve the poor estimability of the proposed PTAM. However, it is not easy to collect this information, and therefore, we shall continue delving in this investigation by simulations.

The PTAM to be simulated has the following true parameter values: $h_1 = 0.025$, $h_m = 0.515$, $s = 1$, $m = 50$, $\psi = 31.25$; and so, $\lambda = 50/31.25 = 1.6$. The corresponding pdf, survival function, hazard rate, and dying rate are exhibited in Figure 6.5.

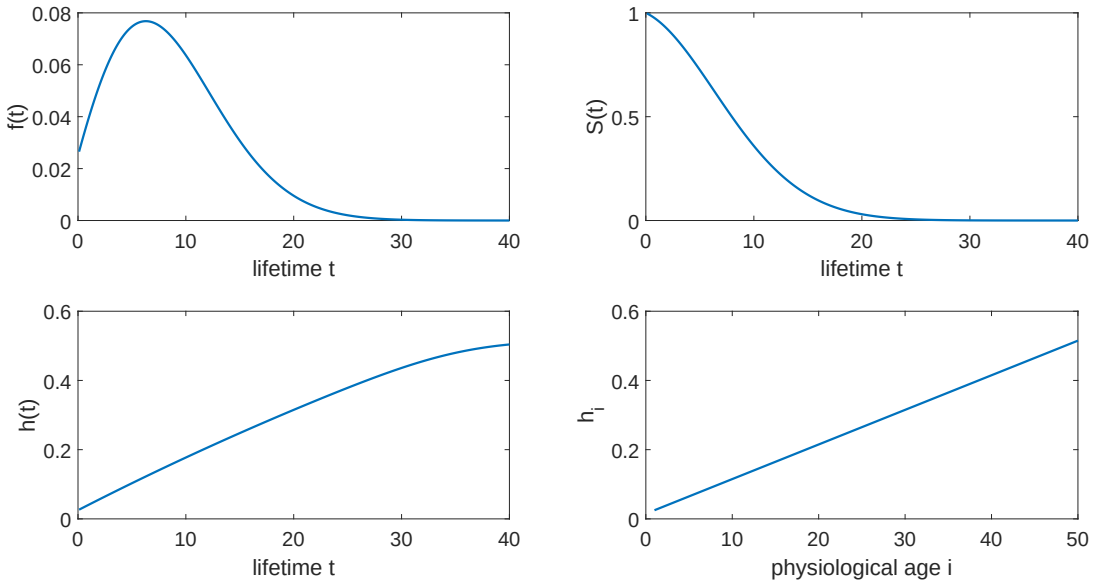

Figure 6.5: Some plots related to the proposed PTAM with paramater values $h_1 = 0.025$, $h_m = 0.515$, $s = 1$, $m = 50$, $\psi = 31.25$. Top-left: pdf; Top-right: survival function; Bottom-left: hazard rate; Bottom-right: dying rate at physiological age.

One thousand phase-type ageing processes are simulated from the true model under the following scenarios.

- Complete information – the complete path is observable;
- Partial information – the physiological age can be observed every k years;
- Partial information with noise – the physiological age can be measured every k years with known measurement error;
- Partial information with unknown noise – the physiological age can be measured every k years with unknown measurement error;
- No information – the physiological age is unobservable and only death time is observable.

Complete information

Suppose we observe the complete paths for n individuals. For the ith individual, let $t_{i,j}$ be the time spent in state j and n_i be the last transient state before absorption. The likelihood for the ith individual is

$$L(\boldsymbol{\theta}; t_{i,\bullet}) = \prod_{j=1}^{n_i-1} \lambda_j e^{-(\lambda_j+h_j)t_{i,j}} h_{n_i} e^{-(\lambda_{n_i}+h_{n_i})t_{i,n_i}},$$

where $\lambda_i = \lambda$ when $i = 1, ..., m-1$ and $\lambda_m = 0$. The log-likelihood for the ith individual is then

$$l(\boldsymbol{\theta}; t_{i,\bullet}) = \begin{cases} (n_i - 1)\log\lambda + \log h_{n_i} - \sum_{j=1}^{n_i}(\lambda + h_j)t_{i,j}, \text{ when } n_i \neq m; \\ (m-1)\log\lambda + \log h_m - \sum_{j=1}^{m-1}(\lambda + h_j)t_{i,j} - h_m t_{i,m}, \text{ when } n_i = m. \end{cases}$$

Then, the log-likelihood is

$$l(\boldsymbol{\theta}; \boldsymbol{t}) = \sum_{i=1}^{n} l(\boldsymbol{\theta}; t_{i,\bullet}) \tag{6.4.2}$$

It is necessary for the parameter MLEs to satisfy the system

$$\begin{cases} \frac{\partial l}{\partial \lambda} = \sum_{i=1}^{n} \frac{(n_i-1)}{\lambda} - \sum_{i=1}^{n} \sum_{j=1}^{\min(n_i, m-1)} t_{i,j} = 0 \\ \frac{\partial l}{\partial h_1} = \sum_{i=1}^{n} \frac{1}{h_{n_i}} \frac{\partial h_{n_i}}{\partial h_1} - \sum_{i=1}^{n} \sum_{j=1}^{n_i} t_{i,j} \frac{\partial h_j}{\partial h_1} = 0 \\ \frac{\partial l}{\partial h_m} = \sum_{i=1}^{n} \frac{1}{h_{n_i}} \frac{\partial h_{n_i}}{\partial h_m} - \sum_{i=1}^{n} \sum_{j=1}^{n_i} t_{i,j} \frac{\partial h_j}{\partial h_m} = 0 \\ \frac{\partial l}{\partial s} = \sum_{i=1}^{n} \frac{1}{h_{n_i}} \frac{\partial h_{n_i}}{\partial s} - \sum_{i=1}^{n} \sum_{j=1}^{n_i} t_{i,j} \frac{\partial h_j}{\partial s} = 0. \end{cases}$$

Example 6.4. We sample 100 sets of data from the true model. Each set of data has 1,000 complete paths. The MLE, using each data set, can be numerically obtained through (6.4.2). The summary statistics of one hundred MLEs via the mean, standard deviation, 5% quantile and 95% quantile for each parameter are presented in Table 6.2.

Table 6.2: Some statistics for 100 MLEs under the complete information case.

	mean	standard deviation	Q(0.05)	Q(0.95)	true value
h_1	0.0248	0.0048	0.0175	0.0322	0.025
h_m	0.5012	0.0488	0.4319	0.5758	0.515
ψ	30.1094	1.4235	27.395	31.5449	31.25
s	0.9961	0.1288	0.78021	1.2044	1.00
m	48	2	44	50	50

Remark 6.3. *For Tables 6.2 – 6.7. Q(0.05) and Q(0.95) are the 5th and 95th quantiles (or more specifically, percentiles), respectively. The mean and standard deviation for m are rounded off to the nearest integer.*

The one hundred MLEs shown in Figure 6.6 can be used to approximate the distribution of the MLE with 1,000 individuals or more specifically, ageing processes for the complete information case. The solid red line in each subgraph is the true parameter value, and the left green dotted line is the 5th percentile of the empirical MLEs, which is the fifth smallest MLE. The green dotted line to the right is the 95th percentile of the empirical MLEs, which is the fifth largest MLE. All solid red lines are between the two green dotted lines, except that the red line overlaps with the right green line in the distribution of estimated m.

From 6.6, the 90% confidence interval for each parameter can capture the true value, and the associated standard deviation is relatively small. The empirical distribution of each estimate is bell-shaped, except for the estimates of m and ψ. This evidence shows that the log-likelihood is symmetric when complete information is observed. Furthermore, 50 out of the 100 estimated m's are equal to the true value 50, and the smallest estimated m is 42. Recall that for each sample the size is 1,000, which is relatively small when fitting PTAMs. However, the estimates are quite promising, especially the estimate of m.

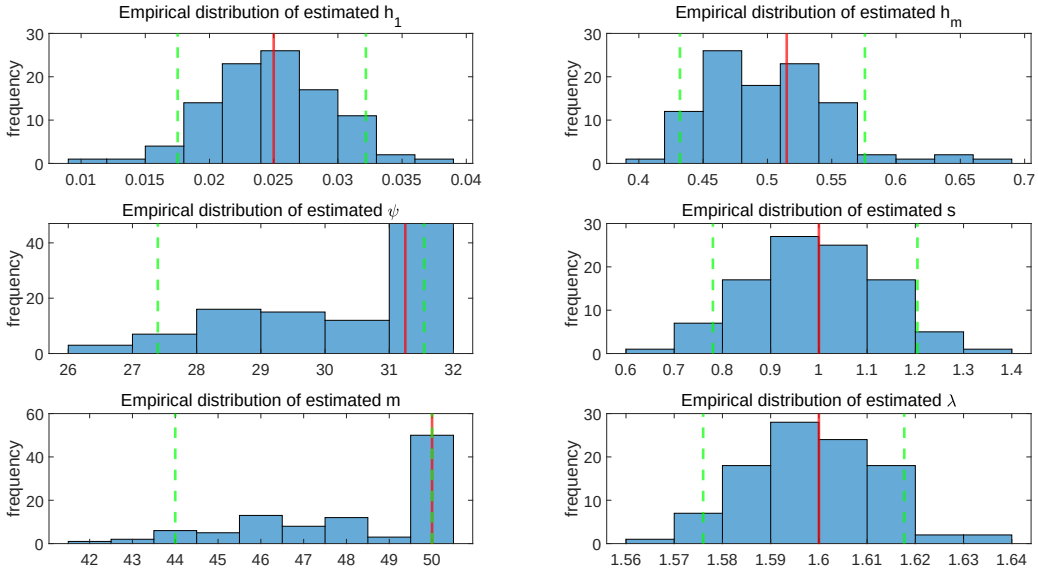

Figure 6.6: One hundred MLEs for the simulation study under the scenario of complete information.

△

Remark 6.4. *In Figures 6.6 – 6.11, the red solid lines are the true values, the green dotted lines to the left are the* 5*th percentiles, and green dotted lines to the right are the* 95*th percentiles.*

Partial information

Suppose we observe the states every k years for n individuals. For the ith individual, the states are observed at time $\boldsymbol{t}_{i,\bullet} = 0, k, 2k, ..., n_i k, n_i k + \Delta t_i$ and the corresponding observed states are $\boldsymbol{y}_{i,\bullet} = y_{i,1}, y_{i,2}, ..., y_{i,n_i+1}, y_{i,n_i+2}$. Particularly, $y_{i,1} = 1$ and $y_{i,n_i+2} = m + 1$. The likelihood for the

ith individual is

$$L(\boldsymbol{\theta}; \boldsymbol{t}_{i,\bullet}, \boldsymbol{y}_{i,\bullet}) = \prod_{j=1}^{n_i} \boldsymbol{P}_{y_{i,j},y_{i,j+1}} \boldsymbol{\alpha}_{\boldsymbol{n_i+1}} e^{\boldsymbol{\Lambda}\Delta t_i} \boldsymbol{h},$$

where $\boldsymbol{P}_{i,j}$ is the (i, j) element of the $m \times m$ matrix $e^{\boldsymbol{\Lambda}k}$. The dimension of the vector $\boldsymbol{\alpha}_{\boldsymbol{n_i+1}}$ is $1 \times m$, where the $(n_i + 1)st$ element is 1 and 0 elsewhere. The log-likelihood is

$$l(\boldsymbol{\theta}; \boldsymbol{t}_{i,\bullet}, \boldsymbol{y}_{i,\bullet}) = \sum_{j=1}^{n_i} \log \boldsymbol{P}_{y_{i,j},y_{i,j+1}} + \log\left(\boldsymbol{\alpha}_{\boldsymbol{n_i+1}} e^{\boldsymbol{\Lambda}\Delta t_i} \boldsymbol{h}\right)$$

The sample log-likelihood is

$$l(\boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log \boldsymbol{P}_{y_{i,j},y_{i,j+1}} + \sum_{i=1}^{n} \log\left(\boldsymbol{\alpha}_{\boldsymbol{n_i+1}} e^{\boldsymbol{\Lambda}\Delta t_i} \boldsymbol{h}\right). \tag{6.4.3}$$

Example 6.5. We sample 100 data sets from the true PTAM. Each data set has 1,000 individual lifetimes, whose ageing states can be observed every 3 years ($k = 3$). The MLE for each data set can be numerically obtained using (6.4.3). The summary statistics for the one hundred MLEs are displayed in Table 6.3.

Recall that the maximum lifetime for the true model is around 30. Hence, the number of physiological ages that can be observed for the individuals who survive to the maximum lifetime in this population is around 10. If we linearly scale the lifetime variable in this population to the human lifetime whose maximum age is 120, the state is scaled to be observed every 12 years, which is a reasonable time for insurance companies to update the physiological age for each individual.

Table 6.3: Some statistics for 100 MLEs under the partial information case.

	mean	standard deviation	Q(0.05)	Q(0.95)	true value
h_1	0.0278	0.0059	0.0182	0.0375	0.025
h_m	0.4730	0.0494	0.3916	0.5432	0.515
ψ	32.8910	0.6500	32.0755	34.214	31.25
s	1.2166	0.16250	0.9748	1.5085	1
m	48	1	48	50	50

In Figure 6.7, one hundred MLEs are used to approximate the distribution of the MLEs. Each MLE emanates from 1,000 data points for the partial-information case. All solid red lines are between two green dotted lines, except for ψ and λ. Recall the fact that $\lambda = m/\psi$. Therefore, the estimate of λ will be biased when either the estimate of m or the estimate of ψ is biased. Compared with Figure 6.6, the estimability of ψ is relatively weaker.

△

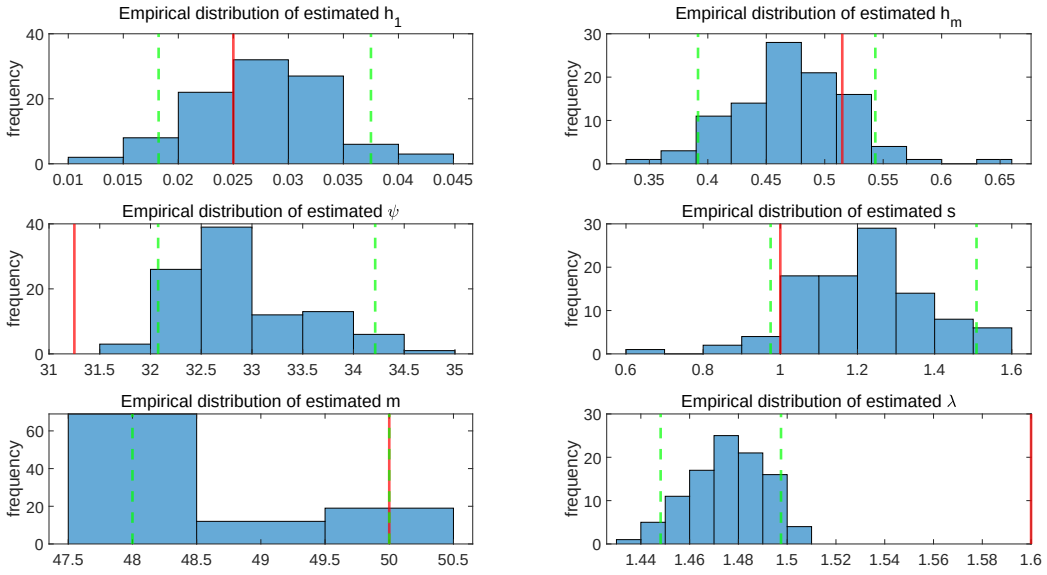

Figure 6.7: Histogram of one hundred MLEs for the simulation stud under the scenario of partial information.

Partial information with noise

In practice, the physiological age (state) is unobservable at any chronological age. However, we can estimate the physiological age from some health-related information. Using this approach to estimate the physiological age for each individual incurs measurement errors. Suppose we observe the physiological age every k years with measurement errors for n individuals. For the ith individual, the states are observed at time $\boldsymbol{t_{i,\bullet}} = (0, k, 2k, ..., n_i k, n_i k + \Delta t_i)$, and the corresponding observed values are $\boldsymbol{c_{i,\bullet}} = (c_{i,1}, c_{i,2}, ..., c_{i,n_i+1}, c_{i,n_i+2})$. Let $\boldsymbol{y_{i,\bullet}} = (y_{i,1}, y_{i,2}, ..., y_{i,n_i+1}, y_{i,n_i+2})$ be the corresponding actual states. Furthermore, it is reasonable to assume that

$$C_{i,j} = Y_{i,j} + \epsilon_{i,j},$$

where $C_{i,j}$ and $Y_{i,j}$ are the random variables with values $c_{i,j}$ and $y_{i,j}$, respectively. Recall that we cannot measure the true physiological age $Y_{i,j}$, but can only get the state information through a measurement $C_{i,j}$. Then, $\epsilon_{i,j}$ is the measurement error and assumed as

$$\epsilon_{i,j} \sim N\left(0, \sigma_{i,j}^2\right),$$

where $N\left(0, \sigma_{i,j}^2\right)$ represents a Gaussian distribution with mean 0 and standard deviation $\sigma_{i,j}$. Additionally, it is reasonable to assume that the measurement errors are homogeneous so that $\sigma_{i,j} = \sigma$. By such assumptions, the conditional distribution for $C_{i,j}$ given $Y_{i,j} = y_{i,j}$ is

$$C_{i,j}|_{Y_{i,j}=y_{i,j}} \sim N\left(y_{i,j}, \sigma^2\right).$$

This tells us that the probability the individual is in state $y_{i,j}$ with its observed value $c_{i,j}$ is

$$\begin{aligned} P(Y_{i,j} = y_{i,j},\ C_{i,j} = c_{i,j}) &= P(Y_{i,j} = y_{i,j})P(C_{i,j} = c_{i,j}|Y_{i,j} = y_{i,j}) \\ &= P(Y_{i,j} = y_{i,j})\frac{1}{\sigma}\phi\left(\frac{c_{i,j} - y_{i,j}}{\sigma}\right). \end{aligned}$$

It is worth noting that one can extend the relation between $C_{i,j}$ and $Y_{i,j}$ to a function of physiological age (state). For example, Govorun et al. (2018) assumed that $C_{i,j} = aY_{i,j} + b$, where a and b are parameters estimated from health-related data. Additionally, $C_{i,j}$ can be treated as a health-related random variable after making appropriate assumptions on the relation between $C_{i,j}$ and $Y_{i,j}$. Govorun et al. (2018) used the extended health benefits data for a Canadian employee and retiree group. The health-related information is the health costs. That is, $C_{i,j}$'s are taken as health costs. One can even extend $C_{i,j}$ to other health-related data by proposing a suitable function $f(\bullet)$ for $C_{i,j} = f(Y_{i,j})$.

