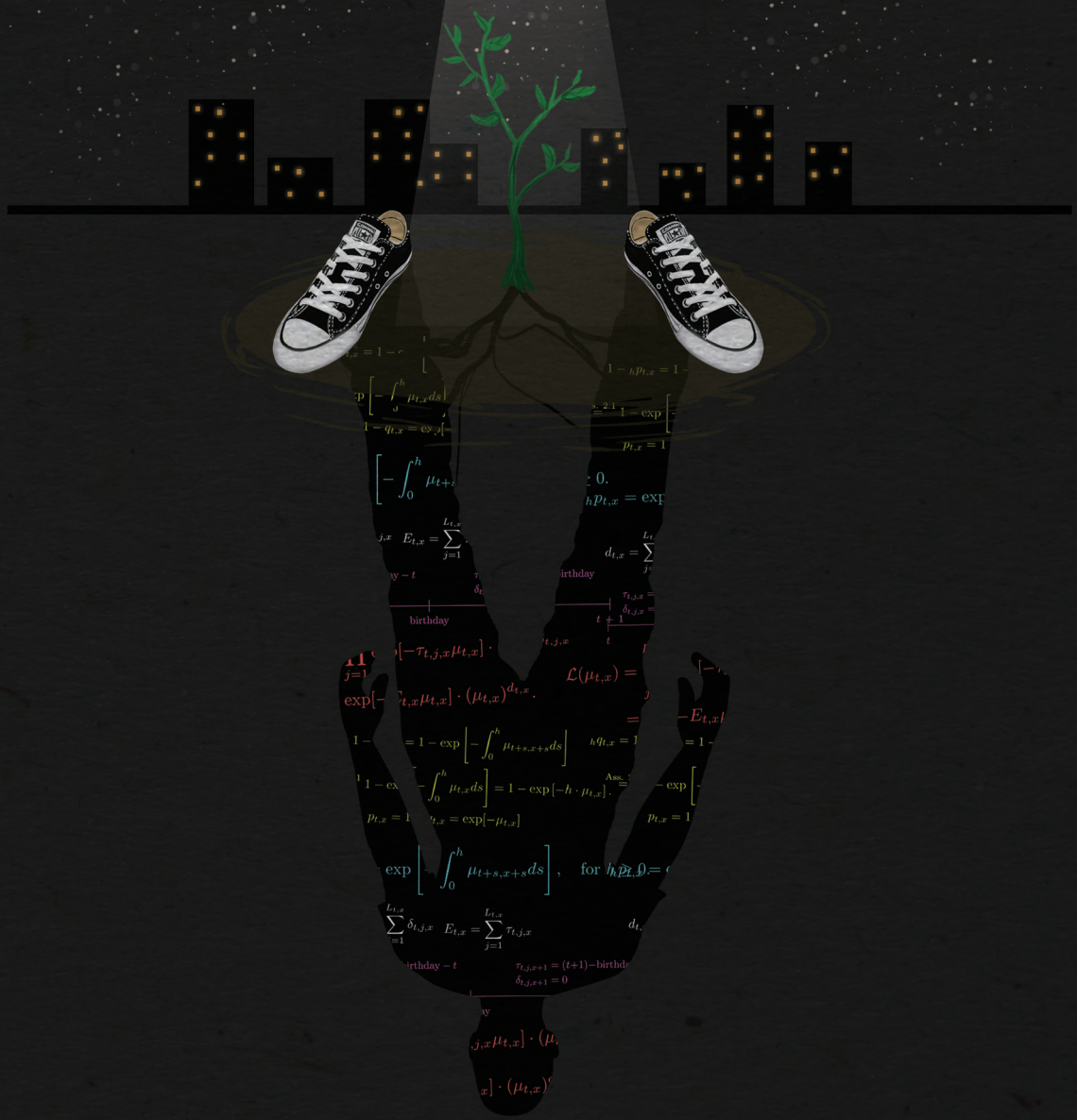
Models for population-wide and portfolio-specific mortality

van Berkum, F.

Publication date
2018
Document Version
Final published version
License
Other

MODELS FOR POPULATION-WIDE AND PORTFOLIO-SPECIFIC MORTALITY



Frank van Berkum

Models for population-wide and portfolio-specific mortality

Frank van Berkum

Models for population-wide and portfolio-specific mortality

This work has been partially funded by Netspar (Network for Studies on Pensions, Aging and Retirement).

Cover design by Evelien Jagtman.
Printed by Gildeprint Drukkerijen.

ISBN: 978-94-6233-883-3

Models for population-wide and portfolio-specific mortality

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 28 maart 2018, te 13.00 uur

door

Frank van Berkum

geboren te Woerden

Promotiecommissie

Promotor:	Prof. dr. ir. M.H. Vellekoop, Universiteit van Amsterdam
Copromotor:	Dr. K. Antonio, Universiteit van Amsterdam
Overige leden:	Prof. dr. A.M.B. De Waegenaere, Tilburg University
	Prof. dr. C.G.H. Diks, Universiteit van Amsterdam
	Prof. dr. R. Kaas, Universiteit van Amsterdam
	Dr. T. Kleinow, Heriot-Watt University
	Prof. dr. R.J.A. Laeven, Universiteit van Amsterdam

Faculteit Economie en Bedrijfskunde

Acknowledgments

It has taken almost seven years of part-time research for me to complete this thesis at the University of Amsterdam, but even at the end I find performing research rewarding. For this, I would like to express my gratitude to a number of people.

I am indebted to my promotor prof. dr. ir. Michel H. Vellekoop and co-promotor dr. Katrien Antonio for their encouragement and support. Thank you both for remaining calm, even when we sometimes strongly disagreed. Every note that I drafted was approached with an open and positive mind, and you always challenged me to push myself. It has been a pleasure working with you during the last few years, and I hope we can continue our successful collaboration.

My sincere thanks go out to the members of the committee, prof. dr. Anja M.B. De Waegenaere, prof. dr. Cees G.H. Diks, prof. dr. Rob Kaas, dr. Torsten Kleinow and prof. dr. Roger J.A. Laeven. Thank you for taking the time to read this thesis, and for critically challenging me during my defense.

I would like to thank my other colleagues from the section Actuarial Science & Mathematical Finance: Rob Kaas, Roger Laeven, Servaas van Bilsen, Tim Boonen, Leendert van Gastel, Umut Can, Hans Schumacher, and fellow PhD-students Jan de Kort, Jitze Hooijsma, Zhenzhen Fan, Yuan Tue, Merrick Zhen Li and Rob Sperna Weiland. Rob (Kaas), thank you for introducing me to non-life insurance (although that is not the subject of my PhD research), and for making me enthousiastic about R. Jan and Jitze, thank you for being able to share my struggles with the combination of parttime research and the more 'regular' work.

My research was partially funded by Netspar (Network for Studies on Pensions, Aging and Retirement). Netspar organized several conferences at which I could present my research, and I am thankful for the valuable feedback that I received which helped me further improve the quality of the research.

Besides working on my thesis for two days a week, I had the pleasure of working at PwC the other three days a week. Working at PwC gives me the opportunity to learn which research areas are relevant for practical issues, but also to apply my research results in practice. I want to thank all PAIS colleagues for keeping me on edge ('hoe gaat het met je werkstuk?'), and for the many activities we do together (getting

drinks after work, winter sports trips, nights out, ...). There are two persons I am especially grateful to: Jan-Huug Lobregt and Lars Janssen. Jan-Huug, thank you for giving me the opportunity to combine research and practice, and for your permanent encouragement and belief in me. Lars, we have studied together at the University of Amsterdam, after which I followed you in working at PwC. After everything we have experienced together you have become a true friend.

Challenging research has to be balanced with sufficient leisure time with friends. Thank you Freek, Carline, Lisette, Michel, Reimer, Maaike, Roman, Erika, Maarten and Rick for the nights out, playing pool at SPDC, for the parties, and for the holiday weekends! I am also thankful to my teammates at TTV Woerden: Roman, Mario and Dennis, I always look forward to our next match of table tennis.

Lastly, I would like to thank my family. Without the support of my parents I would not be where I am today. Susanne, Tom, Melanie and Jeffrey, thank you for always making me feel like home. And Sabine, thank you for putting up with the grumpy version of me the last few months, and for (literally) adding color to my life.

Frank van Berkum Amsterdam January, 2018.

Contents

Acknowledgments **v**

List of Symbols **xi**

1 General introduction **1**
1.1 Longevity risk . . . 1
1.2 Heterogeneity in populations . . . 2
1.3 Contribution of this thesis . . . 3

2 Introduction to mortality modeling **5**
2.1 Mortality observations . . . 5
2.1.1 Deaths and exposures . . . 5
2.1.2 Mortality rates . . . 7
2.2 Stochastic mortality models . . . 9
2.2.1 The Lee-Carter model . . . 9
2.2.2 Extensions of the Lee-Carter model . . . 12
2.3 Forecasting mortality rates . . . 15
2.3.1 Forecasting LC model with a RWD process . . . 15
2.3.2 Other forecasting approaches . . . 17
2.4 Bayesian implementation of the LC-model . . . 22
2.4.1 Bayesian inference and Markov Chain Monte Carlo methods. . . 23
2.4.2 Bayesian estimation of the LC-model . . . 27
2.4.3 Derivation of posterior distributions . . . 29
2.4.4 Case study: Dutch males . . . 32
2.5 Portfolio-specific mortality . . . 35
2.5.1 Multiple-population mortality models . . . 37
2.5.2 Explaining portfolio mortality using risk factors . . . 38

3 The impact of multiple structural changes on mortality predictions 43
3.1 Introduction . . . 43
3.2 Proposed forecasting method . . . 44
3.2.1 Forecasting period effects . . . 44
3.2.2 Forecasting cohort effects . . . 50
3.3 Results . . . 50
3.3.1 Model fit . . . 51
3.3.2 Out-of-sample performance . . . 54
3.4 Conclusion . . . 58

4 A Bayesian joint model for population and portfolio-specific mortality 65
4.1 Introduction . . . 65
4.2 Bayesian portfolio-specific mortality . . . 67
4.2.1 Model formulation and implementation . . . 68
4.2.2 Prior distributions . . . 70
4.3 Empirical study . . . 72
4.3.1 CMI assured lives - original dataset . . . 73
4.3.2 CMI assured lives - reduced portfolio size . . . 78
4.4 Conclusion . . . 83
4.A Posterior distributions . . . 85
4.A.1 Age parameters for population mortality . . . 85
4.A.2 Period parameters for population mortality . . . 87
4.A.3 Portfolio-specific mortality - Gamma prior . . . 89
4.A.4 Portfolio-specific mortality - lognormal prior . . . 90

5 Unraveling relevant risk factors explaining pension fund mortality 93
5.1 Introduction . . . 93
5.2 Data . . . 96
5.3 A framework for statistical modeling of portfolio mortality . . . 100
5.3.1 Distributional assumptions and model estimation . . . 100
5.3.2 A strategy for working with large datasets . . . 104
5.3.3 Model assessment . . . 106
5.4 Results . . . 110
5.4.1 Estimation results . . . 110
5.4.2 Predicted cohort life expectancies . . . 116
5.4.3 Financial backtest . . . 118
5.4.4 Valuation of liabilities . . . 121
5.5 Conclusion . . . 122
5.A Detailed description of the financial backtest . . . 125
5.A.1 Population mortality model . . . 125

5.A.2 Assumptions to compute the expected present value of life annuities 126

Summary **129**

Samenvatting **131**

List of symbols

$D_{t,x}$	The random number of deaths with an age at death in $[x, x+1)$ and time of death in $[t, t+1)$.
$d_{t,x}$	The observed number of deaths with an age at death in $[x, x+1)$ and time of death in $[t, t+1)$.
$E_{t,x}$	The exposure-to-risk for people that may die at age in $[x, x+1)$ in the time interval $[t, t+1)$.
$m_{t,x}$	The central death rate i.e. $d_{t,x}/E_{t,x}$.
${}_hq_{t,x}$	The probability that someone aged exactly x at the beginning of year t dies within the coming period h. The mortality rate $q_{t,x} = {}_1q_{t,x}$ is a special case that is often used in practice.
$\mu_{t,x}$	The force of mortality (or instantaneous mortality rate), $\lim_{h \downarrow 0} \frac{{}_hq_{t,x}}{h}$.
$p_{t,x}$	The one-year survival probability: the probability that someone aged exactly x at the beginning of calendar year t survives for one year, so it equals $1 - q_{t,x}$.
$\tau_{t,j,x}$	The period of calendar year t that individual j survives while having an age in $[x, x+1)$.
$\delta_{t,j,x}$	An indicator variable that equals 1 if individual j dies during year t while having an age in $[x, x+1)$, and 0 otherwise.

Chapter 1

General introduction

In this thesis we develop and analyze possible improvements for human mortality models. The fact that people die, on average, at higher ages is well-known, and the uncertainty surrounding this process is known as longevity risk. It has become a major issue in financial risk management for insurance companies and pension funds. This chapter gives a general introduction to longevity risk and introduces the topics that are investigated in later chapters.

1.1 Longevity risk

Until the early nineties, pension funds and insurance companies valued their liabilities using mortality tables that were based on an average of observed mortality over some historical period (e.g. five years). Possible future developments were not taken into account. As a result, insurance companies and pension funds recognized that mortality was improving, but they were lagging behind the most recent developments.

In 1992, a seminal mortality model was developed by Lee and Carter (1992). In their article, mortality is modeled using a random walk with drift, and future mortality developments are explicitly taken into account. Around the same time, insurance companies in the Netherlands noticed that their periodically updated mortality tables suggested that mortality-linked products were profitable and longevity-linked products led to losses. They concluded that mortality projections which include a mortality trend were needed. The Dutch Association of Insurers (Verbond van Verzekeraars) set up a committee to investigate methods that could be used for the pricing of life insurance products, and they came up with a model that shows similarities with the Lee-Carter model (Tornij (2004)).

For a long time, having some losses due to mortality and longevity risk did not pose problems for insurance companies and pension funds. Insurance companies generated large profits due to high stock returns, and pension funds showed high funding ratios. In the Netherlands and in many other countries, some pension funds even gave premium holidays. This meant that the boards of some pension funds decided

that members could accrue new pension rights without having to pay extra premiums since the funding ratios were so high.

Then, in the beginning of the new millennium, stock markets crashed, while interest rates had been decreasing since the early nineties. As a result, solvency ratios of insurance companies and funding ratios of pension funds deteriorated substantially. Further, it became widely accepted that even though developments in mortality had been taken into account, previous predictions had underestimated the improvements in mortality. Oeppen and Vaupel (2002) show that experts defined upper limits to life expectancy which continued to be broken. For example, they illustrate how the worldwide record for male life expectancy increased steadily over the past 160 years by a quarter of a year per year (on average), which is much faster than was anticipated by the experts.

In recent years, more empirical research has been performed on mortality projections for populations, and the methods used by public institutions seem to converge. New mortality forecasts showed steeper mortality trends than before, which led to higher liability values for pension funds.

January 1st 2016 Solvency II, the new legislative framework for European insurance companies, has become effective. This is prudential legislation that prescribes the amount of capital that insurance companies in the European Union have to hold to reduce the risk of insolvency. The rules that are used to determine the required solvency capital under Solvency II (the so-called Standard Formula) are more complex than they were in Solvency I, though still so simple that they can easily be implemented. The Standard Formula, however, does not always closely represent the specific risks faced by an insurer. Insurance companies are therefore allowed to build an Internal Model for risks, and they can use this Internal Model to determine their required solvency capital. This option has stimulated large(r) insurers to look more closely at the different risks that they face in their insurance products, and this is specifically the case for longevity risk of life insurers.

1.2 Heterogeneity in populations

People with a different socioeconomic background and different life style experience a different level of mortality. Pension funds may wish to account for this when valuing their liabilities, e.g. by distinguishing different risk profiles when defining their mortality assumptions.

Chetty et al. (2016) find, in a study on historically observed mortality in the USA, that the difference in (period) life expectancy between the richest 1% and poorest 1% of individuals was approximately 14.6 years, and that differences can be substantial between different geographic locations. Villegas and Haberman (2014) construct five socioeconomic classes based on seven characteristics such as income deprivation and

education. They find that differences in socioeconomic classes may lead to differences in annuity prices up to 20%.

Based on these results, it is clear that forecasts for the future survival rates of policyholders in insurance companies and participants in pension funds will benefit if both historical data on the level of the whole population as well as portfolio-specific information is explicitly taken into account.

1.3 Contribution of this thesis

We provide an introduction to stochastic mortality modeling in Chapter 2, to provide the reader with some background knowledge and to define the notation that is used in the following chapters. We first introduce stochastic mortality models and show how mortality forecasts can be obtained from datasets using time series methods. We illustrate this by applying the Lee-Carter model to a dataset for the whole population of Dutch males and females. We also analyze the impact of parameter uncertainty in mortality forecasts using a Bayesian implementation of the same model. Finally, we provide an overview of different approaches used to obtain forecasts of portfolio-specific mortality rates.

To describe time-varying effects (usually called 'period effects' in the actuarial literature), linear time series models are used. When applied to single population models, the resulting forecasts do not always show an intuitive continuation of historical observations. In Chapter 3 we therefore propose an approach to detect whether there are structural changes present in the observed period effects. If this is the case, we use a different forecasting approach to ensure the forecasts are in line with the historical observations. We show that the predictive performance of our approach compares favorably with other approaches used in the literature.

In Chapter 4 we separate different sources of uncertainty in portfolio mortality: parameter uncertainty, individual mortality risk, and uncertainty in future mortality rates. We introduce a model to jointly estimate mortality in the England & Wales population and in the CMI dataset on assured lives, which is a subset of that population. We use a Bayesian approach to calibrate our combined model, which allows us to separate parameter uncertainty from individual mortality risk. We also illustrate the relative importance of different sources of uncertainty when projecting mortality for a subpopulation. This allows us to characterize the impact of portfolio size on the uncertainty of future mortality, and to conclude when modeling of subpopulations is useful and when not.

Finally, in Chapter 5 we study observed mortality in a large Dutch pension fund. We use Poisson regression in generalized additive models to estimate smooth effects of risk factors (as opposed to only linear effects). We use information criteria and proper scoring rules to determine how much the various risk factors contribute to the

explanation of historically observed mortality. For a pension fund (or insurance company), however, it is even more relevant to accurately predict liabilities. Therefore, we also investigate which risk factors contribute most to the accuracy of predictions for future dynamics in a pension fund's value of liabilities. We introduce a novel financial backtest specifically for this purpose, and show that this may lead to different conclusions than other, more classical, backtesting procedures.

Chapter 2

Introduction to mortality modeling

In this chapter we provide a short introduction to stochastic mortality modeling and introduce our notation.[1] First, we discuss how mortality variables on the individual and the aggregated level are constructed from raw mortality data in Section 2.1. In Section 2.2 we introduce the seminal stochastic mortality model introduced by Lee and Carter (1992) and extensions from that model that have been developed since then. In Section 2.3 we discuss the different forecasting methods that have been proposed in the literature and illustrate how mortality predictions can be obtained using the Lee-Carter model. In Section 2.4 we formulate a Bayesian version of the Lee-Carter model. We show how this allows us to characterize parameter uncertainty and its effect on mortality forecasts. Finally, in Section 2.5, we discuss different approaches to obtain mortality predictions for a specific pension fund or insurance portfolio.

2.1 Mortality observations

We introduce the following notation that will be used throughout this thesis unless stated otherwise. We usually consider integer ages $x \in \mathcal{X} = \{x_1, x_1 + 1 \ldots, x_X\}$ and integer calendar years $t \in \mathcal{T} = \{t_1, t_1 + 1, \ldots, t_T\}$; we thus have X ages and T years. A person aged x in calendar year t is said to be from cohort $c = t - x$; this is the year of birth for that person.

2.1.1 Deaths and exposures

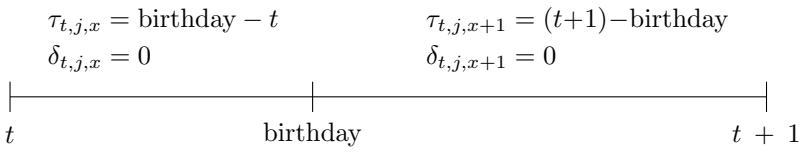
For a population we define the observed number of deaths with an age in $[x, x + 1)$ at death and time of death in $[t, t + 1)$ (i.e. during calendar year t) as $d_{t,x}$. The corresponding exposure-to-risk of people aged x during calendar year t is $E_{t,x}$. This exposure is the total number of 'person-years' in a population over a calendar year, and Pitacco et al. (2009) interpret it as the average number of individuals in the

[1] Section 2.1 is largely based on Chapter 3 from Pitacco et al. (2009), sections 2.2 and 2.3 are based on the introductory part from van Berkum et al. (2016), Section 2.4 is based on van Berkum et al. (2017a), and Section 2.5 is based on van Berkum et al. (2017a) and van Berkum et al. (2017b).

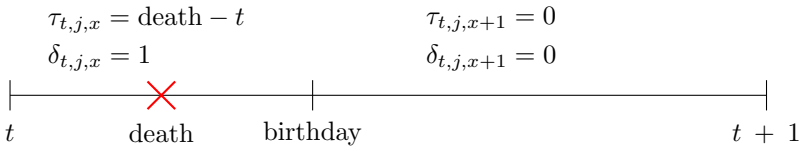
population over a calendar year, adjusted for the length of time they are in the population. For an individual person j we denote the individual death observation by $\delta_{t,j,x}$ and individual exposure by $\tau_{t,j,x}$. Below we illustrate how these variables are determined on the individual level, and we show how to construct aggregated observations from the individual ones.

Individual level. We consider integer years t, integer and non-integer ages x and $\tilde{x}$ respectively.[2] An individual aged $\tilde{x} = x + \iota$ at the beginning of year t with $\iota \in [0, 1)$, celebrates his birthday at $t+(1-\iota)$. There are three possible outcomes with respect to survival in year t, and the corresponding death and exposure observations are defined as follows:

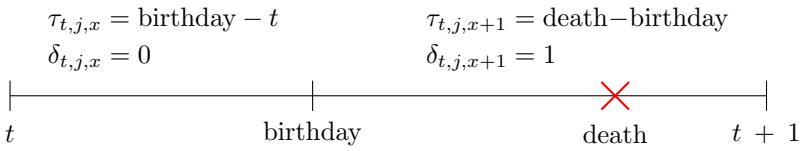
- Individual j survives year t:



- Individual j dies during year t, *before* his birthday:



- Individual j dies during year t, *at or after* his birthday:



For individual j aged $\tilde{x}$ at the beginning of year t we thus obtain observations on death counts $\delta_{t,j,x}$ and $\delta_{t,j,x+1}$ and observations on exposures $\tau_{t,j,x}$ and $\tau_{t,j,x+1}$. In the figures above we see that if individual j survives year t, the contribution to observed deaths is zero and the total contribution to the exposure-to-risk is one (=

[2] The distinction between integer and non-integer ages is only needed in this section. In the remainder of this section we only consider integer ages.

$\tau_{t,j,x} + \tau_{t,j,x+1}$). If individual j dies during calendar year t, the total contribution to the risk exposure is less than one, and either $\delta_{t,j,x}$ or $\delta_{t,j,x+1}$ is one, depending on when individual j died.

Aggregated level. Consider $L_{t,x}$ individuals who are alive at the beginning of year t with ages in $[x, x+1)$. Using the observations $\delta_{t,j,x}$ and $\tau_{t,j,x}$ for individual j, the total number of observed deaths $d_{t,x}$ and the total exposure-to-risk $E_{t,x}$ in the population are given by

$$d_{t,x} = \sum_{j=1}^{L_{t,x}} \delta_{t,j,x} \qquad \text{and} \qquad E_{t,x} = \sum_{j=1}^{L_{t,x}} \tau_{t,j,x}. \tag{2.1}$$

For many countries these statistics can be downloaded from the Human Mortality Database (HMD).[3]

In Chapter 5 we construct aggregated mortality statistics from individual mortality data using the methods described above. In all other cases in the remainder of this thesis we always use mortality data from HMD.

2.1.2 Mortality rates

Using the mortality statistics as defined in the previous section we define the central death rate $m_{t,x}$ as

$$m_{t,x} = \frac{d_{t,x}}{E_{t,x}}. \tag{2.2}$$

The central death rate is not restricted to the interval $[0, 1]$ as one might expect. Consider the case where there is only one person aged x in calendar year t and assume this person dies during year t. The observed number of deaths $d_{t,x}$ equals one with a corresponding exposure-to-risk $E_{t,x}$ smaller than one, which leads to $m_{t,x} > 1$. The central death rate should therefore not be interpreted as a mortality probability but rather as a frequency.

The mortality probability $q_{t,x}$ *is* restricted to the interval $[0, 1]$. This is the probability that someone aged exactly x at the beginning of calendar year t dies within the next year. In a similar way, ${}_hq_{t,x}$ is used to denote the probability that someone aged exactly x at the beginning of calendar year t dies within a period h, i.e. before reaching age $x + h$ at time $t + h$.

[3] The Human Mortality Database is a joined project of the University of California at Berkeley (USA) and the Max Planck Institute for Demographic Research (Germany). Data are available at http://www.mortality.org, and see http://www.mortality.org/Public/Docs/MethodsProtocol.pdf for a description of the methods that are used to construct the datasets.

In a continuous time framework, we can also consider the force of mortality $\mu_{t,x}$ with $(t,x) \in (\mathbb{R}^+ \times \mathbb{R}^+)$. This is an instantaneous rate of mortality, and it is defined by

$$\mu_{t,x} = \lim_{h \downarrow 0} \frac{{}_h q_{t,x}}{h}. \tag{2.3}$$

In survival analysis, this quantity is known as the hazard function. The cumulative survival probability ${}_h p_{t,x}$ is the probability that someone with exact age x at the beginning of year t survives a period of length h. We can express the cumulative survival probability as a function of the force of mortality:

$${}_h p_{t,x} = \exp\left[-\int_0^h \mu_{t+s,x+s} ds\right], \quad \text{for } h \geq 0. \tag{2.4}$$

In practice, we do not have mortality data available in continuous time. To facilitate the development of continuous time models that can be estimated using data on discrete intervals, we assume a piecewise constant force of mortality.

Assumption 2.1 (Piecewise constant force of mortality). Let x and t be integers. Then, the force of mortality $\mu_{t+s,x+u}$ satisfies $\mu_{t+s,x+u} = \mu_{t,x}$ for all s and u in $[0,1)$. □

Under Assumption 2.1, the period h mortality probability for $h \in [0,1)$ is given by

$$\begin{aligned} {}_h q_{t,x} &= 1 - {}_h p_{t,x} = 1 - \exp\left[-\int_0^h \mu_{t+s,x+s} ds\right] \\ &\overset{\text{Ass. 2.1}}{=} 1 - \exp\left[-\int_0^h \mu_{t,x} ds\right] = 1 - \exp\left[-h \cdot \mu_{t,x}\right]. \end{aligned} \tag{2.5}$$

Now suppose we have available mortality data $\delta_{t,j,x}$ and $\tau_{t,j,x}$ for each individual $j \in \{1, 2, \ldots, L_{t,x}\}$. We can estimate $\mu_{t,x}$ under Assumption 2.1 using a likelihood method, if no further modeling assumptions are made on the forces of mortality. The contribution of individual j to the likelihood can then be written as

$$p_{t,x} = 1 - q_{t,x} = \exp[-\mu_{t,x}] \tag{2.6}$$

if he survives, and

$${}_{\tau_{t,j,x}} p_{t,x} \cdot \mu_{t+\tau_{t,j,x},x+\tau_{t,j,x}} \overset{(2.5)}{=} \exp[-\tau_{t,j,x}\mu_{t,x}]\mu_{t,x}. \tag{2.7}$$

These two expressions can be combined into a single expression, and under the assumption of independence for individual lifetimes given the forces of mortality we can

then define the total likelihood as

$$\begin{aligned}\mathcal{L}(\mu_{t,x}) &= \prod_{j=1}^{L_{t,x}} \exp[-\tau_{t,j,x}\mu_{t,x}] \cdot (\mu_{t,x})^{\delta_{t,j,x}} \\ &= \exp[-E_{t,x}\mu_{t,x}] \cdot (\mu_{t,x})^{d_{t,x}}. \end{aligned} \tag{2.8}$$

The last expression is proportional to a Poissonian likelihood for a stochastic number of deaths $D_{t,x} \sim \text{Poisson}(E_{t,x}\mu_{t,x})$. This result is used in later chapters to justify using the Poisson likelihood when modeling death counts.

The maximum likelihood estimate (MLE) for $\mu_{t,x}$ is obtained by maximizing $\mathcal{L}(\mu_{t,x})$ with respect to $\mu_{t,x}$, and the solution is given by

$$\hat{\mu}_{t,x} = \frac{d_{t,x}}{E_{t,x}} = m_{t,x}. \tag{2.9}$$

In this section we have defined three different mortality-related quantities: the central death rate $m_{t,x}$, the mortality probability $q_{t,x}$, and the force of mortality $\mu_{t,x}$. When we discuss mortality-related topics but the discussion is not specific to either one of the quantities, we may refer to all these quantities in general as 'mortality rates'.

2.2 Stochastic mortality models

In the previous section we have introduced definitions of mortality observations. Figure 2.1 shows the development of historical central death rates for Dutch males and females. We observe that mortality is not constant through time, and that it develops differently for different ages. This led Lee and Carter (1992) to define a new model to explain observed mortality and predict future mortality. We now discuss this model and several extensions.

2.2.1 The Lee-Carter model

Lee and Carter (1992) introduce the following mortality model to explain observed death rates:

$$\ln m_{t,x} = \alpha_x + \beta_x \kappa_t + \eta_{t,x}, \tag{2.10}$$

with the $\eta_{t,x}$ iid stochastic variables with mean zero. The variable α_x represents some baseline mortality level for age x, the variable κ_t captures the dynamics over time in a general level of mortality in the population, and the variable β_x represents the sensitivity of age x for changes in this general level of mortality. The Lee-Carter model is a single-factor model since it contains only one factor that evolves over time; this implies that mortality improvements at all ages are perfectly correlated.

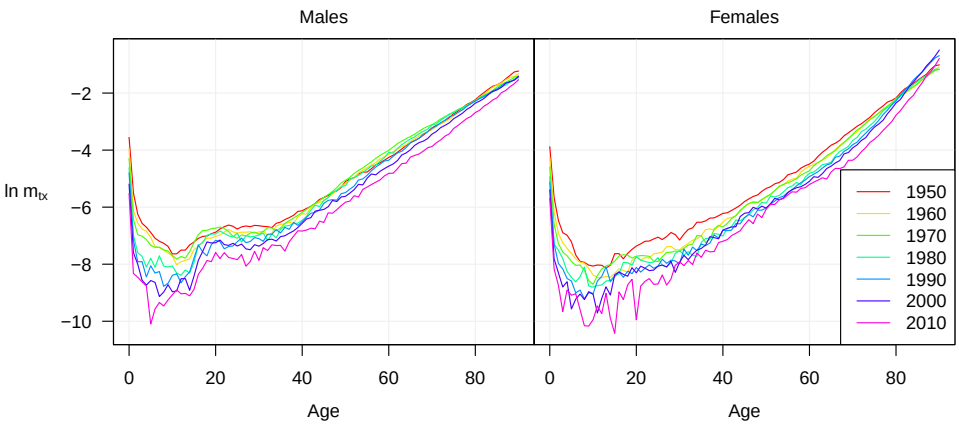


Figure 2.1: Observed central death rates $m_{t,x}$ for Dutch males (left) and females (right) aged 0-90 for selected years between 1950 and 2010.

Lee and Carter estimate this model using a Singular Value Decomposition (SVD), and they model the period effect κ_t using a random walk with drift to generate mortality projections:

$$\kappa_t = \kappa_{t-1} + \delta + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2), \tag{2.11}$$

with $\eta_{t,x}$ and ε_t assumed independent.

Brouhns et al. (2002) explicitly take into account that the deaths $d_{t,x}$ contain two sources of uncertainty: the uncertainty due to the dynamics in the force of mortality $\mu_{t,x}$, and the uncertainty in the deaths given this force of mortality. They model the force of mortality instead of the death rate $m_{t,x}$:

$$D_{t,x}|\mu_{t,x} \sim \text{Poisson}(E_{t,x}\mu_{t,x}), \qquad \text{with} \qquad \ln \mu_{t,x} = \alpha_x + \beta_x \kappa_t. \tag{2.12}$$

This model can be estimated by maximizing the likelihood of the observations $d_{t,x}$ of $D_{t,x}$ (and the exposures $E_{t,x}$), using numerical optimization techniques such as the Newton-Raphson algorithm.

The Lee-Carter model is not uniquely identified. Consider for example the following two parameter sets:

$$\left\{ \begin{array}{c} \alpha_x \\ \beta_x \\ \kappa_t \end{array} \right\} \quad \text{and} \quad \left\{ \begin{array}{c} \tilde{\alpha}_x \\ \tilde{\beta}_x \\ \tilde{\kappa}_t \end{array} \right\} = \left\{ \begin{array}{c} \alpha_x - cd\beta_x \\ c\beta_x \\ \frac{\kappa_t}{c} + d \end{array} \right\}.$$

The log force of mortality for both parameter sets equals $\ln \mu_{t,x} = \alpha_x + \beta_x \kappa_t$. Hence, linear transformations of the parameters α_x, β_x and κ_t can lead to the same force of

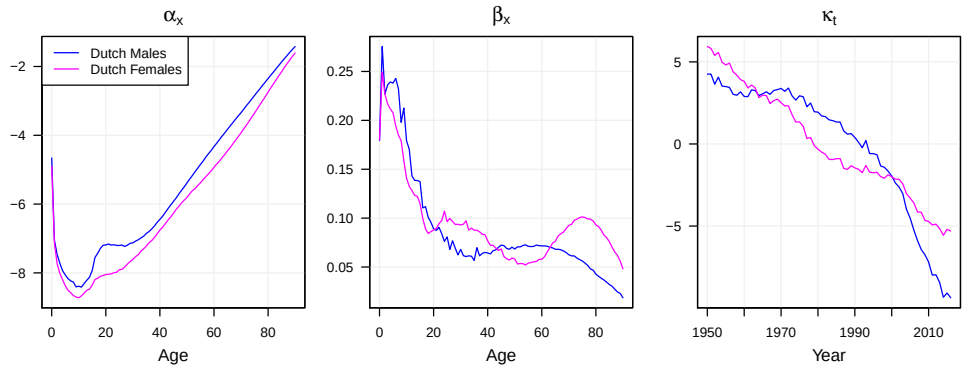


Figure 2.2: Parameter estimates for the Lee-Carter model applied to mortality data on Dutch males and females aged 0-90 in the years 1950-2016.

mortality. The following parameter constraints are applied to avoid this identification problem:

$$\sum_t \kappa_t = 0, \qquad \text{and} \qquad \sum_x \beta_x^2 = 1. \tag{2.13}$$

Under these parameter constraints, we have that $\hat{\alpha}_x = \frac{1}{T}\sum_{t\in\mathcal{T}} \ln \mu_{t,x}$, which means that α_x represents the average logarithmic force of mortality for age x during the observed period.

Example 2.2 (Lee-Carter applied to Dutch mortality data). In Figure 2.2 we show parameter estimates for the Lee-Carter model applied separately to Dutch males (blue) and females (magenta) aged 0-90 in the years 1950-2000. These parameter estimates are obtained by maximizing the Poisson likelihood specification as in (2.12) using the Newton-Raphson algorithm. In the left panel we note several commonly observed characteristics in the pattern of α_x (the mean log force of mortality). Mortality for the first life year is relatively high, but decreases rapidly for youngsters. After the age of 10, the logarithmic force of mortality increases almost linearly up to the highest ages. Only for the ages around 20 years, we observe a 'hump' in mortality which is often referred to as the 'accident hump'. This hump is more pronounced for males than for females, because males tend to live more dangerously than females at this age.

The middle panel shows that – for the dataset considered – mortality for young males improves at a faster rate than for older males (for the highest ages there is actually almost no improvement in mortality). The pattern is different for females. Improvements are highest for the youngest ages, but also for higher ages mortality improvements are still substantial.

Table 2.1: An overview of model specifications used in this thesis. The variables $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and γ_{t-x} represent age, period and cohort effects respectively, $(\bar{x}-x)^+ = \max\{\bar{x}-x, 0\}$ and $\operatorname{logit} q = \ln(q/(1-q))$. Variables used in this table but not defined before: $\bar{x}$ is the average of the ages in $\mathcal{X}$, $b(x) = \left((x-\bar{x})^2 - \frac{1}{n}\sum_{x_i\in\mathcal{X}}(x_i-\bar{x})^2\right)$, $c(x) = (\bar{x}-x)^+ + [(\bar{x}-x)^+]^2$, and x_c is a constant which can be chosen a priori or can be estimated. We estimate the model for all possible $x_c \in \mathcal{X}$, and the value of x_c is chosen such that the likelihood is maximized.

Model	Formula
M1	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$
M1A	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)}$
M2	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}$
M2A	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)} + \gamma_{t-x}$
M3	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_{t-x}$
M5	$\operatorname{logit} q_{t,x} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}$
M6	$\operatorname{logit} q_{t,x} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$
M7	$\operatorname{logit} q_{t,x} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + b(x)\kappa_t^{(3)} + \gamma_{t-x}$
M8	$\operatorname{logit} q_{t,x} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x_c-x)\gamma_{t-x}$
M9	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)} + \gamma_{t-x}$
M10	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)}$
M11	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)} + b(x)\kappa_t^{(4)} + \gamma_{t-x}$
M12	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)} + (x_c-x)\gamma_{t-x}$
M13	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + c(x)\kappa_t^{(3)} + \gamma_{t-x}$

Finally, the right panel shows the development in the general mortality trend over all ages for Dutch males and females, as captured by the time series κ_t. For males there seems to be a 'kink' in the time series around 2000, whereas for females it looks more like a straight line since 1950. □

2.2.2 Extensions of the Lee-Carter model

The Lee-Carter model introduced in the previous section is used in many papers as the reference model since for many datasets it results in plausible mortality rate projections (the projection of mortality rates is discussed in the next section), and in general because of its simplicity. However, the Lee-Carter model also has some

well-known limitations:

- The model contains a single period effect, and as a result, mortality improvements are perfectly correlated over different ages. In Figure 2.1 however, we see that historical mortality developments are different at different ages;

- From Figure 2.3 we observe that the model is not always able to fit the historical observations equally well for all ages. Therefore, mortality forecasts may not always connect nicely with recent observations;

- The model contains many parameters: two for each age included and one for each calendar year included. A mortality calibration exercise therefore easily requires more than 200 parameters.

Several extensions of the Lee-Carter model have been introduced to overcome these limitations. All extensions of the Lee-Carter model contain more than one time-dependent effect and therefore no longer result in mortality improvements that are perfectly correlated over different ages. In Table 2.1 we list the extensions of the Lee-Carter model that will be considered in Chapter 3, and we discuss the motivation for some of these below.

Improving the model fit. Lee and Carter (1992) estimate their model using a Singular Value Decomposition (SVD). In that model, the eigenvector corresponding to the largest singular value is used to create a single factor (the period effect κ_t). The rationale for this approach is that the first eigenvector explains a large part of the observed variation. Renshaw and Haberman (2003) consider a natural extension of the Lee-Carter model in which multiple eigenvectors are used, and they illustrate their approach using two factors (this model is denoted M1A in Table 2.1).

Renshaw and Haberman (2006) observe that residuals from the Lee-Carter model contain a ripple effect when considered over the year-of-birth-axis. They extend the Lee-Carter model with a cohort effect γ_c with $c = t-x$ to take care of this phenomenon (model M2).[4] Currie (2006) considers the Age-Period-Cohort model (APC, model M3 in Table 2.1) that is often used in the biostatistics literature. The APC model differs from the other models considered, since it does not contain bilinear terms. With the introduction of a cohort effect there is no longer a discernible pattern in the residuals. However, the estimated cohort effect is often far from linear and recent cohorts contain only a few datapoints. It is therefore difficult to generate plausible mortality forecasts for future cohorts.

[4]The model introduced in Renshaw and Haberman (2006) has $\beta_x^{(3)} = 1$ for all x, but the authors generalized the model in Renshaw and Haberman (2006) by the introduction of an age-dependent $\beta_x^{(3)}$.

Table 2.2: Overview of the parameter constraints imposed on the models.

Model	Constraints				
M1	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$			
M1A	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_x \beta_x^{(3)} = 1$	$\sum_t \kappa_t^{(3)} = 0$	
M2	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_x \beta_x^{(3)} = 1$	$\sum_{t,x} \gamma_{t-x} = 0$	
M2A	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_x \beta_x^{(3)} = 1$	$\sum_t \kappa_t^{(3)} = 0$	$\sum_{t,x} \gamma_{t-x} = 0$
M3	$\sum_t \kappa_t^{(2)} = 0$	$\sum_{t,x} \gamma_{t-x} = 0$			
M5	(no constraints needed)				
M6	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$			
M7	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_c c^2\gamma_c = 0$		
M8	$\sum_{t,x} \gamma_{t-x} = 0$				
M9	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$		
M10	$\sum_t \kappa_t^{(1)} = 0$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_t \kappa_t^{(3)} = 0$		
M11	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_c c^2\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$	
M12	$\sum_{t,x} \gamma_{t-x} = 0$				
M13	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$		

Reducing the number of parameters. Cairns et al. (2006) note that the force of mortality for pensioners is approximately log linear. They use this property to define a model with prescribed functional forms for the age effects so they only have to estimate period effects (model M5). Cairns et al. (2009) generalize M5 by introducing a cohort effect and other age effects (models M6-M8). However, since these models are designed for mortality of pensioners, one should be careful when applying them to other age ranges.

Improving the model fit while keeping the number of parameters limited. Plat (2009a) defines a model that is appropriate for a wider age range. To capture different mortality dynamics at different ages, he introduces period effects for specific age groups (model M9). To ensure that the number of parameters does not increase too much, he multiplies the period effects for specific age groups by parametric functions of age.

Haberman and Renshaw (2011) investigate several stochastic mortality models, and they propose models in which characteristics from the models in Cairns et al. (2009) and from M9 are combined (models M10-M12). O'Hare and Li (2011) define a variant of M9 in which the age effect for specific groups is quadratic instead of linear (model M13).

Nearly all extensions of the Lee-Carter model exhibit similar identification problems. Unless stated otherwise, we apply the parameter constraints shown in Table 2.2 when estimating these models. Note that for the models M2, M2A, M3, M8 and M12 the parameter restriction is written as a summation over both t and x, whereas for the models M6, M7, M9, M11 and M13 the parameter restriction is written as a summation over $c = t - x$. In the first case, the weighted average of the cohort effect should be zero (weighted over the frequency of the cohorts in the dataset), and in the second case the average of the cohort effect should be zero (each cohort counted once with equal weight).

The fit on historical data is improved by the introduction of a cohort effect and period effects for specific age groups. However, the introduction of extra time-dependent parameters complicates mortality forecasting. Prediction of the cohort effect proves to be especially difficult, and Haberman and Renshaw (2011) therefore even avoid prediction of γ_c. In the next section we discuss the prediction of period and cohort effects.

2.3 Forecasting mortality rates

The mortality models defined in the previous section contain one or multiple period effects $\kappa_t^{(i)}$ and possibly a cohort effect γ_c. (If the model contains only a single period effect we drop the superfluous superscript i.) If projections of mortality rates are needed, one needs to project these time-dependent variables into the future. In this section we first show the approach used most often to forecast mortality rates from the Lee-Carter model, then we discuss other approaches used for forecasting of the period effect(s) and the cohort effect in mortality models.

2.3.1 Forecasting LC model with a RWD process

Recall from Section 2.2.1 that Lee and Carter (1992) use a random walk with drift to predict the variable κ_t:

$$\kappa_t = \kappa_{t-1} + \delta + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2), \tag{2.14}$$

with δ the drift parameter and σ_ε^2 the variance of the error terms. Define $\Delta\kappa_t = \kappa_t - \kappa_{t-1}$ for $t \in \{t_1 + 1, \ldots, t_T\}$. The maximum likelihood estimates for these parameters are given by

$$\hat{\delta} = \frac{1}{T-1} \sum_{t=t_1+1}^{t_T} \Delta\kappa_t = \frac{\kappa_{t_T} - \kappa_{t_1}}{T-1} \quad \text{and} \quad \hat{\sigma}_\varepsilon^2 = \frac{1}{T-1} \sum_{t=t_1+1}^{t_T} (\Delta\kappa_t - \hat{\delta})^2. \tag{2.15}$$

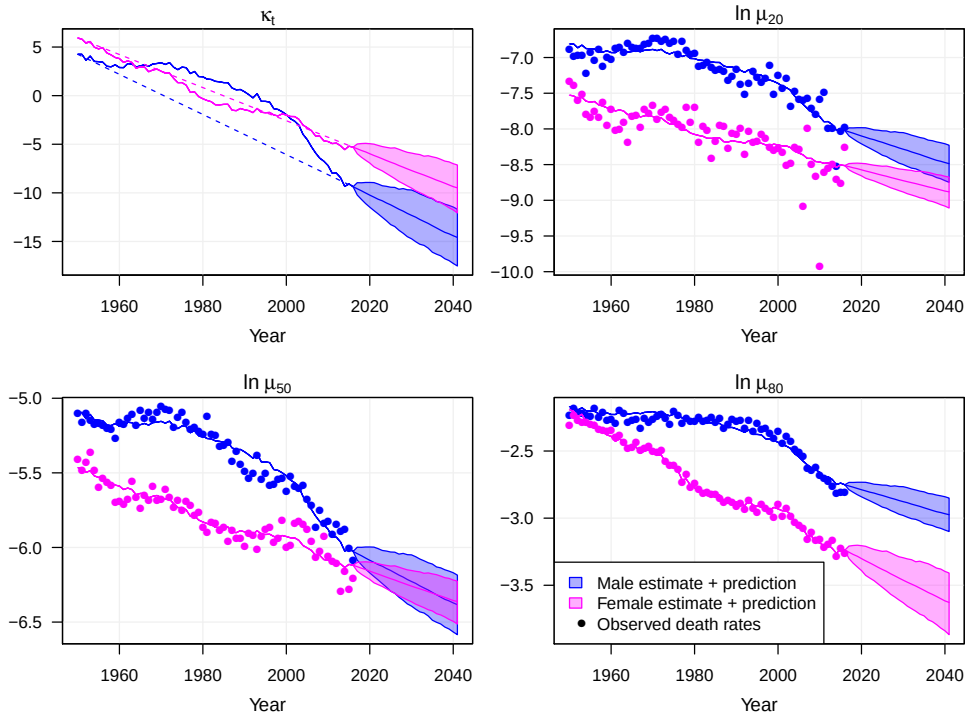


Figure 2.3: Top left panel: prediction for κ_t from Figure 2.2 (2.5th, 50th and 97.5th percentiles are shown). Other panels: fitted and predicted mortality intensities $\mu_{t,x}$ and observed death rates $m_{t,x}$ for Dutch males and females for the ages 20, 50 and 80.

