

# A Study of Statistical Methods for Modelling Longevity and Climate Risks

by

Yiping Guo

A thesis  
presented to the University of Waterloo  
in fulfillment of the  
requirement for the degree of  
Doctor of Philosophy  
in  
Actuarial Science

Waterloo, Ontario, Canada, 2025

© Yiping Guo 2025

## Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Wenjun Zhu  
Associate Professor, Nanyang Business School  
Nanyang Technological University

Supervisor(s): Johnny S.-H. Li  
Professor, Dept. of Finance  
Chinese University of Hong Kong

Ben Feng  
Associate Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Internal Member: Kenneth Q. Zhou  
Associate Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Lisa Gao  
Assistant Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Internal-External Member: Justin W.L. Wan  
Professor, David R. Cheriton School of Computer Science  
University of Waterloo

### **Author's Declaration**

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

In recent years, two pivotal risks have emerged and taken a significant position in modern actuarial science: longevity risk and climate risk. Longevity risk, or the risk of individuals living longer than expected, poses a severe challenge to both private insurance companies and public pension systems, potentially destabilizing financial structures built on assumptions of life expectancy. On the other hand, climate risk, associated with fluctuations and extreme conditions in weather, has substantial implications for various sectors such as agriculture, energy, and insurance, particularly in the era of increasing climate change impacts. The Society of Actuaries (SOA) has recognized the growing importance of these risks, advocating for innovative research and solutions to manage them effectively. Furthermore, statistical modelling plays an indispensable role in understanding, quantifying, and managing these risks. The development of sophisticated and robust statistical methods enables practitioners and researchers to capture complex risk patterns and make reliable predictions, thereby informing risk management strategies. This thesis, composed of four distinct projects, explores statistical methods for modelling longevity and weather risk, contributing valuable insights to these fields.

The first part in this thesis studies the statistical methods for modelling longevity risk, and in particular, modelling mortality rates. In the first chapter, we study parameter estimation of the Lee-Carter model and its multi-population extensions. Although the impact of outliers on stochastic mortality modelling has been examined, previous studies on this topic focus on how outliers in the estimated time-varying indexes may be detected and/or modelled, with little attention being paid to the adverse effects of outliers on estimation robustness, particularly that pertaining to age-specific parameters. In this chapter, we propose a robust estimation method for the Lee-Carter model, through a reformulation of the model into a probabilistic principal component analysis with multivariate  $t$ -distributions and an efficient expectation-maximization algorithm for implementation. The proposed method yields significantly more robust parameter estimates, while preserving the fundamental interpretation for the bilinear term in the model as the first principal component and the flexibility of pairing the estimated time-varying parameters with any appropriate time-series process. We also extend the proposed method for use with multi-population

generalizations of the Lee-Carter model, allowing for a wider range of applications such as quantification of population basis risk in index-based longevity hedges. Using a combination of real and pseudo datasets, we demonstrate that the superiority of the proposed method relative to conventional estimation approaches such as singular value decomposition and maximum likelihood.

Next, we move onto parameter estimation of the Renshaw-Haberman model, a cohort-based extension to the Lee-Carter model. In mortality modelling, cohort effects are often taken into consideration as they add insights about variations in mortality across different generations. Statistically speaking, models such as the Renshaw-Haberman model may provide a better fit to historical data compared to their counterparts that incorporate no cohort effects. However, when such models are estimated using an iterative maximum likelihood method in which parameters are updated one at a time, convergence is typically slow and may not even be reached within a reasonably established maximum number of iterations. Among others, the slow convergence problem hinders the study of parameter uncertainty through bootstrapping methods. In this chapter, we propose an intuitive estimation method that minimizes the sum of squared errors between actual and fitted log central death rates. The complications arising from the incorporation of cohort effects are overcome by formulating part of the optimization as a principal component analysis with missing values. Using mortality data from various populations, we demonstrate that our proposed method produces satisfactory estimation results and is significantly more efficient compared to the traditional likelihood-based approach.

The third part of this thesis continues our exploration of the efficient computational algorithm of the Renshaw-Haberman model. Existing software packages and estimation algorithms often rely on maximum likelihood estimation with iterative Newton-Raphson methods, which can be computationally intensive and prone to convergence issues. In this chapter, we present the R package `RHa1s`, offering an efficient alternative with an alternating least squares method for fitting a generalized class of Renshaw-Haberman models, including configurations with multiple age-period terms. We extend this method to multi-population settings, allowing for shared or population-specific age effects under various configurations. The full modelling workflow and functionalities of `RHa1s` are demonstrated using mortality data from England and Wales.

Lastly, we turn to modelling climate risk in the final chapter of the thesis. The use of weather index insurances is subject to spatial basis risk, which arises from the fact that the location of the user's risk exposure is not the same as the location of any of the weather stations where an index can be measured. To gauge the effectiveness of weather index insurances, spatial interpolation techniques such as kriging can be adopted to estimate the relevant weather index from observations taken at nearby locations. In this chapter, we study the performance of various statistical methods, ranging from simple nearest neighbor to more advanced trans-Gaussian kriging, in spatial interpolations of daily precipitations with data obtained from the US National Oceanic and Atmospheric Administration. We also investigate how spatial interpolations should be implemented in practice when the insurance is linked to popular weather indexes including annual consecutive dry days (*CDD*) and maximum five-day precipitation in one month (*MFP*). It is found that although spatially interpolating the raw weather variables on a daily basis is more sophisticated and computationally demanding, it does not necessarily yield superior results compared to direct interpolations of *CDD/MFP* on a yearly/monthly basis. This intriguing outcome can be explained by the statistical properties of the weather indexes and the underlying weather variables.

## Acknowledgements

First, I would like to express my deepest gratitude to my Ph.D. supervisors, Professor Johnny Li and Professor Ben Feng, for their invaluable support throughout my Ph.D. journey. I am also sincerely grateful to Professor Lisa Gao, Professor Kenneth Zhou, Professor Justin Wan, and Professor Wenjun Zhu for their time and effort in examining this thesis.

A special note of thanks goes to my supervisor, Professor Li. I first met him at the University of Melbourne in 2018 and was immediately inspired by his passion for research and teaching. Even as a master's student back then, he generously offered opportunities for academic exploration. In 2021, I was fortunate to begin my Ph.D. under his supervision at the University of Waterloo. His guidance has been instrumental in helping me transition from a background in statistics to actuarial research, while also allowing me the freedom to explore my interests. This work would not have been possible without his wisdom, encouragement, and unwavering support.

Next, I want to express my heartfelt thanks to my parents. They have always been there, trusting and supporting me unconditionally, and encouraging me to pursue what I love. I am truly blessed to have such supportive parents in my life.

Moving forward, I would like to extend sincere thanks to my girlfriend, Ruiqi. Throughout the last two years, I faced unexpected changes of plan, peer pressure, and uncertainties about my future career, and I have often found myself under tremendous pressure. However, her unconditional love, understanding, and encouragement have been a constant source of momentum in my life.

Further, I would like to extend my gratitude to my friends, whose companionship and support have been invaluable throughout this journey. My sincere thanks also go to the faculty and staff in the department, particularly Ms. Mary Lou Dufton, Mr. Greg Preston, Mr. Carlos Mendes, Ms. Carla Daniels, and Ms. Shaleen Mathur, for their tremendous administrative assistance.

Lastly, I gratefully acknowledge the financial support from my supervisors, the Department of Statistics and Actuarial Science, and the James C. Hickman Scholar program of the Society of Actuaries, which allowed me to focus fully on my research.

# Table of Contents

Examining Committee Membership	ii
Author's Declaration	iii
Abstract	iv
Acknowledgements	vii
List of Figures	xiii
List of Tables	xv
<b>1 Introduction</b>	<b>1</b>
1.1 Background: The Rising Importance of Longevity and Climate Risks in Actuarial Science . . . . .	1
1.2 Objectives and Outline of the Thesis . . . . .	4
<b>2 Robust Parameter Estimation for the Lee-Carter Family: A Probabilistic Principal Component Approach</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.2 Background . . . . .	11

2.2.1	The Lee-Carter Model . . . . .	11
2.2.2	Singular Value Decomposition . . . . .	12
2.2.3	Poisson Maximum Likelihood . . . . .	13
2.2.4	Forecasting . . . . .	14
2.2.5	The Need for Robustifying the Lee-Carter Model . . . . .	14
2.3	Standard PPCA . . . . .	15
2.3.1	Basic Formulation . . . . .	16
2.3.2	Formulation via a Latent Variable Structure . . . . .	17
2.4	Proposed Method . . . . .	18
2.4.1	Multivariate $t$ -Distributions . . . . .	19
2.4.2	Model Formulation . . . . .	19
2.4.3	The EM Algorithm . . . . .	21
2.5	Multi-Population Extensions . . . . .	24
2.5.1	The Augmented Common Factor Model . . . . .	24
2.5.2	Common Age-Effect Model . . . . .	26
2.6	Numerical Illustrations . . . . .	28
2.6.1	Impact of World War II . . . . .	28
2.6.2	Simulated Pandemic Effects . . . . .	32
2.7	Concluding Remarks . . . . .	41
<b>3</b>	<b>Fast Estimation of the Renshaw-Haberman Model and Its Variants</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	The Lee-Carter Model . . . . .	48
3.2.1	Specification . . . . .	48

3.2.2	Least Squares Estimation . . . . .	49
3.2.3	Maximum Likelihood Estimation . . . . .	50
3.3	The Renshaw-Haberman Model and Its Variants . . . . .	50
3.3.1	Specification . . . . .	50
3.3.2	Estimation . . . . .	51
3.3.3	Existing Methods for Expediting Estimation . . . . .	52
3.4	The Proposed Method . . . . .	54
3.4.1	Motivation . . . . .	54
3.4.2	Main Optimization: Alternating Minimization . . . . .	55
3.4.3	Updating the Age-Cohort Parameters: PCA with Missing Values via an Iterative SVD . . . . .	57
3.5	Integrating the Proposed Method with the Existing Methods . . . . .	60
3.5.1	Implementing with the H1 Model . . . . .	60
3.5.2	Implementing with the Hunt-Villegas Method . . . . .	61
3.6	Numerical Illustrations . . . . .	64
3.6.1	Comparing Least Squares with Poisson MLE . . . . .	64
3.6.2	Sharpness of Objective Functions . . . . .	69
3.6.3	Implementing with the H1 Model . . . . .	72
3.6.4	Quantifying Parameter Uncertainty . . . . .	75
3.7	Concluding Remarks . . . . .	79
<b>4</b>	<b>RHals: An R Package for Efficient Alternating Least Squares Estimation of the Renshaw-Haberman Model and Its Extensions</b>	<b>82</b>
4.1	Introduction . . . . .	82
4.2	The Generalized Renshaw-Haberman Model . . . . .	86

4.2.1	Model Formulation . . . . .	86
4.2.2	Alternating Least Squares Estimation of the Generalized Renshaw-Haberman Model . . . . .	89
4.3	Main Functionalities of the RHals Package . . . . .	91
4.3.1	Model Fitting . . . . .	91
4.3.2	Model Selection . . . . .	97
4.3.3	Uncertainty Estimation . . . . .	102
4.3.4	Forecasting . . . . .	105
4.4	Multi-Population Extensions . . . . .	107
4.4.1	Model Formulation and Fitting . . . . .	107
4.4.2	Implementation by RHals . . . . .	110
4.5	Concluding Remarks . . . . .	114
<b>5</b>	<b>Kriging Methods for Modelling Spatial Basis Risk in Weather Index Insurances: A Technical Note</b>	<b>115</b>
5.1	Introduction . . . . .	115
5.2	Data Description and Visualization . . . . .	119
5.3	Methodology . . . . .	124
5.3.1	Nearest Neighbor . . . . .	124
5.3.2	Inverse Distance Weighting . . . . .	125
5.3.3	Ordinary Kriging . . . . .	125
5.3.4	Universal Kriging . . . . .	128
5.3.5	Trans-Gaussian Kriging . . . . .	129
5.4	Numerical Analysis . . . . .	131
5.4.1	Interpolating Daily Precipitations . . . . .	132

5.4.2	Interpolating Precipitation <i>MFP</i> and <i>CDD</i> . . . . .	137
5.5	Concluding Remarks . . . . .	143
<b>6</b>	<b>Conclusion and Future Research</b>	<b>145</b>
	<b>References</b>	<b>148</b>
	<b>APPENDICES</b>	<b>161</b>
<b>A</b>	<b>Relevant Properties of Multivariate Normal Distributions and Multivariate <i>t</i>-Distributions</b>	<b>162</b>
A.1	Normal-Normal Hierarchy . . . . .	162
A.2	Normal-Gamma Hierarchy . . . . .	163
<b>B</b>	<b>Derivation of the EM Algorithm</b>	<b>164</b>
B.1	The Complete Log-Likelihood . . . . .	164
B.2	Expectations in the E-Step . . . . .	165
B.2.1	$\langle u_t \rangle$ (2.22) . . . . .	165
B.2.2	$\langle \log u_t \rangle$ (2.23) . . . . .	166
B.2.3	$\langle z_t \rangle$ (2.24) . . . . .	166
B.2.4	$\langle u_t z_t \rangle$ (2.25) . . . . .	167
B.2.5	$\langle u_t z_t^2 \rangle$ (2.26) . . . . .	167
B.3	Updating Formulas in the M-Step . . . . .	168
<b>C</b>	<b>Theoretical Properties of the Iterative SVD Algorithm</b>	<b>169</b>
C.1	Convergence of the Iterative SVD Algorithm . . . . .	170
C.2	The Iterative SVD Algorithm Minimizes the Target Loss Function . . . . .	171

# List of Figures

2.1	Estimates of $a_x$ (top), $b_x$ (middle) and $k_t$ (bottom) obtained from actual US mortality experience over 1940-2019 (left) and 1970-2019 (right) . . . . .	30
2.2	Estimates of $\mathbf{a}$ , $\mathbf{b}$ and $\mathbf{k}$ based on US mortality experience in 1970-2019, before and after a synthetic outlier is added to 1970-1972, produced by SVD (top), Poisson MLE (middle), and $t$ -PPCA (bottom). . . . .	34
3.1	Parameter estimates derived from the E&W male dataset, RH-MLE and RH-LS. . . . .	66
3.2	Parameter estimates derived from the US male dataset, RH-MLE and RH-LS. . . . .	66
3.3	Poisson maximum likelihood estimates of the parameters in the Renshaw-Haberman model for different tolerance levels: $10^{-6}$ , $10^{-7}$ and $10^{-8}$ ; US male dataset. . . . .	70
3.4	Least squares estimates of the parameters in the Renshaw-Haberman model for different tolerance levels: $10^{-6}$ , $10^{-7}$ and $10^{-8}$ ; US male dataset. . . . .	70
4.1	Parameter estimates of the RH model fitted to the E&W male dataset. . . . .	94
4.2	Parameter estimates of the generalized RH model ( $m = 2$ ) fitted to the E&W male dataset. . . . .	95
4.3	Heat-maps of residuals for different mortality models fitted to the E&W male dataset. . . . .	99

4.4	Bootstrapped parameters of the standard RH model fitted to the E&W dataset. The shaded areas represent the 50%, 80% and 95% prediction intervals, respectively. . . . .	104
4.5	Forecast of the period effects $b_x^{(1)}$ and cohort effects $\gamma_{t-x}$ of the standard RH model fitted to the E&W dataset. The light and dark shaded areas represent the 80% and 95% prediction intervals. . . . .	107
5.1	Daily precipitations $P$ (mm) from the NOAA data set on selected days in 1993 . . . . .	120
5.2	Daily maximum temperatures $T$ ( $^{\circ}$ F) from the NOAA data set on selected days in 1993 . . . . .	121
5.3	Histograms of daily precipitations $P$ (mm) and daily maximum temperatures $T$ ( $^{\circ}$ F) in different seasons . . . . .	122

# List of Tables

2.1	Average relative changes in the $t$ -PPCA estimates of $\mathbf{a}$ and $\mathbf{b}$ arising from shocks to initial values for the EM algorithm . . . . .	31
2.2	US COVID-19 deaths in 2020 . . . . .	33
2.3	Average RMAE and RRMSE of $\hat{a}$ produced by SVD, Poisson MLE and $t$ -PPCA for different outlier durations. . . . .	38
2.4	Average RMAE and RRMSE of $\hat{b}$ produced by SVD, Poisson MLE and $t$ -PPCA for different outlier durations. . . . .	39
2.5	Average RMAE and RRMSE of $\hat{k}$ produced by SVD, Poisson MLE and $t$ -PPCA for different outlier durations. . . . .	40
3.1	$L^2$ errors, log-likelihood values, and computation times for RH-MLE, RH-LS and RH-MLE-HV, based on E&W male and US male datasets. . . . .	65
3.2	Computation times for RH-MLE, RH-LS and RH-MLE-HV, based on all of the ten datasets under consideration, with an age range of 60-89. . . . .	68
3.3	Computation times for RH-MLE, RH-LS and RH-MLE-HV, based on all of the ten datasets, with an age range of 0-89. . . . .	68
3.4	$L^2$ errors, log-likelihoods, and computing times for RH-MLE and RH-LS when three different tolerance levels are used, US male. . . . .	71
3.5	$L^2$ errors, log-likelihoods, and computation times for H1-MLE, H1-MLE-HV, H1-LS, and H1-LS-HV, E&W male and US male datasets. . . . .	72

3.6	Computation times for H1-MLE, H1-MLE-HV, H1-LS, and H1-LS-HV, based on all of the ten datasets under consideration. . . . .	74
3.7	Standard errors of selected parameter estimates for RH-MLE, RH-LS and RH-MLE-HV, based on the EW male dataset. . . . .	77
3.8	Standard errors of selected parameter estimates for RH-MLE, RH-LS and RH-MLE-HV, based on the US male dataset. . . . .	78
4.1	Summary of parameters, constraints and the effective number of parameters $\nu$ for the generalized LC, H1, and RH models. . . . .	101
4.2	Summary of log-likelihood $\ell$ , effective number of the parameters $\nu$ , AIC and BIC for the six fitted single-population models. . . . .	102
4.3	Summary of log-likelihood $\ell$ , effective number of the parameters $\nu$ , AIC and BIC for the six fitted multi-population models. . . . .	113
5.1	Variables defined in Section 5.2 . . . . .	123
5.2	Root mean squared errors (RMSE) and mean absolute errors (MAE) in the 10-fold cross-validations for different spatial interpolation methods applied to daily precipitations $P$ . . . . .	133
5.3	Average sample skewness and kurtosis for daily precipitations $P$ , transformed daily precipitations $\sqrt[3]{P}$ , and daily maximum temperatures $T$ . . . . .	136
5.4	Root mean squared errors (RMSE) and mean absolute errors (MAE) calculated from the cross-validations of the spatial interpolations for $MFP$ , implemented with the direct and two-stage approaches and different spatial interpolation methods. . . . .	139
5.5	Root mean squared errors (RMSE) and mean absolute errors (MAE) calculated from the cross-validations of the spatial interpolations for $CDD$ , implemented with the direct and two-stage approaches and different spatial interpolation methods. . . . .	140
5.6	Sample skewness and kurtosis for $MFP$ , $CDD$ and their transformed values. . . . .	142

# Chapter 1

## Introduction

### 1.1 Background: The Rising Importance of Longevity and Climate Risks in Actuarial Science

The importance of both longevity and climate risk in actuarial science has seen a significant increase in recent times, in response to consistent trends in rising life expectancy and the accelerating impact of climate change. The Society of Actuaries (SOA) frequently publishes research projects and reports focusing on risk management<sup>1</sup>, and their current offerings include two vital subjects: Climate and U.S. population. The clear emphasis on climate and longevity risk shows how important these risks have become in actuarial science. It also encourages professionals and researchers to study these risks closely and come up with financial strategies to protect against them.

Longevity risk, which refers to the financial risk related to increasing life expectancy, has tremendous implications for pension funds and life insurance companies. When people live longer than anticipated, the implications extend to longer payout periods for annuities and a surge in benefit payments for pension plans (Blake and Burrows, 2001). It is the quantifiable impact of these unexpected increases in life expectancy that truly emphasizes the importance of managing longevity risk.

---

<sup>1</sup><https://www.soa.org/research/topics/research-emerging-topics/>

The large scale of longevity risk exposure clearly shows the seriousness of the problem. Biffis and Blake (2014) estimated that this exposure amounts to approximately \$25 trillion (USD) for pension funds and insurance companies combined. Just to give an illustration of the scale, according to the World Bank data, the Gross World Product in 2019 was around \$87.55 trillion (USD), which means that the longevity risk exposure stands at almost a third of the global economic output. This comparison not only emphasizes the significant size of this exposure but also its critical role in the global economy.

Moreover, longevity risk has been drawing more attention due to the steady increase in life expectancy. For instance, a report<sup>2</sup> by the United Nations (2019) noted that global average life expectancy increased from 67 years in 2005 to 72.6 years in 2019 and is projected to reach 77 years by 2050. The same report states that the number of people aged 60 years and above is expected to double by 2050 and triple by 2100, rising from 962 million globally in 2017 to 2.1 billion in 2050 and 3.1 billion in 2100. This upward trend implies that the challenges of managing longevity risk will grow, making its management even more important.

Given the importance and the complexity of longevity risk, various statistical tools and techniques have been developed to understand and manage it. Stochastic mortality modelling is a key tool that allows us to measure and manage longevity risk. It offers a statistical structure that can capture the uncertainties of mortality rates over time. Historically, two main families of stochastic mortality models have dominated the field. The first is the seminal Lee-Carter model (Lee and Carter, 1992), and it provides a simple yet effective way to account for unpredictable changes in mortality rates over time. Despite its straightforward structure, the model has proven useful for capturing basic trends in mortality rates. The second is the Cairns-Blake-Dowd (CBD) model (Cairns et al., 2006), which offers a two-factor structure, allowing it to capture more complex patterns in age and period effects on mortality. This added flexibility makes the CBD model particularly useful in understanding how mortality rates improve over time. In recent years, the field of actuarial science has started to explore the potential of modern machine learning techniques like neural networks in mortality modelling, for example, Hainaut (2018); Ni-

---

<sup>2</sup>World Population Ageing 2019: <https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf>

gri et al. (2019); Richman and Wüthrich (2021). These advanced computational methods offer a flexible and powerful tool for capturing complex patterns and non-linear relationships in mortality data. The adoption of such methods represents a new direction in the field, enhancing our ability to predict mortality rates accurately and providing a deeper understanding of longevity risk.

Just as longevity risk is a significant challenge, climate risk has become a major concern in actuarial science. Climate risk refers to the potential financial losses caused by changes in climate patterns, such as temperature, precipitation, wind patterns, and severe weather events. Climate risk is important because it affects not only the insurance industry, but also many other parts of society like agriculture, real estate, public health, and infrastructure. The actuarial science field is giving increasing attention to climate risk, as can be seen from the numerous reports and studies published by the Society of Actuaries (SOA) on this topic. Highlighting this focus, the SOA has even developed a measure called the Actuaries Climate Index (ACI), which serves as a gauge of climate-related extremes and changes over recent decades. This tool reflects the SOA's commitment to providing robust and practical resources for understanding and managing climate risk.

Similar to longevity risk, climate risks are systematic, meaning they affect broad sections of society and the economy, and cannot be easily diversified. From an actuarial point of view, managing climate risk involves understanding its potential financial impacts and devising mechanisms to transfer these risks. One such risk-transfer mechanism is weather derivatives. These are financial instruments that provide compensation based on specific weather events or conditions. The payouts of weather derivatives are determined based on objective, third-party weather data. They offer a means to hedge against weather risks that traditional insurance policies might not cover, thus providing an additional layer of financial protection against the unpredictable impacts of climate change.

Modelling climate risk indeed presents a significant challenge, mainly due to the nature of climate data. Climate patterns involve extreme values and exhibit complex spatio-temporal dependencies. For example, a severe storm or an unusually hot summer can cause massive financial losses, and these types of extreme events are hard to predict because they do not follow normal distribution patterns. Similarly, climate phenomena are highly dependent on both location and time. Temperature patterns, for instance, can vary greatly

from one geographical area to another and from one season to the next. Thus, it is crucial to develop robust models that can account for these unique characteristics of climate data.

Extreme Value Theory (EVT) is a statistical approach commonly used to address the issue of extreme values in climate data. EVT focuses on the tail behaviour of probability distributions, allowing us to estimate the likelihood of extreme events that are outside the range of ordinary observations. The use of EVT in climate risk modelling can help insurers and other stakeholders understand the potential financial impact of extreme weather events and inform their risk management strategies.

In addition to EVT, spatio-temporal modelling plays a critical role in handling the complex dependencies in climate data. Spatio-temporal models are capable of capturing the relationships of climate variables across space and over time, which is particularly useful when considering the impacts of climate change. This can help actuaries understand how risks might correlate or spread across different geographical areas and time periods, which is important for pricing insurance products and managing risk.

In conclusion, while modelling climate risk poses considerable challenges due to the nature of climate data, statistical approaches like EVT and spatio-temporal modelling provide powerful tools to tackle these challenges. Actuarial science, with its rigorous mathematical and statistical foundations, is well-positioned to take on the task of climate risk modelling and contribute to the broader efforts to manage the financial impacts of climate change.

## **1.2 Objectives and Outline of the Thesis**

The main goal of this thesis is to deeply study several statistical methods for modelling longevity and climate risks. The thesis contains four research papers, each contributing a unique methodology or new application to the existing body of knowledge.

In Chapter 2, we propose a novel robust method for estimating parameters for the Lee-Carter model and its variants. The traditional approach of acquiring parameter estimates through singular value decomposition (SVD) is vulnerable to outlier events, such as pandemics, with Covid-19 serving as a notable example. To begin, we uncover the underlying

relationship between the Lee-Carter model and probabilistic principal component analysis (PPCA), a version of principal component analysis (PCA) with a probabilistic framework, is a statistical machine learning tool that has found wide applications across various fields including pattern recognition. Next, we introduce our proposed method, which formulates the Lee-Carter model by PPCA along with multivariate  $t$ -distributions, a type of distributions frequently used to handle outliers. We also derive an EM algorithm to facilitate an efficient implementation of this technique. We further demonstrate the flexibility of our method by discussing its applicability to several extensions of the Lee-Carter model designed for multi-population mortality modelling. We wrap up this chapter by a series of numerical studies based on U.S. mortality data. These studies illustrate that our proposed method provides more robust parameter estimates compared to traditional approaches.

In Chapter 3, we extend our investigation of parameter estimation within the Lee-Carter model family by focusing on the Renshaw-Haberman model, a cohort-based expansion of the Lee-Carter model. Traditionally, the Renshaw-Haberman model is set up as a Poisson regression, and its parameters are estimated using an iterative Newton-Raphson algorithm, known for its slow convergence rate and high computational demands. In this chapter, we propose a new approach to parameter estimation. Our method directly minimizes the sum of squared errors between the observed and predicted log mortality rates, which is in line with the PCA interpretation of the standard Lee-Carter model. First, we introduce an alternating minimization scheme to tackle the main optimization problem. Then, we demonstrate how the challenging part of this problem - estimating the cohort parameters - can be formulated as a PCA with missing values, and we develop an iterative SVD algorithm to solve it. The chapter concludes with numerical studies using mortality data from England and Wales (EW) and the U.S., where we compare our proposed method with the traditional approach. The comparison illustrates the improved efficiency and comparable accuracy of our approach in estimating the parameters of the Renshaw-Haberman model.

Building on the methodological advancements in Chapter 3, where we addressed the computational challenges of the Renshaw-Haberman model, Chapter 4 focuses on translating these insights into a practical and user-friendly software implementation. We develop the `RHa1s` R package, which employs the alternating least squares approach to efficiently fit the Renshaw-Haberman model and its extensions. By leveraging the alternating least

squares algorithm developed in Chapter 3, the package significantly reduces computational time compared to traditional methods. In addition to improving estimation efficiency, the `RHaIs` package supports flexible model specifications, including multi-population settings, and integrates seamlessly with existing tools for mortality forecasting and analysis. This chapter outlines the design, functionalities, and practical applications of the package, demonstrating its utility through numerical experiments using real-world mortality data.

In Chapter 5, we move our focus onto climate risk modelling, with a special emphasis on managing spatial basis risk in precipitation index insurance. We begin by describing and visualizing our dataset, sourced from the US National Oceanic and Atmospheric Administration National Climatic Data Center. We draw attention to the unique characteristics of precipitation variables - right-skewness and non-normality. These properties add to the modelling and forecasting challenges, making it more complicated than dealing with temperature variables, a more commonly studied area. Next, we detail several spatial interpolation techniques, from simpler methods like the nearest neighbor approach to more advanced techniques like trans-Gaussian kriging. We apply these methods to our dataset, and the results reveal that trans-Gaussian kriging provides the most precise spatial interpolations. This finding underlines the importance of considering normality when carrying out kriging for precipitation-related factors. We also explore the practical aspects of spatially interpolating precipitation indexes. We focus on two commonly used indexes - annual consecutive dry days and maximum five-day precipitation in one month, both of which are components of the Actuaries Climate Index. Interestingly, our findings suggest that for the purpose of spatially interpolating such indexes, direct interpolation could deliver better results with lower computational demand, eliminating the need to interpolate the daily raw data. This surprising result can be traced back to the distinct statistical properties of the precipitation indexes and the underlying precipitation variable.

## Chapter 2

# Robust Parameter Estimation for the Lee-Carter Family: A Probabilistic Principal Component Approach

### 2.1 Introduction

Introduced in 1992, the Lee-Carter model (Lee and Carter, 1992) has been one of the most widely used stochastic mortality models in actuarial studies (Deaton et al., 2001). Over the years, the model has been generalized to accommodate different features of mortality dynamics, such as jumps (Chen and Cummins, 2010; Chen, 2013; Deng et al., 2012; Liu and Li, 2015), trend changes (Coelho and Nunes, 2011; Li et al., 2011), and cohort effects (Renshaw and Haberman, 2006). It has also been extended to versions that are designed for specific purposes, such as modelling mortality dynamics of multiple related populations in tandem (Kleinow, 2015; Li and Lee, 2005; Zhou et al., 2013).

Given that the Lee-Carter model has a remarkable popularity, its estimation and statistical properties have attracted considerable attention. Shortly after the model was published, Wilmoth (1993) developed a Poisson maximum likelihood approach to estimate the model. His work was then complemented by Brouhns et al. (2002), who introduced

a parallel Poisson log-bilinear regression approach to fitting the model. With an aim to enhance goodness-of-fit, Renshaw and Haberman (2003) extended the estimation method in the original work of Lee and Carter (1992) to incorporate multiple principal components. Over-dispersion that is associated with heterogeneity within the population of individuals being modelled was studied by Delwarde et al. (2007) and Li et al. (2009), who proposed using a negative binomial distributional assumption to ameliorate the problem. Contrary to the conventional Lee-Carter implementation, in which estimation and forecasting are executed in separate stages, Pedroza (2006) proposed a Bayesian method whereby mortality forecasts are obtained directly from the predictive posterior distribution. Li et al. (2019) further extended the work of Pedroza (2006) to permit Lee-Carter estimation when the dataset in question involves missing values. Recent years have also seen a number of studies on the inference consistency for the Lee-Carter model (Liu et al., 2019; Li et al., 2021).

While the aforementioned studies addressed various statistical attributes, robustness of the Lee-Carter model remains relatively unexplored. In the context of stochastic mortality modelling, the notion of robustness was first raised by Cairns et al. (2009), who argued that robustness of parameter estimates is an important model selection criterion. In their quantitative comparison of stochastic mortality models, they particularly considered the models' robustness relative to changes in the range of data employed, measured by the extent to which the resulting parameter estimates change when the period of data used to fit a given model is extended or shrunk.

In this chapter, we consider robustness to outliers, which refers to the ability of a statistical method or estimator to remain relatively unaffected by extreme values in the data. This definition of robustness is closely related to that used by Cairns et al. (2009), as any change in the range of data employed may affect the number outliers included in the fit. In the context of our study, outliers are defined as exceptionally high (or low) mortality that may arise due to infrequent but impactful events such as pandemics or wars. In the language of time-series outlier analysis (Chen and Liu, 1993), the outliers considered in this study are Additive Outliers (AO) and Temporary Change (TC), both of which are short-term in nature.<sup>1</sup> They also resemble the transitory mortality jumps studied by Chen

---

<sup>1</sup>An AO affects only one single observation. A TC affects a series at a given time, and its effect decays

and Cummins (2010), Chen (2013), Deng et al. (2012), Liu and Li (2015) and Zhou et al. (2013). Robustness to such outliers is highly relevant to models like the Lee-Carter that are devised for making long-term mortality forecasts, as long-term projections should not be affected by short-term anomalies in the data sample.

In previous studies of outliers in Lee-Carter mortality forecasting, the treatment of outliers is notably performed *after* the Lee-Carter model is estimated. In more detail, the original Lee-Carter model is fitted to historical data using a conventional estimation method; then an outlier analysis is applied to the fitted values of the time-varying parameters (typically denoted by  $k_t$ ), and a suitably adapted time-series process is prescribed for the dynamics of  $k_t$ . For instance, Li and Chan (2005, 2007) applied a systematic outlier detection and re-estimation process for several series of  $k_t$  that are obtained by fitting the Lee-Carter model to data from UK, US, Canada, and Scandinavian countries, through the singular value decomposition (SVD) estimation method that was used in the original work of Lee and Carter (1992). Another example is the work of Wang et al. (2011) in which extreme values in  $k_t$  are accommodated by replacing Gaussian innovations in the time-series process of  $k_t$  with heavy-tailed ones. Chen and Cummins (2010), Chen (2013), Deng et al. (2012), Liu and Li (2015) and Zhou et al. (2013) also handled transitory mortality jumps in the Lee-Carter framework in a similar manner. Although these studies offer invaluable insights about the frequency and severity distributions of mortality outliers, they completely overlook the possible influence of mortality outliers on the age-response to  $k_t$  (measured by age-specific parameters, typically denoted by  $b_x$ ). This limitation is significant, because the primary purpose of the Lee-Carter model is to produce long-term mortality forecasts, and any distortion in  $b_x$  can affect such forecasts.

To fill this research gap, in this chapter we investigate robustness to outliers for the Lee-Carter model. Specifically, we propose a robust parameter estimation method for the model, which is developed using a probabilistic principal component analysis (PPCA) model with multivariate  $t$ -distributions. The method is motivated by the fact that the SVD estimation method used in the original work of Lee and Carter (1992) is equivalent to a principal component analysis (PCA), which is well-known for its lack of robustness since the structure of the sample covariance matrix can be profoundly influenced by extreme values exponentially.

ues. First proposed by Tipping and Bishop (1999), PPCA reformulates the conventional non-parametric PCA framework as a parametric Gaussian latent model, wherein the principal components can be exactly recovered by maximum likelihood estimation (MLE). Surprisingly, the PPCA formulation of the Lee-Carter model bears great resemblance to its state-space representation, which naturally inspires us to improve the robustness of Lee-Carter estimation by robustifying the PPCA. The first robust PPCA model was proposed by Archambeau et al. (2006), who replaced the Gaussian structure in PPCA with multivariate  $t$ -distributions. Recently, Guo and Bondell (2023) introduced a more general formulation of multivariate  $t$ -distribution-based PPCA and the corresponding Monte-Carlo expectation-maximization algorithm. Our proposed method draws on the contributions of Guo and Bondell (2023).

In terms of theoretical contributions, we first contribute a PPCA representation of the original work of Lee and Carter (1992). Through this representation, we develop a  $t$ -PPCA estimation method for the Lee-Carter model, which intends to enhance the model’s robustness to outliers. Then, we derive an efficient expectation-maximization (EM) algorithm for implementing the proposed  $t$ -PPCA method. Finally, we extend the proposed  $t$ -PPCA method to two multi-population extensions of the Lee-Carter model: the augmented common factor model (Li and Lee, 2005) and the common age effect model (Kleinow, 2015).

To evaluate the performance of the proposed robust PPCA-based Lee-Carter model, we design real data experiments and simulations using data from the Human Mortality Database (2023). In one experiment, the proposed method is applied to US mortality data over different calibration windows, and compared against two conventional estimation methods: SVD and Poisson maximum likelihood. In another experiment, we create hypothetical outliers by referencing to excess COVID deaths in the US. These hypothetical outliers are then used to generate pseudo datasets for testing the performance of the proposed method in a wide range of scenarios. The results of both experiments suggest that our proposed estimation method outperforms the conventional SVD and Poisson maximum likelihood methods in terms of robustness to outliers.

The remainder of the chapter is organized as follows. Section 2.2 offers a brief review of the Lee-Carter model and its conventional estimation methods. Section 2.3 describes

the standard PPCA and its connection to the Lee-Carter model. Drawing on the foundation established in Section 2.3, Section 2.4 outlines the proposed method and details the EM algorithm for its implementation. Section 2.5 extends our proposed method to multi-population settings. Section 2.6 presents two numerical experiments that compare the performance of our proposed method with conventional estimation methods. Finally, concluding remarks are provided in Section 2.7.

## 2.2 Background

In this section, we briefly review the Lee-Carter model and its conventional estimation methods, SVD and Poisson MLE. The importance of robustifying the original Lee-Carter model is then highlighted.

### 2.2.1 The Lee-Carter Model

We let  $y_{x,t} := \log(m_{x,t})$  be the log central rate of death for age group  $x$  in calendar year  $t$ . Throughout this chapter, it is assumed that the model is estimated to data sets that span  $p$  age groups  $x_1, \dots, x_p$ , and  $n$  calendar years  $t_1, \dots, t_n$ .

The Lee-Carter model expresses  $y_{x,t}$  in the following bilinear form:

$$y_{x,t} := \log(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}. \quad (2.1)$$

In this model,  $a_x$  is an age-specific parameter that reflects the average level of the log central rate of death over time for age group  $x$ ,  $k_t$  is a time-varying parameter that describes the overall mortality level in year  $t$ ,  $b_x$  another age-specific parameter that measures the sensitivity of  $y_{x,t}$  with respect to  $k_t$ , and  $\varepsilon_{x,t}$  is the error term.

When the model is fitted to datasets from the developed world, the estimated series of  $k_t$  typically follows a downward trend, which indicates a steady reduction in mortality. In this case, age groups with larger values of  $b_x$  tend to experience more rapid mortality reductions compared to other age groups.

The model parameters cannot be uniquely identified without adequate parameter constraints. In this chapter, we use the following constraints to stipulate parameter uniqueness:

$$\sum_{x=x_1}^{x_p} b_x = 1, \quad \text{and} \quad \sum_{t=t_1}^{t_n} k_t = 0. \quad (2.2)$$

These constraints are used in the original work of Lee and Carter (1992). We use  $\hat{a}_x$ ,  $\hat{b}_x$  and  $\hat{k}_t$  to represent estimates of parameters  $a_x$ ,  $b_x$  and  $k_t$ , respectively.