The likelihood for the ith individual is

$$L(\boldsymbol{\theta}; \boldsymbol{t}_{i,\bullet}, \boldsymbol{c}_{i,\bullet}) = \boldsymbol{\alpha}\left(\prod_{j=1}^{n_i} \boldsymbol{P}^* \boldsymbol{C}_{i,j}\right)\boldsymbol{h},$$

where $\boldsymbol{P}^* = e^{\boldsymbol{\Lambda} k}$ is an $m \times m$ matrix and

$$\boldsymbol{C}_{i,j} = \frac{1}{\sigma}\begin{bmatrix} \phi\left(\frac{c_{i,j}-1}{\sigma}\right) & \phi\left(\frac{c_{i,j}-1}{\sigma}\right) & \dots & \phi\left(\frac{c_{i,j}-1}{\sigma}\right) \\ \phi\left(\frac{c_{i,j}-2}{\sigma}\right) & \phi\left(\frac{c_{i,j}-2}{\sigma}\right) & \dots & \phi\left(\frac{c_{i,j}-2}{\sigma}\right) \\ & & \ddots & \\ \phi\left(\frac{c_{i,j}-m}{\sigma}\right) & \phi\left(\frac{c_{i,j}-m}{\sigma}\right) & \dots & \phi\left(\frac{c_{i,j}-m}{\sigma}\right) \end{bmatrix},$$

which is $m \times m$. The $\phi(\bullet)$ is the standard normal pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

The log-likelihood for the ith individual is

$$l(\boldsymbol{\theta}; \boldsymbol{t}_{i,\bullet}, \boldsymbol{c}_{i,\bullet}) = \log\left(\boldsymbol{\alpha}\left\{\prod_{j=1}^{n_i} \boldsymbol{P}^* \boldsymbol{C}_{i,j}\right\}\boldsymbol{h}\right),$$

and the log-likelihood for the sample is

$$l(\boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{c}) = \sum_{i=1}^{n} \log\left(\boldsymbol{\alpha}\left\{\prod_{j=1}^{n_i} \boldsymbol{P}^* \boldsymbol{C}_{i,j}\right\}\boldsymbol{h}\right). \tag{6.4.4}$$

Example 6.6. We sample 100 sets of data from the true PTAM. Each set of data has 1,000 individuals, whose states can be observed, with measurement error, every 3 years. Additionally, the true value for the standard deviation σ is 1. The MLE for each set of data can be numerically obtained using (6.4.4).

Table 6.4: Summary statistics for 100 MLEs under the partial-information case with known measurement error.

	mean	standard deviation	Q(0.05)	Q(0.95)	true value
h_1	0.0180	0.0068	0.0067	0.0283	0.025
h_m	0.4204	0.0424	0.3625	0.5029	0.515
ψ	30.9048	2.4115	26.4233	34.2845	31.25
s	1.3227	0.1621	1.0494	1.6019	1
m	46	4	39	51	50

The investigation starts with the easiest scenario – the true value for σ is known. One hundred MLEs are summarised by the mean, standard error, 5% quantile and 95% quantile for each parameter in Table 6.4. The empirical distribution of 100 MLEs are depicted in Figure 6.8.

Compared with Figure 6.7, the parameter estimability is relatively weaker owing to the fact both the variance for each parameter is larger than that of the partial-information case, and the 90% confidence intervals for h_m and s cannot capture their true values. This result is reasonable because the measurement error contributes additional uncertainty to the parameter estimates.

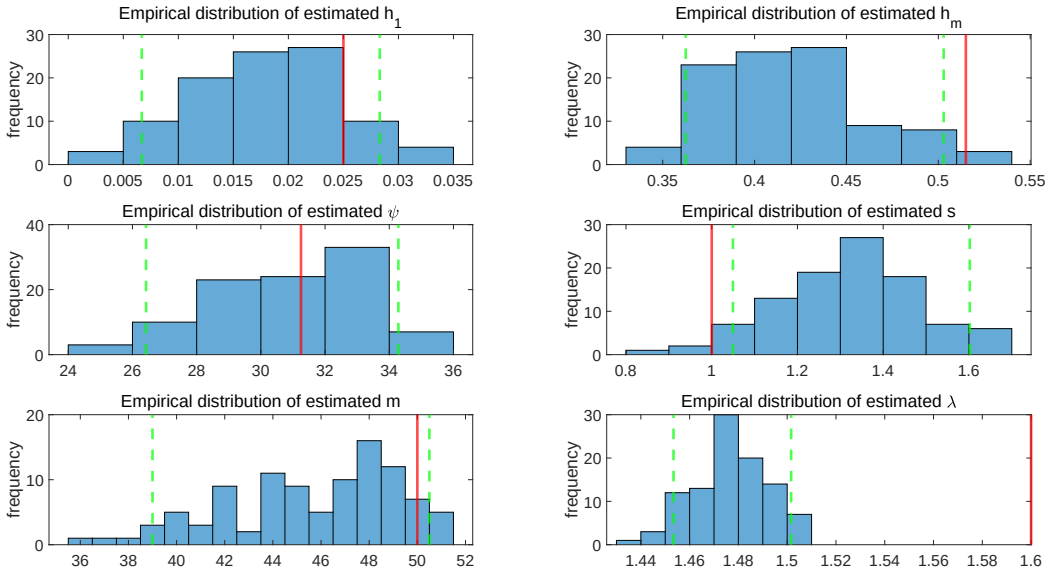

Figure 6.8: Histogram of one hundred MLEs for the simulation study under the scenario of partial information and known measurement error.

The ensuing investigation deals with the scenario with unknown σ but to be estimated from the data. Under this scenario, Table 6.5 contains the summary statistics of one hundred MLEs. Figure 6.8 features the distributions of various MLEs showing that the bias of h_1 and h_m are relatively higher because fewer estimated values are close to their true values. As expected, the unknown standard deviation makes it harder to learn the true parameter values. The standard

error of estimates for h_1 and h_m are marginally smaller than that of the partial-information case with known measurement error. This may be due to the biased estimation.

Table 6.5: Some statistics for the 100 MLEs for the case that states can be observed every 3 years with unknown measurement error.

	mean	standard deviation	Q(0.05)	Q(0.95)	true value
h_1	0.0180	0.0059	0.0089	0.0268	0.025
h_m	0.4198	0.03893	0.3656	0.4835	0.515
ψ	30.6299	2.4358	26.1289	33.73067	31.25
s	1.3113	0.1667	1.0155	1.5738	1.000
m	46	4	39	50	50
σ	1.1147	0.0494	1.0414	1.1978	1.000

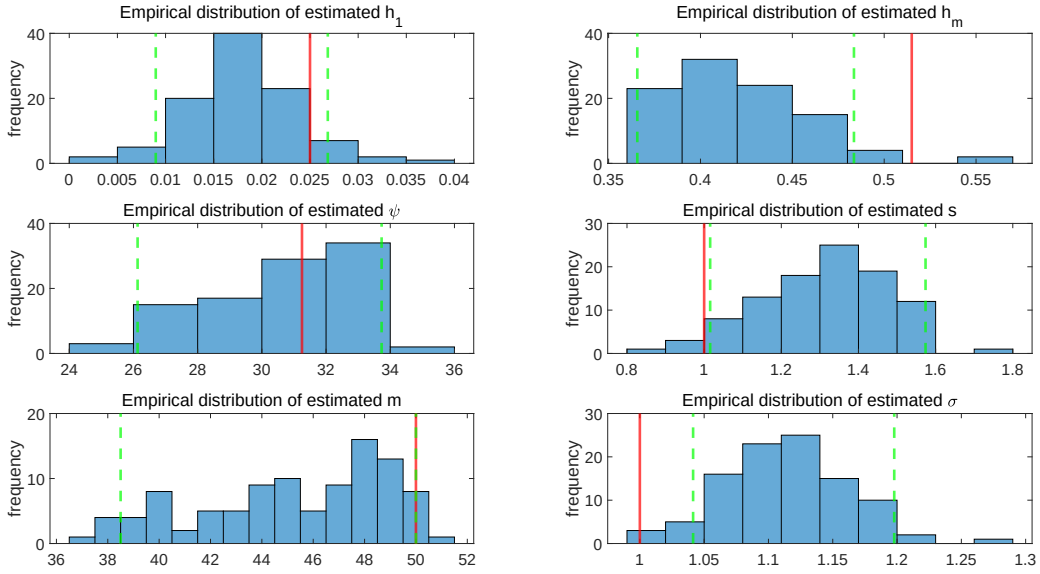

Figure 6.9: Histogram of one hundred MLEs for the simulation study under the scenario of partial information and unknown measurement error.

△

We set $k = 3$ in previous investigations. *What is then the impact on the estimability if using a larger k?* In other words, the effect of longer duration between two time points in the collection of health-pertinent information is also a matter of interest. To demonstrate the impact by a larger k, we set $k = 7$, meaning the states are observed every 7 years for each individual. Furthermore, we assume σ is unknown, which is a more realistic occurrence. Similar to the previous investigation, the parameter σ needs to be estimated from the data.

Example 6.7. We sample 100 data sets from the true PTAM. Each data set has 1,000 individuals, whose states are reported, with measurement error, every 7 years. The true standard deviation of the measurement error is 1.

Table 6.6: Summary statistics for 100 MLEs under the scenario where states can be observed every 7 years with unknown measurement error.

	mean	standard deviation	Q(0.05)	Q(0.95)	true value
h_1	0.0323	0.0058	0.0215	0.0412	0.025
h_m	0.4296	0.0377	0.3757	0.5036	0.515
ψ	24.9278	0.6944	24.2117	25.7927	31.25
s	0.9289	0.1535	0.6668	1.18466	1.000
m	37	1	37	38	50
σ	1.2700	0.2660	0.5785	1.5751	1.000

The mean, standard error, 5th quantile, and 95th quantile for one hundred MLEs under this scenario are given in Table 6.6, and the corresponding distribution is depicted in Figure 6.10. As $C_{i,j}$ is measured with a longer frequency, the estimates have more bias especially for m and ψ. This finding is revealed by the locations of the red lines in Figures 6.9 – 6.10. However, the MLE for ψ and m has less variability even when $C_{i,j}$ is gathered less frequently. One explanation may be due to the underestimation of ψ and m when health-related information is collected every 7 years.

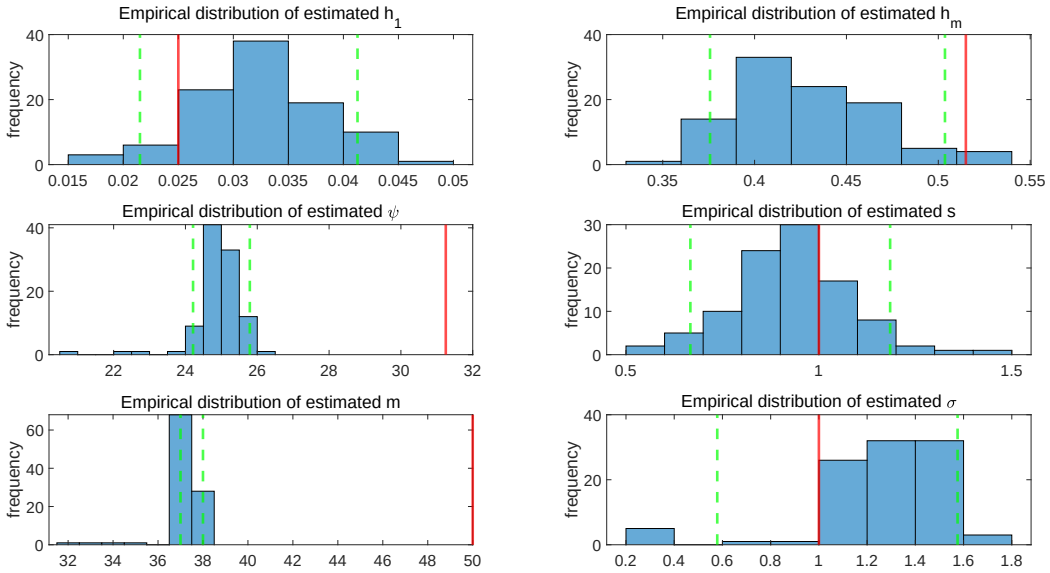

Figure 6.10: Histogram of one hundred MLEs for the simulation study under the scenario where states are observed every 7 years with an unknown measurement error.

△

No information

Recall that when the observed lifetimes are $(t_1, ..., t_n)$, the likelihood can be calculated as

$$L(\boldsymbol{\theta}; \boldsymbol{t}) = \prod_{j=1}^{n} \boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t_j} \boldsymbol{h}.$$

and the log-likelihood is

$$l(\boldsymbol{\theta}; \boldsymbol{t}) = \sum_{j=1}^{n} \log\left(\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t_j} \boldsymbol{h}\right), \tag{6.4.5}$$

Example 6.8. We sample 100 sets of data from the true PTAM. Each set of data has 1,000 individuals, and we can only observe each death time. The MLE for each set of data can be numerically obtained using (6.4.5). Furthermore, since we need to fix m before estimating the other parameters, the values of m are set to $30, 40, 50, 60$ and 70. For each set of data, we can get the MLE for each fixed value of m, and the MLE yielding the highest likelihood is considered as the ultimate MLE.

Table 6.7: Summary statistics for 100 MLEs under the no-information case.

	mean	standard error deviation	Q(0.05)	Q(0.95)	true value
h_1	0.0342	0.0055	0.0245	0.0416	0.025
h_m	2.0356	1.4802	0.2451	3.9998	0.515
ψ	53.4666	27.3086	10.8214	99.9988	31.25
s	0.6121	0.3173	-0.0045	1.0985	1.000
m	44	16	30	70	50

In Table 6.7, summary statistics for one hundred MLE's is provided under the no-information situation. Figure 6.11 is the histogram of 100 MLEs, with each MLE obtained from 1,000 data samples. The standard deviations for the parameters are relatively higher compared with the previous cases. In particular, around 70% of the estimates of m amass at the boundaries (either at 30 or 70), which means the likelihood may increase when setting m equal to the value beyond the boundaries. Therefore, the poorest estimability happens when no information is observed, especially the estimability of m. This is expected because it is the hardest scenario to learn the true parameter value with no information concerning the underlying process.

△

Assessment of estimability by data cloning

We assess the estimability under each scenario by comparing the distribution of the estimate of each parameter. Based on Figure 6.12 , the standard deviations for the estimates getting larger the less information we have for the physiological age, implying a more problematic estimability.

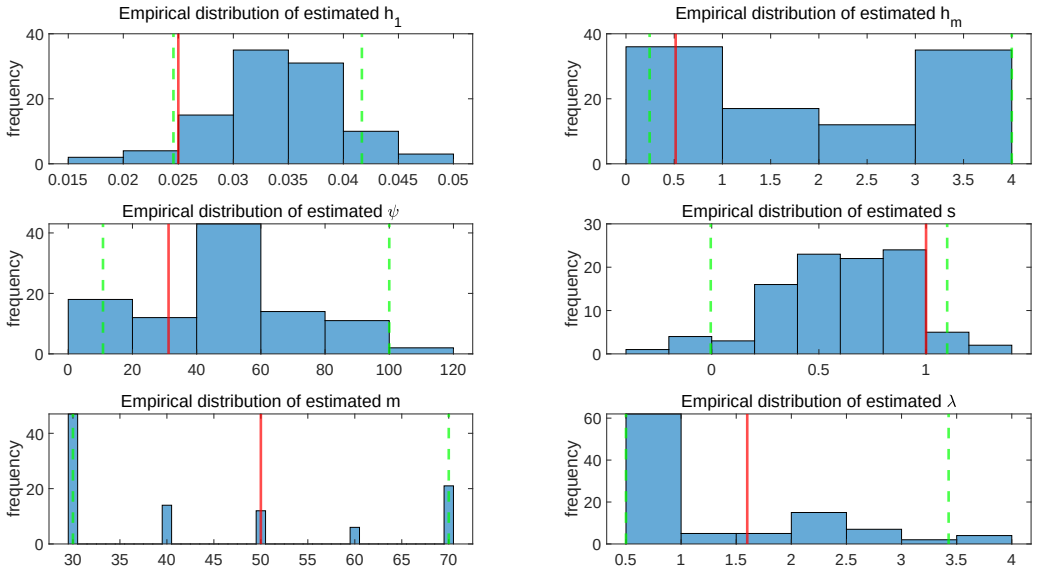

Figure 6.11: Histogram of one hundred MLEs for the simulation study under the scenario no information.

We simulate observations in each scenario. One thousand complete paths of individual's ageing process are simulated from the true model. The only difference in each scenario is the observed information. Since we are going to explore the posterior variance with different number of clones, we need to select a prior. It is reasonable to assume the prior for each parameter in the PTAM is independent and each prior is uniformly distributed. This is because posterior calculation is easiest when the priors for each estimate are independent. For each set of data with K clones, the posterior is obtained by the MCMC using the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Once the posterior is obtained, it is straightforward to get the covariance matrix of the posterior, which can be used to approximate the theoretical covariance matrix.

According to Theorem 6.2, the model is estimable if the largest eigenvalue of the covariance matrix of the posterior converges to 0 as the number of clones approaches infinity. Additionally, Lele et al. (2010) proved that when the model is estimable, the convergence rate for the standardised largest eigenvalue $\lambda_K^s = \lambda_K^*/\lambda_1^*$ is about $\frac{1}{K}$, where λ_K^* and λ_1^* are the largest eigenvalues of the posterior covariance matrices for cloning K times and for the original data, respectively. Therefore, we can graphically compare the standardised largest eigenvalues with increasing K and the line $\frac{1}{K}$ under each scenario. Faster convergence implies better model estimability. This is because the data contains enough information to generate a distinct peak around the MLE, when the model is estimable for such a data set. Data cloning is applied to exaggerate the curvature of the likelihood around the MLE, when model estimability is relatively poor for a given data set. The results under each scenario are summarised in Figure 6.12. There is a clear pattern that the convergence rate for λ_K^s is slower when there is less physiological age information to observe. Apparently, λ_K^s does not converge to 0 as K increases to a fairly large number, when no physiological age information is provided; see Figure 6.12. These data-cloning results

are consistent with the previous conclusion; that is, the model estimabiltiy is relatively poorer when less information is observed.

In summary, the estimability analysis using data cloning supports the conclusion that health-related (state-related) information can improve model estimability. Even observing the physiological age only every few years, despite the unknown measurement error, can improve the model estimability significantly. According to our experiments, we highly recommend incorporating some health-related information when calibrating the PTAM. On the other hand, if no health-related information is available, the PTAM's estimability is relatively poor when both m and ψ are free parameters. One way to improve the model estimability under such a situation is to fix or estimate both m and ψ before estimating the other parameters. This was shown in the application covering the Channing House data (i.e.,fixing m and ψ based on prior knowledge) and the simulation study involving the Le Bras model (i.e., estimating ψ from the data).