We observe that the estimate for the drift parameter depends only on the first and last observation. As a result, the drift parameter can be highly dependent on the calibration period chosen. Including or excluding one year from the calibration period may sometimes lead to very different mortality forecasts.

Example 2.2 (Lee-Carter applied to Dutch mortality data – continued). In the top-left panel of Figure 2.3 we show projections for the period effect κ_t in Figure 2.2 for males and females. The dashed line is the expected value for the years 1951-1999 given the observations in 1950 and 2000. The slope for males is more negative than for females, since mortality improved more for males than for females.

In the top-right, bottom-left and bottom-right panel we show observations $\ln m_{t,x}$ and fitted values $\ln \mu_{t,x}$ during the calibration period, and predicted forces of mortality for the ages 20, 50 and 80 respectively. For $x = 20$ the observations $\ln m_{t,x}$ are the most volatile. However, mortality at $x = 20$ is low, and therefore the expected number

of deaths is low. Our explicit incorporation of Poisson noise in observed numbers of deaths ensures that this part of the volatility does not end up in our predictions, which forecast the force of mortality. For $x = 50$ and $x = 80$ the fitted values seem to correspond more closely to the observed values. Two noteworthy observations:

1. Male mortality at $x = 50$ is predicted to become lower than female mortality. Historically, female mortality has been lower than male mortality, and it is generally assumed that this will continue to be the case in the future. Cairns et al. (2011b) introduce a multiple population model in which such assumptions can be imposed through the model specification.

2. At $x = 80$, male mortality hardly improved for many years, while female mortality did improve steadily. This may be explained by differences in smoking habits between males and females. Further, the sharp decrease in male and female mortality which starts around the year 2000 can be explained by increased health care budgets provided by the Dutch government, see Stoeldraijer et al. (2013). The uncertainty in male mortality predictions is almost twice as low as in female mortality projections. This can be explained using Figure 2.2: β_{80} is close to zero for males but clearly positive for females. As a result, the prediction of $\mu_{t,80}$ for males is much less affected by the prediction of κ_t than the prediction for females.

A life insurer can use the mean of the mortality scenarios shown in Figure 2.3 to assign a single value to the liabilities, and stochastic scenarios can be used to determine the uncertainty in this value. □

The random walk with drift is a special case within the ARIMA time series framework.[5] In the next section we discuss this and other approaches used to forecast mortality rates.

2.3.2 Other forecasting approaches

We give an overview of standard ARIMA time series models, extensions to these models, and other time series models and approaches that have been used for forecasting both period and cohort effects in mortality models.

Standard ARIMA-models. Cairns et al. (2011a) fit the models M1 to M5, M7 and M8 to England and Wales data from the years 1961 to 2004. For the period effects they fit a (single or multivariate) random walk with drift. For the cohort effects they estimate different ARIMA(p, d, q)-specifications. The specifications used in backtesting are based on the BIC and on biological reasonableness of the projections.

[5] See Tsay (2010) for a thorough introduction to time series modeling.

They define biological reasonableness as a concept that is intended to cover a wide range of subjective criteria. The authors suggest that the following question can be used to determine whether mortality projections are biologically reasonable: "*What mixture of biological factors, medical advances and environmental changes would have to happen to cause a particular set of forecasts?*". If this question is difficult to answer for mortality forecasts produced by a certain model, the projections from that model are said to be biologically unreasonable.

Cairns et al. (2011a) consider second order differencing of the cohort effect ($d = 2$), but they find that this leads to large confidence intervals which the authors find less plausible. For the data under consideration a mean reverting process (AR(1)) or an ARIMA$(1, 1, 0)$ process (both including a constant) is found to be most appropriate for the cohort effects.

Plat (2009a) introduces M9 and includes it in a comparative study of mortality models fitted to data from the United States (1961 to 2005), England & Wales (1961 to 2005), and the Netherlands (1951 to 2005). In his approach the first period effect ($\kappa_t^{(1)}$ in Table 2.1) is the main effect, and a random walk with drift is used to project this factor. For the other period effects ($\kappa_t^{(2)}$ and $\kappa_t^{(3)}$ in Table 2.1), a non-stationary ARIMA process such as a random walk with drift is argued to be unsuitable for projection, because this may lead to biologically unreasonable projections. He therefore assumes a mean reverting process with non-zero mean (AR(1) with a constant) to project the other period effects. O'Hare and Li (2011) introduce M13 and apply it to data from a whole range of developed countries from 1950 to 2006. The proposed model is a modification of Plat's model, and they therefore use the same ARIMA-specifications for the period and cohort effects as in Plat (2009a).

Plat (2009a) considers two approaches for calibrating cohort effects: (i) estimate the effect for all cohorts available, and (ii) estimate the effect only for cohorts older than 1946. He argues that for cohort effects that are estimated only on younger ages, it is not sure whether they will persist in the future, and that they therefore should not be used to project mortality rates for the elderly. The cohort effect is then projected using a mean reverting process with mean zero, so it contains no trend. It is not included for younger ages when calibrating the model, while in projecting mortality rates it is. A possible solution to this inconsistency is to multiply the cohort effect with a parametric function of age. We do not investigate this in this thesis; we set the cohort effects equal to zero for the models M9 and M13 when there are no observations available related to age 60 or higher, corresponding to the choice in Plat (2009a).

Haberman and Renshaw (2011) consider the models listed in Table 2.1, except for M2A and M13, and they consider the Lee-Carter model extended with a cohort effect instead of the M3 specification. The models are fitted on England and Wales data from 1961 to 2007. To project mortality these authors fit a multivariate random walk with drift for all period effects. They argue that extrapolation of the cohort effect

should be avoided, because there is no justification to treat the cohort effect and the period effect independently. Therefore, they focus on modeling life expectancy and annuity values for *existing* cohorts. They conclude that the presence of a mild degree of curvature in the main period effect poses projection problems for some of the mortality models investigated.

Lovász (2011) considers several models for Finnish (1950 to 2009) and Swedish (1950 to 2008) data. He models the period effects as in Cairns et al. (2011a) and Haberman and Renshaw (2011) by assuming a multivariate random walk with drift. For the cohort effects he chooses the ARIMA(p, d, q)-process that is optimal in terms of BIC. He considers the combinations $d \in \{0, 1, 2\}$ and $(p, q) \in \{0, 1, 2\}$, and for those datasets the optimal ARIMA specifications are always integrated, possibly with a lag included (ARIMA$(p, 1, 0)$); two times differencing is never found to be optimal.

Kleinow and Richards (2016) investigate different ARIMA-models for the period effect in the Lee-Carter model. They select an ARIMA(p, d, q)-specification using the AIC, and show that this leads to more plausible forecasts than when using a random walk with drift. However, they further show that complex ARIMA-specifications may lead to instable projections, and they therefore conclude that a simpler model may be preferred if this results in more stable projections.

The papers mentioned above all use a random walk with constant drift for the first period effect, and often also for the other period effects. However, different calibration periods are used in different studies, and projections based on a random walk with constant drift are potentially highly sensitive towards the calibration period as reported in Booth et al. (2002) and Denuit and Goderniaux (2005). Furthermore, factors like medical advances (Bots and Grobbee (1996)) and health system reforms (Moreno-Serra and Wagstaff (2010)) have an impact on the speed of the mortality improvements, and Yang et al. (2014) find a correlation between the economic factors gross domestic product and unemployment rates on the one hand and life expectancy and health prevalence rates on the other hand. Dropping the assumption that the drift in the random walk must be *constant* may therefore be a way to improve model performance, and several authors proposed different methods to reduce the sensitivity with respect to the calibration period.

Optimal calibration period. Booth et al. (2002) note that a random walk with constant drift may not be appropriate over the whole period of available mortality data, so they propose to restrict the calibration period. The last year is determined by the most recent data available, and the first year is chosen by optimizing the fit of the random walk with drift model relative to the fit of the Lee-Carter model. They note that age effects may change through time and that by optimizing the calibration period, the age effects are chosen more appropriately for the purpose of projecting mortality rates.

Denuit and Goderniaux (2005) approximate the period effect κ_t by a straight line

which is fitted using OLS, and then choose the calibration period which optimizes the corresponding adjusted R^2 of the OLS fit. Li et al. (2015) include the length of the calibration period in their parameter space in a Bayesian framework. Projections resulting from different calibration periods are weighted by their posterior distribution likelihoods, to arrive at one single projection.

Regime switching models. Milidonis et al. (2011) calibrate the Lee-Carter model on US data for the ages 0-100 in the years 1901-2005. They propose a regime switching model with two regimes for the differences of the κ_t. The two regimes are allowed to have different means as well as different variances, and the estimation results reveal that the variance differs substantially between the two regimes. Based on information criteria and a likelihood ratio test they conclude that for the data set considered, the regime switching model outperforms the random walk with drift.

Hainaut (2012) extends the regime switching model to model M1A and applies it to French data for the ages 20-100 in the years 1946-2007. He concludes that the improvement in log likelihood is significant compared to the standard Lee-Carter model and the extension in Milidonis et al. (2011).

Structural changes in trend stationary models. Li et al. (2011) calibrate the Lee-Carter model on England & Wales and US data for the ages 0-99 in the years 1950-2006 (males and females combined). They perform a unit root test on the time series κ_t, which means that they test the null hypothesis of a random walk with constant drift versus the alternative hypothesis of a broken-trend stationary model. The broken-trend stationary model implies that the mortality trend κ_t fluctuates around a deterministic trend. The deterministic trend is piecewise linear and is estimated by regressing κ_t on t and an intercept. Dummy variables are used to allow the trend to change once in the calibration period, but the trend does not have to be continuous.

The authors use the test introduced in Zivot and Andrews (1992) to determine whether κ_t follows either a difference stationary process or a broken-trend stationary process. For both data sets they conclude that a broken-trend stationary process is preferred, and they use the latest trend for predictions. Since this is a trend stationary process, predictions from this model do not lead to confidence intervals that become wider over time.

Sweeting (2011) calibrates the original CBD-model (M5) on England & Wales data for the ages 60-89 in the years 1841-2005. He assumes a broken-trend stationary model as in Li et al. (2011), but allows for multiple structural changes and imposes continuity on the trend. He then fits distributions for the frequency and the severity of the changes in the trend, and uses these distributions for forecasting. Structural changes are tested for significance using the Chow test (Chow (1960)). Since changes in the mortality trend are included in forecasting, the prediction intervals for life expectancies at age 65 are much wider than generally found in the literature.

Structural change in difference stationary models. Coelho and Nunes (2011) consider the Lee-Carter model for a variety of countries for the ages 0-99 in the years after 1950.[6] They perform a unit root test as suggested by Harvey et al. (2009) and Harris et al. (2009) that allows for a single structural change both in the trend stationary and in the difference stationary model, where Li et al. (2011) only allow for a single structural change in the trend stationary model. They analyze 18 countries for both males and females. For all these data sets, the trend stationary model with a possible structural change is rejected 33 out of 36 times in favor of a difference stationary model with a possible structural change. In 21 out of 36 data sets a structural change is detected.

O'Hare and Li (2015) investigate the impact of a single structural change on mortality models beyond Lee-Carter. They apply the methodology for difference stationary time series to the models M1 (Lee-Carter), M5 (CBD), M9 (Plat) and M13 (O'Hare and Li). They find that in mortality models other than the Lee-Carter model, a structural change is often detected as well, and that allowing for a structural change can substantially improve the quality of forecasts, if this quality is measured in Mean Absolute Error or Root Mean Squared Error.

The papers above investigate the presence of structural changes in the period effects, and the latest structural change is then used for projection purposes. Hári et al. (2008a) define the Lee-Carter using a random walk with *time-varying drift* in a state space framework. The drift term is allowed to change over time, but they assume the drift parameter reverts back to a mean that is to be estimated from the data. This approach differs from the previously mentioned approaches, since the drift parameter is allowed to slightly change every year instead of changing sporadically as a result of structural change. An effect of allowing for a time-varying drift is that the prediction intervals become significantly wider. They find that their model performs similar to the original Lee-Carter model in terms of the cumulative sum of squared deviations of one period ahead in-sample forecasts.

Discussion of suggested forecasting approaches. When regime switching models are applied to mortality models, it is assumed that regimes observed in the past will occur in the future. Changes in mortality dynamics may be a result of (among others) changes in lifestyle and in health care systems. For example, in the Netherlands changes in smoking habits have been an important driver of changes in mortality, which contributed to increasing (1950-1970) and decreasing (from 1970 onwards) mortality rates (Janssen et al. (2007)). Since it is difficult to predict whether the same regimes of the past will occur again in the future, we prefer not to use regime switching models for the prediction of period and cohort effects.

Optimization of the calibration period as in Booth et al. (2002) and Denuit and

[6] The dataset depends on the data availability per country.

Goderniaux (2005) has appealing characteristics. Since older data points are excluded, the age effects are based on more recent data and are therefore more appropriate for forecasting than when all data is included. However, this approach may lead to short calibration periods which gives more volatile parameter estimates and projections. Further, by excluding data the researcher implicitly chooses not to explain part of the available data. Finally, these methods have been applied to the Lee-Carter model, but they are not easily transferable to multi-factor models, since different factors may suggest different calibration periods.

In Chapter 3 we extend the work of Coelho and Nunes (2011). We will use recent information on mortality dynamics, but we use the entire calibration period to estimate the variability in the mortality dynamics. We allow for multiple structural changes in the estimation of a random walk with drift, and we use the most recent drift for prediction purposes.

2.4 Bayesian implementation of the LC-model

In the previous sections, we had to estimate two different sets of parameters before we could create mortality forecasts for the Lee-Carter model. First we estimated α_x, β_x and κ_t, and afterwards we estimated parameters for the time series model. In that approach the κ_t are treated as parameters that need to be estimated in the first step, whereas in the second step we consider them as observations for which we estimate a model. However, the period effect κ_t is actually a latent effect with its own distribution (specified through the time series model), and therefore both the mortality model and the time series model should be estimated simultaneously.

It is possible to estimate the two parts simultaneously in a frequentist setting, but in a full maximal likelihood method the latent variables κ_t should then not be maximized over (as free parameters), but integrated out (over all their possible values). Due to the high dimension of the latent time series, this is often not feasible in practice. Therefore, it is more common to use a Bayesian implementation for the joint estimation problem. A Bayesian approach has the additional advantage that information on parameter uncertainty is also obtained after the model calibration, as we will show later on.

We first provide an introduction on Bayesian inference and the Markov Chain Monte Carlo methods that we use for this purpose. Then, we define the priors and derive the posterior distributions for the Lee-Carter model in a Bayesian setting, and we show the results when this model is applied to mortality data of Dutch males.

2.4.1 Bayesian inference and Markov Chain Monte Carlo methods.

Suppose that we have observed the data points $\boldsymbol{x} = (x_1, \ldots, x_n)$, and the outcomes of the random variables X_i are distributed as $X_i \sim p(x_i|\boldsymbol{\theta})$, so $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ are the parameters that specify the distribution of X_i. We specify a *prior distribution* on $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ are hyperparameters needed to specify the prior distribution. The prior distribution can be interpreted as the distribution of the parameter when we have not seen the data.

For inference purposes, we need to determine the distribution of $\boldsymbol{\theta}$ *given* the data, $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$, also called the posterior distribution. Using Bayes' rule we find that this can be represented as

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})}{p(\boldsymbol{x}|\boldsymbol{\alpha})} \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha}). \tag{2.16}$$

The right-hand side of (2.16) does not involve complicated integrals, but we still cannot easily sample from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$. However, we can use a method known as Markov Chain Monte Carlo (MCMC) to construct a sequence of samples $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^n, \ldots$ with a distribution that converges to the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$.

Markov Chain Monte Carlo. The following is based on Smith and Roberts (1993). Suppose we want to sample from the distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and $\boldsymbol{\theta} \in \mathcal{Z} \subseteq \mathbb{R}^p$ but cannot do this directly. Instead, we can construct a Markov chain with state space $\mathcal{Z}$, which is straightforward to simulate and whose equilibrium distribution is $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$. If we run the chain for a long time, simulated values of the chain can be used to study features of interest of $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$. Such features can be the expected value and quantiles of the distribution itself, but also of the distribution of transformations of the parameters $\boldsymbol{\theta}$, which may be more relevant in practice.

Under suitable regularity conditions, asymptotic results exist which clarify how the sample output from a chain with equilibrium distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$ can be used to mimic a random sample from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$ or to estimate the expected value, with respect to $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$, of a function $f(\boldsymbol{\theta})$. Consider a realization $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^n, \ldots$ from an appropriate chain. Typical asymptotic results include

$$\boldsymbol{\theta}^n \xrightarrow[n\to\infty]{d} \boldsymbol{\Theta} \sim p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha}), \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}^i) \xrightarrow[n\to\infty]{} E_{p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{\alpha})}\{f(\boldsymbol{\theta})\}.$$

We will therefore use MCMC to generate samples of the parameters with a distribution that converges to the posterior distribution of the parameters. We now introduce two algorithms that can generate realizations for the Markov chain: Gibbs sampling and Metropolis(-Hastings) sampling.

Gibbs sampling. For all parameters we have the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$ based on (2.16). In some cases this will lead to an expression that is proportional to a known distribution, and in that case we can directly sample from that distribution. Define the posterior distribution for θ_i conditional on the observed data, the hyperparameters and all other parameters as $p(\theta_i|\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\theta}_{-i})$ with $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1} \ldots, \theta_p)$, and suppose we pick some arbitrary starting values $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_p^0)$. We can then successively generate samples from the full conditional distributions $p(\theta_i|\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\theta}_{-i})$ for $i = 1, \ldots, p$ as follows:

$$
\begin{aligned}
&\text{sample} \quad \theta_1^1 \quad \text{from} \quad p(\theta_1|\boldsymbol{x}, \boldsymbol{\alpha}, \theta_2^0, \theta_3^0, \ldots, \theta_p^0);\\
&\text{sample} \quad \theta_2^1 \quad \text{from} \quad p(\theta_2|\boldsymbol{x}, \boldsymbol{\alpha}, \theta_1^1, \theta_3^0, \ldots, \theta_p^0);\\
&\qquad\qquad \vdots\\
&\text{sample} \quad \theta_p^1 \quad \text{from} \quad p(\theta_p|\boldsymbol{x}, \boldsymbol{\alpha}, \theta_1^1, \theta_2^1, \ldots, \theta_{p-1}^1).
\end{aligned}
$$

At the end of this procedure, we have obtained a new realization for the parameter vector $\boldsymbol{\theta}$.

Metropolis(-Hastings) sampling. If the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$ is not proportional to a known distribution, we can use Metropolis-Hastings sampling (MH). Again, we are ultimately interested in generating a Markov chain $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^n, \ldots$ with state space $\mathcal{Z} \subseteq \mathbb{R}^p$ and limiting distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$.

Suppose we have current values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. The MH algorithm constructs a transition probability from $\boldsymbol{\theta}^n = \boldsymbol{\theta}$ to $\boldsymbol{\theta}^{n+1}$ as follows. Let $g(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denote a transition probability function or candidate generating function, such that, if $\boldsymbol{\theta}^n = \boldsymbol{\theta}$ is the current value, $\boldsymbol{\theta}'$ a sample drawn from the distribution $g(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is considered a candidate for $\boldsymbol{\theta}^{n+1}$. However, the candidate $\boldsymbol{\theta}'$ is not accepted with probability 1 but with probability $\phi(\boldsymbol{\theta}, \boldsymbol{\theta}')$. If the candidate is accepted, we set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}'$, otherwise we reject the candidate and set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}$.

There is much flexibility in what distribution to use as a candidate generating function, as long as the state space $\mathcal{Y}$ of the candidate generating function contains the state space $\mathcal{Z}$ of the parameter vector $\boldsymbol{\theta}$. In many cases, the random walk process is used to generate candidates, which means that $\boldsymbol{\theta}'$ is drawn from a Gaussian distribution with mean $\boldsymbol{\theta}$ and some variance s.

In order for the sequence to be a Markov chain with limiting distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$, the following should hold (Hastings (1970)):[7]

$$
p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})p(\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\theta}'|\boldsymbol{x}, \boldsymbol{\alpha})p(\boldsymbol{\theta}', \boldsymbol{\theta}). \tag{2.17}
$$

[7] Consider a Markov chain in a discrete state space with limiting distribution $\boldsymbol{\pi}$ and transition matrix $\mathbf{P}$. Then, the transition matrix $\mathbf{P}$ should satisfy the *reversibility condition* $\pi_i p_{ij} = \pi_j p_{ji}$ for all i and j. This property ensures that $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ holds (or equivalently $\sum_i \pi_i p_{ij} = \pi_j$ for all j).

Here, $p(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is a transition probability within the Markov chain defined by

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} g(\boldsymbol{\theta}, \boldsymbol{\theta}')\phi(\boldsymbol{\theta}, \boldsymbol{\theta}') & \text{if } \boldsymbol{\theta}' \neq \boldsymbol{\theta} \\ 1 - \sum_{\boldsymbol{\theta}''} g(\boldsymbol{\theta}, \boldsymbol{\theta}'')\phi(\boldsymbol{\theta}, \boldsymbol{\theta}'') & \text{if } \boldsymbol{\theta}' = \boldsymbol{\theta}. \end{cases}$$

The relationship in (2.17) holds when the acceptance probability is set as

$$\phi(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} \min\left\{ \dfrac{p(\boldsymbol{\theta}'|\boldsymbol{x}, \boldsymbol{\alpha})g(\boldsymbol{\theta}', \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})g(\boldsymbol{\theta}, \boldsymbol{\theta}')}, 1 \right\} & \text{if } p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})g(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0, \\ 1 & \text{if } p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})g(\boldsymbol{\theta}, \boldsymbol{\theta}') = 0. \end{cases}$$

This acceptance probability contains the expression $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$. However, in (2.16) we showed that this is proportional to $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\alpha})p(\boldsymbol{\theta}|\boldsymbol{\alpha})$. Further, since the term $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$ appears both in the numerator and in the denominator, we can often simplify the expression for the acceptance probability to a large extent. Finally, if the candidate generating function is symmetric (i.e. $g(\boldsymbol{\theta}, \boldsymbol{\theta}') = g(\boldsymbol{\theta}', \boldsymbol{\theta})$), the acceptance probability simplifies to

$$\phi(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} \min\left\{ \dfrac{p(\boldsymbol{\theta}'|\boldsymbol{x}, \boldsymbol{\alpha})}{p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})}, 1 \right\} & \text{if } p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})g(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0, \\ 1 & \text{if } p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})g(\boldsymbol{\theta}, \boldsymbol{\theta}') = 0. \end{cases}$$

Using a symmetric candidate generating function is a special case of MH sampling, which is called Metropolis sampling.

Implementation issues. In the previous paragraphs we provided the requirements to use the MCMC method to approximate the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha})$. When implementing the MCMC method in practice, there are several issues that need to be dealt with.

Burn-in period: When starting the Markov chain, we have to choose some initial values for the parameters. These values may be far outside the center of the desired posterior distribution, and it may take some time before the sample converges to the posterior distribution. Further, as we discuss below, some parameters need to be calibrated to ensure that the chain performs efficiently. Therefore, we disregard the first samples of the chain, which is referred to as the *burn-in period*.

Thinning: In a Markov chain, successive values of $\boldsymbol{\theta}^n$ will be correlated. If we need independent samples for inference purposes, we cannot directly use the Markov chain $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^n, \ldots$. Instead, we can apply *thinning* to the Markov chain by only taking every nth value, e.g. every 10th realization, in order to reduce correlation.

Candidate generating function: The candidate generating function should be chosen in such a way that the entire parameter space $\mathcal{Z}$ can be explored within the Markov chain. For parameters that are unrestricted, the random walk approach often works well. However, if parameters are restricted (e.g. correlation parameters), then more attention should be paid to the choice of an appropriate candidate generating function.

Spread of the candidate generating function: The current value in a Markov chain is commonly used as the mean of the candidate generating function. The *variance* of the candidate generating function is a parameter that needs to be calibrated with care. If the variance is set too low, the Markov chain may take longer to traverse the entire support of the density and low probability regions may be undersampled, whereas if the variance is set too high samples may be generated that have a high probability of being rejected. Both cases are likely to result in high autocorrelation across the samples.

Acceptance probability: In order to ensure the Markov chain traverses the entire support of the density in an efficient manner, Roberts et al. (1997) show that the acceptance probability should be approximately 25%, so we shall aim at an acceptance probability between 20% en 30%.

Checking convergence: It is important to check whether the chain has converged and whether the final samples (after thinning and disregarding the burn-in period) are no longer autocorrelated. Different approaches are used for this purpose, and we discuss some of these below.

1. We can check the autocorrelation function to check whether there is any autocorrelation left within the chain. Also, if the sequence of samples is plotted there should be no observable drift in the chain (such a plot is referred to as a traceplot);
2. We can use multiple chains with different starting values and create a number of Markov chains. Once the chains have converged, we can compare the distribution of the different chains. The starting values should not influence the final samples. Therefore, we shall compare the distribution of the parameters from different chains, and these density plots should give comparable results.
3. When we use different chains it is possible that one chain has already covered the entire posterior distribution whereas another chain has not yet explored a certain region. In that case, the variances within the different chains will be different. Suppose we use M chains of length N, with θ_{mn} the nth sample in the mth chain, and let $\hat{\theta}_m$ and $\hat{\sigma}_m^2$ be the sample posterior mean and variance of the mth chain, and $\hat{\theta}$ the mean of the MN samples.

Gelman and Rubin (1992) define the between-chains and within-chains variances by

$$B = \frac{N}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta})^2 \quad \text{and} \quad W = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}_m^2.$$

Next, they define the pooled variance as

$$\hat{V} = \frac{N-1}{N}W + \frac{M+1}{MN}B,$$

and they show that this is an unbiased estimator for the posterior variance of the parameters. The *potential scale reduction factor* (hereafter: Gelman-Rubin statistic) is defined as the ratio of the pooled variance $\hat{V}$ and the within-chains variance W, and if the chains have converged to the posterior distribution this factor should be close to 1.

2.4.2 Bayesian estimation of the LC-model

We now define a possible approach to implement the Lee-Carter model in a Bayesian framework using the MCMC algorithm as discussed above. We consider the Lee-Carter model as specified in (2.12) and (2.14):

$$D_{t,x}|\mu_{t,x} \sim \text{Poisson}(E_{t,x}\mu_{t,x}), \qquad \text{with} \qquad \ln\mu_{t,x} = \alpha_x + \beta_x\kappa_t,$$

and

$$\kappa_t = \kappa_{t-1} + \delta + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \text{N}(0,\sigma_\varepsilon^2).$$

In most papers the parameter constraints applied to the Lee-Carter model are $\sum_t \kappa_t = 0$ and $\sum_x \beta_x = 1$, see Table 2.2. However, if parameter constraints are defined in a Bayesian setting (for example to facilitate convergence of the algorithm), then these parameter constraints should be satisfied in the specification of the prior distributions. In this section we use the same parameter constraints as those used in Example 2.2 (but different from those listed in Table 2.2), since these constraints are more easily satisfied in prior distributions. Specifically, we use the constraints $\kappa_{t_1} = 0$ and $\sum_x \beta_x^2 = 1$.

We use the following prior distributions for the parameters in the Lee-Carter model.

Prior distribution for α_x. Following Czado et al. (2005) and Antonio et al. (2015) we use the following prior for α_x with $x \in \mathcal{X} := x_1, \ldots, x_X$:

$$e_x := \exp(\alpha_x) \overset{\text{iid}}{\sim} \text{Gamma}(a_x, b_x). \tag{2.18}$$

Prior distribution for $\boldsymbol{\beta} = (\beta_{x_1}, \ldots, \beta_{x_X})$. For the vector of parameters $\boldsymbol{\beta} = (\beta_{x_1}, \ldots, \beta_{x_X})$ we choose a prior distribution that automatically satisfies the constraint $\sum_x \beta_x^2 = 1$: the Von Mises-Fisher distribution, which has its origins in directional statistics (von Mises (1918) and Fisher (1953)). To the best of our knowledge, Antoniadis et al. (2004) were the first to use this distribution as a prior in Bayesian analysis. The prior distribution for $\boldsymbol{\beta}$ is denoted by

$$\boldsymbol{\beta} \sim \text{vMF}(\boldsymbol{\mu}_\beta, c_\beta), \tag{2.19}$$

for constants $\boldsymbol{\mu}_\beta$ (the mean direction vector) and c_β (the concentration parameter) with $\|\boldsymbol{\mu}_\beta\| = 1$ and $c_\beta > 0$. The probability density function is given by

$$f_X(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, c_\beta) = C_X(c_\beta) \exp\left(c_\beta \boldsymbol{\mu}_\beta^T \boldsymbol{\beta}\right), \tag{2.20}$$

where the normalization constant $C_X(c)$ equals

$$C_X(c) = \frac{c^{X/2-1}}{(2\pi)^{X/2} I_{X/2-1}(c)}, \tag{2.21}$$

with I_v the modified Bessel function of the first kind with order v. See Hoff (2009) for details on how to sample from this distribution.

Note that our approach differs from what is usually done in the actuarial literature (see for example Czado et al. (2005); Li (2014); Antonio et al. (2015)), in the sense that often transformations are applied in a Metropolis-Hastings step after a sample has been accepted. In our approach, every sample already satisfies the necessary constraints because of our choice of the priors.

Prior distribution for κ_t. We assume a random walk with drift for $t \in \mathcal{T} \backslash \{t_1\} = t_1 + 1, \ldots, t_T$ for the period effect κ_t:

$$\kappa_t = \kappa_{t-1} + \delta + \varepsilon_t, \quad \text{with} \quad \varepsilon_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma_\varepsilon^2) \quad \text{and} \quad \kappa_{t_1} = 0. \tag{2.22}$$

The prior distributions for the drift and variance parameters are specified by

$$\delta \sim \text{N}(\mu_\delta, \sigma_\delta^2), \tag{2.23}$$

$$\sigma_\varepsilon \sim \text{Uniform}(0, A_\varepsilon). \tag{2.24}$$

For variance hyperparameters, we follow Gelman (2006) who suggests the use of a uniform prior on σ instead of an Inverse-Gamma(ς, ς) prior which is often proposed for this parameter, because if the estimate of σ is close to zero, the posterior density will then be less sensitive to the choice of ς.

2.4.3 Derivation of posterior distributions

We now derive the posterior distributions for the model specified above. When these have been derived, we can implement the MCMC algorithm in 2.4.4. For convenience, we define the following variables

$$\begin{aligned} \boldsymbol{D} &= (D_{t_1,x_1}, \ldots, D_{t_T,x_X}), & \boldsymbol{E} &= (E_{t_1,x_1}, \ldots, E_{t_T,x_X}), \\ \boldsymbol{e} &= (e_{x_1}, \ldots, e_{x_X}), & \boldsymbol{\beta} &= (\beta_{x_1}, \ldots, \beta_{x_X}), & \boldsymbol{\kappa} &= (\kappa_{t_1}, \ldots, \kappa_{t_T}), \end{aligned}$$

and we define the collection $\boldsymbol{\Lambda}$ that contains both data and parameters:

$$\boldsymbol{\Lambda} = \{\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2\}.$$

Gibbs sampling for α_x. The α_x values are independent for different values of x. Therefore, the posterior distribution for a single $e_x = \exp(\alpha_x)$ with $x \in \mathcal{X}$ is given by

$$\begin{aligned} f(e_x|\boldsymbol{\Lambda}\backslash\{e_x\}) &= P(\boldsymbol{\Lambda})/P(\boldsymbol{\Lambda}\backslash\{e_x\}) \propto P(\boldsymbol{\Lambda}) \\ &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2) f(e_x) \\ &\propto \prod_{t\in\mathcal{T}} e^{-E_{t,x} e_x \exp[\beta_x \kappa_t]} \frac{(E_{t,x} e_x \exp[\beta_x \kappa_t])^{D_{t,x}}}{D_{t,x}!} \\ &\quad \times \frac{b_x^{a_x}}{\Gamma(a_x)} e_x^{a_x - 1} \exp[-b_x e_x] \\ &\propto \exp[-(b_x + d_x) e_x] \cdot e_x^{a_x + D_{\bullet x} - 1}, \end{aligned} \tag{2.25}$$

with $d_x = \sum_{t\in\mathcal{T}} E_{t,x} \exp[\beta_x \kappa_t]$ and $D_{\bullet x} = \sum_{t\in\mathcal{T}} D_{t,x}$. The last line is proportional to a Gamma$(a_x + D_{\bullet x}, b_x + d_x)$ distribution, with a_x and b_x as defined in (2.22). Therefore, we can use Gibbs sampling to draw a new value of e_x, which can subsequently be transformed into a new value of α_x.

Metropolis sampling for β_x. The posterior distribution for $\boldsymbol{\beta}$ is given by

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{\Lambda}\backslash\{\boldsymbol{\beta}\}) &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2) f(\boldsymbol{\beta}) \\ &\propto \prod_{x\in\mathcal{X}} \prod_{t\in\mathcal{T}} e^{-E_{t,x}^{\text{pop}} e_x \exp[\beta_x \kappa_t]} \frac{(E_{t,x} e_x \exp[\beta_x \kappa_t])^{D_{t,x}}}{D_{t,x}!} \times \exp(c_\beta \boldsymbol{\mu}_\beta^T \boldsymbol{\beta}). \end{aligned} \tag{2.26}$$

Given a current value $\tilde{\boldsymbol{\beta}}$ and scaling parameter d_β, we sample a proposal $\hat{\boldsymbol{\beta}}$ from the distribution vMF$(\tilde{\boldsymbol{\beta}}, d_\beta)$. The proposal distribution is symmetric, and the acceptance probability is thus given by:

$$\phi = \min\left\{ \frac{f(\hat{\boldsymbol{\beta}}|\boldsymbol{\Lambda}\backslash\{\hat{\boldsymbol{\beta}}\})}{f(\tilde{\boldsymbol{\beta}}|\boldsymbol{\Lambda}\backslash\{\tilde{\boldsymbol{\beta}}\})}, 1 \right\}.$$

Metropolis sampling for κ_t. Define $\boldsymbol{\kappa}_{-t} = (\kappa_{t_1}, \ldots, \kappa_{t-1}, \kappa_{t+1}, \ldots, \kappa_{t_T})$. The posterior distribution of κ_t for $t \in \mathcal{T}\backslash\{t_1\}$ is given by

$$\begin{aligned} f(\kappa_t|\boldsymbol{\Lambda}\backslash\{\kappa_t\}) &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2) f(\boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2) \\ &\propto \prod_{x \in \mathcal{X}} \exp\left(-E_{t,x} e_x \exp[\beta_x \kappa_t]\right) \exp\left(D_{t,x} \beta_x \kappa_t\right) \\ &\quad \times f(\kappa_t|\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2). \end{aligned} \tag{2.27}$$

The expression in the last line can be simplified:

- for $t = t_1 + 1, \ldots, t_T - 1$:

$$\begin{aligned} f(\kappa_t|\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2) &= P(\boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2) / P(\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2) \propto P(\boldsymbol{\kappa}, \delta, \sigma_\varepsilon^2) \\ &= f(\kappa_{t_2}|\delta, \sigma_\varepsilon^2) \cdots f(\kappa_{t_T}|\kappa_{t_T-1}, \delta, \sigma_\varepsilon^2) \cdot p(\delta, \sigma_\varepsilon^2) \\ &\propto f(\kappa_t|\kappa_{t-1}, \delta, \sigma_\varepsilon^2) f(\kappa_{t+1}|\kappa_t, \delta, \sigma_\varepsilon^2) \\ &= \frac{1}{2\pi\sigma_\varepsilon^2} \exp\left[-\frac{(\kappa_t - \kappa_{t-1} - \delta)^2}{2\sigma_\varepsilon^2}\right] \cdot \exp\left[-\frac{(\kappa_{t+1} - \kappa_t - \delta)^2}{2\sigma_\varepsilon^2}\right] \\ &\propto \exp\left[-\frac{\kappa_t^2 - 2\kappa_t\kappa_{t-1} - 2\kappa_t\delta}{2\sigma_\varepsilon^2}\right] \cdot \exp\left[-\frac{\kappa_t^2 - 2\kappa_t\kappa_{t+1} + 2\kappa_t\delta}{2\sigma_\varepsilon^2}\right] \\ &\propto \exp\left[-\frac{(\kappa_t - [\frac{1}{2}(\kappa_{t-1} + \kappa_{t-1})])^2}{2(\frac{1}{2}\sigma_\varepsilon^2)}\right]. \end{aligned}$$

  This last line is proportional to the pdf of a random variable $Y \sim \mathrm{N}\left(\frac{1}{2}(\kappa_{t-1} + \kappa_{t+1}), \frac{1}{2}\sigma_\varepsilon^2\right)$.

- for $t = t_T$:

$$\begin{aligned} f(\kappa_t|\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2) &\propto p(\kappa_{t_2}|\delta, \sigma_\varepsilon^2) \cdots p(\kappa_{t_T}|\kappa_{t_T-1}, \delta, \sigma_\varepsilon^2) \cdot p(\delta, \sigma_\varepsilon^2) \\ &\propto f(\kappa_t|\kappa_{t-1}, \delta, \sigma_\varepsilon^2). \end{aligned}$$

  This last line is proportional to the pdf of a random variable $Y \sim \mathrm{N}\left(\kappa_{t-1} + \delta, \sigma_\varepsilon^2\right)$.

Given a current value $\tilde{\kappa}_t$ and Metropolis sampling variance $s_{\kappa_t}^2$, we sample a proposal $\hat{\kappa}_t$ from the distribution $\mathrm{N}(\tilde{\kappa}_t, s_{\kappa_t}^2)$. This proposal distribution is symmetric, and the acceptance probability is thus given by

$$\phi = \min\left\{\frac{f(\hat{\kappa}_t|\boldsymbol{\Lambda}\backslash\{\hat{\kappa}_t\})}{f(\tilde{\kappa}_t|\boldsymbol{\Lambda}\backslash\{\tilde{\kappa}_t\})}, 1\right\}.$$

Gibbs sampling for δ. Define $\Delta\kappa_t = \kappa_t - \kappa_{t-1}$. The posterior distribution of δ is given by

$$\begin{aligned} f(\delta|\boldsymbol{\Lambda}\backslash\{\delta\}) &\propto f(\boldsymbol{\kappa}|\delta,\sigma_\varepsilon^2)f(\delta) \\ &\propto \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{t=t_1+1}^{t_T}(\Delta\kappa_t-\delta)^2\right]\cdot\exp\left[-\frac{1}{2\sigma_\delta^2}(\delta-\mu_\delta)^2\right] \\ &\propto \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left((T-1)\delta^2-2\delta\sum_{t=t_1+1}^{t_T}\Delta\kappa_t\right)-\frac{1}{2\sigma_\delta^2}\left(\delta^2-2\delta\mu_\delta\right)\right] \\ &\propto \exp\left[-\frac{1}{2\sigma_\varepsilon^2\sigma_\delta^2}\left(\delta^2\left[(T-1)\sigma_\delta^2+\sigma_\varepsilon^2\right]-2\delta\left(\sigma_\delta^2\sum_{t=t_1+1}^{t_T}\Delta\kappa_t+\mu_\delta\sigma_\varepsilon^2\right)\right)\right] \\ &\propto \exp\left[-\frac{1}{2a_\delta}\left(\delta-b_\delta\right)^2\right], \end{aligned} \tag{2.28}$$

with

$$a_\delta = \frac{\sigma_\varepsilon^2\sigma_\delta^2}{(T-1)\sigma_\delta^2+\sigma_\varepsilon^2}, \qquad \text{and}$$
$$b_\delta = \frac{(T-1)\sigma_\delta^2}{(T-1)\sigma_\delta^2+\sigma_\varepsilon^2}\cdot\left(\frac{1}{(T-1)}\sum_{t=t_1+1}^{t_T}\Delta\kappa_t\right)+\frac{\sigma_\varepsilon^2}{(T-1)\sigma_\delta^2+\sigma_\varepsilon^2}\cdot\mu_\delta.$$

The last line in (2.28) is proportional to the pdf of a random variable $Y \sim \mathrm{N}(b_\delta, a_\delta)$, so we can use Gibbs sampling to draw a new sample for δ.

Gibbs sampling for σ_ε^2. The prior distribution for σ_ε^2 is specified as $\sigma_\varepsilon \sim \mathrm{Uniform}(0, A_\varepsilon)$. Define $X = \sigma_\varepsilon$ and $Y = \sigma_\varepsilon^2 = X^2 = w^{-1}(X)$. The pdf for σ_ε is given by $f_X(x) = A_\varepsilon^{-1}\cdot 1_{[0\le x\le A_\varepsilon]}$. Using the transformation theorem, we derive the pdf for σ_ε^2 as follows:

$$f_Y(y) = f_X(w(y))\cdot|w'(y)| = A_\varepsilon^{-1}\cdot 1_{[0\le\sqrt{y}\le A_\varepsilon]}\cdot\tfrac{1}{2}y^{-\frac{1}{2}}.$$

Substituting $y = \sigma_\varepsilon^2$, we find that the pdf for σ_ε^2 is given by:

$$f(\sigma_\varepsilon^2) = A_\varepsilon^{-1}\cdot 1_{[0\le\sigma_\varepsilon\le A_\varepsilon]}\cdot\tfrac{1}{2}\sigma_\varepsilon^{-1} \propto \sigma_\varepsilon^{-1}\cdot 1_{[0\le\sigma_\varepsilon\le A_\varepsilon]}.$$

The posterior distribution of σ_ε^2 is given by

$$\begin{aligned} f(\sigma_\varepsilon^2|\boldsymbol{\Lambda}\backslash\{\sigma_\varepsilon^2\}) &\propto f(\boldsymbol{\kappa}|\delta,\sigma_\varepsilon^2)f(\sigma_\varepsilon^2) \\ &= \left[\prod_{t=t_1+1}^{t_T}\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\exp\left[-\frac{[\Delta\kappa_t-\delta]^2}{2\sigma_\varepsilon^2}\right]\right]\times\sigma_\varepsilon^{-1}\cdot 1_{[0\le\sigma_\varepsilon\le A_\varepsilon]} \\ &\propto (\sigma_\varepsilon^{-2})^{\frac{T}{2}}\exp\left[-(\sigma_\varepsilon^{-2})\cdot\frac{1}{2}\sum_{t=t_1+1}^{t_T}(\Delta\kappa_t-\delta)^2\right]. \end{aligned} \tag{2.29}$$

We again apply the transformation theorem, but this time using $X = \sigma_\varepsilon^2 = w(Y)$ and $Y = \sigma_\varepsilon^{-2}$, so $w(y) = y^{-1}$ and $|w'(y)| = y^{-2}$. Using the posterior distribution of σ_ε^2 we derive that the posterior distribution of σ_ε^{-2} is given by

$$\begin{aligned} f(\sigma_\varepsilon^{-2}|\mathbf{\Lambda}\backslash\{\sigma_\varepsilon^2\}) &= f(\sigma_\varepsilon^2|\mathbf{\Lambda}\backslash\{\sigma_\varepsilon^2\}) \cdot (\sigma_\varepsilon^{-2})^{-2} \\ &\propto (\sigma_\varepsilon^{-2})^{\frac{T-2}{2}-1} \cdot \exp\left[-(\sigma_\varepsilon^{-2}) \cdot \frac{1}{2} \sum_{t=t_1+1}^{t_T} (\Delta\kappa_t - \delta)^2\right]. \end{aligned} \tag{2.30}$$

Since the last line in (2.30) is proportional to the pdf of a random variable $Y \sim \text{Gamma}\left(\frac{T-2}{2}, \frac{1}{2}\sum_{t=t_1+1}^{t_T}(\Delta\kappa_t - \delta)^2\right)$, we can use Gibbs sampling to draw new values of σ_ε^{-2} which can be transformed into values for σ_ε^2.

2.4.4 Case study: Dutch males

We now apply the Bayesian setting of the Lee-Carter model for Dutch males aged 0-90 in the years 1950-2000. We compare the Bayesian estimates and projections with those from a frequentist setting as illustrated in Section 2.2 and 2.3.

Details of the MCMC algorithm. We run four MCMC chains in parallel. For the parameters α_x, β_x and κ_t we use the frequentist estimates $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{\kappa}_t$ to define starting samples. These estimates are obtained using Poisson maximum likelihood estimation based on (2.12). However, for this calibration we used the restrictions $\sum_x \beta_x^2 = 1$ and $\kappa_{t_1} = 0$ to ensure the restrictions can be imposed on the specification of the prior distribution. In each chain we add some random Gaussian noise to these parameter estimates to obtain different starting values. Using the starting values for β_x and κ_t, we obtain maximum likelihood estimates in each chain for σ_β^2, δ and σ_ε^2, and we use these as initial values for the sampling of the hyperparameters. The constants that complete the specification of the prior distributions and the sampling variances used in the Gibbs and Metropolis sampling algorithms are chosen as follows:

- To ensure the prior does not contain too much information, we use $a_x = b_x \cdot \exp(\hat{\alpha}_x)$ and $b_x = 0.01$, see Antonio et al. (2015). This way, $\mathbb{E}[\exp(\alpha_x)] = \exp(\hat{\alpha}_x)$ with large variance.
- For $\boldsymbol{\beta}$ we use $\boldsymbol{\mu}_\beta = \frac{1}{\sqrt{X}} \cdot \mathbf{1}_X$ with $\mathbf{1}_X$ a vector with ones of length X, and $c_\beta = 0.01$.
- We use $\mu_\delta = \hat{\delta}$ (the maximum likelihood estimate of the drift, as obtained from the frequentist approach) and $\sigma_\delta^2 = 0.5^2$. For the variance hyperparameter we use $A_\varepsilon = 10$.
- For the scale parameters used in the proposal densities, we start with $d_\beta = 10^5$ and $s_{\kappa_t}^2 = 0.05^2$.