## 2.2.2 Singular Value Decomposition

Lee and Carter (1992) found the least squares solution to (2.1), which implies  $\hat{a}_x = \bar{y}_x := \sum_{t=t_1}^{t_n} y_{x,t}/n$ . They proposed to estimate the remaining parameters,  $b_x$  and  $k_t$ , by a first-order singular value decomposition (SVD), the solution to which can be interpreted as the first principal component of the matrix of historical log central rates of death.

For convenience, let us define the following vector/matrix notation:

$$\left\{ \begin{array}{l} \mathbf{y}_t := (y_{x_1,t}, \dots, y_{x_p,t})^T \\ \bar{\mathbf{y}} := \sum_{t=t_1}^{t_n} \mathbf{y}_t / n \\ \mathbf{Y} := (\mathbf{y}_{t_1}, \dots, \mathbf{y}_{t_n}) \\ \tilde{\mathbf{Y}} = (\mathbf{y}_{t_1} - \bar{\mathbf{y}}, \dots, \mathbf{y}_{t_n} - \bar{\mathbf{y}}) \\ \mathbf{a} := (a_{x_1}, \dots, a_{x_p})^T \\ \mathbf{b} := (b_{x_1}, \dots, b_{x_p})^T \\ \mathbf{k} := (k_{t_1}, \dots, k_{t_n})^T \\ \boldsymbol{\varepsilon}_t := (\varepsilon_{x_1,t}, \dots, \varepsilon_{x_p,t})^T \end{array} \right. ,$$

where  $\mathbf{y}_t$ ,  $\bar{\mathbf{y}}$ ,  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\boldsymbol{\varepsilon}_t$  are  $p$ -dimensional column vectors,  $\mathbf{k}$  is an  $n$ -dimensional column vector, and  $\mathbf{Y}$  and  $\tilde{\mathbf{Y}}$  are  $p \times n$  matrices. It follows that we can express (2.1) as

$$\mathbf{y}_t = \mathbf{a} + \mathbf{b}k_t + \boldsymbol{\varepsilon}_t. \quad (2.3)$$

As such, we can interpret the original least squares optimization as a minimization of the squared reconstruction errors:

$$\min_{\mathbf{a}, \mathbf{b}, \mathbf{k}} \sum_{x,t} (y_{x,t} - (a_x + b_x k_t))^2 = \min_{\mathbf{a}, \mathbf{b}, \mathbf{k}} \sum_t \|\mathbf{y}_t - (\mathbf{a} + \mathbf{b}k_t)\|_2^2, \quad (2.4)$$

where  $\|\cdot\|_2$  is the Euclidean norm ( $L^2$  norm).

We let  $\hat{\mathbf{a}}$ ,  $\hat{\mathbf{b}}$  and  $\hat{\mathbf{k}}$  be estimates of parameter vectors  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{k}$ , respectively. Standard PCA theory (Bishop and Nasrabadi, 2006) gives the following solution to (2.4):

$$\hat{\mathbf{a}} = \bar{\mathbf{y}}, \quad \hat{\mathbf{b}} = \frac{\mathbf{u}}{\mathbf{1}^T \mathbf{u}}, \quad \hat{\mathbf{k}} = (\mathbf{1}^T \mathbf{u}) \cdot \tilde{\mathbf{Y}}^T \mathbf{u}, \quad (2.5)$$

where  $\mathbf{1} = (1, \dots, 1)^T$  is a  $p$ -dimensional column vector of ones, and  $\mathbf{u}$  is the first left-singular vector of the matrix of mean-centered log central death rates  $\tilde{\mathbf{Y}}$  with unit length  $\|\mathbf{u}\|_2 = 1$ . The normalizing constant  $\mathbf{1}^T \mathbf{u}$  is introduced to satisfy the identifiability constraint  $\sum_{x=x_1}^{x_p} \hat{b}_x = 1$ , or equivalently  $\mathbf{1}^T \hat{\mathbf{b}} = 1$ . It can also be verified easily that  $\hat{\mathbf{k}}$  satisfies the other identifiability constraint  $\sum_{t=t_1}^{t_n} \hat{k}_t$ .

Lee and Carter (1992) suggested that the estimates of  $k_t$  obtained by (2.5) can be adjusted to match the total death count each year. That is, for each  $t = t_1, \dots, t_n$ , the estimate of  $k_t$  is adjusted such that the following equation is satisfied:

$$D_t = \sum_{x=x_1}^{x_p} D_{x,t} = \sum_{x=x_1}^{x_p} \left( N_{x,t} \cdot e^{\hat{a}_x + \hat{b}_x k_t} \right), \quad (2.6)$$

where  $D_t$  is the total number of deaths in year  $t$ ,  $D_{x,t}$  and  $N_{x,t}$  represent the number of deaths and exposures-to-risk for age group  $x$  in year  $t$ , respectively. We can solve equation (2.6) numerically using standard root-finding methods such as a one-dimensional line search. Note that the expression for  $\hat{\mathbf{k}}$  in (2.5) would be redundant should one prefers estimates of  $k_t$  that match aggregate death counts as specified in (2.6).

### 2.2.3 Poisson Maximum Likelihood

The Poisson maximum likelihood estimation (MLE) method for the Lee-Carter model was originally proposed by Wilmoth (1993). In Poisson MLE, it is assumed that the death count in each age-time cell follows a Poisson distribution, with a mean that equals to the expected number of deaths implied by the Lee-Carter model structure. That is,

$$D_{x,t} \sim \text{Poisson}(N_{x,t} m_{x,t}), \quad \text{with } \log(m_{x,t}) = a_x + b_x k_t. \quad (2.7)$$

This distributional assumption leads to the following log-likelihood function:

$$\ell(\mathbf{a}, \mathbf{b}, \mathbf{k}) = \sum_{x=x_1}^{x_p} \sum_{t=t_1}^{t_n} (D_{x,t}(a_x + b_x k_t) - N_{x,t} e^{a_x + b_x k_t}) + \text{constant}, \quad (2.8)$$

which is then maximized to obtain estimates of  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{k}$ . The solution to the optimization problem can be found by using a modified iterative Newton-Raphson method (Goodman, 1979), in which parameters are updated one at a time until convergence is achieved.

## 2.2.4 Forecasting

Once the estimates of  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{k}$  are obtained (by either SVD, Poisson MLE, or other methods), a time-series process is fitted to  $\mathbf{k}$ . Very often, a random walk with drift is used:

$$k_t = k_{t-1} + \theta + e_t, \quad (2.9)$$

where  $\theta$  is the drift term, and  $e_t$  is the innovation term, which is usually assumed to be normally distributed. The values of  $k_t$  are then extrapolated beyond the data sample through the time-series process to obtain forecasts of future central death rates.

This chapter focuses on the estimation of the Lee-Carter parameters, and thus the time-series modelling of  $k_t$  is not discussed. For recent studies of time-series modelling for  $k_t$ , we refer readers to Lin et al. (2015) and Neves et al. (2017).

## 2.2.5 The Need for Robustifying the Lee-Carter Model

The conventional estimation methods described earlier are sensitive to outliers. Huber and Ronchetti (2011) pointed out that the standard PCA (which corresponds to the SVD estimation method) is highly susceptible to outliers (Huber and Ronchetti, 2011). He argued that even a single extreme point may significantly distort the quality of the low-dimensional approximation. Similarly, as noted by Künsch et al. (1989) and Morgenthaler (1992), Poisson maximum likelihood is also sensitive to outliers. Therefore, it can be

anticipated that the Lee-Carter estimates produced by SVD or Poisson MLE are not robust to outliers. This point is demonstrated in the numerical experiments presented in Section 6.

What is the problem with a lack of robustness to outliers? The Lee-Carter model is designed for obtaining *long-term* projections of future mortality. In actuarial practice, *long-term* mortality projections are often expressed in the form of scale factors (see, e.g., Li et al., 2010; Li and Liu, 2020). For a given Lee-Carter model, the scale factor that describes the expected change in log central death rate from year  $t - 1$  to  $t$  ( $t > t_n$ ) is given by

$$b_x \cdot \mathbb{E}[k_t - k_{t-1} | \mathcal{F}_{t_n}], \quad (2.10)$$

where  $\mathcal{F}_{t_n}$  denotes the information up to and including year  $t_n$  (the forecast origin). If the dynamics of  $k_t$  is modelled by (2.9), then the scale factor can be simplified to  $b_x \theta$ .

While the existing methods for handling mortality outliers in the Lee-Carter framework address the impact of outliers on  $k_t$ , i.e., the expectation in (2.10), they completely ignore how mortality outliers may affect  $b_x$  as they are applied to the estimates of  $k_t$  *after* the model is fitted. Nevertheless, the impact of  $b_x$  cannot be overlooked, as they have critical influence on the distribution of the expected mortality improvement over age. In other words, if the estimates of  $b_x$  are affected by outliers, long-term mortality rates for some age groups may be over-(under-)estimated.

## 2.3 Standard PPCA

In this section, we describe the standard PPCA as well as its connection to the standard PCA and the conventional SVD estimation method for the Lee-Carter model. The materials presented in this section form a foundation upon which our proposed method, robust multivariate  $t$ -PPCA, is developed.

### 2.3.1 Basic Formulation

The standard PPCA with one principal component is the most relevant to the Lee-Carter model, and therefore in the following presentation we focus on this special case. For more general descriptions of the standard PPCA, we refer readers to Tipping and Bishop (1999) and Bishop and Nasrabadi (2006).

As established in Section 2.2.1, estimating  $\mathbf{a}$  and  $\mathbf{b}$  in the Lee-Carter model is equivalent to solving a PCA, which can be achieved via a SVD as outlined in (2.5). Specifically,  $\hat{\mathbf{a}} = \bar{\mathbf{y}}$  is the average log central death rates over the calibration window, and  $\hat{\mathbf{b}} = \mathbf{u}/\mathbf{1}^T\mathbf{u}$  is a normalized first left-singular vector of the matrix of centered log central rates of death  $\tilde{\mathbf{Y}}$ .

Interestingly, this SVD solution can also be expressed as a maximum likelihood estimate of a probabilistic latent variable model. This alternative method, known as the standard PPCA, was first introduced by Tipping and Bishop (1999). It is based on the following probabilistic model:

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{a}, \mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I}), \quad (2.11)$$

where  $\mathbf{y}_t$ ,  $\mathbf{a}$  and  $\mathbf{b}$  are all  $p \times 1$  vectors as defined in Section 2.2.1, and  $(\mathbf{a}, \mathbf{b}, \sigma^2)$  are the collection of model parameters. It can be shown that the MLE of  $(\mathbf{a}, \mathbf{b}, \sigma^2)$  in (2.11) has the following closed form solution:

$$\hat{\mathbf{a}} = \bar{\mathbf{y}}, \quad \hat{\mathbf{b}} = \mathbf{u}\sqrt{\lambda_1 - \hat{\sigma}^2}, \quad \hat{\sigma}^2 = \frac{1}{p-1} \sum_{i=2}^p \lambda_i, \quad (2.12)$$

where  $\bar{\mathbf{y}}$  is the average log central rates of death over the calibration window,  $\mathbf{u}$  is the first eigenvector of the sample covariance matrix  $\mathbf{S} := \tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T/n$ , and  $\lambda_i$  is the  $i$ -th largest eigenvalue of  $\mathbf{S}$ . We refer readers to Tipping and Bishop (1999) for a proof of (2.12).

Importantly,  $\hat{\mathbf{b}} = \mathbf{u}\sqrt{\lambda_1 - \hat{\sigma}^2}$  in (2.12) and  $\hat{\mathbf{b}} = \mathbf{u}/(\mathbf{1}^T\mathbf{u})$  in (2.5) are both constant multiples of  $\mathbf{u}$ , with  $\sqrt{\lambda_1 - \hat{\sigma}^2}$  and  $1/(\mathbf{1}^T\mathbf{u})$  as the scaling factors, respectively. In other words,  $\hat{\mathbf{b}}$  derived from the standard PPCA in (2.12) spans the same one-dimensional principal subspace as its counterpart in the SVD/PCA solution shown in (2.5). This suggests that, after normalizing  $\hat{\mathbf{b}}$  under the identification constraint  $\mathbf{1}^T\hat{\mathbf{b}} = 1$ , the standard PPCA solution (2.12) provides the same estimates of  $\mathbf{a}$  and  $\mathbf{b}$  as the SVD/PCA solution (2.5) does.

Notably, this alternative view of the original SVD/PCA solution to the Lee-Carter estimation problem enables us to improve the quality of estimation. In particular, the probabilistic framework allows us to target specific statistical properties, such as robustness to outliers, which the conventional PCA/SVD method cannot provide.

### 2.3.2 Formulation via a Latent Variable Structure

There exists an equivalent formulation of the probabilistic model specified in (2.11) via a latent variable structure:

$$\mathbf{y}_t | z_t \sim \mathcal{N}(\mathbf{a} + \mathbf{b}z_t, \sigma^2 \mathbf{I}), \quad z_t \sim \mathcal{N}(0, 1), \quad (2.13)$$

where  $z_t$  represents an unobserved latent variable that follows a standard normal distribution. Marginalizing out the latent variable  $z_t$  recovers the marginal distribution of  $\mathbf{y}_t$  described in (2.11). The equivalence can be proved straightforwardly by applying the conditioning and marginalization properties of multivariate normal distributions, as laid out in (A.1) and (A.2) in Appendix A. Readers are directed to Appendix A for additional details about the properties of multivariate normal distributions, which are frequently referenced in Section 2.4 and Appendix B.

The latent model structure in (2.13) may seem redundant, as  $z_t$  is neither an observed variable nor a model parameter. Nevertheless, the conditional structure lays the groundwork for developing an expectation-maximization (EM) algorithm for solving the MLE. The EM algorithm is especially useful when it is impossible to solve the MLE directly, a situation that may arise in some extensions of the PPCA including the one we adopt in our proposed method.

The PPCA estimates  $\mathbf{a}$  and  $\mathbf{b}$  in the Lee-Carter model. Having obtained estimates of  $\mathbf{a}$  and  $\mathbf{b}$ , estimates of  $\mathbf{k}$  can be obtained via death count matching using (2.6).

It is important to understand the role of “normality” in the application of the standard PPCA to the Lee-Carter model. Although, as shown in (2.11), the standard PPCA assumes that log central rates of death are normally distributed, whether normality is strictly satisfied is not the main point, as our ultimate goal is not to fit a multivariate normal

model to the data. Our true aim here is to obtain Lee-Carter parameter estimates that are identical to the PCA/SVD estimates within a more flexible probabilistic framework, so that a generalization that suits our needs can be developed using the framework. A similar logic is seen in the context of linear regression modelling: ordinary least squares estimates of a linear regression do not depend on any specific distributional assumption; yet, they happen to be identical to the ML estimates of the regression model when normality is assumed.

## 2.4 Proposed Method

The standard PPCA is a mere reformulation of the PCA/SVD estimation method. Hence, it does not address the lack of robustness to outliers.

One way to enhance the robustness of the standard PPCA is to replace the marginal distribution in the latent variable structure from multivariate normal to multivariate  $t$ , a distribution that is widely used in robust statistical modelling problems; for instance, maximum likelihood estimates from probabilistic models involving  $t$ -distributions typically exhibit stronger robustness against extreme observations (Lange et al., 1989). Studies conducted by Archambeau et al. (2006), Chen et al. (2009) and Guo and Bondell (2023) have shown, by both theoretical and numerical means, that the PPCA based on a multivariate  $t$ -distribution is more resilient against outliers compared to the standard version. Inspired by these studies, we propose a  $t$ -PPCA method for estimating Lee-Carter parameters, with a goal to improve robustness to outliers.

In what follows, we first define the multivariate  $t$ -distribution that is used in our theoretical work. We then detail the  $t$ -PPCA formulation, with a hierarchical structure that facilitates estimation. Finally, we derive an efficient EM algorithm for practical implementation.

### 2.4.1 Multivariate $t$ -Distributions

Let  $\mathbf{y}$  be a general  $p$ -dimensional random vector following a multivariate  $t$ -distribution:

$$\mathbf{y} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where  $\boldsymbol{\mu}$  is the  $p$ -dimensional mean vector,  $\boldsymbol{\Sigma}$  is the  $p \times p$  symmetric positive definite scale matrix, and  $\nu > 0$  represents the degrees of freedom. The probability density function of  $\mathbf{y}$  is as follows (Kibria and Joarder, 2006):

$$f(\mathbf{y}) = \frac{\Gamma[(\nu + p)/2]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \left[ 1 + \frac{1}{\nu}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right]^{-(\nu+p)/2}, \quad (2.14)$$

where the  $|\boldsymbol{\Sigma}|$  is the determinant of  $\boldsymbol{\Sigma}$ , and  $\Gamma(\cdot)$  denotes the gamma function. When  $\nu$  tends to infinity, the distribution degenerates to the multivariate normal distribution  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . A few properties of multivariate  $t$ -distributions are presented in Appendix A. We draw on these properties in our theoretical work.

### 2.4.2 Model Formulation

The first step in the development of our proposed method is to replace the multivariate normal distribution in the standard PPCA representation of the Lee-Carter model (2.11) with a multivariate  $t$ -distribution. That is, for  $t = t_1, \dots, t_n$ ,

$$\mathbf{y}_t \sim t_\nu(\mathbf{a}, \mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I}). \quad (2.15)$$

The next step is to obtain the ML estimates of  $\mathbf{a}$  and  $\mathbf{b}$ , on the basis of (2.15). To obtain ML estimates, a natural starting point is to write down the probability density function of  $\mathbf{y}$  by combining (2.15) with (2.14):

$$f(\mathbf{y}_t) = \frac{\Gamma[(\nu + p)/2]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \left[ 1 + \frac{1}{\nu}(\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_t - \mathbf{a}) \right]^{-(\nu+p)/2}. \quad (2.16)$$

However, a direct maximization of the log-likelihood from (2.16) is computationally intractable. A strategy that is commonly used in MLE involving multivariate  $t$ -distributions

is to adopt their scale mixture Gaussian representations and then apply an EM algorithm with tractable computation in each iteration (Liu and Rubin, 1995). For the purpose of this study, we propose the following equivalent hierarchical structure, which is then used to derive the EM algorithm in Section 2.4.3.

**Proposition 1.**

$$\mathbf{y}_t \sim t_\nu(\mathbf{a}, \mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I}) \iff \begin{cases} \mathbf{y}_t|z_t, u_t \sim \mathcal{N}\left(\mathbf{a} + \mathbf{b}z_t, \frac{\sigma^2\mathbf{I}}{u_t}\right) \\ z_t|u_t \sim \mathcal{N}\left(0, \frac{1}{u_t}\right) \\ u_t \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \end{cases}, \quad (2.17)$$

where  $z_t$  and  $u_t$  are latent variables.

*Proof.* Proposition 2.1 can be proved by applying the results presented in Appendix A.

First, using (A.1) and (A.2), we can show that  $\mathbf{y}_t|z_t, u_t$  is multivariate normally distributed:

$$\mathbf{y}_t|z_t, u_t \sim \mathcal{N}\left(\mathbf{a} + \mathbf{b}z_t, \frac{\sigma^2\mathbf{I}}{u_t}\right), z_t|u_t \sim \mathcal{N}\left(0, \frac{1}{u_t}\right) \iff \mathbf{y}_t|u_t \sim \mathcal{N}\left(\mathbf{a}, \frac{\mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I}}{u_t}\right). \quad (2.18)$$

Then, using (A.4) and (A.5), we can show that  $\mathbf{y}_t$  is multivariate  $t$ -distributed, which elaborates (2.17):

$$\mathbf{y}_t|u_t \sim \mathcal{N}\left(\mathbf{a}, \frac{\mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I}}{u_t}\right), u_t \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \iff \mathbf{y}_t \sim t_\nu(\mathbf{a}, \mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I}). \quad (2.19)$$

■

Finally, we estimate  $\mathbf{k}$  given the ML estimates of  $\mathbf{a}$  and  $\mathbf{b}$ . This step can be accomplished easily by matching death counts via (2.6).

We conclude this subsection with a remark on the interpretation of the  $t$ -distribution in (2.15). The  $t$ -distribution is employed to obtain ML estimates of  $\mathbf{a}$  and  $\mathbf{b}$  that are more robust to outliers, while still preserving the essential information about the (first)

principal component. The  $t$ -distribution is used *not* on the ground of statistical fit, and hence whether log central death rates strictly follow a multivariate  $t$ -distribution is not a critical issue.

### 2.4.3 The EM Algorithm

As mentioned in the previous subsection, the ML estimates of  $\mathbf{a}$  and  $\mathbf{b}$  can be obtained via an EM algorithm, a method that was first introduced by Dempster et al. (1977). General information about this method can be found in machine learning texts such as those authored by Hastie et al. (2009) and Bishop and Nasrabadi (2006).

In this subsection, we present an EM algorithm for obtaining the ML estimates of  $\mathbf{a}$  and  $\mathbf{b}$  in the proposed  $t$ -PPCA Lee-Carter model specified in (2.15). Instead of maximizing the original log-likelihood function derived from (2.16), the EM algorithm utilizes the hierarchical structure specified in (2.17), in which each conditional distribution is easier to deal with analytically. The presentation in this subsection includes the implementation procedure and key updating formulas only. Detailed derivations of the formulas involved, including (2.21) to (2.30), are provided in Appendix B.

To derive the EM algorithm, we first write down the complete log-likelihood function in terms of the latent variables  $z_t$  and  $u_t$ :

$$L_c = \sum_{t=t_1}^{t_n} \log[f(\mathbf{y}_t, z_t, u_t)] = \sum_{t=t_1}^{t_n} \log[f(\mathbf{y}_t|z_t, u_t)f(z_t|u_t)f(u_t)], \quad (2.20)$$

where  $f(\mathbf{y}_t|z_t, u_t)$ ,  $f(z_t|u_t)$  and  $f(u_t)$  are the probability density functions corresponding to the distributions established in (2.17).

Let  $\langle \cdot \rangle = \mathbb{E}[\cdot|\mathbf{y}_t]$  be the conditional expectation operator. In the E-step, we need to find the conditional expectation of the complete log-likelihood  $\langle L_c \rangle$  conditioning on  $\mathbf{y}_t$ . Substituting the expressions of  $f(\mathbf{y}_t|z_t, u_t)$ ,  $f(z_t|u_t)$ , and  $f(u_t)$  into (2.20), we obtain

$$\begin{aligned} \langle L_c \rangle = & - \sum_{t=t_1}^{t_n} \left[ \frac{p}{2} \log \sigma^2 + \frac{\langle u_t \rangle}{2\sigma^2} (\mathbf{y}_t - \mathbf{a})^T (\mathbf{y}_t - \mathbf{a}) - \frac{1}{\sigma^2} \langle u_t z_t \rangle \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}) \right. \\ & \left. + \frac{1}{2\sigma^2} \mathbf{b}^T \mathbf{b} \langle u_t z_t^2 \rangle - \frac{\nu}{2} \left( \log \frac{\nu}{2} + \langle \log u_t \rangle - \langle u_t \rangle \right) + \log \Gamma \left( \frac{\nu}{2} \right) \right] + \text{constant}. \quad (2.21) \end{aligned}$$

It turns out that all the posterior expectations in (2.21) have analytical forms, making the E-step computationally efficient:

$$\langle u_t \rangle = \frac{\nu + p}{\nu + (\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{y}_t - \mathbf{a})}, \quad (2.22)$$

$$\langle \log u_t \rangle = \psi \left( \frac{\nu + p}{2} \right) - \log \left( \frac{\nu + (\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{y}_t - \mathbf{a})}{2} \right), \quad (2.23)$$

$$\langle z_t \rangle = (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}), \quad (2.24)$$

$$\langle u_t z_t \rangle = \langle u_t \rangle \langle z_t \rangle, \quad (2.25)$$

$$\langle u_t z_t^2 \rangle = \sigma^2 (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} + \langle u_t \rangle \langle z_t \rangle^2, \quad (2.26)$$

where  $\psi(\cdot) = \Gamma(\cdot)/\Gamma'(\cdot)$  is the digamma function.

In the M-step, we maximize  $\langle L_c \rangle$  with respect to  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$ , and  $\nu$ , by setting all of the first order partial derivatives to zero. This results in the following updating equations:

$$\mathbf{a} \leftarrow \frac{\sum_{t=t_1}^{t_n} \langle u_t \rangle (\mathbf{y}_t - \mathbf{b} \langle z_t \rangle)}{\sum_t \langle u_t \rangle}, \quad (2.27)$$

$$\mathbf{b} \leftarrow \left[ \sum_{t=t_1}^{t_n} \langle u_t z_t^2 \rangle \right]^{-1} \left[ \sum_{t=t_1}^{t_n} (\mathbf{y}_t - \mathbf{a}) \langle u_t z_t \rangle \right], \quad (2.28)$$

$$\sigma^2 \leftarrow \frac{1}{np} \sum_{t=t_1}^{t_n} \left[ \langle u_t \rangle (\mathbf{y}_t - \mathbf{a})^T (\mathbf{y}_t - \mathbf{a}) - 2 \langle u_t z_t \rangle \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}) + \mathbf{b}^T \mathbf{b} \langle u_t z_t^2 \rangle \right], \quad (2.29)$$

$$1 + \log \frac{\nu}{2} - \psi \left( \frac{\nu}{2} \right) + \frac{1}{n} \sum_t (\langle \log u_t \rangle - \langle u_t \rangle) = 0. \quad (2.30)$$

The solution to  $\nu$  in (2.30) can be found by using a one-dimensional line search.

The two-stage EM algorithm for finding the ML estimates of  $\mathbf{a}$  and  $\mathbf{b}$  is implemented by alternating the E-step and M-step until convergence. Convergence is defined as the point at which the absolute difference in the log-likelihood between two successive iterations is less than a small threshold, which is set to  $10^{-4}$  throughout this chapter. After convergence, we rescale the estimates of  $\mathbf{b}$  so that the identifiability constraint  $\sum_{x=x_1}^{x_p} \hat{b}_x = 1$  is met. Having obtained estimates of  $\mathbf{a}$  and  $\mathbf{b}$ , an estimate of  $\mathbf{k}$  can be obtained by matching death counts via (2.6). The procedure for estimating the parameters in the robust multivariate  $t$ -PPCA Lee-Carter model is summarized in Algorithm 1.

---

**Algorithm 1** Robust Multivariate  $t$ -PPCA Lee-Carter Estimation

---

1. Initialize  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$ .
  2. Estimate  $\mathbf{a}$  and  $\mathbf{b}$  by the EM algorithm described in Section 2.4.3:
    - (a) E-step: Given the current estimates of parameters  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$ , compute the posterior conditional expectations (2.22)-(2.26) in the E-step.
    - (b) M-step: Update the estimates of  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$  by (2.27)-(2.30) in the M-step.
    - (c) Repeat Step 2(a) and Step 2(b) until convergence.
    - (d) Normalize the estimate of  $\mathbf{b}$  to satisfy the identifiability constraint  $\sum_{x=x_1}^{x_p} \hat{b}_x = 1$ .
  3. Estimate  $\mathbf{k}$  by matching death counts via (2.6).
- 

The algorithm for estimating the  $t$ -PPCA model requires initial values of  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$ . Given that the  $t$ -PPCA model is an extension of the standard PPCA model, in which ML estimates for  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\sigma^2$  can be estimated analytically, it is natural to adopt such estimates as initial values for the algorithm. Following Guo and Bondell (2023), we set the initial value of  $\nu$  to 3 in Algorithm 1.

The updating equations for  $\mathbf{a}$  and  $\mathbf{b}$  provide an insightful explanation as to why the resulting estimates are more robust to outliers. On the one hand, the updating equations for these two parameters, (2.27) and (2.28), can be understood as a weighted average and weighted least squares solution, respectively, with  $u_t$  being the weight on  $\mathbf{y}_t$ . On the other hand, it can be inferred from (2.22) that the updated value of  $u_t$  tends to be small if the observed value of  $\mathbf{y}_t$  is an outlier. In more detail, if the observed value of  $\mathbf{y}_t$  is a potential outlier, then its distance from  $\mathbf{a}$ , which represents the center of  $\mathbf{y}_t$ , tends to be large. As a consequence, the value of  $(\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{y}_t - \mathbf{a})$  in the denominator of (2.22), the updating equation for  $u_t$ , tends to be large as well, thereby resulting in a small updated value of  $u_t$  and thus the aberrant observation of  $\mathbf{y}_t$  is downweighted in the update of  $\mathbf{a}$  and  $\mathbf{b}$ .

## 2.5 Multi-Population Extensions

The proposed method can be applied to a wide range of Lee-Carter extensions that feature age-period interaction parameters. The versatility of the proposed method is demonstrated in this section, which articulates how the proposed method can be applied to two well-known multi-population extensions of the Lee-Carter model.

### 2.5.1 The Augmented Common Factor Model

We first consider the augmented common factor (ACF) model proposed by Li and Lee (2005) for forecasting mortality for multiple populations. Let us suppose that the model is applied to  $I$  populations, and the dataset spans  $p$  age groups and  $n$  calendar years.

The ACF assumes that  $m_{x,t,i}$ , the log central rate of death for individuals in population  $i$  with an age of  $x$  in calendar year  $t$ , is given by

$$y_{x,t,i} := \log(m_{x,t,i}) = a_{x,i} + \underbrace{b_x k_t}_{\text{common factor}} + \underbrace{b_{x,i} k_{t,i}}_{\text{population-specific factor}} + \varepsilon_{x,t,i}, \quad (2.31)$$

where  $a_{x,i}$ ,  $b_{x,i}$  and  $k_{t,i}$  are parameters that are specific to population  $i$ ,  $b_x$  and  $k_t$  are parameters that are shared by all  $I$  populations, and  $\varepsilon_{x,t,i}$  is the error term.

We let

$$\left\{ \begin{array}{l} \mathbf{y}_{t,i} := (y_{x_1,t,i}, \dots, y_{x_p,t,i})^T \\ \mathbf{a}_i := (a_{x_1,i}, \dots, a_{x_p,i})^T \\ \mathbf{b} := (b_{x_1}, \dots, b_{x_p})^T \\ \mathbf{b}_i := (b_{x_1,i}, \dots, b_{x_p,i})^T \\ \mathbf{k} := (k_{t_1}, \dots, k_{t_n})^T \\ \mathbf{k}_i := (k_{t_1,i}, \dots, k_{t_n,i})^T \\ \boldsymbol{\varepsilon}_{t,i} := (\varepsilon_{x_1,t,i}, \dots, \varepsilon_{x_p,t,i})^T \end{array} \right. , \quad (2.32)$$

and express (2.31) in vector form as follows:

$$\mathbf{y}_{t,i} := \log(\mathbf{m}_{t,i}) = \mathbf{a}_i + \mathbf{b}k_t + \mathbf{b}_i k_{t,i} + \boldsymbol{\varepsilon}_{t,i}. \quad (2.33)$$

In using the ACF model, parameters  $\mathbf{a}_i$ ,  $\mathbf{b}$ ,  $\mathbf{b}_i$ ,  $\mathbf{k}$  and  $\mathbf{k}_i$  in equation (2.33) are estimated from historical data, and then the estimated values of  $\mathbf{k}$  and  $\mathbf{k}_i$  are fitted to appropriate time-series processes, which are extrapolated to obtain forecasts of future mortality.

In the original work of Li and Lee (2005), the estimation of parameters  $\mathbf{a}_i$ ,  $\mathbf{b}$ ,  $\mathbf{b}_i$ ,  $\mathbf{k}$  and  $\mathbf{k}_i$  is performed with the following procedure:

1. For each population  $i = 1, \dots, I$ , set the estimate of  $\mathbf{a}_i$  to the average log central rates of death for population  $i$  over the calibration window:

$$\hat{\mathbf{a}}_i = \bar{\mathbf{y}}_i := \frac{1}{n} \sum_{t=t_1}^{t_n} \mathbf{y}_{t,i}. \quad (2.34)$$

The vector of centered log central rates of death for population  $i$  in calendar year  $t$  is also computed:  $\tilde{\mathbf{y}}_{t,i} = \mathbf{y}_{t,i} - \bar{\mathbf{y}}_i$ .

2. Estimate the common factor  $\mathbf{b}$  and  $\mathbf{k}$  by applying a first-order SVD to the following matrix:

$$\tilde{\mathbf{Y}}^c = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_n), \quad (2.35)$$

where  $\tilde{\mathbf{y}}_t := \sum w_i \tilde{\mathbf{y}}_{t,i}$  represents the weighted average of centered log central rates of death over the populations under consideration, with  $w_i$  being the weight on population  $i$ , determined on the basis on the size of population  $i$  relative to the other populations. The solutions to the SVD estimates of  $\mathbf{b}$  and  $\mathbf{k}$  are identical to those in (2.5), except that  $\tilde{\mathbf{Y}}$  is replaced with  $\tilde{\mathbf{Y}}^c$ .

3. For each population  $i = 1, \dots, I$ , estimate the population-specific factors  $\mathbf{b}_i$  and  $\mathbf{k}_i$  from the residuals in the previous step. Specifically, a first-order SVD is applied to the following matrix:

$$\tilde{\mathbf{Y}}^r = (\tilde{\mathbf{y}}_1^r, \dots, \tilde{\mathbf{y}}_n^r), \quad (2.36)$$

where

$$\tilde{\mathbf{y}}_t^r := \underbrace{\mathbf{y}_{t,i} - \hat{\mathbf{a}}_i}_{\tilde{\mathbf{y}}_{t,i}} - \hat{\mathbf{b}} \hat{k}_t \quad (2.37)$$

is the vector of time- $t$  residuals obtained in Step 2, and  $\hat{\mathbf{b}}$  and  $\hat{k}_t$  are the estimates of  $\mathbf{b}$  and  $k_t$  obtained in Step 2, respectively. The solutions to the SVD estimates of

$\mathbf{b}_i$  and  $\mathbf{k}_i$  are identical to those for  $\mathbf{b}$  and  $\mathbf{k}$  in (2.5), respectively, except that  $\tilde{\mathbf{Y}}$  is replaced with  $\tilde{\mathbf{Y}}^r$ .

To robustify this estimation procedure, we propose to replace the SVD in Steps 2 and 3 with the  $t$ -PPCA approach. It is worthwhile that the previously described  $t$ -PPCA set-up is constructed for log central rates of death, via  $\mathbf{y}_t \sim t_\nu(\mathbf{a}, \mathbf{b}\mathbf{b}^T + \sigma^2\mathbf{I})$  in which  $\mathbf{a}$  needs to be estimated as the mean parameter vector. However, we do not need to estimate any mean parameter vector in Steps 2 and 3 of the ACF model estimation procedure, as matrices  $\tilde{\mathbf{Y}}^c$  and  $\tilde{\mathbf{Y}}^r$  are mean-centered. In other words, the  $t$ -PPCA for Steps 2 and 3 should be a restricted  $t$ -PPCA with a zero mean vector. Accordingly, when applying the EM algorithm to these two steps, we set the mean parameter vector to zero and do not update it in the M-step.

## 2.5.2 Common Age-Effect Model

The second multi-population extension of the Lee-Carter model we consider is the common age effect (CAE) model, proposed by Kleinow (2015). Compared to the ACF model, the CAE model is more parsimonious in the sense that it has only one bilinear term, which is formed by a common age effect and a population-specific time (period) effect.

Let us suppose again that the model is applied to  $I$  populations, and the dataset spans  $p$  age groups and  $n$  calendar years. The CAE model assumes that the log central rate of death for individuals in population  $i$  with an age of  $x$  in calendar year  $t$  is given by

$$y_{x,t,i} := \log(m_{x,t,i}) = a_{x,i} + b_x k_{t,i} + \varepsilon_{x,t,i}, \quad (2.38)$$

where  $a_{x,i}$  and  $k_{t,i}$  are population-specific parameters,  $b_x$  is the common age effect parameter, and  $\varepsilon_{x,t,i}$  is the error term.

Using the notation defined in (2.32), we can express (2.38) in the following vector form:

$$\mathbf{y}_{t,i} := \log(\mathbf{m}_{t,i}) = \mathbf{a}_i + \mathbf{b}k_{t,i}. \quad (2.39)$$

In using the CAE model, parameters  $\mathbf{a}_i$ ,  $\mathbf{b}$  and  $\mathbf{k}_i$ , for  $i = 1, \dots, I$ , are estimated to historical data. Time-series processes are then fitted to  $\mathbf{k}_i$ , for  $i = 1, \dots, I$ , and are extrapolated to obtain mortality forecasts.

Kleinow (2015) demonstrated that the estimation for the CAE model can be treated as a special case of a method called common principal component analysis, which relies on certain specific numerical algorithms for finding the solutions, as noted by Clarkson (1988) for instance. However, in this study we discover that the CAE model can be estimated more easily by augmenting the data matrix and then applying a first-order SVD to the augmented matrix. This estimation procedure is summarized below.

1. For each population  $i = 1, \dots, I$ , estimate  $\mathbf{a}$  as  $\hat{\mathbf{a}}_i = \bar{\mathbf{y}}_i$  and compute the vector of centered log central rates of death at time  $t$  as  $\tilde{\mathbf{y}}_{t,i} = \mathbf{y}_{t,i} - \bar{\mathbf{y}}_i$  for every  $t$  in the calibration window  $[t_1, t_n]$ . This step is the same as Step 1 in the estimation procedure for the ACF model.
2. Construct an augmented data matrix  $\tilde{\mathbf{Y}}^*$  by combining the vectors of centered log central rates of death  $\tilde{\mathbf{y}}_{t,i}$  for  $i = 1, \dots, I$  and  $t = t_1, \dots, t_n$  in the following manner:

$$\tilde{\mathbf{Y}}^* = \underbrace{(\tilde{\mathbf{y}}_{1,1}, \dots, \tilde{\mathbf{y}}_{n,1})}_{\tilde{\mathbf{Y}}_1} \underbrace{(\tilde{\mathbf{y}}_{1,2}, \dots, \tilde{\mathbf{y}}_{n,2})}_{\tilde{\mathbf{Y}}_2} \dots \underbrace{(\tilde{\mathbf{y}}_{1,I}, \dots, \tilde{\mathbf{y}}_{n,I})}_{\tilde{\mathbf{Y}}_I}, \quad (2.40)$$

where  $\tilde{\mathbf{Y}}_i = (\tilde{\mathbf{y}}_{1,i}, \dots, \tilde{\mathbf{y}}_{n,i})$  represents the matrix of centered log central rates of death for population  $i$ .

3. Apply a first-order SVD to the augmented matrix  $\tilde{\mathbf{Y}}^*$  to obtain an estimate of  $\mathbf{b}$ . The solution is the same as that for  $\mathbf{b}$  in (2.5), except that  $\tilde{\mathbf{Y}}$  is replaced with  $\tilde{\mathbf{Y}}^*$ .
4. Estimate  $\mathbf{k}_i$  by matching aggregate death counts. Specifically, the estimate of  $k_{t,i}$ , for  $i = 1, \dots, I$  and  $t = t_1, \dots, t_n$ , is obtained by solving the following equation for  $k_{t,i}$ :

$$D_{t,i} = \sum_{x=x_1}^{x_p} D_{x,t,i} = \sum_{x=x_1}^{x_p} \left( N_{x,t,i} \cdot e^{\hat{a}_x + \hat{b}_{x,i} k_{t,i}} \right), \quad (2.41)$$

where  $D_{t,i}$  is the total number of deaths for population  $i$  in year  $t$ ,  $D_{x,t,i}$  represents the number of deaths for population  $i$ , age group  $x$  and year  $t$ , and  $N_{x,t,i}$  is the corresponding number of exposures. The equation above is parallel to (2.6) for the original Lee-Carter model, and the solution to it can be obtained by a one-dimensional line search.