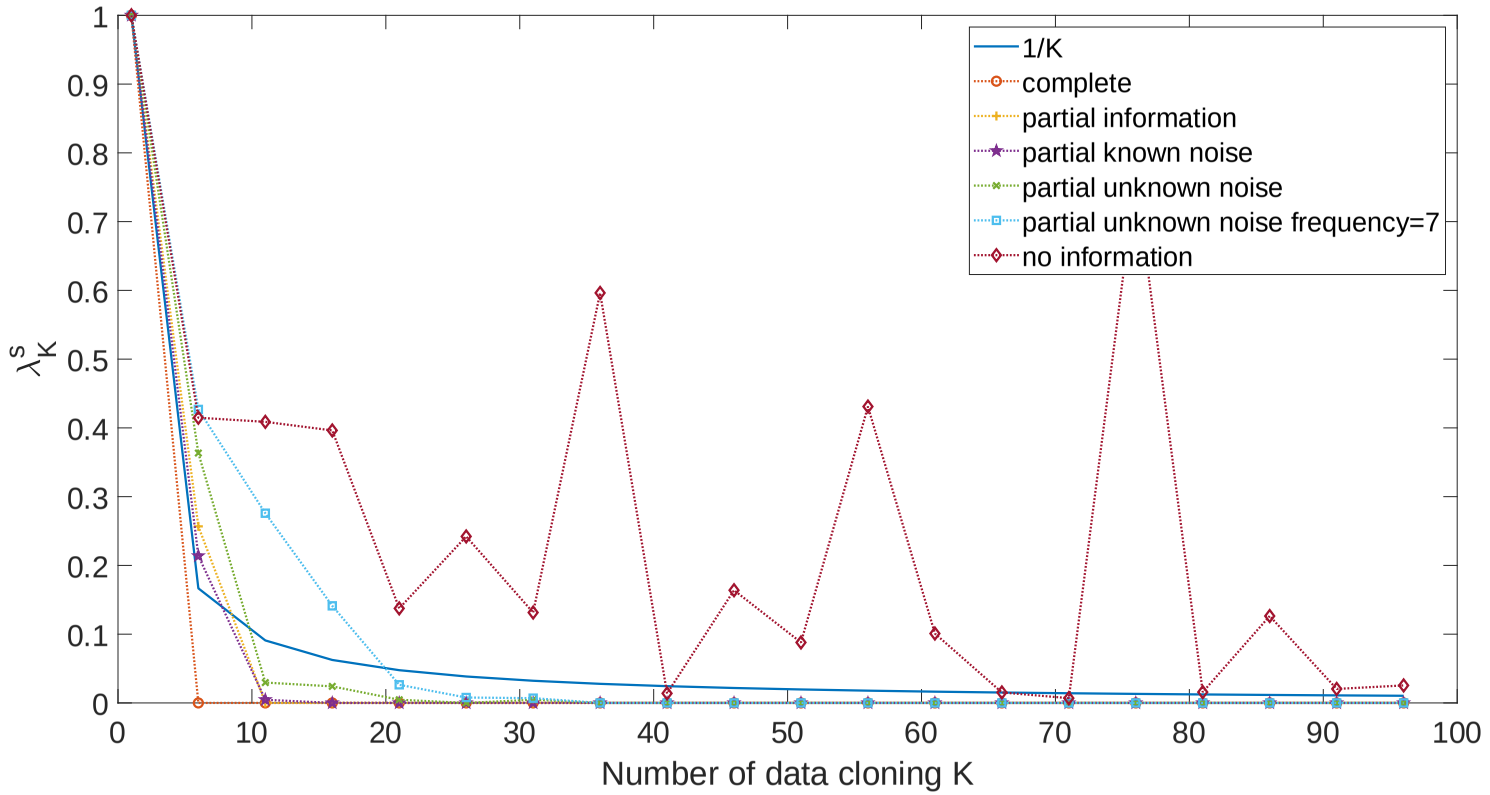

Figure 6.12: Illustrating the standardised largest eigenvalue of the posterior covariance matrix converging to 0 as $K \to \infty$ under various scenarios.

Chapter 7

Conclusion

7.1 Summary of contributions

In this thesis, we proposed a class of distributions, under the so-called the Phase-Type Ageing Model (PTAM), to capture the human ageing process. The PTAM has an embedded Markov chain in which

- each individual's ageing process starts in state 1;
- the process is irreversible, meaning the transition can only move to the next transient state or to the absorbing state from each transient state;
- the transition and absorbing rates have their own structural forms.

The model structure is aligned to our prior knowledge of the human ageing process, which is progressive, essentially irreversible, personalised and highly correlated with mortality process. The PTAM is useful for insurance company to emphasise the ageing experience in the cohort population. In doing so, the pricing strategy can be adapted and tailored to the individual health profile.

In terms of the resulting lifetime distribution, the PTAM as detailed in Chapter 3 can approximate well a variety of lifetime distributions including the Gamma distribution, Weibull distribution, Pareto distribution, a convolution of exponential distributions, a convolution of Weibull distributions, Gompertz-Makeham model, and the Makeham's second extension of the Gompertz distribution. As a result, the lifetime distribution of the PTAM was demonstrated to offer extensive flexibility. Two applications of PTAM, fitting the lifetime data, were considered. The first application is on a data set from the Channing House, a retirement community in Palo Alto, California; the second application is based on a simulated data set from the Le Bras model. Results are promising, when we compare the fitted PTAM with the Kaplan-Meier estimates for the Channing House data set and with the true distribution generating the data for the simulation study. We also illustrated the flexibility of the proposed structure on the absorption rate with a numerical result indicating that the structure can replicate the pattern estimated by Lin and Liu's model for Swedish cohort data. So, the proposed PTAM can be used for mortality modelling by considering the death time as the terminal time of the ageing process.

To quantify the heterogeneity in the lifetime dynamic process, we proposed an index called the physiological age. This concept has an easy interpretation as a comparable value similar to the chronological age. We demonstrated the analysis of the physiological age on the Channing House data set and the Le Bras model simulation data set. The estimated physiological age distribution at any age informs the variability of ageing effect amongst cohorts – the process of an individual with a lower ageing rate progresses slower than that of an individual with a higher ageing rate. In Chapter 4, we proved that the physiological age converges to the chronological age as the number of states in the Markov chain goes to infinite. Such Markov chain with infinite states can be treated as the limit case of the Markov chain with a large number of states. In reality, we can apply this property when the total number of states is large, e.g., $m > 1000$. Thus, a very large number of states is not advisable if we expect some variability on the physiological age at any chronological age. Otherwise, the PTAM loses the heterogeneity of the states. Nonetheless, a very small number of states is not ideal either because the PTAM does not have enough number of states to mimic the lifetime process. We selected the total number of states based on certain educated beliefs including the incorporation of some health-related information to aid in choosing of the the total number of states. In Chapter 5, the model parameter estimation was carried out using the MLE approach. There is no analytical formula for the MLE of the PTAM, and numerical optimisation is required. The numerical procedure involves a large number of likelihood evaluations, which are time-consuming by the traditional method using matrix exponential. We developed a method that is faster and more accurate than the traditional method for likelihood calculation. We also put forward a procedure that is likely to locate the global maximum of the log-likelihood function.

It was shown in Chapter 6 that the model estimability of the PTAM is poor if only the death time is observable, under which at least one of the parameter estimates has a high variance even with a large sample-size observation data. However, the model estimability can be improved significantly even if we can only measure the state every couple years with unknown measurement errors. We examined the scenarios that the state information can be measured every 3 and 7 years. We used 3 and 7 years because age 30 is a limit age that the population could unlikely survive to, and we wanted to mimic the case that the state-related information is collected every 10 years and 30 years in human lifespan scale. Recall that the limit age for human is around 120. In reality, it is impractical to collect the state-related information very frequently due to cost, resources, and time limitations. We believe the frequency of every 10 and 30 years are practically doable based on current technology. In the PTAM calibration, we recommend not only to use the lifetime data but also health variables that could reflect the current state of the individual.

7.2 Future research directions

In the course of this study, several important questions were still left unanswered and they could be pursued as part of future research.

- Some health variables reflecting the current states could be beneficial in calibrating the PTAM. For example, a healthier individual tends to have a stronger grip strength and and a balanced body mass index. It is natural to ask, *what kind of observable health variables*

do we need? What are presently available? And, how information could be extracted efficiently and dynamically from these observed variables?

- We proposed a class of distributions called the GPTAM, which is dense in the field of variables with continuous survival function and non-negative domain. GPTAM's enormous flexibility brings about a challenging parameter-estimation problem. Additional information may be helpful in addressing this issue. For example, the time between transitions and the path of each individual process could be collected in the ideal-world setting. The recovery of parameters then becomes straightforward under various scenarios. However, we are not in the ideal world. *Thus, what additional information will be useful?*

- The PTAM can be used to compare the lifetime processes between populations. The population here represents a group of homogeneous individuals. For example, the process for cohorts 100 years ago could be significantly different from the process for cohorts 100 years in the future; The ageing process for different countries could also have distinct variations. The process for different factors (genders, races, living habits, etc) could be different as well. So, *how should PTAM be re-designed or extended to incorporate the above-mentioned factors in the comparison of the ageing process between populations?* One possible way is to set h_1, h_m and s in the proposed PTAM as functions of the observed factors.

Bibliography

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3):227–243.

Aalen, O. O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, 22(4):447–463.

Aalen, O. O. and Gjessing, H. K. (2001). Understanding the shape of the hazard rate: A process point of view (with comments and a rejoinder by the authors). *Statistical Science*, 16(1):1–22.

Al-Mohy, A. H. and Higham, N. J. (2009). A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3):970–989.

Alalouf, I., Styan, G. P., et al. (1979). Characterizations of estimability in the general linear model. *The Annals of Statistics*, 7(1):194–200.

Albrecher, H. and Bladt, M. (2018). Inhomogeneous phase–type distributions and heavy tails. *arXiv preprint arXiv:1812.04139*.

Asmussen, S. (1992). Phase-type representations in random walk and queueing problems. *Annals of Probability*, 20(2):772–789.

Asmussen, S. (2000). Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics*, 27(2):193–226.

Asmussen, S., Avram, F., and Pistorius, M. R. (2004). Russian and American put options under exponential phase-type Lévy models. *Stochastic Processes and their Applications*, 109(1):79–111.

Asmussen, S., Laub, P. J., and Yang, H. (2019). Phase-type models in life insurance: Fitting and valuation of equity-linked benefits. *Risks*, 7(1):17.

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441.

Atzmon, G., Schechter, C., Greiner, W., Davidson, D., Rennert, G., and Barzilai, N. (2004). Clinical phenotype of families with longevity. *Journal of the American Geriatrics Society*, 52(2):274–277.

Augustin, R. and Büscher, K.-J. (1982). Characteristics of the COX-distribution. *ACM Sigmetrics Performance Evaluation Review*, 12(1):22–32.

Ausın, M., Wiper, M. P., and Lillo, R. E. (2004). Bayesian estimation for the M/G/1 queue using a phase-type approximation. *Journal of Statistical Planning and Inference*, 118(1-2):83–101.

Badescu, A. L., Cheung, E. C., and Landriault, D. (2009). Dependent risk models with bivariate phase-type distributions. *Journal of Applied Probability*, 46(1):113–131.

Beekman, M., Blauw, G. J., Houwing-Duistermaat, J. J., Brandt, B. W., Westendorp, R. G., and Slagboom, P. E. (2006). Chromosome 4q25, microsomal transfer protein gene, and human longevity: Novel data and a meta-analysis of association studies. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 61(4):355–362.

Bellman, R. E. and Kalaba, R. E. (1959). *Dynamic Programming and Feedback Control*. RAND Corporation. Santa Monica, California.

Bickart, T. A. (1968). Matrix exponential: Approximation by truncated power series. *Proceedings of the IEEE*, 56(5):872–873.

Bobbio, A. and Cumani, A. (1992). ML estimation of the parameters of a PH distribution in triangular canonical form. *Computer Performance Evaluation*, 22:33–46.

Bolch, G., Greiner, S., De Meer, H., and Trivedi, K. S. (2006). *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons. Chicester.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Bunch, D. S. (1991). Estimability in the multinomial probit model. *Transportation Research Part B: Methodological*, 25(1):1–12.

Bunke, H. and Bunke, O. (1974). Identifiability and estimability. *Statistics: A Journal of Theoretical and Applied Statistics*, 5(3):223–233.

Campbell, D. and Lele, S. (2014). An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems. *Computational Statistics & Data Analysis*, 70:257–267.

Capri, M., Salvioli, S., Monti, D., Caruso, C., Candore, G., Vasto, S., Olivieri, F., Marchegiani, F., Sansoni, P., and Baggio, G. (2008). Human longevity within an evolutionary perspective: the peculiar paradigm of a post-reproductive genetics. *Experimental Gerontology*, 43(2):53–60.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, volume 2. Duxbury. Pacific Grove, California.

Cavanaugh, J. E. and Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3):e1460.

Cevenini, E., Invidia, L., Lescai, F., Salvioli, S., Tieri, P., Castellani, G., and Franceschi, C. (2008). Human models of aging and longevity. *Expert Opinion on Biological Therapy*, 8(9):1393–1405.

Cheng, B., Jones, B., Liu, X., and Ren, J. (2021). The mathematical mechanism of biological aging. *North American Actuarial Journal*, accepted:DOI:10.1080/10920277.2020.1775654.

Choudhury, A., Choudhury, D., Roy, B., and Mandal, A. (1968). On the evaluation of $e^{A\tau}$. *Proceedings of the IEEE*, 56(6):1110–1111.

Cody, W., Meinardus, G., and Varga, R. (1969). Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems. *Journal of Approximation Theory*, 2(1):50–65.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(2):313–319.

Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics Reliability*, 22(3):583–602.

David, A. and Larry, S. (1987). The least variable phase type distribution is erlang. *Stochastic Models*, 3(3):467–473.

Deiana, L., Ferrucci, L., Pes, G., Carru, C., Delitala, G., Ganau, A., Mariotti, S., Nieddu, A., Pettinato, S., and Putzu, P. (1999). AKEntAnnos. The Sardinia study of extreme longevity. *Aging Clinical and Experimental Research*, 11(3):142–149.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dong, X., Milholland, B., and Vijg, J. (2016). Evidence for a limit to human lifespan. *Nature*, 538(7624):257.

Duan, Q. and Liu, J. (2015). Modelling a bathtub-shaped failure rate by a Coxian distribution. *IEEE Transactions on Reliability*, 65(2):878–885.

Eberly, L. E. and Carlin, B. P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine*, 19(17-18):2279–2294.

Elliott, R. J. and Mamon, R. S. (2002). An interest rate model with a Markovian mean reverting level. *Quantitative Finance*, 2(6):454–458.

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197.

Fackrell, M. (2009). Modelling healthcare systems with phase-type distributions. *Health Care Management Science*, 12(1):11.

Fackrell, M. W. (2003). *Characterisation of matrix-exponential distributions.* PhD thesis, School of Applied Mathematics, Faculty of Engineering, Computer and Mathematical Sciences, The University of Adelaide, South Australia, Australia.

Faddy, M. (1993). A structured compartmental model for drug kinetics. *Biometrics*, 49(1):243–248.

Faddy, M. (1994). Examples of fitting structured phase–type distributions. *Applied Stochastic Models and Data Analysis*, 10(4):247–255.

Faddy, M. (1998). On inferring the number of phases in a coxian phase-type distribution. *Stochastic Models*, 14(1-2):407–417.

Faddy, M. (2002). Penalised maximum likelihood estimation of the parameters in a coxian phase-type distribution. In Latouche, G. and Taylor, P., editors, *Matrix-Analytic Methods: Theory and Applications*. World Scientific. Singapore.

Faddy, M. and McClean, S. (1999). Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15(4):311–317.

Faddy, M. and Wilson, R. (2000). Compartmental modelling of equipment subject to partial repair. *Mathematical and Computer Modelling*, 31(10-12):115–120.

Feng, R. (2009). A matrix operator approach to the analysis of ruin-related quantities in the phase-type renewal risk model. *Schweizerische Aktuarvereinigung Mitteilungen*, 19(1):71.

Fisher, F. M. (1966). *The Identification Problem in Econometrics*. McGraw-Hill. New York.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suñer, D., Cigudosa, J. C., Urioste, M., and Benitez, J. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, 102(30):10604–10609.

Franceschi, C. and Bonafè, M. (2003). *Centenarians as a Model for Healthy Aging*. Portland Press Limited. London.

Franceschi, C., Monti, D., Barbieri, D., Grassilli, E., Troiano, L., Salvioli, S., Negro, P., Capri, M., Guido, M., Azzi, R., et al. (1995a). Immunosenescence in humans: Deterioration or remodelling? *International Reviews of Immunology*, 12(1):57–74.

Franceschi, C., Monti, D., Sansoni, P., and Cossarizza, A. (1995b). The immunology of exceptional individuals: the lesson of centenarians. *Immunology Today*, 16(1):12–16.

Friedman, J. H. (1997). On bias, variance, 0/1 — loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77.

Fukui, H. H., Ackert, L., and Curtsinger, J. W. (1996). Deceleration of age-specific mortality rates in chromosomal homozygotes and heterozygotes of Drosophila melanogaster. *Experimental Gerontology*, 31(4):517–531.

Fukui, H. H., Xiu, L., and Curtsinger, J. W. (1993). Slowing of age-specific mortality rates in Drosophila melanogaster. *Experimental Gerontology*, 28(6):585–599.

Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons. Chicester.

Gantmacher (1959). *Applications of the Theory of Matrices*. Interscience Publishers, Inc., New York.

Garg, L., McClean, S., Meenan, B., and Millard, P. (2010). A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. *Health Care Management Science*, 13(2):155–169.

Gerasimov, I. and Ignatov, D. Y. (2004). Age dynamics of body mass and human lifespan. *Journal of Evolutionary Biochemistry and Physiology*, 40(3):343–349.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics: Developed with Special Reference to the Rational Foundation of Thermodynamics*. C. Scribner's Sons. 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. In a letter to Francis Baily, Esq. FRS &c. *Philosophical Transactions of the Royal Society of London*, 115:513–583.

Govorun, M., Jones, B. L., Liu, X., and Stanford, D. A. (2018). Physiological age, health costs, and their interrelation. *North American Actuarial Journal*, 22(3):323–340.

Gross, D. and Miller, D. R. (1984). The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Operations Research*, 32(2):343–361.

Gueresi, P., Troiano, L., Minicuci, N., Bonafé, M., Pini, G., Salvioli, G., Carani, C., Ferrucci, L., Spazzafumo, L., Olivieri, F., et al. (2003). The MALVA (MAntova LongeVA) study: An investigation on people 98 years of age and over in a province of northern Italy. *Experimental Gerontology*, 38(10):1189–1197.