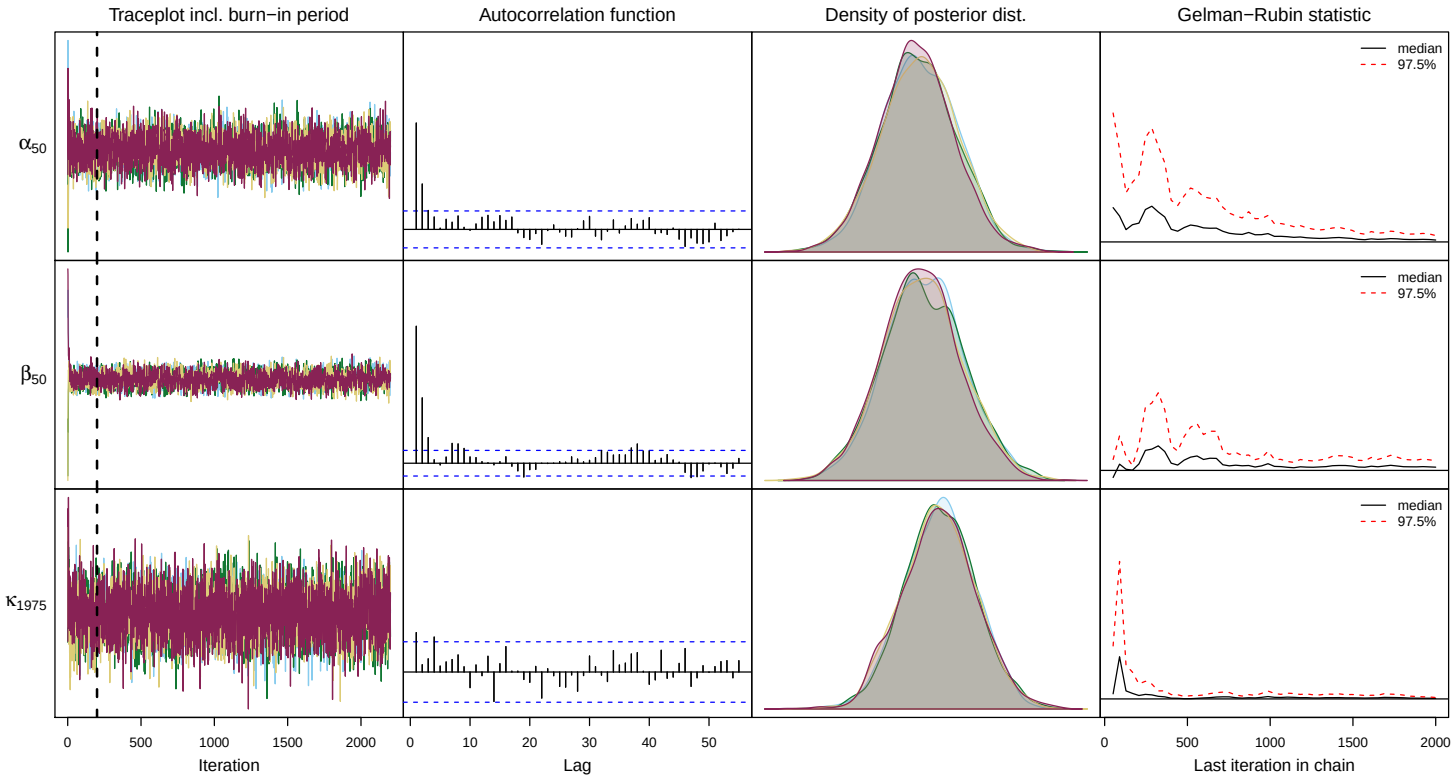


Figure 2.4: Convergence diagnostics for selected variables. First column: traceplots for the complete sample from the MCMC procedure; the vertical dashed line separates the burn-in period (left) and the final sample (right). Second column: the autocorrelation function for the final sample from the first chain. Third column: density plots from the final sample for all four chains. Fourth column: Gelman-Rubin statistic showing the convergence between the different chains.

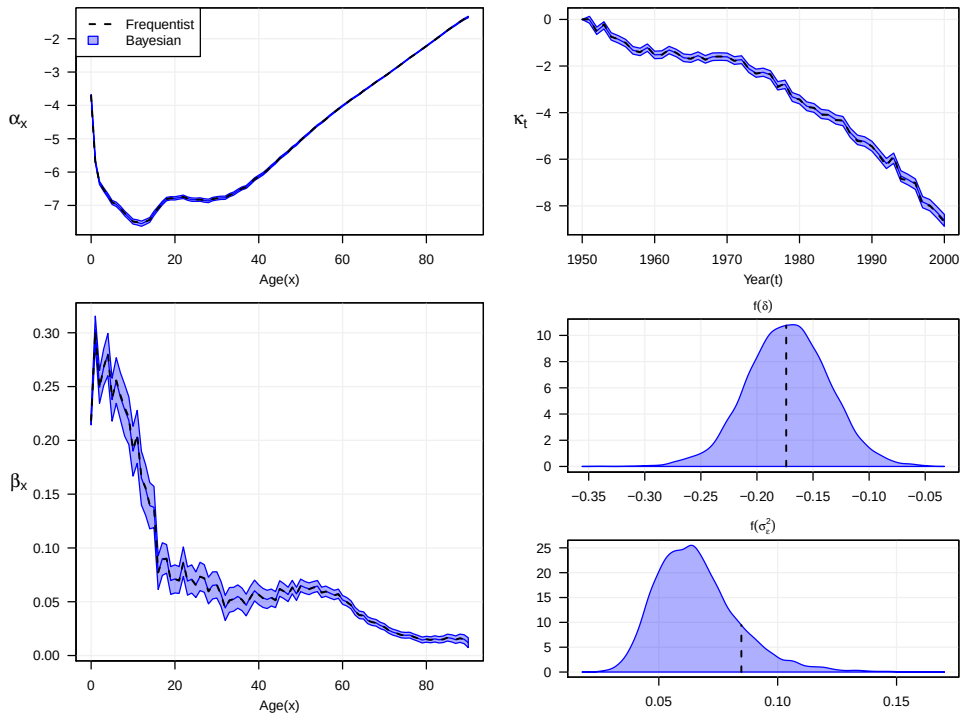


Figure 2.5: Parameter estimates for the Lee-Carter model applied to mortality data on Dutch males aged 0-90 in the years 1950-2000. For the frequentist method we show the maximum likelihood estimates, and for the Bayesian method we show the 95% credible interval (equal-tailed) of the posterior distributions.

Convergence diagnostics. We run 1,100,000 iterations in each chain of the MCMC algorithm. We save every 500th iteration, and during the first 100,000 iterations we calibrate the scale parameters of the proposal distributions every 100th iteration.[8] Our final sample size is 8,000. Our trace plots show good mixing properties, the calculated Gelman and Rubin statistics converge rapidly towards one, and density plots of the parameters in different chains overlap almost perfectly. Some convergence statistics are shown in Figure 2.4.

Estimation results. Figure 2.5 shows the parameter estimates for the frequentist and the Bayesian approach. The frequentist estimates are represented by black dashed

[8]The large number of required iterations is due to the high dimension of our model. However, since our Metropolis-Hastings algorithm for $\boldsymbol{\beta}$ consists of only one step, instead of the usual loop over all ages (see e.g. Czado et al. (2005) and Antonio et al. (2015)), using the Von Mises-Fisher distribution as proposal density speeds up the algorithm considerably.

lines, and for the Bayesian approach we show the 95% equal-tailed credible intervals.[9] We observe that the parameter estimates for α_x, β_x and κ_t generated by the two approaches are similar. Hence, estimating the mortality model and the time series model simultaneously instead of using a two-step procedure has little impact on the estimates in this case. Also, the frequentist estimate of δ is close to the median and mean of the corresponding posterior distribution. However, the frequentist estimate of σ_ε^2 is somewhat higher than the mean and median of the posterior distribution.

Forecasting results. In Figure 2.6 we show predictions of κ_t and $\ln \mu_{t,x}$ for $x = \{20, 50, 80\}$. In the frequentist case, we simulate 800,000 scenarios for future values of κ_t using fixed parameter estimates but with random values of ε_t. In the Bayesian case, we also take parameter uncertainty into account in the generation of future mortality scenarios. This is achieved by generating 100 simulations for κ_t in each MCMC sample, resulting in a total of 800,000 mortality scenarios.

In the top left graph in Figure 2.6 we see that for short prediction horizons the uncertainty in future values of κ_t for $t > T$ increases rapidly. While part of this uncertainty originates from uncertainty in κ_T, most of the uncertainty in the short term is caused by uncertainty in future values of ε_t. As a result, in the short term there is little difference in the forecasts of κ_t between the frequentist and the Bayesian case. Further, the posterior distribution for σ_ε^2 shows both smaller and larger values of σ_ε^2, and the uncertainty in future values of ε_t can thus be smaller and larger. In the top left graph in Figure 2.6 we observe that the uncertainty in σ_ε^2 has little impact on the width of prediction intervals. For projections of mortality rates however, uncertainty in the parameter β_x results in wider Bayesian prediction intervals.

On the longer time horizon, the prediction intervals from the Bayesian approach do become wider than in the frequentist case. This is caused by uncertainty in the drift parameter δ. In the short term this parameter has little impact, but smaller or larger values of δ have a cumulative effect that becomes apparent on the longer time horizon.

In Chapter 4 we will investigate what the impact is of the different sources of uncertainty on predicted numbers of deaths.

2.5 Portfolio-specific mortality

In Section 1.2 we mentioned that remaining life expectancy depends on factors such as income, education and lifestyle. The participants in a pension fund (or policyholders

[9] The equal-tailed interval is defined in such a way that the probability of being below the interval is as high as being above it. Another type of intervals often used in the Bayesian context is the highest posterior density interval. This is defined as the narrowest interval that captures the required probability mass of the posterior distribution. If a posterior distribution is symmetric, the equal-tailed and highest posterior density intervals are the same.

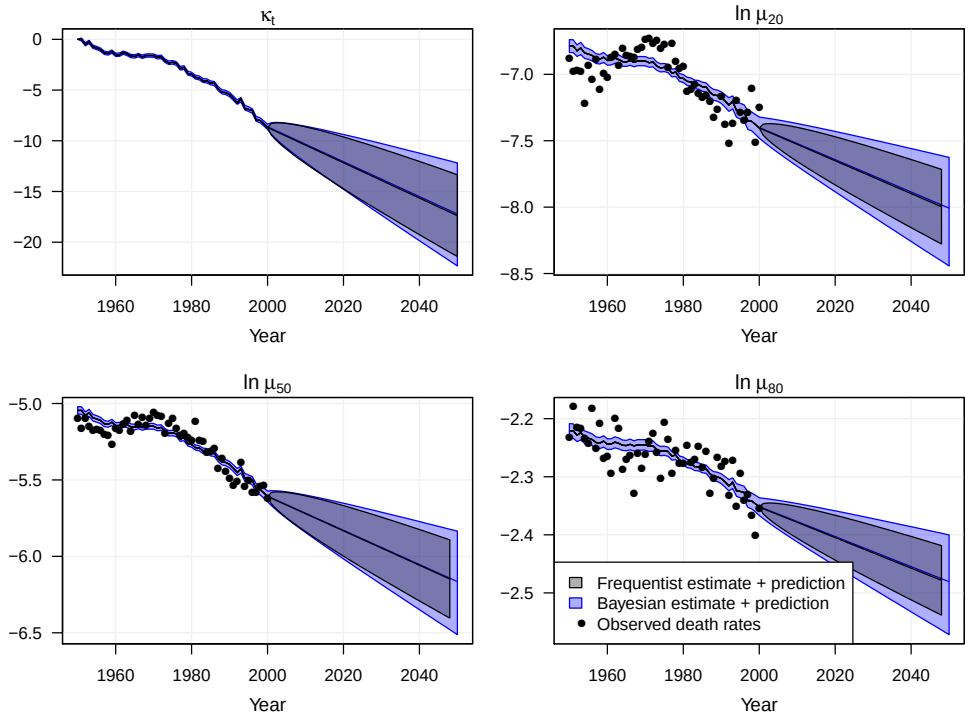


Figure 2.6: Estimates of and projections for κ_t and $\ln \mu_{t,x}$ for $x = \{20, 50, 80\}$ using the results from Figure 2.5. For the frequentist method we show (in gray) the maximum likelihood estimates and projections, taking into account uncertainty in ε_t, and for the Bayesian method we show (in blue) the 95% credible interval (equal-tailed) of the posterior distributions and projections, taking into account uncertainty in all parameters. We show the 2.5th, 50th and 97.5th percentile of the posterior distributions and prediction intervals.

in an insurance company) share several risk factors with other participants in the fund, but there may be great differences with respect to the general population. Pension funds should take these differences into account for pricing and reserving purposes. This means that they cannot simply use population-wide mortality forecasts, but they have to adjust these mortality forecasts to ensure they are appropriate for the portfolio for which the mortality forecasts are used.

Historically, crude methods such as age-shifting were used, which means that the mortality rate q_x is replaced with q_{x+s} with s either positive or negative, see Pitacco et al. (2009). A different approach used in practice is to make mortality rates depend on the time since inception of the contract, e.g. using $\tilde{q}_x = q_x \cdot \rho(x - s, s)$ with s the time since inception of the contract and $x - s$ the age at inception of the contract.

Such crude methods were sufficiently accurate in the past, because funding ratios were high and the focus in risk management was mainly on the best estimate of liability values. Nowadays, with funding ratios closer to 100%, there is more attention for the uncertainty in the liabilities. This requires information about the uncertainty in portfolio-specific mortality rates, and statistically sound models are needed to obtain it.

The methods that are suggested in the literature can be divided into two main categories. In Section 2.5.1 we discuss the approach in which the single population mortality models are extended to multiple populations, and in Section 2.5.2 we discuss how historically observed portfolio mortality is explained using risk factors.

2.5.1 Multiple-population mortality models

Mortality developments in a population can be strongly time-varying. Periods of small mortality improvements may be followed by periods of larger ones, and a rapidly changing mortality trend is difficult to project. Therefore, models have been proposed to incorporate information from different but comparable populations in the estimation process. This can lead to a more stable, global mortality trend, which also provides insight in population-specific deviations from the general pattern. A disadvantage is that a sufficiently large historical data is needed to analyze such population-specific deviations. If there is only a limited historical dataset available for a portfolio, application of the multiple-population approach to portfolio data must therefore contain a careful analysis of the uncertainty in the estimates.

The Lee-Carter model is designed for a single country, but this model can be extended to multiple countries. For example, Li and Lee (2005) propose an augmented common factor model for multiple populations (which are indexed by i)

$$\ln m^i_{t,x} = \alpha^i_x + B_x K_t + \beta^i_x \kappa^i_t + \epsilon_{t,x,i}, \qquad \epsilon_{t,x,i} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_i). \tag{2.31}$$

Under the restriction that $\sum_t K_t = \sum_t \kappa^i_t = 0$ for all i (similar to the original Lee-Carter restriction), α^i_x is the average logarithmic central death rate for age x in population i over time. The term $B_x K_t$ represents the common factor for all populations, and the term $\beta^i_x \kappa^i_t$ is a population-specific, age-dependent mortality development. Li and Lee (2005) estimate this model using Singular Value Decomposition, whereas Antonio et al. (2015) use a Bayesian framework. For a related alternative, where different populations share a common age-effect for mortality improvements, see Kleinow (2015). He finds that mortality forecasts are more accurate when the age-effect is the same for all countries rather than estimated differently for all countries.

The model proposed by Li and Lee (2005) is intended for mortality of different countries, and no country is assumed to be dominant in the mortality developments of other countries. Dowd et al. (2011) investigate mortality for two populations where

the populations are not necessarily considered equal. They propose the following 'gravity' model:

$$\ln m_{t,x}^{(i)} = \alpha_x^{(i)} + \kappa_t^{(i)} + \gamma_{t-x}^{(i)}, \qquad i = 1, 2, \tag{2.32}$$

with $\gamma_{t-x}^{(i)}$ a term representing a cohort effect. The first population is assumed to be the independent population, and mortality in the second population converges to mortality in the first population. They impose that the period effects $\kappa_t^{(i)}$ gravitate to one another, and the gravitational pull is characterized by the following time series specification:

$$\begin{aligned}
\kappa_t^{(1)} &= \kappa_{t-1}^{(1)} + \mu^{(1)} + C^{(11)} Z_t^{(1)}, \\
\kappa_t^{(2)} &= \kappa_{t-1}^{(2)} + \phi(\kappa_{t-1}^{(1)} - \kappa_{t-1}^{(2)}) + \mu^{(2)} + C^{(21)} Z_t^{(1)} + C^{(22)} Z_t^{(2)}.
\end{aligned}$$

The term ϕ represents the gravitational pull: the spread between the two period effects, $\kappa_{t-1}^{(1)} - \kappa_{t-1}^{(2)}$, reduces over time if $0 < \phi < 1$, and the pull is stronger for larger ϕ. Cairns et al. (2011b) estimate parameters for a similar gravity model using a Bayesian approach. By defining the dependence between the two populations slightly differently, they arrive at a specification that can be used for a combination of a dominant and a subordinate population, but also for a combination of two equal-sized populations. This makes it suitable to model mortality in different countries but also for mortality in a country and in a large pension fund.

The papers discussed above consider multiple-population models mainly to describe the evolution of mortality for a collection of countries. Villegas and Haberman (2014) consider mortality in different groups *within* a single country. Specifically, they consider five different socioeconomic classes in England. Mortality for the reference population is described using an extension of the Lee-Carter model, and mortality for different socioeconomic classes is defined relative to the population. Haberman et al. (2014) use a similar approach for an insurance portfolio but consider a wide collection of models. For portfolios with large exposures and sufficient historical observations, the authors suggest to use M7 as defined in Table 2.1 for the general population and M5 for the difference between the population and the portfolio.

All these approaches require that there is a sufficiently large dataset. Otherwise, time-dependent effects may be difficult to forecast, and it may be more appropriate to assume no time dynamics in the portfolio.

2.5.2 Explaining portfolio mortality using risk factors

As an alternative for the multiple-population approach, we may try to explain the relative difference between the baseline mortality rate and mortality observed in the portfolio using observable risk factors.

Dutch life insurers and pension funds often define mortality rates as a function of insured amounts or liabilities. The total value of the liabilities is determined as the product of the insured amount or accrued rights and an appropriate annuity factor (summed over all participants). As a result, the liabilities are more sensitive to mortality among participants with large insured amounts. It is therefore more relevant to accurately estimate the mortality rates for participants with large insured amounts than for participants with smaller insured amounts.

The following example is borrowed from Koninklijk Actuarieel Genootschap (2012). Consider an insurance portfolio with two types of policyholders: 100 of type A and 100 of type B. During year t, six policyholders of type A die and two of type B. Since there are no risk factors to distinguish mortality between both types of policyholders, the death ratio is estimated at $8/200 = 0.04$.

Suppose now that we know that each policyholder of type A had a present value of pension benefits equal to 1,000, and for each policyholder of type B the present value equals 5,000. Then, the total liabilities at the beginning of the year equal $100 \cdot 1,000 + 100 \cdot 5,000 = 600,000$. Given an equal number of deaths for both types of policyholders, the total release of provision (ignoring any benefit payments) equals $6 \cdot 1,000 + 2 \cdot 5,000 = 16,000$. If we determine the mortality rate as the fraction of the total provision released in a year, the death ratio in this example equals $16,000/600,000 = 0.027$.

The above example illustrates that death rates weighted by insured amounts can be substantially different from death rates based on numbers of deaths. Therefore, Plat (2009b) defines portfolio-specific factors as the ratio between mortality in the population and mortality in a portfolio as

$$P_{t,x} = \frac{m^A_{t,x}}{m^{\text{pop}}_{t,x}}, \tag{2.33}$$

where $m^A_{t,x}$ is the observed death rate in the portfolio based on insured amounts, and $m^{\text{pop}}_{t,x}$ is the observed death rate in the population. As an example, he models realized portfolio-specific factors assuming a linear effect in age:

$$P_{t,x} = a_t + b_t x + \varepsilon_{t,x}, \qquad \varepsilon_{t,x} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2_\varepsilon). \tag{2.34}$$

The values of a_t and b_t are estimated using regression techniques, and portfolio-specific factors for future years are obtained by projecting a_t and b_t using time series models. For small portfolios, the observed portfolio-specific factors $P_{t,x}$ may become volatile due to individual mortality risk, which may complicate drawing conclusions on the significance of parameters.

Richards et al. (2013) model the force of mortality using individual observations, and therefore their model takes individual mortality risk into account in an appropriate manner. They use a time-varying version of the Makeham-Beard law to specify

the force of mortality:

$$\mu^i_{t,x} = \frac{e^{\epsilon_i} + e^{\alpha_i+\beta_i x+\delta(t-2000)}}{1 + e^{\alpha_i+\rho_i+\beta_i x+\delta(t-2000)}}. \tag{2.35}$$

The term e^{ϵ_i} is a constant rate of mortality independent of age or time, and is referred to as Makeham's constant. Further, the term α_i is the overall mortality level for individual i, and β_i is the rate of increase in mortality with age for individual i. Finally, the term ρ_i is a heterogeneity parameter, and δ represents a constant rate of change in overall mortality levels over time, normalized to the year 2000. They estimate the parameters on five years of historical portfolio data for individual lives. Since the model is based on observations of individual survival, their approach cannot be used when only aggregated portfolio data are available.

Gschlössl et al. (2011) use aggregated instead of individual mortality data. They only have five years of historical data, and they therefore do not include time dynamics in their model for portfolio-specific mortality. First, they estimate a smooth baseline force of mortality μ^b_i on portfolio data which depends on age only. Remaining heterogeneity is then captured by observable risk factors in a Poisson GLM framework:

$$D_i \sim \text{Poisson}(E_i\mu_i), \tag{2.36}$$

with

$$\ln \mu_i = \beta_0 + \beta_1 \ln \mu^b_i + \sum_{j=2}^{r+1} \beta_j x_{ij}. \tag{2.37}$$

In this specification, individual mortality risk is also appropriately taken into account through the Poisson specification. However, since the baseline mortality rate is estimated on portfolio data without a time trend, mortality rate forecasts are not easily obtained.

Olivieri (2011) view portfolio-specific mortality in a different perspective, namely in a Bayesian setting of the form

$$D_{t,x} \sim \text{Poisson}(E_{t,x}q^*_{t,x}Z_{t,x}). \tag{2.38}$$

Here, $q^*_{t,x}$ is a best estimate mortality rate published by an independent institution, and $Z_{t,x} \sim \text{Gamma}(\alpha_{t,x}, \beta_{t,x})$ is a random adjustment. Starting with values for $\alpha_{0,x}$ and $\beta_{0,x}$, subsequent values of $\alpha_{t,x}$ and $\beta_{t,x}$ can be computed in closed form when new mortality observations become available, since the Gamma distribution is used, which is the conjugate of the Poisson distribution. Kan (2012) considers a similar framework but uses the Lee-Carter model calibrated to the Dutch population to specify the baseline mortality rate $q^*_{t,x}$. This model provides insight in how the distribution of the portfolio-specific factors may evolve over time if new observations

become available. However, it is not trivial to extend this framework to incorporate other risk factors.

The papers discussed above often take population mortality rates as given or do not consider time dynamics. But when population mortality and portfolio-specific mortality are not modeled simultaneously, it is unclear how portfolio-specific mortality forecasts should be constructed and how uncertainty in those forecasts has to be evaluated. In Chapter 4 we combine the ideas from this section with the multiple-population approach and introduce a new method to simultaneously estimate population and portfolio-specific mortality. Further, in Chapter 5 we use the regression approach as introduced below to explain mortality in a Dutch pension fund.

Chapter 3

The impact of multiple structural changes on mortality predictions

This chapter is based on F. van Berkum, K. Antonio, and M. Vellekoop. The impact of multiple structural changes on mortality predictions. *Scandinavian Actuarial Journal*, 7:581 – 603, 2016.

3.1 Introduction

Mortality rates have improved substantially during the last century as has been shown in, for example, Cairns et al. (2008), Barrieu et al. (2012) and Antonio et al. (2017). Life insurance companies and pension funds therefore need to monitor and predict mortality improvements for proper pricing and reserving. It is also important for them to quantify the uncertainty in future mortality rates for regulatory purposes such as Solvency II.

Constructing mortality rate projections consists of two steps, namely (i) calibrating a mortality model on historical data, and (ii) forecasting future values for the time dependent parameters obtained in (i). The seminal paper by Lee and Carter (1992) introduced a stochastic mortality model that describes mortality improvements. This is a single factor model with age and period effects, but in Section 2.2 we discussed different extensions to the Lee-Carter model.

The projection of time-dependent effects in mortality models receives relatively little attention in the recent literature. The period and cohort effects are often projected using ARIMA-models, see Section 2.3.2. However, when structural changes are present, the time-dependent effects cannot be captured by standard ARIMA-models. The resulting mortality forecasts are also highly sensitive to the calibration period.

Alternatives have been proposed to tackle this problem, e.g. Booth et al. (2002) and Denuit and Goderniaux (2005) use a frequentist approach and Li et al. (2015) a Bayesian approach to choose an optimal calibration period, Milidonis et al. (2011) introduce regime switching models to mortality modeling, and Li et al. (2011), Sweeting (2011) and Coelho and Nunes (2011) introduce structural changes in trend stationary

and difference stationary processes.

In this chapter we extend the approach of Coelho and Nunes (2011). They allow for a single structural change in period effects. However, multiple structural changes may have occurred, as has been suggested for trend stationary processes by Sweeting (2011). We focus on the class of difference stationary processes and extend the approach of Coelho and Nunes (2011) by allowing for multiple structural changes in the period effects. We determine the structural changes in an objective manner (Bai and Perron (1998)) using the Bayesian Information Criterion. To evaluate the performance of this approach, we compare the projections using our approach to those obtained when no structural changes or a single structural change is allowed using the Dawid-Sebastiani scoring rule (Riebler et al. (2012)). Whereas the aforementioned papers often focus on a specific mortality model, we show results for Dutch and Belgian mortality data that are calibrated to a wide variety of mortality models. We include both models with and without cohort effects since recent results by Coelho and Nunes (2013) show that evidence of structural changes in models without cohort effects may disappear once cohort effects have been included.

The chapter is organized as follows. In Section 3.2 we present our approach for mortality forecasting when allowing for multiple structural changes in the period effects. We investigate the estimation and backtesting results in Section 3.3, and Section 3.4 concludes.

3.2 Proposed forecasting method

3.2.1 Forecasting period effects

In Section 2.3 we discussed a variety of approaches to forecasting mortality. When using regime switching models, it is assumed that mortality dynamics observed in the past will occur in the future. Consider for example the effect of smoking which severely affected mortality improvements in the Netherlands for many years. People are smoking less frequently nowadays than they did in the past, and therefore we do not expect smoking to have a similar impact on future mortality developments, so we will not use regime switching models. Optimization of the calibration period as in Booth et al. (2002) and Denuit and Goderniaux (2005) has appealing characteristics. For example, only the most recent data is used for calibration, and 'old' data which may not be appropriate for projection of future mortality, is disregarded. However, this method is not easily transferable to multi-factor models because for two period effects different calibration periods may be optimal, so we will not optimize the calibration period. Instead, we will use recent information on mortality dynamics, and we use the entire calibration period to estimate the variability in the mortality dynamics.

Following the findings from Coelho and Nunes (2011) and the fact that a random walk with drift seems to provide the best calibration results in the mortality literature,

we focus on the difference stationary process. However, we extend the approach of Coelho and Nunes (2011) and the work of O'Hare and Li (2015) in such a way that multiple structural changes can be detected, as multiple events in the past may have affected the speed of mortality improvements.

We assume a multivariate random walk with drift for the period effects. Each univariate series may experience multiple structural changes during the calibration period. We determine the optimal number of structural changes separately for each time series using an optimization criterion. The period effects are then projected using the latest drift parameters and the estimated covariance structure.

To determine the number of structural changes and their corresponding dates, we follow the methodology introduced in Bai and Perron (2003). Suppose we have at our disposal different period effects (indexed by i) $\kappa_t^{(i)}$ $(t = t_1, \ldots, t_T)$ and define the first-order differences $\Delta\kappa_t^{(i)} = \kappa_t^{(i)} - \kappa_{t-1}^{(i)}$ for $t = t_2, \ldots, t_T$. We estimate a random walk with a piecewise constant drift:

$$\Delta\kappa_t^{(i)} = \begin{cases} \beta_1 + \varepsilon_t, & t \leq c_1 \\ \ldots & \\ \beta_j + \varepsilon_t & c_{j-1} < t \leq c_j \\ \ldots & \\ \beta_{m+1} + \varepsilon_t, & c_m < t \end{cases} \tag{3.1}$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ are independent over time. We estimate this model using OLS, hence, we minimize the sum of squared residuals (SSR):

$$\text{SSR}(c_1, \ldots, c_m) = \sum_{j=1}^{m+1} \sum_{t=c_{j-1}+1}^{c_j} [\Delta\kappa_t^{(i)} - \beta_j]^2, \tag{3.2}$$

where $c_0 = t_1$ and $c_{m+1} = t_T$. In the model specification above, we distinguish m break points that divide the time series into $m+1$ periods with different drifts. Both the number of break points (m) and the dates of the break points $(c_1, \ldots, c_m)$ are unknown.

Let $\beta(C_m)$ denote the estimates $\{\beta_1, \ldots, \beta_{m+1}\}$ based on a given m-partition $(c_1, \ldots, c_m)$ denoted C_m. If we substitute $\beta(C_m)$ into (3.2), the estimated break points $(\hat{c}_1, \ldots, \hat{c}_m)$ are chosen in such a way that $(\hat{c}_1, \ldots, \hat{c}_m) = \text{argmin}_{c_1,\ldots,c_m} \text{SSR}(c_1, \ldots, c_m)$, where the minimization is taken over all partitions $(c_1, \ldots, c_m)$ for which $c_j - c_{j-1} \geq h$. The parameter h corresponds to the minimum period that a regime should last, and is to be chosen up front. Bai and Perron (2003) describe an efficient algorithm to determine the optimal break points for a given m.

If we set h too low it is possible that spurious effects are picked up, which is undesirable. On the other hand, if we set h too high, it is possible that we miss break points. We take $h = 5$ which is in line with Zeileis et al. (2003) and Harvey et al. (2009), who suggest to set h equal to 10% of the sample size.

Using the method described above, we can determine the optimal break points $(c_1, \ldots, c_m)$ for an *a priori* given number of break points m. We then have to determine what the optimal number of break points, say m^*, is. In general there are two ways to choose the optimal number of break points: (i) using an information criterion like the BIC, and (ii) performing F-tests to test the significance of the improvement in fit when adding one or multiple break points.

If the information criterion is used, one determines the BIC for all $m \in \{0, \ldots, 5\}$, see Zeileis et al. (2003).[1] Denote BIC(m) as the BIC corresponding to the optimal break points for a given m. The optimal number of break points is then defined by $m^* = \arg\max \text{BIC}(m)$.

As in Bai and Perron (1998, 2003), we could also consider two F-tests. The first is the test of $m = l$ versus $m = l+1$ break points. This is a sequential procedure: one starts with the null hypothesis of $m = 0$ versus the alternative hypothesis of $m = 1$ break points. If the null hypothesis of no break points is rejected, one continues testing for the significance of two break points versus the null hypothesis of one break point, and so on. The F-statistic is a function of the restricted sum of squared residuals (RSSR) and the unrestricted sum of squared residuals (USSR):

$$F = \frac{(\text{RSSR} - \text{USSR})/(p_1 - p_0)}{\text{USSR}/(n - p_1)}, \tag{3.3}$$

where p_0 is the number of parameters in the model under the null hypothesis, p_1 the number of parameters in the model under the alternative hypothesis, and n is the number of observations. Since the dates of the structural changes are unknown, we cannot use the standard critical values of the F-distribution as used in Sweeting (2011), but critical values have to be obtained through simulation (see Andrews (1992)). If a break point is significant, this break point is fixed and one searches for a new break point. The old break point is not allowed to move, which may be suboptimal when searching for more than one break point. Therefore, we shall not use the sequential F-test.

The second F-test is based on the null hypothesis of no break point ($m = 0$) versus the alternative hypothesis of $m = k$ break points. To determine the optimal number of break points, we determine the F-statistic as defined in (3.3) for all $k \in \{1, \ldots, 5\}$ which we denote by $F(k)$. We define the UDmax test statistic as the maximum value of those F-statistics:

$$\text{UDmax} = \max_k F(k) \tag{3.4}$$

Since the number and dates of the break points are unknown, critical values have to be obtained through simulation. If the observed UDmax test statistic is larger

[1] We consider at most five structural changes. In exploratory analysis we allowed for more than five structural changes, but the optimal number of structural changes never exceeded three.

than the critical value, the number of break points is equal to $k^* = \arg\max F(k)$. If the test statistic is smaller than the critical value, there is insufficient proof for a structural change.

The latter F-test is close to using the BIC, because an optimal model is chosen while considering all model specifications. Yao (1988) shows that the number of break points that follows from optimizing the BIC is a consistent estimator of the true number of break points, and Bai and Perron (2003) note that the BIC performs well in the absence of serial correlation. We will therefore use the BIC to choose the number of break points. In the following paragraph we apply the method to Dutch male mortality data.

Illustration - the Lee-Carter model. We consider the period effect of the Lee-Carter model, estimated on Dutch male mortality data for the period 1960 to 2008, for the ages 60 to 89. We also show results of the optimal calibration period strategy of Denuit and Goderniaux (2005). The top left graph in Figure 3.1 shows the parameter estimates for $\kappa_t^{(2)}$. A random walk with constant drift does not seem appropriate, because of apparent structural changes around 1972 and 2000. This is confirmed in the bottom left graph. The black lines correspond to projections from a random walk with constant drift and these projections do not seem consistent with the observations. The blue lines correspond to projections when one structural change is allowed; the break point is dated at 1993. These projections are not unreasonable, but the drift of the period effect does not appear to be piecewise constant before and after the break point. We obtain the projections represented by the red lines if we allow for multiple structural changes. The break points are estimated to occur at 1972 and 2002. The projections look reasonable, because the drift of the period effect is piecewise constant between the different break points.

The graphs on the right-hand side of Figure 3.1 show the projections for the period effects from the Lee-Carter model calibrated on different periods. We compare scenarios without structural changes, with a single structural change and with multiple structural changes. Allowing for a single structural change leads to more robust projections with respect to the calibration period, and if we allow for multiple structural changes, projections become even more robust.

Figure 3.2 shows the first order differences of the estimated period effect from Figure 3.1 (top left). In the upper right graph we observe that the first break point is accurately estimated, since the confidence interval (shown by the red line) is narrow.[2] The lower left graph in Figure 3.2 shows the confidence intervals for the case of two break points. The second break point (around the year 2002) is estimated accurately, but the confidence interval corresponding to the first break point is wide. This can

[2]See Bai and Perron (1998) for a description how these confidence intervals are derived. Our algorithm for detecting structural changes makes use of the `R`-package `strucchange` (Zeileis et al. (2002)).

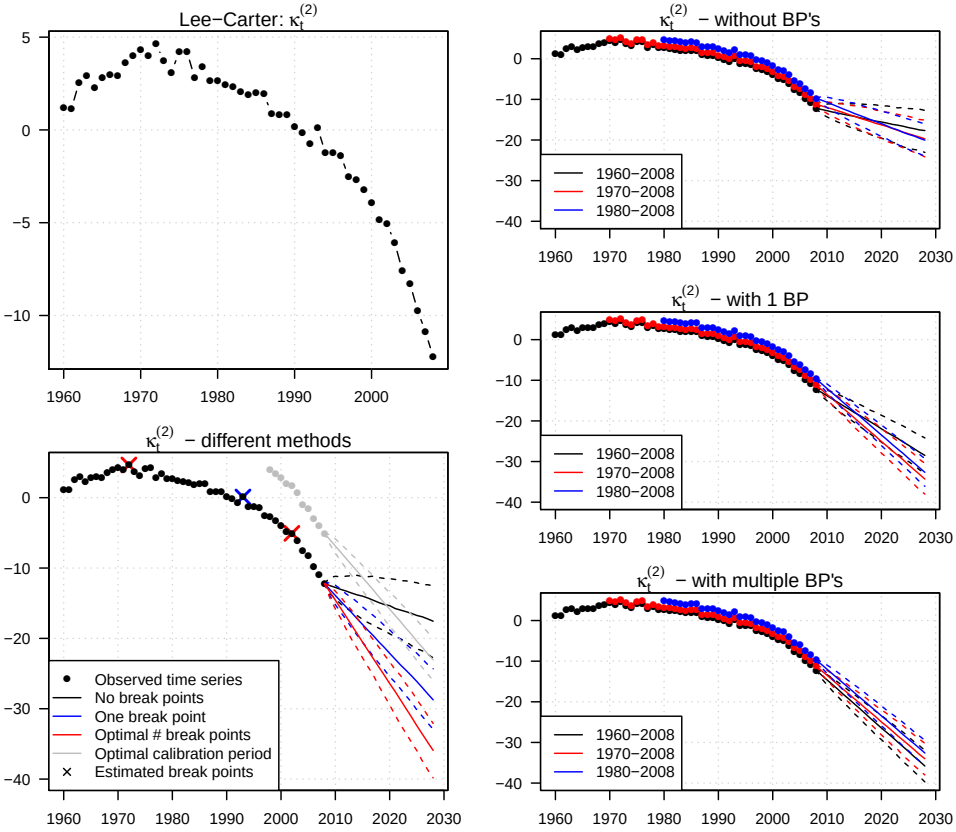


Figure 3.1: Top left: parameter estimates of $\kappa_t^{(2)}$ in the Lee-Carter model, calibrated on data from Dutch males aged 60-89 in the period 1960-2008. Bottom left: projections for the period effect using different projection methods. Top right through bottom right: projections of the period effect for different calibration periods without allowing for structural changes, allowing for one structural change and allowing for multiple structural changes. Dots are estimated parameters, solid lines are the 50th percentiles and dashed lines are the 5th and 95th percentiles of the projections.

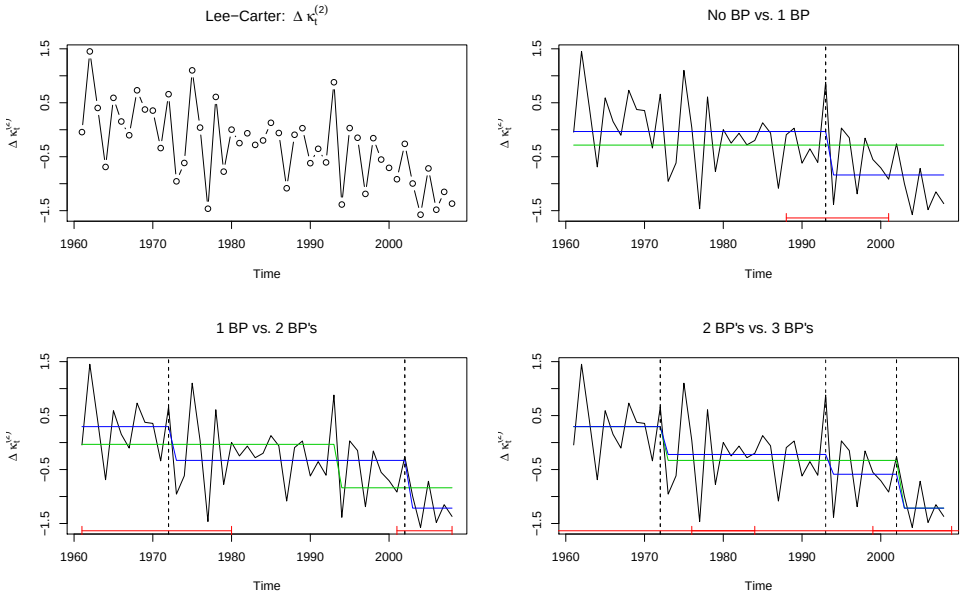


Figure 3.2: Confidence intervals for estimated break points for $\kappa_t^{(2)}$ in the Lee-Carter model, calibrated on Dutch males aged 60-89 in the years 1960-2008. In the plots (i) BP's vs. $(i+1)$ BP's the green lines represent the mean of $\Delta\kappa_t^{(2)}$ for the different periods when (i) BP's are allowed, and the blue lines represent the mean of $\Delta\kappa_t^{(2)}$ when $(i+1)$ BP's are allowed. The red lines represent the confidence intervals corresponding to the break points.

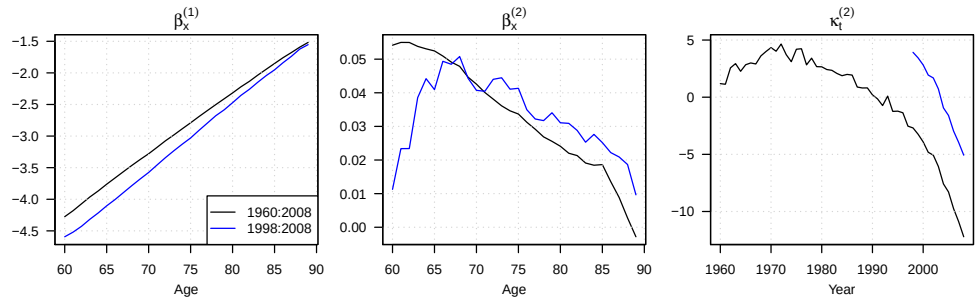


Figure 3.3: Parameter estimates of the Lee-Carter model, calibrated on data from Dutch males aged 60-89 using the large calibration period 1960-2008, and the optimal calibration period 1998-2008 according to the method of Denuit and Goderniaux (2005).

be explained by the values before and after the year 1972. However, allowing for the second break point leads to an improvement in fit over the whole observation period. This is illustrated by the differences between the green and blue lines in Figure 3.2. The bottom right graph shows the confidence intervals for the case of three break points. The confidence intervals overlap and they are much larger than for the case of two break points.

Figure 3.1 (bottom left) also shows the estimated period effect if the calibration period is chosen according to the procedure proposed in Denuit and Goderniaux (2005). We calibrate the Lee-Carter model to the entire calibration period, and then estimate OLS on different subsets of the period effect while keeping the end date fixed. The optimal calibration period is chosen where the adjusted R^2 is maximal. Using that calibration period we recalibrate the Lee-Carter model, and the result is plotted here in gray.[3] In line with Denuit and Goderniaux (2005) we enforce that the calibration period must be larger than ten years, and in this example the optimized calibration period turns out to be of minimal length, in contrast to the findings of Denuit and Goderniaux (2005) for Belgian data.

In this recent calibration period the period effect shows little variability which is translated into narrower confidence intervals than when we would have required the model to explain the entire dataset. Figure 3.3 shows the parameter estimates of the Lee-Carter model based on the entire and the optimal calibration period. Given the parameter restrictions, $\beta_x^{(1)}$ is the mean mortality rate, which explains the downward shift. The estimates for $\beta_x^{(2)}$ differ substantially and those for the optimal (and shorter) calibration period are less smooth.

3.2.2 Forecasting cohort effects

Section 2.3.2 contains an overview of different approaches to project the cohort effect. Imposing an ARIMA-specification up front can lead to biologically unreasonable forecasts. Therefore, we use the BIC to select the optimal specification, but we only consider ARIMA(p, d, q)-specifications for $d \in \{0, 1\}$ and $(p, q) \in \{0, 1, 2\}$. We do not consider the case $d = 2$, because from Cairns et al. (2011a) we conclude that using a second order differencing model leads to implausibly large confidence intervals.

3.3 Results

In this section we calibrate the mortality models from Table 2.1 to Dutch and Belgian mortality data, and we use the parameter constraints listed in Table 2.2 to uniquely identify the models. As in Brouhns et al. (2002), we assume a Poisson distribution for the number of deaths within a year, i.e. $D_{t,x} \sim \text{Poisson}(E_{t,x}\mu_{t,x})$.

[3] The estimated period effect on the optimal calibration period is shifted upwards due to the parameter restrictions.

The models M5 to M8 assume linearity for the age effects for pensioner ages. That linearity does not hold for lower and higher ages, and as a result these models are appropriate for pensioner ages only (60-89). Therefore, we calibrate the models M5-M8 only on the ages 60-89, whereas the other models are calibrated both for the ages 20-89 and the ages 60-89.

After having calibrated the models, we perform out-of-sample backtests to investigate the predictive qualities of the models while allowing for no, a single or multiple structural change(s).

3.3.1 Model fit

We calibrate the models on male and female mortality data from the Netherlands and Belgium for the years 1950 to 2008.[4] Earlier data is excluded so there are no world wars in the data set. We consider the ages 20-89, because mortality rates for younger ages are not relevant for insurers and pension funds, and mortality rates for ages above 89 are less reliable and are therefore excluded. If mortality rates are needed for higher ages, techniques are available to close mortality tables; see e.g. Vaupel (1990), Lindbergson (2001) and Denuit and Goderniaux (2005).