To robustify this estimation procedure, we propose to replace the SVD in Step 3 with the  $t$ -PPCA model. Similar to the  $t$ -PPCA for the ACF model, the input matrix  $\tilde{\mathbf{Y}}^*$  in this application is mean-centered. Therefore, we fix the mean parameter vector in the  $t$ -PPCA to zero and do not update it in the M-step of the EM algorithm.

## 2.6 Numerical Illustrations

In this section, we present two experiments to illustrate the proposed  $t$ -PPCA estimation method. The first experiment is based entirely on historical data, with an aim to highlight how the  $t$ -PPCA estimation method responds to the impact of World War II on mortality. The second experiment leverages simulated effects of COVID-alike pandemics, with a goal to fully examine the performance of the  $t$ -PPCA estimation method when outliers emerge at different time points in the calibration window. For both experiments, we benchmark the proposed  $t$ -PPCA estimation method against two conventional estimation methods: SVD and Poisson MLE. The latter benchmark method is implemented using `StMoMo` package in R.

### 2.6.1 Impact of World War II

In this experiment, we consider the actual mortality experience of US males and females. The data we use is obtained from the Human Mortality Database.

Using the estimation methods under consideration, we fit the Lee-Carter model to two calibration windows: 1940-2019 and 1970-2019. The former calibration window covers World War II, whereas the latter does not. We intentionally exclude data beyond 2019 (when the outbreak of COVID-19 started), so that the experiment can focus on a single major mortality outlier. Throughout the experiment, the age range used is 0-100.

Figure 2.1 shows the estimation results. For the shorter calibration window that begins in 1970, the three estimation methods yield highly similar parameter estimates. This outcome is anticipated, as this shorter calibration window does not cover World War II. However, for the longer calibration window that begins in 1940, the  $t$ -PPCA method yields

significantly different estimates of  $b_x$  compared to the two benchmark methods. In more detail, as the calibration window is widened, the  $t$ -PPCA preserves the curvature of  $b_x$  over young and middle ages, but the curvature is flattened out when the two benchmark methods are used.

The estimation results indicate that the  $t$ -PPCA method is more robust to outliers compared to the two benchmark methods. We believe that the parameter estimates produced by the  $t$ -PPCA method are more reliable for producing long-term mortality forecasts. This is because, as we now explain, the flattening of  $b_x$  observed for the two benchmark methods is due most likely to the short-term impact of World War II instead of permanent changes in the pattern of mortality decline.

According to Smith (1947), in World War II, almost half of the US servicemen were under 26 years of age, and 42.6% were between 26 and 37, with only a scant 7.5% aged 38 or over. On the other hand, the US male population in 1940 is quite balanced across age groups, with only 29% under 26 and 32.8% aged 38 or older. From these facts, we can infer that younger age groups account for most of the WWII-related deaths in the country. As the war ended, the mortality experience of the younger age groups reverted to, approximately, its original level. For estimation methods that are not robust to outliers, the reversion is mis-classified as a long-term improvement in mortality and therefore results estimates of  $b_x$  that are higher than what they should be. For the robust  $t$ -PPCA method, the short-term extreme effect on mortality due to World War II is down-weighted so that it is less influential on the estimates of  $b_x$ .

While other events over the period of 1940-1969 may contribute to the changes in  $b_x$ , we believe that the contribution from World War II is dominant for the following reasons. First, while World War II resulted in approximately 405,399 American deaths, other wars during the period resulted in much fewer: 36,574 for Korean War and 58,220 for Vietnam War (Congressional Research Service, 2020). Second, according to the Center for Disease Control and Prevention (CDC), no significant pandemic occurred during the period of 1940-1969.

We conclude this subsection with a sensitivity test for the initial values used in the EM algorithm for the  $t$ -PPCA method. Recall that the algorithm requires initial values

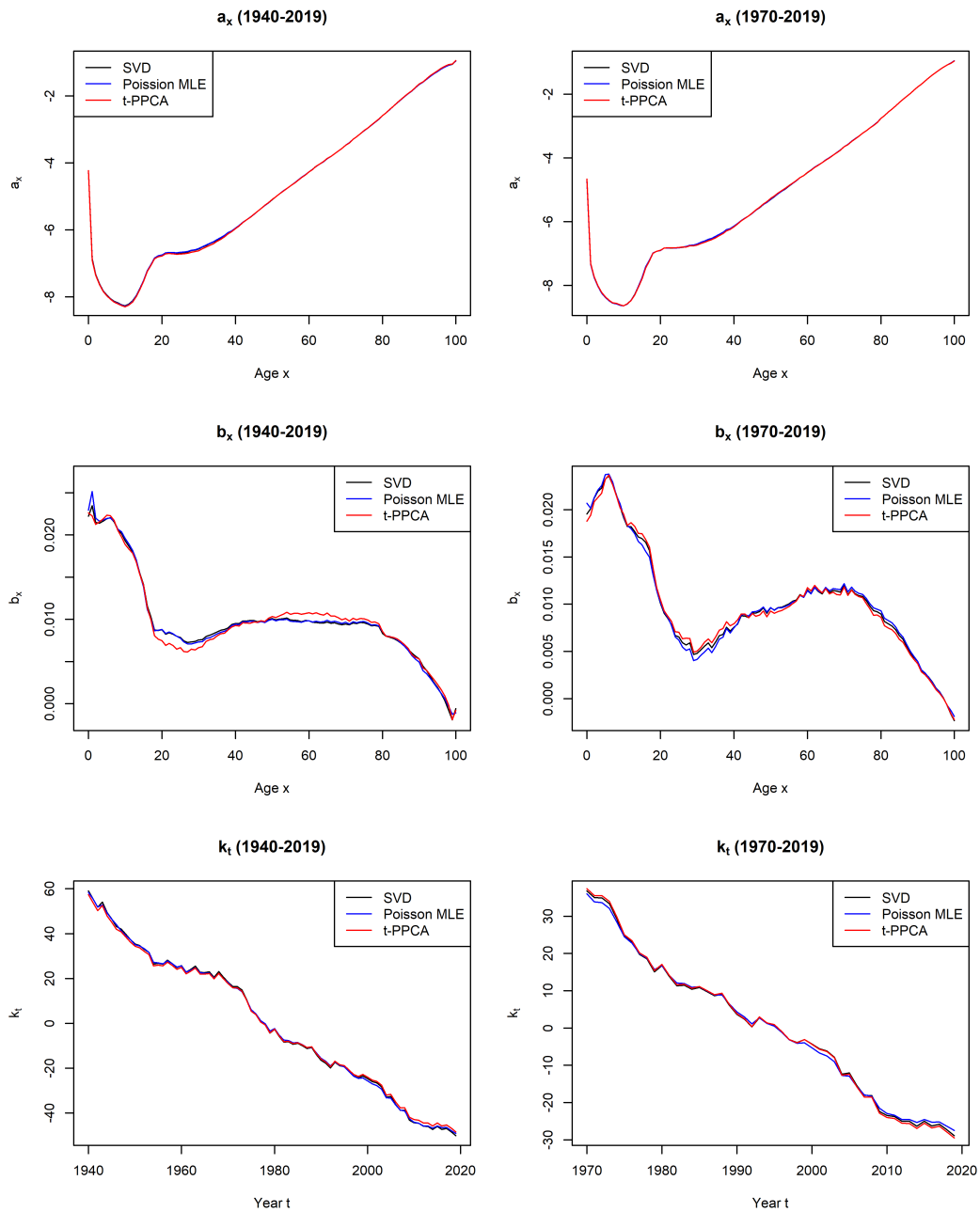


Figure 2.1: Estimates of  $a_x$  (top),  $b_x$  (middle) and  $k_t$  (bottom) obtained from actual US mortality experience over 1940-2019 (left) and 1970-2019 (right)

Table 2.1: Average relative changes in the  $t$ -PPCA estimates of  $\mathbf{a}$  and  $\mathbf{b}$  arising from shocks to initial values for the EM algorithm

Parameter	Initial value shock	Calibration window			
		1940-2019		1970-2019	
		$\Delta a_x$	$\Delta b_x$	$\Delta a_x$	$\Delta b_x$
$\mathbf{a}$	+10%	$8.2 \times 10^{-3}$	$5.6 \times 10^{-5}$	$5.4 \times 10^{-3}$	$5.0 \times 10^{-5}$
	-10%	$3.2 \times 10^{-3}$	$2.4 \times 10^{-5}$	$3.3 \times 10^{-3}$	$6.2 \times 10^{-4}$
$\mathbf{b}$	+10%	$6.4 \times 10^{-4}$	$4.0 \times 10^{-6}$	$2.7 \times 10^{-4}$	$5.1 \times 10^{-4}$
	-10%	$1.7 \times 10^{-3}$	$7.0 \times 10^{-6}$	$4.0 \times 10^{-4}$	$4.2 \times 10^{-5}$
$\sigma^2$	$\times 0.5$	$4.6 \times 10^{-5}$	$2.8 \times 10^{-7}$	$4.2 \times 10^{-5}$	$4.6 \times 10^{-6}$
	$\times 2$	$8.3 \times 10^{-5}$	$5.2 \times 10^{-7}$	$7.0 \times 10^{-5}$	$7.9 \times 10^{-6}$
$\nu$	$\nu = 1.5$	$7.5 \times 10^{-7}$	$4.9 \times 10^{-9}$	$2.7 \times 10^{-7}$	$8.3 \times 10^{-8}$
	$\nu = 10$	$1.4 \times 10^{-6}$	$8.4 \times 10^{-9}$	$6.8 \times 10^{-7}$	$7.5 \times 10^{-8}$

for  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$ . In our baseline results, the initial values for  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\sigma^2$  are taken as their corresponding (analytically calculated) standard PPCA estimates, and that for  $\nu$  is set to 3. To examine the sensitivity of the  $t$ -PPCA estimation results relative to the initial values, we shock the initial values for the four parameters one at a time as follows, and re-obtain the  $t$ -PPCA parameter estimates:

- $\mathbf{a}$  and  $\mathbf{b}$ : the whole vector of default initial values are shocked by  $\pm 10\%$ ;
- $\sigma^2$ : the default initial value is shocked by multiplicative factors of 0.1 and 10;
- $\nu$ : significantly lower and higher initial values (1.5 and 10) are considered.

Table 2.1 presents the sensitivity of the  $t$ -PPCA estimates of  $\mathbf{a}$  and  $\mathbf{b}$  relative to the initial values, measured by the average relative change in the two parameters:

$$\Delta a_x = \frac{1}{p} \sum_{x=x_1}^{x_p} \left| \frac{\hat{a}_x^{(A)} - \hat{a}_x^{(B)}}{\hat{a}_x^{(B)}} \right|, \quad \Delta b_x = \frac{1}{p} \sum_{x=x_1}^{x_p} \left| \frac{\hat{b}_x^{(A)} - \hat{b}_x^{(B)}}{\hat{b}_x^{(B)}} \right|, \quad (2.42)$$

where  $\hat{a}_x^{(A)}$  and  $\hat{b}_x^{(A)}$  represent estimates based on shocked initial values, and  $\hat{a}_x^{(B)}$  and  $\hat{b}_x^{(B)}$  denote estimates that are based on the default initial values. The test result reveals that the  $t$ -PPCA estimation is extremely insensitive to the initial values of  $\sigma^2$  and  $\nu$ , as the resulting parameter estimates remain almost identical even when these initial values are shocked significantly. The sensitivity to the initial values of  $\mathbf{a}$  and  $\mathbf{b}$  is a slightly greater, but within an acceptable range ( $< 1\%$ ). This modest sensitivity is not a concern in practice, because it is natural to use the SVD or standard PPCA estimates of  $\mathbf{a}$  and  $\mathbf{b}$  as initial values for these two parameter vectors, and there is little reason to deviate from these choices.

## 2.6.2 Simulated Pandemic Effects

We now conduct a simulation study to compare the performance of the estimation methods under consideration when mortality outliers that resemble the effect of a pandemic arise at different time points and for different durations in the calibration window. The study is based on pseudo data sets that are created by superimposing synthetic mortality outliers to US mortality experience (male and female combined) from 1970 to 2019, a period that is free of extreme mortality fluctuations. As in the previous experiment, the age range used is 0-100.

The synthetic mortality outliers are created by making reference to US COVID-19 deaths in 2020 (Table 2.2).<sup>2</sup> The COVID-19 death counts provided are arranged by broad age groups. To distribute them across individual ages, we assume that at each age, the COVID-19 mortality rate is proportional to the all-cause mortality rate. This assumption is in line with the “proportionality hypothesis” that is adopted by Cairns et al. (2020, 2024).

As an initial illustration, we consider a scenario where a synthetic COVID-like pandemic occurred for three years from 1970 to 1972. To mimic this scenario in the data set, we add the COVID-19 deaths (distributed across individual ages) to the death counts

---

<sup>2</sup>Source: Centers of Disease Control and Prevention (CDC). <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku>

Table 2.2: US COVID-19 deaths in 2020

Age Group	< 1	1-4	5-14	15-24	25-34	35-44
Deaths	52	25	68	615	2,621	6,785
Age Group	45-54	55-64	65-74	75-84	> 85	
Deaths	18,327	45,572	82,286	106,259	122,820	

in 1970, 1971 and 1972. This arrangement means that we are making a simplifying assumption that the impact of the synthetic pandemic on death counts over the three years remained the same. The three estimation methods under consideration are then applied to this data set.

- Estimates of  $\mathbf{a}$ :

For all three estimation methods under consideration, the estimates of  $\mathbf{a}$  are essentially unaffected by the introduction of the synthetic outlier to the data sample. This observation is in line with the result from the previous experiment (Figure 2.1, left panel). The insensitivity of  $\hat{\mathbf{a}}$  relative to mortality outliers is further confirmed in the full simulation study performed later in this subsection.

- Estimates of  $\mathbf{b}$ :

When the proposed  $t$ -PPCA method is used, the estimate of  $\mathbf{b}$  is almost unaffected by the introduction of the synthetic outlier, indicating a strong robustness provided by the method.

Nevertheless, for the two conventional methods, the estimates of  $\mathbf{b}$  are visibly different as the synthetic outlier is added to the data sample. The changes in the estimates of  $\mathbf{b}$  can be explained as follows. As the deaths associated with the synthetic mortality outlier are concentrated at older ages (see Table 2.2), old-age mortality rates at the time when the synthetic outlier is introduced (the beginning of the calibration window) are abnormally high. If the estimation method fails to acknowledge the observation over this period as a short-term outlier, it would fully incorporate the abnormally high rates of old-age mortality at the beginning of the calibration window

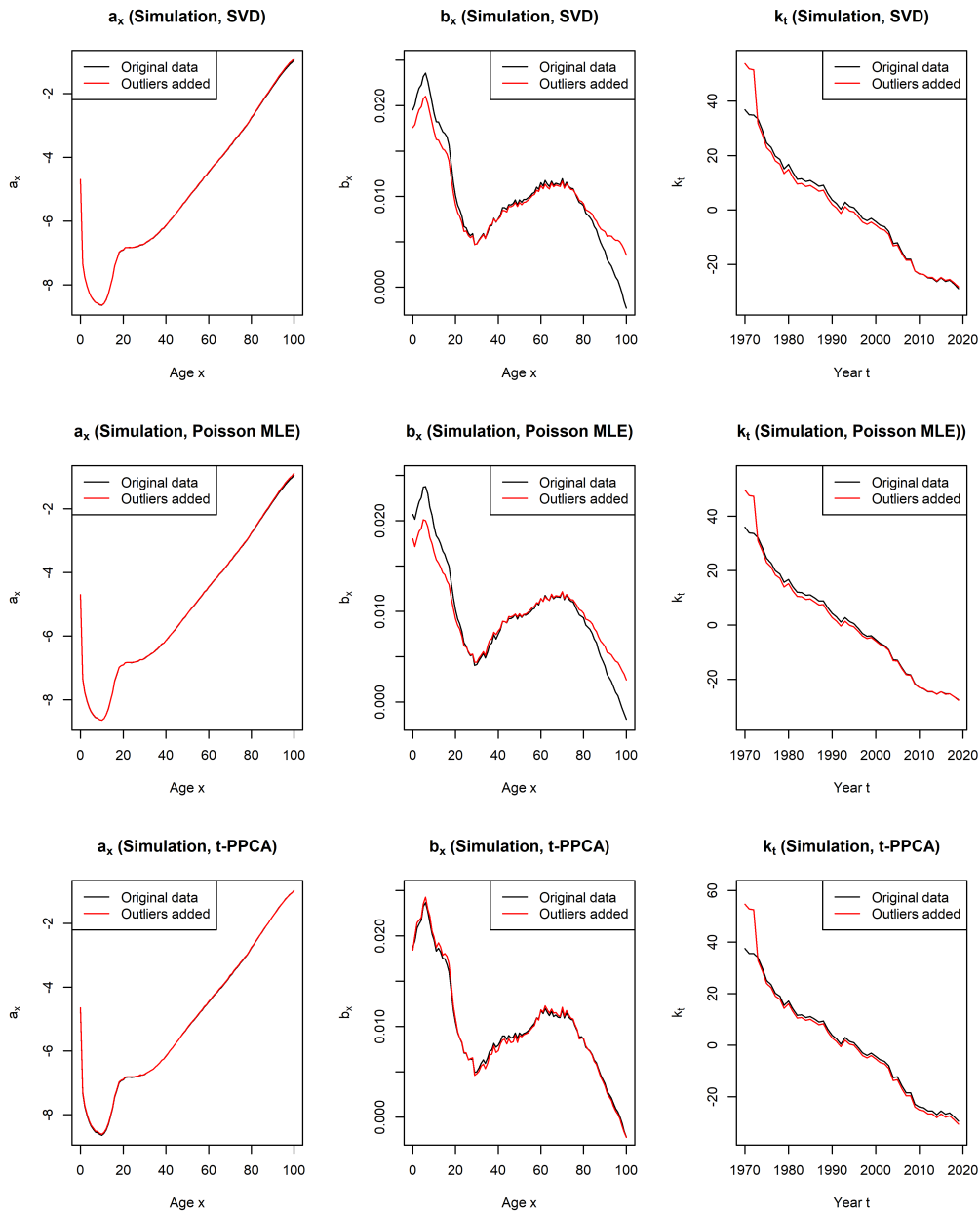


Figure 2.2: Estimates of  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{k}$  based on US mortality experience in 1970-2019, before and after a synthetic outlier is added to 1970-1972, produced by SVD (top), Poisson MLE (middle), and  $t$ -PPCA (bottom).

and mistreat the fall in mortality immediately after the synthetic pandemic as a part of long-term mortality improvement. As a consequence, the resulting estimates of  $b_x$  (which captures the long-term rate of mortality reduction at age  $x$  relative to other ages) for this age range are higher than they should be. On the other hand, the identifiability constraint  $\sum_{x=x_1}^{x_p} \hat{b}_x = 1$  has to be maintained, so that the estimates of  $b_x$  for younger ages must be reduced accordingly.

- Estimates of  $\mathbf{k}$

For all three estimation methods, the estimates of  $k_t$  for  $t = 1970, 1971, 1972$  are significantly higher to reflect the anomaly induced by the synthetic outlier that is superimposed to the original data sample in that three years. This outcome is expected, and is desirable from a modelling perspective as the spike in the estimates of  $k_t$  would not affect long-term forecasts if it is handled by proper time-series processes.

However, it is noteworthy that for  $t > 1972$ , the estimates of  $k_t$  produced by the two conventional methods are still somewhat different as the synthetic outlier is introduced. This is not a desirable outcome as the synthetic outlier, by design, affects the mortality in 1970-1972 only.

On the other hand, for the  $t$ -PPCA method, there are no visible changes to the estimates of  $k_t$  for  $t > 1972$ . This desirable outcome is a natural consequence of the insensitivity of  $\hat{\mathbf{b}}$  to the outliers. Recall that  $k_t$  is estimated by death count matching via (2.6). If there is no change to death counts ( $D_t$  and  $D_{x,t}$ ), exposure counts  $N_{x,t}$ , and the estimates of  $a_x$  and  $b_x$  for  $x = x_1, \dots, x_p$ , then there should be no change to the estimate of  $k_t$ .

Next, we perform a simulation study to evaluate the performance of the estimation methods under consideration, when the outlier has different lifetimes and is located at different positions in the calibration window. To this end, we consider three scenarios, in which the synthetic mortality outlier is assumed to last for one, three, and five years, respectively. For each scenario, we allow the synthetic outlier to emerge at different time points over the calibration window. Specifically, for the scenario where the synthetic outlier is assumed to last for one year, we use  $M = 50$  pseudo data sets, created by superimposing

the synthetic outlier to the original data sample (US mortality, 1970-2019) in 1970, 1971, ..., 2019, respectively. In a similar manner,  $M = 48$  and  $M = 46$  pseudo data sets are created for the scenarios when the synthetic outlier is assumed to last for three and five years, respectively.

For each pseudo data set, we obtain estimates of  $a_x$ ,  $b_x$ , and  $k_t$ , denoted by  $\hat{a}_x^{**}$ ,  $\hat{b}_x^{**}$  and  $\hat{k}_t^{**}$ , respectively. These estimates are compared against their counterparts that are derived from the original outlier-free dataset, denoted by  $\hat{a}_x^*$ ,  $\hat{b}_x^*$  and  $\hat{k}_t^*$ , respectively. The comparison is made on the basis of two metrics, Relative Mean Absolute Error (RMAE) and Relative Root Mean Square Error (RRMSE). For  $\mathbf{a}$  and  $\mathbf{b}$ , we have

$$\text{RMAE}(\hat{\mathbf{a}}) = \frac{1}{p} \sum_{x=x_1}^{x_p} \left| \frac{\hat{a}_x^{**} - \hat{a}_x^*}{\hat{a}_x^*} \right|, \quad \text{RRMSE}(\hat{\mathbf{a}}) = \sqrt{\frac{1}{p} \sum_{x=x_1}^{x_p} \left( \frac{\hat{a}_x^{**} - \hat{a}_x^*}{\hat{a}_x^*} \right)^2}, \quad (2.43)$$

and

$$\text{RMAE}(\hat{\mathbf{b}}) = \frac{1}{p} \sum_{x=x_1}^{x_p} \left| \frac{\hat{b}_x^{**} - \hat{b}_x^*}{\hat{b}_x^*} \right|, \quad \text{RRMSE}(\hat{\mathbf{b}}) = \sqrt{\frac{1}{p} \sum_{x=x_1}^{x_p} \left( \frac{\hat{b}_x^{**} - \hat{b}_x^*}{\hat{b}_x^*} \right)^2}, \quad (2.44)$$

respectively, in which the errors are averaged over the entire age range of  $[x_1, x_p]$ . For  $\mathbf{k}$ , we have

$$\text{RMAE}(\hat{\mathbf{k}}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left| \frac{\hat{k}_t^{**} - \hat{k}_t^*}{\hat{k}_t^*} \right|, \quad \text{RRMSE}(\hat{\mathbf{k}}) = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left( \frac{\hat{k}_t^{**} - \hat{k}_t^*}{\hat{k}_t^*} \right)^2}, \quad (2.45)$$

where  $|\mathcal{T}|$  denotes the cardinality of  $\mathcal{T}$ , a notation representing the set of calendar years that are not affected by the synthetic outlier. For instance, for the pseudo data set with a 3-year synthetic outlier that spans 1973-1975, we have  $\mathcal{T} = \{1970, 1971, 1972, 1976, 1977, \dots, 2019\}$  and  $|\mathcal{T}| = 50 - 3 = 47$ . The RMAE and RRMSE for  $\mathbf{k}$  are defined in this way to exclude the changes in the estimates of  $k_t$  for the years that are affected by the synthetic outlier. As previously mentioned, such changes are anticipated, and are irrelevant to the robustness we desire.

Finally, we average the RMAE and RRMSE over all  $M$  pseudo data sets to obtain the average RMAE and RRMSE reported in Tables 2.3 to 2.5. We have the following comments concerning the average RMAE and RRMSE for the three parameter vectors:

- For all estimation methods under consideration, the estimate of  $\mathbf{a}$  is rather insensitive to outliers, with average RMAE and RRMSE values being less than 0.1% consistently. With respect to the estimate of  $\mathbf{a}$  only, the marginal benefit brought by the  $t$ -PPCA approach is negligible. The estimation of  $\mathbf{a}$  is not our primary focus, because  $\mathbf{a}$  is irrelevant if the Lee-Carter model is used for the purpose of deriving long-term improvement scale factors (see (2.10)).
- With respect to  $\mathbf{b}$ , the  $t$ -PPCA method provides significantly more robust estimation results compared to the two benchmark methods. For instance, for the scenario in which the synthetic outlier is assumed to last for three years, the  $t$ -PPCA method reduces the average RMAE by 75% compared to Poisson MLE and 61% compared to SVD. The reduction in RRMSE offered by the  $t$ -PPCA method is even more remarkable. Given that RRMSE imposes a harsher penalty on larger errors compared to RMAE, the strong reduction in RRMSE indicates that the  $t$ -PPCA method effectively controls large estimation errors. These results make a strong case for using the  $t$ -PPCA method for estimating the Lee-Carter model.
- The  $t$ -PPCA method also provides significantly more robust estimation results for  $k_t$  ( $t \in \mathcal{T}$ ) compared to the two benchmark methods. This improvement is also important, because for  $t \in \mathcal{T}$  the synthetic outlier should not cause the estimates of  $k_t$  to change. Any changes in the estimates of  $k_t$  for  $t \in \mathcal{T}$  would create an erroneous input for the time-series modelling of  $\{k_t\}$  and ultimately affects the forecasting results from the Lee-Carter model.

Table 2.3: Average RMAE and RRMSE of  $\hat{a}$  produced by SVD, Poisson MLE and  $t$ -PPCA for different outlier durations.

Outlier duration	Estimation method	Average RMAE ( $\hat{a}$ )	Average RRMSE ( $\hat{a}$ )
1 year	SVD	0.0010	0.0018
	Poisson MLE	0.0013	0.0027
	$t$ -PPCA	0.0006	0.0008
3 years	SVD	0.0029	0.0054
	Poisson MLE	0.0038	0.0079
	$t$ -PPCA	0.0017	0.0023
5 years	SVD	0.0048	0.0089
	Poisson MLE	0.0060	0.0124
	$t$ -PPCA	0.0028	0.0038

Table 2.4: Average RMAE and RRMSE of  $\hat{b}$  produced by SVD, Poisson MLE and  $t$ -PPCA for different outlier durations.

Outlier duration	Estimation method	Average RMAE ( $\hat{b}$ )	Average RRMSE ( $\hat{b}$ )
1 year	SVD	0.0448	0.1890
	Poisson MLE	0.0705	0.3001
	$t$ -PPCA	0.0170	0.0472
3 years	SVD	0.1220	0.5120
	Poisson MLE	0.1919	0.8200
	$t$ -PPCA	0.0479	0.1326
5 years	SVD	0.1851	0.7730
	Poisson MLE	0.2909	1.2463
	$t$ -PPCA	0.0746	0.2028

Table 2.5: Average RMAE and RRMSE of  $\hat{k}$  produced by SVD, Poisson MLE and  $t$ -PPCA for different outlier durations.

Outlier duration	Estimation method	Average RMAE ( $\hat{k}$ )	Average RRMSE ( $\hat{k}$ )
1 year	SVD	0.0725	0.1647
	Poisson MLE	0.0532	0.1027
	$t$ -PPCA	0.0379	0.0905
3 years	SVD	0.2155	0.4878
	Poisson MLE	0.1640	0.3149
	$t$ -PPCA	0.1240	0.2934
5 years	SVD	0.3554	0.8002
	Poisson MLE	0.2711	0.5178
	$t$ -PPCA	0.2187	0.5135

## 2.7 Concluding Remarks

Robustness to outliers is an important property of stochastic mortality models that are devised to produce long-term mortality projections. Considering the Lee-Carter model, we have argued, through scale factors shown in (2.10), that long-term life table projections can be biased if the estimates of  $b_x$  in the model are sensitive to outliers, even though the time-series modelling of  $\{k_t\}$  is adjusted for the effect of outliers.

Conventional estimation methods including SVD and Poisson MLE are not robust to outliers. To address this problem, in this chapter we have introduced the  $t$ -PPCA method for estimating the Lee-Carter model, developed by a careful reformulation of the SVD method in the original work of Lee and Carter (1992) into a PPCA setting, and a robustification of the PPCA with multivariate  $t$ -distributions. We have also developed a computationally efficient EM algorithm for implementing the  $t$ -PPCA method. Through illustrations with real and synthetic data, we have demonstrated that the  $t$ -PPCA method significantly enhances parameter robustness, particularly for  $b_x$ .

We have further extended our proposed  $t$ -PPCA method to multi-population generalizations of the Lee-Carter, including the augmented common factor model and the CAE model. In actuarial practice, these models are often used to quantify population basis risk in index-based longevity hedges. Such risk arises from the difference in mortality experience between the population of individuals in the hedger's portfolio and the (national) population to which the hedging instrument is linked (Cairns and El Boukfaoui, 2021; Li and Hardy, 2011). The risk is long-term in nature, and heavily dependent on the age-response parameters in the assumed model. For instance, if the augmented common factor model is assumed, then the projected mortality differentials between two populations, say  $i$  and  $j$ , in logarithmic scale is given by

$$\log(m_{x,t,i}) - \log(m_{x,t,j}) = \hat{a}_{x,i} - \hat{a}_{x,j} + \hat{b}_{x,i}\hat{k}_{t,i} - \hat{b}_{x,j}\hat{k}_{t,j}, \quad (2.46)$$

which clearly depends on the estimates of age-response parameters  $b_{x,i}$  and  $b_{x,j}$ . It is also noteworthy that hedge ratios calibrated for index-based longevity hedges are heavily dependent on age-response parameters in the assumed mortality model (Zhou and Li, 2020, 2021). Our proposed  $t$ -PPCA method enhances the robustness of the relevant age-response

parameters to outliers, and thus improves the reliability of hedge ratio and population basis risk calculations for index-based longevity hedges.

When gauging the uncertainty involved in a mortality projection, it is important not to ignore parameter uncertainty (Cairns, 2000). For the Lee-Carter and many others stochastic mortality models, one may estimate parameter uncertainty using a bootstrapping approach. For instance, when the residual bootstrap (Koissi et al., 2006) is used, parameter uncertainty is estimated from empirical distributions of parameter estimates, which are obtained by fitting the mortality model in question to a large number, say 5,000, pseudo samples that are generated by sampling residual vectors with replacement. Our proposed  $t$ -PPCA estimation method is fully compatible with bootstrapping methods including the residual bootstrap. Given that the method can be implemented with an efficient EM algorithm, the bootstrapping procedure can be completed within a reasonable amount of time even though it encompasses multiple model re-estimations.

We stress that while the  $t$ -PPCA method enhances robustness to short-term mortality outliers, it does *not* ignore the information contained in such outliers. As a matter of fact, as we have demonstrated in Section 2.6, the  $t$ -PPCA method allows short-term mortality outliers to be fully reflected in the estimates of  $k_t$  over the years when the data sample is contaminated with outliers. Extreme values in  $k_t$  can then be captured by an outlier-adjusted time-series process (Li and Chan, 2005) or a jump process (Chen and Cummins, 2010; Chen, 2013) for  $\{k_t\}$ . The true aim of the  $t$ -PPCA method is to reduce the sensitivity to short-term outliers for the estimates of  $b_x$  (which govern the age distribution of long-term mortality reduction and therefore should not be affected by short-term extreme fluctuations) and  $k_t$  for years in which the data sample is outlier-free (which should be insensitive to irrelevant outliers). It is worth-noting that the insensitivity of the estimates of  $k_t$  over the years when there is no outlier also makes the input for the time-series modelling of  $\{k_t\}$  more reliable.

# Chapter 3

## Fast Estimation of the Renshaw-Haberman Model and Its Variants

### 3.1 Introduction

One important concept in modelling and management of longevity risk is cohort effects. In the context of longevity risk, cohort effects refer to the impact of a person's birth year or generation on their health and mortality outcomes. The significance of cohort effects has long been recognized by demographers (Hobcraft et al., 1985; Wilmoth, 1990) and actuaries (Willets, 2004).

Cohort effects can be attributed to various factors such as changes in lifestyle, medical advancements, etc. Their strength varies across geographical regions, although it is widely acknowledged that they are particularly strong in the United Kingdom, where the “golden generation” who were born in the early 1930s experienced significantly higher mortality improvement. It is important to note that cohort effects are not purely historical. For example, in 2022, New Zealand announced an anti-smoking law that bans the sale of tobacco

to anyone born on or after January 1, 2009.<sup>1</sup> Shortly after New Zealand’s announcement, UK also unveiled plans to phase out smoking for young generations. Such legislations are expected to result in cohort effects in mortality improvement, as they prevent younger generations from being exposed to the negative effects of tobaccos.

To incorporate cohort effects into stochastic mortality modelling, Renshaw and Haberman (2006) extend the seminal work of Lee and Carter (1992) to develop the Renshaw-Haberman model. It adds to the original Lee-Carter model a bi-linear term, which captures the variation of mortality across years-of-birth and the interaction between such variation with age. It is also closely connected to the classical age-period-cohort (APC) model (Hobcraft et al., 1985), as it degenerates to the APC model when some of its age-specific parameters are eliminated. When estimated to historical mortality data, the model is able to absorb part of the remaining variation that is not captured by models with age and period (time-related) effects only, leaving residuals that exhibit a more random pattern. Recently, the Renshaw-Haberman model has been generalized to incorporate socioeconomic differences in mortality (Villegas and Haberman, 2014), making it applicable to an even wider range of insurance and pension applications.

In the literature, including the original work of Renshaw and Haberman (2006), the Renshaw-Haberman model is often estimated with maximum likelihood (ML). When fitting the Renshaw-Haberman with ML, a log-likelihood function is derived on the basis of a distributional assumption, typically Poisson, made on observed death counts; then, parameter estimates are obtained by maximizing the log-likelihood function. Given that the Renshaw-Haberman model has a large number of parameters, the maximization is customarily performed with an iterative Newton-Raphson method, in which parameters are updated one batch at a time. Unfortunately, ML estimation for the Renshaw-Haberman model is slow and sometimes unstable. Depending on the dataset in question, the iterative algorithm may not even converge given the desired convergence criterion. This problem is noted by a number of researchers, including Cairns et al. (2009, 2011) and Haberman and Renshaw (2009, 2011).

While a slow convergence might be acceptable is model estimation is a one-off task,

---

<sup>1</sup>The law was repealed in early 2024 for economic reasons.

it may render applications that require repeated model estimation time-prohibitive. Such applications include the following.

- *Assessment of parameter uncertainty via bootstrapping*

Any model-based mortality projection is subject to parameter uncertainty, as the parameters used for extrapolating future death rates are estimates rather than exact. One way to gauge parameter uncertainty is bootstrapping (Brouhns et al., 2005; D’Amato et al., 2012; Koissi et al., 2006). In a bootstrap, a large number pseudo datasets are generated by, for example, resampling residuals (residual bootstrapping); then, the model is re-estimated using the pseudo data sets. The procedure results in empirical distributions of model parameters, from which parameter uncertainty can be inferred. The bootstrapping procedure involves a large number of (re-)estimations, and cannot be executed in practice if the estimation is slow.

- *Calculation of Solvency Capital Requirements*

Under Solvency II, solvency capital requirement (SCR) is based on the Value-at-Risk at a 99.5% confidence level over a one-year horizon (Zhou et al., 2014). In lieu of the prescribed standard formula, an insurer may opt to calculate SCR by simulating from an approved internal model. Taking re-calibration risk<sup>2</sup> (Cairns, 2013) into account, the simulation procedure for estimating longevity Value-at-Risk encompasses the following steps: (1) simulate  $M_1$  mortality scenarios in one year from a model that is fitted to historical data; (2) for each mortality scenario, re-estimate the model to an updated dataset that includes the simulated mortality scenario, and use the re-estimated model to simulate  $M_2$  sample paths of mortality (for year 2 and beyond), from which the expected value of the liability at the end of year 1 can be calculated. Step (2) yields a distribution of liabilities at the end of year 1, which can be used to infer the 99.5% Value-at-Risk. Typically,  $M_1$  is large, so that the procedure includes a large number of model re-estimations.

- *Identification of ultimate mortality improvement rates*

---

<sup>2</sup>Re-calibration risk arises because model parameter estimates may become different if the model in question is fitted to an updated data set.

In recent years, two-dimensional mortality improvement scales have been promulgated by major actuarial professional organizations (see, e.g., Society of Actuaries, 2021). A two-dimensional mortality improvement scale is composed of relatively high short-term scale factors, which are blended into lower long-term (ultimate) scale factors through an interpolative mid-term scale. One possible way to estimate the ultimate scale factors is to fit a parametric model to absorb all transient period and cohort effects that are present in the historical data, leaving a long-term pattern from which the ultimate scale factors can be inferred (Li et al., 2020). This method requires the modeler to experiment different model structures, some of which, ideally, include multiple age-cohort interaction terms. A slow convergence rate plagues the use of this method; in particular, it hinders the consideration of models with additional age-cohort interaction terms.

So far as we aware, two attempts have been made to mitigate the estimation issues of the Renshaw-Haberman model. The first attempt is made by Renshaw and Haberman (2006), who consider a number of restricted versions of the Renshaw-Haberman model which may take less time to estimate given that they have fewer free parameters. Most notably, they propose the H1 model, which still incorporates cohort effects but assumes that such effects do not interact with age. The second attempt is made by Hunt and Villegas (2015), who argue that the problem of slow convergence is due possibly to an approximate identification issue that is applicable to the Renshaw-Haberman model. To mitigate the issue, they recommend imposing an additional parameter constraint to stabilize the estimation process and enhance algorithmic robustness. It is noteworthy that both approaches are based on a reduction in parameter space. That said, they improve estimation efficiency at the expense of goodness-of-fit to the historical data.

In this chapter, we attack the problem of estimation efficiency for the Renshaw-Haberman model from a different angle. Instead of building on the commonly used maximum likelihood approach, we consider a least squares method in which parameters are estimated by minimizing the sum of squared errors between the actual and fitted log central death rates. The idea of using a least squares approach to estimate stochastic mortality models is not new. As a matter of fact, when the original Lee-Carter and Cairns-Blake-Dowd

models were first proposed, the authors estimated them with least squares methods (Lee and Carter, 1992; Cairns et al., 2006).