Hajek, B. (2015). *Random Processes for Engineers*. Cambridge University Press. Cambridge.

Hamilton, J. B. (1951). Patterned loss of hair in man: Types and incidence. *Annals of the New York Academy of Sciences*, 53(3):708–728.

Hastings, W. K. (1970). *Monte Carlo Sampling Methods using Markov Chains and their Applications*. Oxford University Press. Oxford.

Healey, M. (1973). Study of methods of computing transition matrices. *Proceedings of the Institution of Electrical Engineers*, 120(8):905–912.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107(1):49–80.

Hengl, S., Kreutz, C., Timmer, J., and Maiwald, T. (2007). Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618.

Herskind, A. M., McGue, M., Holm, N. V., Sörensen, T. I., Harvald, B., and Vaupel, J. W. (1996). The heritability of human longevity: A population-based study of 2872 Danish twin pairs born 1870–1900. *Human Genetics*, 97(3):319–323.

Higham, N. J. (2005). The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193.

Hoem, J. M. (1969). Markov chain models in life insurance. *Blätter der DGVFM*, 9(2):91–107.

Hoem, J. M. (1977). A Markov chain model of working life tables. *Scandinavian Actuarial Journal*, 1977(1):1–20.

Hsiao, C. (1983). Identification. *Handbook of Econometrics*, 1:223–283.

Huzurbazar, A. V. (1999). Flowgraph models for generalized phase type distributions having non-exponential waiting times. *Scandinavian Journal of Statistics*, 26(1):145–157.

Jacquez, J. A. and Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2):201–227.

Jensen, A. (1953). Markoff chains as an aid in the study of Markoff processes. *Scandinavian Actuarial Journal*, 1953(sup1):87–91.

Johnson, M. A. (1993). Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and Erlang distributions. *ORSA Journal on Computing*, 5(1):69–83.

Jones, B. L. (1994). Actuarial calculations using a Markov model. *Transactions of the Society of Actuaries*, 46:227–250.

Jones, H. B. (1956). A special consideration of the aging process, disease, and life expectancy. *Advances in Biological and Medical Physics*, 4:281–337.

Källström, C. G. (1973). *Computing exp (A) and* $\int$ *exp (As) ds*. Division of Automatic Control, Lund Institute of Technology, Lund.

Karlin, S. (2014). *A First Course in Stochastic Processes*. Academic Press. Cambridge, Massachusetts.

Kirchner, R. (1967). An explicit formula for eat. *American Mathematical Monthly*, 74(10):1200–1204.

Klimenok, V. and Dudin, A. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, 54(4):245–259.

Kodell, R. and Nelson, C. (1980). An illness-death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics*, 36(2):267–277.

Kullback, S. (1997). *Information Theory and Statistics*. Courier Corporation. Chelmsford, Massachusetts.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*, volume 5. SIAM. Philadelphia.

Le Bras, H. (1976). Lois de mortalité et âge limite. *Population (French Edition)*, pages 655–692.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media. New York.

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492):1617–1625.

Leroi, A. M., Bartke, A., De Benedictis, G., Franceschi, C., Gartner, A., Gonos, E., Feder, M. E., Kivisild, T., Lee, S., Kartal-Özer, N., et al. (2005). What evidence is there for the existence of individual genes with antagonistic pleiotropic effects? *Mechanisms of Ageing and Development*, 126(3):421–429.

Li, J. S. and Ng, A. C. (2008). "Markov Aging Process and Phase-Type Law of Mortality," X. Sheldon Lin and Xiaoming Liu, October 2007. *North American Actuarial Journal*, 12(1):90–94.

Liedo, P., Orozco, D., and Vaupel, J. (1992). Slowing of mortality rates at older ages in large medfly cohorts. *Science*, 258(5081):457–461.

Lin, X. S. and Liu, X. (2007). Markov aging process and phase-type law of mortality. *North American Actuarial Journal*, 11(4):92–109.

Liou, M. (1966). A novel method of evaluating transient response. *Proceedings of the IEEE*, 54(1):20–23.

Liu, X. and Lin, X. S. (2012). A subordinated Markov for stochastic mortality. *European Actuarial Journal*, 2(1):105–127.

Makeham, W. M. (1860). On the law of mortality and construction of annuity tables. *Journal of the Institute of Actuaries*, 8(6):301–310.

Makeham, W. M. (1890). On the Further Development of Gompertz's Law. *Journal of the Institute of Actuaries*, 28(4):316–332.

Mamon, R. S. and Elliott, R. J. (2007). *Hidden Markov Models in Finance*. Springer. New York.

Markov, A. A. (1954). The theory of algorithms. *Trudy Matematicheskogo Instituta Imeni VA Steklova*, 42:3–375.

Markov, A. A. (2006). An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. *Science in Context*, 19(4):591–600.

Marshall, A. H. and Zenga, M. (2009a). Recent developments in fitting coxian phase-type distributions in healthcare. In *ASMDA. Proceedings of the International Conference Applied Stochastic Models and Data Analysis*, volume 13, page 482. Department of Construction Economics, Vilnius Gediminas Technical University, Vilnius.

Marshall, A. H. and Zenga, M. (2009b). Simulating Coxian phase-type distributions for patient survival. *International Transactions in Operational Research*, 16(2):213–226.

Marshall, A. H. and Zenga, M. (2012). Experimenting with the Coxian phase-type distribution to uncover suitable fits. *Methodology and Computing in Applied Probability*, 14(1):71–86.

McLean, K. A. and McAuley, K. B. (2012). Mathematical modelling of chemical processes—obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Canadian Journal of Chemical Engineering*, 90(2):351–366.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Miettinen, O. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, 103(2):226–235.

Moler, C. and Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49.

Moriguchi, S. and Murota, K. (2012). On discrete Hessian matrix and convex extensibility. *The Operations Research Society of Japan*, 55:48–62.

Mu, J. (2019). Exploring the estimability of mark-recapture models with individual, time-varying covariates using the scaled logit link function. *Electronic Thesis and Dissertation Repository*, 6385. The University of Western Ontario, Canada.

Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Neuts, M. F. (1975). *Probability distributions of phase type*. Liber Amicorum Prof. Emeritus H. Florin, Department of Mathematics, University of Louvain, Belgium.

Neuts, M. F. (1978). Markov chains with applications in queueing theory, which have a matrix-geometric invariant probability vector. *Advances in Applied Probability*, 10(1):185–212.

Neuts, M. F. (1982). Explicit steady-state solutions to some elementary queueing models. *Operations Research*, 30(3):480–489.

Olsson, M. (1996). Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, pages 443–460.

Pandhi, D. and Khanna, D. (2013). Premature graying of hair. *Indian Journal of Dermatology, Venereology, and Leprology*, 79(5):641.

Pérez-Ocón, R. and Castro, J. R. (2004). Two models for a repairable two-system with phase-type sojourn time distributions. *Reliability Engineering & System Safety*, 84(3):253–260.

Perls, T. T., Wilmoth, J., Levenson, R., Drinkwater, M., Cohen, M., Bogan, H., Joyce, E., Brewster, S., Kunkel, L., and Puca, A. (2002). Life-long sustained mortality advantage of siblings of centenarians. *Proceedings of the National Academy of Sciences*, 99(12):8442–8447.

Prelog, M. (2006). Aging of the immune system: A risk factor for autoimmunity? *Autoimmunity Reviews*, 5(2):136–139.

Putzer, E. J. (1966). Avoiding the jordan canonical form in the discussion of linear systems with constant coefficients. *American Mathematical Monthly*, 73(1):2–7.

Ramírez-Cobo, P., Lillo, R. E., and Wiper, M. P. (2010). Nonidentifiability of the two-state Markovian arrival process. *Journal of Applied robability*, 47(3):630–649.

Ricklefs, R. E. (1998). Evolutionary theories of aging: confirmation of a fundamental prediction, with implications for the genetic basis and evolution of life span. *The American Naturalist*, 152(1):24–44.

Rizk, J., Burke, K., and Walsh, C. (2019). An alternative formulation of Coxian phase-type distributions with covariates: Application to emergency department length of stay. *arXiv preprint arXiv:1907.13489*.

Rodríguez Valiente, A., Trinidad, A., García Berrocal, J., Górriz, C., and Ramírez Camacho, R. (2014). Extended high-frequency (9–20 khz) audiometry reference thresholds in 645 healthy subjects. *International Journal of Audiology*, 53(8):531–545.

Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. L. (1998). *Stochastic Processes for Insurance and Finance*. John Wiley & Sons. Chicester.

Ross, S. M. (2014). *Introduction to Probability Models*. Academic Press. Cambridge, Massachusetts.

Ruiz-Castro, J. E., Pérez-Ocón, R., and Fernández-Villodre, G. (2008). Modelling a reliability system governed by discrete phase-type distributions. *Reliability Engineering & System Safety*, 93(11):1650–1657.

Saff, E. (1971). The convergence of rational functions of best approximation to the exponential function. *Transactions of the American Mathematical Society*, 153:483–493.

Salvioli, S., Olivieri, F., Marchegiani, F., Cardelli, M., Santoro, A., Bellavista, E., Mishto, M., Invidia, L., Capri, M., Valensin, S., et al. (2006). Genes, ageing and longevity in humans: problems, advantages and perspectives. *Free Radical Research*, 40(12):1303–1323.

Sansoni, P., Vescovini, R., Fagnoni, F., Biasini, C., Zanni, F., Zanlari, L., Telera, A., Lucchini, G., Passeri, G., and Monti, D. (2008). The immune system in extreme longevity. *Experimental Gerontology*, 43(2):61–65.

Schoenmaker, M., de Craen, A. J., de Meijer, P. H., Beekman, M., Blauw, G. J., Slagboom, P. E., and Westendorp, R. G. (2006). Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden longevity study. *European Journal of Human Genetics*, 14(1):79–84.

Sengupta, B. (1989). Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Advances in Applied Probability*, 21(1):159–180.

Sherris, M. and Zhou, Q. (2014). *Model risk, mortality heterogeneity, and implications for solvency and tail risk*. Pension Research Council.

Sidje, R. B. (1998). Expokit: A software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, 24(1):130–156.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press. Princeton.

Su, S. and Sherris, M. (2012). Heterogeneity of Australian population mortality and implications for a viable life annuity market. *Insurance: Mathematics and Economics*, 51(2):322–332.

Szilard, L. (1959). On the nature of the aging process. *Proceedings of the National Academy of Sciences*, 45(1):30–45.

Thompson, D. (2018). Does human life span really have a limit? `https://www.webmd.com/healthy-aging/news/20180628/does-human-life-span-really-have-a-limit#1`.

Thummler, A., Buchholz, P., and Telek, M. (2006). A novel approach for phase-type fitting with the EM algorithm. *IEEE Transactions on Dependable and Secure Computing*, 3(3):245–258.

Thurstan, S. A., Gibbs, N. K., Langton, A. K., Griffiths, C. E., Watson, R. E., and Sherratt, M. J. (2012). Chemical consequences of cutaneous photoageing. *Chemistry Central Journal*, 6(1):34.

Tijms, H. C. (1994). *Stochastic models: An algorithmic approach*. John Wiley & Sons. Chichester.

Titman, A. C. and Sharples, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics*, 66(3):742–752.

Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1(3):306–307.

Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press. Cambridge.

U.S. Department of Health and Human Services (2016). NIDCD Fact Sheet Hearing and Balance: Hearing Loss and Older Adults. `https://www.nidcd.nih.gov/sites/default/files/Documents/health/hearing/HearingLossOlderAdults.pdf`, last updated march 2016.

Vary Jr, J. C. (2016). Selected disorders of skin appendages – acne, alopecia, hyperhidrosis. *The Medical Clinics of North America*, 99(6):1195–1211.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228.

Ward, R. C. (1977). Numerical computation of the matrix exponential with accuracy estimate. *SIAM Journal on Numerical Analysis*, 14(4):600–610.

Willcox, B. J., Willcox, D. C., He, Q., Curb, J. D., and Suzuki, M. (2006). Siblings of Okinawan centenarians share lifelong mortality advantages. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 61(4):345–354.

Woodbury, M. A. and Manton, K. G. (1977). A random-walk model of human mortality and aging. *Theoretical Population Biology*, 11(1):37–48.

Yashin, A. I., Arbeev, K. G., Ukraintseva, S. V., Akushevich, I., and Kulminski, A. (2012). Patterns of aging-related changes on the way to 100. *North American Actuarial Journal*, 16(4):403–433.

Yashin, A. I., Iachine, I. A., and Begun, A. S. (2000). Mortality modeling: A review. *Mathematical Population Studies*, 8(4):305–332.

Yashin, A. I., Manton, K. G., and Vaupel, J. W. (1985). Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables. *Theoretical Population Biology*, 27(2):154–175.

Yashin, A. I., Vaupel, J. W., and Iachine, I. A. (1994). A duality in aging: the equivalence of mortality models based on radically different concepts. *Mechanisms of Ageing and Development*, 74(1-2):1–14.

Zhang, C. and Vahidi, A. (2011). Predictive cruise control with probabilistic constraints for eco driving. In *ASME 2011 dynamic systems and control conference and bath/ASME symposium on fluid power and motion control*, pages 233–238. American Society of Mechanical Engineers Digital Collection. New York.

Zhang, Y. (2016). Actuarial Modelling with Mixtures of Markov Chains. *Electronic Thesis and Dissertation Repository*, 4026.

Zimmer, C. (2016). *What's the longest humans can live? 115 years, new study says*. The New York Times. Retrieved 6 October 2016.

Appendix A

Proofs

This Appendix includes the proof of the denseness of the GPTAM in the continuous non-negative-valued distribution, the justification of the state distribution for the GPTAM, and the proofs of Theorem 3.1, Lemma 3.1 and Lemma 4.1.

A.1 Proof of Theorem 3.1

Let $\mathcal{F}$ be the distribution family of all GPTAMs whose absorbing rate in state i is equal to $h_i \geq 0$ for $i = 1, ..., m$; transition rate from state j to $j+1$ is equal to λ for $j = 1, ..., m-1$; and the total number of states m is a positive integer. Given a non-negative-valued distribution with domain $(0, T)$, assume its survival function $S(t)$ is continuous. For any $\epsilon > 0$, there is a GPTAM in $\mathcal{F}$ such that its resulting survival function $S^*(t)$ satisfies $|S^*(t) - S(t)| < \epsilon$ for any $t \in (0, T)$. Therefore, $\mathcal{F}$ is dense in the field of all continuous non-negative-valued distributions.

Proof. The proof follows by adapting the approach employed in Rolski et al. (1998) and Tijms (1994). Let $S(t)$ be any survival function on $\mathbb{R}^+$ with $S(0) = 1$ and $\lim_{t\to T} S(t) = 0$. We are letting $T = \infty$ when the distribution with non-negative value domain is the point of interest.

For any fixed $t \in (0, T)$, consider the following approximation of $S(t)$:

$$S_n(t) = \sum_{k=0}^{\infty} e^{-nt} \frac{(nt)^k}{k!} S\left(\frac{k}{n}\right) = \mathbb{E}\left(S\left(\frac{X_n(t)}{n}\right)\right),$$

where $X_n(t)$ follows a Poisson distribution with rate nt. Since $S(t) \leq 1$ is bounded and continuous at t, for each $\epsilon > 0$, there is $\delta > 0$ such that $|S(x) - S(t)| \leq \epsilon/2$ whenever $|x - t| \leq \delta$. Thus,

$$\begin{aligned}
&|S_n(t) - S(t)| \\
&= \sum_{k:\left|\frac{k}{n}-t\right|\leq\delta} \left|S\left(\frac{k}{n}\right) - S(t)\right| P(X_n(t) = k) + \sum_{k:\left|\frac{k}{n}-t\right|>\delta} \left|S\left(\frac{k}{n}\right) - S(t)\right| P(X_n(t) = k) \\
&\leq \frac{\epsilon}{2} + 2P\left(\left|\frac{X_n(t)}{n} - t\right| > \delta\right) \leq \frac{\epsilon}{2} + \frac{2t}{\delta^2 n},
\end{aligned}$$

where the first inequality holds because of the following facts:

- $\left|S\left(\frac{k}{n}\right) - S(t)\right| \leq \epsilon/2$, when $\left|\frac{k}{n} - t\right| \leq \delta$;
- $\sum_{\left|\frac{k}{n}-t\right| \leq \delta} P(X_n(t) = k) \leq 1$;
- $\left|S\left(\frac{k}{n}\right) - S(t)\right| \leq 2$ always holds;
- $\sum_{\left|\frac{k}{n}-t\right| > \delta} P(X_n(t) = k) = P\left(\left|\frac{X_n(t)}{n} - t\right| > \delta\right)$.

The last inequality holds by the Chebyshev's inequality, $\mathbb{E}\left(\frac{X_n(t)}{n}\right) = t$, and $\mathrm{Var}\left(\frac{X_n(t)}{n}\right) = \frac{t}{n}$. Therefore, when $n \geq \frac{4t}{\delta^2\epsilon}$ or $\frac{2t}{\delta^2 n} \leq \frac{\epsilon}{2}$, we have

$$|S_n(t) - S(t)| \leq \epsilon. \tag{A.1.1}$$

With the uniformisation method and (5.3.7), for any m, the resulting survival function of a GPTAM with m transient states is

$$S^*(t) = \sum_{i=1}^{m}\left(\sum_{k=0}^{\infty} P_{1,i}^{(k)} e^{-\nu t}\frac{(\nu t)^k}{k!}\right) = \sum_{k=0}^{\infty}\left(\left(\sum_{i=1}^{m} P_{1,i}^{(k)}\right)e^{-\nu t}\frac{(\nu t)^k}{k!}\right),$$

where ν can be any number greater than $\max_{i=1,\ldots,m-1}(\lambda + h_i, h_m)$, and the $P_{1,i}^{(k)}$ is the first row's ith element of $\mathbf{P}^k$. The $\mathbf{P}$ matrix is

$$\mathbf{P} = \begin{bmatrix} 1 - \frac{(\lambda+h_1)}{\nu} & \frac{\lambda}{\nu} & & & \\ & 1 - \frac{(\lambda+h_2)}{\nu} & \frac{\lambda}{\nu} & & \\ & & \ddots & & \\ & & & 1 - \frac{(\lambda+h_{m-1})}{\nu} & \frac{\lambda}{\nu} \\ & & & & 1 - \frac{h_m}{\nu} \end{bmatrix}.$$

Let $\nu = n$ when $n > \lambda + h$ and $h = \max_{i=1,\ldots,m}(h_i)$. Hence,

$$S^*(t) = \sum_{k=0}^{\infty}\left(\left(\sum_{i=1}^{m} P_{1,i}^{(k)}\right)e^{-nt}\frac{(nt)^k}{k!}\right),$$

and

$$\mathbf{P} = \begin{bmatrix} 1 - \frac{(\lambda+h_1)}{n} & \frac{\lambda}{n} & & & \\ & 1 - \frac{(\lambda+h_2)}{n} & \frac{\lambda}{n} & & \\ & & \ddots & & \\ & & & 1 - \frac{(\lambda+h_{m-1})}{n} & \frac{\lambda}{n} \\ & & & & 1 - \frac{h_m}{n} \end{bmatrix}.$$

On the other hand, for any fixed n and any $\xi > 0$ there is a K such that $\sum_{k=K}^{\infty} e^{-nt}\frac{(nt)^k}{k!} < \xi/2$, because $\sum_{k=K}^{\infty} e^{-nt}\frac{(nt)^k}{k!}$ is the tail probability of a Poisson distribution with rate nt.