We present the estimation results for Dutch and Belgian males for ages 20-89 in Table 3.1 and for ages 60-89 in Table 3.2. These tables show the effective number of parameters that is estimated in each of the models, and the corresponding AIC and BIC that we define as AIC $= \ln \mathcal{L} - k$ and BIC $= \ln \mathcal{L} - \frac{1}{2} k \cdot \ln n$, where $\ln \mathcal{L}$ is the log likelihood, n is the number of observations, and k is the effective number of parameters.[5] A higher AIC or BIC means that the model is better able to explain the data. The difference between the AIC and the BIC is that the BIC imposes a higher penalty for the number of parameters used. Mortality models contain many parameters and we therefore believe the BIC to be a more appropriate information criterion. However, the ranking based on a fit on historical data does not predict whether a model will produce good mortality projections.

For the age range 20-89, the models with a cohort effect and interaction between age and period effects have the highest AIC and BIC. As expected, models that score well on AIC but which have many parameters, score worse on BIC; M11 is the clearest example of this. The ranking of the models for Dutch males is similar to the ranking for Belgian males. However, some models that score well on the age range 20-89 score worse for the age range 60-89 (M9, M11, M12 and M13) and vice versa (M2 and M3). The ranking of the models for the age range 60-89 is again similar for the Dutch and Belgian males.

[4] Mortality data is downloaded from the Human Mortality Database, which is a joined project of the University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Data are available at http://www.mortality.org.

[5] The effective number of parameters is the total number of parameters that is included in the

Table 3.1: Results of estimated models on Dutch and Belgian male mortality data for the ages 20-89 and the years 1950-2008. The numbers in brackets represent the ranking of the models for a specific dataset. The top three models are shown in blue boldface.

		The Netherlands		Belgium	
Model	Pars	AIC	BIC	AIC	BIC
M1	197	-22,000 (10)	-22,623 (10)	-22,332 (10)	-22,955 (10)
M1A	324	-19,535 (8)	-20,559 (8)	-20,122 (8)	-21,146 (8)
M2	385	-18,425 (5)	-19,642 (5)	-19,129 (6)	-20,345 (6)
M2A	513	-18,438 (6)	-20,060 (7)	-18,994 (5)	-20,616 (7)
M3	246	-18,947 (7)	-19,724 (6)	-19,538 (7)	-20,315 (5)
M9	327	-18,359 (4)	-19,392 **(2)**	-18,885 (4)	-19,919 **(2)**
M10	244	-19,905 (9)	-20,676 (9)	-21,419 (9)	-22,190 (9)
M11	422	-18,258 **(1)**	-19,591 (4)	-18,810 **(1)**	-20,144 (4)
M12	364	-18,289 **(2)**	-19,439 **(3)**	-18,840 **(2)**	-19,990 **(3)**
M13	327	-18,358 **(3)**	-19,392 **(1)**	-18,873 **(3)**	-19,907 **(1)**

Table 3.2: Results of estimated models on Dutch and Belgian male mortality data for the ages 20-89 and the years 1950-2008. Notes as in Table 3.1.

		The Netherlands		Belgium	
Model	Pars	AIC	BIC	AIC	BIC
M1	117	-11,035 (14)	-11,355 (14)	-10,421 (14)	-10,741 (14)
M1A	204	-9,204 (12)	-9,762 (13)	-9,665 (12)	-10,223 (12)
M2	225	-8,797 (8)	-9,412 (4)	-8,991 (6)	-9,606 (4)
M2A	313	-8,820 (9)	-9,675 (11)	-8,995 (7)	-9,850 (9)
M3	166	-8,941 (11)	-9,395 **(3)**	-9,101 (10)	-9,555 **(2)**
M5	118	-9,345 (13)	-9,668 (10)	-9,912 (13)	-10,235 (13)
M6	196	-8,732 **(1)**	-9,268 **(1)**	-8,935 **(1)**	-9,471 **(1)**
M7	254	-8,735 **(2)**	-9,429 (5)	-8,938 **(2)**	-9,632 (5)
M8	198	-8,792 (7)	-9,333 **(2)**	-9,031 (9)	-9,572 **(3)**
M9	284	-8,752 (4)	-9,528 (8)	-8,942 (4)	-9,719 (7)
M10	204	-8,908 (10)	-9,465 (6)	-9,347 (11)	-9,905 (11)
M11	342	-8,771 (5)	-9,706 (12)	-8,965 (5)	-9,900 (10)
M12	284	-8,783 (6)	-9,560 (9)	-9,002 (8)	-9,778 (8)
M13	284	-8,748 **(3)**	-9,524 (7)	-8,939 **(3)**	-9,716 (6)

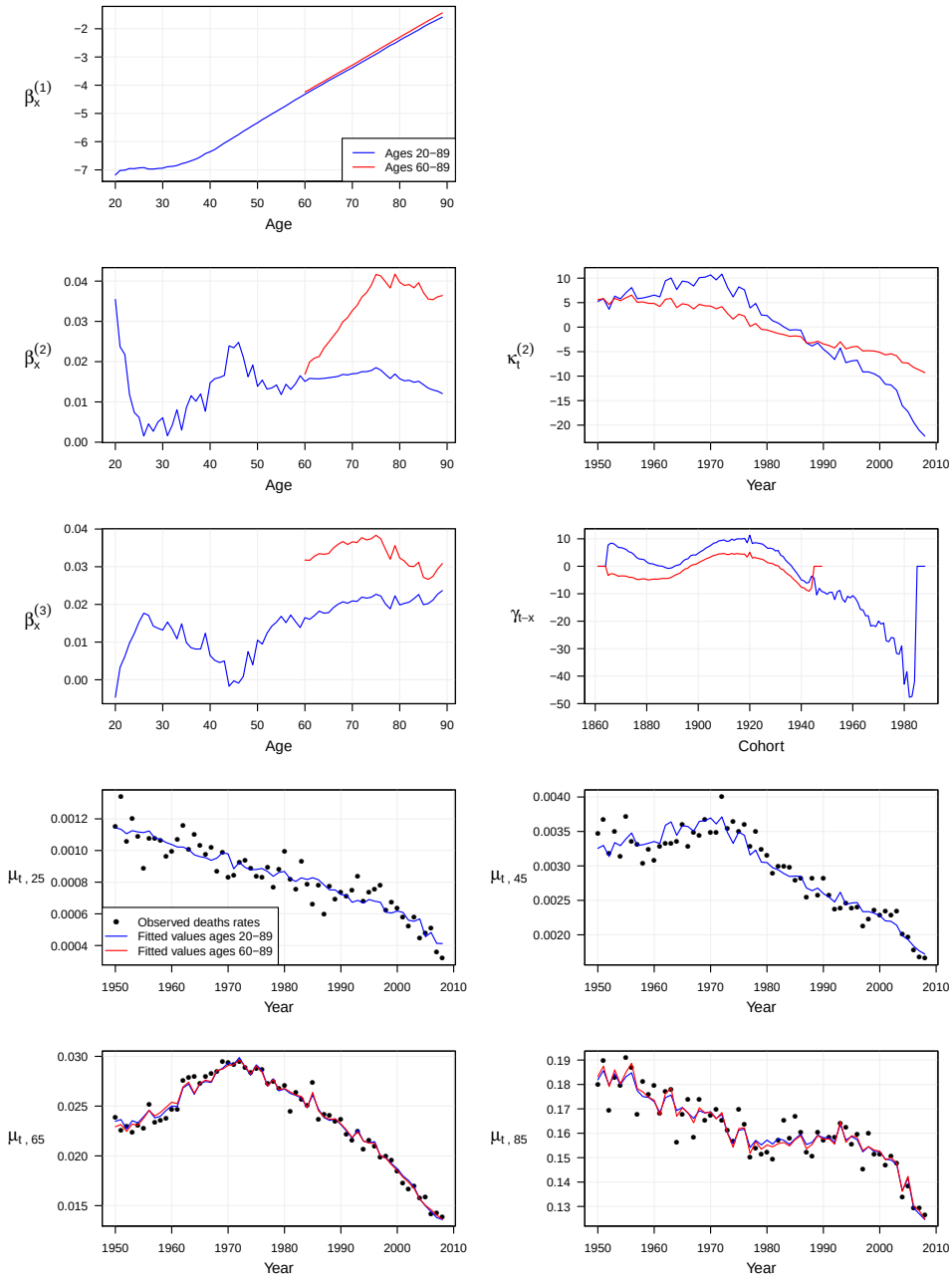


Figure 3.4: The first five panels show the parameter estimates for M2 calibrated on Dutch male mortality in the years 1950 to 2008 on the ages 20-89 and 60-89. The last four panels show realized mortality rates (dots) and fitted mortality rates for $x = \{25, 45, 65, 85\}$ (calibrated on ages 20-89 and ages 60-89)

Table 3.3: Optimal values for x_c in M8 and M12 when $x_c \in \{60, \ldots, 89\}$ or $x_c \in \{20, \ldots, 89\}$, based on the calibration period 1950-2008.

		The Netherlands		Belgium	
Model	Ages	Males	Females	Males	Females
M8	60-89	60	60	60	60
M12	60-89	60	89	89	89
M12	20-89	20	89	20	26

For the models M8 and M12 the impact of the cohort effect on the mortality rates for age x depends on the parameter x_c. The cohort effect γ_{t-x} is multiplied with $(x_c - x)$, so it has a larger impact on mortality rates for ages farther away from x_c. From Table 3.3 we conclude that, for the datasets considered, 4 out of 12 times the cohort effect mainly affects younger ages ($x_c = 89$), and 8 out of 12 times the cohort effect mainly affects the elderly.

We present the parameter estimates for M2 estimated on Dutch male mortality data in Figure 3.4 since this model fits the data well for both age ranges. The parameter estimates for the two age ranges are similar and the fitted mortality rates differ only marginally. In order to forecast mortality, the parameter $\kappa_t^{(2)}$ needs to be projected into the future, and for new cohorts we also have to project the cohort effect γ_{t-x}. Since the time-dependent parameters are different, it is possible and even likely that mortality projections resulting from the two different age ranges are different, regardless of the similar in-sample fit.

3.3.2 Out-of-sample performance

We now evaluate the predictive power of the models under consideration. We calibrate the models using data from 1950 to 2000 and then simulate forces of mortality for the years 2001 to 2008. This leads to a predictive distribution for the forces of mortality $\mu_{t,x}$ with $x = x_1, \ldots, x_X$ and $t = t_{T+1}, \ldots, t_{T+S}$; there are thus mortality rates for X ages and S years in the future. As in Riebler et al. (2012), we obtain the mean $\mathbb{E}(\mu_{t,x})$ and variance $\mathrm{Var}(\mu_{t,x})$ of future forces of mortality from the simulated predictive distribution. With $D_{t,x} \sim \mathrm{Poisson}(E_{t,x}\mu_{t,x})$ and using the law of total expectation it follows that for $t > T$ the expected death counts are

$$\widehat{d}_{t,x} = \mathbb{E}(D_{t,x}) = E_{t,x}\mathbb{E}(\mu_{t,x}) \tag{3.5}$$

model minus the number of parameter constraints that are used to identify the model.

and the variance of the death counts is

$$\begin{aligned}\sigma_{t,x}^2 = \text{Var}(D_{t,x}) &= \mathbb{E}(\text{Var}(D_{t,x}|\mu_{t,x})) + \text{Var}(\mathbb{E}(D_{t,x}|\mu_{t,x})) \\ &= \mathbb{E}(E_{t,x}\mu_{t,x}) + \text{Var}(E_{t,x}\mu_{t,x}) \\ &= E_{t,x}\mathbb{E}(\mu_{t,x}) + E_{t,x}^2\text{Var}(\mu_{t,x}),\end{aligned} \tag{3.6}$$

since we assume the exposure $E_{t,x}$ is given.[6] In the evaluation of the out-of-sample performance we consider the differences between observations and projections (hereafter: calibration of the projections) and the width of the confidence intervals of the projections (hereafter: sharpness of the projections), see also Gneiting and Raftery (2007). We compare the calibration of the mortality models using the root mean squared error (RMSE), both with and without the possibility of structural changes:

$$\text{RMSE} = \sqrt{\frac{1}{X \cdot S}\sum_{t,x}\left(d_{t,x} - \widehat{d}_{t,x}\right)^2}. \tag{3.7}$$

The RMSE only accounts for differences between observations and predictions, but not for differences in scale. A typical problem for mortality data is summarizing the quality of the forecasts for different ages and years in a single statistic. The death counts under consideration differ in scale for different ages and years due to different forces of mortality and exposures. The Dawid-Sebastiani scoring rule (DSS) introduced by Gneiting and Raftery (2007) is a statistic that evaluates the calibration and the sharpness of the projections, and also takes the scale of the observations into account. We compute the average DSS ($\overline{\text{DSS}}$) as introduced by Riebler et al. (2012), which allows us to summarize the quality of all forecast death counts into a single statistic:

$$\overline{\text{DSS}} = \frac{1}{X \cdot S}\sum_{t,x}\left[\left(\frac{d_{t,x} - \widehat{d}_{t,x}}{\sigma_{t,x}}\right)^2 + \log\sigma_{t,x}^2\right]. \tag{3.8}$$

Table 3.4 and 3.6 show the backtesting results for Dutch and Belgian females for the ages 20-89 and 60-89 respectively[7], and Table 3.5 and 3.7 show similar results for Dutch and Belgian males. For some models the statistics are lower when structural changes are incorporated (the bold figures in the tables), which means that allowing for structural changes has improved the quality of the mortality forecasts; especially the decrease in RMSE can be large. For other models however, the statistics are higher (the red figures), which means that the quality of the forecasts has worsened. Allowing for structural changes has little effect on the ranking of the models based on RMSE or $\overline{\text{DSS}}$, but the ranking of the models based on the backtest is markedly different from the ranking based on the fit on historical data in Table 3.1 and 3.2:

[6] We shall not simulate the population size, because then assumptions must be made on immigration and emigration.

[7] In Table 3.4, the results for M2 applied to Belgian females are implausible due to unrealistic cohort projections and are therefore not included in the table.

- For the age range 20-89 the models M2A and M11 perform well in the backtest, whereas the models M11, M12 and M13 explain the historical data well;

- For the age range 60-89 the models M1A, M2 and M2A perform well in the backtest, whereas the models M6 and M7 explain the historical data well.

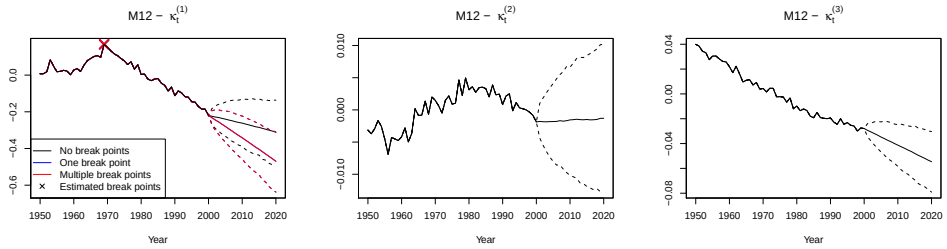
Figure 3.5a shows projections of the period effects for M12 applied to Dutch females aged 20-89 and Figure 3.5b shows resulting mortality projections. The non-monotone behavior observed in the red and gray projections is due to the estimated cohort effect. This effect is not visible for q_{80} because for Dutch females aged 20-89 we found $x_c = 89$, which implies that the cohort effect hardly affects the highest ages. From Figure 3.5a we observe that the projections of $\kappa_t^{(1)}$ are more convincing if we allow for structural changes, and in Figure 3.5b the mortality projections with structural changes are more convincing as well. This is confirmed in Table 3.4 as both the RMSE and the $\overline{\text{DSS}}$ have improved substantially.

Similar results are shown in Figure 3.6a and 3.6b for model M9 applied to Dutch females aged 20-89. The projections for $\kappa_t^{(2)}$ are more plausible when structural changes are allowed, but the projections for $\kappa_t^{(3)}$ are still implausible. The last fitted cohort effect is the cohort 1935[8], and later cohort effects are projected using an appropriate ARIMA-process. The cohort effect needed for projections for $x = 30$ are therefore projected over 35 years into the future[9], while for $x = 60$ the cohort effect is projected only few years into the future and for $x = 80$ it is available from the model calibration. This explains the relatively large confidence interval for q_{30} in Figure 3.6b. The projections for q_{80} including the structural change in $\kappa_t^{(2)}$ do not follow the realized mortality improvements, while the projections without structural changes do follow the realized mortality rates closely. Hence, even though the projected period effect is more plausible when structural changes are accounted for, the resulting mortality projections can be implausible for certain ages leading to worse backtesting results in Table 3.4.

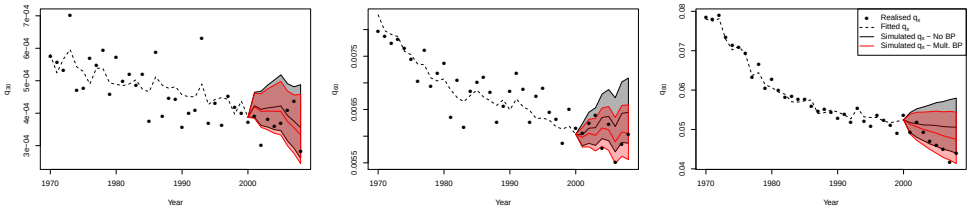
The most interesting example is M7 applied to Dutch females aged 60-89. In Table 3.6 we see that both the RMSE and $\overline{\text{DSS}}$ worsen if we allow for a single structural change, but the statistics improve if we allow for multiple structural changes. Figure 3.7 shows the projections for the period effects while allowing for no, one or multiple structural changes. The projections for $\kappa_t^{(1)}$ with a single structural change are less convincing than when no structural changes are allowed, because the last structural change has not been identified. When we allow for multiple structural changes we

[8]For M9 and M13 the cohort effect is set equal to zero if there are no observations related to the age 60 or higher. For the age range 20-89 and the calibration period 1950-2000 this means that the last estimated cohort is $2000 - 65 = 1935$.

[9]The cohort effect needed in 2001 for $x = 30$ is for the cohort 1971. The last estimated cohort effect is for the cohort 1935. Hence, the cohort effect for the cohort 1971 is projected 36 years from the last estimated cohort effect.

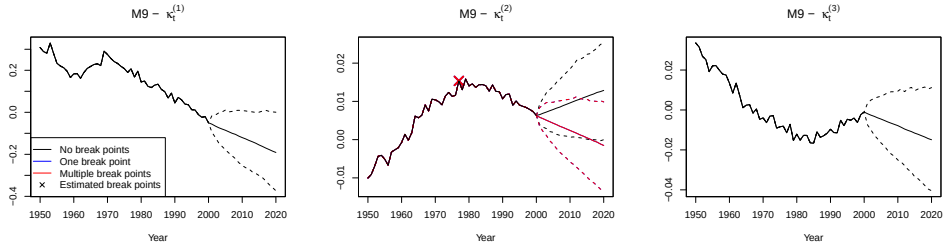


(a) Projections for the period effects $\kappa_t^{(i)}$ taking into account no, a single or multiple structural changes.

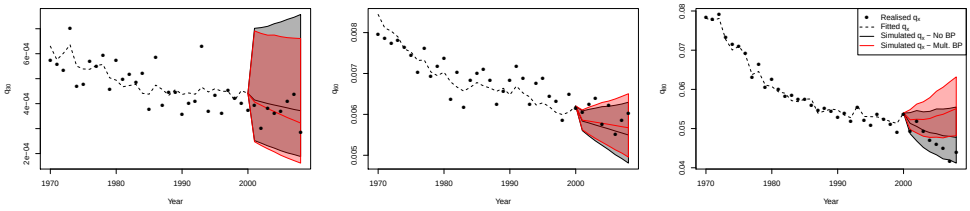


(b) Mortality projections for $x = \{30, 60, 80\}$. The black and red lines represent projections without and with multiple structural changes, respectively, at the 5th, 50th and 95th percentile.

Figure 3.5: Illustration of modeling approach for the period effects of M12 applied to Dutch females aged 20-89 in the period 1950-2000.



(a) Projections for the period effects $\kappa_t^{(i)}$ taking into account no, a single or multiple structural changes.



(b) Mortality projections for $x = \{30, 60, 80\}$. Notes: see Figure 3.5b.

Figure 3.6: Illustration of modeling approach for the period effects of M9 applied to Dutch females aged 20-89 in the period 1950-2000.

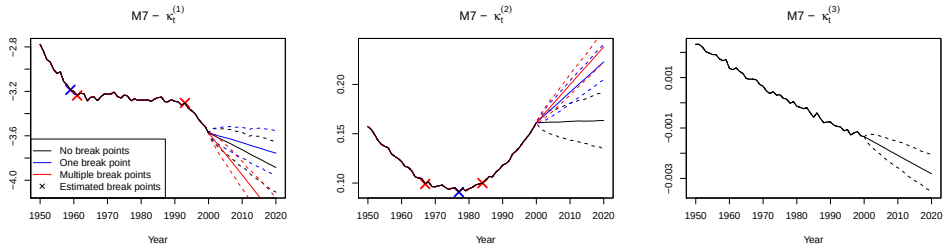


Figure 3.7: Projections for the period effects $\kappa_t^{(i)}$ of M7 applied to Dutch females aged 60-89 in the period 1950-2000.

detect two structural changes, and the projections for the period effects are more plausible. The projections for $\kappa_t^{(2)}$ are also most convincing if we allow for multiple structural changes. This example illustrates the added value when allowing multiple structural changes.

3.4 Conclusion

In this chapter we calibrate a selection of stochastic mortality models on historical mortality data from the Netherlands and Belgium. To create mortality projections, we project the period and the cohort effects. The cohort effects are projected using an ARIMA(p, d, q)-specification, where (p, d, q) are chosen such that the BIC is optimal. The period effect is projected using a modeling strategy that allows for objective detection of multiple structural changes in the difference stationary process. We observe that projections of the period effects are more robust with respect to the calibration period if we allow for multiple structural changes.

We compare the impact of allowing a single or multiple structural changes. We often find evidence for one structural change, and sometimes even multiple structural changes are estimated. We also find that allowing for structural changes can lead to improved backtesting results. But this is not always the case, because apparent structural changes may not be identified until sufficient evidence for their existence has accumulated, i.e. the improvement in fit from including a structural change may not be sufficient yet to to overcome the penalty in the BIC caused by the extra parameter. Another explanation for why backtesting results may not have improved, is that changes in age effects have not been accounted for.

The model we propose relaxes the assumption that all parameter values remain constant over the considered time period. We check for different mortality trends in the period effects and use the latest trend to project mortality. In that sense it resembles methods in which the calibration period is restricted to a particular subset

of recent data points which is chosen to provide the best model fit. Such alternative methods also allow that other, age-dependent, parameters are only fitted for this restricted period and this may improve fit for the most recent observations.

Our approach has the advantage that it can still be used when one requires that a model structure describes the entire collection of data points. This would for example be the case if we want to compare the performance of different model structures for a given dataset. If such structures involve more than one stochastic factor, we do not have to exclude the possibility that one of the multiple time series undergoes a structural change while the others remain the same as before and we do not need to adjust the overall calibration period as a result of such a change.

Each approach therefore has its advantages and disadvantages, but it is reassuring that our numerical example for a single factor model suggests that estimates generated by the two methods will differ in their fit over the most recent years, but not too substantially.

Table 3.4: Results for Dutch and Belgian female mortality rates for the ages 20-89 calibrated on the years 1950-2000, backtested on the years 2001-2008.

		RMSE			$\overline{\text{DSS}}$		
Country	Model	0	1	> 1	0	1	> 1
NL	M1	67.7	67.7	67.7	7.79	7.79	7.79
	M1A	75.9	75.9	75.9	7.76	7.76	7.76
	M2	71.7	71.7	71.7	8.73	8.73	8.73
	M2A	79.7	**82.1**	**82.1**	7.62	**7.69**	**7.69**
	M3	118.1	118.1	118.1	8.69	8.69	8.69
	M9	81.9	**166.9**	**166.9**	8.50	**9.34**	**9.34**
	M10	92.5	92.5	92.5	8.71	8.71	8.71
	M11	64.6	64.6	64.6	8.19	8.19	8.19
	M12	121.9	**76.2**	**76.2**	9.06	**8.21**	**8.21**
	M13	91.7	91.7	91.7	8.54	8.54	8.54
BE	M1	56.1	56.1	56.1	7.36	7.36	7.36
	M1A	57.7	57.7	57.7	7.46	7.46	7.46
	M2♣	-	-	-	-	-	-
	M2A	39.0	**36.8**	**36.8**	7.24	**7.31**	**7.31**
	M3	82.2	82.2	82.2	7.72	7.72	7.72
	M9	61.4	**81.0**	**81.0**	8.31	**8.62**	**8.62**
	M10	86.9	86.9	86.9	9.62	9.62	9.62
	M11	38.0	38.0	38.0	7.05	7.05	7.05
	M12	72.9	72.9	72.9	7.32	7.32	7.32
	M13	60.8	**61.2**	**61.2**	8.04	**8.16**	**8.16**

Note: Mortality forecasts are backtested for the years 2001-2008 using different forecasting methods for the period effects. "0", "1" or "> 1" means we allow for no, a single or multiple structural changes, respectively. Bold numbers indicate improved backtesting results with respect to no structural changes; red numbers indicate worsened results with respect to no structural changes.

♣*:* Projections of the cohort effect for M2 for Belgian females lead to unreliable predictions. For this model, the results are therefore not included in this table.

Table 3.5: Results for Dutch and Belgian male mortality rates for the ages 20-89 calibrated on the years 1950-2000, backtested on the years 2001-2008.

		RMSE			$\overline{\text{DSS}}$		
Country	Model	0	1	> 1	0	1	> 1
NL	M1	266.4	**243.4**	**243.4**	22.01	**21.43**	**21.43**
	M1A	222.9	222.9	222.9	14.35	14.35	14.35
	M2	105.2	105.2	105.2	9.41	9.41	9.41
	M2A	164.9	164.9	164.9	10.76	10.76	10.76
	M3	145.4	145.4	145.4	10.06	10.06	10.06
	M9	176.7	**120.3**	**120.3**	9.71	**8.99**	**8.99**
	M10	193.7	**159.4**	**159.4**	10.55	**9.87**	**9.87**
	M11	187.2	187.2	187.2	10.25	10.25	10.25
	M12	178.1	**118.2**	**118.2**	12.69	**11.31**	**11.31**
	M13	164.4	**111.8**	**111.8**	9.59	**8.96**	**8.96**
BE	M1	147.2	147.2	147.2	10.56	10.56	10.56
	M1A	124.1	**113.3**	**113.3**	9.52	**9.42**	**9.42**
	M2	84.7	84.7	84.7	8.79	8.79	8.79
	M2A	87.3	**61.7**	**61.7**	8.46	**8.25**	**8.25**
	M3	71.8	71.8	71.8	8.77	8.77	8.77
	M9	79.7	79.7	79.7	8.61	8.61	8.61
	M10	117.1	**83.4**	**83.4**	9.32	**9.45**	**9.45**
	M11	93.3	93.3	93.3	8.38	8.38	8.38
	M12	45.3	45.3	45.3	8.58	8.58	8.58
	M13	72.3	72.3	72.3	8.67	8.67	8.67

Note: See Table 3.4

Table 3.6: Results for Dutch and Belgian female mortality rates for the ages 60-89 calibrated on the years 1950-2000, backtested on the years 2001-2008.

		RMSE			$\overline{\text{DSS}}$		
Country	Model	0	1	> 1	0	1	> 1
NL	M1	128.2	128.2	128.2	10.38	10.38	10.38
	M1A	112.8	112.8	112.8	9.71	9.71	9.71
	M2	102.5	102.5	102.5	10.03	10.03	10.03
	M2A	201.8	**124.7**	**124.7**	10.06	**9.83**	**9.83**
	M3	160.7	160.7	160.7	11.32	11.32	11.32
	M5	134.1	134.1	134.1	12.75	12.75	12.75
	M6	339.5	**412.0**	**412.0**	15.22	**15.36**	**15.36**
	M7	517.3	**719.3**	**421.9**	19.88	**23.39**	**16.77**
	M8	141.2	**88.6**	**88.6**	10.03	**10.38**	**10.38**
	M9	114.4	114.4	114.4	10.08	10.08	10.08
	M10	113.1	113.1	113.1	9.93	9.93	9.93
	M11	137.0	137.0	137.0	10.06	10.06	10.06
	M12	151.5	151.5	151.5	10.51	10.51	10.51
	M13	135.5	**218.6**	**218.6**	10.16	**11.66**	**11.66**
BE	M1	78.2	78.2	78.2	9.36	9.36	9.36
	M1A	87.8	87.8	87.8	9.58	9.58	9.58
	M2	61.8	61.8	61.8	9.49	9.49	9.49
	M2A	154.7	**84.4**	**84.4**	9.95	**9.68**	**9.68**
	M3	111.4	111.4	111.4	9.97	9.97	9.97
	M5	101.9	101.9	101.9	13.28	13.28	13.28
	M6	177.8	177.8	177.8	10.92	10.92	10.92
	M7	399.8	**500.7**	**470.5**	15.33	**15.54**	**14.17**
	M8	149.1	149.1	149.1	10.49	10.49	10.49
	M9	86.6	86.6	86.6	9.55	9.55	9.55
	M10	86.0	86.0	86.0	9.56	9.56	9.56
	M11	83.3	83.3	83.3	9.40	9.40	9.40
	M12	98.5	98.5	98.5	9.66	9.66	9.66
	M13	87.8	87.8	87.8	9.40	9.40	9.40

Note: See Table 3.4

Table 3.7: Results for Dutch and Belgian male mortality rates for the ages 60-89 calibrated on the years 1950-2000, backtested on the years 2001-2008.

		RMSE			$\overline{\text{DSS}}$		
Country	Model	0	1	> 1	0	1	> 1
NL	M1	296.9	296.9	296.9	16.30	16.30	16.30
	M1A	297.0	297.0	297.0	15.53	15.53	15.53
	M2	120.2	120.2	120.2	10.58	10.58	10.58
	M2A	166.9	166.9	166.9	10.97	10.97	10.97
	M3	200.2	200.2	200.2	11.63	11.63	11.63
	M5	286.5	286.5	286.5	13.86	13.86	13.86
	M6	232.3	232.3	232.3	13.59	13.59	13.59
	M7	202.4	202.4	202.4	12.53	12.53	12.53
	M8	386.6	**284.7**	**284.7**	15.47	**14.14**	**14.14**
	M9	207.9	207.9	207.9	12.00	12.00	12.00
	M10	283.4	283.4	283.4	13.61	13.61	13.61
	M11	283.2	283.2	283.2	13.42	13.42	13.42
	M12	343.7	**227.5**	**227.5**	14.09	**12.41**	**12.41**
	M13	233.1	233.1	233.1	12.56	12.56	12.56
BE	M1	160.6	160.6	160.6	10.99	10.99	10.99
	M1A	173.3	173.3	173.3	11.19	11.19	11.19
	M2	77.3	77.3	77.3	10.11	10.11	10.11
	M2A	112.1	**80.5**	**80.5**	10.17	**10.15**	**10.15**
	M3	91.7	91.7	91.7	9.81	9.81	9.81
	M5	166.6	166.6	166.6	10.68	10.68	10.68
	M6	163.1	163.1	163.1	10.63	10.63	10.63
	M7	132.3	132.3	132.3	10.23	10.23	10.23
	M8	209.4	209.4	209.4	11.35	11.35	11.35
	M9	148.3	148.3	148.3	10.41	10.41	10.41
	M10	161.5	161.5	161.5	10.62	10.62	10.62
	M11	174.5	174.5	174.5	10.99	10.99	10.99
	M12	154.6	154.6	154.6	10.51	10.51	10.51
	M13	198.8	198.8	198.8	11.65	11.65	11.65

Note: See Table 3.4

Chapter 4

A Bayesian joint model for population and portfolio-specific mortality

This chapter is based on F. van Berkum, K. Antonio, and M. Vellekoop. A Bayesian joint model for population and portfolio-specific mortality. *ASTIN Bulletin*, 47(3): 681 – 713, 2017a.

4.1 Introduction

Earlier in this thesis, we stressed that life insurance companies and pension funds need to value their liabilities using mortality rates appropriate for their portfolio. For many countries projections of mortality rates are available for the entire population, but substantial heterogeneity in mortality rates exists between individuals within a population. This is caused, amongst others, by differences in socioeconomic classes, see Villegas and Haberman (2014). Lantz et al. (1998) argue that individuals with a higher education tend to live more healthily, which may help to explain these differences in mortality.

Heterogeneity in mortality also occurs between individuals when they have different motivations to buy insurance. Finkelstein and Poterba (2002) show that substantial differences in mortality even exist between individuals with voluntary annuities, compulsory annuities or without annuities. Pitacco et al. (2009) discuss the presence of select mortality when individuals are subject to medical tests when starting a life insurance policy. Policyholders with a longer duration since the test will on average experience higher mortality than policyholders that have been accepted more recently. Therefore, an insurance company or pension fund cannot use mortality projections for the whole population without making any adjustments. The difference between mortality in a population and a portfolio is often called basis risk, see for example Barrieu et al. (2012).

In practice, portfolio-specific mortality rates are often constructed by multiplying projections of country-wide mortality rates with portfolio-specific factors. These portfolio-specific factors, also called experience factors, thus represent the relative

difference between the mortality rates of the population and the portfolio under consideration. In Solvency II, insurance companies are obliged to derive portfolio-specific mortality rate projections and analyze the uncertainty in these projections.

We propose a model to estimate population and portfolio-specific mortality simultaneously. To account for yearly fluctuations in small portfolios we use a Poisson distribution to model individual deaths for a given realization of hazard rates, as in Brouhns et al. (2002). We view the portfolio as part of the population and use a baseline mortality trend for the population. The larger dataset for the population allows us to generate reliable estimates for its dynamics of mortality. The relative difference between the population and the portfolio is modeled using a portfolio-specific and age-dependent random effect. Such random effects reflect the remaining heterogeneity among policyholders which is not captured by the observable risk factors. See Denuit et al. (2007) and Antonio and Zhang (2014) for similar examples in pricing models for non-life insurance, where policy(holder)-specific behavior is captured by such a random effect.

We use the Lee-Carter model for population mortality. In our Bayesian setting, we consider two prior distributions for the portfolio-specific factors. The first prior distribution assumes independent factors for different ages and independence between groups (the portfolio and the rest). The second prior distribution is an autoregressive smoothing prior which implies dependence between ages but independence between the factor for our own portfolio and the factor for the rest of the population.

We describe population mortality and portfolio-specific mortality simultaneously, in contrast to the multi-step method that is required in a frequentist approach. This helps to distinguish between volatility in the time series for the population, parameter uncertainty in the model for the population, and parameter uncertainty in the portfolio-specific factors.

To illustrate this point, Figure 4.1 shows observed portfolio-specific factors for the CMI dataset on assured lives in England & Wales.[1] These factors are the ratio of death rates in the CMI portfolio and death rates in the whole of England and Wales, for different years and ages. The observations are very volatile when considered as a function of age and they can fluctuate wildly over consecutive years. These fluctuations are mainly due to the randomness in the number of individual deaths for a given fixed mortality rate. In order to take this into account, we will explicitly model the noise in the outcomes that we can actually observe (the number of deaths), by specifying that these follow a Poisson distribution when conditioned on the unobserved hazard rates that contain the unknown portfolio-specific factors that we are ultimately interested in. In a case study based on this dataset for England and Wales, we will show that parameter uncertainty in portfolio-specific factors can be substantial but that its impact on mortality projections is relatively small compared to the impact of

[1] See Section 4.3 for a description of the CMI dataset on assured lives.

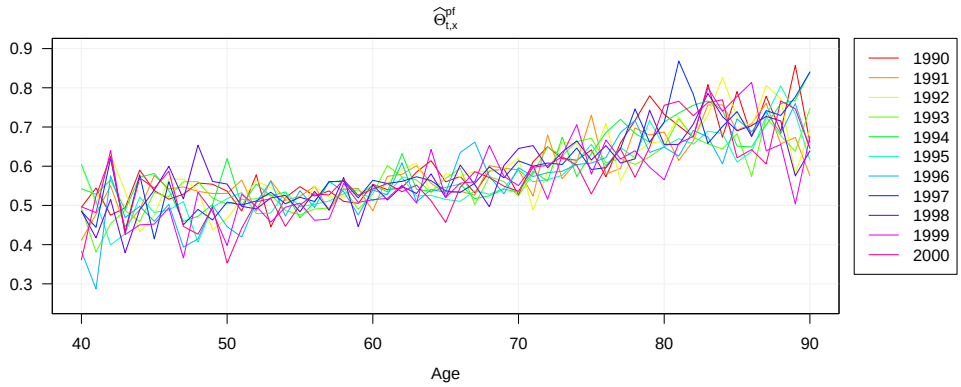


Figure 4.1: Observed portfolio-specific factors (the ratio of death rates in a portfolio and death rates in the whole country) for the CMI portfolio of assured male lives in England & Wales.

the Poisson noise in individual deaths.

In Section 4.2 we introduce our method and describe the prior distributions that are used in our Bayesian setting. Section 4.3 contains the illustration of our approach using the dataset on assured male lives from England & Wales, and Section 4.4 concludes.

4.2 Bayesian portfolio-specific mortality

In Section 2.5 we discussed different approaches to modeling portfolio-specific factors, which are suitable for different types of datasets. We consider the situation where only limited historical portfolio data is available, which hinders reliable estimation of mortality developments if only portfolio data would be used. We therefore simultaneously estimate mortality in the population and the portfolio-specific factors in a Bayesian setting.

Let the observed number of deaths for group i during calendar year t for ages in $[x, x+1)$ be $d^i_{t,x}$, and denote the exposure in this group for that period by $E^i_{t,x}$. The groups we consider are the entire population of a country ('pop'), the portfolio under investigation ('pf'), and the part of the population which is not included in the portfolio under consideration (hereafter referred to as the 'rest'), so $i \in \{\text{pop}, \text{pf}, \text{rest}\}$. The observed portfolio and the rest thus form the total population and we have that $d^{\text{pf}}_{t,x} + d^{\text{rest}}_{t,x} = d^{\text{pop}}_{t,x}$ and $E^{\text{pf}}_{t,x} + E^{\text{rest}}_{t,x} = E^{\text{pop}}_{t,x}$. We need to define the rest group explicitly, to ensure that we always consider all information available in the population.

To estimate parameters, we extend the portfolio dataset with observations of the total population. In the dataset of the portfolio we consider X ages and S years, and in the population X ages and T years. We define the set of cells (t, x) for which we

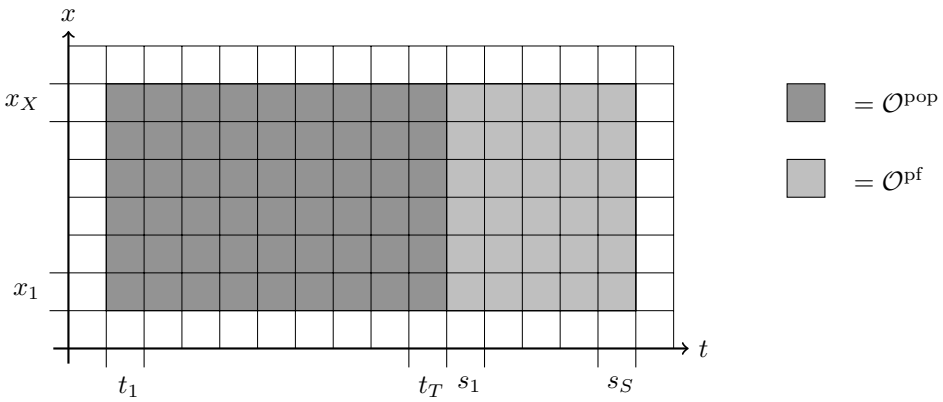


Figure 4.2: Illustration of $\mathcal{O}^{\text{pf}}$ and $\mathcal{O}^{\text{pop}}$.

have observations from both our portfolio and the rest (the light gray cells in Figure 4.2) as $\mathcal{O}^{\text{pf}} = \mathcal{S} \times \mathcal{X}$ with $\mathcal{S} = \{s_1, s_1 + 1, \ldots, s_S\}$ and $\mathcal{X} = \{x_1, x_1 + 1 \ldots, x_X\}$. We can only measure the differences between the portfolio and the rest on *this* set of observations. The set for which we have observations from the population but not from our portfolio (the dark gray cells in Figure 4.2) is defined by $\mathcal{O}^{\text{pop}} = \mathcal{T} \times \mathcal{X}$, with $\mathcal{T} = \{t_1, t_1 + 1 \ldots, t_T\}$ and $t_1 \leq t_T < t_T + 1 = s_1 \leq s_S$, and by construction $\mathcal{O}^{\text{pop}} \cap \mathcal{O}^{\text{pf}} = \emptyset$.[2] In total, $T^* = T + S$ years are included in the dataset.

We introduce indicator variables that will turn out to be useful when working with likelihoods:

$$I^{\text{pf}}_{t,x} = I^{\text{rest}}_{t,x} = \begin{cases} 1 & \text{if } (t,x) \in \mathcal{O}^{\text{pf}} \\ 0 & \text{otherwise,} \end{cases} \qquad I^{\text{pop}}_{t,x} = \begin{cases} 1 & \text{if } (t,x) \in \mathcal{O}^{\text{pop}} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

4.2.1 Model formulation and implementation

We assume that people in our own portfolio and the rest of the population share a baseline force of mortality which is denoted by $\mu_{t,x}$. Heterogeneity between groups is captured by a random effect Θ^i_x which depends on age. This leads to the following specification:

$$D^{\text{pop}}_{t,x} | \mu_{t,x} \sim \text{Poisson}(E^{\text{pop}}_{t,x} \mu_{t,x}), \qquad \text{for } (t,x) \in \mathcal{O}^{\text{pop}} \tag{4.2}$$

[2] We could use a wider age range for the population, but we choose to use the same set of ages in the population as we have available for the portfolio. This way, we ensure that the parameters α_x, β_x and κ_t are most appropriate for projection of mortality for the portfolio.

and

$$D^{\text{pf}}_{t,x}|(\mu_{t,x},\Theta^{\text{pf}}_x) \sim \text{Poisson}(E^{\text{pf}}_{t,x}\mu_{t,x}\Theta^{\text{pf}}_x), \qquad \text{for } (t,x)\in\mathcal{O}^{\text{pf}} \tag{4.3}$$

$$D^{\text{rest}}_{t,x}|(\mu_{t,x},\Theta^{\text{rest}}_x) \sim \text{Poisson}(E^{\text{rest}}_{t,x}\mu_{t,x}\Theta^{\text{rest}}_x), \tag{4.4}$$

with the Lee-Carter model for the baseline

$$\ln \mu_{t,x} = \alpha_x + \beta_x\kappa_t. \tag{4.5}$$

This implies that we consider all deaths in the population for every cell (t,x), by either using $D^{\text{pop}}_{t,x}$ or both $D^{\text{pf}}_{t,x}$ and $D^{\text{rest}}_{t,x}$.[3]

The random effects Θ^i_x are independent between groups i, but there may be dependence for different ages x.[4] In Section 4.2.2 we consider two prior specifications for Θ^i_x, a Gamma prior and a lognormal prior. In the first one, we assume independence between ages x and between groups i, but in the second one we assume dependence between ages and independence between groups. Given the baseline force of mortality $\mu_{t,x}$ and the portfolio-specific factors Θ^i_x, the Poisson distributed numbers of deaths are independent between ages, calendar years and groups.

To project mortality into the future, we need to impose a time series model on the period effect κ_t. Two time series specifications that are often used for projecting the period effect in the Lee-Carter model are a trend stationary and a difference stationary model (also known as a random walk with possibly a drift). As discussed in Chapter 3 we believe a difference stationary model to be more appropriate to model the period effect for a single country so that is what we will use in this chapter.

In order to generate samples of posterior distributions, we use the Markov chain Monte Carlo method (MCMC) with a burn-in period which allows the chain to move towards the desired distribution before we start taking samples, see Section 2.4 for an introduction on the MCMC method. Since the MCMC algorithm requires Metropolis(-Hastings) sampling for our model, the burn-in period is also used to calibrate scale parameters for the distribution to propose new samples. We calibrate the scale parameters in such a way that the acceptance probabilities are within the interval [20%, 30%]. Only samples that are found after the burn-in period are used for inference on parameters and for prediction purposes.

[3] We use the Lee-Carter model to specify the baseline mortality, but our model can easily be extended to include e.g. a cohort effect as in Renshaw and Haberman (2006). Also note that our model differs from the augmented common-factor model as in (2.31) since we do not include an extra dynamic factor when modeling subpopulations. The absence of such a term makes our model similar to the common factor model also discussed in Li and Lee (2005).

[4] We use time-independent portfolio factors, because estimating time dynamics on a few historical years can lead to spurious forecasting results. As a result, mortality improvements are perfectly correlated between groups, and this makes our model less appropriate for assessing basis risk in longevity hedging as in Haberman et al. (2014).

In a frequentist setting, parameter constraints are needed to uniquely identify the Lee-Carter model, since linear transformations can be applied which change the value of the parameters α_x, β_x and κ_t without changing the forces of mortality. In a Bayesian framework, parameters are random variables so there is no identifiability problem due to the introduction of priors. However, the presence of identification problems in a frequentist setting suggests that we may also encounter convergence problems for the MCMC algorithm in a Bayesian implementation of the same model. We therefore apply two parameter constraints:

$$\kappa_{t_1} = 0 \qquad \text{and} \qquad \|\boldsymbol{\beta}\|^2 = \sum_{x \in \mathcal{X}} \beta_x^2 = 1, \tag{4.6}$$

that are implemented through the specification of the prior distributions. Further, for $t \geq s_1$ we have more information: two observations per cell (t, x). As a result, in a frequentist setting the forces of mortality for the portfolio and the rest group are invariant under the following transformation:

$$\Theta_x^{\text{pf}} \to \Theta_x^{\text{pf}} \cdot \exp(c\beta_x), \quad \Theta_x^{\text{rest}} \to \Theta_x^{\text{rest}} \cdot \exp(c\beta_x) \quad \text{and} \quad \kappa_t \to \kappa_t - cI_{[t \in \mathcal{S}]}.$$

We observed that we may encounter convergence problems in κ_t and Θ_x^i if no additional constraint is imposed, and therefore we impose

$$\kappa_{s_1} = \kappa_{t_T}. \tag{4.7}$$

With this constraint we further ensure that the parameters α_x and β_x can be used for both $\mathcal{O}^{\text{pop}}$ and $\mathcal{O}^{\text{pf}}$. The parameter constraint in (4.7) is appropriately taken into account in the prior specification and in the derivation of the posterior distributions.