It is not straightforward to efficiently estimate the Renshaw-Haberman model with a least squares method. This is because the model involves an additional (year-of-birth) dimension that is not orthogonal to the age and time dimensions, rendering the efficient singular value decomposition (SVD) technique that is used for fitting the original Lee-Carter model inapplicable. Recently, SriDaran et al. (2022) has discussed the least squares implementation of the generalized APC model as a Gaussian generalized linear model, but it may have similar convergence issues as the classic MLE estimation. To overcome the optimization challenge, we develop an alternating minimization scheme which sequentially updates one group of parameters at a time. We also formulate the update of the age-cohort component in the model as a principal component analysis (PCA) problem with missing values, so that it can be accomplished effectively using an iterative SVD algorithm. Using data from various national populations, we demonstrate that our proposed least squares method significantly outperforms the ML approach in terms of estimation efficiency, without sacrificing goodness-of-fit to historical data.

Our proposed least squares method offers several advantages over the ML approach. First, given the same convergence criterion, our proposed method takes less computation time. We argue that the improvement in estimation efficiency is due to a sharper objective function, and empirically verify this argument with a numerical experiment. Second, unlike the ML approach, the objective function in our proposed approach is not built on a specific distributional assumption, thereby avoiding the potential problems associated with choosing such an assumption. Finally, our proposed method can be implemented seamlessly with the two methods that are previously proposed by Renshaw and Haberman (2006) and Hunt and Villegas (2015) to further improve estimation efficiency.

The remainder of this chapter is organized as follows. Section 3.2 presents an overview of the Lee-Carter model, with a focus on the estimation methods for the model that are relevant to this study. Section 3.3 reviews the Renshaw-Haberman model and its estimation challenges. Section 3.4 details our proposed method, including its motivation, theoretical support, and execution. Section 3.5 explains how our proposed method can be implemented simultaneously with the two methods that are previously proposed by

Renshaw and Haberman (2006) and Hunt and Villegas (2015). Section 3.6 documents the numerical experiments that validate the advantages of our proposed method. Finally, concluding remarks are provided in Section 3.7.

## 3.2 The Lee-Carter Model

### 3.2.1 Specification

This section presents a concise review of the Lee-Carter model (Lee and Carter, 1992), with a focus on two commonly used methods for estimating the model. We let  $m_{x,t}$  be the central rate of death for age  $x$  and year  $t$ , and  $y_{x,t} := \log(m_{x,t})$  for notational convenience. The Lee-Carter model assumes that

$$y_{x,t} := \log(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}, \quad (3.1)$$

where  $a_x$  and  $b_x$  are age-specific parameters,  $k_t$  is a time-varying index, and  $\varepsilon_{x,t}$  is the error term. In the model,  $a_x$  captures ‘age effects’ (the age pattern of mortality),  $k_t$  captures ‘period effects’ (changes in the overall mortality level over time), and  $b_x$  measures the interaction between age and period effects. Throughout this chapter, we assume that the data set in question covers  $p$  ages,  $x \in [x_1, \dots, x_p]$ , and  $n$  calendar years,  $t \in [t_1, \dots, t_n]$ .

Within the Lee-Carter model (3.1), there is a deterministic part,  $a_x + b_x k_t$ , and a stochastic error term,  $\varepsilon_{x,t}$ . The error term is generally assumed to be independently and identically (i.i.d.) distributed with a normal distribution. The deterministic part, or the estimate  $\hat{y}_{x,t}$  of the true log mortality rate  $y_{x,t}$ , is where  $a_x$  depicts the shape of log mortality rates across ages,  $k_t$  indicates the overall mortality trend across time, and  $b_x$  adjusts the trend of  $k_t$  in relation to ages.

The Lee-Carter model is subject to an identifiability problem. It can be shown that two parameter constraints are required to stipulate parameter uniqueness. In the literature (including the original work of Lee and Carter (1992)), the following two parameter

constraints are typically imposed:

$$\sum_{x=x_1}^{x_p} b_x = 1 \quad \text{and} \quad \sum_{t=t_1}^{t_n} k_t = 0. \quad (3.2)$$

### 3.2.2 Least Squares Estimation

In their original work, Lee and Carter (1992) estimated (3.1) using a least squares approach, in which parameter estimates are chosen such that they minimize the sum of squared errors between the observed and fitted log central mortality rates. In more detail, let us rewrite the model in vector form as follows:

$$\mathbf{y}_t = \mathbf{a} + \mathbf{b}k_t + \boldsymbol{\varepsilon}_t, \quad (3.3)$$

where we use the notations  $\mathbf{y}_t = (y_{x_1,t}, \dots, y_{x_p,t})^T$ ,  $\mathbf{a} = (a_{x_1}, \dots, a_{x_p})^T$ ,  $\mathbf{b} = (b_{x_1}, \dots, b_{x_p})^T$  and  $\boldsymbol{\varepsilon}_t = (\varepsilon_{x_1,t}, \dots, \varepsilon_{x_p,t})^T$ . In using the least squares approach, the estimates of  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{k}$  are obtained by solving the following optimization:

$$\min_{\mathbf{a}, \mathbf{b}, \mathbf{k}} \sum_{x,t} (y_{x,t} - (a_x + b_x k_t))^2 = \min_{\mathbf{a}, \mathbf{b}, \mathbf{k}} \sum_t \|\mathbf{y}_t - (\mathbf{a} + \mathbf{b}k_t)\|_2^2, \quad (3.4)$$

where  $\mathbf{k} = (k_{t_1}, \dots, k_{t_n})^T$  and  $\|\cdot\|_2$  denotes the Euclidean norm (or  $L^2$ -norm) of a vector. When the identification constraints specified in (3.2) are applied, the optimization problem specified in (3.4) is equivalent to a special case of principal component analysis (PCA) with one principal component. Its solution can thus be obtained by performing a singular value decomposition (SVD) on the mean-centered log mortality data matrix,  $\mathbf{Y} - \bar{\mathbf{Y}} := (\mathbf{y}_{t_1} - \bar{\mathbf{y}}, \dots, \mathbf{y}_{t_n} - \bar{\mathbf{y}})$ , where  $\bar{\mathbf{Y}} = (\bar{\mathbf{y}}, \dots, \bar{\mathbf{y}})$  and  $\bar{\mathbf{y}} = \frac{1}{n} \sum_{t=t_1}^{t_n} \mathbf{y}_t$ . The solution has the following closed form:

$$\hat{\mathbf{a}} = \bar{\mathbf{y}}, \quad \hat{\mathbf{b}} = \frac{\mathbf{u}}{\mathbf{1}^T \mathbf{u}}, \quad \hat{\mathbf{k}} = (\mathbf{1}^T \mathbf{u}) \cdot (\mathbf{Y} - \bar{\mathbf{Y}})^T \mathbf{u}, \quad (3.5)$$

where  $\mathbf{u}$  is the first left-singular vector of  $\mathbf{Y} - \bar{\mathbf{Y}}$  and  $\mathbf{1} = (1, \dots, 1)^T$ . In this solution, the term  $\mathbf{1}^T \mathbf{u}$  normalizes the standard PCA solution due to the imposed constraint  $\sum_x b_x = 1$ . Additionally, it is easy to check that the constraint  $\sum_t k_t = 0$  is also met. We refer readers to Bishop and Nasrabadi (2006) and Hastie et al. (2009) for a comprehensive overview of the theory of PCA.

### 3.2.3 Maximum Likelihood Estimation

In contrast to the least squares approach, maximum likelihood estimation (MLE) requires a distributional assumption. Estimation of the Lee-Carter model using maximum likelihood was first accomplished by Wilmoth (1993), who assumes that the observed death count in each age-time cell follows a Poisson distribution. We let  $D_{x,t}$  be the observed number of deaths for age  $x$  and year  $t$ , and  $N_{x,t}$  be the corresponding exposure-to-risk. The method of Poisson-MLE assumes that

$$D_{x,t} \sim \text{Poisson}(N_{x,t}m_{x,t}), \text{ with } \log(m_{x,t}) = a_x + b_x k_t. \quad (3.6)$$

Parameter estimates are obtained by maximizing the following log-likelihood function:

$$\ell(\mathbf{a}, \mathbf{b}, \mathbf{k}) = \sum_{x,t} (D_{x,t}(a_x + b_x k_t) - N_{x,t}e^{a_x + b_x k_t}) + \text{const}. \quad (3.7)$$

The optimization problem can be solved via an iterative Newton-Raphson method (Goodman, 1979).

## 3.3 The Renshaw-Haberman Model and Its Variants

### 3.3.1 Specification

The focus of this chapter is the Renshaw-Haberman model (Renshaw and Haberman, 2006), which extends the Lee-Carter model by incorporating cohort effects. The Renshaw-Haberman model assumes that

$$y_{x,t} := \log(m_{x,t}) = a_x + b_x k_t + c_x \gamma_{t-x} + \varepsilon_{x,t}, \quad (3.8)$$

In the above,  $\gamma_{t-x}$  is an index that is linked to year-of-birth ( $t - x$ ), thereby capturing cohort effects. Parameter  $c_x$  captures the sensitivity of the log central death rate at each age to cohort effects. The interpretations of  $a_x$ ,  $b_x$  and  $k_t$  in (3.8) are the same as those in the Lee-Carter model.

The Renshaw-Haberman model is also subject to an identifiability problem. In addition to the two constraints specified in (3.2), two constraints on  $c_x$  and  $\gamma_{t-x}$  are needed. Following Renshaw and Haberman (2006), the additional constraints we use are

$$\sum_{x=x_1}^{x_p} c_x = 1, \quad \sum_{t-x=t_1-x_p}^{t_n-x_1} \gamma_{t-x} = 0. \quad (3.9)$$

It is worth-noting that the constraint for  $\gamma_{t-x}$  may be formulated differently. For instance, as mentioned by Renshaw and Haberman (2006), another possible choice is  $\gamma_{t_1-x_p} = 0$ . We choose to use  $\sum_{t-x=t_1-x_p}^{t_n-x_1} \gamma_{t-x} = 0$ , because it is commonly adopted in the literature (e.g., Cairns et al., 2009) and used in the `StMoMo` package in `R` (Villegas et al., 2018). The choice of the constraints makes no difference to the goodness-of-fit.

### 3.3.2 Estimation

Estimation of Renshaw-Haberman model is well-known to be challenging. While the Lee-Carter model can be estimated readily using a least squares approach, a parallel least squares method for estimating the Renshaw-Haberman model remains largely unexplored in the literature. The least squares solution to the Renshaw-Haberman estimation problem is not easy to obtain, because the incorporation of cohort effects expands the dimension of the problem. This challenge is succinctly described by Fung et al. (2019):

*“Under the Lee–Carter original approach, one might consider modelling the crude death rate with cohort effects as follows:*

$$\log(\tilde{m}_{x,t}) = \alpha_x + \beta_x \kappa_t + \beta_x^\gamma \gamma_{t-x} + \varepsilon_{x,t}.$$

*However the dimension of the cohort index would cause difficulty for the SVD estimation approach.”*

In the literature, the Renshaw-Haberman model is often estimated using maximum likelihood. Assuming Poisson death counts, the log-likelihood function for the Renshaw-Haberman model is given by

$$\ell(\mathbf{a}, \mathbf{b}, \mathbf{k}, \mathbf{c}, \boldsymbol{\gamma}) = \sum_{x,t} (D_{x,t}(a_x + b_x k_t + c_x \gamma_{t-x}) - N_{x,t} e^{a_x + b_x k_t + c_x \gamma_{t-x}}) + \text{constant}, \quad (3.10)$$

where  $\mathbf{c} = (c_{x_1}, \dots, c_{x_p})$  and  $\boldsymbol{\gamma} = (\gamma_{t_1-x_p}, \dots, \gamma_{t_n-x_1})$ . This objective function is maximized through an iterative Newton-Raphson method to obtain parameter estimates. Although Poisson-MLE is technically feasible for the Renshaw-Haberman model, computational efficiency represents a significant concern to users. It is widely reported that Poisson-MLE for the Renshaw-Haberman model takes a lot of iterations to converge (Cairns et al., 2009, 2011; Haberman and Renshaw, 2009, 2011). The problem is investigated more deeply by Currie (2016), who emphasized the importance of using appropriate starting values in the estimation process. More recently, SriDaran et al. (2022) attempt to obtain least squares estimates of Renshaw-Haberman parameters by changing the distributional assumption in MLE to Gaussian.<sup>3</sup> However, their attempt still entails a computationally demanding iterative Newton-Raphson algorithm, and does not aim to solve the convergence problem.

### 3.3.3 Existing Methods for Expediting Estimation

So far as we aware, there have been two major attempts to expedite estimation for the Renshaw-Haberman model. These methods are reviewed in this subsection.

#### The H1 Model

Renshaw and Haberman (2006) attempt to improve estimation efficiency by simplifying the structure of the Renshaw-Haberman model. Specifically, they consider setting  $c_x$  in the original Renshaw-Haberman Model to  $1/p$ , where  $p$  represents the number of ages covered by the data set. The resulting model, given by

$$y_{x,t} := \log(m_{x,t}) = a_x + b_x k_t + \frac{1}{p} \gamma_{t-x} + \varepsilon_{x,t} \quad (3.11)$$

is often referred to as the *H1 model*, and is further discussed by Haberman and Renshaw (2011). The H1 model may be further reduced by setting  $b_x = 1/p$ . This further simplification would result in the classical age-period-cohort (APC) model (Hobcraft et al., 1985):

$$y_{x,t} := \log(m_{x,t}) = a_x + \frac{1}{p} k_t + \frac{1}{p} \gamma_{t-x} + \varepsilon_{x,t}. \quad (3.12)$$

---

<sup>3</sup>The relationship between least squares and a Gaussian assumption in MLE is discussed in Section 3.7.

Reducing the model structure may result in a faster convergence; however, a reduced model structure may no longer provide an adequate fit.

### The Hunt-Villegas Method

Hunt and Villegas (2015) argue that the slow convergence of the MLE for the Renshaw-Haberman model is due possibly to an approximate identifiability issue.

Specifically, Hunt and Villegas (2015) show that if  $k_t$  in (3.8) follows a perfect straight line, then there exists an approximately invariant parameter transformation. In other words, parameters are not unique even if the four parameter constraints specified in (3.2) and (3.9) are imposed.

Empirically, the estimates of  $k_t$  typically exhibit a steady downward trend due to mortality improvements, but the trend is not perfectly linearly. As such, this identification problem is ‘approximate’ rather than ‘exact’. The approximate identification problem means that there exist different sets of parameters that would lead to different allocations between the time effect and cohort effect but approximately the same fit to the historical data. This phenomenon could potentially make the optimization procedure slow and unstable.

To resolve the approximate identifiability issue, Hunt and Villegas (2015) suggest imposing an additional constraint:

$$\sum_{s=t_1-x_p}^{t_n-x_1} (s - \bar{s})\gamma_s = 0, \tag{3.13}$$

where  $\bar{s}$  represents the average year-of-birth over the years-of-birth covered by the data set. This constraint ensures that  $\gamma_s$  does not follow a linear trend over the years-of-birth covered by the data set. To see why this is true, we can treat (3.13) as a requirement that the sample covariance between  $\gamma_s$  (which has a zero mean due to another identifiability constraint) and year-of-birth  $s$  is zero. Hunt and Blake (2021) mention that the additional constraint has significant demographic significance. Specifically, it is conceivable that  $\gamma_s$  is approximately trendless, because systematic changes in mortality over time should have been captured by  $k_t$ .

Imposing the additional constraint can mitigate the approximate identification issue. It also shrinks the parameter space over which the Newton-Raphson’s algorithm has to cover, thereby stabilizing and accelerating the optimization.

## 3.4 The Proposed Method

### 3.4.1 Motivation

The existing methods for expediting Renshaw-Haberman estimation are both based on the MLE framework, and therefore the requirement of a distributional assumption (which may turn out to be wrong) remains. Also, both methods rely on a reduction in parameter space, so that the improve estimation efficiency at the expense of goodness-of-fit.

The aforementioned limitations motivate us to tackle the estimation challenge from a different angle. Specifically, we develop a least squares approach for the Renshaw-Haberman model, in which computational efficiency is achieved through some closed-form SVD solutions. The proposed approach has the following merits:

- *A sharper objective function*

Compared to MLE, the proposed least squares method is based on a different objective function. We show empirically in Section 3.6 that the objective function in our proposed method is sharper, thereby resulting in a faster convergence. Also, the objective function is optimized in part by some SVD closed-form solutions, so that our proposed method is more computational efficient.

- *Less reliant on distributional assumptions*

The MLE approach requires a strict distributional assumption. The commonly used Poisson death count assumption is not without criticism. For instance, the over-dispersion problem arising from population heterogeneity would render the Poisson assumption inappropriate. Although this problem may be mitigated by assuming a more flexible death count distribution, such as negative binomial (Li et al., 2009),

the number of parameter would increase and consequently model estimation may be even slower. In contrast, in our proposed method, the objective function, defined as the sum of squared differences between observed and fitted log mortality rates, is not developed upon any specific distributional assumption.

- *Seamless integration with existing methods for expediting estimation*

The proposed estimation method applies to not only the original Renshaw-Haberman model but also its reduced versions including the H1 model. It can also be implemented with the Hunt-Villegas method to further improve computational efficiency.

### 3.4.2 Main Optimization: Alternating Minimization

When a least squares approach is used to estimate the Renshaw-Haberman model, the optimization problem can be expressed as

$$\min_{\mathbf{a}, \mathbf{b}, \mathbf{k}, \mathbf{c}, \gamma} \sum_{x,t} (y_{x,t} - (a_x + b_x k_t + c_x \gamma_{t-x}))^2. \quad (3.14)$$

The parameter constraints specified in (3.2) and (3.9) are imposed to stipulate a unique solution.

Unlike the the least squares optimization problem (3.4) for the Lee-Carter model, (3.14) is a much more challenging problem that cannot be easily solved. To overcome the estimation challenge, we can consider an *alternating minimization* strategy, in which the shape parameter vector  $\mathbf{a}$ , the age-period component  $(\mathbf{b}, \mathbf{k})$  and the age-cohort component  $(\mathbf{c}, \gamma)$  are updated in turns, while other components are held fixed.

The core iteration is outlined in Algorithm 2. Each cycle of iteration is composed of several steps. Step 2 is simple, as it updates the shape vector  $\mathbf{a}$  through an explicit averaging formula. Step 3 is also straightforward as it just fits a Lee-Carter structure, through a SVD, to the residual after the removal of the shape vector and age-cohort effects. Note that the updated values of  $k_t$  from Step 3 sum to zero, because the input residual matrix  $[y_{x,t} - a_x - c_x \gamma_{t-x}]_{x,t}$  is row-centered following the implementation of (3.15) in Step 2.

---

**Algorithm 2** The main iteration

---

1. Set initial values of  $\boldsymbol{\theta} := (\mathbf{a}, \mathbf{b}, \mathbf{k}, \mathbf{c}, \boldsymbol{\gamma})$ .
2. Update  $\mathbf{a}$  while  $\mathbf{b}, \mathbf{k}, \mathbf{c}$  and  $\boldsymbol{\gamma}$  are fixed:

$$\min_{\mathbf{a}} \sum_{x,t} \underbrace{((y_{x,t} - b_x k_t - c_x \gamma_{t-x}) - a_x)^2}_{\text{given}}. \quad (3.15)$$

This sub-optimization is accomplished with the following explicit solution:

$$a_x := \frac{1}{n} \sum_{t=t_1}^{t_n} (y_{x,t} - b_x k_t - c_x \gamma_{t-x}) = \frac{1}{n} \sum_{t=t_1}^{t_n} (y_{x,t} - c_x \gamma_{t-x}). \quad (3.16)$$

The last step is the above originates from the identification constraint  $\sum_{t=t_1}^{t_n} k_t = 0$ .

3. Fixing  $\mathbf{a}, \mathbf{c}$  and  $\boldsymbol{\gamma}$ , update  $\mathbf{b}$  and  $\mathbf{k}$ :

$$\min_{\mathbf{b}, \mathbf{k}} \sum_{x,t} \underbrace{((y_{x,t} - a_x - c_x \gamma_{t-x}) - b_x k_t)^2}_{\text{given}}. \quad (3.17)$$

This sub-optimization is accomplished by applying a first-order SVD to the matrix of  $y_{x,t} - a_x - c_x \gamma_{t-x}$ .

4. Update  $\mathbf{c}$  and  $\boldsymbol{\gamma}$  while  $\mathbf{a}, \mathbf{b}$  and  $\mathbf{k}$  are fixed:

$$\min_{\mathbf{c}, \boldsymbol{\gamma}} \sum_{x,t} \underbrace{((y_{x,t} - a_x - b_x k_t) - c_x \gamma_{t-x})^2}_{\text{given}}, \quad (3.18)$$

This sub-optimization is accomplished by an *iterative SVD algorithm*, described in Section 3.4.2. Then, the estimates of  $\mathbf{a}$  and  $\boldsymbol{\gamma}$  are adjusted so that the identifiability constraint  $\sum_{s=t_1-x_p}^{t_n-x_1} \gamma_s = 0$  is satisfied:

$$\boldsymbol{\gamma} := \boldsymbol{\gamma} - \bar{\boldsymbol{\gamma}}, \quad \mathbf{a} := \mathbf{a} + \mathbf{c}\bar{\boldsymbol{\gamma}}, \quad \text{where } \bar{\boldsymbol{\gamma}} = \frac{1}{n+p-1} \sum_{s=t_1-x_p}^{t_n-x_1} \gamma_s. \quad (3.19)$$

5. Repeat Steps 2-4 until the convergence criterion is satisfied.
-

Step 4, however, represents a much more complex optimization challenge, because it cannot be directly translated into a traditional PCA problem. Therefore, efficiently solving (3.18) in Step 4 is a crucial milestone of our research question. In the next subsection, we show that Step 4 can be formulated as a PCA problem with missing values, which can be efficiently solved in an iterative manner.

Finally, Step 5 requires a convergence criterion. In this chapter, convergence is achieved when the relative change in the objective function, as defined by (3.14), falls below a pre-determined small threshold. Algorithm 2 always converges, since each of Steps 2-4 in the algorithm consistently decreases the objective function, and the objective function ( $L^2$  error) is inherently bounded below by zero.

### 3.4.3 Updating the Age-Cohort Parameters: PCA with Missing Values via an Iterative SVD

In this subsection, we develop a method to overcome the optimization challenge in Step 4 of Algorithm 2.

First, let us explain in more detail the optimization challenge we are facing. Let  $z_{x,t} := y_{x,t} - a_x - b_x k_t$  be the input for the sub-optimization problem in Step 4. We may arrange the values of  $z_{x,t}$  in a  $p \times n$  age-period (age-time) matrix as follows:

$$\mathbf{Z}_{ap} := \begin{bmatrix} z_{x_1,t_1} & z_{x_1,t_2} & \cdots & \cdots & z_{x_1,t_{n-1}} & z_{x_1,t_n} \\ z_{x_2,t_1} & z_{x_2,t_2} & \cdots & \cdots & z_{x_2,t_{n-1}} & z_{x_2,t_n} \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ z_{x_{p-1},t_1} & z_{x_{p-1},t_2} & \cdots & \cdots & z_{x_{p-1},t_{n-1}} & z_{x_{p-1},t_n} \\ z_{x_p,t_1} & z_{x_p,t_2} & \cdots & \cdots & z_{x_p,t_{n-1}} & z_{x_p,t_n} \end{bmatrix} \quad (3.20)$$

We are unable to update the age-cohort component  $(\mathbf{c}, \boldsymbol{\gamma})$  by applying a SVD directly to this age-period matrix, because age and cohort are not orthogonal in  $\mathbf{Z}_{ap}$ .

To solve the sub-optimization, we first rearrange the input values in a  $p \times (n + p - 1)$

age-cohort data matrix:

$$\mathbf{Z}_{ac} := \begin{bmatrix} \times & \times & \cdots & \cdots & \times & z_{x_1, t_1} & \cdots & z_{x_1, t_{n-1}} & z_{x_1, t_n} \\ \times & \times & \cdots & \cdots & z_{x_2, t_1} & z_{x_2, t_2} & \cdots & z_{x_2, t_n} & \times \\ \vdots & \vdots & & \ddots & & & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & & & \ddots & & \vdots & \vdots \\ \times & z_{x_{p-1}, t_1} & \cdots & z_{x_{p-1}, t_{n-1}} & z_{x_{p-1}, t_n} & \cdots & \cdots & \times & \times \\ z_{x_p, t_1} & z_{x_p, t_2} & \cdots & z_{x_p, t_n} & \times & \cdots & \cdots & \times & \times \end{bmatrix}. \quad (3.21)$$

In  $\mathbf{Z}_{ac}$ , each  $\times$  represents a missing value, which arise because the oldest and youngest cohorts are not completely observed. For instance, for the youngest cohort of individuals who are born in year  $t_n - x_1$ , only one observed value ( $z_{x_1, t_n}$ ) is available. Similar arrangements of mortality data are also made by researchers such as Basellini and Camarda (2022). In the spirit of this rearrangement, the sub-optimization problem can be expressed as

$$\min_{c, \gamma} \sum_{x, s \in \mathcal{O}} (z_{x, s} - c_x \gamma_s)^2, \quad (3.22)$$

where  $s := t - x$  represents year-of-birth and  $\mathcal{O}$  is the set of indices of the observed values.

If  $\mathbf{Z}_{ac}$  contains no missing value, then we can solve (3.22) readily by applying a SVD to  $\mathbf{Z}_{ac}$ . However, given the presence of missing values, the sub-optimization boils down to a first-order PCA with missing values.

Handling missing values in PCA is a complex problem in statistics and machine learning. The technical challenges entailed are discussed in the work of Ilin and Raiko (2010). The same paper also provides a comprehensive review of the classical algorithms for PCA with missing values. In the modern statistics and machine learning literature, there exist advanced techniques for handling missing data in PCA, such as matrix completion with a nuclear norm regularization (Mazumder et al., 2010). However, such advanced techniques are designed for extremely large-scale and sparse matrices. Additionally, their primary goal is to predict the missing values (matrix completion) rather than finding the optimal least squares solution.

In this study, we utilize a method called the *iterative SVD algorithm*. This algorithm begins with an imputation of the missing values, typically with row-wise means of the

---

**Algorithm 3** The Iterative SVD Algorithm
 

---

1. Obtain the initial approximate complete matrix  $\mathbf{Z}_{ac}^*$  by imputing the missing values in  $\mathbf{Z}_{ac}$  with row-wise means of the observed values in  $\mathbf{Z}_{ac}$ .
2. Apply a SVD to the approximate complete matrix  $\mathbf{Z}_{ac}^*$ . Incorporating the identifiability constraint  $\sum_{x=x_1}^{x_p} c_x = 1$ , the updated estimates of  $\mathbf{c}$  and  $\boldsymbol{\gamma}$  are given by the following expressions

$$\mathbf{c} := \frac{\mathbf{u}_c}{\mathbf{1}^T \mathbf{u}_c}, \quad \boldsymbol{\gamma} := (\mathbf{1}^T \mathbf{u}_c) \cdot \mathbf{Z}_{ac}^{*T} \mathbf{u}_c, \quad (3.23)$$

where  $\mathbf{u}_c$  is the first left-singular vector of the approximate complete matrix  $\mathbf{Z}_{ac}^*$ . Note that the constraint  $\sum_{t-x} \gamma_{t-x} = 0$  is incorporated in Algorithm 2 through (3.19).

3. Update the missing values in  $\mathbf{Z}_{ac}$  by a PCA reconstructions with the estimates of  $\mathbf{c}$  and  $\boldsymbol{\gamma}$  obtained from Step 2. In particular, the missing values in  $\mathbf{Z}_{ac}$  are imputed as  $\mathbf{c}\boldsymbol{\gamma}^T$ , while the observed values in  $\mathbf{Z}_{ac}$  remain unchanged.
  4. Repeat Steps 2 and 3 until the relative change in the objective function specified by (3.22) is smaller than a certain pre-determined tolerance level.
- 

observed values in the input matrix. This creates an approximate complete matrix, to which a PCA can be applied to obtain singular vectors (parameter estimates). Then, a PCA reconstruction is employed to generate an improved imputation of the missing values. The process is repeated until convergence is achieved. The implementation of the iterative SVD algorithm in the context of our research is presented in Algorithm 3.

While the iterative SVD algorithm appears to be a suitable method for solving PCA with missing values, it is not immediately clear why this algorithm addresses our specific  $L^2$  minimization problem with missing values. To elucidate this, in Appendix C, we prove that the iterative SVD algorithm minimizes the target loss function specified in (3.22), and that the iterative SVD algorithm always converges.

## 3.5 Integrating the Proposed Method with the Existing Methods

We may implement our proposed method with one or both of the the existing methods (H1 and Hunt-Villegas) to further boost estimation speed.

### 3.5.1 Implementing with the H1 Model

Our proposed method can be applied to the variants of the Renshaw-Haberman model, including the H1 model discussed in Section 3.3. For the H1 model, least squares estimation can be formulated as the following the optimization problem:

$$\min_{a,b,k,\gamma} \sum_{x,t} \left( y_{x,t} - \left( a_x + b_x k_t + \frac{1}{p} \gamma_{t-x} \right) \right)^2, \quad (3.24)$$

and the following three identification constraints can be used to stipulate parameter uniqueness:

$$\sum_{x=x_1}^{x_p} b_x = 1, \quad \sum_{t=t_1}^{t_n} k_t = 0, \quad \sum_{t-x=t_1-x_p}^{t_n-x_1} \gamma_{t-x} = 0. \quad (3.25)$$

The main algorithm of our proposed method for the H1 model is identical to Algorithm 2 for the Renshaw-Haberman model, except that  $c_x$  is always set to  $= 1/p$ . Interestingly, as explained below, further computational simplifications can be achieved when our proposed method is applied to the H1 model.

For the H1 model, we can update  $\gamma$  using explicit formulas, thereby eliminating the need for iterative algorithms. To explain, we first express the sub-optimization problem for updating  $\gamma$  (Step 4 in Algorithm 2) in the H1 model as follows:

$$\min_{\gamma} \sum_{x,t} \left( \underbrace{(y_{x,t} - a_x - b_x k_t)}_{\text{given}} - \frac{1}{p} \gamma_{t-x} \right)^2. \quad (3.26)$$

The above can be rewritten in an age-cohort dimension as

$$\min_{\gamma} \sum_{x,s \in \mathcal{O}} \left( z_{x,s} - \frac{1}{p} \gamma_s \right)^2, \quad (3.27)$$

where  $z_{x,s} := y_{x,s} - a_x - b_x k_s$  denotes the residual from Step 3 in Algorithm 2,  $s := t - x$  represents year-of-birth, and  $\mathcal{O}$  is the set of the indices for the observed values in  $\mathbf{Z}_{ac}$  (the matrix of  $z_{x,s}$  in age-cohort dimension).

Noticing that (3.27) is separable, we can rewrite it as:

$$\min_{\gamma} \sum_s \sum_{x \in \mathcal{O}_s} \left( z_{x,s} - \frac{1}{p} \gamma_s \right)^2, \quad (3.28)$$

where  $\mathcal{O}_s$  denotes the set of the indices for the observed values in column  $s$  of  $\mathbf{Z}_{ac}$ . Note that this convenient separability does not hold for the general Renshaw-Haberman model with  $c_x \neq 1/p$ , since each summand  $(z_{x,s} - c_x \gamma_s)^2$  depends on both age  $x$  and year-of-birth  $s$ .

The separability enables us to solve the target optimization problem (3.24) by solving the following for each  $s$ :

$$\min_{\gamma_s} \sum_{x \in \mathcal{O}_s} \left( z_{x,s} - \frac{1}{p} \gamma_s \right)^2. \quad (3.29)$$

For a given  $s$ , (3.29) is a simple linear regression with no intercept and a slope of

$$\hat{\gamma}_s = \frac{p}{n_s} \sum_{x \in \mathcal{O}_s} z_{x,s}, \quad (3.30)$$

where  $n_s := |\mathcal{O}_s|$  is the cardinality of  $\mathcal{O}_s$ . Applying (3.30) for every year-of-birth  $s$  covered by the data set yields an update of  $\gamma$ .

### 3.5.2 Implementing with the Hunt-Villegas Method

This subsection explains how our proposed method can be utilized with the H1 model and the Hunt-Villegas method.

Recall that the Hunt-Villegas method originates from an approximate identifiability problem of the Renshaw-Haberman model and its variants. For the H1 model, Hunt and Villegas (2015) show that if  $k_t$  follows a perfect straight line, i.g.,  $k_t = K(t - \bar{t})$ , where  $K$  is a constant that is less than zero and  $\bar{t} = (t_n + t_1)/2$  represents the mid-point of the calibration window, then there exists the following invariant transformation that is equivalent to  $\{a_x, b_x, k_t, \gamma_s\}$ :

$$\left\{ a_x + \frac{g}{p}(x - \bar{x}), \frac{K}{K - g}b_x - \frac{g}{p(K - g)}, \frac{K - g}{K}k_t, \gamma_s + g(s - \bar{s}) \right\}, \quad (3.31)$$

where  $\bar{x} = (x_p + x_1)/2$  represents the mid-point of the age range under consideration and  $g$  is a real constant. In practice, the trend in  $k_t$  close to but not perfectly linear, so that an *approximate* identifiability problem exists. This approximate identifiability problem may adversely affect convergence of the estimation algorithm. Hunt and Villegas (2015) propose to mitigate approximate identifiability problem by imposing the extra constraint specified in (3.13).

Hunt and Villegas (2015) proposed a modified Newton-Raphson method to impose (3.13) in Poisson ML estimation of model parameters. Specifically, in each iteration of the Newton-Raphson algorithm, they determine the values of  $K$  and  $g$  in the invariant transformation such that (3.13) is satisfied; then, the invariant transformation is applied to adjust the parameter estimates.

However, it turns out that the modified Newton-Raphson method is not applicable to our alternating minimization scheme. To explain, let us suppose that in one iteration we have updated the value of  $\boldsymbol{\gamma}$  using the closed-form solution provided in (3.30). This update is guaranteed to decrease the value of the overall objective function specified in (3.24). However, if we adjust the estimates of  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\mathbf{k}$ , and  $\boldsymbol{\gamma}$  using the approximate invariant transformation specified in (3.31) to make (3.13) hold, then the resulting estimates may lead to a higher (less optimal) value of (3.24), since the transformation is only approximately (rather than exactly) invariant. If the value of the objective function increases in some iterations, the alternating minimization algorithm may diverge.

We propose to incorporate the additional constraint specified in (3.31) by using a Lagrange multiplier in the update of  $\boldsymbol{\gamma}$  in model H1. Incorporating (3.31), we aim to solve

the following constrained optimization problem in the update of  $\gamma$ :

$$\min_{\mathbf{c}, \gamma} \sum_{x, s \in \mathcal{O}} (z_{x, s} - c_x \gamma_s)^2, \quad \text{s.t.} \quad \sum_{s=t_1-x_p}^{t_n-x_1} \gamma_s (s - \bar{s}) = 0. \quad (3.32)$$

Then, the Lagrangian can be written as:

$$\mathcal{L}(\gamma, \lambda) = \sum_{x, s \in \mathcal{O}} \left( z_{x, s} - \frac{1}{p} \gamma_s \right)^2 + 2\lambda \sum_{s=t_1-x_p}^{t_n-x_1} \gamma_s (s - \bar{s}), \quad (3.33)$$

where  $\lambda$  represents the Lagrange multiplier.<sup>4</sup>

Unlike the unconstrained case in which objective function is separable, the minimization of (3.33) is non-separable because the Lagrange multiplier  $\lambda$  applies to all years-of-birth  $s = t_1 - x_p, \dots, t_n - x_1$ . To obtain the solution to (3.32), we derive the first-order partial derivatives of  $\mathcal{L}(\gamma, \lambda)$  with respect to  $\gamma$  and  $\lambda$ , and set them to zero:

$$\frac{\partial \mathcal{L}}{\partial \gamma_s} = -\frac{2}{p} \cdot \sum_{x \in \mathcal{O}_s} \left( z_{x, s} - \frac{1}{p} \gamma_s \right) + 2\lambda (s - \bar{s}) = 0, \quad s = t_1 - x_p, \dots, t_n - x_1. \quad (3.34)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{s=t_1-x_p}^{t_n-x_1} \gamma_s (s - \bar{s}) = 0. \quad (3.35)$$

From (3.34), we obtain the following expression of  $\gamma_s$  in terms of  $\lambda$ :

$$\gamma_s = \frac{p}{n_s} \cdot \left[ \left( \sum_{x \in \mathcal{O}_s} z_{x, s} \right) - p\lambda (s - \bar{s}) \right], \quad (3.36)$$

for  $s = t_1 - x_p, \dots, t_n - x_1$ . Plugging (3.36) into (3.35), we get the optimal solution for  $\lambda$ :

$$\hat{\lambda} = \frac{1}{p} \cdot \left[ \sum_{s=t_1-x_p}^{t_n-x_1} \frac{(s - \bar{s})^2}{n_s} \right]^{-1} \cdot \sum_{s=t_1-x_p}^{t_n-x_1} \left[ \frac{s - \bar{s}}{n_s} \cdot \sum_{x \in \mathcal{O}_s} z_{x, s} \right]. \quad (3.37)$$

Plugging (3.37) back into (3.36) gives the solution to  $\gamma_s$  for  $s = t_1 - x_p, \dots, t_n - x_1$ .

---

<sup>4</sup>We multiply  $\lambda$  by two for computational convenience. The use of  $2\lambda$  instead of  $\lambda$  makes no difference in the final solution.

## 3.6 Numerical Illustrations

In this section, we present various experiments to illustrate our proposed least square method for estimating the Renshaw-Haberman model. The data used are obtained from the Human Mortality Database (2023). They cover a calibration window of 1950-2019 and an age range of 60-89. All of the experiments are performed using a desktop with an Intel Core i9-10900 CPU at 2.80 GHZ, 16 GB of RAM, and Windows 11 Education (64 bits).

All estimation methods under consideration involve an iterative procedure. While it is usual to base the convergence criterion of an iterative procedure on the *absolute* change in the objective function in each iteration, we consider the *relative* change instead, because the objective functions of Poisson ML estimation and the proposed least squares estimation have rather different magnitudes. Basing the convergence criterion on relative changes allows us to compare the two streams of estimation methods more fairly.

We use  $\delta$  to represent the tolerance level used in main estimation algorithms. The choice of  $\delta$  is admittedly subjective. The `StMoMo` package, by default, uses a tolerance level of  $10^{-4}$  and the absolute change in the log-likelihood function as the convergence criterion when fitting the Renshaw-Haberman model. Considering the size of the datasets we are using, the values of the maximized log-likelihood functions (when models are fitted using Poisson MLE) have a magnitude of  $10^4$ . Since we are using basing our convergence criterion on relative changes, the baseline value of  $\delta$  is set to  $10^{-8}$  to match the standard used in the `StMoMo` package.

### 3.6.1 Comparing Least Squares with Poisson MLE

We first compare the following three methods for fitting the Renshaw-Haberman model:

- *RH-MLE*: The Renshaw-Haberman model estimated with Poisson MLE;
- *RH-MLE-HV*: The Renshaw-Haberman model estimated with Poisson MLE and the Hunt-Villegas method;

- *RH-LS*: The Renshaw-Haberman model estimated with our proposed least squares method.