Next, we need to prove there exists a set of parameter value $(h_1, ..., h_m, \lambda, m)$ such that $\sum_{i=1}^{m} P_{1,i}^{(k)} = S\left(\frac{k}{n}\right)$ for any $k = 0, ..., K-1$.

We know that $P_{1,1} = 1 - \frac{\lambda+h_1}{n}$, $P_{1,2} = \frac{\lambda}{n}$ and $P_{1,j} = 0$ for $j = 3, ..., m$. Suppose $P^{(\ell)}_{1,i}$ is a function of $(h_1, ..., h_\ell, \lambda)$ when $i = 1, .., \ell$, $P^{(\ell)}_{1,\ell+1} = \left(\frac{\lambda}{n}\right)^\ell$ and $P_{1,j} = 0$ for $j > \ell + 1$. Evaluating the matrix power, we have

$$P^{(k)}_{1,1} = \left(1 - \frac{\lambda + h_1}{n}\right)^k,$$
$$P^{(n+1)}_{1,i} = \frac{\lambda}{n} P^{(n)}_{1,i-1} + \left(1 - \frac{\lambda + h_i}{n}\right) P^{(n)}_{1,i}, \text{ for } 2 \le i < m.$$

For any $1 < i \le \ell$,

$$\begin{aligned} P^{(\ell+1)}_{1,i} &= \frac{\lambda}{n} P^{(\ell)}_{1,i-1} + \left(1 - \frac{\lambda + h_i}{n}\right) P^{(\ell)}_{1,i} \\ &= \frac{\lambda}{n}(P^{(\ell)}_{1,i-1} - P^{(\ell)}_{1,i}) + \left(1 - \frac{h_i}{n}\right) P^{(\ell)}_{1,i}, \end{aligned}$$

which is a function of $(h_1, ..., h_\ell, \lambda)$ because $P^{(\ell)}_{1,i}$ is a function of $(h_1, ..., h_\ell, \lambda)$. For $i = \ell + 1$,

$$\begin{aligned} P^{(\ell+1)}_{1,\ell+1} &= \frac{\lambda}{n} P^{(\ell)}_{1,\ell} + \left(1 - \frac{\lambda + h_{\ell+1}}{n}\right) P^{(\ell)}_{1,\ell+1} \\ &= \frac{\lambda}{n} P^{(\ell)}_{1,\ell} + \left(1 - \frac{\lambda + h_{\ell+1}}{n}\right)\left(\frac{\lambda}{n}\right)^\ell, \end{aligned}$$

which is a function of $(h_1, ..., h_\ell, h_{\ell+1}, \lambda)$. In particular, the coefficient of $h_{\ell+1}$ is $-\frac{\lambda^\ell}{n^{\ell+1}}$, which is non-zero. For $i = \ell + 2$,

$$P^{(\ell+1)}_{1,\ell+2} = \frac{\lambda}{n} P^{(\ell)}_{1,\ell+1} + \left(1 - \frac{\lambda + h_{\ell+2}}{n}\right) P^{(\ell)}_{1,\ell+2} = \left(\frac{\lambda}{n}\right)^{\ell+1}$$

For $\ell + 2 < i \le m$,

$$P^{(\ell+1)}_{1,i} = \frac{\lambda}{n} P^{(\ell)}_{1,i-1} + \left(1 - \frac{\lambda + h_i}{n}\right) P^{(\ell)}_{1,i} = 0,$$

because $P^{(\ell)}_{1,i-1}$ and $P^{(\ell)}_{1,i}$ are equal to 0. By induction, $P^{(\ell)}_{1,i}$ is a function of $(h_1, ..., h_\ell, \lambda)$ when $i = 1, .., \ell$, $P^{(\ell)}_{1,\ell+1} = \left(\frac{\lambda}{n}\right)^\ell$ and $P_{1,j} = 0$ for $j > \ell + 1$ for any $\ell = 1, ..., m$. Furthermore, $\sum_{i=1}^m P^{(k)}_{1,i}$ is the probability that the process under the uniformisation is in any transient state given the process transits k times, whose value is determined by $(h_1, ..., h_k, \lambda)$. As a result, the number of parameters in $\sum_{i=1}^m P^{(k)}_{1,i}$ is $k + 1$.

Consequently, with $m \ge K - 2$, there exists a set of parameter values $(h_1, ..., h_m, \lambda)$ such that $\sum_{i=1}^m P^{(k)}_{1,i} = S\left(\frac{k}{n}\right)$ for any $k = 1, ..., K - 1$. Meanwhile, $\sum_{i=1}^m P^{(0)}_{1,i} = S\left(\frac{0}{n}\right) = 1$ for any parameter value $(h_1, ..., h_m, \lambda)$.

Therefore, for any $n > \max(\frac{4t}{\delta^2\epsilon}, \lambda + h)$ and $m \geq K-2$, there exists a set of parameter values $(h_1, ..., h_m, \lambda)$ such that $\sum_{i=1}^{m} P_{1,i}^{(k)} = S\left(\frac{k}{n}\right)$ for any $k = 0, ..., K-1$ and

$$\begin{aligned}
|S^*(t) - S_n(t)| &= \left| \sum_{k=0}^{\infty} e^{-nt} \frac{(nt)^k}{k!} \left(\sum_{i=1}^{m} P_{1,i}^{(k)} - S\left(\frac{k}{n}\right) \right) \right| \\
&= \left| \sum_{k=K}^{\infty} e^{-nt} \frac{(nt)^k}{k!} \left(\sum_{i=1}^{m} P_{1,i}^{(k)} - S\left(\frac{k}{n}\right) \right) \right| \\
&\leq \sum_{k=K}^{\infty} e^{-nt} \frac{(nt)^k}{k!} \left| \sum_{i=1}^{m} P_{1,i}^{(k)} - S\left(\frac{k}{n}\right) \right| \\
&\leq 2 \sum_{k=K}^{\infty} e^{-nt} \frac{(nt)^k}{k!} \leq \xi,
\end{aligned}$$

by recalling $\sum_{k=K}^{\infty} e^{-nt} \frac{(nt)^k}{k!} < \xi/2$. Combining with (A.1.1),

$$\begin{aligned}
|S^*(t) - S(t)| &= |S^*(t) - S_n(t) + S_n(t) - S(t)| \\
&\leq |S^*(t) - S_n(t)| + |S_n(t) - S(t)| \leq \epsilon + \xi.
\end{aligned}$$

Since ϵ and ξ are arbitrarily small, we successfully constructed a GPTAM in $\mathcal{F}$ whose resulting survival function $S^*(t)$ can approximate $S(t)$ well. This shows the distribution family $\mathcal{F}$ is dense in the field of all continuous and non-negative-valued distributions. ■

A.2 Proof of Lemma 3.1

Consider an m−state GPTAM with transition rate from one transient state to the next transient state equal to λ, and the absorption rate in state i is h_i for $i = 1, ..., m$. For any $t > 0$, let Y_t be the state variable at time t. Then $P(Y_t \geq k | Y_t \in E)$ is an increasing function of t for $k = 1, ..., m$.

Proof. The $m \times m$ degenerated transition matrix of the GPTAM is

$$\mathbf{\Lambda} = \begin{bmatrix}
-(\lambda + h_1) & \lambda & & & \\
& -(\lambda + h_2) & \lambda & & \\
& & \ddots & & \\
& & & -(\lambda + h_{m-1}) & \lambda \\
& & & & -h_m
\end{bmatrix},$$

where λ is the transition rate from state i to $i+1$ and h_i is the absorbing rate at state i. For any fixed $t > 0$, let $Y_{\ell,t}$ be the state random variable at time t assuming starting in state ℓ at 0, particularly $Y_{1,t} = Y_t$. For any $k = 1, ..., m$,

$$P(Y_t \geq k | Y_t \in E) = \frac{\sum_{j=k}^{m} P(Y_{1,t} = j)}{\sum_{j=1}^{m} P(Y_{1,t} = j)} = \frac{\boldsymbol{\alpha} e^{\mathbf{\Lambda} t} \boldsymbol{e_k}}{\boldsymbol{\alpha} e^{\mathbf{\Lambda} t} \boldsymbol{e}},$$

where $\boldsymbol{e_k}$ is a $m\times 1$ column vector with first $k-1$ elements equal to 0 and the remaining elements are equal to 1. The quantity $P(Y_t \geq k|Y_t \in E)$ is the probability that the state of the individual at time t is greater than or equal to k conditional on this individual being alive. We note that $P(Y_t \geq 1|Y_t \in E) = 1$. When $k \neq 1$, the first derivative of $P(Y_t \geq k|Y_t \in E)$ with respect to t is

$$\frac{\mathrm{d}P(Y_t \geq k|Y_t \in E)}{\mathrm{d}t} = \frac{\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t}(\boldsymbol{\Lambda e_k \alpha} - \boldsymbol{e_k \alpha \Lambda}) e^{\boldsymbol{\Lambda} t}\boldsymbol{e}}{(\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t}\boldsymbol{e})^2}.$$

Let $g(t) = \boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t}(\boldsymbol{\Lambda e_k \alpha} - \boldsymbol{e_k \alpha \Lambda}) e^{\boldsymbol{\Lambda} t}\boldsymbol{e}$. The (i, j) entry of $e^{\boldsymbol{\Lambda} t}$, denoted by $P_{i,j}(t)$, is the probability of being in state j at time t given that the starting state is i at time 0. One may verify that

$$\begin{aligned}
\boldsymbol{\alpha} e^{\boldsymbol{\Lambda} t} &= (P_{1,1}(t), ..., P_{1,m}(t)) \text{ a } 1 \times m \text{ row vector} \\
e^{\boldsymbol{\Lambda} t}\boldsymbol{e} &= \left(\sum_{j=1}^{m} P_{1,j}(t), ..., \sum_{j=1}^{m} P_{m,j}(t)\right)^{\intercal} \text{ a } m \times 1 \text{ column vector} \\
\boldsymbol{\Lambda e_k \alpha} &= \begin{bmatrix} 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots \\ \lambda & 0 & 0 & \cdots \\ -h_k & 0 & 0 & \cdots \\ -h_{k+1} & 0 & 0 & \cdots \\ \vdots & & & \\ -h_m & 0 & 0 & \cdots \end{bmatrix} \text{ a } m \times m \text{ matrix} \\
\boldsymbol{e_k \alpha \Lambda} &= \begin{bmatrix} 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ -(\lambda + h_1) & \lambda & 0 & \cdots \\ -(\lambda + h_1) & \lambda & 0 & \cdots \\ \vdots & & & \\ -(\lambda + h_1) & \lambda & 0 & \cdots \end{bmatrix} \text{ a } m \times m \text{ matrix,}
\end{aligned}$$

where λ is the $(k-1, 1)$ entry of $\boldsymbol{\Lambda e_k \alpha}$ and the entire first $(k-1)$th rows in $\boldsymbol{e_k \alpha \Lambda}$ are 0.

Hence, $\boldsymbol{\Lambda e_k \alpha} - \boldsymbol{e_k \alpha \Lambda} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \mathbf{0}, ..., \mathbf{0})^{\intercal}$, where $\boldsymbol{\beta_2}$ is a $m \times 1$ column vector with first $k-1$ elements equal to 0 and remaining elements equal to $-\lambda$ and $\boldsymbol{\beta_1}$ is a $m \times 1$ column vector with first $k-2$ elements equal to 0, $(k-1)$th element equal to λ and ℓth($\ell \geq k$) element equal to

$\lambda + h_1 - h_\ell$. In matrix form,

$$\boldsymbol{\Lambda e_k \alpha} - \boldsymbol{e_k \alpha \Lambda} = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots \\ \lambda & 0 & 0 & \cdots \\ \lambda + h_1 - h_k & -\lambda & 0 & \cdots \\ \lambda + h_1 - h_{k+1} & -\lambda & 0 & \cdots \\ \vdots & & & \\ \lambda + h_1 - h_m & -\lambda & 0 & \cdots \end{bmatrix}.$$

Then $g(t)$ can be simplified as

$$\begin{aligned} g(t) &= (P_{1,1}(t), ..., P_{1,m}(t)) \begin{bmatrix} 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots \\ \lambda & 0 & 0 & \cdots \\ \lambda + h_1 - h_k & -\lambda & 0 & \cdots \\ \lambda + h_1 - h_{k+1} & -\lambda & 0 & \cdots \\ \vdots & & & \\ \lambda + h_1 - h_m & -\lambda & 0 & \cdots \end{bmatrix} \left(\sum_{j=1}^{m} P_{1,j}(t), ..., \sum_{j=1}^{m} P_{m,j}(t) \right)^\intercal \\ &= \left(\left(\lambda P_{1,k-1}(t) + \sum_{\ell=k}^{m} (\lambda + h_1 - h_\ell) P_{1,\ell}(t) \right), -\lambda \sum_{\ell=k}^{m} P_{1,\ell}(t), 0, ..., 0 \right) \left(\sum_{j=1}^{m} P_{1,j}(t), ..., \sum_{j=1}^{m} P_{m,j}(t) \right)^\intercal \\ &= \left(\lambda P_{1,k-1}(t) + \sum_{\ell=k}^{m} (\lambda + h_1 - h_\ell) P_{1,\ell}(t) \right) \sum_{j=1}^{m} P_{1,j}(t) - \sum_{\ell=k}^{m} \lambda P_{1,\ell}(t) \sum_{j=1}^{m} P_{2,j}(t). \end{aligned}$$

On the other hand, we know that

$$\frac{\mathrm{d}e^{\boldsymbol{\Lambda} t}}{\mathrm{d}t} = \boldsymbol{\Lambda} e^{\boldsymbol{\Lambda} t} = e^{\boldsymbol{\Lambda} t} \boldsymbol{\Lambda},$$

yielding

$$\begin{aligned} \frac{\mathrm{d}P_{1,k}(t)}{\mathrm{d}t} &= -(\lambda + h_1) P_{1,k}(t) + \lambda P_{2,k}(t) = \lambda P_{1,k-1}(t) - (\lambda + h_k) P_{1,k}(t) \text{ for } k = 2, ..., m-1 \\ \frac{\mathrm{d}P_{1,m}(t)}{\mathrm{d}t} &= -(\lambda + h_1) P_{1,m}(t) + \lambda P_{2,m}(t) = \lambda P_{1,k-1}(t) - h_m P_{1,m}(t). \end{aligned}$$

Rearranging further, we have

$$\begin{aligned} (h_1 - h_k) P_{1,k}(t) &= \lambda (P_{2,k}(t) - P_{1,k-1}(t)) \text{ for } k = 2, ..., m-1 \\ (\lambda + h_1 - h_m) P_{1,m}(t) &= \lambda (P_{2,m}(t) - P_{1,m-1}(t)) \end{aligned}$$

The first part of $g(t)$ can be written as

$$\begin{aligned}
&\lambda P_{1,k-1}(t) + \sum_{\ell=k}^{m}(\lambda + h_1 - h_\ell)P_{1,\ell}(t) \\
&= \lambda P_{1,k-1}(t) + \sum_{\ell=k}^{m-1}\lambda P_{1,\ell}(t) + \sum_{\ell=k}^{m-1}(h_1 - h_\ell)P_{1,\ell}(t) + (\lambda + h_1 - h_m)P_{1,m}(t) \\
&= \lambda P_{1,k-1}(t) + \sum_{\ell=k}^{m-1}\lambda P_{1,\ell}(t) + \lambda\sum_{\ell=k}^{m}(P_{2,\ell}(t) - P_{1,\ell-1}(t)) \\
&= \lambda\sum_{\ell=k}^{m}P_{2,\ell}(t).
\end{aligned}$$

Now, $g(t)$ can be further simplified to

$$g(t) = \lambda\left(\sum_{\ell=k}^{m}P_{2,\ell}(t)\sum_{j=1}^{m}P_{1,j}(t) - \sum_{\ell=k}^{m}P_{1,\ell}(t)\sum_{j=1}^{m}P_{2,j}(t)\right),$$

which shows that $g(t) \geq 0$ if and only if

$$P(Y_{1,t} \geq k|Y_{1,t} \in E) = \frac{\sum_{\ell=k}^{m}P_{1,\ell}(t)}{\sum_{j=1}^{m}P_{1,j}(t)} \leq \frac{\sum_{\ell=k}^{m}P_{2,\ell}(t)}{\sum_{j=1}^{m}P_{2,j}(t)} = P(Y_{2,t} \geq k|Y_{2,t} \in E),$$

with the equality holds if and only if $g(t) = 0$. Similarly, $\frac{\mathrm{d}P(Y_{\ell,t}\geq k|Y_{\ell,t}\in E)}{\mathrm{d}t} \geq 0$ if and only if

$$P(Y_{\ell,t} \geq k|Y_{\ell,t} \in E) \leq P(Y_{\ell+1,t} \geq k|Y_{\ell+1,t} \in E). \tag{A.2.2}$$

The inequality is trivial when $\ell = m - 1$ because $P(Y_{m,t} \geq k|Y_{m,t} \in E) = 1$ and $P(Y_{\ell,t} \geq k|Y_{\ell,t} \in E) \leq 1$. Recall that $Y_{\ell,t}$ is the state variable at time t assuming the starting state is ℓ at time 0. By letting $t = 0$, it could be verified that $P(Y_{\ell,t=0} \geq k|Y_{\ell,t=0} \in E) = 1$ when $\ell \geq k$ and $P(Y_{\ell,t=0} \geq k|Y_{\ell,t=0} \in E) = 0$ when $\ell < k$.

Since (A.2.2) holds when $\ell = m - 1$, we have $\frac{\mathrm{d}P(Y_{m-1,t}\geq k|Y_{m-1,t}\in E)}{\mathrm{d}t} \geq 0$, or $P(Y_{m-1,t} \geq k|Y_{m-1,t} \in E)$ is increasing with respect to t.