4.2.2 Prior distributions

We will now describe the prior distributions for parameters and hyperparameters, to complete the Bayesian specification of the model.

Population mortality parameters

Prior distribution for α_x. In line with Section 2.4 we use the following prior for α_x with $x = x_1, \ldots, x_X$:

$$e_x = \exp(\alpha_x) \overset{\text{iid}}{\sim} \text{Gamma}(a_x, b_x). \tag{4.8}$$

Prior distribution for β_x. In line with Section 2.4 we use the following prior for the vector of parameters $\boldsymbol{\beta} = (\beta_{x_1}, \ldots, \beta_{x_X})$:

$$\boldsymbol{\beta} \sim \text{vMF}(\boldsymbol{\mu}_\beta, c_\beta), \tag{4.9}$$

where $\text{vMF}(\boldsymbol{\mu}, c)$ is a Von Mises-Fisher distribution with constants $\boldsymbol{\mu}$ (the mean direction vector) and c (the concentration parameter) with $\|\boldsymbol{\mu}\| = 1$ and $c > 0$.

Prior specification for κ_t. In line with Section 2.4 and Chapter 3 we assume a random walk with drift for the period effect κ_t. The prior distribution is specified by

$$\delta \sim \mathrm{N}(\mu_\delta, \sigma_\delta^2) \tag{4.10}$$

$$\sigma_\varepsilon \sim \mathrm{Uniform}(0, A_\varepsilon) \tag{4.11}$$

$$\kappa_t = \kappa_{t-1} + \delta + \varepsilon_t, \quad \text{with} \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2) \quad \text{for} \quad t > t_1 \quad \text{and} \quad t \neq s_1 \quad \text{and} \quad \kappa_{t_1} = 0, \quad \kappa_{s_1} = \kappa_{t_T}. \tag{4.12}$$

Note that we have included the constraint defined in (4.7).

Portfolio-specific factors. The portfolio-specific factor Θ_x^i represents the ratio between the hazard rate for group i at age x and the hazard rate for the whole population at age x, where $i \in \{\mathrm{pf}, \mathrm{rest}\}$. We do not want to make a priori assumptions on whether mortality in a group is on average higher or lower than the baseline mortality, and therefore we impose $\mathbb{E}(\Theta_x^i) = 1$ $(\forall x, \forall i)$. We consider two prior distributions for Θ_x^i, an independent (Gamma) prior and a lognormal prior.

Gamma prior. The Gamma prior on the age-dependent factors for group i is given by

$$\Theta_x^i \sim \mathrm{Gamma}(c_x^i, c_x^i), \quad \text{for } x_1 \leq x \leq x_X. \tag{4.13}$$

The factors are independent over ages x and between groups i. By choosing equal values for the two parameters in the Gamma distribution we ensure that $\mathbb{E}(\Theta_x^i) = 1$ for all x and i, and the variance of the prior can be controlled by the choice of c_x^i.

Lognormal prior. In this specification we assume a mean reverting process (AR(1)) for the logarithm of the age-dependent factors. This ensures that the factors are non-negative. The lognormal prior on the age-dependent factors for group i is given by

$$\mathrm{logit}\,(\rho_i) \sim \mathrm{N}(\mu_{\rho_i}, \sigma_{\rho_i}^2) \tag{4.14}$$

$$\sigma_i \sim \mathrm{Uniform}(0, A_i) \tag{4.15}$$

$$\ln \Theta_x^i = \mu_i + \rho_i \ln \Theta_{x-1}^i + \eta_x^i, \tag{4.16}$$

$$\text{with } \eta_x^i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_i^2(1-\rho_i^2)) \quad \text{for } x_1 < x \leq x_X, \tag{4.17}$$

$$\text{and } \ln \Theta_{x_1}^i \overset{\text{iid}}{\sim} \mathrm{N}(-\tfrac{1}{2}\sigma_i^2, \sigma_i^2), \tag{4.18}$$

and all the $\ln \Theta_{x_1}^i$ and η_x^i are independent. Note that this implies that there may be dependence between group-specific mortality factors for different ages x, while factors for different groups i are independent. Due to the autoregressive structure in (4.17) this prior is also often referred to as an autoregressive smoothing prior.

The prior for ρ_i is chosen in such a way that it is restricted to the interval $(-1, 1)$, and the prior for σ_i is taken in line with the other variance prior specifications. The mean parameter equals $\mu_i = -\frac{1}{2}(1-\rho_i)\sigma_i^2$ and we again make sure that $\mathbb{E}(\Theta_x^i) = 1$ for all x and i, see Denuit et al. (2005).

In the next section we specify the constants that are needed to finalize the specification of the prior distributions for the parameters and hyperparameters. With the definition of the prior distributions our model is completely specified, and all posterior distributions can thus be calculated. They can be found in Appendix 4.A.

4.3 Empirical study

In this section we apply our model to data from the Continuous Mortality Investigation (CMI), which contains mortality statistics of assured male lives in England & Wales. We use the years $s \in \mathcal{S} = \{s_1 = 1990, \ldots, s_S = 2000\}$ and the ages $x \in \mathcal{X} = \{x_1 = 40, \ldots, x_X = 90\}$. Dowd et al. (2011) also use the CMI dataset when estimating a two-population mortality model, but they use the years 1961-2005 and the ages 60-84.

We extend the dataset with mortality data on the England & Wales population for the years $t \in \mathcal{T} = \{t_1 = 1950, \ldots, t_T = 1989\}$ and the same ages $x \in \mathcal{X}$ to ensure we obtain mortality forecasts consistent with population mortality forecasts.[5] We use population mortality data for $t \in \mathcal{S}$ to construct the rest group by subtracting portfolio deaths and exposures from the population deaths and exposures in those cells (t, x) for which portfolio data are available.

The size of the portfolio as a portion of the population, measured in observed deaths and observed exposures, is shown in Figure 4.3. In total there were around 28.5 million years of exposure and 159,029 observed deaths. If mortality in the portfolio were similar to that in the population we would expect the ratio of observed deaths and observed exposures to be of similar size. However, the ratios of observed deaths and observed exposures clearly differ, and we see that mortality in the portfolio is lower than in the population as whole.

We estimate four different models:[6]

1. The Lee-Carter model is used for population mortality for the England & Wales population for $t \in \{\mathcal{T}, \mathcal{S}\}$ and $x \in \mathcal{X}$, and parameters are estimated using maximum likelihood. This method is referred to as POP(f);

[5] Population mortality data are obtained from the Human Mortality Database. The Human Mortality Database is a joint project of the University of California, Berkeley (USA) and the Max Planck Institute for Demographic Research (Germany). Data are available at http://www.mortality.org.

[6] For POP(f) and POP(B) we only apply the parameter restrictions in (4.6), and for PF(B-G) and PF(B-logN) we apply the parameter restrictions in (4.6) and (4.7).

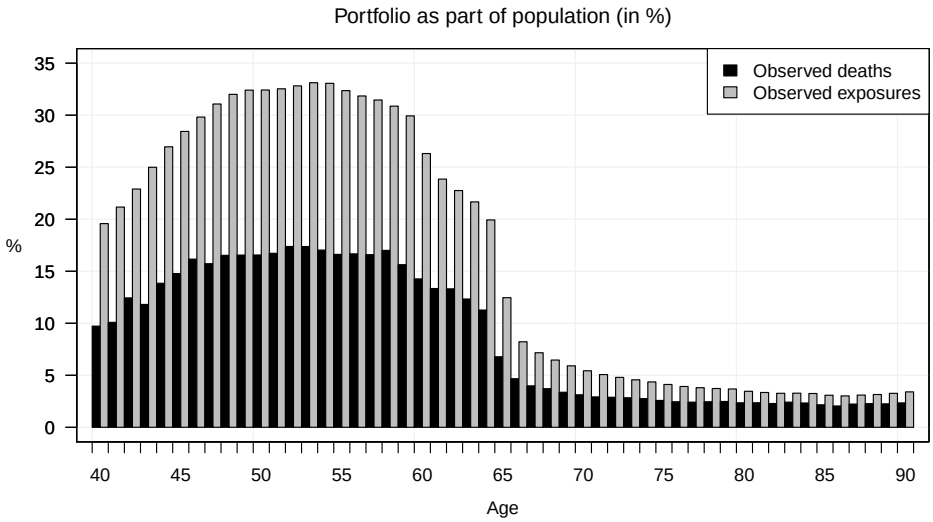


Figure 4.3: The relative size of the portfolio in terms of observed deaths and observed exposures. For each age the relative size is computed as $\sum_t d^{\text{pf}}_{t,x} / \sum_t d^{\text{pop}}_{t,x}$ and $\sum_t E^{\text{pf}}_{t,x} / \sum_t E^{\text{pop}}_{t,x}$, where each summation is over $t \in \mathcal{S}$.

2. The Lee-Carter model is used for population mortality for the England & Wales population for $t \in \{\mathcal{T}, \mathcal{S}\}$ and $x \in \mathcal{X}$, and parameters are estimated in a Bayesian framework. This method is referred to as POP(B);

3. The model described in Section 4.2.1 is used, with a Gamma prior for Θ^i_x. Population and group-specific mortality are estimated simultaneously in a Bayesian framework. This method is referred to as PF(B-G);

4. The model described in Section 4.2.1 is used, with a lognormal prior for Θ^i_x. Population and group-specific mortality are estimated simultaneously in a Bayesian framework. This method is referred to as PF(B-logN).

4.3.1 CMI assured lives - original dataset

In this section we consider the original CMI dataset, and we use the ages and years as described above. In the next section we will artificially reduce the size of the CMI dataset, to investigate the effect of portfolio size on the posterior distribution of the parameters.

Prior distributions. To complete the specifications of the prior distributions, we have to choose the constants that are used in these specifications. We do this in such

a way that the priors contain little information about our prior beliefs, i.e. such that the prior variance is large.

We run four MCMC chains in parallel. For the population mortality parameters α_x, β_x and κ_t we use frequentist estimates $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{\kappa}_t$ as starting points, but in each chain we add some Gaussian noise to obtain different starting values. Using the starting values for β_x and κ_t, we obtain maximum likelihood estimates in each chain for σ^2_β, δ and σ^2_ε, and we use these as initial values for the hyperparameters. For the portfolio-specific factors Θ^i_x we take the initial sample of the MCMC simulations equal to one. For the hyperparameters of Θ^i_x we start with $\rho_i = 0.8$ and $\sigma^2_i = 1$. The constants that complete the specification of the prior distributions and the sampling variances used in the Gibbs and Metropolis(-Hastings) sampling algorithms are chosen as follows:

- To ensure that the prior does not contain much information, we use $a_x = b_x \cdot \exp(\hat{\alpha}_x)$ and $b_x = 0.01$. This way, $\mathbb{E}[\exp(\alpha_x)] = \exp(\hat{\alpha}_x)$ with large variance.
- For $\boldsymbol{\beta}$ we use $\boldsymbol{\mu}_\beta = \frac{1}{\sqrt{X}} \cdot \mathbf{1}_X$ with $\mathbf{1}_X$ a vector with ones of length X, and $c_\beta = 0.01$.
- We use $\mu_\delta = \hat{\delta}$ (the maximum likelihood estimate of the drift, as obtained from the frequentist estimates of the κ_t) and $\sigma^2_\delta = 0.5^2$. For the variance hyperparameter we use $A_\varepsilon = 10$.
- For the Gamma prior on the portfolio-specific factors we use $c^i_x = 1$ for all x and for $i \in \{\text{pf}, \text{rest}\}$. As a result, the prior 95% confidence interval for Θ^i_x is approximately $(0, 4)$.
- For the lognormal prior on the portfolio-specific factors we use $\mu_{\rho_i} = 0$ and $\sigma^2_{\rho_i} = 1$, and for the variance hyperparameter we use $A_i = 10$ for $i \in \{\text{pf}, \text{rest}\}$.
- For the scale parameters used in the proposal densities, we start with $d_\beta = 10^5$, $s^2_{\kappa_t} = 0.05^2$, $s^2_{\Theta^i_x} = 2^2$, $s^2_{\rho_i} = 0.05^2$ and $s^2_{\sigma^2_i} = 0.5^2$. For the definition of other scale parameters, we refer to Appendix 4.A.

Convergence diagnostics. We run 1,100,000 iterations in each chain of the MCMC algorithm. We save every 500th iteration, and during the first 100,000 iterations we calibrate the scale parameters of the proposal distributions every 100th iteration.[7] Our final sample size is 8,000. Our trace plots show good mixing properties, the

[7] The large number of required iterations is due to the high dimension of our model. However, using the Von Mises-Fisher distribution as proposal density speeds up the algorithm considerably, since our Metropolis-Hastings algorithm for $\boldsymbol{\beta}$ consists of only one step, instead of the usual loop over all ages (see e.g. Czado et al. (2005) and Antonio et al. (2015)).

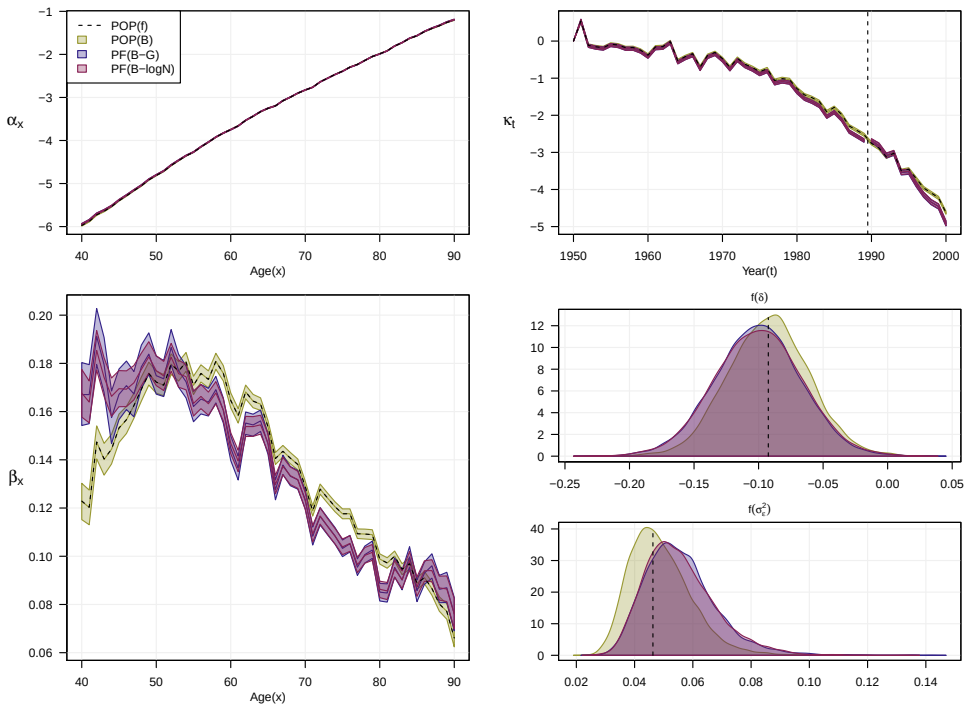


Figure 4.4: Parameter estimates for the CMI and England & Wales datasets using portfolio data for 1990-2000 and ages 40-90. For the frequentist method (POP(f)) we show the maximum likelihood estimates, and for the Bayesian methods (POP(B), PF(B-G) and PF(B-logN)) we show the 95% credible interval (equal-tailed) of the posterior distributions.

Gelman and Rubin statistics converge rapidly towards one, and density plots of the parameters in different chains overlap almost perfectly.[8]

Estimation results. Figure 4.4 shows frequentist and Bayesian estimation results for the population mortality parameters. The parameter estimates for POP(f) are represented by dashed black lines, and the median and the 95% equal-tailed credible intervals derived from the posterior distributions for POP(B), PF(B-G) and PF(B-logN) by respectively green, blue and red lines and areas.

The estimates for POP(f) and POP(B) overlap which means that estimating the Lee-Carter model and the time series model simultaneously gives practically the same best estimates as a two-step frequentist approach. The specification of the hyperpa-

[8]Convergence diagnostics are available in an online appendix, see http://dx.doi.org/10.1017/asb.2017.17.

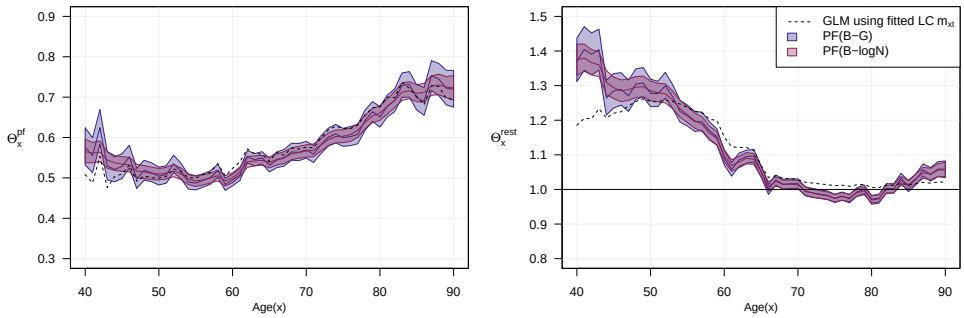


Figure 4.5: Parameter estimates for Θ_x^{pf} and Θ_x^{rest} using the original CMI portfolio.

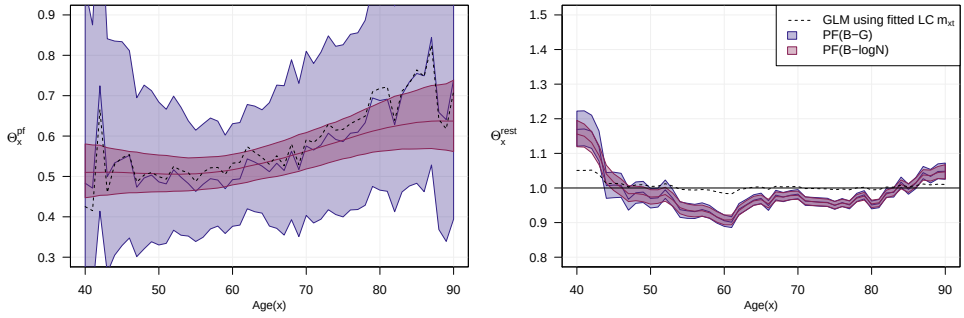


Figure 4.6: Parameter estimates for Θ_x^{pf} and Θ_x^{rest} when the CMI portfolio is reduced by a factor of 100.

rameters therefore seems to have a limited effect on the posterior distributions of the parameters. In models PF(B-G) and PF(B-logN) we also include portfolio data. The credible intervals for α_x and κ_t are similar to the ones found for POP(B). For β_x, however, we observe differences for all ages. The prior specification for Θ_x^i (Gamma versus lognormal) does not have a large effect on the credible interval for β_x. The posterior distributions for the hyperparameters δ and σ_ε^2 are also similar for all model specifications.

Figure 4.5 shows estimates for the portfolio-specific factors using the different methods. The dashed black line represents a frequentist method that corresponds to the method that is most often used in practice. First, the Lee-Carter model is estimated on population mortality. A Poisson GLM with age-dependent factors is then estimated, in which the deaths in the portfolio are explained using the portfolio exposure and the fitted population mortality rate as offset:

$$D_{t,x}^i \sim \text{Poisson}(E_{t,x}^i \hat{\mu}_{t,x}^{\text{LC}} \cdot \Theta_x^i) \tag{4.19}$$

The blue and red areas again correspond to the 95% equal-tailed credible intervals for PF(B-G) and PF(B-logN).

The factors for the portfolio are all below one, implying that mortality in the portfolio is lower than the baseline mortality rate, and in the rest group the factors are generally above one. The baseline mortality rate $\mu_{t,x}$ in our model is estimated using $\mathcal{O}^{\text{pop}} \cup \mathcal{O}^{\text{pf}}$. Therefore, $\Theta_x^{\text{pf}} < 1$ does not automatically imply $\Theta_x^{\text{rest}} > 1$ or vice versa, since both Θ_x^{pf} and Θ_x^{rest} apply only to $\mathcal{O}^{\text{pf}}$. Estimated portfolio factors below one are in line with the results in Dowd et al. (2011) where estimates for the CMI dataset are shown to be significantly lower than for the England and Wales population.

The estimated factors from PF(B-G) show equally irregular behavior as the frequentist estimates for the factors. We find different estimated Lee-Carter parameters for POP(f) on the one hand and PF(B-G) or PF(B-logN) on the other hand. This leads to different baseline hazard rates $\mu_{t,x}$ which explains why the frequentist portfolio-specific factor estimates differ slightly from their Bayesian counterparts. The estimated factors for PF(B-logN), which incorporate dependence between ages within a group, are much smoother than the ones for PF(B-G), where we assume independence.

The posterior means of the mean reversion coefficients for the lognormal prior specification of Θ_x^i are $\rho_{\theta^{\text{pf}}} = 0.997$ and $\rho_{\theta^{\text{rest}}} = 0.999$.[9] We see in Figure 4.5 that the posterior distributions of Θ_x^{pf} have smaller credible intervals than the posterior distributions of Θ_x^{rest} for most ages. This can be explained by the fact that the portfolio is apparently more homogeneous than the remainder of the population for those ages.

We have investigated what happens if we take different constants for the prior distributions, but the estimated effects are hardly affected. Therefore, we conclude that any differences in parameter estimates are caused by differences in the model and the prior specification instead of the prior constants.

Forecasting mortality. Figure 4.7a shows projections of mortality rates from 1) a combination of POP(f) and frequentist estimates of portfolio-specific factors (hereafter indicated by PF(f)), 2) PF(B-G), and 3) PF(B-logN). These mortality projections are constructed as follows. For each MCMC sample, we generate 100 scenarios for future κ_t's using κ_T, δ and σ_ε^2. The mortality rates are then constructed using the other parameters α_x, β_x and Θ_x^i from that sample. Hence, a total of 800,000 scenarios are used to construct the prediction intervals in Figure 4.7a.

In these graphs we only show fitted mortality rates for observations that are included in the likelihood, which means we consider the population for $t < 1990$ and the

[9] Parameters ρ_i close to 1 imply that a random walk (with drift) model might be more appropriate for $\log \Theta_x^i$. For this alternative approach, see Congdon (2009). However, since the estimates of the parameters ρ_i are already close to one, we expect that the posterior distributions for other parameters would not differ significantly when we would use a random walk specification.

portfolio and the rest group for $t \geq 1990$. Projected mortality rates for the portfolio are less uncertain than the ones for the rest group in absolute terms, but not when the uncertainty is expressed as a percentage of the best estimate.

Projections of mortality rates in a Bayesian setting using the two different prior distributions for Θ_x^i show little difference; both the medians and standard deviations of the projections are similar.

We further observe that the prediction intervals from PF(B-G) and PF(B-logN) are similar to those from PF(f), though only the first two include parameter uncertainty. Our projections include uncertainty in the variance parameter σ_ϵ^2 in the time series model, and a higher variance leads to wider prediction intervals whereas a lower variance leads to narrower prediction intervals. Including the uncertainty in the variance parameter therefore does not necessarily lead to wider prediction intervals. The slightly wider prediction intervals further in the future are mainly caused by uncertainty in the drift parameter δ.

4.3.2 CMI assured lives - reduced portfolio size

The CMI dataset is much larger than any portfolio for a single insurance company. Haberman et al. (2014) consider a minimum annual exposure of 25,000 life years and a minimum of eight years of observations sufficient to estimate a mortality model on the portfolio book itself. Chen et al. (2017) investigate the impact of population size on parameter uncertainty and the resulting mortality forecasts, and they find that prediction intervals become wider if the population size decreases. To investigate how these results change if population and portfolio-specific mortality are modeled simultaneously, and to assess how well our model performs on smaller datasets, we artificially reduce the size of the CMI portfolio. We divide observed deaths and exposures by a factor of 100, and the resulting deaths are subsequently rounded to the nearest integer. This ensures that the crude portfolio-specific factors remain largely the same as in the original dataset, which facilitates a comparison of the outcomes. The resulting dataset has on average 25,000 life years annually. We have again defined the rest group in such a way that the population is the disjoint union of the portfolio and the rest group for $(t, x) \in \mathcal{O}^{\text{pf}}$.

We use the same constants to define the prior distributions, and the same initial values and settings in the MCMC algorithm, as in the previous subsection. Convergence diagnostics again show good behavior; they are available in an online appendix.

Estimation results. The posterior distributions for the Lee-Carter parameters are similar to those in Figure 4.4, so we do not show these. Figure 4.6 shows the portfolio-specific factors when estimated for the reduced portfolio. Since there are now fewer lives in our dataset, the observed portfolio death rates show more irregular behavior over the years.

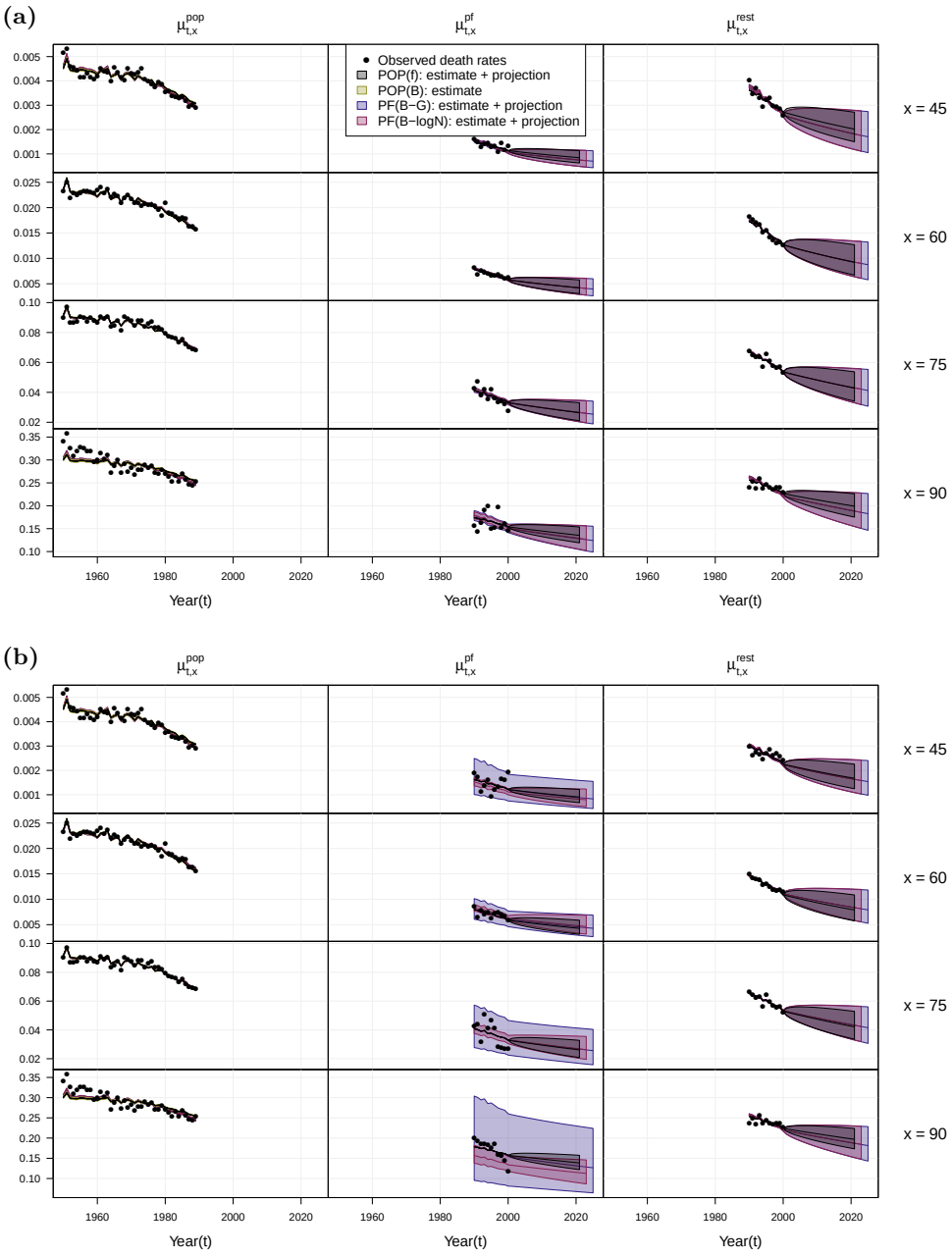


Figure 4.7: Estimated and projected mortality rates from POP(f) in combination with frequentist estimates of group-specific factors (black lines and gray areas), and from PF(B-G) and PF(B-logN) (blue and red areas respectively). Top panel is for the original CMI portfolio, and the bottom panel for the reduced CMI portfolio.

Table 4.1: Predictive mean and standard deviation for future numbers of deaths for different projection horizons, selected ages and for the sum over all ages, for different models that include different sources of uncertainty.

(a) Original CMI portfolio.

	Year	PF(f) TS	PF(f) TS + Pois	PF(B-G) TS + Pois + PU	PF(B-logN) TS + Pois + PU
$x = 45$	2001	(58.1 ; 1.9)	(58.1 ; 7.9)	(54.4 ; 7.8)	(54.6 ; 7.8)
	2010	(51.4 ; 5.4)	(51.4 ; 9.0)	(47.2 ; 9.5)	(47.3 ; 9.6)
	2025	(41.9 ; 7.0)	(41.9 ; 9.5)	(37.4 ; 11.1)	(37.4 ; 11.2)
$x = 65$	2001	(277.5 ; 9.2)	(277.5 ; 19.0)	(270.7 ; 19.5)	(274.1 ; 19.4)
	2010	(245.4 ; 25.8)	(245.4 ; 30.2)	(238.5 ; 32.8)	(241.7 ; 33.2)
	2025	(199.9 ; 33.4)	(199.9 ; 36.2)	(194.1 ; 44.0)	(196.6 ; 44.8)
$x = 85$	2001	(190.4 ; 3.6)	(190.4 ; 14.3)	(187.7 ; 14.8)	(189.6 ; 14.6)
	2010	(177.1 ; 10.7)	(177.1 ; 17.1)	(174.0 ; 18.4)	(175.6 ; 18.5)
	2025	(157.0 ; 15.1)	(157.0 ; 19.6)	(153.7 ; 23.1)	(154.8 ; 23.5)
$40 - 90$	2001	(10302.8 ; 307.8)	(10302.9 ; 324.1)	(10077.4 ; 334.0)	(10077.1 ; 332.7)
	2010	(9225.2 ; 868.2)	(9225.2 ; 873.5)	(8995.6 ; 970.7)	(9002.3 ; 972.8)
	2025	(7682.1 ; 1134.5)	(7682.1 ; 1137.9)	(7484.7 ; 1416.9)	(7493.6 ; 1425.5)

(b) Reduced portfolio size.

	Year	PF(f) TS	PF(f) TS + Pois	PF(B-G) TS + Pois + PU	PF(B-logN) TS + Pois + PU
$x = 45$	2001	(0.6 ; 0.0)	(0.6 ; 0.8)	(0.6 ; 0.8)	(0.6 ; 0.8)
	2010	(0.6 ; 0.1)	(0.6 ; 0.7)	(0.6 ; 0.8)	(0.5 ; 0.7)
	2025	(0.4 ; 0.1)	(0.4 ; 0.7)	(0.5 ; 0.7)	(0.4 ; 0.6)
$x = 65$	2001	(2.7 ; 0.1)	(2.7 ; 1.7)	(2.7 ; 1.7)	(2.9 ; 1.7)
	2010	(2.4 ; 0.3)	(2.4 ; 1.6)	(2.4 ; 1.6)	(2.5 ; 1.6)
	2025	(2.0 ; 0.3)	(2.0 ; 1.4)	(2.0 ; 1.5)	(2.1 ; 1.5)
$x = 85$	2001	(2.1 ; 0.0)	(2.1 ; 1.4)	(2.1 ; 1.5)	(1.7 ; 1.3)
	2010	(1.9 ; 0.1)	(1.9 ; 1.4)	(1.9 ; 1.5)	(1.6 ; 1.3)
	2025	(1.7 ; 0.2)	(1.7 ; 1.3)	(1.7 ; 1.4)	(1.4 ; 1.2)
$40 - 90$	2001	(103.2 ; 3.1)	(103.2 ; 10.6)	(104.7 ; 11.1)	(102.3 ; 10.9)
	2010	(92.4 ; 8.7)	(92.4 ; 12.9)	(94.3 ; 14.2)	(92.0 ; 14.0)
	2025	(77.0 ; 11.3)	(77.0 ; 14.3)	(79.3 ; 17.5)	(77.5 ; 17.4)

The right graph of Figure 4.6 shows the estimated factors for the rest group, Θ_x^{rest}. Now that the portfolio has become smaller, this group constitutes a larger part of the general population. As a result, the posterior means are in general closer to one and the posterior credible intervals are slightly smaller. In the left graph, we observe how this results in frequentist estimates for Θ_x^{pf} which fluctuate more over the years (see the black line). The estimates for Θ_x^{pf} are also more irregular in the PF(B-G) model (in blue) when compared with the original portfolio, and the corresponding posterior credible interval is much wider. In the PF(B-logN) model (in red), the estimates are smoother than in Figure 4.5 while the posterior credible interval is again wider than before, but much less so than for PF(B-G). This is due to the smoothing characteristic of PF(B-logN): information from ages near x influences the estimates for Θ_x^i. The PF(B-logN) prior is more parsimonious than the one for PF(B-G) (it is a 'shrinkage prior') and the effect of the prior specification on the posterior distribution is stronger if less data is available. We believe it is reasonable to assume that portfolio-specific factors for ages close to each other are related, which makes the posterior credible intervals from PF(B-logN) more plausible than those from PF(B-G).

Forecasting mortality. Figure 4.7b shows projections of mortality rates from PF(f), PF(B-G) and PF(B-logN) using the reduced CMI portfolio, which can be compared to the mortality rate projections in Figure 4.7a where the whole CMI portfolio is used.

The credible intervals and prediction intervals from PF(B-G) for mortality rates in the reduced CMI portfolio are much wider, which is caused by the wider credible intervals for Θ_x^{pf}, see Figure 4.6. The intervals for mortality rates for PF(B-logN) are a bit wider when the smaller CMI portfolio is used, but much less so than the ones for PF(B-G). Based on Figure 4.6 and 4.7b we thus conclude that parameter uncertainty in portfolio-specific factors can be substantial for small portfolios, and that the level of uncertainty may strongly depend on the prior specification for the portfolio-specific factors.

Hoem (1973) already identified different sources of uncertainty in mortality predictions but no quantification was given for the relative impact of these differences. In Table 4.1 we investigate the relative impacts by showing the mean and standard deviation of the predicted numbers of deaths for future times $t \in \{2001, 2010, 2025\}$ based on observations until $s_S = 2000$, for $x = \{45, 65, 85\}$ and for all ages combined. To get the appropriate comparison, we use the exposures at time s_S for later times as well, so we take $E_{t,x}^{\text{pf}} = E_{s_S,x}^{\text{pf}}$ for $t \geq s_S$. The mortality scenarios correspond to the ones used in Figure 4.7a and 4.7b, and the figures presented in Table 4.1a and 4.1b are constructed using four different methods:

- Using PF(f) we predict mortality rates $\mu_{t,x}^{\text{pf}}$, taking into account uncertainty in the projection of the time series κ_t, but we do not include parameter uncertainty.

These mortality rates are then multiplied by the exposures $E^{\text{pf}}_{t,x} = E^{\text{pf}}_{T,x}$ to generate the expected number of deaths given the scenario for mortality rates, i.e. $\mathbb{E}[D^{\text{pf}}_{t,x} \mid \mu^{\text{pf}}_{t,x}]$ (first column with intervals);

- Using PF(f) we predict mortality rates $\mu^{\text{pf}}_{t,x}$, taking into account uncertainty in the projection of κ_t. We still do not include parameter uncertainty but for each generated mortality scenario, we include Poisson randomness[10] by drawing random numbers of deaths $D^{\text{pf}}_{t,x} \sim \text{Poisson}(E^{\text{pf}}_{t,x}\mu^{\text{pf}}_{t,x})$ (second column);

- Using PF(B-G) we predict mortality rates $\mu^{\text{pf}}_{t,x}$, taking into account uncertainty in the projection of κ_t. Parameter uncertainty is now included, since we use the MCMC samples. For each generated mortality scenario, we draw random numbers of deaths $D^{\text{pf}}_{t,x} \sim \text{Poisson}(E^{\text{pf}}_{t,x}\mu^{\text{pf}}_{t,x})$ (third column);

- For PF(B-logN) our approach is similar to that for PF(B-G) (fourth column).

If only uncertainty in the evolution of the time series κ_t is taken into account, the uncertainty in the conditional expectation of $D^{\text{pf}}_{t,x}$ given $\mu^{\text{pf}}_{t,x}$ can be very small for small portfolios, see the first column in Table 4.1b. For larger portfolios the uncertainty is much larger due to the higher exposures, as shown in Table 4.1a. The uncertainty also becomes larger if uncertainty in the individual number of deaths (Poisson noise) is added. This is shown in the second column, and the effect is of course stronger for the smaller portfolio.

If we compare the results with parameter uncertainty (third and fourth column) and without parameter uncertainty (second column), we conclude that the impact on the predicted numbers of deaths is negligible compared to the impact of the Poisson noise due to individual deaths. The mortality prediction intervals for the smaller portfolio size with the Gamma prior are very wide, as shown in Figure 4.7b. However, the uncertainty in the future numbers of deaths is similar to the cases where we only include time series and Poisson uncertainty. Therefore, we conclude that for the time horizons considered here, the Poisson noise due to individual deaths is more important than parameter uncertainty in the portfolio-specific factors, and this turns out to be true for the smaller but also for the larger portfolio.

Given fixed mortality rates, death numbers at different ages are independent (but of course not identically distributed). Therefore, one might expect that the relevance of Poisson randomness disappears if we consider a whole insurance portfolio and thus look at the sum of random death numbers over all ages. Predicted means and standard deviations for this sum over all ages are shown in the bottom rows in Table 4.1a and 4.1b. We observe that for the small portfolio, including Poisson uncertainty

[10] When sampling the individual deaths in practice (e.g. for portfolio valuation purposes), one may prefer to use the Bernoulli distribution. Here we use the Poisson distribution to remain consistent with the approach used for estimation.

leads to larger uncertainty even at the portfolio level, but this is not true for large portfolios. Comparing the second to the third and fourth column, we further see that parameter uncertainty has little impact on the short horizon, but the effect of parameter uncertainty increases with the projection horizon, regardless of portfolio size. This is explained by the uncertainty in the mortality trend parameter δ which has an effect on mortality rates that cumulates over time. We conclude:

- For large portfolios, individual mortality risk (modeled through Poisson noise) is important for individual ages, but not at the portfolio level. For small portfolios, individual mortality risk is important both for individual ages and at the portfolio level;
- For both large and small portfolios, parameter uncertainty in portfolio-specific factors is not relevant since it is overshadowed by Poisson noise;
- For both large and small portfolios, parameter uncertainty in the mortality trend is not relevant in the short term, but of increasing importance if the projection horizon increases.

The results regarding trend uncertainty are in line with the results for small portfolios in Haberman et al. (2014). Further, Hári et al. (2008b) conclude that individual mortality risk (which we capture through Poisson uncertainty) may be just as important for small portfolios as time series uncertainty and parameter uncertainty, whereas individual mortality risk is relatively small for large portfolios. Our results indicate that even at the portfolio level Poisson and trend uncertainty cannot be ignored.

4.4 Conclusion

Proper risk management for portfolios in life insurance companies or pension funds requires a reliable method to estimate the distribution of future deaths in such portfolios. This involves the modeling of population-wide mortality trends, a specification of portfolio-specific deviations from this trend, and the conditional distribution for the individual deaths in a portfolio given its mortality rates. In this chapter we use Bayesian inference to analyze these three sources of uncertainty in life insurance portfolio data. This may help to generate scenarios for survival in a portfolio in which these three different components in the predictions can be explicitly distinguished.

The law of large numbers implies that the last component will be relatively small for very large portfolios. But when the portfolio under consideration is small or when observations have only been available for a limited number of years, it may be difficult to know a priori what part of the fluctuations in the observations over age and time should be assigned to genuine changes in mortality over time, to noise in the observations and to parameter uncertainty. For those cases, we believe that our

method may turn out to be a useful alternative to what has been proposed in the actuarial literature so far.

By using both the CMI dataset of assured male lives and a scaled version of that dataset, we show that estimates of the difference between country-wide and portfolio-specific hazard rates strongly depend on a priori assumptions about the age-dependence of that difference. Assuming that there is no dependence for different ages can give unrealistically large posterior credible intervals for portfolio-specific factors in small portfolios, while an alternative based on an autoregressive smoothing prior gives much more satisfactory results.

However, the impact of uncertainty in the portfolio-specific factors on the predictive distributions of future number of deaths in the portfolio is negligible compared to the Poisson noise that is added by individual deaths, regardless of the projection horizon. As the projection horizon increases, the effect of parameter uncertainty on predictive distributions becomes increasingly relevant, which is solely caused by uncertainty in the mortality trend. This reinforces our conclusion that a full analysis for small portfolios must always be based on an explicit description of the different sources of uncertainty in the predictive distributions of future deaths.

Appendix

4.A Posterior distributions

We derive the posterior distribution for all parameters in the model. For convenience, we define the following variables

$$\begin{aligned}
&\boldsymbol{D} = \{\boldsymbol{D}^{\text{pop}}, \boldsymbol{D}^{\text{pf}}, \boldsymbol{D}^{\text{rest}}\}, && \boldsymbol{E} = \{\boldsymbol{E}^{\text{pop}}, \boldsymbol{E}^{\text{pf}}, \boldsymbol{E}^{\text{rest}}\}, && \\
&\boldsymbol{\alpha} = \{\alpha_{x_1}, \ldots, \alpha_{x_X}\}, && \boldsymbol{\beta} = \{\beta_{x_1}, \ldots, \beta_{x_X}\}, && \boldsymbol{\kappa} = \{\kappa_{t_1}, \ldots, \kappa_{t_T}\}, \\
&\boldsymbol{\Theta} = \{\boldsymbol{\Theta}^{\text{pf}}, \boldsymbol{\Theta}^{\text{rest}}\}, && \boldsymbol{\rho}_\theta = \{\rho_{\text{pf}}, \rho_{\text{rest}}\}, && \boldsymbol{\sigma}^2_\theta = \{\sigma^2_{\text{pf}}, \sigma^2_{\text{rest}}\},
\end{aligned}$$

with $(D^{\text{pop}}_{t,x}, E^{\text{pop}}_{t,x})$ defined on $(t,x) \in \mathcal{O}^{\text{pop}}$ and $(D^{i}_{t,x}, E^{i}_{t,x})$ defined on $(t,x) \in \mathcal{O}^{\text{pf}}$ for $i \in \{\text{pf}, \text{rest}\}$. See Section 4.2.1 for the definition of $\mathcal{O}^{\text{pop}}$ and $\mathcal{O}^{\text{pf}}$.

We further define the set $\boldsymbol{\Lambda}$ that contains both data and parameters:

$$\boldsymbol{\Lambda} = \{\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \delta, \sigma^2_\varepsilon, \boldsymbol{\Theta}, \boldsymbol{\rho}_\theta, \boldsymbol{\sigma}^2_\theta\},$$

and remark that $\boldsymbol{\rho}_\theta$ and $\boldsymbol{\sigma}^2_\theta$ are not needed when we use a Gamma prior for the portfolio-specific factors.

4.A.1 Age parameters for population mortality

Gibbs sampling for α_x

The individual α_x's are independent. Therefore, the posterior distribution for a single $e_x = \exp(\alpha_x)$ with $x_1 \leq x \leq x_X$ is given by

$$\begin{aligned}
f(e_x|\boldsymbol{\Lambda}\backslash\{e_x\}) &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\Theta}^{\text{pf}}, \boldsymbol{\Theta}^{\text{rest}}) f(e_x) \qquad (4.20) \\
&\propto \prod_{t \in \mathcal{T}} \left(e^{-E^{\text{pf}}_{t,x} e_x \exp[\beta_x \kappa_t] \Theta^{\text{pf}}_x} \frac{(E^{\text{pf}}_{t,x} e_x \exp[\beta_x \kappa_t] \Theta^{\text{pf}}_x)^{D^{\text{pf}}_{t,x}}}{D^{\text{pf}}_{t,x}!} \right)^{I^{\text{pf}}_{t,x}} \\
&\quad \times \prod_{t \in \mathcal{T}} \left(e^{-E^{\text{rest}}_{t,x} e_x \exp[\beta_x \kappa_t] \Theta^{\text{rest}}_x} \frac{(E^{\text{rest}}_{t,x} e_x \exp[\beta_x \kappa_t] \Theta^{\text{rest}}_x)^{D^{\text{rest}}_{t,x}}}{D^{\text{rest}}_{t,x}!} \right)^{I^{\text{rest}}_{t,x}}
\end{aligned}$$

$$
\begin{aligned}
&\times \prod_{t\in\mathcal{T}} \left(e^{-E^{\text{pop}}_{t,x} e_x \exp[\beta_x\kappa_t]} \frac{(E^{\text{pop}}_{t,x} e_x \exp[\beta_x\kappa_t])^{D^{\text{pop}}_{t,x}}}{D^{\text{pop}}_{t,x}!} \right)^{I^{\text{pop}}_{t,x}} \\
&\times \frac{b_x^{a_x}}{\Gamma(a_x)} e_x^{a_x-1} \exp[-b_x e_x] \\
&\propto \exp[-(b_x+d_x)e_x] \cdot e_x^{a_x+D_{\bullet x}-1},
\end{aligned}
$$

with

$$
\begin{aligned}
d_x = \sum_{t\in\mathcal{T}} \Big\{ & I^{\text{pf}}_{t,x} \left(E^{\text{pf}}_{t,x} \exp[\beta_x\kappa_t]\Theta^{\text{pf}}_x \right) + I^{\text{rest}}_{t,x} \left(E^{\text{rest}}_{t,x} \exp[\beta_x\kappa_t]\Theta^{\text{rest}}_x \right) \\
& + I^{\text{pop}}_{t,x} \left(E^{\text{pop}}_{t,x} \exp[\beta_x\kappa_t] \right) \Big\}
\end{aligned}
$$

and

$$
D_{\bullet x} = \sum_{t\in\mathcal{T}} \left\{ I^{\text{pf}}_{t,x} \cdot D^{\text{pf}}_{t,x} + I^{\text{rest}}_{t,x} \cdot D^{\text{rest}}_{t,x} + I^{\text{pop}}_{t,x} \cdot D^{\text{pop}}_{t,x} \right\} = \sum_{t\in\mathcal{T}} D^{\text{pop}}_{t,x}.
$$

The last line in (4.20) is proportional to a Gamma$(a_x + D_{\bullet x}, b_x + d_x)$ distribution. Therefore, we can use Gibbs sampling to draw a new value of e_x, which can subsequently be transformed into a new value of α_x.