The baseline results are obtained using the data from the male populations of England and Wales (E&W) and the US. These data sets are considered in prominent works on stochastic mortality modelling (e.g., Cairns et al., 2009, 2011).

The results are summarized in Table 3.1, from which we observe that RH-LS consumes significantly less computation time compared to RH-MLE. Reductions in computational time are over 90% in general. Figure 3.1 shows that for a given data set, the parameter estimates from RH-LS and RH-MLE are highly similar. It is not surprising that the parameter estimates from the two estimations methods are not identical, because they are based on different objective functions. As expected, RH-LS (which minimizes the  $L^2$  error) yields a lower (less preferred) log-likelihood but a smaller  $L^2$  error compared to RH-MLE.

Table 3.1:  $L^2$  errors, log-likelihood values, and computation times for RH-MLE, RH-LS and RH-MLE-HV, based on E&W male and US male datasets.

Data		<i>RH-MLE</i>	<i>RH-LS</i>	<i>RH-MLE-HV</i>
E&W male	$L^2$ error	0.578	<b>0.565</b>	0.581
	Log-likelihood	− <b>12843</b>	−12890	−12853
	Computing time (seconds)	336.68	38.72	<b>20.39</b>
US male	$L^2$ error	0.472	<b>0.465</b>	0.473
	Log-likelihood	− <b>17736</b>	−17828	−17744
	Computing time (seconds)	228.72	<b>10.44</b>	25.95

We also observe from Table 3.1 that RH-MLE-HV takes shorter computational times relative to RH-MLE, and is comparable to RH-LS in terms of computational efficiency. However, it is important to note that RH-MLE-HV results in less desirable  $L^2$  errors and log-likelihoods compared to both RH-MLE and RH-LS, because RH-MLE-HV entails an additional constraint which makes the model more restrictive. That said, RH-MLE-HV improves computational efficiency at the expense of goodness-of-fit.

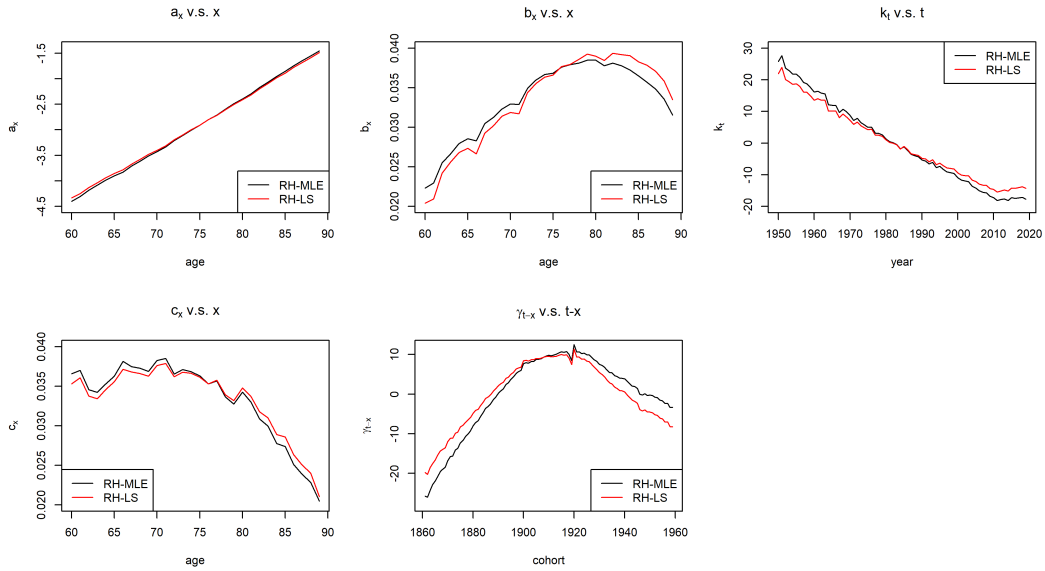


Figure 3.1: Parameter estimates derived from the E&W male dataset, RH-MLE and RH-LS.

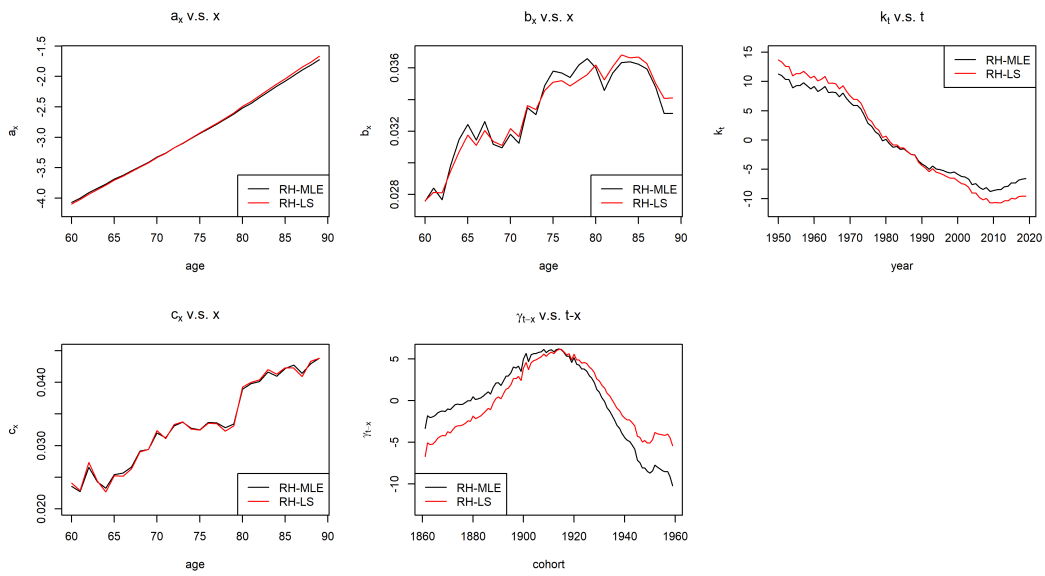


Figure 3.2: Parameter estimates derived from the US male dataset, RH-MLE and RH-LS.

To demonstrate the consistency in computation time reduction, we compare RH-LS with RH-MLE using eight alternative data sets: E&W female, US female, Australia male and female, Canada male and female, and the Netherlands male and female. Reported in Table 3.2, the results show that RH-LS takes a significantly shorter computation time compared to RH-MLE for all of the ten datasets under consideration. While it is more computationally efficient, our proposed method preserves goodness-of-fit in the sense that it results in smaller  $L^2$  errors and similar log-likelihood values compared to the Poisson ML approach.

As a robustness check, we repeat the numerical experiment using a wider age range of 0-89. The resulting computation times for all of the ten datasets under consideration are presented in Table 3.3. As expected, both estimation methods take longer to converge due to the increased parameter space introduced by the expanded age range. For RH-MLE, the increase is significantly higher, leading computational times ranging from approximately ten minutes to one hour per estimation. This level of fitting speed renders tasks that require repeated model re-fitting, such as uncertainty estimation via the bootstrapping, practically infeasible. In contrast, for RH-LS, the computation times needed for the extended age range remain modest.

Table 3.2: Computation times for RH-MLE, RH-LS and RH-MLE-HV, based on all of the ten datasets under consideration, with an age range of 60-89.

Computation time (s)			
Data	<i>RH-MLE</i>	<i>RH-LS</i>	<i>RH-MLE-HV</i>
E&W male	336.68	38.72	<b>20.39</b>
E&W female	132.21	36.28	<b>10.26</b>
US male	228.72	<b>10.44</b>	25.95
US female	30.23	<b>26.18</b>	42.41
Australia male	53.10	<b>5.76</b>	18.28
Australia female	59.12	<b>10.81</b>	12.76
Canada male	118.51	<b>16.17</b>	29.98
Canada female	59.02	<b>11.71</b>	43.31
The Netherlands male	101.22	<b>13.58</b>	15.34
The Netherlands female	39.25	24.52	<b>5.62</b>

Table 3.3: Computation times for RH-MLE, RH-LS and RH-MLE-HV, based on all of the ten datasets, with an age range of 0-89.

Computation time (s)			
Data	<i>RH-MLE</i>	<i>RH-LS</i>	<i>RH-MLE-HV</i>
E&W male	891.71	<b>22.23</b>	204.25
E&W female	262.38	<b>31.31</b>	166.38
US male	3210.81	210.22	<b>32.03</b>
US female	1320.94	280.12	<b>199.01</b>
Australia male	2015.08	212.40	<b>32.53</b>
Australia female	715.67	194.17	<b>12.22</b>
Canada male	1188.53	<b>73.21</b>	132.10
Canada female	128.22	<b>22.89</b>	68.99
The Netherlands male	601.22	116.12	<b>49.32</b>
The Netherlands female	900.29	<b>98.23</b>	148.42

### 3.6.2 Sharpness of Objective Functions

One may wonder why the proposed least squares method is more computationally efficient than the Poisson maximum likelihood approach, while producing a comparable goodness-of-fit. In this sub-section, we attempt to account for the superiority of our proposed approach by considering the sharpness of the objective functions used in each of the candidate estimation methods.

In a study of maximum likelihood estimation of various stochastic mortality models, Cairns et al. (2009) mentioned that “the likelihood function will be close to flat in certain dimensions.” As a result of such flatness, over the iterative estimation process, parameter estimates tend to stray around the area of parameter space over which the resulting log-likelihood values are similar, thereby resulting in a slow convergence. It follows that a faster convergence can be achieved if the objective function is sharper, in the sense that the change in its value in each iteration of the estimation algorithm tends to be larger.

Should there be flatness in certain dimensions of the objective function, parameter estimates tend to be sensitive to the tolerance level  $\delta$  used in the iterative estimation process. To compare the sharpness of the objective functions for RH-MLE and RH-LS, we estimate the Renshaw-Haberman model with Poisson MLE and the proposed least squares methods to the US male dataset, for different tolerance levels:  $10^{-6}$ ,  $10^{-7}$  and  $10^{-8}$ .

Figure 3.3 reveals that parameter estimates obtained from Poisson MLE are quite sensitive to the tolerance level. The reduction in tolerance level does not materially improve the log-likelihood value, but comes with a substantially longer computation time as reported in Table 3.4. In contrast, Figure 3.4 shows that the parameter estimates obtained from the proposed least squares method are more robust with respect to the tolerance level. Additionally, compared to Poisson MLE, the increase in computational time as the tolerance level reduces is moderate, as also shown in Table 3.4. These outcomes suggest that the objective function for the proposed method is sharper, offering a reason as to why the proposed method is more computationally efficient.

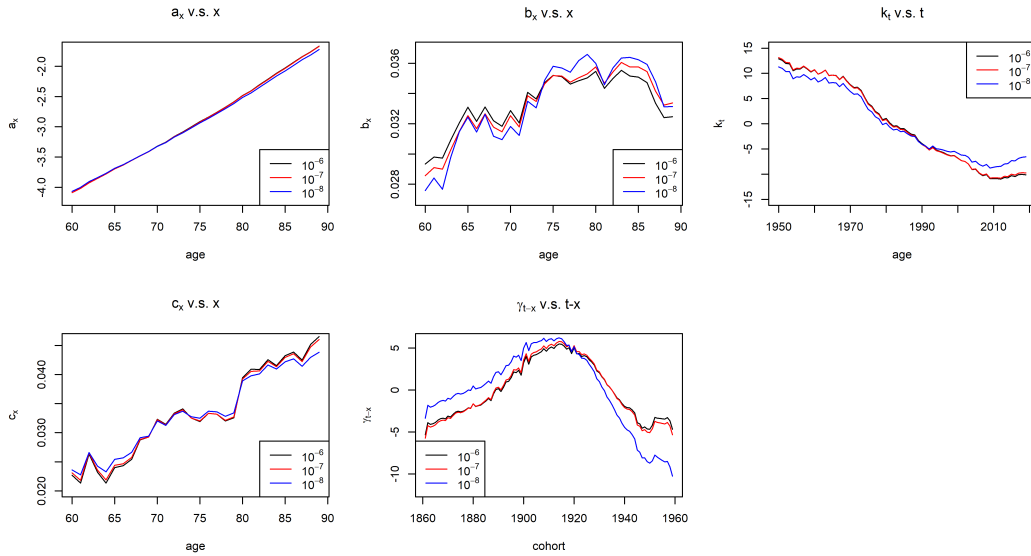


Figure 3.3: Poisson maximum likelihood estimates of the parameters in the Renshaw-Haberman model for different tolerance levels:  $10^{-6}$ ,  $10^{-7}$  and  $10^{-8}$ ; US male dataset.

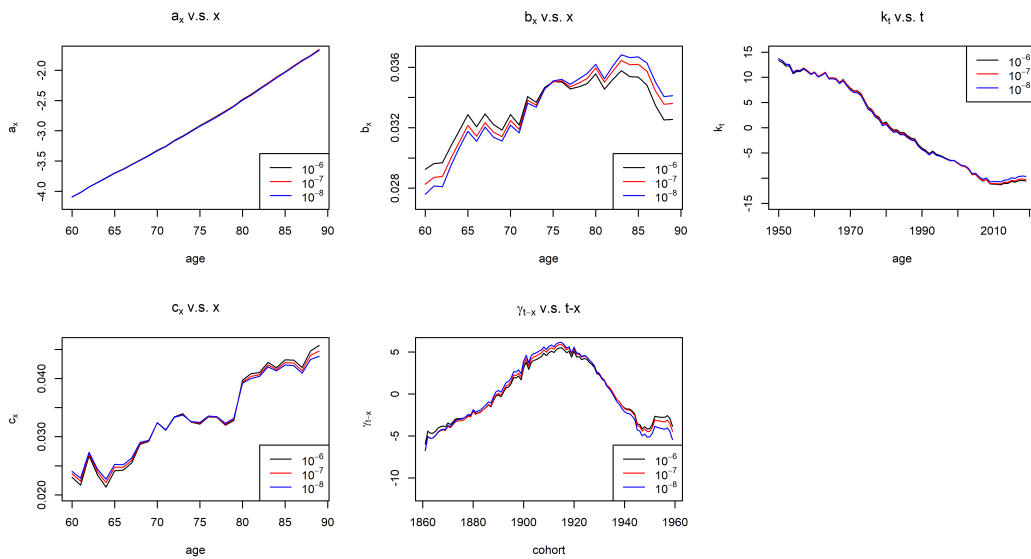


Figure 3.4: Least squares estimates of the parameters in the Renshaw-Haberman model for different tolerance levels:  $10^{-6}$ ,  $10^{-7}$  and  $10^{-8}$ ; US male dataset.

Table 3.4:  $L^2$  errors, log-likelihoods, and computing times for RH-MLE and RH-LS when three different tolerance levels are used, US male.

	Tolerance Level	<i>RH-MLE</i>	<i>RH-LS</i>
$L^2$ error	$10^{-6}$	0.4734	0.4661
	$10^{-7}$	0.4727	0.4654
	$10^{-8}$	0.4720	0.4652
Log-likelihood	$10^{-6}$	-17749	-17828
	$10^{-7}$	-17745	-17827
	$10^{-8}$	-17737	-17827
Time (s)	$10^{-6}$	12.28	3.96
	$10^{-7}$	20.67	4.92
	$10^{-8}$	283.87	11.28

### 3.6.3 Implementing with the H1 Model

In this subsection, we implement our proposed least squares estimation method with the H1 model and/or the Hunt-Villegas method to further boost estimation efficiency. The following four settings are considered:

- *H1-MLE*: The H1 model estimated with Poisson MLE;
- *H1-MLE-HV*: The H1 model estimated with Poisson MLE plus the Hunt-Villegas method;
- *H1-LS*: The H1 model estimated with the proposed least squares method;
- *H1-LS-HV*: The H1 model estimated with the proposed least squares method plus the Hunt-Villegas method.

Table 3.5:  $L^2$  errors, log-likelihoods, and computation times for H1-MLE, H1-MLE-HV, H1-LS, and H1-LS-HV, E&W male and US male datasets.

Data		<i>H1-MLE</i>	<i>H1-LS</i>	<i>H1-MLE-HV</i>	<i>H1-LS-HV</i>
E&W male	$L^2$ error	0.682	<b>0.663</b>	0.721	0.697
	Log-likelihood	<b>-13149</b>	-13208	-13247	-13321
	Computation time (s)	62.25	9.79	19.05	<b>3.16</b>
US male	$L^2$ error	0.557	<b>0.545</b>	0.557	0.545
	Log-likelihood	<b>-18692</b>	-18817	-18699	-18816
	Computation time (s)	138.88	2.56	18.30	<b>2.12</b>

Table 3.5 presents the results for the four settings above, derived from the E&W male and US datasets. Comparing the results for H1-MLE (Table 3.5) and RH-MLE (Table 3.1), we notice that using the H1 model (a reduced version of the original Renshaw-Haberman model) helps reduce computation time. Nevertheless, compared to the full Renshaw-Haberman model, the H1 model yields lower log-likelihood values and higher  $L^2$

errors, suggesting that it produces a reduced goodness-of-fit. This outcome is expected, as the H1 model is a restricted version of the Renshaw-Haberman model with  $p$  fewer parameters.

On the other hand, from Table 3.5 we observe that the computation times for H1-LS are significantly less than those for H1-MLE, suggesting that the proposed least square estimation methods also offers an improvement in estimation efficiency when a restricted version of the Renshaw-Haberman model is considered. Finally, from Table 3.5 we notice that H1-LS-HV requires the least computation time among all settings under consideration. For the US male dataset, H1-LS-HV takes just slightly over 2 seconds, which is less than 1% of the computation time required when we estimate the original Renshaw-Haberman model with Poisson MLE.

For a more comprehensive analysis, we study the four settings with the eight alternative datasets considered in Section 3.6.1. Tabulated in Table 3.6, the results indicate the superiority of H1-LS over H1-MLE in terms of computation efficiency for all of the eight datasets under consideration. We also observe from Table 3.6 that for certain datasets, such as US female, Canada female, and the Netherlands female, fitting the H1 model is very time consuming (even though the H1 model is a restricted version of the original Renshaw-Haberman model), suggesting convergence issues that are possibly caused by the approximate identification problem discussed in Section 3.3. In these cases, using the Hunt-Villegas method could significantly reduce the computation time, and switching from Poisson MLE to the proposed least squares approach could lower the computation time even more.

Table 3.6: Computation times for H1-MLE, H1-MLE-HV, H1-LS, and H1-LS-HV, based on all of the ten datasets under consideration.

Data	Computing time (s)			
	<i>H1-MLE</i>	<i>H1-LS</i>	<i>H1-MLE-HV</i>	<i>H1-LS-HV</i>
E&W male	62.25	9.79	19.05	<b>3.16</b>
E&W female	58.23	1.12	7.58	<b>0.97</b>
US male	138.88	2.56	18.30	<b>2.12</b>
US female	702.34	89.93	3.58	<b>1.01</b>
Australia male	18.23	7.24	12.09	<b>2.55</b>
Australia female	9.12	5.58	6.61	<b>1.68</b>
Canada male	18.51	<b>2.69</b>	15.23	3.42
Canada female	162.26	78.62	7.28	<b>2.33</b>
The Netherlands male	21.22	2.76	9.38	<b>1.59</b>
The Netherlands female	425.69	85.51	2.85	<b>0.36</b>

### 3.6.4 Quantifying Parameter Uncertainty

One important application of stochastic mortality models is quantifying the uncertainty involved in mortality projections. A part of such uncertainty is parameter risk, which arises because parameters used in projecting future mortality are only estimates rather than known constants.

In this subsection, we demonstrate the advantage of our proposed method in the context of parameter uncertainty quantification. To this end, we consider a residual bootstrapping method, originally proposed by Koissi et al. (2006) and discussed in the review paper of Li (2014). The residual bootstrapping method is implemented using the following algorithm:

1. Estimate the Renshaw-Haberman model to the original dataset, and calculate the residuals of fit as  $y_{x,t} - \hat{y}_{x,t}$  for all  $x$  in the age range and  $t$  in the calibration window.
2. Sample residuals calculated in Step 1 with replacement, and generate a pseudo dataset by adding the sampled residuals to the fitted log mortality rates  $\hat{y}_{x,t}$  over the entire age range and calibration window.
3. Fit the model to the pseudo dataset produced in the immediate previous step. A collection of parameter estimates are then obtained.
4. Repeat Steps 2 and 3  $M$  times, where  $M$  is a large integer, say 1000. This step yields, for each parameter in the Renshaw-Haberman model, an empirical distribution of parameter estimates. From the empirical distribution, measures of parameter uncertainty such as standard error can be calculated.

We implement the algorithm with RH-MLE, RH-LS and RH-MLE-HV, using the EW male and US male datasets (with an age range of 60-89). The resulting standard errors for a subset of parameters are presented in Tables 3.7 and 3.8, respectively. It is observed that the three estimation methods produce standard errors of similar orders of magnitude.

Although the three estimations methods produce similar standard errors, they demand significantly different amounts of time to produce such standard errors. Since the residual

bootstrapping algorithm entails  $M$  re-estimations, the runtime of the algorithm is directly proportional to the amount of time needed for each estimation. When the US male dataset is considered and  $M$  is set to 1000, the amounts of time required by RH-MLE and RH-LS are 3812 minutes (2.65 days) and 174 minutes, respectively. This result underscores the advantage of our proposed estimation method in practical applications that involve repeated model estimation.

Table 3.7: Standard errors of selected parameter estimates for RH-MLE, RH-LS and RH-MLE-HV, based on the EW male dataset.

Parameter		<i>RH-MLE</i>	<i>RH-LS</i>	<i>RH-MLE-HV</i>
$a_x$	$a_{60}$	0.0114	0.0048	0.0039
	$a_{75}$	0.0072	0.0063	0.0065
	$a_{89}$	0.0034	0.0049	0.0054
$b_x$	$b_{60}$	0.0006	0.0007	0.0008
	$b_{75}$	0.0003	0.0003	0.0002
	$b_{89}$	0.0011	0.0010	0.0011
$k_t$	$k_{1950}$	0.3835	0.3710	0.3803
	$k_{1985}$	0.1954	0.1834	0.1805
	$k_{2019}$	0.3585	0.3622	0.3315
$c_x$	$c_{60}$	0.0013	0.0012	0.0015
	$c_{75}$	0.0004	0.0004	0.0004
	$c_{89}$	0.0012	0.0018	0.0013
$\gamma_{t-x}$	$\gamma_{1861}$	1.9783	1.3747	1.9718
	$\gamma_{1910}$	0.3678	0.3166	0.3650
	$\gamma_{1959}$	0.7992	0.8069	0.8508

Table 3.8: Standard errors of selected parameter estimates for RH-MLE, RH-LS and RH-MLE-HV, based on the US male dataset.

Parameter		<i>RH-MLE</i>	s.e. <i>RH-LS</i>	<i>RH-MLE-HV</i>
$a_x$	$a_{60}$	0.0096	0.0066	0.0038
	$a_{75}$	0.0058	0.0051	0.0045
	$a_{89}$	0.0201	0.0096	0.0053
$b_x$	$b_{60}$	0.0010	0.0007	0.0009
	$b_{75}$	0.0004	0.0003	0.0003
	$b_{89}$	0.0015	0.0010	0.0012
$k_t$	$k_{1950}$	0.9492	0.3231	0.2767
	$k_{1985}$	0.1819	0.1875	0.1743
	$k_{2019}$	0.9811	0.4943	0.5510
$c_x$	$c_{60}$	0.0012	0.0011	0.0015
	$c_{75}$	0.0005	0.0007	0.0006
	$c_{89}$	0.0024	0.0027	0.0027
$\gamma_{t-x}$	$\gamma_{1861}$	1.5197	0.7801	0.8251
	$\gamma_{1910}$	0.3313	0.2789	0.2901
	$\gamma_{1959}$	1.8623	0.8211	0.6808

## 3.7 Concluding Remarks

In this chapter, we introduce a least squares method for estimating the Renshaw-Haberman model. Our proposed approach obtains parameter estimates by minimizing the total  $L^2$  error, which measures the sum of squared errors between the observed and fitted log central mortality rates. To overcome the optimization challenge, we develop an alternating minimization scheme which sequentially updates one group of parameters at a time. We also formulate the update of the age-cohort component as a PCA problem with missing values, so that it can be accomplished effectively using an iterative SVD algorithm.

Through a number of numerical experiments, we demonstrate that our proposed method significantly outperforms the traditional Poisson MLE in terms of computation time, while producing a better goodness-of-fit in terms of  $L^2$  error and a similar goodness-of-fit in terms of log-likelihood.

Our proposed method can be applied to the H1 model, a reduced version of the Renshaw-Haberman model that is designed to improve estimation efficiency. It can also be implemented in tandem with the Hunt-Villegas method, which reduces computation time through an extra parameter constraint. Our numerical experiments indicate that computation time can be reduced further if our proposed method is used with the H1 model and/or the Hunt-Villegas method.

In Section 3.6.4, we demonstrate that our proposed estimation method can be implemented with a residual bootstrap to generate measures of parameter uncertainty in mortality forecasting. Without the improvement in efficiency brought by our proposed estimation method, the residual bootstrap would have taken a few days to complete. Similar benefits are also seen in the application to solvency capital requirement calculations under Solvency II. When re-calibration risk is taken into consideration, the solvency capital requirement of a liability can be obtained with the following algorithm:

1. simulate  $M_1$  realizations of mortality in the following year;
2. for each of the  $M_1$  realizations obtained from the previous step,

- (a) expand the original data set to include the realization of mortality in the following year;
  - (b) re-estimate the mortality model with the updated dataset;
  - (c) using the re-estimated model, simulate  $M_2$  sample paths of mortality (for year 2 and beyond);
  - (d) calculate the expected value of the liability at the end of year 1 using the  $M_2$  sample paths;
3. obtain an empirical distribution of liabilities at the end of year 1;
  4. calculate the solvency risk capital (Value-at-Risk at 99.5% confidence level) from the empirical distribution.

The outer loop of the algorithm above entails  $M_1$  model re-estimation. Typically,  $M_1$  needs to be large enough (say 5000) to that the tail risk measure can be prudently estimated. Assuming  $M_1 = 5000$  and the US male dataset is considered, it would take 13.2 days when RH-MLE is used, but only 14.5 hours when our proposed RH-LS is used.

Further, the computational efficiency provided by our estimation method enables researchers to consider extensions of the Lee-Carter model with multiple cohort effects, such as one that generalizes the Renshaw-Haberman model with the following specification:

$$\log(m_{x,t}) = a_x + \sum_{i=1}^P b_x^{(i)} k_t^{(i)} + \sum_{j=1}^Q c_x^{(j)} \gamma_{t-x}^{(j)} + \varepsilon_{x,t}, \quad (3.38)$$

so that  $P$  period effects and  $Q$  cohort effects are captured. When estimated with the traditional ML method, this generalization would be subject to even more severe convergence problem compared to the Renshaw-Haberman model, as it comes with a larger parameter space. However, the computational efficiency of our proposed method is unaffected by the generalized model structure, as estimates of the additional model parameters can be obtained easily by replacing the first-order SVD in Steps 3 and 4 of Algorithm 2 with a  $P$ -order SVD and  $Q$ -order SVD, respectively.

The model specified in equation (3.38) may be used to identify long-term (ultimate) scale factors in two-dimensional mortality improvement scales, a mortality projection

method that has been promulgated by major actuarial professional organizations in recent years. Defined as mortality improvement rates that are not subject to any transient period and cohort effects, such scale factors may be derived from the model specified in equation (3.38), with  $P$  and  $Q$  that are chosen in such a way that transient period and cohort effects are fully filtered by the model structure. The identification of  $P$  and  $Q$  in equation (3.38) is in principle similar to that for a GARCH( $P, Q$ ) process in the context of time-series analysis. The implementation of this generalized model with  $Q = 1$  is presented in Chapter 4, and the full version is left for future development.

## Chapter 4

# RHals: An R Package for Efficient Alternating Least Squares Estimation of the Renshaw-Haberman Model and Its Extensions

### 4.1 Introduction

Accurate modelling and forecasting of mortality rates are critical components of actuarial science and demography. Among the foundational models in mortality forecasting is the Lee-Carter model (Lee and Carter, 1992), which utilizes singular value decomposition (SVD) to capture primary mortality patterns via age-specific parameters and a time-varying mortality index. The model's simplicity and robustness have established it as a benchmark, prompting the development of numerous extensions and variants aimed at enhancing model fit and capturing additional influences on mortality trends.

A significant advancement in mortality modelling is the incorporation of cohort effects, which account for variations in mortality attributable to the experiences of specific birth cohorts. Cohort effects reflect generational influences, such as shifts in lifestyle, medical

advancements, economic conditions, and environmental exposures, that distinctly impact mortality rates for different cohorts. The importance of cohort effects has long been recognized by demographers and actuaries (Hobcraft et al., 1985; Wilmoth, 1990; Willets, 2004), and empirical evidence suggests that models incorporating cohort effects provide better fits compared to those that do not (Cairns et al., 2009).

The Renshaw-Haberman (RH) model (Renshaw and Haberman, 2006) extends the Lee-Carter framework by introducing an age-cohort interaction term, effectively integrating cohort effects into the model. Traditional estimation methods for the RH model are based on maximum likelihood estimation (MLE), typically assuming a Poisson or binomial distribution for death counts and maximizing a complex likelihood function using iterative algorithms such as Newton-Raphson.

Despite the theoretical appeal of the Renshaw-Haberman model, fitting these models involves considerable computational challenges. It is well documented in the literature, such as Cairns et al. (2009, 2011); Haberman and Renshaw (2009, 2011), that MLE for the RH model suffers from serious convergence issues and can be prohibitively slow. These challenges significantly limit the practical applicability of such models, particularly in contexts that require repeated estimation or real-time computations. Examples include bootstrapping for parameter uncertainty assessment, which requires multiple re-estimations to generate empirical parameter distributions (Brouhns et al., 2005; D’Amato et al., 2012; Koissi et al., 2006), and calculating solvency capital requirements under Solvency II, where repeated model re-estimations are needed for scenario-based risk assessments (Cairns, 2013; Zhou et al., 2014).

To address the computational difficulties of the RH model and its extensions, several approaches have been proposed in the literature. Renshaw and Haberman (2006) suggest simplified versions of the RH model to mitigate slow convergence. Additionally, Hunt and Villegas (2015) identify an approximate identification issue inherent to the RH model and its extensions, recommending an additional parameter constraint to stabilize the estimation process. However, both approaches remain within the conventional MLE framework and focus on reducing the parameter space, thereby improving computational efficiency at the expense of model fit.

Several R packages have been developed related to the Renshaw-Haberman model and its variants. Notable among these are:

- **demography** (Hyndman et al., 2014): Implements the original Lee-Carter model along with variants such as those proposed by Lee and Miller (2001); Booth et al. (2002); Hyndman and Ullah (2007). The package focuses on methods for smoothing and forecasting demographic data, but does not include fitting the Renshaw-Haberman model.
- **ilc** (Butt et al., 2014): Provides implementations of the Lee-Carter model and its cohort extension, the Renshaw-Haberman model, under a Poisson regression framework.
- **LifeMetrics**<sup>1</sup> (available at [www.macs.hw.ac.uk/~andrewc/lifemetrics](http://www.macs.hw.ac.uk/~andrewc/lifemetrics)): Offers implementations of the Lee-Carter model, the standard Cairns-Blake-Dowd model (Cairns et al., 2006), extended CBD models (Cairns et al., 2009), the traditional age-period-cohort model (Osmond, 1985), and the Renshaw-Haberman model, all under a Poisson regression framework.
- **StMoMo** (Villegas et al., 2018): A more modern and comprehensive package that implements a wide range of generalized age-period-cohort framework within the generalized linear/non-linear model framework. It provides tools for model fitting, forecasting, simulation, and includes functionality for assessing goodness-of-fit and incorporating parameter uncertainty through bootstrapping. It also includes the method in Hunt and Villegas (2015) to fit the Renshaw-Haberman model with the additional approximate constraint.

In Chapter 3, we have explored an alternating least squares estimation method for fitting the RH model, which directly minimizes the sum of squared differences between observed and fitted log mortality rates. This approach iteratively updates parameter groups, leveraging SVD for age-period components and principal component analysis (PCA) with

---

<sup>1</sup>Note that the **LifeMetrics** package has been discontinued since 2013.

missing values for age-cohort components. This method effectively handles the complexities introduced by cohort effects, significantly reducing computation time compared to traditional MLE methods.

Building upon our previous work in Chapter 3, this chapter introduces the R package `RHa1s`, which implements the proposed efficient least squares fitting method for the RH model and its extensions. The key contributions of the `RHa1s` package include:

- *Efficient estimation*: By employing an alternating optimization algorithm that leverages SVD and PCA with missing values, the package significantly reduces computation time and enhances numerical stability when fitting the Renshaw-Haberman model and its extensions.
- *Flexible model specification*: Our approach naturally and straightforwardly generalizes to models with multiple age-period, allowing for more flexible and nuanced modelling of mortality patterns. Traditional MLE methods face increasing computational burdens with additional terms, whereas the least squares approach efficiently handles the complexity by extracting multiple principal components during each iteration without substantial extra computation.
- *Comprehensive modelling workflow and integration with existing tools*: Beyond model fitting, the `RHa1s` package is designed for seamless integration with the `StMoMo` package for visualization, forecasting, and uncertainty estimation, supporting a full modelling cycle from data preparation to analysis and prediction.
- *Model selection tools*: The package includes methods for model evaluation and selection based on information criteria, which are commonly used in maximum likelihood estimation. These tools allow users to compare different model specifications systematically under our general model setting and select the most appropriate model for their data and objectives.
- *Multi-population extension*: The `RHa1s` package supports the implementation of the Renshaw-Haberman model and its extensions for multiple populations. With efficient fitting algorithms and model selection methods, users can analyze and compare

mortality trends across different populations, facilitating studies on shared mortality patterns and differences between groups.

The remainder of this chapter is structured as follows. In Section 4.2, we provide a detailed overview of the Renshaw-Haberman model and its extensions, including the model formulation, identifiability constraints, and common special cases. Section 4.3 describes the main functionalities of the `RHa1s` package, covering the least squares fitting method, mortality forecasting, parameter uncertainty estimation via bootstrapping, and model selection techniques. Section 4.4 explores the extension of the least squares estimation method to multi-population models, discussing how the package accommodates shared parameters and the corresponding model selection procedure. Finally, Section 4.5 concludes the chapter by giving some further remarks.

## 4.2 The Generalized Renshaw-Haberman Model

### 4.2.1 Model Formulation

We first provide an overview of the general mortality modelling framework employed in our study, including important special cases such as the Lee-Carter model and the Renshaw-Haberman model. This foundation is essential for understanding the estimation techniques and functionalities implemented in the `RHa1s` package.

We assume that the dataset covers  $p$  ages,  $x \in \{x_1, \dots, x_p\}$ , and  $n$  calendar years,  $t \in \{t_1, \dots, t_n\}$ . The corresponding cohorts (year of birth) are defined as  $t - x \in \{x_1 - t_n, \dots, x_p - t_n\}$ . Following the frameworks of Villegas et al. (2018) and Hunt and Blake (2021), the generalized stochastic mortality model we employ represents mortality rates as a sum of multiple age-period terms and a cohort effect. Specifically, the model assumes:

$$\log m_{x,t} = a_x + \sum_{i=1}^m b_x^{(i)} k_t^{(i)} + b_x^{(0)} \gamma_{t-x} + \varepsilon_{x,t}, \quad (4.1)$$

which we refer to as the **generalized Renshaw-Haberman model** throughout this chapter.

In (4.1), the components are defined as follows:

- $m_{x,t}$  is the central mortality rate at age  $x$  in year  $t$ .
- $a_x$  is the age-specific intercept, representing the average log-mortality over time.
- $b_x^{(i)}$  are the age-specific factor loadings for the  $i$ -th period effect.
- $k_t^{(i)}$  are the  $i$ -th period-specific indices that capture temporal trends in mortality.
- $m$  is a positive integer representing the number of age-period terms in the model.
- $b_x^{(0)}$  is the age-specific factor loading for the cohort effect.
- $\gamma_{t-x}$  is the cohort-specific index that captures generational effects, with  $t-x$  denoting the cohort (year of birth).
- $\varepsilon_{x,t}$  is the error term.

### Special Cases of the Generalized RH Model

Several well-established mortality models are special cases of the generalized Renshaw-Haberman model described above.

#### 1. Lee-Carter model:

The Lee-Carter model (Lee and Carter, 1992), one of the most influential models in mortality modelling, is a special case of the generalized framework with no cohort effect and a single age-period term:

$$\log m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}. \quad (4.2)$$

In addition, Renshaw and Haberman (2003) added an extra age-period term to the Lee-Carter model:

$$\log m_{x,t} = a_x + b_x^{(1)} k_t^{(1)} + b_x^{(2)} k_t^{(2)} + \varepsilon_{x,t}. \quad (4.3)$$

2. **Renshaw-Haberman model:** The Renshaw-Haberman model (Renshaw and Haberman, 2006) extends the Lee-Carter model by incorporating a cohort effect to capture generational influences. The RH model is obtained by setting  $m = 1$  in the generalized formulation:

$$\log m_{x,t} = a_x + b_x k_t + b_x^{(0)} \gamma_{t-x} + \varepsilon_{x,t}. \quad (4.4)$$

3. **H1 model:** To address the computational difficulties of the RH model, simplified variants have been proposed in the literature. A notable example is the H1 model (Renshaw and Haberman, 2006; Haberman and Renshaw, 2011), where the age-specific factor for the cohort term is set to 1, i.e.,  $b_x^{(0)} = 1$ . The H1 model formulation is:

$$\log m_{x,t} = a_x + b_x k_t + \gamma_{t-x} + \varepsilon_{x,t}. \quad (4.5)$$

## Identifiability Constraints

The generalized Renshaw-Haberman model is subject to identifiability issues, as the parameters are not uniquely determined without additional constraints. To ensure uniqueness of the parameter estimates, the following constraints are imposed:

$$\sum_{x=x_1}^{x_p} b_x^{(i)} = 1, \quad \sum_{t=t_1}^{t_n} k_t^{(i)} = 0, \quad \sum_{x=x_1}^{x_p} c_x = 1, \quad \sum_{t-x=t_1-x_p}^{t_n-x_1} \gamma_{t-x} = 0, \quad (4.6)$$

for all  $i = 1, \dots, m$ . These constraints address the potential identifiability issues arising from the redundancy between parameters and ensure a meaningful decomposition of the mortality rates.

Note that different choices of constraints on  $\gamma_{t-x}$  are possible. The constraint adopted here,  $\sum_{t-x=t_1-x_p}^{t_n-x_1} \gamma_{t-x} = 0$ , controls the mean of the cohort effect and is widely used in practice, as seen in Cairns et al. (2009) and the **StMoMo** package (Villegas et al., 2018). Alternatively, other common approaches involve setting the boundary values of the cohort effect to zero, as discussed in Renshaw and Haberman (2006) and Hunt and Villegas (2015):

- $\gamma_{t_1-x_p} = 0$  (first cohort effect set to zero), or

- $\gamma_{t_n-x_1} = 0$  (last cohort effect set to zero).