The next step is to prove $P(Y_{m-2,t} \geq k|Y_{m-2,t} \in E) \leq P(Y_{m-1,t} \geq k|Y_{m-1,t} \in E)$ by contradiction. Suppose

$$P(Y_{m-2,t} \geq k|Y_{m-2,t} \in E) > P(Y_{m-1,t} \geq k|Y_{m-1,t} \in E) \tag{A.2.3}$$

By (A.2.2), we have $\frac{\mathrm{d}P(Y_{m-2,t}\geq k|Y_{m-2,t}\in E)}{\mathrm{d}t} < 0$, or $P(Y_{m-2,t} \geq k|Y_{m-2,t} \in E)$ is decreasing with respect to t. Therefore, for any $t > 0$,

$$\begin{aligned}
P(Y_{m-2,t} \geq k|Y_{m-2,t} \in E) &< P(Y_{m-2,t=0} \geq k|Y_{m-2,t=0} \in E) \\
&\leq P(Y_{m-1,t=0} \geq k|Y_{m-1,t=0} \in E) \\
&< P(Y_{m-1,t} \geq k|Y_{m-1,t} \in E),
\end{aligned}$$

which conflicts with (A.2.3). The first inequality holds because $P(Y_{m-2,t} \geq k|Y_{m-2,t} \in E)$ is decreasing with respect to t. The second inequality holds because $P(Y_{\ell,t=0} \geq k|Y_{\ell,t=0} \in E)$ is

equal to 1 when $\ell \geq k$; otherwise, it is equal to 0, so that $P(Y_{m-1,t=0} \geq k | Y_{m-1,t=0} \in E)$ must be 1 if $P(Y_{m-2,t=0} \geq k | Y_{m-2,t=0} \in E) = 1$ for any k. The third inequality holds because we proved $P(Y_{m-1,t} \geq k | Y_{m-1,t} \in E)$ is an increasing function of t.

Hence, (A.2.3) is false, and it must be that

$$P(Y_{m-2,t} \geq k | Y_{m-2,t} \in E) \leq P(Y_{m-1,t} \geq k | Y_{m-1,t} \in E),$$

which shows $P(Y_{m-2,t} \geq k | Y_{m-2,t} \in E)$ is increasing with respect to t by (A.2.2). Following the same procedure recursively, it could be shown that $P(Y_{1,t} \geq k | Y_{1,t} \in E) \leq P(Y_{2,t} \geq k | Y_{2,t} \in E)$. Therefore, $g(t) \geq 0$ and $\frac{\mathrm{d}P(Y_t \geq k | Y_t \in E)}{\mathrm{d}t} \geq 0$, implying that $P(Y_t \geq k | Y_t \in E)$ is an increasing function of t. ■

A.3 Proof of Lemma 4.1

Let $f(s) = \frac{\log(ab^s + (1-a)c^s)}{s}$ *when* $s \neq 0$ *and* $f(s) = a\log(b) + (1-a)\log(c)$ *when* $s = 0$*, where* $0 \leq a \leq 1$*,* $b \geq 0$*, and* $c \geq 0$ *but* b, c *cannot be* 0 *at the same time. Then* $f(s)$ *is an increasing function with respect to* s*.*

Proof. The function $f(s)$ is a continuous function since

$$\lim_{s \to 0} \frac{\log\left(ab^s + (1-a)c^s\right)}{s} = a\log(b) + (1-a)\log(c).$$

In the first case, either b or c equals 0. Suppose $c = 0$ ($b = 0$ is the same as $c = 0$ by substituting a for $1 - a$). Then

$$f(s) = \frac{\log(ab^s)}{s} = \frac{\log(a)}{s} + b.$$

Since a is in $[0, 1]$, $\log(a) \leq 0$, and $f(s)$ is an increasing function of s,

The second case is both b and c are positive. When $s \neq 0$, the first derivative of $f(s)$ is

$$\frac{\mathrm{d}f(s)}{\mathrm{d}s} = \frac{-b^s g(a,d)}{s^2\left(ab^s + (1-a)c^s\right)},$$

where $d = \left(\frac{c}{b}\right)^s > 0$ and $g(a,d) = (a + (1-a)d)\log(a + (1-a)d) - (1-a)d\log(d)$. For any $0 \leq a_0 \leq 1$, The first partial derivatives of $g(a_0, d)$ with respect to d is

$$\frac{\partial g}{\partial d} = (1-a_0)\log\left(\frac{a_0}{d} + (1-a_0)\right).$$

When $0 < d < 1$,

$$\frac{\partial g}{\partial d} > (1-a_0)\log(a_0 + (1-a_0)) = 0,$$

and when $d > 1$,

$$\frac{\partial g}{\partial d} < (1 - a_0) \log(a_0 + (1 - a_0)) = 0.$$

So, for any $a_0 \in [0, 1]$, $\max g(a_0, d) = g(a_0, 1) = 0$, yielding $g(a_0, d) \leq 0$ for $d > 0$. Since a_0 is arbitrary, we have $g(a, d) \leq 0$ for $0 \leq a \leq 1$ and $d > 0$. Hence, $\frac{\mathrm{d}f(s)}{\mathrm{d}s} \geq 0$ and $f(s)$ is an increasing function with respect to s. ■

Appendix B

Selection of typical lifespan ψ for a PTAM with a large number of states m

This Appendix relates to Subsection 3.4.4 of Chapter 3. It addresses the issue that poor identifiability may occur if ψ and m are not appropriate. Let T be the maximum lifetime, which is not necessarily equal to the lifespan ψ. It is more reasonable to assume $T \leq \psi$.

B.1 When ψ is much greater than T

Let us assume $h_1 \leq h_m$. Suppose ϵ is a fixed calculation tolerance. When $\psi > T\frac{\log(\frac{h_m}{h_1})}{\log(1+\epsilon)}$, we have

$$\left(\frac{h_m}{h_1}\right)^{\frac{T}{\psi}} < 1 + \epsilon.$$

By Theorem 4.6, the resulting hazard function $h(t)$ asymptotically converges to $h_1\left(\frac{h_m}{h_1}\right)^{\frac{t}{\psi}}$, and

$$h_1\left(\frac{h_m}{h_1}\right)^{\frac{t}{\psi}} < h_1\left(\frac{h_m}{h_1}\right)^{\frac{T}{\psi}} < h_1(1+\epsilon),$$

as $m \to \infty$. For some small ϵ, the numerical values of $h(t)$ could be made close to h_1, resulting to the same distribution as an exponential distribution with rate h_1. Let $f_m(t)$ be the pdf of an $m-$state PTAM. Then, for any $\xi > 0$, there is an M such that for any $m > M$,

$$\left|f_m(t) - h_1 e^{-h_1 t}\right| < \xi.$$

Therefore, for any $m > M$, the numerical values for $f_m(t)$ are identical to $h_1 e^{-h_1 t}$, in which m, h_m, s and ψ are non-identifiable based on a numerical evaluation.

On the other hand, when m is too small compared to the value of ψ, it is typical to have $\frac{m}{\psi} \approx 0$, because

$$\psi > T\frac{\log\left(\frac{h_m}{h_1}\right)}{\log(1+\epsilon)}.$$

The right-hand side is typically much larger than m due to the fact that $\log(1+\epsilon) \approx \epsilon$ is very small. The probability of absorption conditional on one transition occurring in state 1 is

$$P(Y_{t+\Delta} = m+1 | Y_t = 1, Y_{t+\Delta} \neq 1) = \frac{h_1}{\lambda + h_1} = \frac{h_1}{m/\psi + h_1} \approx 1,$$

where Δ is a small fraction of time. Hence, each individual starting in state 1 at time 0 moves to the absorbing state after one transition with probability almost equal to 1. The resulting lifetime distribution is an exponential distribution with rate h_1 with probability almost 1, in which, once again, m, h_m, s, and ψ are non-identifiable based on numerical calculation.

Under both situations considered above, the numerical evaluation for the resulting lifetime distribution is identical to an exponential distribution with rate h_1. Given any lifetimes, the numerical value of the likelihood will be dominated by h_1. The likelihood is not sensitive to the other parameter values, in which it is impossible to estimate the true values of (h_m, s, ψ, m) by the MLE method. Identifiability and estimability are extremely poor. To avoid these scenarios, ψ should not be much greater than T in applications.

B.2 When ψ is not much greater than T

The limit of the resulting survival function as $m \to \infty$ is

$$\lim_{m\to\infty} S(t) = \lim_{m\to\infty} e^{-\int_0^t h(u)\mathrm{d}u} = e^{-\int_0^t g(u)\mathrm{d}u}.$$

When $s \neq 0$,

$$\begin{aligned}\lim_{m\to\infty} S(t) &= e^{-\int_0^t \left(\left(h_m^s - h_1^s\right)\frac{u}{\psi} + h_1^s\right)^{1/s}\mathrm{d}u} \\ &= e^{-\frac{1}{ab}\left((at+c)^b - c^b\right)},\end{aligned}$$

where $a = \frac{h_m^s - h_1^s}{\psi}$, $b = \frac{s+1}{s}$ and $c = h_1^s$.

For a set of parameter values (h_1, h_m, s, ψ), we can find another set of parameter values (h_1', h_m', s', ψ') leading to the same value of (a, b, c). This is achievable by setting $h_1' = h_1$, $s' = s$, $\psi' \neq \psi$ and $h_m' = \left(\left(h_m^s - h_1^s\right)\frac{\psi'}{\psi} + h_1^s\right)^{1/s}$. The limit of the resulting survival functions are identical under two different sets of parameter values as $m \to \infty$, in which the parameters (h_1, h_m, s, ψ) are not identifiable, but the parameters $\left(h_1, s, \frac{h_m^s}{\psi}\right)$ are. To mitigate the non-identifiable issue, one should fix/estimate ψ before estimating (h_1, h_m, s).

To validate the formula for the limit of the resulting survival function, we provide a numerical example illustrated in Figure B.1.

Example B.1. One can easily calculate the resulting survival function of an $m-$state PTAM given the parameter values $h_1 = 0.001$, $h_m = 1.275$, $s = -0.073$, $\psi = 55$. Consequently, the values of a, b, c are -0.0122, -12.6986 and 1.6558, respectively. Let $S_1(t)$ be the resulting survival function of an $m-$state PTAM and $S_2(t) = e^{-\frac{1}{ab}((at+c)^b - c^b)}$. The vertical axis refers to the difference of $\max_t |S_1(t) - S_2(t)|$ and the horizontal axis represents the value for m. It is clear that $\max_t |S_1(t) - S_2(t)|$ converges to 0 as $m \to \infty$ in Figure B.1. △

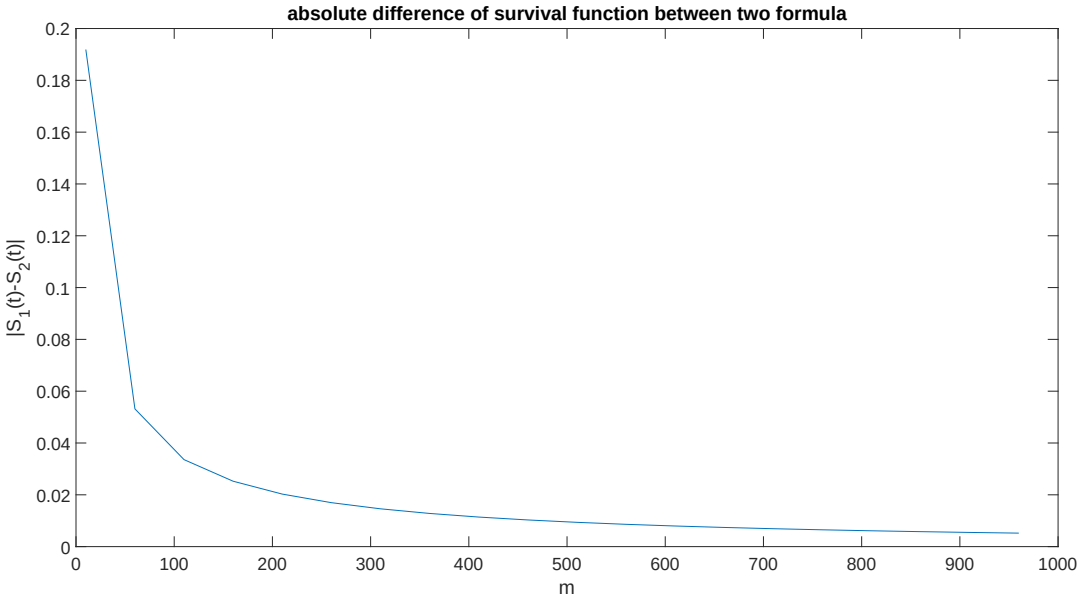

Figure B.1: Validating numerically the formula for the limit of the resulting survival function when $s \neq 0$.

When $s = 0$,

$$\lim_{m\to\infty} S(t) = e^{-\int_0^t h_1^{(1-u/\psi)} h_m^{u/\psi} \mathrm{d}u} = e^{-\frac{1}{a'b'}\left(e^{c't}-1\right)},$$

where $a' = \frac{1}{h_1\psi}$, $b' = \log(h_m/h_1)$ and $c' = \frac{1}{\psi}\log(h_m/h_1)$. There are two different sets of parameter values (h_1, h_m, ψ) and (h_1', h_m', ψ') such that the value of $(a'b', c')$ are the same, where $\psi' \neq \psi$, $h_1' = h_1$ and $h_m' = h_1 (h_m/h_1)^{\frac{\psi'}{\psi}}$. The resulting survival functions under the two sets of parameter values are identical. One may verify that the parameters (h_1, h_m, ψ) are non-identifiable, but the parameters $\left(h_1, (h_m/h_1)^{1/\psi}\right)$ are identifiable. To mitigate the non-identifiability problem, once again, ψ must be fixed or estimated before estimating (h_1, h_m, s)

Example B.2. This example aims to verify the formula for the limit of the resulting survival function. Let $S_2(t) = e^{-\frac{1}{a'b'}\left(e^{c't}-1\right)}$ and $S_1(t)$ be the resulting survival function of an $m-$state PTAM with parameter values $h_1 = 0.01$, $h_m = 1$, $s = 0$, and $\psi = 55$. The corresponding a', b', c' are 1.8182, 4.6052 and 0.0837, respectively. It is clear that $\max_t |S_1(t) - S_2(t)|$ also approaches 0 as $m \to \infty$ in Figure B.2. △

B.3 Consideration summary

We demonstrated that the PTAM's identifiability and estimability are poor when ψ is too large. To improve the model identifiability and the model estimability, the value for ψ cannot be too big.

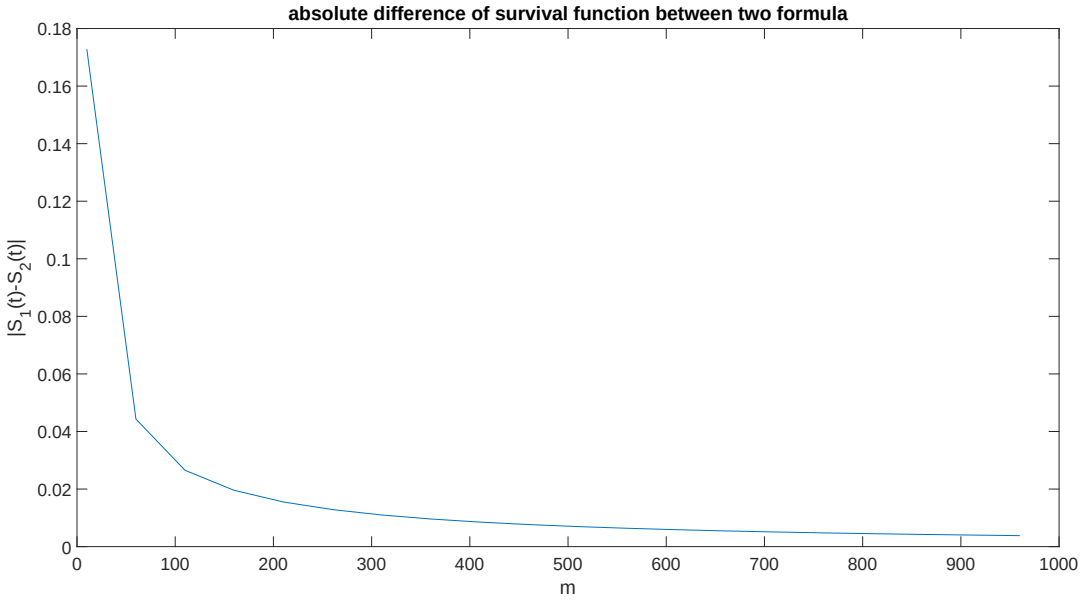

Figure B.2: A Verification of the formula for the limit of the resulting survival function when $s = 0$.

When ψ is a free parameter with a suitable domain (the upper bound cannot be too large), there are at least two different sets of parameter values such that the resulting lifetime distributions are identical numerically when m is too large. Under such a situation, the estimates of h_1, h_m, s and ψ are not unique. To circumvent this issue, ψ must be fixed or estimated before estimating h_1, h_m and s. Ideally, the parameter ψ should be estimated from other health-related information rather than lifetime data only; see Govorun et al. (2018).

The value of m determines the state variability. The bigger the value of m is, the less state variability is for the physiological age t. The value of m can be either estimated from health information or determined by the required state variability from prior knowledge.

Appendix C

Experiments showing the flexibility of the resulting distribution from a PTAM

This Appendix details an analysis of the PTAM's flexibility as described in Chapter 3. This is based on the concept of goodness-of-fit to the target distribution by the calibrated PTAM.

C.1 Some distributions for lifetime modelling

The Gamma distribution has two parameters: α and β. Its associated pdf is

$$f(t) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t),\ \alpha > 0,\ \beta > 0,$$

where $\Gamma(x) = \int_{z=0}^{\infty} z^{x-1} e^{-z} dz$ is the gamma function. When $\alpha > 1$, the Gamma distribution has an increasing hazard rate, whilst the hazard rate for a Gamma distribution when $\alpha < 1$ is decreasing. When $\alpha = 1$, the Gamma distribution reduces to an exponential distribution with a rate $\lambda = \beta$. Since the family of Gamma distributions is rich and capable of reproducing a variety of distribution shapes, it is extensively used in lifetime modelling. The Gamma distribution is one of the potential lifetime distributions utilised to capture lifetime processes of humans and some other living things.

The Weibull distribution has two parameters: k and λ. Its associated pdf is

$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{k}\right),\ \lambda > 0,\ k > 0.$$

When $k > 1$, the Weibull distribution has an increasing hazard rate, whilst the hazard rate is decreasing when $k < 1$. A Weibull distribution with $k = 1$ reduces to an exponential distribution with rate λ. The Weibull distribution is popular in survival analysis and affords a sound basis for modelling lifetime distributions.

The Pareto distribution has two parameters: k and σ. Its associated survival function is

$$S(t) = \left(1 + \frac{kt}{\sigma}\right)^{-\frac{1}{k}},\ k > 0,\ \sigma > 0.$$

It is well-known that Pareto distributions are heavy-tailed.