Metropolis sampling for β_x

The posterior distribution for $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{\Lambda}\backslash\{\boldsymbol{\beta}\}) &\propto f(\boldsymbol{D}|\boldsymbol{E},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa},\boldsymbol{\Theta}^{\text{pf}},\boldsymbol{\Theta}^{\text{rest}}) f(\boldsymbol{\beta}) \qquad (4.21) \\
&\propto \prod_{x\in\mathcal{X}} \prod_{t\in\mathcal{T}} \left(e^{-E^{\text{pf}}_{t,x} \exp[\alpha_x+\beta_x\kappa_t]\Theta^{\text{pf}}_x} \frac{(E^{\text{pf}}_{t,x} \exp[\alpha_x+\beta_x\kappa_t]\Theta^{\text{pf}}_x)^{D^{\text{pf}}_{t,x}}}{D^{\text{pf}}_{t,x}!} \right)^{I^{\text{pf}}_{t,x}} \\
&\quad \times \prod_{x\in\mathcal{X}} \prod_{t\in\mathcal{T}} \left(e^{-E^{\text{rest}}_{t,x} \exp[\alpha_x+\beta_x\kappa_t]\Theta^{\text{rest}}_x} \frac{(E^{\text{rest}}_{t,x} \exp[\alpha_x+\beta_x\kappa_t]\Theta^{\text{rest}}_x)^{D^{\text{rest}}_{t,x}}}{D^{\text{rest}}_{t,x}!} \right)^{I^{\text{rest}}_{t,x}} \\
&\quad \times \prod_{x\in\mathcal{X}} \prod_{t\in\mathcal{T}} \left(e^{-E^{\text{pop}}_{t,x} \exp[\alpha_x+\beta_x\kappa_t]} \frac{(E^{\text{pop}}_{t,x} \exp[\alpha_x+\beta_x\kappa_t])^{D^{\text{pop}}_{t,x}}}{D^{\text{pop}}_{t,x}!} \right)^{I^{\text{pop}}_{t,x}} \\
&\quad \times \exp(c_\beta \boldsymbol{\mu}_\beta^T \boldsymbol{\beta}).
\end{aligned}
$$

Given a current value $\tilde{\boldsymbol{\beta}}$ and scaling parameter d_β, we sample a proposal $\hat{\boldsymbol{\beta}}$ from the distribution vMF$(\tilde{\boldsymbol{\beta}}, d_\beta)$. The proposal distribution is symmetric, and the acceptance probability is thus given by:

$$
\phi = \min \left\{ \frac{f(\hat{\boldsymbol{\beta}}|\boldsymbol{\Lambda}\backslash\{\hat{\boldsymbol{\beta}}\})}{f(\tilde{\boldsymbol{\beta}}|\boldsymbol{\Lambda}\backslash\{\tilde{\boldsymbol{\beta}}\})}; 1 \right\}.
$$

4.A.2 Period parameters for population mortality

Metropolis sampling for κ_t

Define $\boldsymbol{\kappa}_{-t} = \{\kappa_{t_1}, \ldots, \kappa_{t-1}, \kappa_{t+1}, \ldots, \kappa_{s_S}\}$. The posterior distribution of κ_t for $t_1 < t < t_T$ and $s_1 < t \leq s_S$ is given by

$$\begin{aligned} f(\kappa_t|\boldsymbol{\Lambda}\backslash\{\kappa_t\}) &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\Theta}^{\text{pf}}, \boldsymbol{\Theta}^{\text{rest}}) f(\boldsymbol{\kappa}|\kappa_{t_1}, \delta, \sigma_\varepsilon^2) \\ &\propto \prod_{x\in\mathcal{X}} \left[\exp\left(-E_{t,x}^{\text{pf}} \exp[\alpha_x + \beta_x\kappa_t]\Theta_x^{\text{pf}}\right) \exp\left(D_{t,x}^{\text{pf}}\beta_x\kappa_t\right)\right]^{I_{t,x}^{\text{pf}}} \\ &\quad \times \prod_{x\in\mathcal{X}} \left[\exp\left(-E_{t,x}^{\text{rest}} \exp[\alpha_x + \beta_x\kappa_t]\Theta_x^{\text{rest}}\right) \exp\left(D_{t,x}^{\text{rest}}\beta_x\kappa_t\right)\right]^{I_{t,x}^{\text{rest}}} \\ &\quad \times \prod_{x\in\mathcal{X}} \left[\exp\left(-E_{t,x}^{\text{pop}} \exp[\alpha_x + \beta_x\kappa_t]\right) \exp\left(D_{t,x}^{\text{pop}}\beta_x\kappa_t\right)\right]^{I_{t,x}^{\text{pop}}} \\ &\quad \times f(\kappa_t|\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2), \end{aligned} \tag{4.22}$$

and the expression in the last line can be simplified:

- for $t_1 < t < t_T$ and $s_1 < t < s_S$:

$$\begin{aligned} f(\kappa_t|\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2) &\propto f(\kappa_t|\kappa_{t-1}, \delta, \sigma_\varepsilon^2) f(\kappa_{t+1}|\kappa_t, \delta, \sigma_\varepsilon^2) \\ &\sim \text{N}\left(\tfrac{1}{2}(\kappa_{t-1} + \kappa_{t+1}), \tfrac{1}{2}\sigma_\varepsilon^2\right), \end{aligned}$$

- for $t = s_S$:

$$\begin{aligned} f(\kappa_t|\boldsymbol{\kappa}_{-t}, \delta, \sigma_\varepsilon^2) &\propto f(\kappa_t|\kappa_{t-1}, \delta, \sigma_\varepsilon^2) \\ &\sim \text{N}\left(\kappa_{t-1} + \delta, \sigma_\varepsilon^2\right). \end{aligned}$$

For κ_{t_T} and κ_{s_1} we derive the *joint* posterior distribution, since we have applied the restriction $\kappa_{t_T} = \kappa_{s_1}$. It is given by

$$\begin{aligned} f(\kappa_{t_T}, \kappa_{s_1}|\boldsymbol{\Lambda}\backslash\{\kappa_{t_T}, \kappa_{s_1}\}) &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\Theta}^{\text{pf}}, \boldsymbol{\Theta}^{\text{rest}}) f(\boldsymbol{\kappa}|\kappa_{t_1}, \delta, \sigma_\varepsilon^2) \\ &\propto \prod_{t=t_T}^{s_1} \prod_{x\in\mathcal{X}} \left[\exp\left(-E_{t,x}^{\text{pf}} \exp[\alpha_x + \beta_x\kappa_t]\Theta_x^{\text{pf}}\right) \exp\left(D_{t,x}^{\text{pf}}\beta_x\kappa_t\right)\right]^{I_{t,x}^{\text{pf}}} \\ &\quad \times \prod_{t=t_T}^{s_1} \prod_{x\in\mathcal{X}} \left[\exp\left(-E_{t,x}^{\text{rest}} \exp[\alpha_x + \beta_x\kappa_t]\Theta_x^{\text{rest}}\right) \exp\left(D_{t,x}^{\text{rest}}\beta_x\kappa_t\right)\right]^{I_{t,x}^{\text{rest}}} \\ &\quad \times \prod_{t=t_T}^{s_1} \prod_{x\in\mathcal{X}} \left[\exp\left(-E_{t,x}^{\text{pop}} \exp[\alpha_x + \beta_x\kappa_t]\right) \exp\left(D_{t,x}^{\text{pop}}\beta_x\kappa_t\right)\right]^{I_{t,x}^{\text{pop}}} \\ &\quad \times f(\kappa_{t_T}, \kappa_{s_1}|\boldsymbol{\kappa}_{-\{t_T,s_1\}}, \delta, \sigma_\varepsilon^2), \end{aligned} \tag{4.23}$$

and the expression in the last line can be simplified:

$$\begin{aligned} f(\kappa_{t_T}, \kappa_{s_1} | \boldsymbol{\kappa}_{-\{t_T,s_1\}}, \delta, \sigma_\varepsilon^2) &= f(\kappa_{t_T} | \boldsymbol{\kappa}_{-\{t_T,s_1\}}, \delta, \sigma_\varepsilon^2) \\ &\propto f(\kappa_{t_T} | \kappa_{t_T-1}, \delta, \sigma_\varepsilon^2) f(\kappa_{s_1+1} | \kappa_{s_1}, \delta, \sigma_\varepsilon^2) \\ &\sim \mathrm{N}\left(\tfrac{1}{2}(\kappa_{t_T-1} + \kappa_{s_1+1}), \tfrac{1}{2}\sigma_\varepsilon^2\right), \end{aligned}$$

Given a current value $\tilde{\kappa}_t$ and Metropolis sampling variance $s^2_{\kappa_t}$, we sample a proposal $\hat{\kappa}_t$ from the distribution $\mathrm{N}(\tilde{\kappa}_t, s^2_{\kappa_t})$. This proposal distribution is symmetric, and the acceptance probability is thus given by

$$\phi = \min\left\{ \frac{f(\hat{\kappa}_t | \boldsymbol{\Lambda} \backslash \{\hat{\kappa}_t\})}{f(\tilde{\kappa}_t | \boldsymbol{\Lambda} \backslash \{\tilde{\kappa}_t\})}; 1 \right\}.$$

Gibbs sampling for δ

Define $\Delta\kappa_t = \kappa_t - \kappa_{t-1}$. Note that we have applied the restriction $\kappa_{s_1} := \kappa_{s_1-1} = \kappa_{t_T}$, and that summations are therefore not simply over all t. The posterior distribution of δ is given by

$$\begin{aligned} f(\delta | \boldsymbol{\Lambda} \backslash \{\delta\}) &\propto f(\boldsymbol{\kappa} | \kappa_1, \delta, \sigma_\varepsilon^2) f(\delta) \\ &\propto \exp\left[-\sum_{t=t_1+1}^{t_T} \frac{[\Delta\kappa_t - \delta]^2}{2\sigma_\varepsilon^2} - \sum_{t=s_1+1}^{s_S} \frac{[\Delta\kappa_t - \delta]^2}{2\sigma_\varepsilon^2} \right] \cdot \exp\left[-\frac{[\delta - \mu_\delta]^2}{2\sigma_\delta^2} \right] \\ &\propto \exp\left[-\frac{1}{2a_\delta} \left(\delta^2 - 2\delta b_\delta \right) \right] \\ &\sim \mathrm{N}\left(b_\delta, a_\delta\right), \end{aligned} \tag{4.24}$$

with

$$a_\delta = \frac{\sigma_\varepsilon^2}{(T^* - 2) + \sigma_\varepsilon^2/\sigma_\delta^2},$$

and

$$\begin{aligned} b_\delta = &\frac{(T^* - 2)}{(T^* - 2) + \sigma_\varepsilon^2/\sigma_\delta^2} \cdot \left(\frac{1}{(T^* - 2)} \left\{ \sum_{t=t_1+1}^{t_T} \Delta\kappa_t + \sum_{t=s_1+1}^{s_S} \Delta\kappa_t \right\} \right) \\ &+ \frac{\sigma_\varepsilon^2/\sigma_\delta^2}{(T^* - 2) + \sigma_\varepsilon^2/\sigma_\delta^2} \cdot \mu_\delta. \end{aligned}$$

We can use Gibbs sampling to draw a new value for δ.

Gibbs sampling for σ_ε^2

The posterior distribution of σ_ε^2 is given by

$$\begin{aligned} f(\sigma_\varepsilon^2|\boldsymbol{\Lambda}\backslash\{\sigma_\varepsilon^2\}) &\propto f(\boldsymbol{\kappa}|\kappa_1,\delta,\sigma_\varepsilon^2)f(\sigma_\varepsilon^2) \qquad (4.25)\\ &= \prod_{t=t_1+1}^{t_T}\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\exp\left[-\frac{[\Delta\kappa_t-\delta]^2}{2\sigma_\varepsilon^2}\right]\cdot\prod_{t=s_1+1}^{s_S}\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\exp\left[-\frac{[\Delta\kappa_t-\delta]^2}{2\sigma_\varepsilon^2}\right]\\ &\quad\times\sigma_\varepsilon^{-1}\cdot 1_{[0\leq\sigma_\varepsilon\leq A_\varepsilon]}\\ &\propto (\sigma_\varepsilon^{-2})^{\frac{T^*-1}{2}}\exp\left[-(\sigma_\varepsilon^{-2})\cdot\frac{1}{2}\left(\sum_{t=t_1+1}^{t_T}(\Delta\kappa_t-\delta)^2+\sum_{t=s_1+1}^{s_S}(\Delta\kappa_t-\delta)^2\right)\right]. \end{aligned}$$

Therefore, we know that the posterior distribution of σ_ε^{-2} is

$$\begin{aligned} &f(\sigma_\varepsilon^{-2}|\boldsymbol{\Lambda}\backslash\{\sigma_\varepsilon^2\})\\ &\quad\propto (\sigma_\varepsilon^{-2})^{\frac{T^*-1}{2}-1-1}\cdot\exp\left[-(\sigma_\varepsilon^{-2})\cdot\frac{1}{2}\left(\sum_{t=t_1+1}^{t_T}(\Delta\kappa_t-\delta)^2+\sum_{t=s_1+1}^{s_S}(\Delta\kappa_t-\delta)^2\right)\right]\\ &\quad\sim \text{Gamma}\left(\frac{T^*-3}{2},\frac{1}{2}\left\{\sum_{t=t_1+1}^{t_T}(\Delta\kappa_t-\delta)^2+\sum_{t=s_1+1}^{s_S}(\Delta\kappa_t-\delta)^2\right\}\right). \end{aligned}$$

We can use Gibbs sampling to draw new values of σ_ε^{-2} which can be transformed into σ_ε^2.

4.A.3 Portfolio-specific mortality - Gamma prior

The posterior distribution of Θ_x^i for $i\in\{\text{pf},\text{rest}\}$ and $y_1\leq x\leq y_Y$ is given by

$$\begin{aligned} f(\Theta_x^i|\boldsymbol{\Lambda}\backslash\{\Theta_x^i\}) &\propto f(\boldsymbol{D}|\boldsymbol{E},\boldsymbol{\Theta}^{\text{pf}},\boldsymbol{\Theta}^{\text{rest}},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\kappa})f(\Theta_x^i) \qquad (4.26)\\ &\propto \prod_{t\in\mathcal{S}}\left(e^{-E_{t,x}^i\exp[\alpha_x+\beta_x\kappa_t]\Theta_x^i}\frac{(E_{t,x}^i\exp[\alpha_x+\beta_x\kappa_t]\Theta_x^i)^{D_{t,x}^i}}{D_{t,x}^i!}\right)^{I_{t,x}^i}\\ &\quad\times\frac{(c_x^i)^{c_x^i}}{\Gamma(c_x^i)}(\Theta_x^i)^{c_x^i-1}\exp[-c_x^i\Theta_x^i]\\ &\propto \exp[-(c_x^i+f_x^i)\Theta_x^i]\cdot(\Theta_x^i)^{c_x^i+D_{\bullet x}^i-1}, \end{aligned}$$

with

$$f_x^i=\sum_{t\in\mathcal{S}}I_{t,x}^i\cdot E_{t,x}^i\exp[\alpha_x+\beta_x\kappa_t]\quad\text{and}\quad D_{\bullet x}^i=\sum_{t\in\mathcal{S}}I_{t,x}^i\cdot D_{t,x}^i.$$

The last line in (4.26) is proportional to a Gamma$(c_x^i+D_{\bullet x}^i, c_x^i+f_x^i)$ distribution and we can therefore use Gibbs sampling to obtain new values for Θ_x^i. Note that the

posterior mean can be written as

$$\frac{c_x^i}{c_x^i + \sum_{t\in\mathcal{S}} I_{t,x}^i \cdot E_{t,x}^i \mu_{t,x}} \cdot 1 + \frac{\sum_{t\in\mathcal{S}} I_{t,x}^i \cdot E_{t,x}^i \mu_{t,x}}{c_x^i + \sum_{t\in\mathcal{S}} I_{t,x}^i \cdot E_{t,x}^i \mu_{t,x}} \cdot \frac{\sum_{t\in\mathcal{S}} I_{t,x}^i \cdot D_{t,x}^i}{\sum_{t\in\mathcal{S}} I_{t,x}^i \cdot E_{t,x}^i \mu_{t,x}}.$$

If c_x^i is chosen small relative to $\sum_{t\in\mathcal{S}} I_{t,x}^i \cdot E_{t,x}^i \mu_{t,x}$, the posterior mean is close to $\frac{\sum_{t\in\mathcal{S}} I_{t,x}^i \cdot D_{t,x}^i}{\sum_{t\in\mathcal{S}} I_{t,x}^i \cdot E_{t,x}^i \mu_{t,x}}$ which is often used in practice to determine portfolio-specific factors.

4.A.4 Portfolio-specific mortality - lognormal prior

Define $\boldsymbol{\Theta}^i = \{\Theta_{y_1}^i, \ldots, \Theta_{y_Y}^i\}$. The distribution generated by the mean reverting process for the logarithm of the portfolio-specific factors in (4.17) can also be written as a multivariate lognormal distribution (Purcaru et al., 2004, Section 3.3.2):

$$\boldsymbol{\Theta}^i \sim \text{lognormal}(\tilde{\boldsymbol{\mu}}_i, \boldsymbol{\Sigma}_i), \tag{4.27}$$

with $\tilde{\boldsymbol{\mu}}_i = -\frac{1}{2}\sigma_i^2 \mathbf{1}_Y$ where $\mathbf{1}_Y$ is a column vector of ones of length Y and $(\boldsymbol{\Sigma}_i)_{xy} = \rho_i^{|x-y|}\sigma_i^2$. Before we derive the posterior distribution for Θ_x^i and the hyperparameters, we define the following variables and relations:

$$\begin{aligned}
\boldsymbol{\Sigma}_i &= \sigma_i^2 \cdot \boldsymbol{\Gamma}(\rho_i) \\
\boldsymbol{\Sigma}_i^{-1} &= \tfrac{1}{\sigma_i^2} \cdot \boldsymbol{\Gamma}^{-1}(\rho_i) = \tfrac{1}{\sigma_i^2}\tfrac{1}{1-\rho_i^2} \cdot \widetilde{\boldsymbol{\Gamma}}^{-1}(\rho_i) \\
|\boldsymbol{\Sigma}_i| &= \left|\sigma_i^2 \cdot \boldsymbol{\Gamma}(\rho_i)\right| = (\sigma_i^2)^Y \cdot (1-\rho_i^2)^{Y-1} \\
\boldsymbol{\Psi}^i &= \ln \boldsymbol{\Theta}^i - \tilde{\boldsymbol{\mu}}_i = \ln \boldsymbol{\Theta}^i + \tfrac{1}{2}\sigma_i^2 \mathbf{1}_Y,
\end{aligned}$$

with

$$\boldsymbol{\Gamma}(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho^{Y-2} & \rho^{Y-1} \\ \rho & 1 & \cdots & \rho^{Y-3} & \rho^{Y-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{Y-2} & \rho^{Y-3} & \cdots & 1 & \rho \\ \rho^{Y-1} & \rho^{Y-2} & \cdots & \rho & 1 \end{pmatrix}, \quad \widetilde{\boldsymbol{\Gamma}}^{-1}(\rho) = \begin{pmatrix} 1 & -\rho & 0 & \ldots & 0 \\ -\rho & 1+\rho^2 & -\rho & \ddots & 0 \\ 0 & -\rho & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}.$$

Metropolis-Hastings sampling for Θ_x^i

Define $\boldsymbol{\Theta}_{-j}^i = \{\Theta_{y_1}^i, \ldots, \Theta_{j-1}^i, \Theta_{j+1}^i, \ldots, \Theta_{y_Y}^i\}$. The posterior distribution of Θ_x^i for $i \in \{\text{pf}, \text{rest}\}$ and $y_1 \le x \le y_Y$ is given by

$$\begin{aligned}
f(\Theta_x^i | \boldsymbol{\Lambda} \backslash \{\Theta_x^i\}) &\propto f(\boldsymbol{D}|\boldsymbol{E}, \boldsymbol{\Theta}^{\text{pf}}, \boldsymbol{\Theta}^{\text{rest}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}) f(\boldsymbol{\Theta}^i | \rho_i, \sigma_i^2) \\
&\propto \prod_{t\in\mathcal{S}} \left(e^{-E_{t,x}^i \exp[\alpha_x + \beta_x \kappa_t]\Theta_x^i} (\Theta_x^i)^{D_{t,x}^i} \right)^{I_{t,x}^i} \\
&\quad \times f(\Theta_x^i | \boldsymbol{\Theta}_{-x}^i, \rho_i, \sigma_i^2).
\end{aligned} \tag{4.28}$$

In this last equation, we can simplify $f(\Theta_x^i | \boldsymbol{\Theta}_{-x}^i, \rho_i, \sigma_i^2)$ for different x:

- for $x = y_1$:

$$\begin{aligned} f(\Theta_x^i|\mathbf{\Theta}^i_{-x},\rho_i,\sigma_i^2) &\propto f(\Theta_x^i|\rho_i,\sigma_i^2)\cdot f(\Theta_{x+1}^i|\Theta_x^i,\rho_i,\sigma_i^2) \\ &= \frac{1}{\Theta_x^i\sqrt{2\pi\sigma_i^2}}\cdot\exp\left[-\frac{(\ln\Theta_x^i+\frac{1}{2}\sigma_i^2)^2}{2\sigma_i^2}\right] \\ &\quad\times\frac{1}{\Theta_{x+1}^i\sqrt{2\pi\sigma_i^2(1-\rho_i^2)}}\cdot\exp\left[-\frac{(\ln\Theta_{x+1}^i+\frac{1}{2}\sigma_i^2(1-\rho_i)-\rho_i\ln\Theta_x^i)^2}{2\sigma_i^2(1-\rho_i^2)}\right] \\ &\propto \frac{1}{\Theta_x^i\sqrt{2\pi\sigma_i^2(1-\rho_i^2)}}\cdot\exp\left[-\frac{1}{2\sigma_i^2(1-\rho_i^2)}\left(\ln\Theta_x^i+\frac{1}{2}\sigma_i^2-\rho_i(\ln\Theta_{x+1}^i+\frac{1}{2}\sigma_i^2)\right)^2\right] \\ &\sim \text{logN}\left(-\tfrac{1}{2}\sigma_i^2+\rho_i(\ln\Theta_{x+1}^i+\tfrac{1}{2}\sigma_i^2),\sigma_i^2(1-\rho_i^2)\right), \end{aligned}$$

- for $y_1 < x < y_Y$:

$$\begin{aligned} f(\Theta_x^i|\mathbf{\Theta}^i_{-x},\rho_i,\sigma_i^2) &\propto f(\Theta_x^i|\Theta_{x-1}^i,\rho_i,\sigma_i^2)\cdot f(\Theta_{x+1}^i|\Theta_x^i,\rho_i,\sigma_i^2) \\ &\sim \text{logN}\left(-\frac{1}{2}\sigma_i^2+\frac{\rho_i}{1+\rho_i^2}\left(\ln\Theta_{x-1}^i+\ln\Theta_{x+1}^i+\sigma_i^2\right),\sigma_i^2\frac{(1-\rho_i^2)}{(1+\rho_i^2)}\right), \end{aligned}$$

- for $x = y_Y$:

$$\begin{aligned} f(\Theta_x^i|\mathbf{\Theta}^i_{-x},\rho_i,\sigma_i^2) &\propto f(\Theta_x^i|\Theta_{x-1}^i,\rho_i,\sigma_i^2) \\ &\sim \text{logN}\left(-\tfrac{1}{2}\sigma_i^2+\rho_i(\ln\Theta_{x-1}^i+\tfrac{1}{2}\sigma_i^2),\sigma_i^2(1-\rho_i^2)\right). \end{aligned}$$

Given a current $\widetilde{\Theta}_x^i$ and Metropolis-Hastings sampling variance $s^2_{\Theta_x^i}$, we draw a proposal $\widehat{\Theta}_x^i$ from the distribution $\ln\widehat{\Theta}_x^i \sim \text{N}(\ln\widetilde{\Theta}_x^i - \frac{1}{2}s^2_{\Theta_x^i}, s^2_{\Theta_x^i})$. Using this proposal distribution ensures that $\mathbb{E}[\widehat{\Theta}_x^i] = \exp\left[\ln\widetilde{\Theta}_x^i - \frac{1}{2}s^2_{\Theta_x^i} + \frac{1}{2}s^2_{\Theta_x^i}\right] = \widetilde{\Theta}_x^i$. The proposal distribution is not symmetric and the acceptance probability is thus given by

$$\phi = \min\left\{\frac{f(\widehat{\Theta}_x^i|\mathbf{\Lambda}\backslash\{\widehat{\Theta}_x^i\})}{f(\widetilde{\Theta}_x^i|\mathbf{\Lambda}\backslash\{\widetilde{\Theta}_x^i\})}\cdot\frac{g(\widetilde{\Theta}_x^i|\widehat{\Theta}_x^i)}{g(\widehat{\Theta}_x^i|\widetilde{\Theta}_x^i)};1\right\}.$$

Here, $g(.|\Theta_x)$ is the lognormal density which gives the logarithm of the stochastic variable mean $\ln\Theta_x - \frac{1}{2}s^2_{\Theta_x}$ and variance $s^2_{\Theta_x}$.

Metropolis-Hastings sampling for ρ_i

The posterior distribution of ρ_i for $i \in \{\text{pf}, \text{rest}\}$ is given by

$$\begin{aligned} f(\rho_i|\mathbf{\Lambda}\backslash\{\rho_i\}) &\propto f(\mathbf{\Theta}^i|\sigma_i^2,\rho_i)\cdot f(\rho_i) \\ &= \frac{1}{(2\pi)^{Y/2}\Theta_{y_1}^i\cdots\Theta_{y_Y}^i\cdot|\mathbf{\Sigma}_i|^{1/2}} \end{aligned} \tag{4.29}$$

$$
\begin{aligned}
&\times \exp\left[-\frac{1}{2}(\ln \boldsymbol{\Theta}^i - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\ln \boldsymbol{\Theta}^i - \boldsymbol{\mu}_i)'\right] \\
&\times \frac{1}{\sqrt{2\pi\sigma^2_{\rho_i}}} \cdot \exp\left[-\frac{(\text{logit}\,(\rho_i) - \mu_{\rho_i})^2}{2\sigma^2_{\rho_i}}\right] \cdot \frac{1}{\rho_i(1-\rho_i)} \\
\propto\ & \frac{1}{\rho_i(1-\rho_i)(1-\rho_i^2)^{\frac{Y-1}{2}}} \\
&\times \exp\left[-\frac{a^i_\rho}{2\sigma_i^2(1-\rho_i^2)}\left(\rho_i - \frac{b^i_\rho}{a^i_\rho}\right)^2\right] \cdot \exp\left[-\frac{(\text{logit}\,(\rho_i) - \mu_{\rho_i})^2}{2\sigma^2_{\rho_i}}\right],
\end{aligned}
$$

with $a^i_\rho = \sum_{x=y_1+1}^{y_Y-1}(\Psi^i_x)^2$ and $b^i_\rho = \sum_{x=y_1+1}^{y_Y}\Psi^i_{x-1}\Psi^i_x$. This final expression will be used in the Metropolis-Hastings sampling algorithm. For a given current value $\tilde{\rho}_i$ and Metropolis-Hastings sampling variance $s^2_{\rho_i}$, we draw a proposal $\hat{\rho}_i$ from the distribution $\hat{\rho}_i \sim \text{TN}(\tilde{\rho}_i, s^2_{\rho_i}|0,1)$, with $\text{TN}(a, b|c, d)$ a truncated normal distribution with mean a, variance b, lower and upper bound c and d respectively. We use the truncated normal distribution to ensure the proposal is between 0 and 1. The proposal distribution is not symmetric and the acceptance probability is thus given by

$$
\phi = \min\left\{\frac{f(\hat{\rho}_i^2|\boldsymbol{\Lambda}\backslash\{\hat{\rho}_i^2\})}{f(\tilde{\rho}_i^2|\boldsymbol{\Lambda}\backslash\{\tilde{\rho}_i^2\})} \cdot \frac{g(\tilde{\rho}_i^2|\hat{\rho}_i^2)}{g(\hat{\rho}_i^2|\tilde{\rho}_i^2)}; 1\right\},
$$

where g is the density for the truncated normal distribution as described above.

Metropolis-Hastings sampling for σ_i^2

The posterior distribution of σ_i^2 is given by

$$
\begin{aligned}
f(\sigma_i^2|\boldsymbol{\Lambda}\backslash\{\sigma_i^2\}) &\propto f(\boldsymbol{\Theta}^i|\sigma_i^2, \rho_i) \cdot f(\sigma_i^2) \qquad (4.30)\\
&= \frac{1}{(2\pi)^{Y/2}\Theta^i_{y_1}\cdots\Theta^i_{y_Y}|\boldsymbol{\Sigma}_i|^{1/2}} \cdot \exp\left[-\tfrac{1}{2}(\ln \boldsymbol{\Theta}^i - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\ln \boldsymbol{\Theta}^i - \boldsymbol{\mu}_i)\right] \\
&\quad \times \sigma_i^{-1} \\
&\propto \frac{1}{\sigma_i^{Y+1}} \cdot \exp\left[-\sigma_i^{-2}\tfrac{1}{2}(\ln \boldsymbol{\Theta}^i - \boldsymbol{\mu}_i)'\boldsymbol{\Gamma}^{-1}(\rho_i)(\ln \boldsymbol{\Theta}^i - \boldsymbol{\mu}_i)\right].
\end{aligned}
$$

We use the final expression for the Metropolis-Hastings sampling algorithm. Given a current value $\tilde{\sigma}_i^2$ and Metropolis-Hastings sampling variance $s^2_{\sigma_i^2}$, we draw a new candidate $\hat{\sigma}_i^2$ from the proposal distribution $\ln\hat{\sigma}_i^2 \sim \text{N}(\ln\tilde{\sigma}_i^2 - \frac{1}{2}s^2_{\sigma_i^2}, s^2_{\sigma_i^2})$. The proposal distribution is not symmetric and the acceptance probability is thus given by

$$
\phi = \min\left\{\frac{f(\hat{\sigma}_i^2|\boldsymbol{\Lambda}\backslash\{\hat{\sigma}_i^2\})}{f(\tilde{\sigma}_i^2|\boldsymbol{\Lambda}\backslash\{\tilde{\sigma}_i^2\})} \cdot \frac{g(\tilde{\sigma}_i^2|\hat{\sigma}_i^2)}{g(\hat{\sigma}_i^2|\tilde{\sigma}_i^2)}; 1\right\}.
$$

Here, $g(.|\sigma_i^2)$ is the lognormal density which gives the logarithm of the stochastic variable mean $\ln\sigma_i^2 - \frac{1}{2}s^2_{\sigma_i^2}$ and variance $s^2_{\sigma_i^2}$.

Chapter 5

Unraveling relevant risk factors explaining pension fund mortality

This chapter is based on F. van Berkum, K. Antonio, and M. Vellekoop. Unraveling relevant risk factors explaining pension fund mortality: a case study in the Netherlands. *Working paper*, 2017b.

5.1 Introduction

People with different socioeconomic characteristics and different life styles experience a different level of mortality. Pension funds should account for this when valuing their liabilities, e.g. by distinguishing different risk profiles.

Brown and McDaid (2003) review 45 research papers that consider a variety of risk factors to explain differences in mortality. Relevant factors are (besides age and gender) education, income, tobacco and alcohol consumption, and marital status. They find that the reported impacts for the risk factors differ among the studies analyzed, and the estimated effects may differ between males and females or for different ages. Elo and Preston (1996) study adult US mortality in the period 1975-1985. They find that the probability of dying in a five-year period of individuals with many years of education (more than 16 years) is about 40% lower than mortality of individuals with no or a few years of education (less than 8 years), even after correcting for factors like income and race. Chetty et al. (2016) reveal a gap in the remaining period life expectancy at age 40 between the 1% richest and 1% poorest individuals in the US of about 14.6 years for males and 10.1 years for females, based on observations over the period 2001-2014.

Pension funds use adjustments on population mortality forecasts, which can only be based on information about participants that is directly available in their administration. Historically, crude methods such as age-shifting were used, which means that the probability q_x is replaced by q_{x+s} where s can be either positive or negative, see Pitacco et al. (2009). Such approaches are easy to implement, but potentially highly inaccurate.

In actuarial practice, observed mortality in the portfolio is sometimes expressed in terms of accrued rights. Mortality is then characterized by the fraction of the total accrued rights that has been released in a year, see e.g Plat (2009b). Using this alternative definition, the death rates (also referred to as *insured amount weighted death rates*) can be very volatile over the years due to individual mortality risk.[1] When defining a parametric regression model to explain mortality, it is difficult to appropriately account for this source of uncertainty, see van Berkum et al. (2017a).

By using this alternative definition, more weight is given to members with high accrued rights when explaining historical portfolio mortality. Since the liabilities of a pension fund are a function of accrued rights, this approach leads to more accurate predictions than when the accrued rights are neglected. However, if the level of mortality is dependent on the accrued rights, then it is also possible to include accrued rights (or some related risk factor) directly in a regression analysis. This way, the dependence is explicitly modeled, and it is possible to account appropriately for individual mortality risk, for example by using Poisson regression or survival analysis.

Richards et al. (2013) investigate the mortality in a German multi-employer pension scheme using survival models. The survival probabilities are modeled using a variant of the Makeham-Beard mortality law which includes a deterministic, linear mortality trend, and they use risk factors such as scheme size and the health status of members to explain remaining heterogeneity. They conclude that survival models based on individual observations are more flexible than models based on grouped observations, since they argue that those survival models can include a wider variety of risk factors. Gschlössl et al. (2011) use Poisson regression on observed deaths in a German insurance portfolio. Their dataset contains risk factors such as product type, policy duration and insured amount. First, they estimate a baseline mortality rate in the portfolio without time dynamics, then they explain remaining heterogeneity in the portfolio using categorical risk factors in a generalized linear model (GLM). When we compare the approach of Richards et al. (2013) to that of Gschlössl et al. (2011) we notice that survival models are not the only approach in which a wide variety of risk factors can be used to explain historical mortality.

Gschlössl et al. (2011) and Richards et al. (2013) have continuous risk factors available, but they discretize and group these variables to construct risk cells. As an example, Gschlössl et al. (2011) divide the continuous variable duration of the policy into the groups '0-2 years', '3-4 years' and '5+ years', and use this categorical risk factor in regression analysis instead of the continuous risk factor. The grouping of continuous risk factors often involves subjective decisions, and the outcomes are sensitive towards these choices.

With generalized additive models (GAMs) continuous variables can be included in

[1] Individual mortality risk is the uncertainty associated with the binary outcome of survival *given* a fixed mortality rate.

a more direct way. Semi-parametric smooth functions of the continuous variables are added to the linear predictor, and these are estimated, for example, using thin plate regression splines (Hastie and Tibshirani (1986); Wood (2006)). GAMs have been used successfully in the non-life actuarial field to predict claim frequencies for automobile insurance, see e.g. Denuit and Lang (2004), Klein et al. (2014) and Verbelen et al. (2016). We explore the use of this statistical modeling approach to construct better mortality risk profiles.

A different type of models that are often used to model pension fund mortality are multiple population mortality models. Villegas and Haberman (2014) investigate several multiple population extensions of the Lee and Carter (1992) model to predict mortality for five socioeconomic classes within England. These socioeconomic classes are based on characteristics such as income and education. The investigated specifications share a common mortality trend but differ in how mortality evolves within the socioeconomic classes relative to this common trend. Cairns et al. (2016) analyze mortality of Danish males for ten socioeconomic groups which are constructed using an affluence index which is based on wealth and reported income. They use an extension of the gravity model from Dowd et al. (2011) to model mortality in the different affluence groups, and they make the interesting observation that when they used affluence-based deciles they find greater levels of inequality than when they would use, for example, income quartiles.

However, both approaches are based on datasets which contain over 20 years of observations, and they consider separate time dynamics for each group. Such models are no longer feasible if the dataset has only a few years of data available. Moreover, socioeconomic classes are taken as exogenously given, while we would like to investigate which of the variables that can be combined in the definition of such classes are relevant and which are not.

We work on a unique dataset from a large Dutch pension fund covering the period 2006-2011. In this dataset individuals in the fund are followed over time, and their risk characteristics such as age, gender, salary, disability status and postal code are recorded on a yearly basis. This dataset contains too few years to estimate dynamics over time, but we will show how to construct mortality forecasts for individual risk profiles using this dataset. First, we estimate a multiple population mortality model (Koninklijk Actuarieel Genootschap (2014)) which provides us with an appropriate baseline mortality level and at the same time allows us to construct mortality forecasts for the Dutch population. Then, using the baseline, we explain remaining heterogeneity in the portfolio using Poisson regression, as suggested by Gschlössl et al. (2011). We consider a wide variety of possible risk factors (including continuous and spatial variables) in a GAM framework to ensure that all information in these variables is adequately captured in our model.

As a second contribution, we provide a complete framework on how to determine which risk factors should be used to explain historical portfolio mortality data. Our

final contribution is the design of a novel type of backtest to evaluate the performance of mortality models, which is driven by the impact of such models on the valuation of liabilities. As we will see later in this chapter, the (number of) risk factors used to explain portfolio mortality depends on the criteria used.

The remainder of this chapter is organized as follows. We discuss the dataset used in this chapter in Section 5.2. In Section 5.3 we introduce the generalized additive models that we will use to explain observed portfolio mortality, and we discuss how we assess in-sample and out-of-sample performance. We present estimation and backtesting results in Section 5.4 and we conclude in Section 5.5.

5.2 Data

We use a large dataset from a Dutch pension fund which follows individuals during the period 2006 to 2011. At the end of each year it is recorded whether a participant is still alive and the observable risk factors of the participant are updated. In our analysis we include information from active participants[2], pensioners, and people who are fully or partially disabled.

Table 5.1 lists the variables that we constructed from the dataset and Figure 5.1 shows the empirical distribution of these variables. We include the ages 20 to 90; we exclude lower ages because they are not relevant for the liabilities of the pension fund, and higher ages are excluded because their exposure turned out to be negligible in our dataset (less than 0.1% of the total exposure). The dataset contains 11,325,511 individual observations on 2,162,899 unique individuals resulting in a total of 11,304,448 years lived, and during the observed period 41,622 deaths were recorded (see Section 5.3.1 for a description of how we define the death and exposure observations).[3] When participants celebrate their birthday, the risk factor `Age` changes. We split the individual observations into observations before and after the birthday of the participants, which results in 22,632,277 observations with constant risk factors.

As discussed in the introduction, we expect that salary influences the level of mortality. There are several variables available in the dataset that can be used to include this effect: salary earned during a year, the fraction of the year the participant worked (hereafter: parttime factor), and the amount of salary that is earned from

[2]Inactive participants no longer work at a company linked to the pension fund, and these members therefore no longer pay premium and no longer accrue new pension rights.

[3]These numbers are in line with the assumption of a uniform distribution of deaths (UDD) (see Pitacco et al. (2009)). If the UDD assumption holds, then the members who died in the observed period will have contributed on average approximately half a year to the exposure in the year of their death. The total number of individual observations minus half of the total number of observed deaths is therefore approximately equal to the total exposure measured in years lived.

Table 5.1: A description of the mortality observations and risk factors. The percentage on the right shows for which fraction of the observations the information is available.

	Mortality information	
`D`	1 if the participant died at the current age, 0 if the participant survived	(100.0%)
`E`	The fraction of the year lived by the participant at the current age	(100.0%)
	Risk factors	
`Year`	Year of the observation	(100.0%)
`Age`	Age of the participant	(100.0%)
`Gender`	Gender of the participant	(100.0%)
`Sal`	Logarithm of FTE salary on annual basis, normalized per year per age per gender (if applicable: including an allowance for working at irregular hours)	(88.8%)
`IA`	The percentage of the FTE salary earned through an allowance for working at irregular hours	(44.6%)
`DisTime`	The cumulative disability spell of the participant, adjusted for partial disability	(96.0%)
`DisPerc`	The time spent in disability as a percentage of total service years registered	(96.0%)
`AFPP`	The age at which the participant received his first pension payment	(11.6%)
(`Long`, `Lat`)	The longitude and latitude that correspond to the center of the four-digit postal code of the participant	(99.9%)
`Edu`	Average education level at the postal code where the participant resides (obtained from Statistics Netherlands)	(95.3%)

working at irregular hours. We define full time equivalent salary as

$$\text{FTE salary} = \frac{\text{parttime salary earned in a year}}{\text{parttime factor}} + \text{irregularity allowance.}$$

Salaries tend to increase with age and as a result of inflation. Therefore, salaries at different ages and in different years cannot be compared directly, and we believe the full time equivalent salary should not directly be used as a risk factor. Instead, we construct the variable `Sal` which is a normalized version of the logarithm of the FTE salary earned. For each participant, we subtract the mean from the logarithm of the

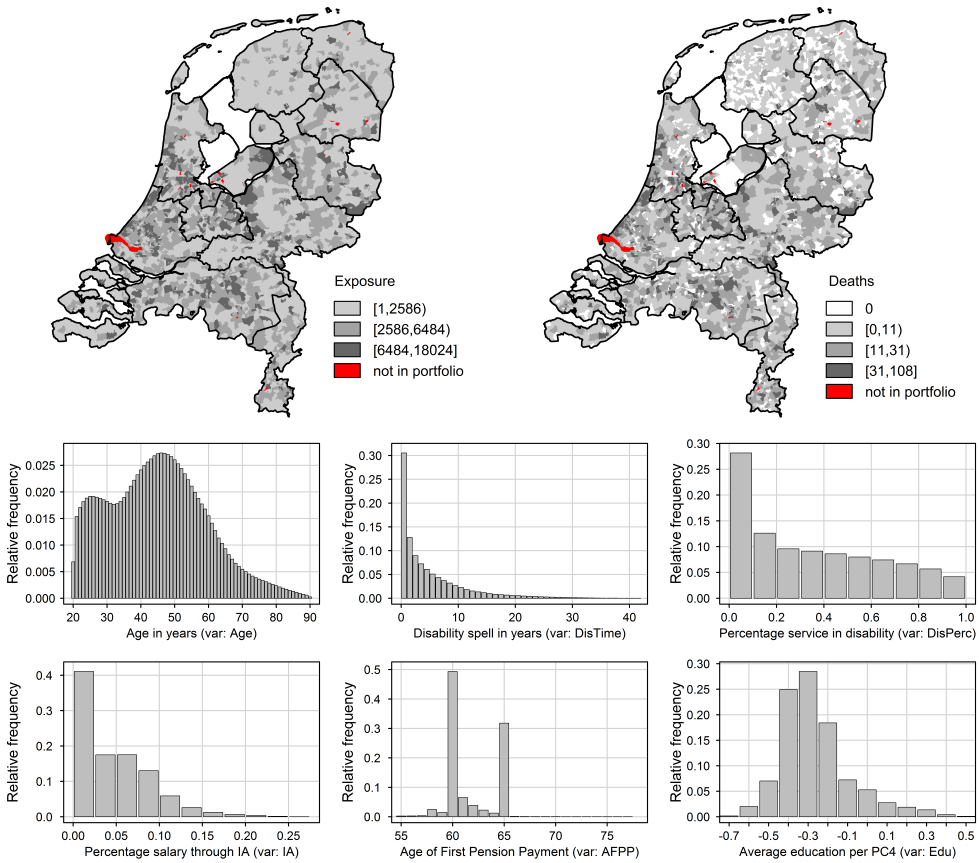


Figure 5.1: Empirical distribution of the observations and variables in the dataset. Missing records are excluded, and for `DisTime` and `DisPer` we also do not show observations equal to zero.

FTE salary and divide the result by its standard deviation, where mean and variance of the log transformed FTE salary are determined per year per age per gender. For most pensioners salary information is not available, and when this information is available for pensioners it corresponds to the latest salary earned as an active participant.

Participants with irregular working hours are more likely to have an irregular sleeping pattern, so participants who earn a larger fraction of their total salary through an irregularity allowance may have a higher mortality rate, see Costa (1996). We define the variable `IA` to include this effect, and it is computed as the irregularity

allowance divided by FTE salary.[4] This variable is available for participants with an irregularity allowance (44.6% of the exposure), and this variable is missing for other participants.

The dataset also contains information on disability. In the Netherlands, people can be classified as being partially disabled. Someone who has been partially disabled at 40% for three years will have a cumulative disability spell of 1.2 years, and the years of service at the company will have increased by 1.8 years. The variable `DisTime` represents the cumulative disability spell for a participant, and `DisPerc` represents the fraction of working years spent in disability. The latter is defined as the cumulative disability spell divided by the sum of the number of service years and the cumulative disability spell.