These alternative constraints also ensure identifiability without affecting the model's fit, as they only shift the location of the cohort effect.

## 4.2.2 Alternating Least Squares Estimation of the Generalized Renshaw-Haberman Model

The core functionality of the `RHa1s` package is the efficient estimation of the RH model and its extensions using an alternating least squares algorithm proposed in Chapter 3. Traditional estimation approaches, such as maximum likelihood estimation (MLE), often face challenges when applied to these models due to their complexity and the high dimensionality of the parameter space.

The proposed least squares method directly minimizes the sum of squared errors between actual and fitted log central death rates, corresponding to the following optimization problem for the generalized RH model (4.1):

$$\min_{\mathbf{a}, \mathbf{b}, \mathbf{k}, \mathbf{b}^{(0)}, \boldsymbol{\gamma}} \sum_{x,t} \left( \log m_{x,t} - \left( a_x + \sum_{i=1}^m b_x^{(i)} k_t^{(i)} + b_x^{(0)} \gamma_{t-x} \right) \right)^2, \quad (4.7)$$

where the vectors  $\mathbf{a} = (a_x)_{x=x_1}^{x_p}$ ,  $\mathbf{b}^{(i)} = (b_x^{(i)})_{i=1, \dots, m; x=x_1, \dots, x_p}$ ,  $\mathbf{k}^{(i)} = (k_t^{(i)})_{t=t_1}^{t_n}$ ,  $\mathbf{b}^{(0)} = (b_x^{(0)})_{x=x_1}^{x_p}$ , and  $\boldsymbol{\gamma} = (\gamma_{t-x})_{t-x=t_1-x_p}^{t_n-x_1}$  represent the model parameters for all ages, periods, and cohorts, respectively.

The optimization of (4.7) follows an alternating least squares (ALS) scheme, where the parameters are updated iteratively in blocks: the intercept  $\mathbf{a}$ , the age-period terms  $(\mathbf{b}, \mathbf{k})$ , and the age-cohort terms  $(\mathbf{b}^{(0)}, \boldsymbol{\gamma})$ . Below, we summarize the key steps of the algorithm within each iteration cycle, referring readers to Chapter 3 for detailed technical explanations:

1. Update  $\mathbf{a}$  (intercept term):

With the other parameters fixed, the intercept  $\mathbf{a}$  is updated by taking the average of the residual log mortality rates across periods and cohorts.

2. Update  $(\mathbf{b}, \mathbf{k})$  (age-period terms):

With  $\mathbf{a}$ ,  $\mathbf{b}^{(0)}$ , and  $\boldsymbol{\gamma}$  fixed, this step performs a SVD of order  $m$  on the residual matrix, after removing the intercept and age-cohort effects.

3. Update  $(\mathbf{b}^{(0)}, \boldsymbol{\gamma})$  (age-cohort terms):

This step involves the core technical challenge of the algorithm. As shown in Section 3.4.3, this step can be reformulated as a **PCA with missing values** problem. The missing values arise in this step since the oldest and youngest cohorts are not fully observable. To visualize, let  $z_{x,t} := \log m_{x,t} - a_x - \sum_{i=1}^m b_x^{(i)} k_t^{(i)}$  be the residuals from the Step 2, we can arrange the input values  $z_{x,t}$  in a  $p \times (n + p - 1)$  age-cohort data matrix:

$$\mathbf{Z}_{ac} := \begin{bmatrix} \times & \times & \cdots & \cdots & \times & z_{x_1,t_1} & \cdots & z_{x_1,t_{n-1}} & z_{x_1,t_n} \\ \times & \times & \cdots & \cdots & z_{x_2,t_1} & z_{x_2,t_2} & \cdots & z_{x_2,t_n} & \times \\ \vdots & \vdots & & \ddots & & & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & & & \ddots & & \vdots & \vdots \\ \times & z_{x_{p-1},t_1} & \cdots & z_{x_{p-1},t_{n-1}} & z_{x_{p-1},t_n} & \cdots & \cdots & \times & \times \\ z_{x_p,t_1} & z_{x_p,t_2} & \cdots & z_{x_p,t_n} & \times & \cdots & \cdots & \times & \times \end{bmatrix}, \quad (4.8)$$

each  $\times$  represents a missing value.

In Chapter 3, we propose to solve this sub-optimization problem using an efficient **iterative SVD algorithm**, summarized as follows:

- (a) **Impute missing values** in the input matrix (typically using row-wise means of observed values). In this context, the input matrix corresponds to the age-cohort residual matrix, which is obtained by transforming the original age-period matrix after removing the intercept and age-period effects.
- (b) **Perform SVD** on the imputed matrix, treating both observed and imputed values. The first pair of singular vectors extracted from the SVD correspond to the updated estimates of  $\mathbf{b}^{(0)}$  and  $\boldsymbol{\gamma}$ .
- (c) **Update the input matrix** using the new estimates of  $\mathbf{b}^{(0)}$  and  $\boldsymbol{\gamma}$  to construct a refined approximation for the next iteration of the SVD. This process continues until convergence.

Throughout the iterative algorithm, simple linear transformations of the parameters will be applied to make sure that the identifiability constraints (4.6) are satisfied.

## 4.3 Main Functionalities of the RHals Package

The RHals package can be installed from the author's GitHub page using:

```
R> install.packages("devtools")
R> devtools::install_github("yiping-guo/RHals")
```

Then the package can be loaded with:

```
R> library(RHals)
```

The package vignette can be accessed via:

```
R> vignette("RHals")
```

### 4.3.1 Model Fitting

The core function for fitting the generalized Renshaw-Haberman model (4.1) in the RHals package is `RHfit`. This function implements the alternating least squares algorithm discussed in Section 4.2.2, offering a computationally efficient alternative to traditional maximum likelihood estimation methods. In this subsection, we demonstrate the use of `RHfit` with real mortality data and provide examples of various model specifications.

The `RHfit` function takes a matrix of central mortality rates  $m_{x,t}$  as the main input, with rows corresponding to ages and columns to years. To facilitate the acquisition of such datasets, the package includes two utility functions:

- `load_EWData`: Loads mortality rates from England and Wales (E&W).
- `load_USData`: Loads mortality rates from the United States (US).

These functions retrieve data from the Human Mortality Database (2023), with flexible options to select the desired age ranges, years, and series (Male, Female, or Total). The full data ranges available are:

- England & Wales: Ages 0–110, Years 1841–2021.
- United States: Ages 0–110, Years 1933–2021.

To illustrate the fitting process, we use the male mortality data from England & Wales for ages 60–89 and years 1960–2009:

```
R> Data <- load_EWData(ages = 60:89, years = 1960:2009, series = "Male")
```

We first fit the standard RH model, described by (4.4), using the `RHfit` function.

```
R> fit_rh <- RHfit(Data, ages = 60:89, years = 1960:2009)
R> print(fit_rh)
```

Generalized Renshaw-Haberman Model fit

Gaussian model with predictor:  $\log m[x,t] = a[x] + b1[x] k1[t] + b0[x] g[t-x]$

Years in fit: 1960 - 2009

Ages in fit: 60 - 89

The printed output confirms the successful fitting of the model, specifying that the predictor is a Gaussian model. While the least squares method used in `RHfit` does not assume any specific distribution, the result is equivalent to a maximum likelihood estimate under a Gaussian assumption for log central death rates. We will revisit this point in Section 4.3.2 on model selection.

The `RHfit` function supports the fitting of generalized RH models with multiple age-period terms, controlled by the parameter `m`. Below, we fit a generalized RH model with  $m = 2$ :

```
R> fit_rh_2 <- RHfit(Data, ages = 60:89, years = 1960:2009, m = 2)
R> print(fit_rh_2)
```

Generalized Renshaw-Haberman Model fit

Gaussian model with predictor:  $\log m[x,t] = a[x] + b1[x] k1[t] +$   
 $b2[x] k2[t] + b0[x] g[t-x]$

Years in fit: 1960 - 2009

Ages in fit: 60 - 89

The parameter estimates from the fitted model are accessible via the output object, which is of class `fitRHals`. Notably, the values of  $b_x^{(i)}$  and  $k_t^{(i)}$  are stored in matrix form to handle models where  $m > 1$ . Below is an example of extracting the fitted values of  $b_x^{(2)}$  from `fit_rh_2` (rounded to eight decimal places):

```
R> fit_rh_2$bx[,2]
      60      61      62      63      64      65
-0.09726380 -0.10186579 -0.11426266 -0.14185281 -0.10971223 -0.10576589
      66      67      68      69      70      71
-0.08348668 -0.10866769 -0.09367565 -0.08458012 -0.07495153 -0.05978307
      72      73      74      75      76      77
-0.05890409 -0.05593078 -0.04932212 -0.02271192 -0.02319578  0.00752348
      78      79      80      81      82      83
 0.02511822  0.04679366  0.10872328  0.13880873  0.14158264  0.20106983
      84      85      86      87      88      89
 0.17891234  0.24133833  0.26096681  0.28182035  0.39633893  0.35693603
```

The `RHals` package integrates seamlessly with the widely-used `StMoMo` package, enabling users to leverage the `plot S3` method directly on fitted objects. This makes it easy to generate high-quality visualizations of the parameter estimates. Below, we demonstrate the visualization for both the fitted standard RH model `fit_rh` and the generalized RH model with two age-period terms `fit_rh_2`, shown in Figures 4.1 and 4.2.

```
R> plot(fit_rh, nCol = 3)
R> plot(fit_rh_2, nCol = 3)
```

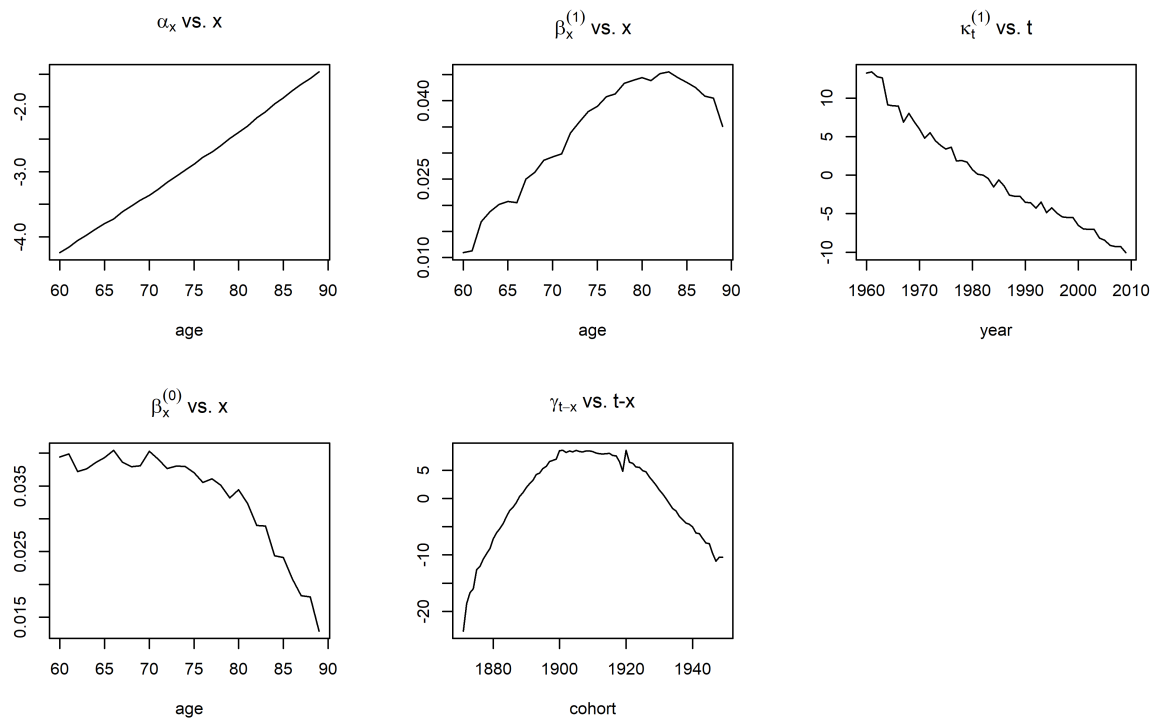


Figure 4.1: Parameter estimates of the RH model fitted to the E&W male dataset.

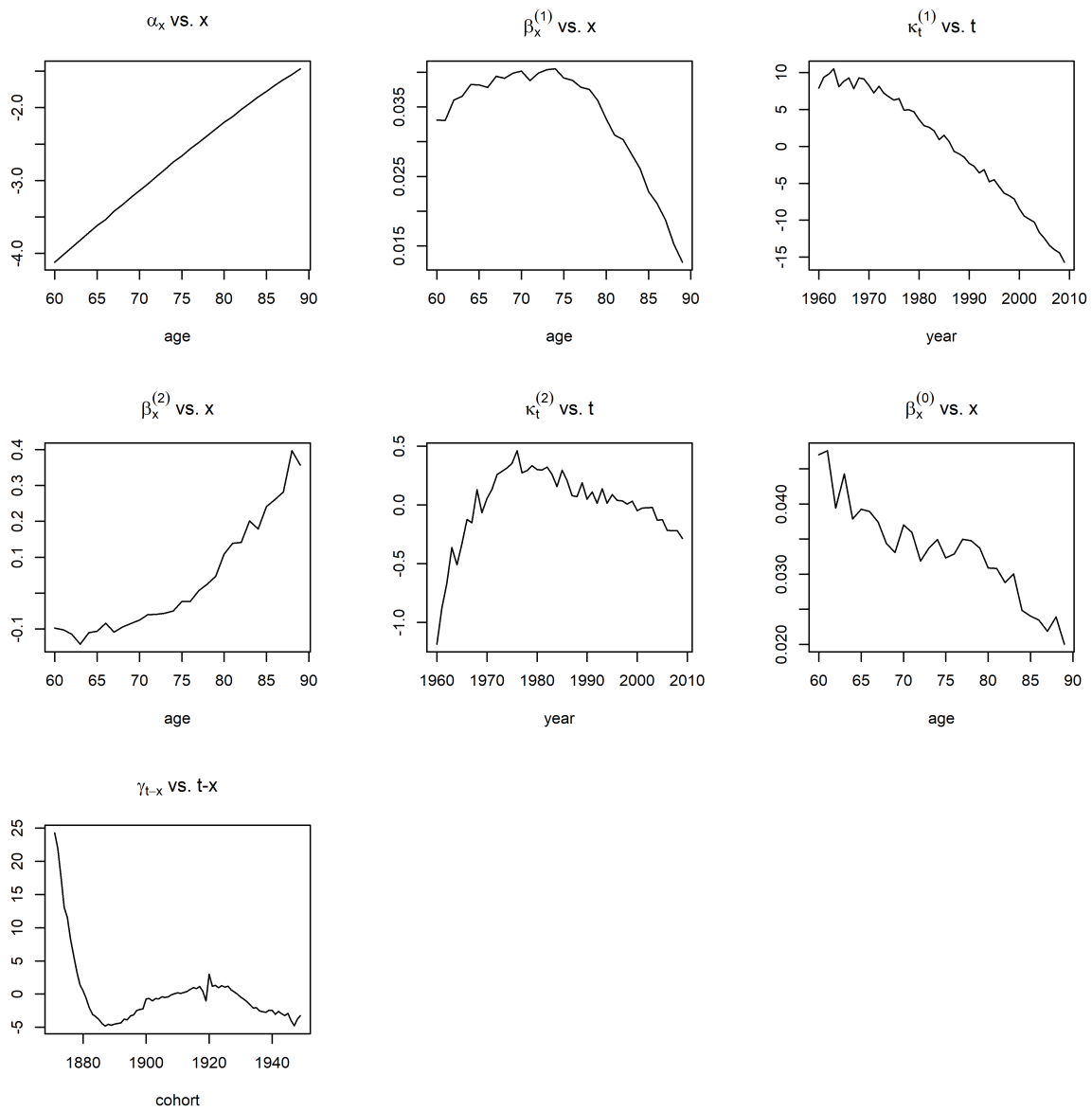


Figure 4.2: Parameter estimates of the generalized RH model ( $m = 2$ ) fitted to the E&W male dataset.

The `RHfit` also supports fitting a wide range of saturated models of the general structure (4.1), described in Section 4.2.1, under the least squares principle. First, the Lee-Carter models (with  $m \geq 1$  age-period terms),

$$\log m_{x,t} = a_x + \sum_{i=1}^m b_x^{(i)} k_t^{(i)} + \varepsilon_{x,t}, \quad (4.9)$$

which correspond to special cases of the generalized RH model without cohort effects, can be fitted by setting the `lc` argument to `TRUE` (set as `FALSE` by default if not specified). The Lee-Carter model is fitted using the SVD, and the following code fits a standard Lee-Carter model and its generalized version with  $m = 2$ :

```
R> fit_lc <- RHfit(Data, ages = 60:89, years = 1960:2009, lc = TRUE)
R> fit_lc_2 <- RHfit(Data, ages = 60:89, years = 1960:2009, lc = TRUE,
  m = 2)
```

Furthermore, to improve computational efficiency, the RH model can be simplified by setting  $b_x^{(0)} = 1$  for all  $x$ :

$$\log m_{x,t} = a_x + \sum_{i=1}^m b_x^{(i)} k_t^{(i)} + \gamma_{t-x} + \varepsilon_{x,t}. \quad (4.10)$$

This specification corresponds to the H1 model when  $m = 1$ . As shown in Chapter 3, the H1 model allows the cohort effect  $\gamma_{t-x}$  to be updated using explicit formulas, eliminating the need for iterative SVD updates. This significantly reduces the computational burden.

Fitting the H1 model or its generalization with multiple age-period terms is straightforward by setting `const.b0x = TRUE`. The codes below fit the H1 model with one and two age-period terms, respectively:

```
R> fit_h1 <- RHfit(Data, ages = 60:89, years = 1960:2009, const.b0x = TRUE)
R> fit_h1_2 <- RHfit(Data, m = 2, ages = 60:89, years = 1960:2009,
  const.b0x = TRUE)
```

Hunt and Villegas (2015) propose an additional identifiability constraint on the cohort indices:

$$\sum_{s=t_1-x_p}^{t_n-x_1} (s - \bar{s})\gamma_s = 0, \quad (4.11)$$

to mitigate an approximate identifiability issue inherent in the RH model and H1 model. In Section 3.5.2, we developed a Lagrange multiplier method to incorporate this constraint into the H1 model under the least squares scheme. This method can be implemented in `RHfit` by setting the `approxConst = TRUE`. Here is how to fit the H1 models ( $m = 1, 2$ ) with the additional constraint (4.11):

```
R> fit_h1_appro <- RHfit(Data, ages = 60:89, years = 1960:2009,
  const.b0x = TRUE, approxConst = TRUE)
R> fit_h1_2_appro <- RHfit(Data, ages = 60:89, years = 1960:2009, m = 2,
  const.b0x = TRUE, approxConst = TRUE)
```

It is important to note that the `approxConst` option can only be used for the H1 model, that is, when setting `const.b0x = TRUE`. Attempting to use the constraint without this condition will result in a warning message. At present, the incorporation of this constraint into the generalized RH model is not available and remains for future research and package development.

### 4.3.2 Model Selection

A key step in evaluating the quality of a mortality model is the analysis of its residuals, which reflect the difference between observed and fitted values. Residuals can highlight any patterns or systematic deviations not captured by the model, indicating areas for potential model improvement. A heat-map provides a powerful visual tool to detect such patterns. In this context, we define the residual for the central mortality rates  $m_{x,t}$  as:

$$d_{x,t} = \log m_{x,t} - \log \hat{m}_{x,t}, \quad (4.12)$$

where  $\hat{m}_{x,t}$  is the fitted mortality rate at age  $x$  and year  $t$ . Blue colors in the heat-map represent positive residuals (overestimated mortality rates), while red colors represent negative residuals (underestimated rates). A well-fitted model should ideally exhibit residuals that are randomly distributed, with no discernible patterns across ages or time periods.

The `RHals` package includes a `heatmap` function, an `S3` method for objects of the class `fitRHals`, to generate heat-maps of residuals. This function utilizes the `plot()` method from the `StMoMo` package to produce a color map of the residuals. Below is the code to generate heat-maps for the standard Renshaw-Haberman model: Below is the code to generate heat-maps for the standard Renshaw-Haberman model:

```
R> heatmap(fit_rh)
```

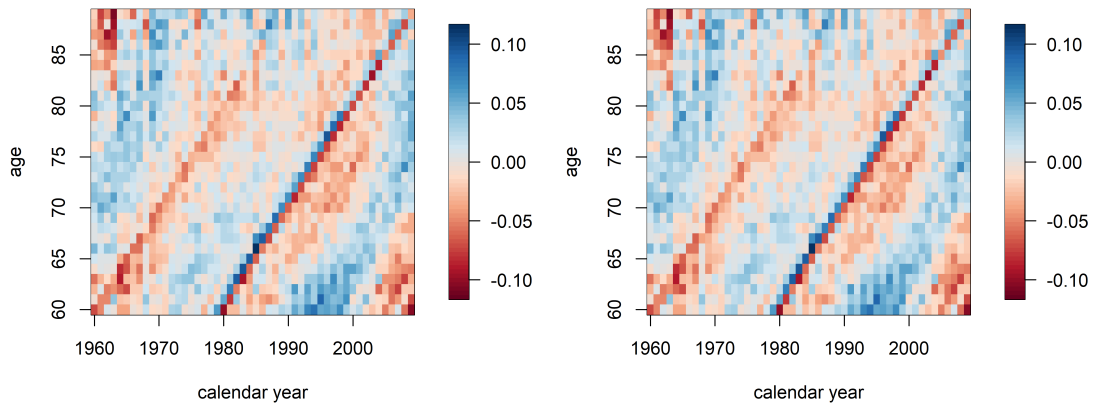
The heat-maps of the standard and generalized Lee-Carter models and the standard and generalized Renshaw-Haberman models, are presented in Figures 4.3(a)-4.3(d), respectively. These plots offer important insights into the strengths and limitations of each model. Below, we summarize the key observations from the heat-maps.

- For both the generalized Lee-Carter model (`fit_lc_2`) and the generalized RH model (`fit_rh_2`), the residuals appear more random than those from their simpler counterparts (`fit_lc` and `fit_rh`). It implies that including an additional age-period term  $b_x^{(2)}k_t^{(2)}$  improves the model's ability to capture complex mortality patterns.

A clear example is seen in the heat-map for the standard RH model, where a red vertical line at year 1969 suggests underestimation of mortality during that year. This pattern disappears in the heat-map for `fit_rh_2`, indicating that the second age-period term successfully captures this temporal effect.

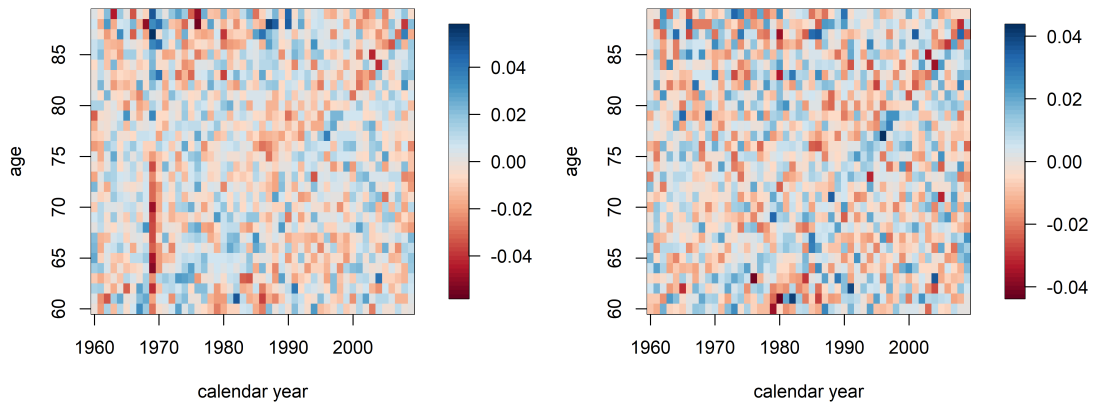
- In the heat-maps for both Lee-Carter models (`fit_lc` and `fit_lc_2`), strong diagonal patterns are visible, suggesting that these models fail to account for cohort-specific mortality trends.

Once the age-cohort term  $b_x^{(0)}\gamma_{t-x}$  is incorporated in the RH models, these diagonal patterns disappear, confirming that the cohort effects have been successfully captured.



(a) Standard Lee-Carter Model

(b) Generalized Lee-Carter Model ( $m = 2$ )



(c) Standard RH Model

(d) Generalized RH Model ( $m = 2$ )

Figure 4.3: Heat-maps of residuals for different mortality models fitted to the E&W male dataset.

Selecting an appropriate model is a critical step in mortality modelling, especially when working with a package like `RHalS` that can fit a variety of models with different levels of complexity. A key challenge in model selection is balancing goodness-of-fit with model simplicity. While more complex models, with a greater number of parameters, tend to fit the data better, they also run the risk of overfitting, capturing noise rather than meaningful patterns. Overfitted models may perform well on historical data but poorly on future forecasts or out-of-sample predictions, reducing their practical value in actuarial or demographic applications.

To address this, information criteria, such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are widely used in the mortality literature (Cairns et al., 2009; Haberman and Renshaw, 2011). These criteria modify the log-likelihood function by imposing a penalty for model complexity, helping to prevent overfitting and ensuring a model generalizes well to new data. Both criteria favor models with lower values, meaning better fits with fewer parameters.

The AIC and BIC are defined as follows:

$$AIC = 2\nu - 2\ell, \quad BIC = \log N \cdot \nu - 2\ell, \quad (4.13)$$

where  $\nu$  is the effective number of parameters,  $\ell$  is the fitted log-likelihood, and  $N$  is the total number of observations, which equals  $n \times p$  in our notations. We make some remarks to (4.13):

- The log-likelihood  $\ell$  measures the fit of the model to the observed data. Even though the least squares framework does not explicitly assume a distribution, the parameter estimates are equivalent to those obtained from Gaussian MLE. Therefore, we can compute the Gaussian log-likelihood as:

$$\ell = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{x,t} (\log m_{x,t} - \log \hat{m}_{x,t})^2, \quad (4.14)$$

where  $\sigma^2$  is estimated by its MLE  $\hat{\sigma}^2 = \sum_{x,t} (\log m_{x,t} - \log \hat{m}_{x,t})^2 / N$ .

- The effective number of parameters  $\nu$  is the total number of parameters being estimated minus the number of constraints imposed to ensure identifiability. It is worth

noting that for the H1 model,  $\nu$  is further reduced by 1 if the additional identifiability constraint (4.11) is imposed. Table 4.1 summarizes the effective number of parameters for each model considered, where  $p$  denotes the number of ages,  $n$  the number of years and  $m$  the number of age-period terms included in the model.

Table 4.1: Summary of parameters, constraints and the effective number of parameters  $\nu$  for the generalized LC, H1, and RH models.

	$a_x$	$b_x^{(i)}$	$k_t^{(i)}$	$b_x^{(0)}$	$\gamma_{t-x}$	Constraints	Eff. parameters $\nu$
LC	$p$	$mp$	$mn$	0	0	$2m$	$p + m(p + n - 2)$
H1	$p$	$mp$	$mn$	0	$n + p - 1$	$2m + 1$	$2p + n - 2 + m(p + n - 2)$
RH	$p$	$mp$	$mn$	$p$	$n + p - 1$	$2m + 2$	$3p + n - 3 + m(p + n - 2)$

The `RHfit` function returns all relevant statistics required for model selection. The effective number of parameters  $\nu$ , the log-likelihood  $\ell$ , and the corresponding AIC and BIC values can be accessed directly from the fitted `fitRHals` object. Below is an example using the standard RH model:

```
> fit_rh$npars      # Effective number of parameters (nu)
[1] 215
> fit_rh$loglik     # Gaussian log-likelihood (ell)
[1] 4210.116
> fit_rh$AIC        # AIC value
[1] -7990.231
> fit_rh$BIC        # BIC value
[1] -6847.889
```

We summarize the relevant statistics for all the six models fitted in our example in Table 4.2, and we observe the following:

- The generalized RH model with  $m = 2$  achieves the lowest AIC, suggesting that it captures complex mortality patterns most effectively. In contrast, the generalized RH

model also performs well under the BIC criterion, though the heavier penalty imposed by BIC results in a slight preference for simpler models, such as the generalized H1 model with  $m = 2$ .

- Additionally, the two Lee-Carter models (with  $m = 1, 2$ ) show much higher AIC and BIC values compared to the models incorporating cohort effects, reinforcing the importance of accounting for cohort effects when modelling this dataset. These findings are consistent with the visual evidence from the residual heat-maps (Figure 4.3), where the cohort effects are clearly visible and only adequately captured by the RH models.

Table 4.2: Summary of log-likelihood  $\ell$ , effective number of the parameters  $\nu$ , AIC and BIC for the six fitted single-population models.

	Log-likelihood $\ell$	Eff. parameters $\nu$	AIC	BIC
LC ( $m = 1$ )	3146.48	108	-6076.96	-5503.132
LC ( $m = 2$ )	3409.85	186	-6447.70	-5459.44
H1 ( $m = 1$ )	4033.50	186	-7695.00	-6706.75
H1 ( $m = 2$ )	4395.81	264	-8263.62	<b>-6860.93</b>
RH ( $m = 1$ )	4210.12	215	-7990.23	-6847.89
RH ( $m = 2$ )	4466.90	293	<b>-8347.81</b>	-6791.03

### 4.3.3 Uncertainty Estimation

An important application of stochastic mortality models is the quantification of uncertainty involved in mortality projections. A key source of such uncertainty is parameter risk, which arises because parameters used to project future mortality rates are estimates, not exact values. Accurately quantifying this uncertainty is critical for making informed decisions in risk management and actuarial planning, as it provides insight into the potential variability of mortality predictions.

To address parameter uncertainty, we employ a residual bootstrap method following Chapter 3. Residual bootstrapping, initially proposed by Lee and Carter (1992) and further developed by Koissi et al. (2006) and Li (2014), allows us to generate empirical distributions for model parameters, thereby providing quantification of the parameter uncertainties such as standard errors. The residual bootstrap algorithm proceeds as follows:

1. First, estimate the generalized RH model on the original dataset, then calculate residuals by  $d_{x,t} = \log m_{x,t} - \log \hat{m}_{x,t}$ , as defined in (4.12), for each age  $x$  and year  $t$  in the calibration window.
2. Draw random samples from these residuals with replacement to generate a pseudo dataset by adding the resampled residuals to the fitted log mortality rates  $\log \hat{m}_{x,t}$ .
3. Refit the model to the pseudo dataset to obtain a new set of parameter estimates.
4. Repeat Steps 2 and 3 a large number of times (e.g., 1000 times). These obtained empirical distributions can then be used to calculate measures of parameter uncertainty.

To facilitate this bootstrap process, the `RHals` package provides a generic `S3` method, `bootstrap`, which operates on objects of class `fitRHals`. The argument `nBoot` defines the number of bootstrap samples. For example, the following code performs 1000 bootstrap iterations on the standard RH model (`fit_rh`):

```
R> boot_rh <- bootstrap(fit_rh, nBoot = 1000)
```

The output is a list of fitted model objects based on each resampled dataset. This fitted object can be further analyzed in multiple ways. First, we can calculate the standard errors of each model parameter across the bootstrap samples using the `seBoot` generic `S3` method. For instance, the standard errors of the parameter  $b_{60}^{(1)}$  (of age 60, the youngest age within our calibration window) can be extracted as follows:

```
R> seBoot(boot_rh)$bx[1]
[1] 4.721251e-05
```

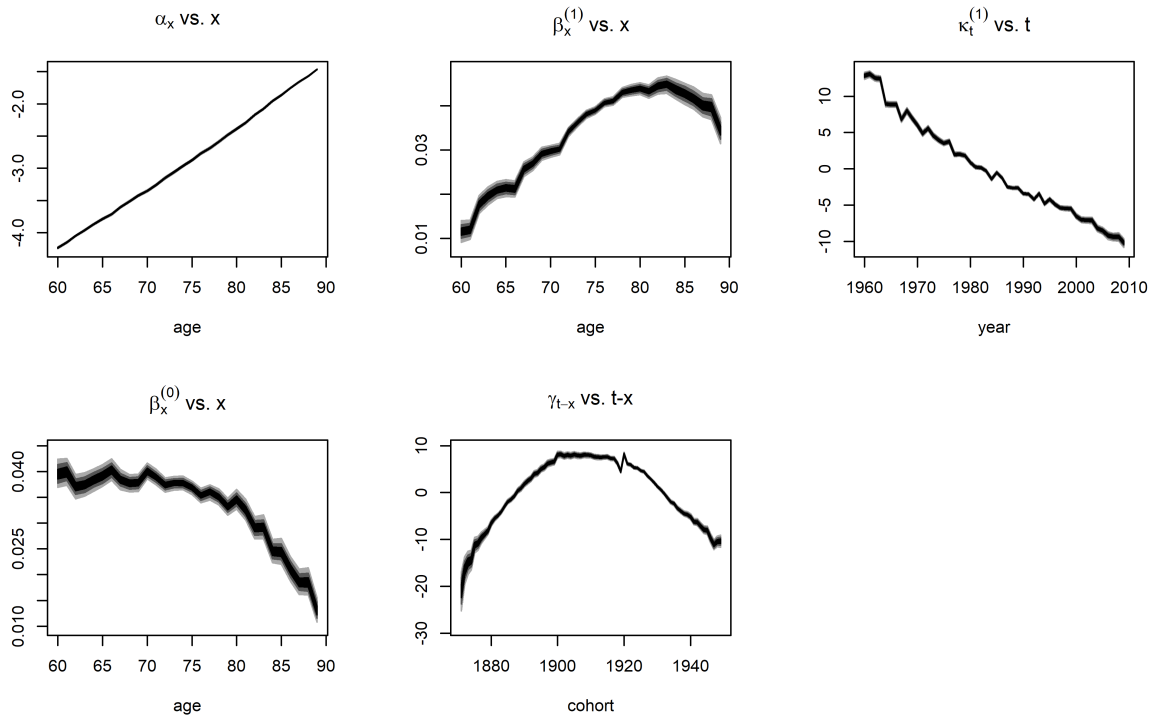


Figure 4.4: Bootstrapped parameters of the standard RH model fitted to the E&W dataset. The shaded areas represent the 50%, 80% and 95% prediction intervals, respectively.

Additionally, the `plot` function, drawing from the `StMoMo` package, allows for visualizing the parameter estimates through fan charts. This approach can show the 50%, 80%, and 95% confidence intervals for each parameter, providing an intuitive sense of the uncertainty range. The following code generates a fan chart for the parameters in the standard RH model. From the fan chart shown in Figure 4.4, we observe that while the uncertainty in  $a_x$ ,  $k_t^{(1)}$  and  $\gamma_{t-x}$  is relatively modest, the uncertainty associated with  $b_x^{(1)}$  and  $c_x^{(1)}$  is more substantial.

```
plot(boot_rh, nCol = 3)
```

The feasibility of implementing residual bootstrapping in the `RHals` package is significantly enhanced by the computational speed of the ALS method used in model fitting.

Since the bootstrapping algorithm requires a large number of repeated estimations, typically several hundred to a thousand or more, the overall runtime depends directly on the time required for each estimation cycle. For example, in our analysis of the E&W male dataset with  $M = 1000$  bootstrap iterations, the ALS approach in `RHaIs` completes the process in around 2.5 hours, whereas the classical maximum likelihood estimation method (implemented in the `StMoMo` package) requires approximately one full day. This efficiency significantly enhances the feasibility of bootstrapping and similar applications that require repeated model estimations in practical settings.

### 4.3.4 Forecasting

The `RHaIs` package is designed for efficient model estimation, with forecasting capabilities enhanced through its seamless integration with the `StMoMo` package. This integration allows users to extend their workflow beyond estimation to include the forecasting stage, an essential part of mortality modelling. In this subsection, we demonstrate basic forecasting functionalities; for a comprehensive guide on the `StMoMo` package, we refer readers to Villegas et al. (2018).

In the generalized Renshaw-Haberman model (4.1), both the period effects  $k_t^{(i)}$  and the cohort effects  $\gamma_{t-x}$  are treated as time series that drive the dynamics of mortality rates. Forecasting these components is an essential step for projecting future mortality trends.

- Period effects  $k_t^{(i)}$ : Commonly modeled using a multivariate drifted random walk, as is standard in actuarial literature (Cairns et al., 2009; Haberman and Renshaw, 2011). Once a suitable time series model is fitted, linear extrapolation is used to forecast future values of the period effects.
- Cohort effect  $\gamma_{t-x}$ : More challenging for modelling since it often lacks a clear linear trend. A common approach, suggested by Renshaw and Haberman (2006) and others, is to model the cohort effect as an ARIMA( $p, q, d$ ) process with drift. Typical choices in the literature include ARIMA(1, 1, 0) (equivalent to an ARMA(1, 1) model) and ARIMA(2, 0, 0) (equivalent to an AR(2) model).

Forecasting with the `RHals` package is straightforward using the `forecast S3` method. This function takes a `fitRHals` object returned by `RHfit`, and allows the user to specify the forecasting horizon and the ARIMA order for the cohort effect.

Below is an example of forecasting the mortality rates for 10 years ahead (2010–2019) based on the fitted RH model `fit_rh`. We model the period effect using a drifted random walk and the cohort effect with an ARIMA(1,1,0) process:

```
R> for_rh <- forecast(fit_rh, h = 10, gc.order = c(1, 1, 0))
```

In this code, the argument `h` specifies the number of years to forecast, and `gc.order` defines the ARIMA( $p, q, d$ ) parameters for the cohort effect.