A convolution of two exponential distributions has two parameters, and the random variable that represents a convolution of two exponential distributions is

$$Y = X_1 + X_2,$$

where X_1, X_2 follow the exponential distributions with rates λ_1 and λ_2, respectively. The convolutions of different distributions have physical interpretation, i.e., the process can be decomposed into several components and the time spent on each component follows a certain distribution. In particular, the physical interpretation for the convolution mechanism is consistent with our cognizance of ageing, being a progressive process. Furthermore, it is often assumed that the time spent in each component follows an exponential distribution, because an exponential distribution matches most observed data. Therefore, a convolution of exponential distributions is a further extension capable of affording more flexibility.

A convolution of two Weibull distributions has four parameters, and the random variable representing a convolution of two Weibull distributions is

$$Z = W_1 + W_2,$$

where W_1 and W_2 follow the Weibull distribution with parameter (λ_1, k_1) and (λ_2, k_2) respectively. This is a simple extension of the previous example by changing the distributions of each component, similar to Huzurbazar (1999). Since the Weibull distribution is utilised in survival analysis, it is plausible for the ageing process that the time spent in each component follows a Weibull distribution as well. It is then worth exploring whether one can use the PTAM to approximate well a convolution of Weibull distributions.

The Gompertz-Makeham distribution has three parameters, and its associated hazard rate is

$$h(t) = \zeta + \xi \exp(\lambda t),\ \zeta \geq -\xi,\ \xi > 0,\ \lambda > 0,$$

where ξ is the growth rate of mortality and ζ represents the background mortality with additional constraint $\zeta > 0$.

The Gompertz-Makeham distribution were proposed by Makeham (1860) and Gompertz (1825) , and it assumes a shifted exponential increasing hazard rate with respect to age. Such a distribution fits human mortality rates from age 30 to 80 well. When we calibrate the PTAM with human lifetime observations only, the PTAM is essentially approximating the Gompertz-Makeham distribution.

The Makeham's second extension of the Gompertz distribution has four parameters, and its survival function is

$$S(t) = \exp\left(-\xi(\exp(\lambda t) - 1) - \xi\theta\lambda t - \xi\alpha(\lambda t)^2\right),\ \xi > 0,\ \lambda > 0,\ \theta > 0,\ \alpha > 0.$$

Makeham (1890) extended the assumptions on the deferential equation for the hazard rate to higher order derivatives, because Makeham (1890) observed that the third differences of empirical hazard rates were much closer to a geometrical progression. Once again, it is likely to observe human lifetimes following the extended Gompertz distribution.

For each example, we select appropriate parameter values to achieve a particular shape of probability density function or hazard function. Once the parameter values are chosen, the target lifetime distributions are fixed and known.

C.2 Calibration criterion

We shall use the Kullback–Leibler divergence (Kullback and Leibler, 1951; Kullback, 1997) to summarise the information loss when using the PTAM to approximate the target lifetime distribution, which is known. When calibrating the proposed PTAM, we minimise the Kullback–Leibler divergence to ensure minimal lost information.

Definition C.1. *For continuous distributions P and Q with respective pdf' $p(x)$ and $q(x)$ defined on a sample probability space, the Kullback-Leibler divergence of Q from P is*

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x.$$

Usually, the distribution P represents the actual model, and the distribution Q represents the distribution used to approximate P. In our case, the distribution P is the target lifetime distribution in each example, and the distribution Q is the proposed PTAM.

A good property for the Kullback-Leibler divergence is its non-negative value, which means

$$D_{KL}(P \parallel Q) \geq 0,$$

with equality if and only if $P = Q$. This result is known as Gibbs' inequality (Gibbs, 1902). One can interpret the divergence value as the extra information required for Q to represent P. Therefore, the smaller the divergence, the less extra information is required when using Q to approximate P and the better goodness of fit is for Q. Furthermore, by its non-negative property, the closer the divergence value to 0, the better the approximation.

In our case, we calibrate the proposed PTAM by minimizing the Kullback-Leibler divergence of the proposed PTAM from the target lifetime distribution. The analytical formula for the PTAM's resulting distribution is quite involved. However, the resulting distribution is fairly easy to evaluate numerically for a given set of parameter values. Thus, we numerically search for the minimum of D_{KL}(lifetime distribution || proposed PTAM), where $p(x)$ is the pdf for the target lifetime distribution, $q(x)$ is the pdf for the PTAM, and

$$D_{KL}(\text{lifetime distribution} \parallel \text{proposed PTAM}) = \int_{0}^{+\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x.$$

The numerical search utilises the optimisation function **fmincon** in MATLAB. The calibration process is similar to previous model calibration in the Le Bras model simulation, except for switching the optimisation criterion from Negative Log-Likelihood to the Kullback-Leibler divergence. The Tail Value at Risk with risk level $0.999(\mathrm{TVaR}_{0.999}(T))$ can be calculated in each example, and the calculated value is used to estimate ψ. The estimated values for the other four parameters are the values that minimise the divergence.

C.3 Estimates of parameters for each distribution

In the Tables C.1–C.10, h_1, h_m, λ, s, m are the estimated values, and D_{KL} is the numerical value for the Kullback–Leibler divergence. Some D_{KL}s are negative because we use the following

approximation to calculate the KL divergence numerically:

$$\int_0^\infty p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x \approx \sum_{i=1}^{N} p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right)\Delta x,$$

where $x_{i+1} - x_i = \Delta x$ is a small number, $x_1 = 0$ and the survival probability to x_N, $S(x_N)$, is very small.

D_{KL}	h_1	h_m	λ	s	m
0.1099	1.51×10^{-11}	99.1353	0.3495	0.2323	10
0.0477	2.12×10^{-12}	99.9896	0.3845	0.2339	11
0.0142	3.27×10^{-11}	99.8473	0.4195	0.2351	12
0.0037	3.12×10^{-9}	99.7605	0.4542	0.2349	13
0.0109	8.39×10^{-10}	83.8905	0.4893	0.2437	14
0.0273	2.90×10^{-8}	21.8657	0.5244	0.3146	15
0.0438	3.98×10^{-10}	10.2016	0.5594	0.3759	16

Table C.1: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Gamma distribution with $\alpha = 4$ and $\beta = 0.5$.

D_{KL}	h_1	h_m	λ	s	m
0.0018	1.2273	0.3788	0.1575	−0.7211	2
−0.0022	3.6463	0.4414	20.2472	−4.4077	257
−0.0022	3.8817	0.4188	30.2441	−3.9887	384
−0.0021	3.9809	0.4045	40.3026	−3.7623	512
−0.0021	3.9919	0.3876	60.5440	−3.5211	768
−0.0021	3.9288	0.3782	80.6907	−3.4003	1024

Table C.2: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Gamma distribution with $\alpha = 0.5$ and $\beta = 0.5$.

D_{KL}	h_1	h_m	λ	s	m
20.7744	0.4994	2.0686	0.3365	−0.4522	2
0.0071	0.0257	2.0078	86.3166	1.9922	513
0.0029	0.0161	2.0065	129.2587	1.9920	768
0.0007	0.0088	2.0056	161.5810	1.9925	960
0.0002	0.0063	2.0049	169.7268	1.9927	1008
8.55×10^{-5}	0.0061	2.0052	171.7103	1.9927	1020
6.40×10^{-5}	0.0061	2.0059	171.9036	1.9927	1022
5.25×10^{-5}	0.0059	2.0052	172.2107	1.9927	1023
4.19×10^{-5}	0.0059	2.0059	172.2599	1.9927	1024

Table C.3: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Weibull distribution with $\lambda = 1.5$ and $k = 1.5$.

D_{KL}	h_1	h_m	λ	s	m
25.6238	8.15×10^{-7}	1.4283	0.6619	0.5127	2
0.0063	2.69×10^{-6}	15.0291	169.8246	0.2261	513
0.0034	9.57×10^{-7}	14.4057	254.2103	0.2328	768
0.0027	6.72×10^{-7}	14.2276	296.6121	0.2346	896
0.0024	6.05×10^{-7}	14.1681	317.7771	0.2353	960
0.0023	4.84×10^{-7}	14.1096	333.7019	0.2360	1008
0.0022	4.93×10^{-7}	14.1064	337.6457	0.2361	1020
0.0022	5.15×10^{-7}	14.1049	338.9924	0.2359	1024

Table C.4: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Weibull distribution with $\lambda = 2$ and $k = 5$.

D_{KL}	h_1	h_m	λ	s	m
0.2133	0.3426	0.0424	0.0157	−0.8731	2
−1.0749	0.8666	0.0341	2.0305	−2.1191	258
−1.0733	0.8849	0.0327	2.5184	−2.0471	320
−1.0654	0.9063	0.0301	4.0294	−1.9219	512
−1.0579	0.9101	0.0285	6.0441	−1.8466	768
−1.0536	0.9092	0.0277	8.0589	−1.8089	1024

Table C.5: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Weibull distribution with $\lambda = 2$ and $k = 0.5$.

D_{KL}	h_1	h_m	λ	s	m
0.0065	0.1851	0.0704	0.0201	−0.1012	2
−0.0014	0.1962	0.0419	2.5954	−0.9922	258
−0.0013	0.1962	0.0418	3.2190	−0.9874	320
−0.0013	0.1963	0.0416	5.1505	−0.9799	512
−0.0013	0.1963	0.0415	7.7257	−0.9758	768
−0.0013	0.1963	0.0414	10.3009	−0.9737	1024

Table C.6: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Pareto distribution with $k = 0.2$ and $\sigma = 5$.

D_{KL}	h_1	h_m	λ	s	m
0.0905	0.1876	0.6259	0.0178	0.1853	2
4.05×10^{-13}	0.0046	0.3549	0.5775	21.9696	65
1.36×10^{-12}	0.0057	0.4277	0.7108	7.8810	80
1.20×10^{-12}	0.0048	0.4824	0.8529	5.6336	96
5.58×10^{-5}	4.10×10^{-8}	0.5622	1.1461	4.1867	129
0.0010	1.67×10^{-9}	0.6755	1.7058	3.2584	192

Table C.7: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the convolution of two exponential distributions with $\lambda_1 = 0.6$ and $\lambda = 0.3$.

D_{KL}	h_1	h_m	λ	s	m
0.0009	7.29×10^{-12}	19.1858	2.1776	0.3584	17
0.0007	4.64×10^{-14}	15.7827	2.3057	0.3793	18
0.0006	4.24×10^{-14}	13.4428	2.4338	0.3982	19
0.0006	8.85×10^{-12}	11.7538	2.5619	0.4154	20
0.0007	7.52×10^{-11}	10.5362	2.6899	0.4303	21
0.0007	7.06×10^{-13}	9.5865	2.8181	0.4441	22
0.0008	8.81×10^{-11}	8.8437	2.9461	0.4564	23
0.0009	5.84×10^{-11}	8.2471	3.0742	0.4677	24

Table C.8: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the convolution of two Weibull distributions with $\lambda_1 = 2, k_1 = 1, \lambda_2 = 1$ and $k_2 = 1.3$.

D_{KL}	h_1	h_m	λ	s	m
1.8724	2.62×10^{-8}	0.0263	0.0185	0.5925	2
5.91×10^{-5}	6.22×10^{-5}	17.9815	4.7413	−0.1775	513
6.53×10^{-6}	1.69×10^{-5}	2.3699	7.0981	−0.0821	768
4.19×10^{-6}	1.60×10^{-5}	2.1037	8.2811	−0.0762	896
3.16×10^{-6}	1.54×10^{-5}	1.9653	9.1684	−0.0726	992
2.88×10^{-6}	1.51×10^{-5}	1.9237	9.4642	−0.0713	1024

Table C.9: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ under each fixed m for the Gompertz-Makeham model with $\zeta = 2.2 \times 10^{-5}, \xi = 2.7 \times 10^{-6}$ and $\lambda = \log 1.125$.

D_{KL}	h_1	h_m	λ	s	m
0.2031	0.0465	0.2800	0.0938	0.8680	2
0.0002	0.0301	1.8184	24.0717	−0.0478	513
9.10×10^{-5}	0.0301	1.7668	36.0371	−0.0411	768
6.96×10^{-5}	0.0301	1.7529	42.0433	−0.0393	896
5.55×10^{-5}	0.0301	1.7439	47.2987	−0.0381	1008
5.46×10^{-5}	0.0301	1.7441	47.6741	−0.0381	1016
5.37×10^{-5}	0.0301	1.7436	48.0495	−0.0381	1024

Table C.10: Estimated values of $(h_1, h_m, \lambda = m/\psi, s)$ for each fixed m under the Markham's second extension of the Gompertz distribution with $\xi = 0.1, \lambda = 0.2, \theta = 0.3$ and $\alpha = 0.4$.

C.4 Fitted Results

In summary, the fitted results for each example are promising in terms of the pdf, survival function and hazard rate. For instance, the Kolmogorov–Smirnov statistic is 0.0032 with p-value equal to 0 for Figure C.1. Thus, the fitted PTAM is not different from the true distribution. The results of the Kolmogorov-Smirnov tests for Figures C.2–C.10 are similar, and thus, the PTAM provides an excellent approximation. The results are promising in approximating the Gompertz model, including its extended version. These demonstrate the PTAM's goodness of fit for human lifetime data. The calibrated PTAM is validated by checking its resulting lifetime distribution, and the proposed PTAM successfully captures the human mortality trend.

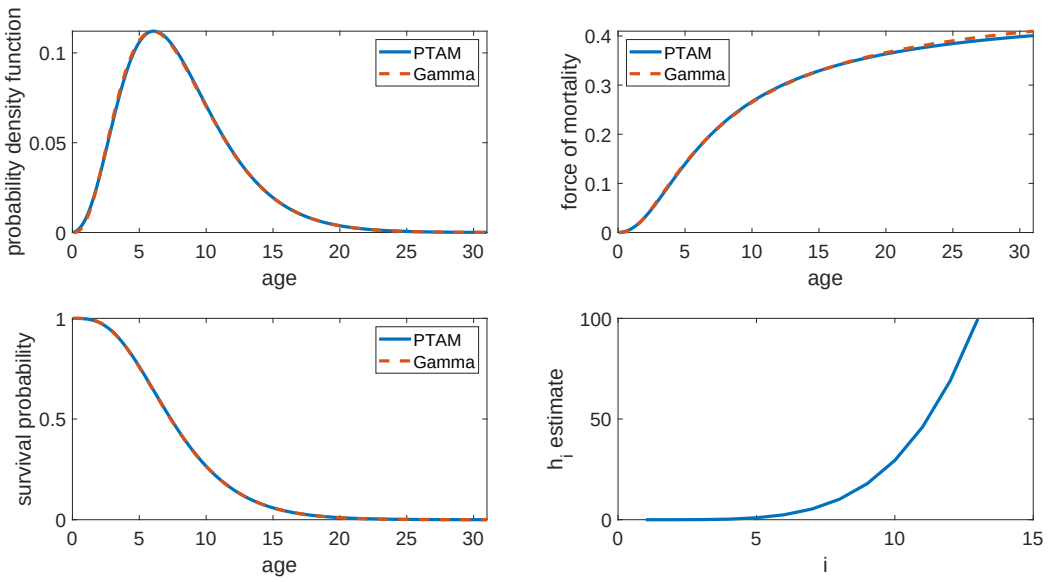

Figure C.1: Proposed PTAM approximating a Gamma distribution with $\alpha = 4$ and $\beta = 0.5$.

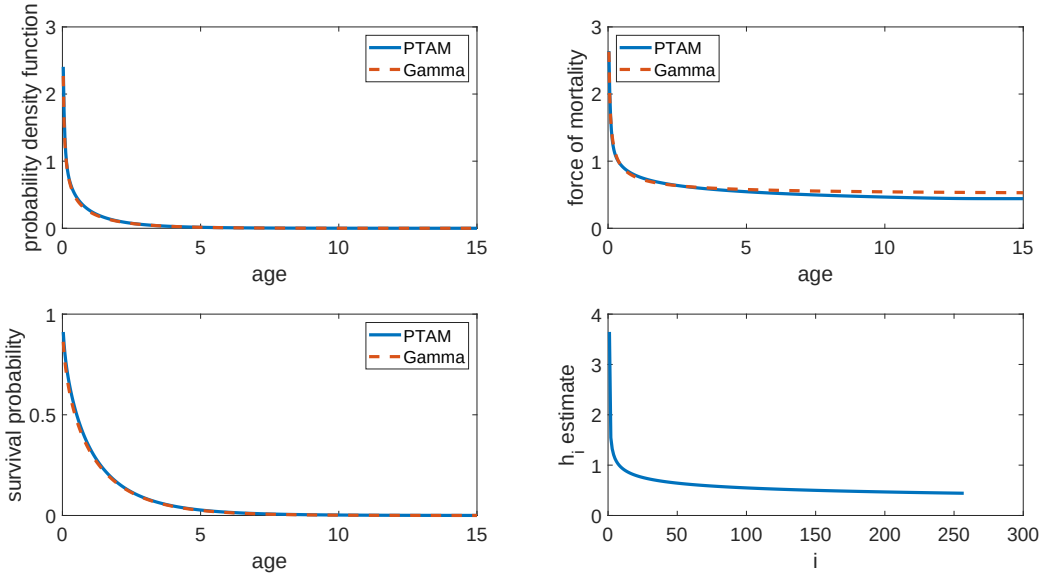

Figure C.2: Proposed PTAM approximating a Gamma distribution with $\alpha = 0.5$ and $\beta = 0.5$.

Meanwhile, it is plausible to use the calibrated model to analyse the human ageing mechanism. For lifetimes that follow the distributions highlighted in Figures C.1 - C.10, our results suggest that the PTAM could be applied to gain insights on the embedded ageing process. For cases with decreasing hazard rate, there are always distinct tails when using the proposed PTAM to approximate the target lifetime distributions. Nevertheless, the fitted results are reasonably good for most values in the domain. The areas where the PTAM cannot approximate well are negligible because the probability in such areas is close to 0.