Both the number of service years and the cumulative disability spell relate only to the active period at the pension fund; if a participant was active at another pension fund before joining this pension fund, that information is not available in this dataset. For the active participants and the fully or partially disabled participants, we either have a positive disability spell (8.8% of the exposure) or a disability spell equal to zero (87.2%). For retired members, information on disability is often missing (4.0%). We will estimate different effects for these three groups (unknown disability status, disability spell or percentage equal to zero, and positive disability spell or percentage).

The official retirement age in the Netherlands during the observed period was 65 years, but many participants received pension payments before the age of 65. To stimulate the inflow of younger workers, many older workers close to the retirement age have had the opportunity to retire early (either fully or partially). We define the variable `AFPP` as the age at which the first pension payment is received, and we use this variable to investigate whether early retirement affects mortality.

In the Netherlands an address is completely specified by a six-digit postal code and a number that specifies the house. Our dataset contains for nearly all participants a four-digit postal code (`PC`), which corresponds to a district in a city. Using the longitude (`Long`) and latitude (`Lat`) of the center of the four-digit postal code we can estimate spatial heterogeneity in mortality rates.

Many studies have shown a link between education and mortality. Our dataset does not contain information on the education level of the participant, but from the website of Statistics Netherlands a dataset is available that contains information on education per four-digit postal code.[5] This dataset shows the fraction (i_c) of residents that has obtained education level c, where $c \in \{\text{Low}, \text{Medium}, \text{High}\}$. From this we construct the variable $\texttt{Edu} = -1 \cdot i_{\text{Low}} + 0 \cdot i_{\text{Medium}} + 1 \cdot i_{\text{High}}$, which is a weighted

[4] We divide by FTE salary because we believe that, for example, someone who works only one day a week at irregular hours should be treated differently from someone who works five days a week at irregular hours.

[5] https://www.cbs.nl/-/media/imported/documents/2013/49/131203-opleiding-regelingen-verdachten-pc4-mw.xls

average of the education attained by the residents in a postal code. The variable ranges from -1 (if every resident is classified as having Low education) to $+1$ (if every resident is classified as having High education). The information on education is not provided in full for all postal codes, and this variable is therefore not available for all postal codes.

Not all variables are applicable to, or available for, all participants. For example, participants who do not work at irregular hours do not have an irregularity allowance, and for those participants the variable is considered missing. Table 5.1 shows that some variables are available for (nearly) all participants (`Year`, `Age`, `Gender`, `Sal`, `DisTime`, `DisPerc`, `PC` and `Edu`), and some variables are rather scarce (`IA` and `AFPP`).

5.3 A framework for statistical modeling of portfolio mortality

In this section we introduce our framework for explaining observed portfolio mortality using risk factors. We specify the likelihood function for our model and discuss how we will optimize this likelihood. We also introduce criteria to assess the performance of estimated models, both in-sample and out-of-sample.

5.3.1 Distributional assumptions and model estimation

For integer t and x, we assume a constant force of mortality μ_{tx} on the interval $[t, t+1) \times [x, x+1)$, and we consider L_{tx} participants who are alive at the beginning of year t and for now, we assume they all have the exact age x. We will consider other ages in $[x, x+1)$ later. Using the method described in Section 2.1.1 we construct the death and exposure observations for all participants. For each participant j we define τ_{tjx} as the fraction of the year lived by participant j in calender year t at age x and define the corresponding indicator variable δ_{tjx} which equals 1 if the participant died at age x and 0 otherwise. The total number of observed deaths d_{tx} and the total exposure-to-risk E_{tx} in the pension fund are then given by

$$d_{tx} = \sum_{j=1}^{L_{tx}} \delta_{tjx} \qquad \text{and} \qquad E_{tx} = \sum_{j=1}^{L_{tx}} \tau_{tjx}. \tag{5.1}$$

From Section 2.1.2 we recall that the total survival likelihood for all participants is defined by

$$\prod_{j=1}^{L_{tx}} \exp[-\tau_{tjx}\mu_{tx}] \cdot (\mu_{tx})^{\delta_{tjx}} = \exp[-E_{tx}\mu_{tx}] \cdot (\mu_{tx})^{d_{tx}}, \tag{5.2}$$

where μ_{tx} is an unknown parameter that should be estimated. The right-hand side is proportional to a Poisson likelihood for μ_{tx}. Therefore, if we want to estimate μ_{tx} we can consider the model specification $D_{tx} \sim \text{Poisson}(E_{tx}\mu_{tx})$.

Likelihood of the model. We define the force of mortality for participant j at time t and age x by $\mu_{tjx} = \mu^{\text{pop}}_{tjx}\eta_{tjx}$. We assume that in calendar year t the baseline population force of mortality $\mu^{\text{pop}}_{tjx} = \mu^{\text{pop},g(j)}_{tx}$ is given for participant j aged x during calendar year t with gender $g(j) \in \{M, F\}$. The portfolio factor η_{tjx} represents the ratio between the population force of mortality and the force of mortality of participant j in calendar year t at age x, and this factor must be estimated from the data.

In the previous paragraph we were not explicit regarding the age of participants at the beginning of calendar year t since we assumed they were all born on January 1st. In the likelihood that we define below, we explicitly take into account that participant j has his birthday at $t + (1 - \iota_j)$. For the following derivation we define η_{tjx} as the portfolio factors for participant j in calendar year t with $x = x(j,t)$ before and $x = x(j,t)+1$ after his birthday in calendar year t, and these subscripts have the same meaning for τ_{tjx}, δ_{tjx} and μ^{pop}_{tjx}.[6] The likelihood then depends on all portfolio factors η_{tjx}, and using the yearly information from all individuals in the fund it is given by:

$$\mathcal{L} = \prod_{t=2006}^{2011} \prod_{j=1}^{L_t} \prod_{x=x(j,t)}^{x(j,t)+1} \exp[-\tau_{tjx}\mu^{\text{pop}}_{tjx}\eta_{tjx}](\mu^{\text{pop}}_{tjx}\eta_{tjx})^{\delta_{tjx}} \tag{5.3}$$

with L_t the number of participants in year t. In total, there are 11,325,511 observations from participants in the different years (summed over t and j), and the likelihood is built up from 22,632,277 contributions from participants in the portfolio (summed over t, j and x).

We use the subscript j to refer to participants and introduce a subscript i to refer to risk profiles, i.e. groups of individuals who share the same values of the risk factors. The likelihood in (5.3) is proportional to a Poisson likelihood, and we aggregate the observed death counts and exposures over all times t and over participants j with the same risk profile i:

$$\tilde{D}_i \sim \text{Poisson}(\tilde{E}_i \cdot \tilde{\eta}_i), \tag{5.4}$$

with $\tilde{D}_i$ the random death counts for risk profile i, and corresponding observed death

[6] Note that the factors $\eta_{t,j,x(j,t)}$ and $\eta_{t,j,x(j,t)+1}$ are the same if age is not included as an explanatory variable.

counts $\tilde{d}_i$ and exposures $\tilde{E}_i$ that are defined by:

$$\tilde{d}_i = \sum_{t=2006}^{2011} \sum_{j=1}^{L_t} \sum_{x=x(j,t)}^{x(j,t)+1} \delta_{tjx} \cdot I(i,t,j,x), \tag{5.5}$$

$$\tilde{E}_i = \sum_{t=2006}^{2011} \sum_{j=1}^{L_t} \sum_{x=x(j,t)}^{x(j,t)+1} \tau_{tjx} \mu^{\text{pop}}_{tjx} \cdot I(i,t,j,x), \tag{5.6}$$

with $I(i,t,j,x) = 1_{\{\eta_{tjx}=\tilde{\eta}_i\}}$ an indicator variable that is 1 if participant j aged x has risk profile i in calendar year t, and 0 otherwise. We use $\tilde{E}_i$ to denote exposure adjusted for population mortality. [7] With the above result we can estimate the portfolio factors $\tilde{\eta}_i$ using Poisson regression.

Estimation of the model. In practice, the population force of mortality μ^{pop}_{tjx} is unknown. Most pension funds and insurance companies in the Netherlands use mortality rates and forecasts published by the Royal Dutch Actuarial Society. For population mortality estimation and forecasting we use the model described in Koninklijk Actuarieel Genootschap (2014) (hereafter referred to as the AG model). The AG model is a variant of the Li and Lee (2005) model, which we use to estimate $\hat{\mu}^{\text{AG},g}_{t,x}$ on population mortality data from 1970-2011, where the starting year (1970) of the calibration period is in line with Koninklijk Actuarieel Genootschap (2014) and the end year (2011) is in line with the endpoint of our dataset. When estimating the model in (5.4), we use the fitted force of mortality from the AG model $\hat{\mu}^{\text{AG},g}_{t,x}$ for the unknown μ^{pop}_{tjx}. This way, we use an appropriate baseline mortality to explain historical portfolio mortality. If we are willing to assume that the portfolio factors do not evolve over time, we can use estimates for $\tilde{\eta}_i$ in combination with population mortality forecasts $\hat{\mu}^{\text{AG},g}_{t,x}$ for $t > 2011$ for pricing and valuation.[8] We use $\mu^{\text{pop}}_{tjx} = \hat{\mu}^{\text{AG},g(j)}_{t,x}$ and τ_{tjx} to construct the aggregated exposures $\tilde{E}_i$ using (5.6).

The factor $\tilde{\eta}_i$ is estimated using generalized additive models (GAMs), introduced by Hastie and Tibshirani (1986) and popularized by Wood (2006). We include p categorical variables x^d_{ik} such as gender $(k = 1, \ldots, p)$[9], q smooth functions $f_l(\cdot)$ of one-dimensional continuous variables x^c_{il} such as age or salary $(l = 1, \ldots, q)$, and a smooth function $g(\cdot,\cdot)$ of a two-dimensional variable $(x^{\text{long}}_i, x^{\text{lat}}_i)$ for postal code based longitude and latitude coordinates. The model is specified as $\tilde{D}_i \sim \text{Poisson}(\tilde{E}_i \cdot \tilde{\eta}_i)$

[7] This can also be interpreted as the expected number of deaths given the exposure of the considered risk cell within the pension fund and the population mortality rate.

[8] We describe the AG model in Appendix 5.A. In the appendix, the final year used for calibration is 2010, but that is specific to the financial backtest that is described there. For general model estimation we use the AG model calibrated on data until 2011.

[9] We use the superscript d to emphasize that we use dummy coding of the categorical variables.

for which the additive predictor is given by:

$$\ln(\mathbb{E}(\tilde{D}_i)) - \ln(\tilde{E}_i) = \ln \tilde{\eta}_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}^d + \sum_{l=1}^{q} f_l(x_{il}^c) + g(x_i^{\text{long}}, x_i^{\text{lat}}), \quad (5.7)$$

and we define $\ln \widetilde{\mathcal{L}}(\boldsymbol{\beta})$ as the corresponding log likelihood on aggregated data, where $\boldsymbol{\beta}$ represents the parameter vector of the unknown parameters β for the categorical variables and for the smooth functions of the continuous variables. We use thin plate regression splines to estimate the smooth functions f and g. This means that a function $f(x)$ is represented as $\sum_{m=1}^{M} \beta_m b_m(x)$ and a function $g(x, y)$ as $\sum_{n=1}^{N} \beta_n \tilde{b}_n(x, y)$, for fixed M and N and known basis functions $b_m(x)$ and $\tilde{b}_n(x, y)$. Through this representation the model is reduced to a GLM. To avoid overfitting a wiggliness penalty is added to the log likelihood, resulting in a penalized log likelihood. The wiggliness penalty is the product of the wiggliness of a function f or g and a corresponding smoothing parameter λ:

$$\begin{aligned} \ln \mathcal{L}^{\text{Pen.}} = \ln \widetilde{\mathcal{L}}(\boldsymbol{\beta}) + \sum_{l=1}^{q} \lambda_l \int [f_l''(x)]^2 dx \\ + \lambda_g \iint \left(\frac{\partial^2 g}{(\partial x^{\text{long}})^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x^{\text{long}} \partial x^{\text{lat}}} \right)^2 + \left(\frac{\partial^2 g}{(\partial x^{\text{lat}})^2} \right)^2 dx^{\text{long}} dx^{\text{lat}}. \end{aligned} \quad (5.8)$$

The unknown parameters β are estimated by optimizing the penalized log likelihood in (5.8). The generalized cross-validation (GCV) and the Akaike information criterion (AIC) are often used to select the smoothing parameters λ_l and λ_g. However, these methods are extremely sensitive to misspecification of the correlation structure in the error terms which may result in over- or underfitting of the data, see e.g. Krivobokova and Kauermann (2007) and Reiss and Ogden (2009). An alternative approach is to treat the smooth functions as random effects, which means that the λ_l can be estimated by maximum marginal likelihood or restricted maximum likelihood (REML), see Wood (2011). We explored both methods and found that REML results in more robust parameter estimates. Therefore, we use REML to select the smoothing parameters.

The estimation of a Poisson GAM is complicated by the presence of missing observations (see Table 5.1). If records for a variable are missing, the smooth effect for that variable is multiplied with a dummy variable that indicates whether the variable is available or not. For example, a model with `Age` and `Sal` is specified as

$$\ln(\mathbb{E}(D_i)) = \ln(\tilde{E}_i) + \beta_0 + f_{\text{Age}}(\texttt{Age}_i) + \text{I}[\texttt{Sal}_i \text{ is available}] \cdot f_{\text{Sal}}(\texttt{Sal}_i),$$

with $\text{I}[\texttt{Sal}_i \text{ is available}] = 1$ if salary is known for risk category i and 0 otherwise. In the remainder of this chapter we suppress this indicator function in our notation.

Identification problems exist for risk factors that are available for all participants. The smooth function for such a risk factor can be shifted with a constant c and

the intercept can be shifted by $-c$ without affecting the model fit. Therefore, for risk factors that are available for all participants the smooth functions are centered around zero. This constraint is not necessary to identify the smooth function for risk factors that are not available for all participants.

For the categorical variables we quantify the uncertainty in the estimates by constructing confidence intervals based on assumed large sample normality of these estimators. For the smooth components of the GAM we construct component-wise Bayesian confidence intervals (the confidence intervals include the uncertainty in the intercept, see Marra and Wood (2012)).

5.3.2 A strategy for working with large datasets

Even though GAM software exists for very large datasets (the function `bam`), it is impossible to estimate GAMs on datasets with millions of observations such as our dataset. Using (5.3)-(5.6) we therefore aggregate death counts and exposures for participants in the same risk cell, which decreases the number of observations. We discretize the continuous variables to facilitate model estimation. The variables related to time (`Year`, `Age`, `DisTime` and `AFPP`) are divided into 6, 71, 42 and 23 classes of width 1, the variables `Sal` and `Edu` are divided into 10 and 13 buckets of width 0.1, and the variables `DisPerc` and `IA` are divided into 10 and 11 buckets of width 10% and 2.5% respectively. For each bucket the middle value of that bucket is taken as the representative value.

As mentioned before, our dataset contains 22,632,277 observations with constant risk factors. After discretizing and aggregating where possible, and if we continue to consider `Year` as a risk factor the dataset has 16,914,481 different risk profiles. If we disregard `Year` as a risk factor, the number of risk profiles decreases to 8,428,419, and if we disregard postal code the number of risk profiles decreases even further to 852,382.[10] Since incorporating a spatial effect creates a number of risk profiles that is too high to process with the `gam` routine, we will cluster similar postal codes to reduce the number of risk categories even further. We will not cluster other variables such as age or salary, because these have fewer classes and the gain from clustering is smaller (and it is not necessary in this chapter).

We use the Fisher-Jenks method to cluster similar postal codes, which minimizes the variance within a cluster, see Fisher (1958) and Henckaerts et al. (2017). Suppose we have an estimated spatial effect $\hat{\eta}_i^{\text{pc}}$, and we wish to group these into m clusters. The numerical values of the estimated effect are then ordered, and for each admissible grouping the variance within a cluster is calculated. The optimal grouping is found using a dynamic programming approach.

[10] One might expect that disregarding `Year` would reduce the dataset by a factor of 6. However, for all participants at least one risk factor changes every year. Therefore, the reduction in the size of the dataset is less than 6.

Similar to (5.7) we define the effect for all risk factors excluding postal code $\tilde{\eta}_i^{-\text{pc}}$ and the effect for postal code $\tilde{\eta}_i^{\text{pc}}$ as

$$\tilde{\eta}_i^{-\text{pc}} = \exp[\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}^d + \sum_{k=1}^{q} f_k(x_{ik}^c)] \quad \text{and} \quad \tilde{\eta}_i^{\text{pc}} = \exp[g(x_i^{\text{long}}, x_i^{\text{lat}})], \tag{5.9}$$

so $\tilde{\eta}_i = \tilde{\eta}_i^{-\text{pc}} \cdot \tilde{\eta}_i^{\text{pc}}$. Then, we use the following steps to estimate $\tilde{\eta}_i$ when postal code is included as a risk factor in the model:

1. **Estimate the effects for all risk factors except postal code.** Since postal code is not included as a risk factor in this step, we can aggregate deaths and exposures over the different postal codes using (5.6), thereby decreasing the size of the dataset. From the model estimation we obtain an estimate for $\tilde{\eta}_i^{-\text{pc}}$;

2. **Estimate a spatial effect for postal code using the longitude and latitude information, taking the estimate of $\tilde{\eta}_i^{-\text{pc}}$ as given.** We aggregate deaths and exposures over all risk factors except postal code using a slightly adjusted version of (5.6), and the resulting dataset has only 4019 observations. Each observation corresponds to a four-digit postal code in the Netherlands for which there is positive exposure. Estimating the spatial effect on this aggregated dataset yields an estimate for $\tilde{\eta}_i^{\text{pc}}$;

3. **Cluster similar postal codes using the Fisher-Jenks method applied to the estimate of $\tilde{\eta}_i^{\text{pc}}$.** We consider $n \in \mathcal{N} = \{2, \ldots, 9\}$ clusters, and for each $n \in \mathcal{N}$ we determine the optimal clusters where a cluster is a group of similar postal codes, see Henckaerts et al. (2017). Then, we estimate a Poisson GLM using the categorical postal code with $n+1$ levels and with the estimate of $\tilde{\eta}_i^{-\text{pc}}$ given. We determine the optimal number of clusters n^* by comparing these GLMs using the Bayesian Information Criterion (Schwarz (1978)).

4. **Estimate a GAM including all risk factors and the clustered postal codes.** Note that we use n^*+1 levels for clustered postal codes, since the postal code is missing for some observations.

This approach greatly reduces the dimensionality of our parameter space. Instead of 4018 different postal codes, we now have n^* parameters for the clustered postal codes. We take the cluster with the lowest estimated effect as the reference cluster, and this is thus included in the intercept. We perform this procedure for all specifications considered, and different numbers of postal code clusters are allowed for different specifications.

5.3.3 Model assessment

First, we introduce several statistical measures to determine the added value of the inclusion of a risk factor when explaining observed mortality. However, for a pension fund it is more relevant to accurately predict the value of its future liabilities than to predict the future numbers of deaths. Therefore, we also introduce a novel back-test based on prediction of the value of the liabilities. This has, to the best of our knowledge, not been used before when testing models for portfolio-specific mortality.

For most model assessments discussed below, we use a point estimate for the portfolio factors $\tilde{\eta}_i$, but when we investigate the robustness of the portfolio factors, we will explicitly take uncertainty in the estimated effect of risk factors into account.

In-sample model fit. We investigate how well different models are able to explain the observations. When models are estimated on the complete dataset, we compute the log likelihood ($\ln \mathcal{L}$) on individual observations as defined in (5.3). Since we view the smooth effects as random effects, we calculate the conditional Akaike information criterion (cAIC) which is defined as

$$\text{cAIC} = -2 \ln \mathcal{L} + 2k. \tag{5.10}$$

If no smooth effect is included, k equals the number of parameters, and if smooth effects are included k equals the estimated degrees of freedom (EDF) adjusted for smoothing parameter uncertainty, see Wood et al. (2016) for details. We also calculate the Bayesian information criterion (BIC) defined as:

$$\text{BIC} = -2 \ln \mathcal{L} + k \ln n, \tag{5.11}$$

with n the number of observations included in the likelihood function (i.e. $n = 22,632,277$), and we use the same estimated degrees of freedom k as in the calculation of the cAIC.

Cross-validation statistics and robustness analysis. Czado et al. (2009) discuss different proper scoring rules that can be used to evaluate out-of-sample performance of models for count data. We focus on the log score which can be interpreted as an out-of-sample log likelihood statistic. Denote by $\mathcal{F}_{-t}$ all observations in the dataset excluding observations from year t. For all $t \in \mathcal{T} = \{2006, \ldots, 2011\}$ we estimate the model using $\mathcal{F}_{-t}$ which yields an estimate $\hat{\eta}_{j,x}^{-t}$ for participant j in year t at age x.

Similar to (2.6) and (2.7), define p_{tj} as the likelihood of observed death or survival for participant $j = 1, \ldots, L_t$ in calendar year t given the predictive distribution that follows from $\mathcal{F}_{-t}$. If τ_{tjx} represents the fraction of year t that participant j was alive

at age x (assuming he was alive at the start of that year), p_{tj} is computed as follows:

$$p_{tj} = \prod_{x=x(j,t)}^{x(j,t)+1} \exp[-\tau_{tjx}\mu_{tjx}^{\text{pop}}\hat{\eta}_{j,x}^{-t}] \left(\tau_{tjx}\mu_{tjx}^{\text{pop}}\hat{\eta}_{j,x}^{-t}\right)^{\delta_{tjx}}. \tag{5.12}$$

The log score for year t is then defined by:

$$\ln \mathrm{S}_t = -\frac{1}{L_t}\sum_{j=1}^{L_t} \ln p_{tj}, \tag{5.13}$$

with $\sum_{t=2006}^{2011} L_t = 11,325,511$, and we summarize this for the different years into the time-averaged value $\ln \mathrm{S} = \frac{1}{6}\sum_{t=2006}^{2011} \ln \mathrm{S}_t$.

We also investigate the robustness of estimated effects. For risk factor l, define $\hat{f}_l^{-t}(x_l)$ as the effect estimated using $\mathcal{F}_{-t}$. We compare the estimated effects $\hat{f}_l^{-t}(x_l)$ with the 80% confidence interval for $\hat{f}_l(x_l)$ estimated on the complete dataset. If the estimated effect is robust (i.e. consistent through time), the estimates $\hat{f}_l^{-t}$ are close to (and show a similar pattern as) $\hat{f}_l(x_l)$.

Predicted life expectancies. We further investigate the impact of different risk factors on remaining life expectancies. We compute remaining life expectancies for different ages and for males and females separately. This means we consider a risk profile as a combination of all risk factors except age and gender and it is therefore represented by k instead of i.[11] We define the remaining cohort life expectancy for risk profile k with age x at the beginning of calendar year t and for gender g as

$$LE_g(k,t,x) \approx \frac{1}{2} + \sum_{y=x+1}^{\infty} S_g(k,t,x,y) \tag{5.14}$$

with cumulative survival probabilities

$$S_g(k,t,x,y) = \exp\left[-\sum_{s=0}^{y-x+1} \hat{\mu}_{t+s,x(k)+s}^{\text{AG},g(k)} \cdot \hat{\eta}_k\right]. \tag{5.15}$$

Although `Age` and `Gender` are not included in k, these risk factors may be included in the regression model.

Financial backtest. For this backtest, we assume that the management of the pension fund at the beginning of the year 2011 wants to predict the value of the liabilities at the end of the year.[12] We use data from the years 2006-2010 as the training

[11] The difference between risk profile i and risk profile k is that the risk factors `Age` and `Gender` are fixed.

[12] For simplicity we ignore any cash flows during the year 2011, but this will not materially affect the results since the liabilities are mostly determined by cash flows further in the future. See Appendix 5.A for a detailed description of the financial backtest.

sample, and data from 2011 as the test sample. On January 1st 2011, we use all information available up to that point to approximate for each participant separately the value of the liabilities on December 31st 2011. Using the risk profile of individuals and the estimated effects from the training sample, we predict the probability that a participant survives the year 2011. The observations in the test sample tell us which participants died in the year 2011 and who survived until December 31st 2011. Using the test set we know the liabilities at the end of the year, and we can compare how well different model specifications are able to predict the liabilities in the test sample. This is a natural way to determine the added value of a risk factor when predicting the value of liabilities.

At the beginning of 2011, 5.7% of the participants was retired, and for 91.1% of the participants the accrued rights are available in the dataset. For all participants with unknown accrued rights we assume a final-pay plan with 1.75% accrual per service year to approximate these rights. We denote the yearly benefit for participant j by b_j, and we assume this benefit remains constant over time. This amount is paid in the middle of year t if $x(j,t) \geq x_r$, with $x(j,t)$ the age of participant j at the beginning of year t and x_r the retirement age. In 2011 the retirement age was 65 years in the Netherlands, so we use $x_r = 65$. For participant j we compute the expected present value of an annuity a_j that pays 1 euro halfway during the year if the participant is still alive at that time and $x(j,t) \geq x_r$. See Appendix 5.A for details.

Define $\hat{\eta}^{-2011}_{j,x}$ as the fitted portfolio factor for participant j of age x during the year 2011. This portfolio factor is estimated using the training sample. We model the uncertainty of participant j surviving the year 2011 using a Bernoulli$(p_{2011,j})$ distributed random variable $Y_{2011,j}$, with $P(Y_{2011,j} = 1) = p_{2011,j} = 1 - q_{2011,j} = \exp[-\mu_{2011,j}]$ which are independent for different j's. Not all participants are born on January 1st which we account for when computing the one-year survival probabilities. Under the assumption of a constant force of mortality $\mu_{t,x}$ on the interval $[t, t+1) \times [x, x+1)$, the one-year survival probability for participant j with his birthday at $t + (1 - \iota_j)$ is given by:

$$\begin{aligned}
\exp[-\mu_{2011,j}] &= \exp\left[- \int_0^{1-\iota_j} \hat{\mu}^{\mathrm{AG},g(j)}_{2011,x(j,2011)} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)} dt \right. \\
&\qquad \left. - \int_{1-\iota_j}^{1} \hat{\mu}^{\mathrm{AG},g(j)}_{2011,x(j,2011)+1} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)+1} dt \right] \\
&= \exp\left[-(1-\iota_j) \cdot \hat{\mu}^{\mathrm{AG},g(j)}_{2011,x(j,2011)} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)} \right. \\
&\qquad \left. - \iota_j \cdot \hat{\mu}^{\mathrm{AG},g(j)}_{2011,x(j,2011)+1} \cdot \hat{\eta}^{-2011}_{j,x(j,2011)+1} \right]. \qquad (5.16)
\end{aligned}$$

The stochastic value of the liabilities Γ on December 31st 2011 is then given by

$$\Gamma = \sum_{j=1}^{L_{2011}} \left(Y_{2011,j} \cdot b_j a_j + (1 - Y_{2011,j}) \cdot 0\right), \tag{5.17}$$

with L_{2011} the number of members alive at the beginning of the year 2011. Given this specification and using only the point estimate $\hat{\eta}_{j,x}^{-2011}$, the expected value and variance of the random liabilities Γ are given by

$$\mathbb{E}(\Gamma|\hat{\eta}_{j,x}^{-2011}) = \sum_{j=1}^{L_{2011}} p_{2011,j} \cdot b_j a_j \tag{5.18}$$

$$\mathrm{Var}(\Gamma|\hat{\eta}_{j,x}^{-2011}) = \sum_{j=1}^{L_{2011}} (b_j a_j)^2 \cdot p_{2011,j} \cdot (1 - p_{2011,j}). \tag{5.19}$$

The latter expression shows that the uncertainty in the evolution in the liabilities over a one-year horizon is mainly caused by participants that have large annuities or large accrued rights, and participants that have a moderate probability of dying (i.e. $q_{2011,j}$ not too close to zero or one).

We define the indicator variable $\mathrm{I}_{2011,j}$ that is 1 if participant j was still alive at December 31st 2011 and zero otherwise. The actual liabilities per December 31st 2011 are then calculated as:

$$\tilde{\Gamma} = \sum_{j=1}^{L_{2011}} \mathrm{I}_{2011,j} \cdot b_j a_j. \tag{5.20}$$

Assuming normality for the liabilities at an aggregate level, we construct the 90% prediction interval for Γ with mean and variance as defined in (5.18) and (5.19). If a model is able to accurately predict the evolution in the liabilities over a one-year horizon, the actual liabilities $\tilde{\Gamma}$ will often lie within the prediction interval for Γ.

Within a predictive distribution for the liabilities, underestimations and overestimations may cancel out. Therefore, we also calculate the mean squared prediction error (MSPE) as:

$$\begin{aligned}\mathrm{MSPE} &= \sum_{j=1}^{L_{2011}} (\mathrm{I}_{2011,j} \cdot b_j a_j - p_{2011,j} \cdot b_j a_j)^2 \Bigg/ \sum_{j=1}^{L_{2011}} (b_j a_j)^2 \\ &= \sum_{j=1}^{L_{2011}} \underbrace{(b_j a_j)^2}_{\text{'weights'}} \underbrace{(\mathrm{I}_{2011,j} - p_{2011,j})^2}_{\text{'errors'}} \Bigg/ \underbrace{\sum_{j=1}^{L_{2011}} (b_j a_j)^2}_{\text{normalizing constant}}.\end{aligned} \tag{5.21}$$

Through this definition of the MSPE, the prediction error on a participants survival is weighted by the value of the liabilities needed for that participant. This way we ensure that participants that contribute more to the liabilities of a pension fund are given more weight in this statistic.

5.4 Results

We estimate portfolio factors using the procedure described in Sections 5.3.1 and 5.3.2, and we define two reference specifications:

- Portfolio mortality equals population mortality ($\eta_i = 1$ for all i);
- The relative difference with population mortality is the same for all participants, i.e. the regression model in (5.7) only includes the constant β_0 ($\eta_i = \exp(\beta_0)$ for all i).

These two reference models allow us to quantify the relative importance of including different risk factors in explaining historical portfolio mortality. Besides these reference models, we consider all single variable specifications and a selection of multiple variable specifications. The variables that we include as explanatory variables are `Gender`, `Age`, `DisTime`, `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and (`Long`,`Lat`). The variable `Year` is not included, because the dataset covers only a few years. In this chapter, our focus is on investigating which in-sample statistic our out-of-sample test leads to optimal risk factor selection. For exploratory purposes, we therefore consider single effects only. The analysis can be extended to include interactions between effects such as gender-specific salary effects. Interactions between continuous variables are also possible, but this greatly increases the number of possible specifications and increases the computational cost.

5.4.1 Estimation results

Table 5.2 shows the model fit (the log likelihood, estimated degrees of freedom, conditional AIC and BIC) and cross validation statistics (the time-averaged log score) for a selection of single and multiple variable specifications.

Model fit and estimated factors. The two reference specifications are shown in the top two rows of Table 5.2, and the model fit improves considerably if we allow for differences between mortality in the population and the portfolio. The cAIC and BIC almost always improve if we add a single variable on top of the constant, but the BIC does not improve if we add `Gender`. Note that gender is already included in the baseline mortality level $\mu_{t,x}^{\text{pop},g}$. It is therefore no surprise that `Gender` has little additional explanatory power. The results for cAIC and BIC are similar, but as is to be expected, there are some cases in which the information criteria suggest different model specifications.

Large improvements in model fit come from adding `DisTime`, `DisPerc` or `Sal`. With `Salary` and `DisPerc` included in the model (for example row 15 in Table 5.2), the information criteria and score function improve considerably compared to the

Table 5.2: Estimation results for a selection of models. For $\ln \mathcal{L}$ larger is better, for the other statistics smaller is better. The horizontal lines separate models with increasing numbers of risk factors. Note that all regression models (row 2 and below) include a constant.

	Model	$\ln \mathcal{L}$	EDF	cAIC	BIC	ln S ($\times 10^{-3}$)
1.	Baseline	-222,772	0.0	445,545	445,545	19.66795
2.	Constant	-222,089	1.0	444,180	444,195	19.60966
3.	Gender	-222,083	2.0	444,170	444,199	19.60912
4.	Age	-221,993	2.6	443,992	444,031	19.60180
5.	DisTime	-220,482	7.0	440,978	441,083	19.46849
6.	DisPerc	-220,501	6.6	441,016	441,114	19.47032
7.	Sal	-221,037	5.3	442,085	442,163	19.51721
8.	IA	-221,962	3.2	443,931	443,979	19.59840
9.	AFPP	-221,977	3.1	443,960	444,007	19.59983
10.	Edu	-222,017	6.9	444,047	444,150	19.60374
11.	PC	-222,026	7.0	444,065	444,170	19.60752
12.	DisPerc-DisTime	-220,480	9.3	440,978	441,117	19.46892
13.	DisPerc-Gender	-220,476	7.6	440,968	441,081	19.46815
14.	DisPerc-Age	-220,442	11.7	440,908	441,084	19.46621
15.	DisPerc-Sal	-219,853	10.9	439,728	439,891	19.41381
16.	DisPerc-IA	-220,477	8.7	440,970	441,100	19.46821
17.	DisPerc-AFPP	-220,416	8.7	440,850	440,981	19.46306
18.	DisPerc-Edu	-220,444	12.0	440,912	441,092	19.46569
19.	DisPerc-PC	-220,446	11.6	440,916	441,089	19.46841
20.	DisPerc-Sal-Gender	-219,834	11.8	439,691	439,868	19.41212
21.	DisPerc-Sal-Age	-219,808	16.2	439,649	439,891	19.41100
22.	DisPerc-Sal-IA	-219,793	13.3	439,613	439,811	19.40865
23.	DisPerc-Sal-AFPP	-219,786	13.6	439,598	439,802	19.40814
24.	DisPerc-Sal-Edu	-219,815	16.2	439,663	439,905	19.41084
25.	DisPerc-Sal-PC	-219,808	15.9	439,648	439,886	19.41212
26.	DisPerc-Sal-IA-AFPP	-219,726	15.9	439,484	439,721	19.40305
27.	DisPerc-Sal-IA-Edu	-219,756	18.5	439,549	439,826	19.40577
28.	DisPerc-Sal-IA-PC	-219,759	16.2	439,550	439,792	19.40732
29.	DisPerc-Sal-IA-AFPP-Edu-PC	-219,661	24.2	439,370	439,731	19.39947

first two rows. If we keep adding variables, the statistics improve further, but the improvements are much smaller than those from adding `Salary` and `DisPerc`.

Figures 5.2a-5.2h show the estimated effects for the models with a single variable, i.e. for the models in row 3-10 of Table 5.2. The gray area in the graphs represents the 80% confidence interval based on the calibration using the complete dataset. The colored lines represent the estimated effect if a single year is left out of the dataset.

The shape of the estimated effects for the variables `Sal` and `IA` are in line with intuition: higher salary leads to lower mortality, and more hours worked at irregular times leads to higher mortality, and the estimated effects for these single risk factor models are strong. The maximal relative difference in the force of mortality for participants with high and low salary is about $\exp[0.2-(-0.6)] \approx 2.2$, and the effect for `IA` is only slightly lower. For `AFPP` most of the exposure is located at `AFPP`=60 and `AFPP`=65, and from Figure 5.2g we derive that retiring at the age of 60 (early retirement) reduces the force of mortality by about 18% compared with retiring at 65 (the official retirement age).

In Figure 5.2e and 5.2f we consider $\texttt{Dis} \in \{\texttt{DisPerc}, \texttt{DisTime}\}$. Participants with a disability spell equal to zero ($\texttt{Dis} = 0$) are included in the reference group, for participants with missing disability information (`Dis` is missing) we include a dummy variable, and for participants with a non-zero disability spell ($\texttt{Dis} > 0$) a smooth effect is included.[13] The effect estimated for the dummy variable and the smooth effect are added to the intercept for the respective groups, and the portfolio factor η_i when only `Dis` is included in the model is thus represented as:

$$\ln \tilde{\eta}_i = \beta_0 + \mathrm{I}[\texttt{Dis is missing}] \cdot \beta_1 + \mathrm{I}[\texttt{Dis} > 0] \cdot f_{\mathrm{Dis}}(\texttt{Dis}_i).$$

The parameters β_0 and β_1 in these models are estimated at -0.42 and 0.28 respectively.[14] The force of mortality for participants with missing disability information is therefore roughly $\exp[0.28]-1 = 33\%$ higher compared to participants with `Dis` equal to zero. This can be explained by the fact that participants with missing disability information are a mixture of participants with `Dis` equal to zero (low mortality) and participants with positive `Dis` (high mortality). The estimated smooth effects for `DisPerc` and `DisTime` are shown in Figure 5.2e and 5.2f, and those effects are on average approximately 0.7. For participants with `Dis` greater than zero the force of mortality is almost twice as large compared to the force of mortality for participants with `Dis` equal to zero.

The estimated effect for `DisPerc` shown in Figure 5.2e is difficult to explain. In 2005 the Dutch government introduced new legislation regarding income protection

[13] If a participant has a cumulative disability spell equal to zero, then both `DisTime` and `DisPerc` are zero. The same principle holds for participants with missing disability information or with a positive disability spell.

[14] These numbers are identical for the models with either `DisPerc` or `DisTime`.

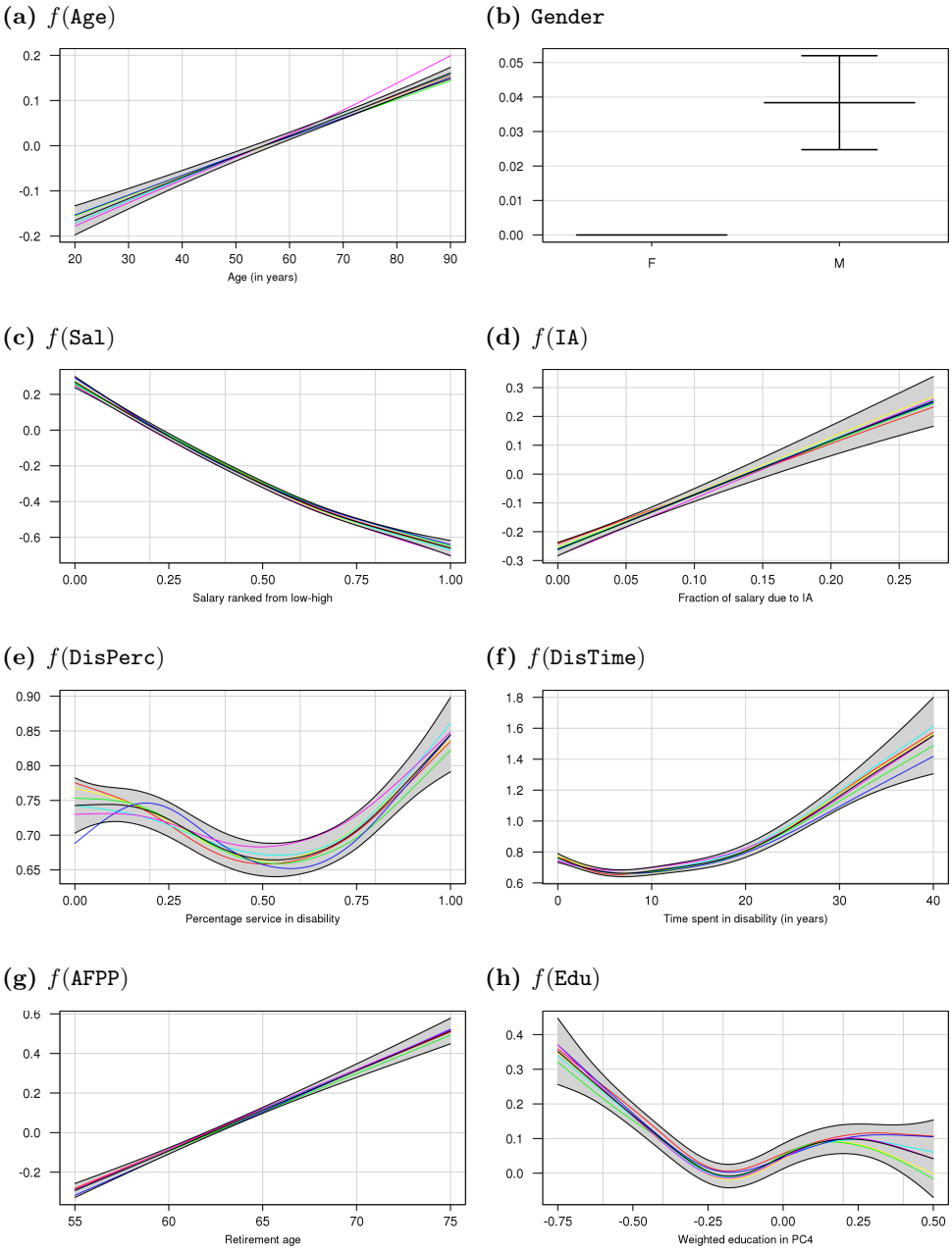


Figure 5.2: Estimated factors for the models shown in row 3-10 in Table 5.2. The gray area bounded by the black lines represents the 80% confidence interval for the effect estimated on the complete dataset, and the colored lines represent the estimated effect if a single observation year is excluded.

provided by the state. Under this new legislation people can be classified as being partially disabled, and for this group a suitable replacement job is searched. Three main classes are distinguished: less than 35% disabled, between 35% and 80% disabled, and above 80% disabled, and the state pension and effort put into searching a suitable replacement job depends on the disability class. `DisPerc` is a variable aggregated over time and can thus not be directly compared with the disability classes distinguished in the legislation. The different treatment of these three disability classes may be the cause for the curvature in the estimated effect for `DisPerc`.

From Figure 5.2f we observe that mortality increases with the length of a disability spell, except for the first few years. The exposure for small values of `DisTime` is large, but also potentially heterogeneous. Participants with a disability spell of one year may have been disabled for one year and lived healthily for many years since then, or they may be in the first year of a disability spell that will eventually become very long.

Models with `Salary` and `DisTime` perform similarly to models with `Salary` and `DisPerc`, but we prefer to use `DisPerc` over `DisTime` since in the dataset `DisPerc` is more uniformly spread over different classes, see Figure 5.1. Also, from the single factor analysis we observe that the confidence interval widens for larger values of `DisTime`, but remains steady for all values of `DisPerc`, see Figures 5.2e and 5.2f.

The model with the risk factors `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC` performs well on the model assessment criteria we introduced in Section 5.3.3 (see row 29 in Table 5.2 for those results). Figure 5.3 shows the estimated effects for this model. In the middle row of Figure 5.3, we observe that the effects of `DisPerc`, `Sal`, `IA` and `Edu` are similar to the effects of factors shown in Figure 5.2. The effects of `Sal` and `IA` are nearly identical, which suggests that the effects of these variables are independent of the effects of the other variables included. It is remarkable that the effect for `AFPP` is reversed. In Figure 5.2g the estimated effect indicates that early retirement decreases the level of mortality, but Figure 5.3 indicates otherwise. Participants may have different motivations for retiring early. Wealthy participants may have chosen to retire early because they no longer *needed* to work, whereas participants with bad health may have retired early because they *could* no longer work. For the first group we expect lower mortality and for the second group we expect higher mortality. In a single factor model, `AFPP` tries to capture these two effects, while in a model with multiple risk factors this effect may be captured, for example, by `Sal` and `DisPerc`.

The top row of Figure 5.3 shows (from left to right) the estimated postcode effect using longitude and latitude, the significance of the estimated spatial effect, and the clusters created by the Fisher-Jenks clustering method (see Section 5.3.2). There are clear regions with higher mortality and regions with lower mortality, and the effect is significantly different from zero at many regions using a 80% confidence level. The resulting clustering of postcodes shows great similarities with the estimated spatial effect, but each postcode within a cluster now has the same impact on the predictor.

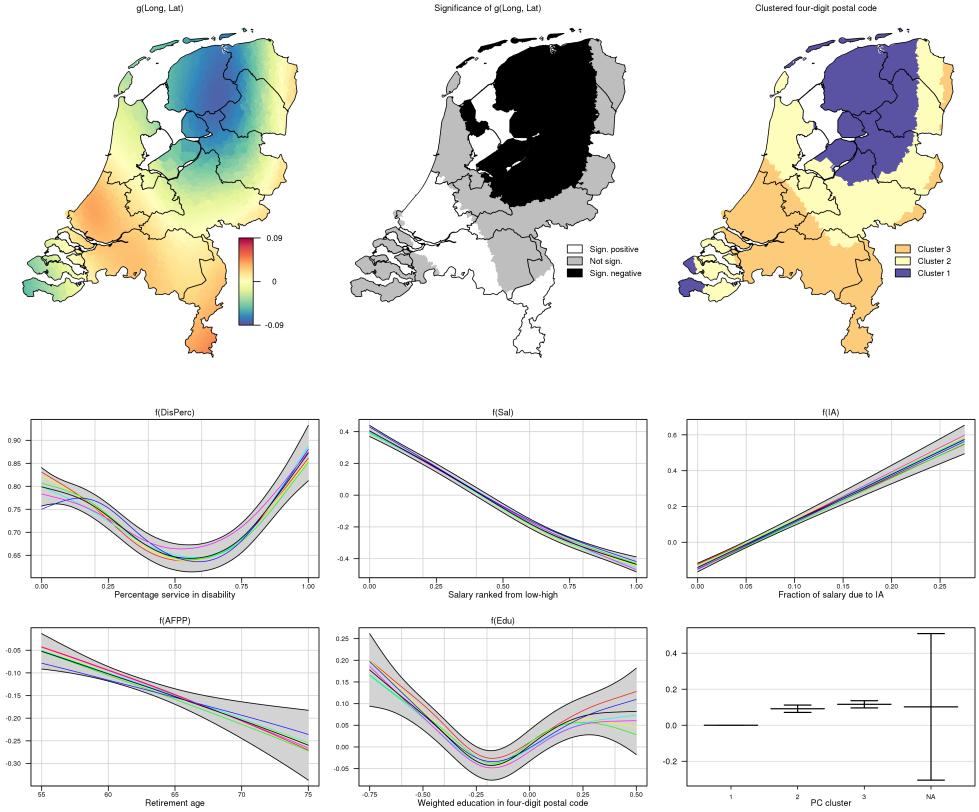


Figure 5.3: Estimated effects for the model with `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC` included (row 29 in Table 5.2). The top row shows the estimated spatial effect from Step 2 as described in Section 5.3.2, the significance of that spatial effect, and the resulting clusters after the Fisher-Jenks method has been applied. The graph at the bottom right shows the estimated effects for the corresponding clusters in the top-right graph. The remaining graphs show the estimated smooth effects for `DisPerc`, `Sal`, `IA`, `AFPP` and `Edu`; the gray area bounded by the black lines represents the 80% confidence interval for the effects estimated on the complete dataset, and the colored lines represent the estimated effects if a single observation year is excluded.