After fitting the time series models, the projected central mortality rates  $m_{x,t}$  can be easily accessed. For instance, the projected mortality rates for age 60 over the years 2010–2019 (rounded to eight decimal places) are:

```
R> for_rh$rates[1,]
      2010      2011      2012      2013      2014      2015
0.00852204 0.00853587 0.00855592 0.00857833 0.00860167 0.00862540
      2016      2017      2018      2019
0.00864932 0.00867335 0.00869746 0.00872165
```

The `plot S3` method can also be used to visualize the projected values of both the period and cohort effects. The following commands generate these visualizations, as shown in Figure 4.5:

```
# Fan chart for project kt
R> plot(for_rh, only.kt = TRUE)
# Fan chart for project gc
R> plot(for_rh, only.gc = TRUE)
```

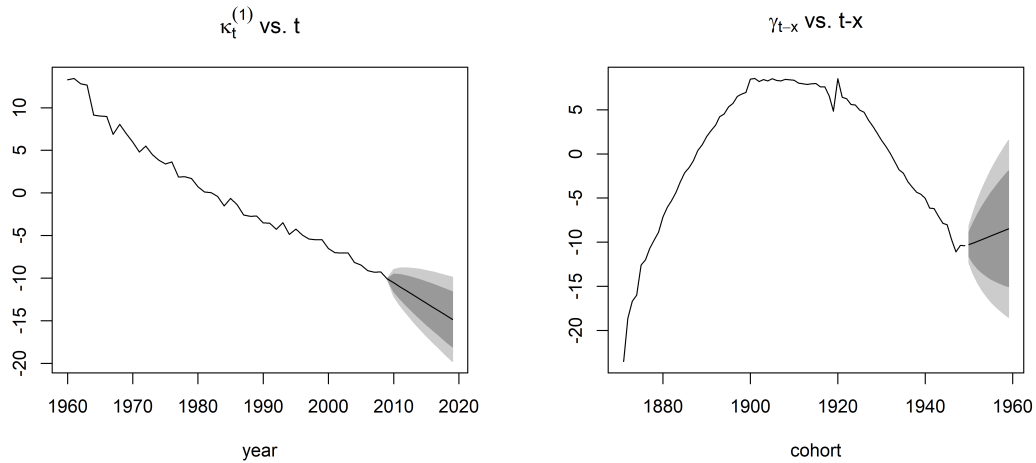


Figure 4.5: Forecast of the period effects  $b_x^{(1)}$  and cohort effects  $\gamma_{t-x}$  of the standard RH model fitted to the E&W dataset. The light and dark shaded areas represent the 80% and 95% prediction intervals.

## 4.4 Multi-Population Extensions

### 4.4.1 Model Formulation and Fitting

The generalized RH model was originally developed to capture mortality dynamics for a single population. However, in practice, mortality rates across different populations - such as neighboring countries, or subpopulations within a country - often exhibit interdependencies due to shared socioeconomic, environmental, and healthcare influences. This dependence has spurred significant research interest in multi-population mortality models, with many extensions based on the foundational Lee-Carter model, of which the generalized RH model is a flexible variant.

One key class of multi-population Lee-Carter extensions is the augmented common factor (ACF) Model proposed by Li and Lee (2005). This model incorporates an additional factor that is shared across populations, leading to “coherent” mortality forecasts where the mortality rates for different populations converge over time. This coherence property is particularly useful in actuarial and demographic forecasting, as it prevents the unrealistic

scenario of population-specific mortality rates diverging indefinitely.

Another influential approach is the common age effect (CAE) model, introduced by Kleinow (2015). The CAE model assumes that populations share a common age-effect term, which captures the general mortality behavior across ages, while allowing each population to retain unique period effects that account for population-specific trends and shocks. This model achieves a balance between capturing shared characteristics and preserving individual population dynamics. Further details on the scope and variations of multi-population mortality models can be found in Villegas et al. (2017) and Enchev et al. (2017).

In this section, we extend the CAE framework to the generalized RH model, illustrating how it can flexibly accommodate different configurations of shared and population-specific age effects across multiple populations. We also demonstrate how this GCAE model can be implemented efficiently using the `RHals` package.

In this section, we extend the CAE framework to the generalized RH model, illustrating how it can flexibly accommodate different configurations of shared and population-specific age effects across multiple populations. We also demonstrate how this GCAE model can be implemented efficiently using the `RHals` package.

Let  $J$  denotes the number of populations, indexed by  $j = 1, \dots, J$ , and let  $m_{x,t,j}$  represent the central death rate at age  $x$  in year  $t$  for population  $j$ . The CAE model for population  $j$  is given by:

$$\log(m_{x,t,j}) = a_{x,j} + b_x k_{t,j} + \varepsilon_{x,t}, \quad (4.15)$$

where the age-effect  $b_x$  is shared across all populations, and each population retains its unique age-specific intercept  $a_{x,j}$  and period effect  $k_{t,j}$ .

This formulation can be extended naturally to the generalized RH model, which we refer as the Generalized Common Age-Effect (GCAE) model. The generalized RH model includes two sets of age effects:  $\{b_x^{(i)}\}_{i=1}^m$  for the period effects  $\{k_t^{(i)}\}_{i=1}^m$ , and  $(b_x^{(0)})$  for the cohort effect  $\gamma_{t-x}$ . With these, we propose three variants of the GCAE model, denoted as *GCAE 1-3*, to capture different configurations of shared and population-specific age effects:

1. *GCAE 1*: Shared  $\{b_x^{(i)}\}_{i=1}^m$  and individual  $b_x^{(0)}$ :

$$\log(m_{x,t,j}) = a_{x,j} + \sum_{i=1}^m b_x^{(i)} k_{t,j}^{(i)} + b_{x,j}^{(0)} \gamma_{t-x,j} + \varepsilon_{x,t}. \quad (4.16)$$

2. *GCAE 2*: Individual  $\{b_x^{(i)}\}_{i=1}^m$  and shared  $b_x^{(0)}$ :

$$\log(m_{x,t,j}) = a_{x,j} + \sum_{i=1}^m b_{x,j}^{(i)} k_{t,j}^{(i)} + b_x^{(0)} \gamma_{t-x,j} + \varepsilon_{x,t}. \quad (4.17)$$

3. *GCAE 3*: Shared both  $\{b_x^{(i)}\}_{i=1}^m$  and  $b_x^{(0)}$ :

$$\log(m_{x,t,j}) = a_{x,j} + \sum_{i=1}^m b_x^{(i)} k_{t,j}^{(i)} + b_x^{(0)} \gamma_{t-x,j} + \varepsilon_{x,t}. \quad (4.18)$$

Fitting the GCAE models is straightforward and follows a similar approach to the CAE models. In the CAE framework, Kleinow (2015) showed that estimation could be formulated as a common principal component analysis (CPCA), which can be solved using specific numerical techniques, as discussed in Clarkson (1988). However, it turns out that iteratively applying SVD to the residual matrices could significantly simplify the estimation process.

Below, we summarize the iterative algorithm for fitting GCAE models. In each iteration cycle, parameters are updated in the following order:

1. Update  $\mathbf{a}$  (intercept terms):

With the other parameters fixed, for each of sub-population  $j = 1, \dots, J$ , the intercept  $a_{x,j}$  is updated by taking the average of the residual log mortality rates across periods and cohorts.

2. Update  $(\mathbf{b}, \mathbf{k})$  (age-period terms):

With the other parameters fixed:

- If the age effects  $b_{x,j}$  is **not shared**:

For each of sub-population  $j = 1, \dots, J$ , perform a SVD of order  $m$  on the residual matrix  $\mathbf{Y}_j$ , after removing the intercept and age-cohort effects. This SVD step updates individual estimates for  $b_{x,j}$  and  $k_{t,j}$ .

- If the age effects  $b_x$  is **shared**:

Construct an augmented residual matrix  $\tilde{\mathbf{Y}} = (\mathbf{Y}_1, \dots, \mathbf{Y}_J)$  by combining the individual residual matrices by column, and then perform a SVD of order  $m$ . The resulting  $b_x$  represents the shared age effect, while the “augmented”  $\mathbf{k}$  provides a column-stacked update for each  $\mathbf{k}_j$  stacked by column.

### 3. Update $(\mathbf{b}^{(0)}, \boldsymbol{\gamma})$ (age-cohort terms):

With the other parameters fixed:

- If the age effects  $b_{x,j}^{(0)}$  is **not shared**:

For each of sub-population  $j = 1, \dots, J$ , perform a SVD of order 1 on the residual matrix  $\mathbf{Z}_j$ , after removing the intercept and age-period effects.

- If the age effects  $b_x^{(0)}$  is **shared**:

Construct an augmented residual matrix  $\tilde{\mathbf{Z}} = (\mathbf{Z}_1, \dots, \mathbf{Z}_J)$  by combining the individual residual matrices by column, and then perform a SVD of order 1.

## 4.4.2 Implementation by RHals

Next, we illustrate how to fit the generalized Common Age-Effect (GCAE) model using the `RHals` package, specifically the `RHMultifit` function. This function is built similarly to the single-population `RHfit` function but accommodates the shared and individual parameter structures across populations, allowing flexible modelling within the GCAE framework.

The `RHMultifit` function includes two additional arguments, controlling the configurations of the shared age effects in the GCAE model:

- `common.bx`: Specifies whether the age-period effects  $\{b_x^{(i)}\}_{i=1}^m$  are shared across populations.

- `common.b0x`: Specifies whether the age-cohort effect  $b_x^{(0)}$  is shared across populations. This argument is only valid for non-parametric  $b_x^{(0)}$ , that is, when `lc = FALSE` and `const.b0x = FALSE`.

Also, we make two comments on `RHMultifit` function:

- In this multi-population setting, the data input for `RHMultifit` is a list of mortality matrices, each representing a distinct population, rather than a single mortality matrix in `RHfit`. Each matrix in this list should have identical dimensions, with rows and columns corresponding to the same ages and years across all populations.
- Currently, the additional identifiability constraint proposed in (4.11) is not yet incorporated within the `RHMultifit` function. Adding this constraint for the GCAE model is left for future research and package expansion.

To demonstrate the functionality of the `RHMultifit` function, we use mortality data for males and females in England and Wales, spanning ages 60–89 and years 1960–2009.

```
R> Data1 <- load_EWData(ages = 60:89, years = 1960:2009, series = "Male")
R> Data2 <- load_EWData(ages = 60:89, years = 1960:2009, series = "Female")
R> MData <- list(Data1, Data2)
```

We first fit the standard RH model (with  $m = 1$ ) for each population individually, which can be achieved implemented using the `RHMultifit` function with ease:

```
# Individual Renshaw-Haberman fitting
R> multifit_RH <- RHMultifit(MData, ages = 60:89, years = 1960:2009)
```

Next, the GCAE model is fitted using three different configurations (4.16) - (4.18) to control which parameters are shared across populations:

```
# GCAE 1 (shared bx only)
R> multifit_GCAE1 <- RHMultifit(MData, ages = 60:89, years = 1960:2009,
```

```

                                common.bx = TRUE)
# GCAE 2 (shared b0x only)
R> multifit_GCAE2 <- RHMultifit(MData, ages = 60:89, years = 1960:2009,
                                common.b0x = TRUE)
# GCAE 3 (shared both bx and b0x)
R> multifit_GCAE3 <- RHMultifit(MData, ages = 60:89, years = 1960:2009,
                                common.bx = TRUE, common.b0x = TRUE)

```

To provide a benchmark, we also fit the LC model by setting `lc = TRUE`. Then, to fit the CAE model, we set `lc = TRUE` and `common.bx = TRUE`, enabling shared age effects across populations:

```

# Individual Lee-Carter fitting
R> multifit_LC <- RHMultifit(MData, ages = 60:89, years = 1960:2009,
                              lc = TRUE)
# CAE model
R> multifit_CAE <- RHMultifit(MData, ages = 60:89, years = 1960:2009,
                              lc = TRUE, common.bx = TRUE)

```

The primary output from the `RHMultifit` function is a list named `fits`, containing `fitRHals` objects for each population. This structure mirrors the single-population `RHfit` output, where each `fitRHals` object provides the model details and fitted model parameters. For instance, to access the fitted RH model for the female population in our example, we can use:

```
R> multifit_RH$fits[[2]]
```

This extracted object can then be used for visualization, uncertainty estimation, and forecasting, just as we did with the single-population model in Section 4.3.

Given the flexibility of the GCAE model, model selection plays a critical role in identifying the optimal configuration. Different combinations of shared age effects offer varying levels of complexity and thus require careful consideration. For this, we use the AIC and

the BIC, as defined in (4.13), to select a model that balances goodness-of-fit and model complexity. Both criteria are available directly from the `RHMultifit` output, and the codes below illustrate how to access the AIC and BIC for the fitted GCAE 3 model (shared both  $b_x^{(1)}$  and  $b_x^{(0)}$ ):

```
R> multifit_GCAE3$AIC
[1] -15867.87
R> multifit_GCAE3$BIC
[1] -13633.5
```

It is worth noting that calculating AIC and BIC in this multi-population context involves adjustments beyond simply summing up the corresponding values for individual populations. While the total log-likelihood  $\ell$  is indeed the sum of log-likelihoods for each population, the total number of effective parameter  $\nu$  must take any shared age effects into account.

Table 4.3: Summary of log-likelihood  $\ell$ , effective number of the parameters  $\nu$ , AIC and BIC for the six fitted multi-population models.

	Log-likelihood $\ell$	Eff. parameters $\nu$	AIC	BIC
LC	6023.87	216	-11615.75	-10318.37
CAE	5611.17	187	-10848.34	-9725.15
RH	8351.97	430	-15843.94	-13261.20
GCAE 1	8339.59	401	<b>-15877.17</b>	-13468.62
GCAE 2	8333.37	401	-15864.75	-13456.19
GCAE 3	8305.93	372	-15867.87	<b>-13633.50</b>

The relevant statistics for each of the six multi-population models we have fitted are summarized in Table 4.3. We see that the GCAE1 model achieves the lowest AIC and the GCAE3 model achieves the lowest BIC. Notably, both AIC and BIC values for the LC and CAE models are significantly higher than those for the RH and GCAE models, suggesting the importance of including cohort effects when modelling the E&W datasets. This result

is consistent with our findings in the single-population setting in Section 4.3. Among models that incorporate cohort effects, the GCAE models consistently outperform the RH model, providing empirical support for the value of common age effects in multi-population mortality modelling.

## 4.5 Concluding Remarks

In this chapter, we introduced the `RHals` R package, designed for efficiently fitting an extended class of Renshaw-Haberman models with multiple age-period components. To address the computational challenges and convergence issues associated with classical maximum likelihood estimation, `RHals` employs an efficient alternating least squares method, as proposed in Chapter 3, implementing the model within a least squares framework. The package supports the full modelling cycle, including visualization, forecasting, uncertainty estimation, and model selection. Furthermore, we extended the alternating least squares algorithm to a multi-population setting, particularly for generalized common age-effect models, which can also be implemented through `RHals`.

While the alternating least squares approach implemented here is less dependent on strict distributional assumptions, unlike MLE-based methods, the parameter estimates it produces align with MLE under the assumption that log central death rates are normally distributed with constant variance over the age range and calibration window. This assumption, however, may not hold in some cases, particularly at older ages (Brouhns et al., 2002). A potential solution could be to adopt a weighted least squares framework to address heteroscedasticity. However, this approach requires solving a weighted PCA with missing values, a problem known to be NP-hard (Gillis and Glineur, 2011). Future work will explore the methodology and package development for this extension.

# Chapter 5

## Kriging Methods for Modelling Spatial Basis Risk in Weather Index Insurances: A Technical Note

### 5.1 Introduction

Weather insurances are often used by farmers and agricultural firms to protect against themselves losses or damages incurred because of adverse, measurable weather conditions. Traditional weather insurance contracts are generally indemnity-based, meaning that their payoffs are based on the insured parties' actual losses. From the insurer's perspective, indemnity-based weather insurances entail relatively high administration costs and are subject to moral hazard (Quiggin et al., 1994; Erhardt and Smith, 2014). These problems must be factored into insurance prices, thereby affecting the affordability of weather insurances to the agricultural sector as a risk management tool. To mitigate the drawback of indemnity-based weather insurances, insurers may choose to offer weather index insurances, the payoffs of which are linked to certain weather indexes that are calculated on the basis of certain common weather variables such as temperature and precipitation. As weather indexes are objective, insurers do not need to validate reported losses from weather index insurances, thereby saving administration costs. The objectivity of weather

indexes also reduces the risk of moral hazard.

As with other index-linked insurances, weather index insurances entail basis risk, the risk that the actual loss incurred by the insured is different than the payout from the policy (Dick et al., 2011). In more detail, such basis risk is composed of two main components. The first component is structural basis risk, which arises from the imperfect relationship between the payoff function and the insured parties' actual losses. In practice, the payoff function is often an indicator function, so that an insured party receives a fixed amount of payoff if the weather index (e.g., maximum temperature or aggregate precipitation) to which the policy is linked exceeds a certain trigger level. However, the true underlying relationship between actual losses and the underlying weather index is much more complicated and unknown. Therefore, instead of an exact indemnification, weather index insurances are generally used for mitigating the uncertainty surrounding a target outcome (e.g., crop yields for farmers), and its effectiveness of such risk mitigation has been empirically evaluated in some literature based on data from different countries including Canada (Turvey, 2001), China (Sun et al., 2014) and the United States (Zhou et al., 2018).

The second component is *spatial basis risk*, the risk that is investigated in this chapter. Spatial basis risk exists because the coverage of weather stations is never perfect (Norton et al., 2012). If the weather stations at which measurements of weather indexes are taken are too far from the location of the insured party's risk exposure, then the mismatch between payoff and actual loss is inevitably deepened. Spatial basis risk is of particular concern in the context of microinsurance, which is commonly seen in developing countries with a low density of weather stations (Hazell et al., 2010). When the target location (location of the insured party's risk exposure) is considerably distanced from weather stations, it becomes necessary to estimate the relevant weather variables at the target location from the nearby observations. This important procedure is known as a *spatial interpolation*.

The most commonly adopted family of spatial interpolation methods is *kriging*, which enables the user to statistically incorporate information from multiple locations into the prediction for the target location. Compared to the fields of geostatistics and spatial statistics, kriging techniques have been much less extensively studied in actuarial science and insurance, particularly in the context of weather risk management, even though they

lend themselves very well to the modelling of spatial basis risk in weather index insurances. Notable previous studies of kriging techniques in actuarial science and insurance domain include the work of Norton et al. (2012) who adopt an empirical approach to study and quantify spatial basis risk that is inherent in weather index insurances using US data, the contribution of Roznik et al. (2019) who compare different universal kriging and generalized additive models for interpolating daily temperature data in the context of agricultural microinsurance, and the paper by Boyd et al. (2019) who further study the impact of kriging daily temperature on spatial basis risk reduction by analysing the correlation between estimated payoffs and reported forage yields.

The literature reviewed in the previous paragraph has only studied temperature variables and their related indexes such as consecutive cooling days. However, apart from temperature, precipitation is also regarded as a crucially important weather variable by the agricultural sector; for example, Murphy (1970) shows that forage yields are heavily impacted by cumulative precipitations within certain time periods. Kriging techniques that perform satisfactorily for temperature data do not necessarily yield the same level of performance for precipitation data. This is because compared to distributions of temperatures, distributions of precipitations are typically heavily skewed and have a significant probability mass at zero. To fill this research gap, in this chapter, we perform a deeper investigation of kriging techniques in the context of weather index insurance, with a focus on precipitations and their related indexes.

We consider daily precipitations, as well as two precipitation indexes that are derived from daily precipitations. The two precipitation indexes we consider are (i) maximum precipitation per month in five consecutive days (*MFP*) and (ii) annual maximum consecutive dry days (*CDD*). Given how they are defined, *MFP* and *CDD* capture changes in the left and right tails of the underlying precipitation distributions. They are therefore well suited as bases of weather index insurances that aim to mitigate the risk associated with extreme weather events which may result in huge losses, for example, significant reductions in agricultural yield (Turvey, 2001). It is noteworthy that *MFP* and *CDD* are component indexes of the Actuaries Climate Index (*ACI*), co-developed by a number of professional actuarial organizations to help inform actuaries, public policymakers, and the general public about climate trends and some of the potential impacts of a changing climate.

The first objective of this chapter is to compare the performance of a range of spatial interpolation methods in spatial interpolations of precipitations and precipitation indexes. We begin with simple methods including nearest neighbor and inverse distance weighting; then we consider kriging methods including standard ordinary kriging, universal kriging, and trans-Gaussian kriging. The performance of the candidate methods for different types of data, including raw daily precipitations and the two mentioned precipitation indexes, is gauged by cross-validated (CV) interpolation errors. It is found that the optimal spatial interpolation methods for raw daily precipitations and the precipitation indexes are different, owing to the differences in their distributional properties.

Our second objective is to investigate how spatial interpolations of precipitation indexes  $MFP$  and  $CDD$  are best implemented in practice. To fix ideas, let us consider a farm owner who wishes to mitigate the uncertainty surrounding the yield of his farm by purchasing a weather index insurance that is linked to the  $CDD$  applicable to the location at which his farm is located. However, the farm is distanced considerably from weather stations at which precipitation measurements can be taken. In this situation, there is a need to estimate the  $CDD$  values at the farm's location (target location) with spatial interpolations. Generally speaking, there are two ways to implement such spatial interpolations. One way is to take a **direct approach** in which  $CDD$  values at the target location are estimated by spatially interpolating  $CDD$  values recorded at nearby weather stations directly. The direct approach can be implemented easily without tracking the raw precipitation observations, and is not computationally demanding as spatial interpolations are performed only on a yearly basis (the frequency at which  $CDD$  is reported). Another way is to take a more sophisticated **two-stage approach**, in which we first, on a daily basis, spatially interpolate raw precipitations recorded at nearby weather stations to obtain an estimate of the raw precipitation at the target location every day, and then compute the  $CDD$  values at the target location on the basis of the daily precipitation estimates obtained in the first stage. Intuitively speaking, the two-stage approach appears to incorporate more information into the resulting estimates. We perform numerical analysis to compare these two approaches. Interestingly, it is found that although the two-stage approach entails a heavier data requirement (as raw daily temperatures measured at all nearby stations are needed) and computationally demanding (as spatial interpolations have

to be performed substantially more frequently), it does not produce any better prediction accuracy compared to direct interpolations of  $CDD/MFP$  on a yearly/monthly basis. We provide a statistical argument to explain this intriguing finding, which may help insurers offering weather index insurances with their risk quantification and management processes.

The remainder of the chapter is organized as follows. Section 5.2 describes the data set used in the chapter. Section 5.3 presents the five spatial interpolation methods we consider. Section 5.4 documents two numerical analyses. The first analysis evaluates the performance of the five spatial interpolation methods in the application to daily precipitation data. Through this analysis, differences between temperatures and precipitations in the context of spatial interpolation are highlighted. The second analysis compares the direct and two-stage approaches for spatially interpolating precipitation indexes  $MFP$  and  $CDD$ . Finally, some concluding remarks are made in Section 5.5.

## 5.2 Data Description and Visualization

The data used in this chapter originated from the US National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center, and is accessed through R package `STRBOOK`. Among various variables in the data set, we consider daily precipitation  $P$  in millimeters (mm) and maximum temperature  $T$  in °F at 138 weather stations in the central USA, recorded between 1990 and 1993, as well as the latitude  $Lat$  and longitude  $Lon$  of each of the 138 weather stations.

Figures 5.1 and 5.2 show the locations of the weather stations and the daily precipitations and maximum temperatures measured on some selected days. It can be seen that certain areas of the region have no weather station. For these areas, one may use spatial interpolation to estimate precipitations. In both figures, we observe a clustering phenomenon that daily precipitations and temperatures at nearby observations tend to be similar. From Figure 5.1 we observe that daily precipitations vary significantly across the region, with some close-to-zero values and some extremely large values. This observation suggests that daily precipitations exhibit a strong non-normality, a statistical property that is well taken care of in the modelling work presented in the next sections. On the contrary, Figure 5.2

shows that daily temperatures are distributed much more regularly with few outliers.

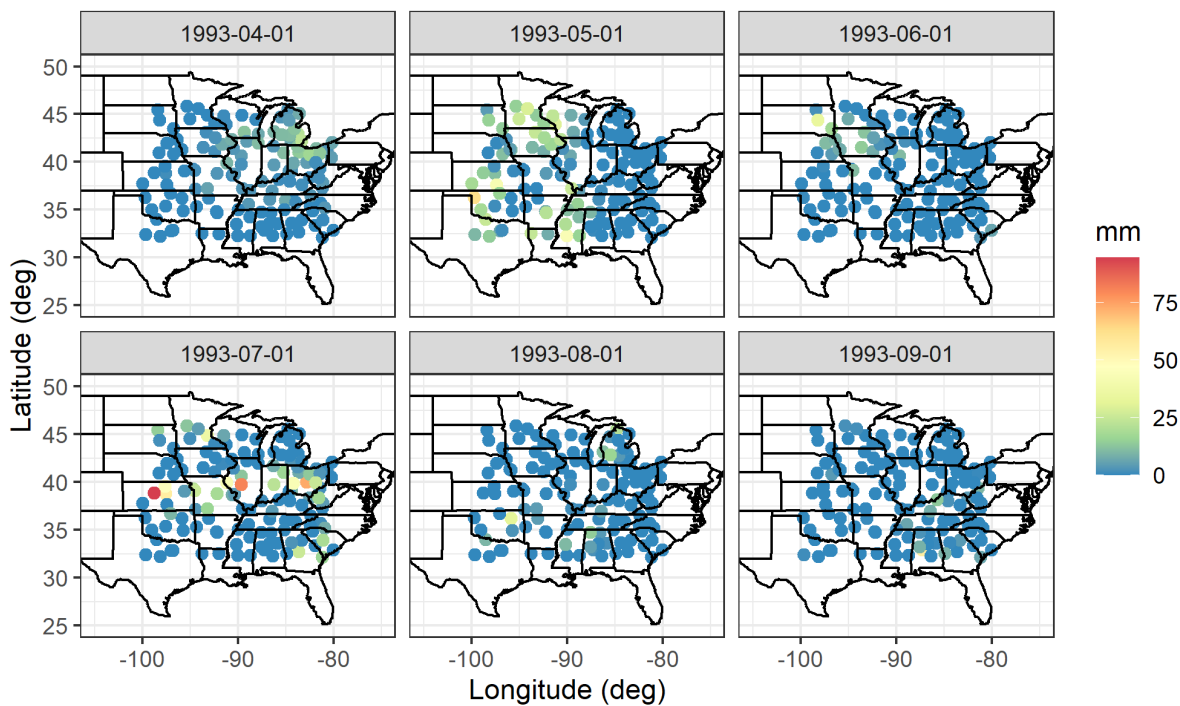


Figure 5.1: Daily precipitations  $P$  (mm) from the NOAA data set on selected days in 1993

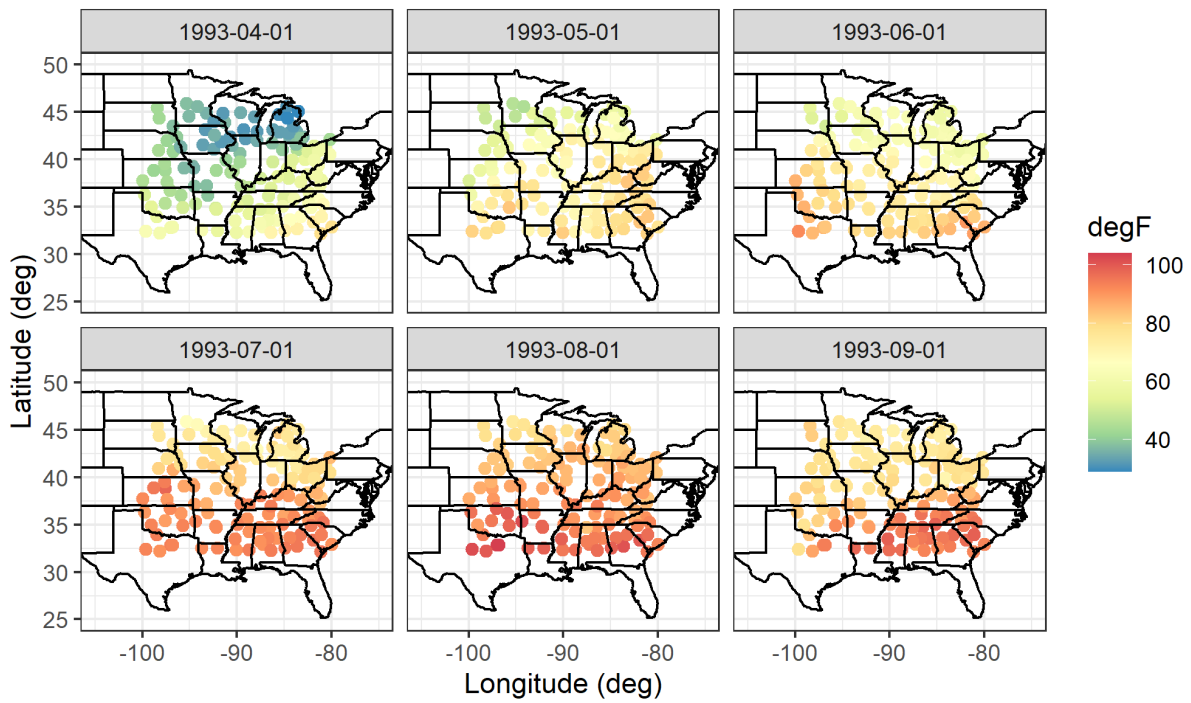


Figure 5.2: Daily maximum temperatures  $T$  ( $^{\circ}\text{F}$ ) from the NOAA data set on selected days in 1993

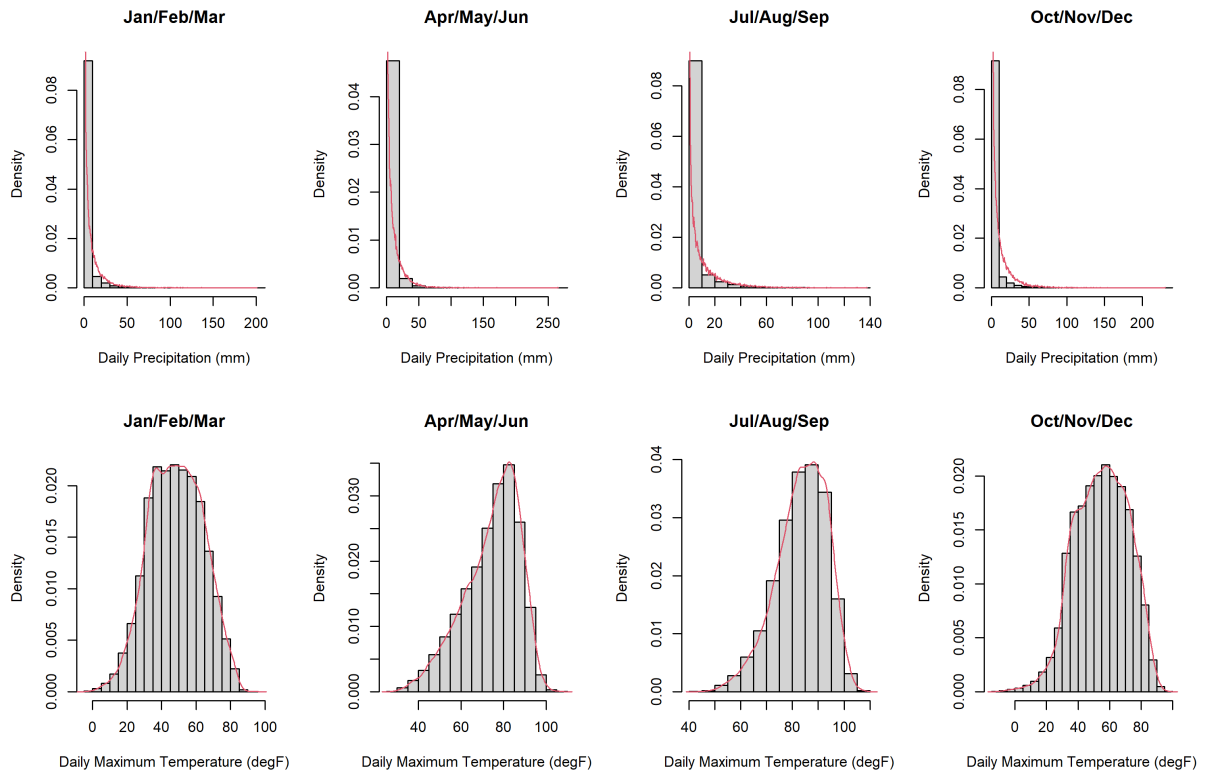


Figure 5.3: Histograms of daily precipitations  $P$  (mm) and daily maximum temperatures  $T$  ( $^{\circ}\text{F}$ ) in different seasons

Table 5.1: Variables defined in Section 5.2

Notation	Name	Data type
<i>Lat</i>	Latitude of the weather station	Continuous
<i>Lon</i>	Longitude of the weather station	Continuous
<i>Elev</i>	Elevation of the weather station	Continuous
<i>P</i>	Daily precipitation (mm)	Continuous
<i>T</i>	Daily maximum temperature (°F)	Continuous
<i>MFP</i>	Maximum precipitation per month in five consecutive days	Continuous
<i>CDD</i>	Annual maximum consecutive dry days	Count

To further compare the two variables, in Figure 5.3 we plot seasonal histograms with estimated densities for daily precipitations and daily maximum temperatures. In line with the observations made in Figures 5.1 and 5.2, we observe from the histograms that daily precipitation data are extremely right-skewed and non-normal, whereas daily maximum temperatures are distributed fairly symmetrically in bell shapes. As elaborated in the next section, normality plays a crucial role in spatial interpolation, and therefore spatial interpolation techniques applicable to a certain weather variable may not be applicable to other weather variables without appropriate adaptations.

In addition to the four variables from the NOAA dataset, we consider the vertical elevation *Elev* in meters of each weather station, as this variable is often taken into account as a covariate for interpolating precipitations in climatology (Phillips et al., 1992; Martínez-Cob, 1996) and actuarial science (Boyd et al., 2019; Roznik et al., 2019). The elevation point data is obtained from the Elevation Point Query Service (EPQS) and the WGS84 coordinate system, through R package `elevatr`.

From the daily precipitation data, we calculate the historical values of two precipitation indexes on which weather insurances may be written. The first precipitation index we consider is the maximum precipitation per month in five consecutive days (*MFP*). The second precipitation index is the annual maximum consecutive dry days (*CDD*). To calculate *CDD*, we follow the definition adopted by Actuaries Climate Index, which regards

one day as a “dry day” when the daily precipitation  $P$  is below 1mm.

For the reader’s convenience, we summarize all of the variables defined earlier in Table 5.1:  $Lat$ ,  $Lon$  and  $Elev$  are fixed covariates,  $P$  and  $T$  are daily observations, and  $MFP$  and  $CDD$  are precipitation indexes.

## 5.3 Methodology

In this section, we present the spatial interpolation techniques considered in this chapter, including basic benchmark algorithms (nearest neighbor and inverse distance weighting), fundamental kriging methods (ordinary and universal kriging), and a more advanced kriging method known as trans-Gaussian kriging. We highlight the differences among these methods and discuss the appropriateness of these methods in the context of spatially interpolating precipitation-related quantities.

Let the response variable and its estimate be  $z(\mathbf{s})$  and  $\hat{z}(\mathbf{s})$  in general, where  $\mathbf{s} = (Lat, Lon)$  denotes the coordinate vector of the target location, with  $Lat$  and  $Lon$  representing the latitude and longitude, respectively. In our numerical analysis,  $z$  can be either a basic weather variable like precipitation  $P$  or a precipitation index like  $CDD$ .

### 5.3.1 Nearest Neighbor

Nearest neighbor (NN) is the simplest spatial interpolation method, which directly takes the observation from the nearest weather station for the target location:

$$\hat{z}(\mathbf{s}^*) = z(\mathbf{s}_c), \tag{5.1}$$

where  $\mathbf{s}^*$  is the vector of the coordinates of the target location where a prediction is made, and  $\mathbf{s}_c$  denotes the coordinate vector of the nearest weather station. Because of its simplicity and transparency, NN is easy to implement and understand and serves as a benchmark for evaluating the performance of more advanced spatial interpolation techniques. The drawback of this method is that it only takes the nearest point into account and ignores all information from other observations. It is therefore expected that NN produces relatively large interpolation errors.

### 5.3.2 Inverse Distance Weighting

Inverse distance weighting (IDW) generalizes NN by taking multiple nearby observations into account and calculating their weighted average as a prediction of the target location  $\mathbf{s}^*$ :

$$\hat{z}(\mathbf{s}^*) = \sum_{i=1}^n w_i \cdot z(\mathbf{s}_i), \quad (5.2)$$

where  $n$  is the number of nearby stations taken into consideration and

$$w_i = \frac{1}{\sum_{i=1}^n \frac{1}{\|\mathbf{s}_i - \mathbf{s}^*\|^p}}$$

is the weight on location  $i$  which reduces at a power rate as the distance  $\|\mathbf{s}_i - \mathbf{s}^*\|$  between location  $i$  and the target location increases.

The pre-determined parameter  $p$  controls rate at which the weight  $w_i$  decreases with the distance  $\|\mathbf{s}_i - \mathbf{s}^*\|$ . A larger  $p$  assigns more weight to closer points. We choose  $p = 2$  (square decay) as suggested by the classical geoscience literature (Li and Heap, 2008).

When applying IDW, it is a common practice to place certain arbitrary constraints on  $n$  (Boyd et al., 2019). In this chapter, we limit  $n$  to 20, which means that we consider a maximum of 20 weather stations that are the closest to the target location.

The weights in (5.2) sum to one, so that IDW estimators are unbiased if the underlying process of  $z(\mathbf{s})$  has a constant mean over all locations. In fact, NN also enjoys the unbiasedness property since it is a special case of IDW when  $n = 1$ . Therefore, NN and IDW are commonly adopted and used as benchmarks to evaluate the performance of more sophisticated spatial interpolation techniques.

### 5.3.3 Ordinary Kriging

The main limitation of NN and IDW is that they restrict the form of spatial correlations to a power function. However, in practice, spatial correlations among observations are highly complicated and data-specific. Therefore, to produce more reliable spatial interpolations, it is necessary to model dependence structures via a more general family of functions

which still retain the desirable properties of NN and IDW predictors such as linearity and unbiasedness. This necessity motivates the use of various kriging methods. This subsection focuses on ordinary kriging, which is foundational to the more sophisticated kriging methods discussed in the next two subsections.