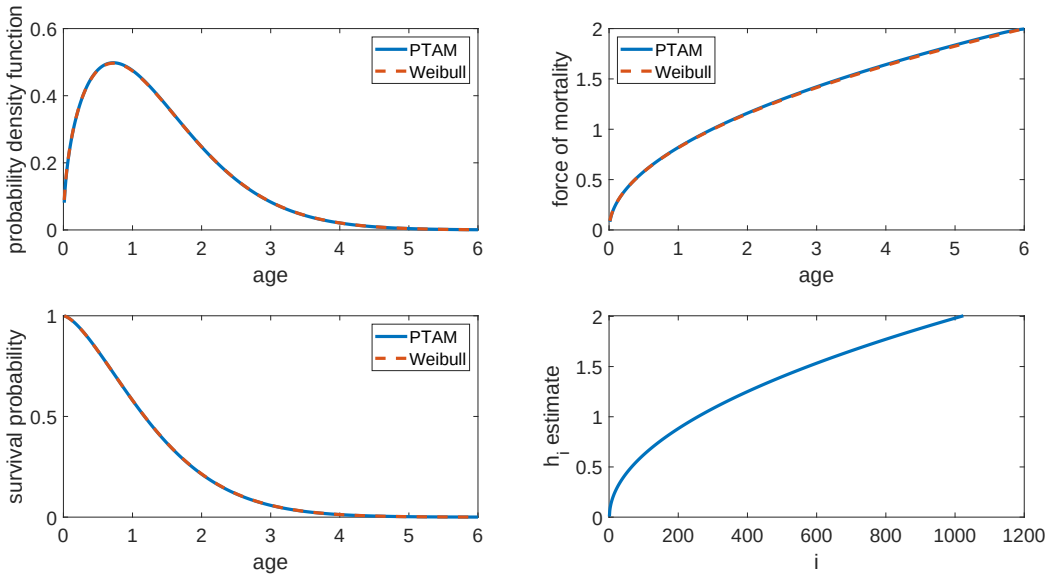

Figure C.3: Proposed PTAM approximating a Weibull distribution with $\lambda = 1.5$ and $k = 1.5$.

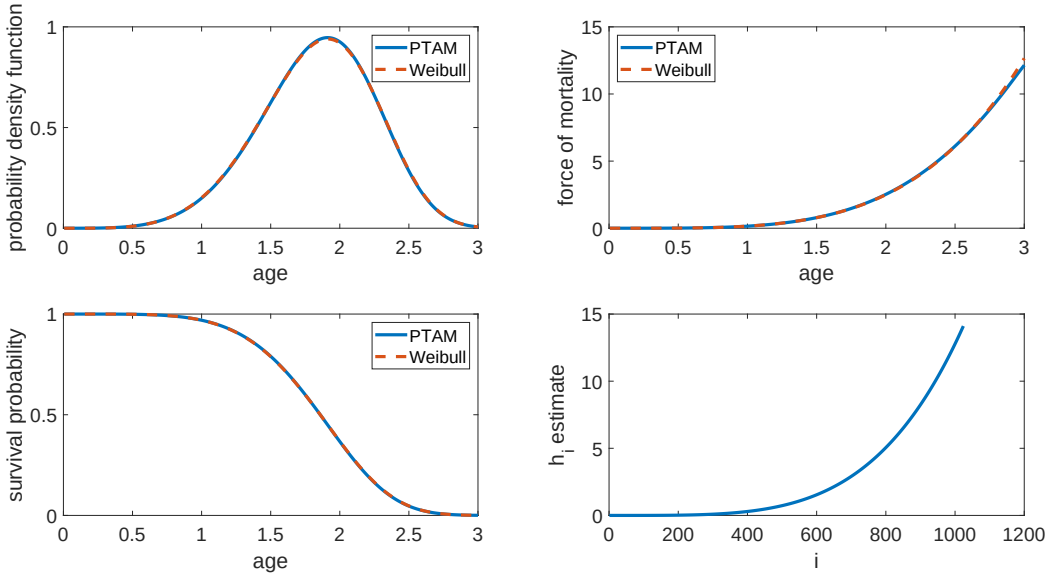

Figure C.4: Proposed PTAM approximating a Weibull distribution with $\lambda = 2$ and $k = 5$.

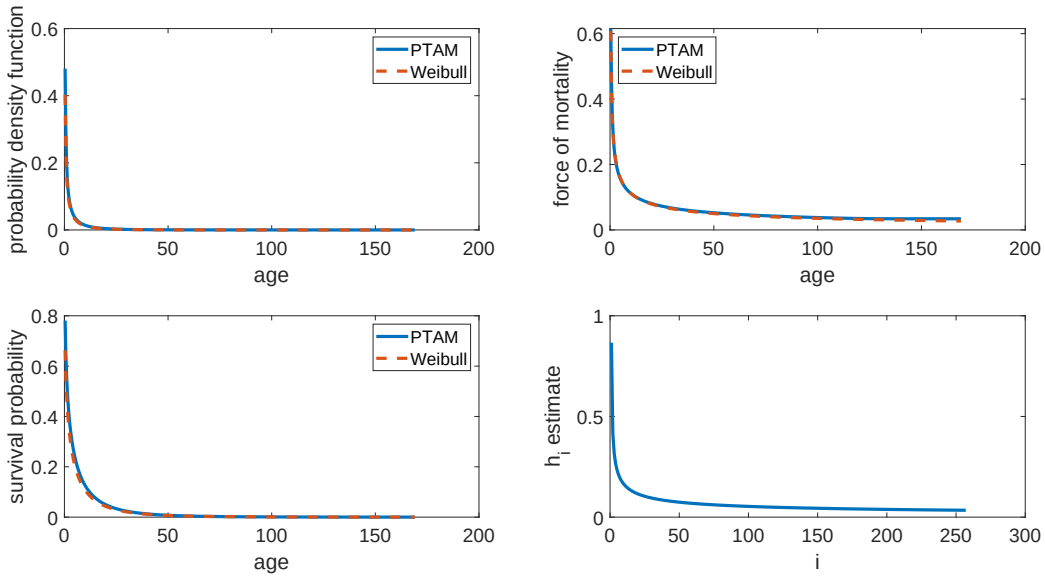

Figure C.5: Proposed PTAM approximating a Weibull distribution with $\lambda = 2$ and $k = 0.5$.

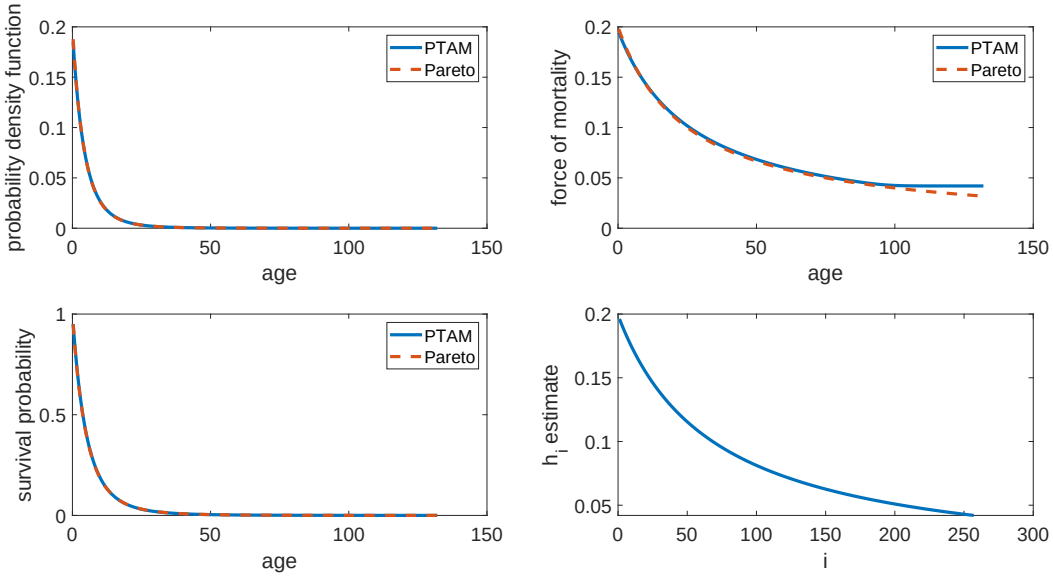

Figure C.6: Proposed PTAM approximating a Pareto distribution with $k = 0.2$ and $\sigma = 5$.

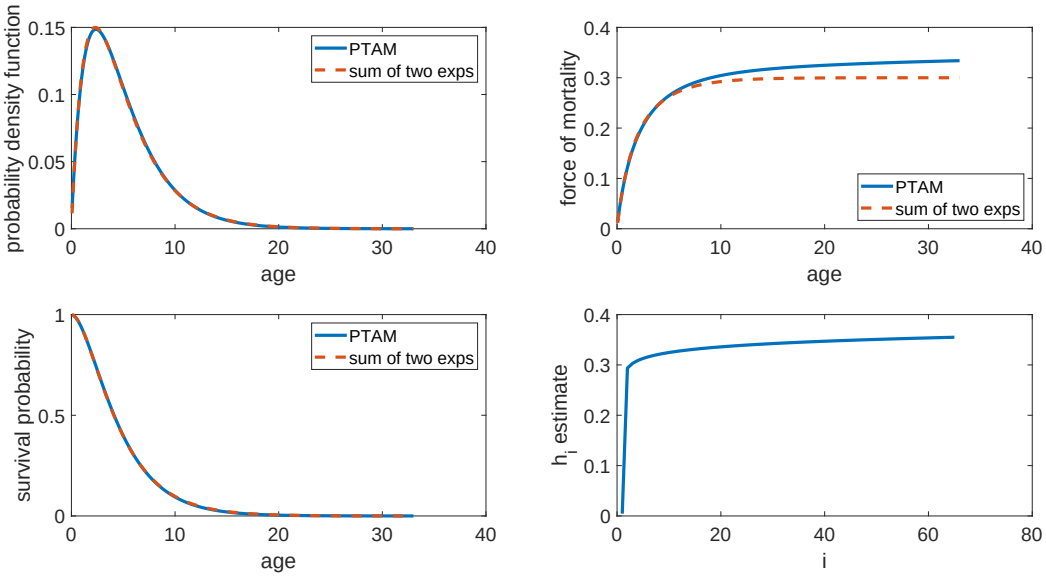

Figure C.7: Proposed PTAM approximating a convolution of two exponential distributions with $\lambda_1 = 0.6$ and $\lambda = 0.3$.

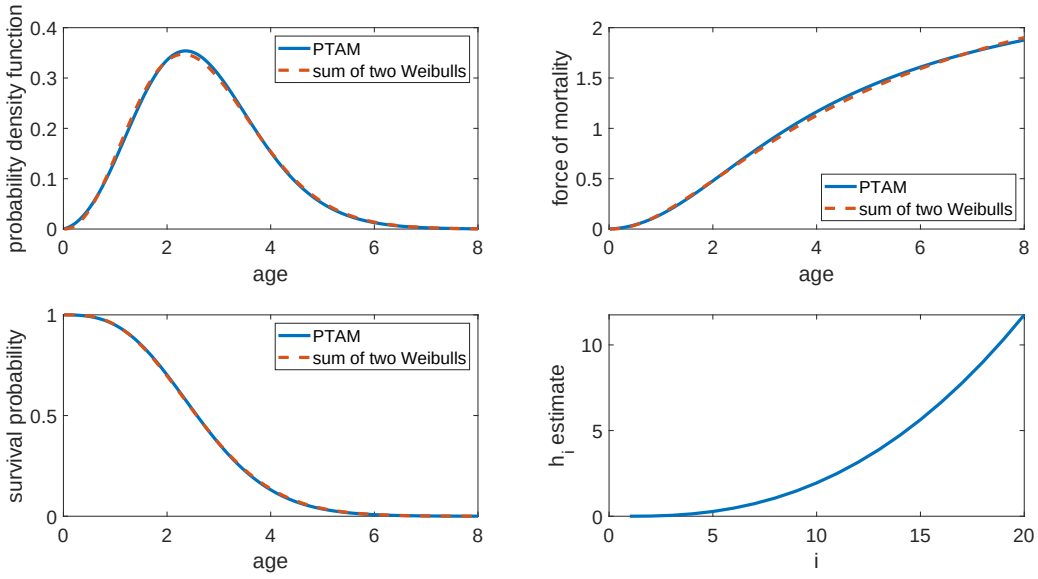

Figure C.8: Proposed PTAM approximating a convolution of two Weibull distributions with $\lambda_1 = 2, k_1 = 1, \lambda_2 = 1$ and $k_2 = 1.3$.

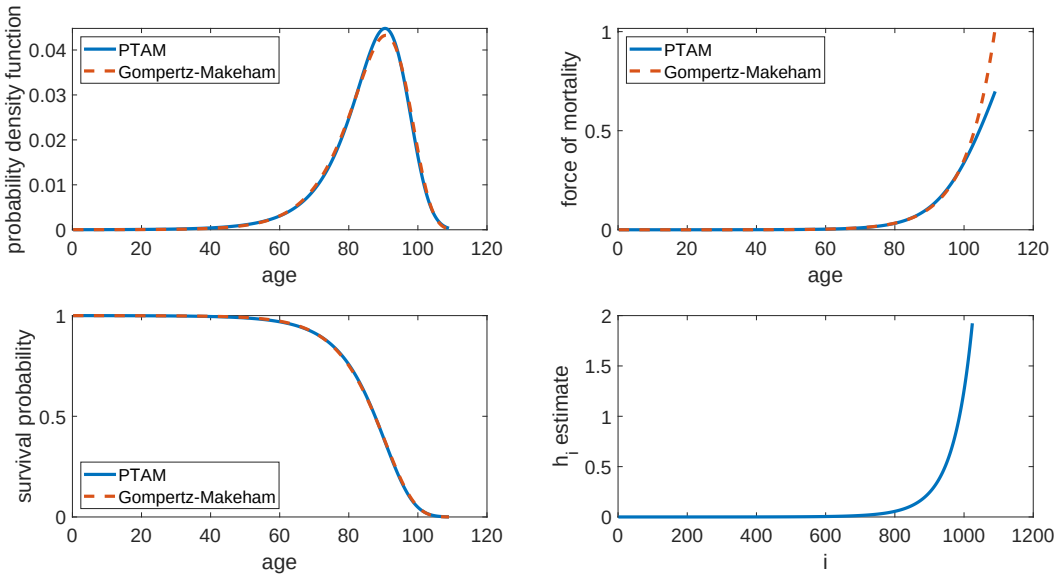

Figure C.9: Proposed PTAM approximating a Gompertz-Makeham Model with $\zeta = 2.2 \times 10^{-5}, \xi = 2.7 \times 10^{-6}$ and $\lambda = \log 1.125$.

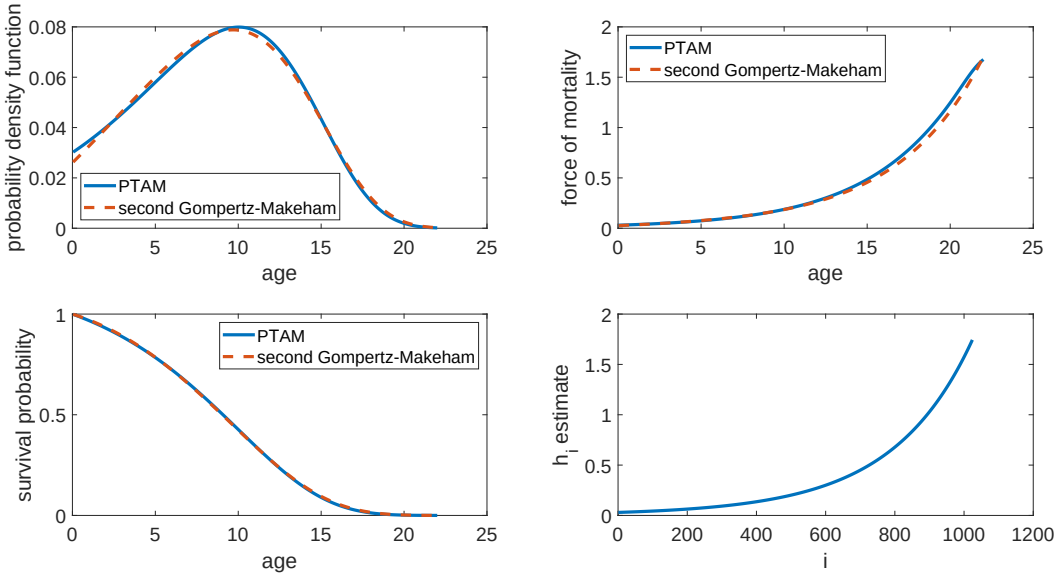

Figure C.10: Proposed PTAM approximating a Makeham's second extension of the Gompertz distribution with $\xi = 0.1, \lambda = 0.2, \theta = 0.3$ and $\alpha = 0.4$.

Appendix D

A proposed algorithm

This is the proposed algorithm written in MATLAB in connection with the proposed algorithm in Subsection 5.3.3 of Chapter 5. This algorithm can efficiently calculate the likelihood and pdf of the proposed PTAM.

```
%% y is a one by n vector and it contains the observation in descending ...
    order
%% delta is a one by n vector, delta=1 is exact death time, delta=0 is ...
    sensor data
%% m is the total number of state
%% lam is a one by one numerical value and it is the constant lambda in ...
    the PTAM
%% epsilon is the numerical tolerance


function [Likelihood]=PTAM_likelihood(y,m,h1,hm,lam,s,delta,epsilon)
%total number of state m
nn=(1:m)';
% hi is the dying rate in the PTAM, it can be modified to Coxian model ...
    with specifying each value of hi. Here use the proposed structure
if abs(s)<10^-3
     hi=h1.^((m-nn)/(m-1)).*hm.^((nn-1)/(m-1));
else
     hi=(h1^s*(m-nn)/(m-1)+hm^s*(nn-1)/(m-1)).^(1/s);
end
%%%
%lambda is a m-1 vector with each value is equal to lambda_i for the PTAM
lambda=lam*ones((m-1),1);
maxx=max(hi+[lambda;0]);
P=zeros(m,1);
P(1)=1-(lambda(1)+hi(1))/maxx;
P(2)=lambda(1)/maxx;
% ex is a n by m vector with the ith row element is the probability in ...
    each state for the ith observation
ex=zeros(size(y,1),m);
ex(:,1)=poisspdf(0,maxx*y(:,1));
w1=[0;lambda]/maxx; w2=1-([lambda;0]+hi)/maxx;
%N is the truncation point for the infinite sum
```

```
N=poissinv(1-epsilon,maxx*y(end,1));
for N_i=1:N
ex=ex+poisspdf(N_i,maxx*y(:,1))*P';
P=w1.*[0;P(1:end-1)]+w2.*P;
end
% Likelihood is a n by one vector. The ith element is the likelihood ...
   for the ith element
Likelihood=ex*hi.*delta+sum(ex,2).*(1-delta);
end
```

Curriculum Vitae

Name:	Boquan Cheng
Post-Secondary Education and Degrees:	PhD in Statistical and Actuarial Sciences, 2016 - 2021 The University of Western Ontario, London, ON, Canada MSc in Statistical and Actuarial Sciences, 2015 - 2016 The University of Western Ontario, London, ON, Canada BSc in Mathematics and Applied Mathematics, 2011 - 2015 South China University of Technology, China
Honours and Awards:	Towers Watson Graduate Scholarship, awarded for academic achievement 2016
Related Work Experience:	Research and Teaching Assistant, 2016 - 2021 The University of Western Ontario

Publications:

Cheng,B., Jones, B., Liu, X., and Ren, J. (2021). The mathematical mechanism of biological aging. *North American Actuarial Journal*, accepted:DOI:10.1080/10920277.2020.1775654