Cross validation and robustness analysis. The last column in Table 5.2 shows the cross validation statistics (smaller ln S is better). There is no penalty on the scoring rule for adding parameters. Therefore, if a variable captures an effect that is consistent through time, including that variable will lead to a more accurate predictive distribution and thus to improved results for the scoring rule, while including variables that represent uninformative noise may deteriorate those results. We see that results for the scoring rule improve if we add variables, so apparently all variables capture effects that are consistent through time.

From Figure 5.2 we observe that the colored lines show a similar pattern as the gray areas. This means that the estimated effects for the risk factors are stable over time. Even in Figure 5.3, where we included many risk factors in a single model, the estimated effects are stable over time. This is not surprising, since we have many observations.

5.4.2 Predicted cohort life expectancies

The model with the risk factors `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC` performs well both in-sample and out-of-sample, and the estimated effects for this model are robust. Using the estimated effects from that model as illustrated in Figure 5.3 and (5.14), we compute remaining cohort life expectancies (LE) for different risk profiles which are shown in Table 5.3. This allows us to quantify the impact of different variables on LE, and to compare our results with existing literature. Note that we do not distinguish between different levels of `IA`, `AFPP` and `Edu` because doing so would complicate providing a clear presentation of the results. These variables are therefore assumed missing in calculating the remaining life expectancies. The most-favorable group in this table is represented by (`PC_Group`=1, `DisPerc`="No", `Sal`=0.90) and the least-favorable group by (`PC_Group`=3, `DisPerc`=5%, `Sal`=0.10).

The results are striking. For the remaining life expectancy at age 25, the difference between low and high salary risk profiles is between 5 and 6 years, and the difference is of similar size for being/having been disabled or not. Chetty et al. (2016) found a difference of 14.6 years between the 1% richest and 1% poorest, but their dataset is not restricted to people in pension funds (they consider all individuals with positive household earnings in the USA). The USA is a more heterogeneous group than the participants within the Dutch pension fund, and we use different quantiles for salary, so the different scale need not surprise us. Further, even the effect of postal code on life expectancy is not negligible since it amounts to approximately one year, and this result is in line with Figure 5 from Chetty et al. (2016).

Cairns et al. (2016) compute the partial period life expectancy from age 55 to 90 for different affluence groups. These affluence groups are defined using a combination of reported wealth and income. They find that the difference in partial period life expectancy between the lowest and highest affluence groups is about 6.5 years, which

Table 5.3: Predicted remaining cohort life expectancies for different risk profiles. The model used for estimating portfolio factors includes the variables `DisPerc`, `Sal`, `IA`, `AFPP`, `Edu` and `PC_Group`, and remaining life expectancies are computed in the year 2012 for males (M) and females (F) at the age of 25 and 65. To limit the dimensions of the risk profiles, we have assumed `IA`, `AFPP` and `Edu` missing in all risk profiles.

PC_Group	DisPerc	Sal	$LE_M(25)$	$LE_F(25)$	$LE_M(65)$	$LE_F(65)$
Lowest mortality (cluster 1)	No	0.90	69.1	71.2	26.2	28.8
		0.50	66.7	68.9	23.7	26.5
		0.10	63.5	66.0	20.6	23.6
	5%	0.90	62.7	65.2	19.9	22.9
		0.50	60.2	62.8	17.6	20.7
		0.10	56.7	59.5	14.8	18.0
Highest mortality (cluster 3)	No	0.90	68.2	70.3	25.2	27.9
		0.50	65.8	68.1	22.8	25.6
		0.10	62.6	65.1	19.7	22.7
	5%	0.90	61.8	64.3	19.0	22.0
		0.50	59.2	61.9	16.7	19.9
		0.10	55.6	58.5	14.0	17.2
Baseline in:	general population		62.5	65.1	19.6	22.7
	portfolio		64.0	66.4	21.0	24.0

is also in line with our results. Further, RIVM (2014) reports differences up to six years in period life expectancy at birth between low and highly educated people. This range is similar to what we observe for low and high salaries.

The difference in remaining life expectancy for $x = 25$ between the most-favorable group in this table and the least-favorable group is 13.5 and 12.7 for males and females respectively. These differences persist over time: at age 65 the differences are 12.2 and 11.6 for males and females respectively. This means that if participants were to retire at age 65, males in the most-favorable group benefit 26.2 years from their pension which is almost twice as long as males in the least-favorable group, who benefit 14.0 years on average. We have not taken into account the effect from all risk factors that can be included in the model, so for more granular risk profiles the differences may be even larger.

In this analysis we have assumed that the variables `Sal`, `DisPerc` and `PC_Group` remain constant throughout the entire lifetime of a participant. Although this is a strong assumption, it is the best assumption we can make based on the available data.

5.4.3 Financial backtest

Figure 5.4 shows the results of the financial backtest. For each model we computed the mean and variance of the value of the liabilities (predicted at January 1st 2011) using (5.18) and (5.19). The mean is represented by the large dot, the horizontal lines represent the 90% prediction intervals under the assumption of normality for the value of the liabilities. The actual values observed at the end of the year (i.e. the target for our predictions at the beginning of the year) are represented by the vertical dashed line. The mean squared prediction error (MSPE) as defined in (5.21) is shown on the right-hand side of the figure.

If we use portfolio factors equal to one (i.e. portfolio mortality is the same as population mortality), we underestimate the liabilities; the actual liabilities are far outside the prediction interval. Predictions improve if we take portfolio mortality into account, so this clearly shows the usefulness of adjustments to population mortality rates. However, the liabilities are still underestimated. It is surprising that if we include the variables `Age`, `Gender`, `DisTime`, `DisPerc`, `AFPP`, `PC` or `IA`, the predicted liabilities are very similar to the predicted liabilities when only an intercept, and thus a fixed correction to population mortality, is included. Note that this does not imply that including these risk factors leads to the same predicted liabilities at individual level.

However, if we include `Sal` as an explanatory variable the prediction of the liabilities improves significantly, and the actual liabilities lie within the prediction interval. Since accrued benefits are correlated with salary, this was to be expected. This is also the reason why practitioners in the actuarial field tend to work with mortality rates weighted by insured amounts, as in Plat (2009b).

The model estimation results in Table 5.2 suggest to use a model which includes `Sal` and either `DisTime` or `DisPerc`, and possibly `IA`, `Edu` and `PC` as additional explanatory variables. However, in Figure 5.4 we see that including information on disability when `Sal` is already taken into account decreases the predicted value of liabilities and the backtesting results worsen (compared to the model where only `Sal` is included). This is further investigated in Figure 5.5 where we show the predicted value of liabilities for specifications with `DisPerc` and/or `Sal` included, but now separately for participants with known and unknown disability and/or salary information:

- If only `DisPerc` is included, liabilities are underestimated for participants without a positive disability spell, because the correlation between mortality and accrued rights is not taken into account. The actual liabilities fall within the prediction interval for participants with a positive disability spell;

- If only `Sal` is included, predicted liabilities increase for participants with known salary information compared to the predicted liabilities for the other single factor regression models. For participants without positive disability spell the

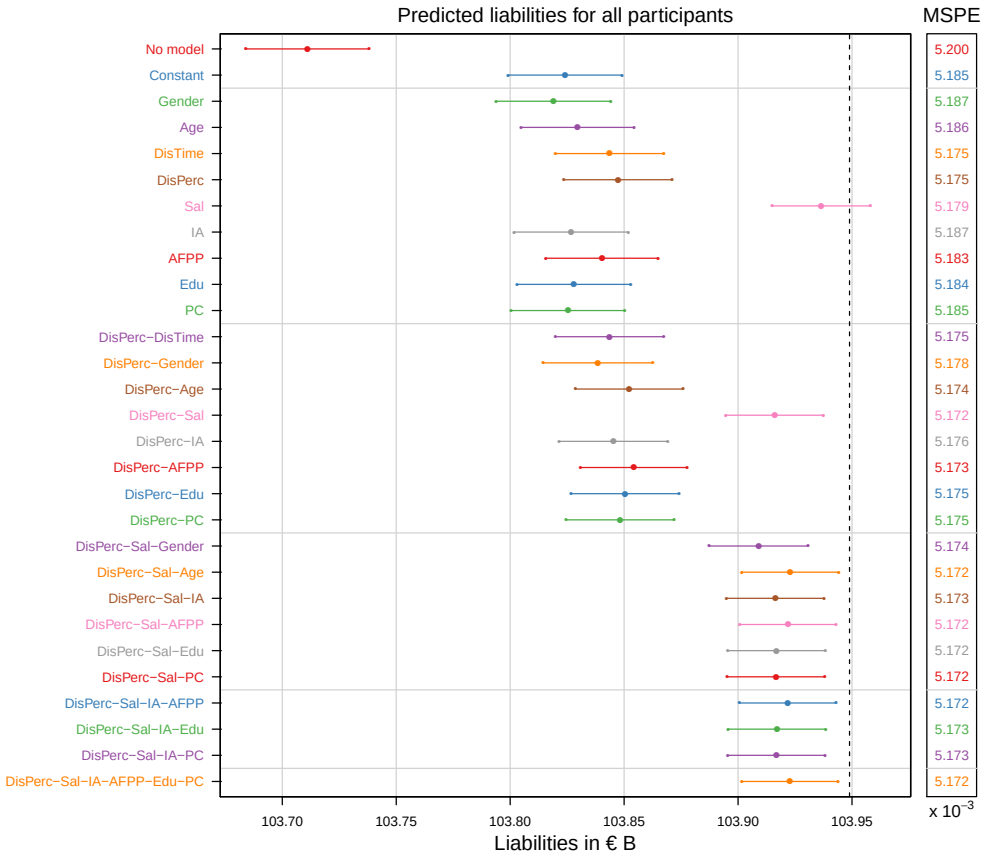


Figure 5.4: Results for the financial backtest. For each model we show the expected value and 90% confidence interval for the predicted liabilities needed on December 31st 2011 (predicted at January 1st 2011). The vertical dotted line represents the liabilities for the participants who were still actually alive on December 31st 2011. The mean squared prediction error is shown for each model on the right-hand side of the figure.

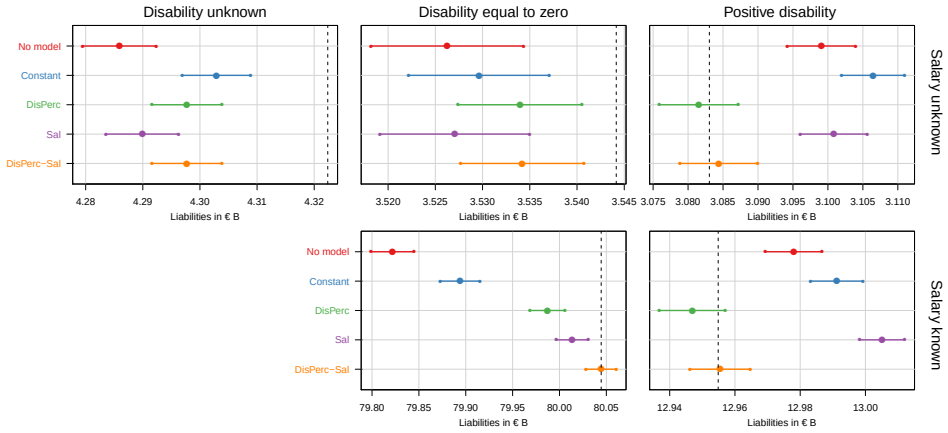


Figure 5.5: The financial backtest split between participants with and without disability and/or salary information. There are no participants with known salary information but unknown disability information. Notes: see Figure 5.4.

liabilities are underestimated, and for participants with positive disability spell the liabilities are overestimated. The actual liabilities fall outside the prediction interval for all groups;

- If both `DisPerc` and `Sal` are included, the predicted liabilities are closer to the actual liabilities for nearly all groups compared to the liabilities predicted using the single factor regression models. For people with known salary the prediction agrees very closely to the actual liabilities.

This example shows that focusing on a single backtest at an aggregate level may lead to suboptimal decisions regarding the selection of risk factors. For one risk profile we may underestimate the liabilities while for another risk profile the liabilities may be overestimated. Therefore, we also calculated the MSPE and the results are included on the right-hand side of Figure 5.4. We calculated the MSPE using liabilities predicted at an individual level, so within the MSPE under- and overestimation cannot cancel out each other.

The MSPE for the model with only `Sal` included is worse than when only `DisPerc` is included. Further, if we include both `DisPerc` and `Sal` the MSPE improves substantially. Once both `Sal` and `DisPerc` have been included, the MSPE improves only marginally when other risk factors are added.

Based on Figures 5.4 and 5.5 we conclude that the MSPE yields more accurate information regarding which models provide accurate liability predictions on an individual level. However, the prediction intervals may provide valuable insights for subgroups of the portfolio on whether liabilities are under- or overestimated. There-

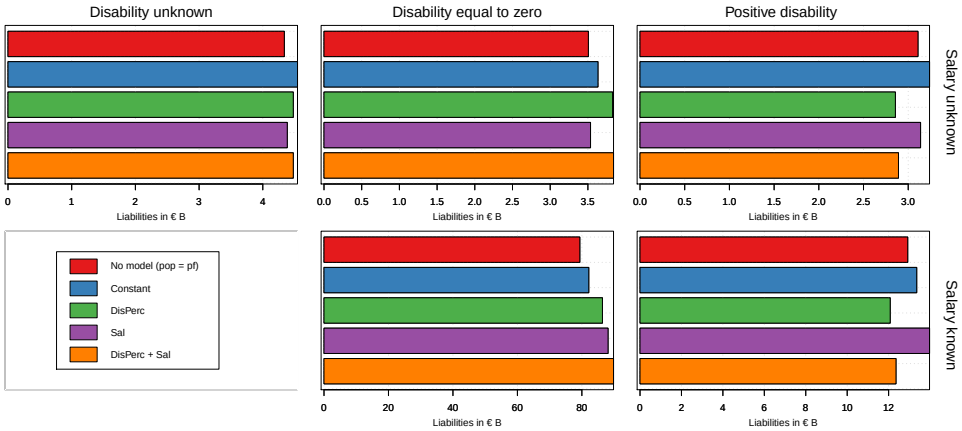


Figure 5.6: The value of the liabilities split between participants with and without disability and/or salary information. For all participants the portfolio factors is used for the entire projection period. There are no participants with known salary information but unknown disability information.

Table 5.4: Liabilities for participants who are alive at January 1st 2011, valued using different specifications for the portfolio-specific factors.

Model	Liabilities in €B
No model	103.24
Constant	106.91
DisPerc	109.62
Sal	113.14
DisPerc-Sal	113.38

fore, we recommend to use both backtesting tools to obtain a complete view on how different models perform relative to each other.

5.4.4 Valuation of liabilities

In this section we quantify the impact on the liabilities when different models are used to explain historically observed portfolio mortality data. In the previous section we used participant-specific mortality adjustments only for the year 2011, and we used the same mortality adjustment for all participants for the years thereafter. We use the same approach here, but we now use the estimated participant-specific mortality adjustments for the entire future instead of only for the year 2011.

In Figure 5.6 we show the liabilities for specifications with `DisPerc` and/or `Sal`

included, but separately for participants with known and unknown disability spell and/or salary. The total liabilities summed over the different subgroups are shown in Table 5.4. By comparing Figure 5.4 and Figure 5.6 we observe that the ranking of the models in terms of liabilities is the same for each subgroup. However, the differences in liabilities between the different models have become larger in absolute terms since the participant-specific mortality adjustments are now used for the entire future. The largest relative difference is observed for the participants with known salary and who have a positive disability spell. The liabilities calculated using the model with only `Sal` included equals €13.97 B, whereas if we also include `DisPerc` the liabilities are only €12.35 B, which is almost 12% lower.

In the financial backtest the total predicted liabilities for the model with both `DisPerc` and `Sal` are lower than those for the model with `Sal` only. However, in Table 5.4 we observe that if we account for the differences in mortality between participants for the entire future that the liabilities when including both `DisPerc` and `Sal` are higher than when only `Sal` is included. If we only include `Sal`, the potential correlation between `Sal` and `DisPerc` is not taken into account. As a result, we value the liabilities lower for the participants with `DisPerc` $= 0$ and known salary (the largest group in Figure 5.6) if we only include `Sal` than when we also include `DisPerc`.

5.5 Conclusion

It is well known that many factors influence the level of individual mortality rates. The increasing sophistication of modern statistical methods allows pension funds to improve their mortality estimates by not just using age and gender. We can now also include other explanatory variables, and these variables may be categorical, continuous or spatial factors.

Using a large Dutch pension portfolio, we show that such extensions are now feasible for the very large datasets involved if the data are aggregated in a particular manner. This allows us to base our conclusions on a large quantity of information, while results that have been reported so far have been based, to the best of our knowledge, on datasets with a much smaller number of pension fund participants.

The scale of our analysis strengthens our conclusion that salary, disability and working at irregular hours all have a particularly strong impact on the level of mortality, while portfolio-specific adjustments to variables such as age, gender and the age at which pension payments start are less important.

By applying a financial backtest which aggregated the effect of individual deaths to the level of the liabilities for the fund as whole, we found that only information on salary and disability is particularly valuable to improve the accuracy of predictions. This shows that one may want to use the modeling framework introduced in this

chapter to distinguish between risk factors which improve the accuracy of predicting individual deaths from risk factors which also improve on the quality of the predictions which are most important from an actuarial perspective.

The application of our framework could be used to include other factors as well if these are available, but this obviously depends on the information that is collected within a pension fund. For older retirees, which form a substantial part of the portfolio, less information on risk factors is available: past salary information is usually missing for example, and only total accrued rights are reported. These accrued rights are determined by the whole history of earned salaries and past service years, and they are therefore difficult to compare between participants in absolute terms. It would be worthwhile to investigate whether they can still be included in an appropriate manner in the regression analysis. Further, in this chapter we only consider single effects, but estimation and backtesting results may improve if combined effects, such as a gender-specific salary effect, are introduced.

Another interesting topic left for future research is related to the derivation of portfolio factors based on insured amount weighted mortality rates. This approach is widely used in practice, but it lacks theoretical justification. In this chapter we have illustrated how a factor such as salary can be included in a regression model to explain observed death counts, but it is worthwhile investigating whether better predictions can be obtained when survival of participants is weighted by their insured amount or the value of their liabilities in a statistically just manner.

Appendix

5.A Detailed description of the financial backtest

Below we first discuss the model used to create population-wide mortality forecasts, then we discuss how the expected present value of annuities is calculated.

5.A.1 Population mortality model

We use the AG2014 model to obtain population mortality forecasts, see Koninklijk Actuarieel Genootschap (2014). This model is an application of the Li and Lee (2005) model to a selection of West-European countries under a Poisson assumption for the observed death counts (Brouhns et al. (2002)).

They consider a group of West-European countries: Austria, Belgium, Denmark, England and Wales, West-Germany, Finland, France, Iceland, Ireland, Luxembourg, the Netherlands, Norway, Sweden and Switzerland. For each country c we download the observed death counts $d^c_{t,x}$ and corresponding risk exposures $E^c_{t,x}$ from the Human Mortality Database[15], and we define $d^{\text{EU}}_{t,x} = \sum_c d^c_{t,x}$ and $E^{\text{EU}}_{t,x} = \sum_c E^c_{t,x}$ for $t = 1970, \ldots, 2010$ and $x = 0, \ldots, 90$.

The mortality model is specified as

$$D^{\text{EU}}_{t,x} \sim \text{Poisson}(E^{\text{EU}}_{t,x} \mu^{\text{EU}}_{t,x}) \tag{5.22}$$
$$D^{\text{NL}}_{t,x} \sim \text{Poisson}(E^{\text{NL}}_{t,x} \mu^{\text{NL}}_{t,x}), \tag{5.23}$$

with

$$\ln \mu^{\text{EU}}_{t,x} = A_x + B_x K_t \tag{5.24}$$
$$\ln \Delta^{\text{NL}}_{t,x} = \alpha_x + \beta_x \kappa_t \tag{5.25}$$
$$\ln \mu^{\text{NL}}_{t,x} = \ln \mu^{\text{EU}}_{t,x} + \ln \Delta^{\text{NL}}_{t,x}. \tag{5.26}$$

[15] http://www.mortality.org

The model is applied to both genders separately, so dependence on gender g is not shown. Here, $\mu_{t,x}^{\text{EU}}$ is the force of mortality for the collection of West-European countries, $\Delta_{t,x}^{\text{NL}}$ is the (relative) difference in force of mortality between West-Europe and the Netherlands, and $\mu_{t,x}^{\text{NL}}$ is the resulting force of mortality for the Netherlands.

For K_t we assume a random walk with drift and for κ_t a mean reverting process that reverts back to zero:

$$K_t = K_{t-1} + \theta + \varepsilon_t, \qquad \varepsilon_t \sim \text{N}(0, \sigma_\varepsilon^2) \tag{5.27}$$

$$\kappa_t = a\kappa_{t-1} + \nu_t, \qquad \nu_t \sim \text{N}(0, \sigma_\nu^2). \tag{5.28}$$

We further assume the error terms ε_t and ν_t to be correlated for fixed t but uncorrelated over different t, so we can use Seemingly Unrelated Regression techniques to estimate this model. To compute annuities (see the next section) we use the most likely mortality path, i.e. we set error terms ε_t and ν_t equal to zero when forecasting mortality. We calibrate our model to the ages 0 to 90, but mortality estimates for higher ages are needed in order to obtain more accurate (complete) estimates of annuities. We use the closure method of Kannistö (1992) to obtain mortality rates for ages higher than 90, using ages $80, \ldots, 90$ in the Kannistö regression.

5.A.2 Assumptions to compute the expected present value of life annuities

The financial backtest uses the expected present value of life annuities a_j for participant j starting at December 31st 2011, as predicted on December 31st 2010. We make the following assumptions when computing the expected present value of the annuities:

1. All participants have integer age at January 1st 2011 (rounded down);
2. Benefits are paid halfway during the year;
3. There are no payments after age 121, i.e. $p_j = 0$ when $x(j,t) \geq 121$;
4. The relative difference between mortality in the population and the pension fund is the same for all participants in the pension fund. This difference is represented by the factor η which is computed as

$$\sum_{tjx} \delta_{tjx} / \sum_{tjx} \tau_{tjx} \hat{\mu}_{tx}^{\text{AG},g(j)},$$

where the summation is taken over all individuals in the training sample $t = \{2006, \ldots, 2010\}$, and including both information before and after the birthday of participant j.

Here, $\hat{\mu}_{tx}^{\mathrm{AG},g}$ is obtained by calibrating the AG2014 model on mortality data from 1970-2010. The portfolio factor η defined above is the same as the one that results from a model with only a constant included in (5.7).

We use the discount curve published by the Dutch Central Bank for December 31st 2010, and we define z_k as the k-year zero rate for $k \geq 0$ and $z_0 = 0$. We define the mid-year discount factor in year k as

$$DF_{k+\frac{1}{2}} = [(1+z_k)^k \cdot (1+z_{k+1})^{k+1}]^{-1/2}, \tag{5.29}$$

with $k \geq 0$.

In calculating the expected present value of the annuities a_j we only consider the risk factors age x and gender g. The one-year survival probability for participant j in year $2011+k$ is given by

$$p_{j,k} = \exp[-\hat{\mu}^{\mathrm{AG},g(j)}_{2011+k,x(j,2011)+k} \cdot \eta], \quad \text{for } k \geq 0,$$

where the force of mortality $\hat{\mu}_{tx}^{\mathrm{AG},g}$ is obtained from a calibration using mortality data up to and including the year 2010. Given the set of assumptions listed above, we compute an annuity a_j as follows:

$$a_j = \sum_{k \geq 0} DF_{k+\frac{1}{2}} \cdot {}_{k+\frac{1}{2}}p_j,$$

with

$${}_{k+\frac{1}{2}}p_j = \begin{cases} (p_{j,0})^{\frac{1}{2}} & \text{for } k = 0, \\ \left(\prod_{l=0}^{k-1} p_{j,l}\right)(p_{j,k})^{\frac{1}{2}} & \text{for } k \geq 1. \end{cases}$$

The expected present value of the annuity a_j is used in (5.18) and (5.19).

Summary

This thesis investigates models for historical mortality data that can be used to construct forecasts for both population mortality and portfolio-specific mortality.

In **Chapter 2** we provide an extensive introduction to stochastic mortality modeling. We discuss the Lee-Carter model and several extensions, and we provide an overview of different approaches to forecast the time series in mortality models. In a Bayesian setting, we demonstrate the impact of parameter uncertainty on prediction intervals for mortality rates.

For nearly all mortality models the random walk with drift is used for the modeling and projection of the underlying time series. However, projections are potentially very sensitive with respect to the calibration period, and when the historical data contains a structural change it is possible that mortality forecasts are not in line with the most recent observations. In **Chapter 3** we propose a modeling approach in which the possible presence of structural changes in the calibrated period effects is explicitly incorporated. We find that allowing for multiple structural changes results in more plausible projections of the period effects, and that this also leads to improved backtesting results in some cases.

In **Chapter 4** we propose a method to estimate population mortality and portfolio-specific mortality simultaneously. We define our model in a Bayesian setting for two reasons: 1) it allows for simultaneous estimation of the population mortality model and the portfolio-specific adjustments, and 2) as a byproduct of the estimation procedure, we obtain information on parameter uncertainty which is essential for risk management purposes. For the population we use the Lee-Carter model, and we use age-dependent factors to adjust mortality in the portfolio. We consider two specifications for the age-dependent portfolio-specific factors, and we find that imposing a correlation structure between these factors for different ages results in a reduction of the uncertainty in their estimates. However, for the future numbers of deaths, we find that individual mortality risk is most relevant on the short term, and as the projection horizon increases, uncertainty in the mortality trend becomes increasingly important.

Finally, in **Chapter 5** we investigate mortality in a large pension fund. We

estimate effects for a variety of risk factors using generalized additive models (GAMs). The use of GAMs means that we do not have to impose beforehand what the structure of the effects should be (e.g. linear or quadratic), since we estimate the structure using thin plate splines. We consider different combinations of risk factors, and compare the model fit using information criteria and proper scoring rules. Of particular interest for pension funds is the impact of the inclusion of different risk factors when predicting the value of the liabilities that are needed at the end of the year. We show that it is crucial to include salary information to ensure that the estimated value of the liabilities is at an appropriate level for the portfolio as a whole. However, when disability information is included additional to salary information, predictions become significantly more accurate for specific risk groups.

Samenvatting

Rond 1850 werden mensen gemiddeld nog geen 40 jaar oud, inmiddels wordt meer dan 60% van de Nederlandse mannen en 75% van de Nederlandse vrouwen die nu geboren worden ouder dan 90 jaar. Pensioenfondsen en levensverzekeraars hebben langlopende verplichtingen aan hun deelnemers respectievelijk polishouders. Voor een nauwkeurige waardering van die verplichtingen is een goede inschatting van de toekomstige ontwikkelingen in sterftekansen essentieel.

Sinds de jaren '90 wordt er veel onderzoek gedaan naar modellen die de historische sterftedata proberen te beschrijven, en die vervolgens gebruikt kunnen worden om projecties van sterftekansen voor de toekomst te maken. Vaak worden daarbij lineaire modellen gebruikt om tijdreeksen te projecteren, zonder rekening te houden met mogelijke veranderingen in de bijbehorende parameters. Dit heeft tot gevolg dat projecties niet altijd goed aansluiten bij de meest recente historie.

In hoofdstuk 3 onderzoeken we modellen waarin structurele veranderingen in tijdreeksen worden geïdentificeerd. Op basis van informatiecriteria bepalen we het meest aannemelijke aantal structurele veranderingen, en we laten zien dat projecties van tijdreeksen het beste aansluiten bij historische waarnemingen wanneer rekening gehouden wordt met meerdere structurele veranderingen. We testen de meerwaarde van het toelaten van structurele veranderingen door de voorspellende kracht van de sterftemodellen te testen. We zien dat die in enkele gevallen verbetert maar soms ook verslechtert. In die gevallen zijn de projecties van sterftekansen minder intuïtief, ook al zijn de projecties van de onderliggende tijdreeksen wel plausibeler.

In een populatie kunnen er tussen verschillende groepen individuen grote verschillen bestaan in sterftekansen. Mensen met een fysiek zwaar beroep overlijden naar verwachting eerder. In de waardering van verplichtingen moet ook hier rekening mee gehouden worden: de projecties van sterftekansen voor de Nederlandse bevolking kunnen niet direct toegepast worden, maar moeten gecorrigeerd worden voor portefeuillespecifieke kenmerken.

In hoofdstuk 4 beschouwen we sterfte in een portefeuille van verzekerden in Engeland en Wales met sterfte in de gehele bevolking in Engeland en Wales. We veronderstellen dat de relatieve verschillen in sterfte tussen beide groepen (de portefeuille-

factoren) leeftijdsafhankelijk en constant over de tijd zijn. Doordat het model in een Bayesiaanse setting geschat wordt, zijn we in staat om drie vormen van onzekerheid van elkaar te onderscheiden: individueel sterfterisico, onzekerheid in de sterfteontwikkelingen, en parameteronzekerheid. We vinden dat de onzekerheid in de portefeuillefactoren groter is wanneer onafhankelijkheid tussen de portefeuillefactoren verondersteld wordt, en dat individueel sterfterisico op de korte termijn een grotere impact heeft op de onzekerheid in sterfteaantallen dan parameteronzekerheid. Op de lange termijn heeft de onzekerheid in de sterftetrend de grootste impact op de onzekerheid in sterfteaantallen.

Tot slot analyseren we in hoofdstuk 5 de sterfte die geobserveerd is in een groot Nederlands pensioenfonds over de jaren 2006 tot en met 2011. De dataset bevat informatie over verschillende risicofactoren zoals salaris en arbeidsongeschiktheid, en deze informatie kan gebruikt worden om verschillen in sterfte te verklaren. We gebruiken een gegeneraliseerd additief model, waardoor we niet-lineaire gladde effecten kunnen schatten voor de risicofactoren; dit in tegenstelling tot de lineaire effecten die vaak met een gegeneraliseerd lineair model geschat worden. Door het gebruik van informatiecriteria en kruisvalidatietesten vinden we dat het salaris, een eventueel arbeidsongeschiktheidsverleden, en een toeslag voor het werken op onregelmatige tijden de sterkste invloed hebben op de sterftekansen. Pensioenfondsen en verzekeraars zijn vooral geïnteresseerd in het accuraat voorspellen van (de vrijval in) de waarde van de verplichtingen, en we vinden dat informatie over het salaris en het arbeidsongeschiktheidsverleden hier cruciaal voor zijn.

Bibliography

D. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1992.

A. Antoniadis, G. Grégoire, and I. McKeague. Bayesian estimation in single-index models. *Statistica Sinica*, 14:1147 – 1164, 2004.

K. Antonio and Y. Zhang. Nonlinear mixed models. In E. Frees, R. Derrig, and G. Meyers, editors, *Predictive Modeling Applications in Actuarial Science*, volume 1, pages 398 – 426. Cambridge University Press, 2014.

K. Antonio, A. Bardoutsos, and W. Ouburg. A Bayesian Poisson log-bilinear model for mortality projections with multiple populations. *European Actuarial Journal*, 5 (2):245 – 281, 2015.

K. Antonio, S. Devriendt, W. de Boer, R. de Vries, A. De Waegenaere, H.-K. Kan, E. Kromme, W. Ouburg, T. Schulteis, E. Slagter, M. van der Winden, C. van Iersel, and M. Vellekoop. Producing the Dutch and Belgian mortality projections: a stochastic multi-population standard. *European Actuarial Journal*, 2017. Available online at: https://doi.org/10.1007/s13385-017-0159-x.

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

P. Barrieu, H. Bensusan, N. E. Karoui, C. Hillairet, S. Loisel, C. Ravanelli, and Y. Salhi. Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian Actuarial Journal*, 3:203 – 231, 2012.

H. Booth, J. Maindonald, and L. Smith. Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3):325 – 336, 2002.

M. Bots and D. Grobbee. Decline of coronary heart disease mortality in the Netherlands from 1978 to 1985: contribution of medical care and changes over time in presence of major cardiovascular risk factors. *Journal of cardiovascular risk*, 3(3): 271 – 276, 1996.

N. Brouhns, M. Denuit, and J. Vermunt. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373 – 393, 2002.

R. Brown and J. McDaid. Factors affecting retirement mortality. *North American Actuarial Journal*, 7(2):24 – 43, 2003.

A. Cairns, D. Blake, and K. Dowd. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73 (4):687 – 718, 2006.

A. Cairns, D. Blake, and K. Dowd. Modelling and management of mortality risk: a review. *Scandinavian Actuarial Journal*, 2008(2-3):79 – 113, 2008.

A. Cairns, D. Blake, K. Dowd, G. Coughlan, D. Epstein, A. Ong, and I. Balevich. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1 – 35, 2009.

A. Cairns, D. Blake, K. Dowd, G. Coughlan, D. Epstein, and M. Khalaf-Allah. Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355 – 367, 2011a.

A. Cairns, D. Blake, K. Dowd, G. Coughlan, and M. Khalaf-Allah. Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, 41(1):25 – 59, 2011b.

A. Cairns, M. Kallestrup-Lamb, C. Rosenskjold, D. Blake, and K. Dowd. Modelling socio-economic differences in the mortality of Danish males using a new affluence index. 2016. Available online at: http://www.macs.hw.ac.uk/~andrewc/.

L. Chen, A. Cairns, and T. Kleinow. Small population bias and sampling effects in stochastic mortality modelling. *European Actuarial Journal*, 7(1):193 – 230, 2017.

R. Chetty, M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler. The association between income and life expectancy in the United States, 2001-2014. *Journal of the American Medical Association*, 315(16):1750 – 1766, 2016.

G. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

E. Coelho and L. Nunes. Forecasting mortality in the event of a structural change. *Journal of the Royal Statistical Society: Series A*, 174(3):713 – 736, 2011.

E. Coelho and L. Nunes. Cohort effects and structural changes in the mortality trend. Working paper, October 2013. Available at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.11/2013/WP_5.1.pdf.

P. Congdon. Life expectancies for small areas: a Bayesian random effects methodology. *International Statistical Review*, 77(2):222 – 240, 2009.

G. Costa. The impact of shift and night work on health. *Applied Ergonomics*, 27(1): 9 – 16, 1996.

I. Currie. Smoothing and forecasting mortality rates with P-splines. Talk given at the Institute of Actuaries, 2006. Available online at: http://www.ma.hw.ac.uk/~iain/research/talks.html.

C. Czado, A. Delwarde, and M. Denuit. Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, 36(3):260 – 284, 2005.

C. Czado, T. Gneiting, and L. Held. Predictive model assessment for count data. *Biometrics*, 65:1254 – 1261, 2009.

M. Denuit and A. Goderniaux. Closing and projecting lifetables using log-linear models. *Bulletin of the Swiss Association of Actuaries*, 1:29 – 49, 2005.

M. Denuit and S. Lang. Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics*, 35(3):627 – 647, 2004.

M. Denuit, J. Dhaene, M. Goovaerts, and R. Kaas. *Actuarial theory for dependent risks*. John Wiley & Sons, 2005.

M. Denuit, X. Maréchal, S. Pitrebois, and J.-F. Walhin. *Actuarial Modelling of Claim Counts*. John Wiley & Sons, Ltd, 2007.

K. Dowd, A. Cairns, D. Blake, G. Coughlan, and M. Khalaf-Allah. A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2):334 – 356, 2011.

I. Elo and S. Preston. Educational differentials in mortality: United States, 1979-85. *Social Science & Medicine*, 42(1):47 – 57, 1996.

A. Finkelstein and J. Poterba. Selection effects in the United Kingdom individual annuities market. *The Economic Journal*, 112(476):28 – 50, 2002.

R. Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 217(1130):295 – 305, 1953.

W. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53:789 – 798, 1958.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515 – 534, 2006.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457 – 472, 1992.

T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359 – 378, 2007.

S. Gschlössl, P. Schoenmaekers, and M. Denuit. Risk classification in life insurance: methodology and case study. *European Actuarial Journal*, 1:23 – 41, 2011.

S. Haberman and A. Renshaw. A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35 – 55, 2011.

S. Haberman, V. Kaishev, P. Millossovich, A. Villegas, S. Baxter, A. Gaches, S. Gunnlaugsson, and M. Sison. Longevity basis risk: A methodology for assessing basis risk. Technical report, Institute and Faculty of Actuaries, December 2014. Available online at: www.actuaries.org.uk/events/pages/sessional-research-programme.

D. Hainaut. Multi dimensional Lee-Carter model with switching mortality processes. *Insurance: Mathematics and Economics*, 50(2):236 – 246, 2012.

D. Harris, D. Harvey, S. Leybourne, and A. Taylor. Testing for a unit-root in the presence of a possible break in trend. *Econometric Theory*, 25:1545 – 1588, 2009.

D. Harvey, S. Leybourne, and A. Taylor. Simple, robust and powerful tests of changing trend hypothesis. *Econometric Theory*, 25:995 – 1029, 2009.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297 – 310, 1986.

W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97 – 109, 1970.

R. Henckaerts, K. Antonio, M. Clijsters, and R. Verbelen. A data driven binning strategy for the construction of insurance tariff classes. 2017. Available online at: https://lirias.kuleuven.be/bitstream/123456789/583471/1/AFI_17115.pdf.

J. Hoem. Levels of error in population forecasts. *Artikler fra Statistisk Sentralbyrå*, 61, 1973.

P. Hoff. Simulation of the matrix Bingham-von-Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computation and Graphical Statistics*, 18(3):438 – 456, 2009.

N. Hári, A. De Waegenaere, B. Melenberg, and T. Nijman. Estimating the term structure of mortality. *Insurance: Mathematics and Economics*, 42(2):492 – 504, 2008a.

N. Hári, A. De Waegenaere, B. Melenberg, and T. Nijman. Longevity risk in portfolios of pension annuities. *Insurance: Mathematics and Economics*, 42(2):505 – 519, 2008b.

F. Janssen, A. Kunst, and J. Mackenbach. Variations in the pace of old-age mortality decline in seven European countries, 1950-1999: The role of smoking and other factors earlier in life. *European Journal of Population*, 23(2):171–188, 2007.

H. Kan. A Bayesian mortality forecasting framework for population and portfolio mortality. MSc thesis, University of Amsterdam, The Netherlands, 2012.

V. Kannistö. *Development of the oldest - old mortality, 1950-1980: evidence from 28 developed countries*. Odense University Press, 1992.

N. Klein, M. Denuit, S. Lang, and T. Kneib. Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55:225 – 249, 2014.

T. Kleinow. A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, 63:147 – 152, 2015.

T. Kleinow and S. Richards. Parameter risk in time-series mortality forecasts. *Scandinavian Actuarial Journal*, pages 1 – 25, 2016. Available online at: http://dx.doi.org/10.1080/03461238.2016.1255655.

Koninklijk Actuarieel Genootschap. Concept Leidraad Ervaringssterfte, 2012. Available online at: http://www.ag-ai.nl/download/26398-Concept_Leidraad_Ervaringssterfte_2_oktober_2012.pdf.

Koninklijk Actuarieel Genootschap. Projection table AG 2014, 2014. Available online at: http://www.ag-ai.nl/view.php?action=view&Pagina_Id=625.

T. Krivobokova and G. Kauermann. A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102(480):1328 – 1337, 2007.

P. Lantz, J. House, J. Lepkowski, D. Williams, R. Mero, and J. Chen. Socioeconomic factors, health behaviors, and mortality: Results from a nationally representative prospective study of US adults. *Journal of the American Medical Association*, 279 (21):1703 – 1708, 1998.

R. Lee and L. Carter. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.

H. Li, A. De Waegenaere, and B. Melenberg. The choice of sample size for mortality forecasting: a Bayesian learning approach. *Insurance: Mathematics and Economics*, 63:153 – 168, 2015.

J. Li. An application of MCMC simulation in mortality projection for populations with limited data. *Demographic Research*, 30(1):1 – 48, 2014.

J.-H. Li, W.-S. Chan, and S.-H. Cheung. Structural changes in the Lee-Carter mortality indexes: detection and implications. *North American Actuarial Journal*, 15 (1):13 – 31, 2011.

N. Li and R. Lee. Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography*, 42(3):575 – 594, 2005.

M. Lindbergson. Mortality among the elderly in Sweden 1988−1997. *Scandinavian Actuarial Journal*, (3):79 – 94, 2001.

E. Lovász. Analysis of Finnish and Swedish mortality data with stochastic mortality models. *European Actuarial Journal*, 1:259–289, 2011.

G. Marra and S. Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53 – 74, 2012.

A. Milidonis, Y. Lin, and S. Cox. Mortality regimes and pricing. *North American Actuarial Journal*, 15(2):266 – 289, 2011.

R. Moreno-Serra and A. Wagstaff. System-wide impacts of hospital payment reforms: evidence from Central and Eastern Europe and Central Asia. *Journal of Health Economics*, 29(4):585 – 602, 2010.

J. Oeppen and J. Vaupel. Broken limits to life expectancy. *Demography*, 296(5570): 1029 – 1031, 2002.

C. O'Hare and Y. Li. Explaining young mortality. *Insurance: Mathematics and Economics*, 50(1):12 – 25, 2011.

C. O'Hare and Y. Li. Identifying structural breaks in stochastic mortality models. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B*, 1 (2):1 – 14, 2015.

A. Olivieri. Stochastic mortality: experience-based modeling and application issues consistent with Solvency 2. *European Actuarial Journal*, 1:S101 – S125, 2011.

E. Pitacco, M. Denuit, S. Haberman, and A. Olivieri. *Modelling Longevity Dynamics for Pensions and Annuity Business*. Oxford University Press, New York, 2009.

R. Plat. On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393 – 404, 2009a.

R. Plat. Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, 45:123 – 132, 2009b.

O. Purcaru, M. Guillén, and M. Denuit. Linear credibility models based on time series for claim counts. *Belgian Actuarial Bulletin*, 4(1):62 – 74, 2004.

P. Reiss and R. Ogden. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B*, 71(2):505 – 523, 2009.

A. Renshaw and S. Haberman. Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255 – 272, 2003.

A. Renshaw and S. Haberman. A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556 – 570, 2006.

S. Richards, K. Kaufhold, and S. Rosenbusch. Creating portfolio-specific mortality tables: a case study. *European Actuarial Journal*, 3:295 – 319, 2013.

A. Riebler, L. Held, and H. Rue. Estimation and extrapolation of time trends registry data - borrowing strength from related populations. *The Annals of Applied Statistics*, 6(1):304 – 333, 2012.

RIVM. Socio-economic differences in remaining life expectancy, 2014. Available online at: http://www.eengezondernederland.nl/Heden_en_verleden/Levensverwachting (accessed on 2017-10-30).

G. Roberts, A. Gelman, and W. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110 – 120, 1997.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

A. Smith and G. Roberts. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 55(1):3 – 23, 1993.

L. Stoeldraijer, C. van Duin, and F. Janssen. Bevolkingsprognose 2012-2060: Model en veronderstellingen betreffende de sterfte. Technical report, Centraal Bureau voor de Statistiek, 2013. Available online at: https://www.cbs.nl/-/media/imported/documents/2013/26/2013-07-b15-art.pdf.

P. Sweeting. A trend-change extension of the Cairns-Blake-Dowd Model. *Annals of Actuarial Science*, 5(2):143 – 162, 2011.

J. Tornij. De ontwikkeling van sterfte in de tijd: daling van sterftekansen gaat onverminderd door. *De Actuaris*, pages 13 – 14, March 2004.

R. Tsay. *Analysis of Financial Time Series*. John Wiley & Sons, 2010.

F. van Berkum, K. Antonio, and M. Vellekoop. The impact of multiple structural changes on mortality predictions. *Scandinavian Actuarial Journal*, 7:581 – 603, 2016.

F. van Berkum, K. Antonio, and M. Vellekoop. A Bayesian joint model for population and portfolio-specific mortality. *ASTIN Bulletin*, 47(3):681 – 713, 2017a.

F. van Berkum, K. Antonio, and M. Vellekoop. Unraveling relevant risk factors explaining pension fund mortality: a case study in the Netherlands. *Working paper*, 2017b.

J. Vaupel. Relatives' risks: frailty models of life history data. *Theoretical population biology*, 37(1):220 – 234, 1990.

R. Verbelen, K. Antonio, and G. Claeskens. Unraveling the predictive power of telematics data in car insurance pricing. 2016. Available online at: https://feb.kuleuven.be/public/ndbaf45/papers/KBI_1624.pdf.

A. Villegas and S. Haberman. On the modeling and forecasting of socioeconomic mortality differentials: An application to deprivation and mortality in England. *North American Actuarial Journal*, 18(1):168 – 193, 2014.

R. von Mises. Über die 'Ganzzahligkeit' der Atomgewicht und verwandte Fragen. *Physikalische Zeitschrift*, 19:490 – 500, 1918.

S. Wood. *Generalized additive models: an introduction with R*. CRC Press, 2006.

S. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73(1):3 – 36, 2011.

S. Wood, N. Pya, and B. Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548 – 1563, 2016.

Y. Yang, A. De Waegenaere, and B. Melenberg. Do Americans live longer and healthier? Forecasting healthy life expectancy by including dynamic evolutions of mortality, health, and macroeconomic variables. 2014. Available at SSRN: https://ssrn.com/abstract=2662978.

Y.-C. Yao. Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181 – 189, 1988.

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. `strucchange`: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002.

A. Zeileis, C. Kleiber, W. Krämer, and K. Hornik. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1–2):109 – 123, 2003.

E. Zivot and D. Andrews. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, 10(3): 251 – 270, 1992.