The formulation of ordinary kriging (OK) was originally proposed by Krige (1951) and a more formal derivation of it was first provided by Davis (1952). The framework of ordinary kriging assumes that an observation  $z(\mathbf{s})$  can be decomposed into an unknown stationary mean  $\mu$  and a spatially correlated zero-mean noise  $\varepsilon(\mathbf{s})$  as follows:

$$z(\mathbf{s}) = \mu + \varepsilon(\mathbf{s}). \quad (5.3)$$

Then, for the target location  $s^*$ , OK aims to find the optimal unbiased predictor  $\hat{z}(\mathbf{s}^*)$ , which takes the form of a homogeneously linear combination of other observations:  $\mathbf{z} = (z(\mathbf{s}_1), \dots, z(\mathbf{s}_n))$ . Optimality is achieved by minimizing the mean squared prediction error subject to the unbiasedness condition:

$$\min_{\boldsymbol{\lambda}} \mathbb{E}[z(\mathbf{s}^*) - \boldsymbol{\lambda}^T \mathbf{z}]^2 \quad \text{with } \boldsymbol{\lambda}^T \mathbf{1} = 1, \quad (5.4)$$

where  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$  is the vector of kriging coefficients (kriging weights) to be determined, and  $\mathbf{1}$  denotes a vector of ones. The solution can be obtained via the Lagrange multiplier method:

$$\hat{z}(\mathbf{s}^*) = \underbrace{\hat{\mu}_{gls}}_{\text{mean estimator}} + \underbrace{\mathbf{c}(\mathbf{s}^*)^T \mathbf{C}^{-1}}_{\text{weight}} \underbrace{(\mathbf{z} - \hat{\mu}_{gls} \cdot \mathbf{1})}_{\text{detrended data}} \quad \text{with} \quad \hat{\mu}_{gls} = \frac{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{z}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}}, \quad (5.5)$$

where  $\mathbf{C} = [\text{Cov}(z(\mathbf{s}_i), z(\mathbf{s}_j))]_{i,j=1,\dots,n}$  is the covariance matrix of the observations (made at nearby locations) and  $\mathbf{c}(\mathbf{s}^*) = (\text{Cov}(z(\mathbf{s}_1), z(\mathbf{s}^*)), \dots, \text{Cov}(z(\mathbf{s}_n), z(\mathbf{s}^*)))$  is the vector of covariances between the predicted value and observations. Although the kriging predictor may be expressed in other equivalent forms such as the form of kriging equations (Cressie, 2015), (5.5) is more interpretable. In (5.5), the predictor  $\hat{z}(\mathbf{s}^*)$  is composed of two parts: (1) the trend term  $\hat{\mu}_{gls}$  which represents the (restricted) generalized least squares estimate of the global mean  $\mu$ , and (2) the mean “correction” term  $\mathbf{c}(\mathbf{s}^*)^T \mathbf{C}^{-1}(\mathbf{z} - \hat{\mu}_{gls} \cdot \mathbf{1})$  that is expressed as a weighted sum of the detrended data  $\mathbf{z} - \hat{\mu}_{gls} \cdot \mathbf{1}$ , where the weights  $\mathbf{c}(\mathbf{s}^*)^T \mathbf{C}^{-1}$  depend on the spatial correlations.

In (5.5), the covariances  $\mathbf{C}$  and  $\mathbf{c}(\mathbf{s}^*)$  are taken as inputs, so they need to be estimated from the data  $\mathbf{z}$ . The principle behind spatial covariability modelling is that a specific family of covariance functions is fitted to the sample covariances, with a common assumption that  $z(\mathbf{s})$  is second-order stationary, that is,  $z(\mathbf{s})$  has a constant mean vector and the covariance between  $z(\mathbf{s}_i)$  and  $z(\mathbf{s}_j)$  for any  $i \neq j$  depends only on the distance between  $\mathbf{s}_i$  and  $\mathbf{s}_j$ . Covariance functions  $C(\mathbf{h})$  that are frequently used include Gaussian, exponential and spherical. They are all decreasing functions of the distance  $\|\mathbf{h}\| = \|\mathbf{s}_i - \mathbf{s}_j\|$ , but have different decaying rates. Following the classical literature in geostatistics for kriging rainfall variables (Goovaerts, 2000), we choose the spherical covariance function:

$$C(\mathbf{h}) = \begin{cases} \sigma^2 \left[ 1 - \frac{3}{2} \cdot \frac{\|\mathbf{h}\|}{\alpha} + \frac{1}{2} \cdot \left( \frac{\|\mathbf{h}\|}{\alpha} \right)^3 \right], & 0 \leq \|\mathbf{h}\| \leq \alpha \\ 0, & \|\mathbf{h}\| > \alpha \end{cases}, \quad (5.6)$$

where  $\alpha$  is the practical range that allows the covariance vanishes if the distance between two observations becomes too large. The usual procedure for implementing OK is to first calculate the sample covariances of the observations, and then use the iterated generalized-least-squares (GLS) method (Cressie, 2015) to estimate the parameters, which are  $\sigma^2$  and  $\alpha$  for the spherical case. For weather variables such as daily maximum temperature which are relatively regularly distributed, the covariance estimation procedure often works well and has a fast convergence rate. Nevertheless, for weather variables such as daily precipitations which feature highly imbalanced distributions with a large point mass at zero, the iteration might converge very slowly. This problem is practically important and is investigated deeper in the empirical analysis presented in Section 5.4.

The covariance fitting procedure is fairly well-developed and can be implemented with comprehensive geo-statistical R packages such as `gstat` and `automap`. It is worth noting that most of these packages fit the so-called variograms instead of directly fitting the covariances. However, the end results of both routes are equivalent, because under the assumption of second-order stationarity, the variogram, defined as  $2\gamma(\mathbf{h}) = \text{Var}(z(\mathbf{s} + \mathbf{h}) - z(\mathbf{s}))$ , has a one-to-one correspondence  $2\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$  with the covariance function  $C(\mathbf{h})$  (Cressie and Wikle, 2011). We choose to present OK in terms of covariance functions instead of the variograms, since covariance functions are more accessible to the

finance community, and in fact, the optimization specified by (5.4) and (5.5) is highly similar to a mean-variance portfolio optimization.

### 5.3.4 Universal Kriging

The crucial underlying assumption behind OK is that the mean of  $z(\mathbf{s})$  is constant over all locations and spatial dependence is completely captured by the residual term  $\varepsilon(\mathbf{s})$ . However, this assumption can be violated if the weather variable  $z(\mathbf{s})$  is highly correlated with certain covariates; for example, daily maximum temperatures are usually strongly correlated with latitude. Such an effect should be removed before modelling the spatial correlation between the residuals (Hudson and Wackernagel, 1994). One way to achieve this is to utilize universal kriging (UK), which assumes that  $z(\mathbf{s})$  can be decomposed into a linear function of location-related covariates  $\mathbf{x}(\mathbf{s})^T \boldsymbol{\beta}$  and a spatially correlated noise  $\varepsilon(\mathbf{s})$ :

$$z(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T \boldsymbol{\beta} + \varepsilon(\mathbf{s}). \quad (5.7)$$

The UK predictor takes a similar form to the OK predictor (5.5). The only difference is that the UK predictor replaces the constant  $\hat{\mu}_{gls}$  by a linear predictor  $\mathbf{x}(\mathbf{s})^T \hat{\boldsymbol{\beta}}_{gls}$  (Cressie and Wikle, 2011):

$$\hat{z}(\mathbf{s}^*) = \underbrace{\mathbf{x}(\mathbf{s}^*)^T \hat{\boldsymbol{\beta}}_{gls}}_{\text{mean estimator}} + \underbrace{\mathbf{c}(\mathbf{s}^*)^T \mathbf{C}^{-1}}_{\text{weight}} \underbrace{(\mathbf{z} - \mathbf{X} \hat{\boldsymbol{\beta}}_{gls})}_{\text{detrended data}}, \quad (5.8)$$

where  $\hat{\boldsymbol{\beta}}_{gls} = (\mathbf{X}^T \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C}^{-1} \mathbf{z}$  is the GLS estimator of  $\boldsymbol{\beta}$  and  $\mathbf{X} = (\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n))^T$  is the matrix of covariates at all locations. In this chapter, we select the latitude, longitude, and elevation of a weather station as covariates for kriging precipitation-related quantities, that is,  $\mathbf{x} = (\text{Lat}, \text{Lon}, \text{Elev})$ .

As in ordinary kriging, the covariance matrix  $\mathbf{C}$  and vector  $\mathbf{c}(\mathbf{s}^*)$  must be estimated. However, behind the same R function, the mechanism of fitting the covariance function is different. In OK the random part of the spatial dependence applies to the original data  $\mathbf{z}$  since a stationary mean is assumed, whereas in UK the randomness comes from the residuals only. Thus, one must “detrend” the data  $\mathbf{z}$  before modelling the spatial covariances. An “optimal” detrending is not feasible, because the GLS estimate  $\hat{\boldsymbol{\beta}}_{gls}$  of

the trend involves the covariance matrix  $\mathbf{C}$ , which should be estimated after detrending. To circumvent this problem, one may obtain an initial estimate of  $\boldsymbol{\beta}$  by a weighted least square (WLS) regression in which the weights might be chosen based on different rules (Pebesma, 2004); then, one can model the covariance function based on the detrended data and recalculate  $\hat{\boldsymbol{\beta}}_{gls}$  for the final universal kriging predictor  $\hat{z}(\mathbf{s}^*)$ .

It is documented in the literature that UK sometimes underperforms OK, even when the covariates for UK are empirically highly correlated to the interested quantity (e.g., latitude for daily maximum temperatures). This outcome is mainly caused by the intrinsically complex structure of the underlying spatial correlations, which are hardly driven by a handful number of covariates in a single linear form. Unless there exists a strong linear relationship between the chosen covariates and the observations, UK may yield an inferior covariance estimation and result in inaccurate kriging predictions.

### 5.3.5 Trans-Gaussian Kriging

As shown from equation (5.3) to (5.8), the derivation of the OK and UK kriging predictors is based on the minimization of a mean squared prediction error, which does not depend on any specific distributional assumption. Despite this fact, it is important to note the relationship between kriging and Gaussian process regression.

Gaussian process regression is a non-parametric approach which aims to determine the posterior distribution of the unobserved data given the observed data. In the derivation of the posterior mean, both observed and unobserved data are assumed to be drawn from a Gaussian process characterized by an unknown mean function and kernel. The posterior mean derived from Gaussian process regression and the kriging predictor derived from the minimization of a mean squared prediction error take the same form, which is represented by a weighted average of the observed values with the weights being determined by the underlying covariance structure. For this reason, the terms kriging and Gaussian process regression are sometimes used interchangeably, even though they are developed in different manners. The relationship between kriging and Gaussian process regression suggests that a kriging model is expected to yield superior prediction performance when the data follow

closely to a Gaussian distribution. In other words, the normality of the underlying data does matter.

As shown in Figures 5.1 and 5.3, distributions of daily precipitations  $P$  are non-Gaussian with heavy right tails, unlike the distributions of daily maximum temperatures  $T$  which appear to be more Gaussian. To avoid large kriging predictive errors due to the non-normality of  $P$ , a simple strategy (Cecinati et al., 2017) is to transform daily precipitation data with the Box-Cox transformation (Box and Cox, 1964):

$$y = \begin{cases} \frac{z^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(z), & \lambda = 0 \end{cases}, \quad (5.9)$$

where  $y$  and  $z$  represent the variables after and before transformation, respectively. The choice of the power parameter  $\lambda$  relies on an empirical judgment. Following Sun et al. (2003), we set  $\lambda = 1/3$  so that  $y = 3(\sqrt[3]{z} - 1)$ . It is worth noting that the Box-Cox transformation can only be applied to non-negative data, but this is not a concern in this study as precipitations are always non-negative.

The approach of performing kriging algorithms on the transformed data  $y$  instead of the original data  $z$  is known as trans-Gaussian kriging (TGK). The implementation of TGK involves two stages. First, a standard non-transformed kriging algorithm is performed on the transformed data  $y$ . As such, a prediction  $\hat{y}(\mathbf{s}^*)$  for the target location  $\mathbf{s}^*$  is obtained under the transformed scale. Second, a prediction  $\hat{z}(\mathbf{s}^*)$  of  $z(\mathbf{s}^*)$  in its original scale is made by back-transforming the transformed prediction  $\hat{y}(\mathbf{s}^*)$ .

In the second stage, a simple inverse transformation is inappropriate due to the fact that  $\mathbb{E}[z(\mathbf{s}^*)] = \mathbb{E}[\phi(y(\mathbf{s}^*))] \neq \phi(\mathbb{E}[y(\mathbf{s}^*)])$  for a non-linear function  $\phi(\cdot)$ . In this chapter, we adopt the approximately unbiased estimator, obtained based on the delta method, recommended by Cressie (2015):

$$\hat{z}(\mathbf{s}^*) = \phi(\hat{y}(\mathbf{s}^*)) + \phi''(\hat{\mu}_Y) \cdot \left( \frac{\sigma_Y^2(\mathbf{s}^*)}{2} - m_Y \right), \quad (5.10)$$

where  $\phi(\cdot)$  is the inverse function of the chosen Box-Cox transformation,  $\phi''(\cdot)$  is the corresponding second-order derivative,  $\hat{\mu}_Y$  is the estimated mean defined in (5.5),  $\sigma_Y^2(\mathbf{s}^*)$  is

the kriging variance<sup>1</sup>, and  $m_Y$  is the estimated Lagrange multiplier. We can obtain  $\hat{\mu}_Y$ ,  $\sigma_Y^2(\mathbf{s}^*)$  and  $m_Y$  from the standard ordinary kriging implementation.

It is worth noting that trans-Gaussian kriging cannot be applied together with universal kriging. As such, in this chapter, we only consider ordinary trans-Gaussian kriging for spatially interpolating daily precipitations.

## 5.4 Numerical Analysis

In this section, we apply the spatial interpolation methods described in Section 5.3 to the NOAA data set, with the aim to answer the following two questions:

1. Which of the spatial interpolation methods is the most appropriate for daily precipitations in terms of interpolation errors?
2. When the weather index insurance under consideration is linked to *MFP/CDD*, would spatial interpolations of the raw precipitation data on a daily basis outperform those of *MFP/CDD* itself on a monthly/yearly basis?

Throughout the analysis, we measure predictive accuracy with a  $K$ -fold cross-validation (CV), an out-of-sample model validation technique that is widely used in geoscience (Hofstra et al., 2008), climatology (Moral, 2010) and actuarial science (Boyd et al., 2019; Roznik et al., 2019) for comparing the performance of different spatial interpolation methods for weather variables. We implement CV with the following procedure. First, we randomly divide the observations to which a spatial interpolation is applied (e.g., daily precipitations recorded at the 138 weather stations on 1993-04-01) into  $K$  equal-sized groups (folds). Second, For each of the  $K$  groups, we predict the precipitation at every weather station in the group, on the basis of a spatial interpolation model that is fitted to the data from the remaining  $K - 1$  groups. So, for each weather station with location  $\mathbf{s}_i$ , we have a predicted

---

<sup>1</sup>The kriging variance is the minimized mean-squared prediction error, that is,  $\mathbb{E}[z(\mathbf{s}^*) - \hat{z}(\mathbf{s}^*)]$ , where  $\hat{z}(\mathbf{s}^*)$  is the kriging predictor. The exact formula can be found in classical spatial statistics texts (e.g. Cressie, 2015).

value  $\hat{z}(\mathbf{s}_i)$  (obtained from the second step), which can be compared against its corresponding actual observed value  $z(\mathbf{s}_i)$ . Finally, the performance of the spatial interpolation is measured by the root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i))^2}{n}}, \quad (5.11)$$

and the mean absolute error (MAE),

$$\text{MAE} = \frac{\sum_{i=1}^n |\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)|}{n}, \quad (5.12)$$

where  $n$  denotes the sample size (the number of weather stations in our context).

The choice of  $K$ , the number of folds, controls the balance between bias and variance. A smaller  $K$  leads to a lower variance but a larger bias, whereas a larger  $K$  results in the opposite. Following classic texts on model validation (Breiman and Spector, 1992; Kohavi et al., 1995), we choose  $K = 10$  to compromise.

The method of CV has been criticized by researchers such as Roberts et al. (2017), who argue that CV may underestimate predictive errors if the observations are not independent. The independence condition is clearly not satisfied in any spatial interpolation, which is by definition devised to capture the dependence of observations on the spatial domain. Thankfully, theoretical support for evaluating spatial interpolation methods with CV has recently been provided by Rabinowicz and Rosset (2022), who rigorously formulate CV for dependent data and explicitly demonstrate the correctness of using CV in spatial interpolations.

### 5.4.1 Interpolating Daily Precipitations

In this subsection, we utilize the previously discussed techniques to spatially interpolate daily temperatures from the NOAA dataset. Through the analysis, we can discern whether more advanced techniques such as UK and TGK can improve precipitation interpolation accuracy over simple benchmark techniques including NN and IDW. The results of this subsection are also useful in various means of weather risk analysis, for example, the

Table 5.2: Root mean squared errors (RMSE) and mean absolute errors (MAE) in the 10-fold cross-validations for different spatial interpolation methods applied to daily precipitations  $P$ .

		RMSE ( $P$ )				
Model	Formula	1990	1991	1992	1993	4-year average
NN	N/A	5.93	6.00	5.44	5.66	5.76
IDW	N/A	4.86	4.89	4.47	4.63	4.71
OK	$P \sim 1$	4.87	4.90	4.51	4.69	4.74
UK	$P \sim Lat + Lon + Elev$	4.90	4.93	4.55	4.73	4.77
TGK	$\sqrt[3]{P} \sim 1$	4.76	4.82	4.38	4.60	4.64

		MAE ( $P$ )				
Model	Formula	1990	1991	1992	1993	4-year average
NN	N/A	2.49	2.53	2.33	2.44	2.44
IDW	N/A	2.32	2.31	2.17	2.26	2.26
OK	$P \sim 1$	2.45	2.47	2.32	2.39	2.41
UK	$P \sim Lat + Lon + Elev$	2.53	2.54	2.38	2.47	2.48
TGK	$\sqrt[3]{P} \sim 1$	2.07	2.11	1.97	2.06	2.05

creation of a high-resolution precipitation risk map that takes spatially interpolated daily precipitations as input. Further, the results in this section are relevant to our next analysis, which investigates whether spatially interpolating raw daily precipitation values may yield superior results compared to a direct interpolation of precipitation indexes such as  $CDD$  and  $MFP$  on a less frequent basis.

We apply NN, IDW, OK, UK and TGK to daily precipitations  $P$  each day and calculate the corresponding RMSE and MAE with a 10-fold cross-validation. The resulting average RMSE and MAE over each year from 1990 to 1993 and the entire four year window are presented in Table 5.2.

Let us first compare the two benchmark methods, NN and IDW. We observe that IDW

produces significantly lower average RMSE and MAE compared to NN in each year and over the whole 4-year window. This result indicates that it is important to draw information from multiple nearby weather stations, and echoes the conclusions from previous studies (Chen et al., 2010; Shope and Maharjan, 2015) that IDW generally serves as a better benchmark compared to NN.

Before analyzing the results for more advanced spatial interpolation methods, let us make a practical note. When applying kriging methods to daily precipitations, it is important to consider the fact that distributions of daily precipitations are highly imbalanced with heavy right tails and a significant point mass at zero. The non-normality may cause potential issues when fitting a kriging model, particularly during the covariance function fitting stage.

As previously mentioned, the parameters in the covariance function of the OK kriging predictor are estimated with a GLS iteration, which may converge slowly when non-normality is present. In the extreme scenario when all of the weather stations under consideration record zero precipitation on a day, then it is simply infeasible to fit any covariance model for the day and the GLS iteration will not converge. On the contrary, the solutions from NN and IDW always exist given their nonparametric formulations. To get around the possible non-convergence problem, we adopt IDW (which is demonstrated to perform better than NN) when an OK, UK or TGK fails to yield a converged estimate of the covariance function.

Next, we turn to OK and UK. Although these techniques aim to capture spatial variability more precisely, in this application they underperform the benchmark method IDW in terms of both the RMSE and MAE. This result immediately raises the question as to whether it is necessary to consider more advanced kriging methods such as OK and UK. Further, this result seems to contradict some previous claims in the literature. For example, Boyd et al. (2019) and Roznik et al. (2019) compare different spatial interpolation methods for mean daily temperatures and find that OK and UK generally produce lower RMSE compared to IDW.

This seemingly anti-intuitive result can be attributed at least in part to the non-normality of daily precipitations. As demonstrated in Section 5.2, distributions of daily

precipitations are far from Gaussian, whereas distributions of daily maximum temperatures are fairly close to normal. As normality is implicitly assumed in OK and UK, it is conceivable that they do not yield promising results when applied to daily precipitations which exhibit significant non-normality but perform satisfactorily when applied to daily maximum temperatures for which normality roughly holds.

Furthermore, we also observe that UK performs even worse than OK in this application. This result might be caused by the possibility that the underlying relationships between daily precipitations  $P$  and the included covariates ( $Lat$ ,  $Lon$ , and  $Elev$ ) are non-linear so that the linear predictor in UK incorrectly detrend the observations and consequently introduce a bias when fitting the covariance function.

Finally, we observe that TGK outperforms all of the other four methods (NN, IDW, OK, UK) consistently. Again, this result can be attributed to the non-normality of daily precipitations, a problem that is well handled by the transformation in TGK and the proper back-transformation specified in (5.10). Another interesting finding concerning TGK is that its improvement over the IDW benchmark is more significant in terms of (percentage reduction in) MAE than RMSE. This outcome is an indication that the improvement produced by TGK is mainly contributed by the better predictive quality for non-extreme points, as by definition (5.11) and (5.12) RMSE penalizes large errors more heavily compared to MAE.

To demonstrate the normalization effect of the Box-Cox transformation, we calculate the sample skewness and kurtosis of the original daily precipitations  $P$  and their corresponding transformed values  $\sqrt[3]{P}$ . The same skewness and kurtosis calculations are also conducted for the daily maximum temperatures  $T$ , to illustrate the distributional differences between daily temperatures and precipitations. The definitions of sample skewness and sample kurtosis we adopt are as follows:

$$\text{Skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}}; \quad (5.13)$$

$$\text{Kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2}. \quad (5.14)$$

Table 5.3: Average sample skewness and kurtosis for daily precipitations  $P$ , transformed daily precipitations  $\sqrt[3]{P}$ , and daily maximum temperatures  $T$ .

Metric	Variable	1990	1991	1992	1993	4-year average
Skewness	$P$	4.32	4.37	4.01	4.11	4.20
	$\sqrt[3]{P}$	2.00	1.96	1.70	1.74	1.85
	$T$	-0.06	-0.21	-0.09	0.03	-0.08
Kurtosis	$P$	28.22	28.87	24.58	25.65	26.83
	$\sqrt[3]{P}$	9.33	9.19	7.11	7.43	8.26
	$T$	2.47	2.65	2.47	2.47	2.51

In the above,  $(x_1, \dots, x_n)$  is the sample vector and  $\bar{x}$  is the corresponding sample mean. The sample skewness is indicative of the symmetry of the underlying distribution, whereas the sample kurtosis reflects the heaviness of the tails of the underlying distribution. Samples from a normal distribution should have a sample skewness that is close to 0 and a sample kurtosis that is close to 3. We compute the sample skewness and kurtosis on a daily basis, and obtain the average values for each year from 1990 to 1993. The results are tabulated in Table 5.3.

Before the transformation, daily precipitations  $P$  has a sample skewness of 4.20 and a sample kurtosis of 26.83, which respectively suggest a significantly positive skewness and heavy tails. After transformation, the distribution of  $\sqrt[3]{P}$  becomes closer to normal, thereby resulting in better kriging accuracy as depicted in Table 5.3. Arguably, the cube root transformation does not produce a perfect normality, as the resulting sample skewness (1.85) and kurtosis (8.26) are still quite different from the normality benchmark (0 for skewness and 3 for kurtosis). As a matter of empirical fact, the choice of transformation should not be based entirely on the proximity to normality after transformation. Cecinati et al. (2017) compare different Gaussian transformation methods for precipitation data, with a focus on selecting the optimal parameter of the Box-Cox transformation. Their results show that a transformation achieving the best normality does not necessarily result in the best kriging performance.

On the other hand, the distribution of daily maximum temperatures is much closer to Gaussian, with an average sample skewness and kurtosis of  $-0.08$  and  $2.51$ , respectively. This result offers an explanation as to why TGK is rarely considered when interpolating temperature-related variables. The proximity to normality also makes the non-convergence problem moot when fitting kriging models to temperature data.

Summing up, IDW serves as a reliable benchmark method as it consistently performs better than NN. In the application to daily precipitations, of which the underlying distribution deviates significantly from Gaussian, kriging methods outperform IDW only if a proper Gaussian transformation is adopted. There exist significant differences between kriging precipitations and temperatures (which are more normally distributed) in terms of both model selection and convergence issues, and therefore analysts should not transfer kriging approaches between different weather variables arbitrarily.

#### 5.4.2 Interpolating Precipitation *MFP* and *CDD*

In this subsection, we focus on spatial interpolations for two precipitation indexes, maximum precipitation per month in five consecutive days (*MFP*) and annual maximum consecutive dry days (*CDD*), to which weather index insurances may be linked. From a practical perspective, there are two ways to spatially interpolate the two indexes:

1. **Direct approach:**

In the direct approach, we perform spatial interpolations on *MFP* and *CDD* directly without considering the raw data from which the indexes are derived. This approach yields spatially interpolated values of *MFP* every month and *CDD* every year in one single step.

2. **Two-stage approach:**

In the two-stage approach, we first perform spatial interpolations on the raw daily precipitations to obtain spatially interpolated precipitations at target locations every day. Then, following the definitions of *MFP* and *CDD*, we compute the predicted values of *MFP* every month and *CDD* every year at target locations from the spatially interpolated daily precipitations obtained in the previous step.

For each approach, we perform a 10-fold CV for the predicted values of *MFP* every month and *CDD* every year. These calculations result in, for each approach, an MAE every month and RMSE every year, which can be used to compare the performance of the direct and two-stage approaches in our application. The (average) values RMSE and MAE in each year from 1990 to 1993 and over the whole 4-year period, derived from both direct and two-stage approaches with the five spatial interpolation methods under consideration, are reported in Tables 5.4 (for *MFP*) and 5.5 (for *CDD*).

Table 5.4: Root mean squared errors (RMSE) and mean absolute errors (MAE) calculated from the cross-validations of the spatial interpolations for *MFP*, implemented with the direct and two-stage approaches and different spatial interpolation methods.

		RMSE ( <i>MFP</i> )				
Approach	Model	1990	1991	1992	1993	4-year average
Two-Stage	NN	28.23	28.52	23.93	26.81	26.87
	IDW	25.58	25.78	21.60	23.12	24.02
	OK	24.56	25.29	21.21	22.37	23.36
	UK	24.11	25.04	20.93	22.21	23.07
	TGK	27.87	28.22	23.58	25.01	26.17
Direct	NN	28.23	28.52	23.93	26.81	26.87
	IDW	23.91	23.94	19.88	21.28	22.25
	OK	23.38	23.70	20.08	21.60	22.19
	UK	23.69	24.12	20.10	21.81	22.43
	TGK	22.87	23.04	19.67	21.40	21.75
		MAE ( <i>MFP</i> )				
Approach	Model	1990	1991	1992	1993	4-year average
Two-Stage	NN	19.93	19.54	17.18	19.01	18.91
	IDW	17.18	16.82	14.91	15.83	16.19
	OK	16.64	16.79	15.01	15.63	16.01
	UK	16.42	16.54	14.89	15.62	15.87
	TGK	18.58	18.58	16.33	17.21	17.68
Direct	NN	19.93	19.54	17.18	19.01	18.91
	IDW	16.68	16.32	14.64	15.27	15.73
	OK	16.37	16.43	14.84	15.59	15.81
	UK	16.87	16.83	14.86	15.89	16.11
	TGK	16.05	15.74	14.41	15.33	15.43

Table 5.5: Root mean squared errors (RMSE) and mean absolute errors (MAE) calculated from the cross-validations of the spatial interpolations for *CDD*, implemented with the direct and two-stage approaches and different spatial interpolation methods.

RMSE ( <i>CDD</i> )						
Approach	Model	1990	1991	1992	1993	4-year average
Two-Stage	NN	6.63	6.72	6.03	4.37	5.94
	IDW	8.97	10.53	8.51	6.73	8.69
	OK	9.10	12.70	8.38	7.54	9.43
	UK	9.25	10.47	7.87	7.50	8.77
	TGK	8.30	11.32	6.85	6.36	8.21
Direct	NN	6.63	6.72	6.03	4.37	5.94
	IDW	5.41	5.99	4.63	3.64	4.92
	OK	5.70	5.87	5.49	3.47	5.13
	UK	5.46	6.19	4.61	3.43	4.92
	TGK	5.58	5.88	5.61	3.48	5.14

MAE ( <i>CDD</i> )						
Approach	Model	1990	1991	1992	1993	4-year average
Two-Stage	NN	4.16	4.96	3.98	2.91	4.00
	IDW	6.21	7.84	5.32	4.30	5.92
	OK	5.89	9.22	5.43	4.84	6.35
	UK	6.44	8.08	5.33	5.02	6.22
	TGK	5.07	7.72	4.13	3.71	5.16
Direct	NN	4.16	4.96	3.98	2.91	4.00
	IDW	3.51	4.51	3.34	2.59	3.49
	OK	3.79	4.47	3.76	2.52	3.64
	UK	3.71	4.89	3.44	2.48	3.63
	TGK	3.72	4.47	3.83	2.53	3.64

From Tables 5.4 and 5.5 we observe that for both  $MFP$  and  $CDD$ , the direct approach produces smaller RMSE and MAE compared to the two-stage approach when the interpolation method used is IDW, OK, UK or TGK. The two approaches yield the same RMSE and MAE when the interpolation method used is NN, as NN makes use of the nearest observation to the target location only. The differences between the predictive errors resulting from the two approaches are particularly remarkable in the application to  $CDD$ .

In addition, Table 5.5 shows that the two-stage approach yields unreasonable results in the application to  $CDD$ : the predictive errors produced by IDW, OK, UK, and TGK are even higher than those from NN, which should not outperform the other four methods as we argued in Section 5.3.

The striking results presented in Tables 5.4 and 5.5 beg explanations. The empirical fact that the two-stage approach under-performs the direct approach even though it is more sophisticated and computationally demanding can be attributed to the definitions (and hence statistical properties) of the precipitation indexes.

For  $MFP$ , the two-stage approach first interpolates daily precipitations  $P$ . As the distribution of  $P$  is not sufficiently close to Gaussian (even after the cube-root transformation in TGK), the kriging predictions of daily precipitations are generally not very satisfactory, and as a consequence, the calculated values of  $MFP$  in the second stage might be inaccurate. In contrast, the direct approach directly interpolates  $MFP$ , the distributions of which are closer to normal. From Table 5.6 we observe that distributions of  $MFP$  have an average skewness of 1.43 and average kurtosis of 6.06, suggesting that they are closer to Gaussian compared to the distributions of daily precipitations  $P$  (with an average skewness of 4.20 and average kurtosis of 26.83; Table 5.3) and transformed daily precipitations  $\sqrt[3]{P}$  (with an average skewness of 1.85 and average kurtosis of 8.26; Table 5.3). The higher proximity to normality can be attributed to an implicit smoothing effect introduced by the definition of  $MFP$ . As  $MFP$  is calculated as the maximum precipitation per month in five consecutive days, as long as heavy precipitation days do not cluster, the distribution of  $MFP$  should feature a less heavy right tail compared to that of  $P$ .

For  $CDD$ , the index is computed as the longest run of dry days within a year through

Table 5.6: Sample skewness and kurtosis for  $MFP$ ,  $CDD$  and their transformed values.

Metric	Variable	1990	1991	1992	1993	4-year average
Skewness	$MFP$	1.53	1.69	1.24	1.28	1.43
	$\sqrt[3]{MFP}$	0.20	0.25	0.09	0.09	0.34
	$CDD$	1.59	1.73	1.83	1.29	1.61
	$\sqrt[3]{CDD}$	1.01	1.08	1.09	0.77	0.99
Kurtosis	$MFP$	6.26	7.44	5.01	5.54	6.06
	$\sqrt[3]{MFP}$	3.27	3.34	3.27	3.08	3.24
	$CDD$	5.72	6.00	7.36	4.93	6.00
	$\sqrt[3]{CDD}$	3.64	3.90	4.69	3.61	3.96

a “rolling window” approach, where a dry day is defined as a day with precipitation that is less than a strict threshold (the threshold used in this chapter is  $P \leq 1\text{mm}$ ). In the two-stage approach, the kriging errors in the first stage (where daily precipitations are spatially interpolated) can easily lead to a large number of misclassified dry days if some true precipitations are very close to the threshold. As just one single misclassification will break a run of dry days, the kriging errors in the first stage will ultimately result in highly inaccurate  $CDD$  predictions.

Next, we compare the five spatial interpolation methods when the direct approach is taken. For  $MFP$ , TGK produces the most accurate predictions, a result that suggests that the Gaussian transformation remains important when spatially interpolating this participation index. However, compared to the application to  $P$ , TGK improves prediction errors over the benchmark method IDW only marginally, an outcome that might be attributed to the empirical fact that the distribution of  $MFP$  is closer to normal compared to that of  $P$  so that the benefit of the transformation is smaller.

Interestingly, for the spatial interpolations of  $CDD$  with the direct approach, TGK does not improve prediction accuracy over both OK and UK, even though  $CDD$  features a similar degree of normality to  $MFP$  as indicated by the sample skewness and kurtosis displayed in Table 5.6. Moreover, all of the three kriging methods (OK, UK and TKG)

underperform the benchmark method IDW. These results are due possibly to a violation of the fundamental assumption of kriging that an observation can be decomposed into a spatial trend plus a spatially correlated error term. To resolve this issue, one might consider more advanced nonlinear kriging methods such as multiple indicator kriging and probability kriging (Cressie, 2015), which are beyond the scope of this chapter.

## 5.5 Concluding Remarks

In this chapter, we study a range of spatial interpolation methods for modelling spatial basis risk inherent in weather index insurances. Our empirical work is supported by weather data obtained from the NOAA in the United States.

We extend the literature in the actuarial science and insurance domain by studying spatial interpolation methods for daily precipitations and precipitation indexes, which possess rather different distributional properties compared to temperature-related quantities that are considered in previous studies. For daily precipitations, we found that TGK is the best spatial interpolation method, an outcome that can be attributed to its Box-Cox transformation which largely eliminates the non-normality in daily temperatures. For precipitation index  $MFP$ , the conclusion remains the same but the improvement of TGK over the benchmark method is only marginal, because  $MFP$  is somewhat more normally distributed compared to daily temperatures.

We also compare two approaches that may be taken to spatially interpolate precipitation indexes including  $MFP$  and  $CDD$  in practice: a direct approach in which  $MFP/CDD$  are interpolated directly on a monthly/annual basis, and a two-stage approach in which the  $MFP/CDD$  values at the target locations are computed from the estimated daily temperatures at the target location. To our knowledge, this study represents the first attempt to study this practically relevant problem. It is found that although the two-stage approach is more sophisticated and computationally demanding, it does not yield superior results compared to a direct interpolation of  $CDD/MFP$  on a yearly/monthly basis. This intriguing outcome can be explained by the statistical properties of the precipitation indexes and their underlying weather variable. Our finding suggests that  $CDD$

and *MFP*, both of which are components of the Actuaries Climate Index, enable spatial interpolations with a lower data requirement and smaller computational effort. This property may be considered in tandem with the measures of effectiveness considered by Pan et al. (2022) when evaluating weather indexes for risk management purposes.

# Chapter 6

## Conclusion and Future Research

This thesis explores key challenges in modelling longevity and climate risks. In this concluding chapter, we summarize the main findings from each chapter, highlighting their contributions, and also discuss directions for future research to build upon these results.

Chapter 2 introduces a robust parameter estimation method for the Lee-Carter model, leveraging a  $t$ -PPCA framework to mitigate the influence of outliers. The method enhances the robustness of  $b_x$ , a key parameter for capturing age-specific sensitivity to mortality trends, and allows seamless integration with time series models for the mortality index  $k_t$ . These advancements contribute to more reliable long-term mortality projections, particularly in the presence of extreme events.

In this study, we focus on short-term outliers only. The proposed  $t$ -PPCA method does not account for long-term outliers and structural changes that have prolonged effects. We believe that long-term outliers and structural changes should be dealt with differently, as they should have an influence on parameters governing long-term mortality reduction. One possible way to appropriately incorporate both short-term and long-term anomalies is to implement the  $t$ -PPCA method simultaneously with the method proposed by Li and Li (2017), in which the longest calibration window without structural changes is identified and adopted. Another possible way is to integrate the  $t$ -PPCA method to mortality models with a regime-switching feature, such as that of Milidonis et al. (2011). We leave these possible extensions to future research.

Chapter 3 introduced a computationally efficient least squares estimation method for the Renshaw-Haberman model, providing a practical alternative to MLE. The method minimizes  $L^2$  error using an alternating minimization framework and leverages PCA with missing values to handle age-cohort interactions. Numerical results validated the method’s capability to significantly reduce computation time while maintaining comparable accuracy to traditional methods.

We have argued that the least squares method is less dependent on distributional assumptions, as the objective function that is optimized to yield parameter estimates is not formulated on the basis of any explicit distributional assumption. Nevertheless, it should be noted that even though no explicit distributional assumption is made, the end result produced by the least squares approach is equivalent to that obtained from an MLE with the assumption that log central death rates are normally distributed with the same variance across the entire age range and calibration window. From this angle, MLE may be regarded as more advantageous than the least squares approach, and it permits the user to specify different distributions for log death rates (or death counts) and incorporate a variance structure that depends on age and/or time. For instance, in Poisson-MLE, the variance of the death count in an age-time cell is the expected number of deaths in the age-time cell implied by the model in question, thereby incorporating heteroskedasticity.

The (implicit) assumption of uniform variance across age and time is admittedly significant, particularly when the data set includes older ages for which variances (of empirical log central death rates) are higher due to reduced exposure counts (Brouhns et al., 2002). One way to mitigate this limitation is to weight the squared errors with their corresponding exposure counts, an approach that was considered by Wilmoth (1993) for estimating the original Lee-Carter model. Unfortunately, it is not straightforward to extend the least squares approach proposed in this chapter to incorporate weights. This is because unlike an unweighted PCA with missing values, which can be efficiently solved using an iterative SVD algorithm (as discussed in Section 3.4.3), a weighted PCA with missing values is significantly more complex and is proven to be NP-hard (Gillis and Glineur, 2011). Future research is warranted to develop an appropriate algorithm for solving the weighted PCA with missing values problem entailed in estimating  $c_x$  and  $\gamma_{t-x}$  in the Renshaw-Haberman model.

Chapter 4 presented the `RHa1s` R package, a robust implementation of the least squares method for efficiently fitting Renshaw-Haberman models. The package leverages alternating least squares techniques to address computational challenges associated with MLE and supports multi-population modelling, offering researchers a flexible and scalable framework for mortality analysis. Future work will focus on expanding the package’s capabilities. Currently, the package allows multiple age-period terms but supports only a single age-cohort term. Extending it to accommodate multiple age-cohort terms is our immediate priority. Additionally, the flexibility of the `RHa1s` framework offers opportunities to integrate alternative methods that address computational challenges in the RH model. One such approach involves imposing an additional identification constraint (4.11), which has been implemented under the least squares framework for the H1 model as in Chapter 3; extending this to the general RH model is a future goal.

Chapter 5 contributed to the modelling of spatial basis risk in weather index insurance by evaluating a range of spatial interpolation methods for daily precipitations and precipitation indexes. The study demonstrated the effectiveness of TGK for daily precipitation and *MFP*, showed the advantages of direct interpolation over two-stage approaches. Furthermore, it is found that for the spatial interpolations of *CDD* with the direct approach, all of the three kriging methods (OK, UK and TKG) underperform the benchmark method IDW. As discussed in Section 5.4.2, this outcome is due possibly to a violation of the fundamental assumption of kriging that an observation can be decomposed into a spatial trend plus a spatially correlated error term. In future research, it would be interesting to explore whether more advanced non-linear kriging methods such as multiple indicator kriging and probability kriging (Cressie, 2015) can mitigate the unveiled issue concerning *CDD*.

# References

- Archambeau, C., Delannay, N., and Verleysen, M. (2006). Robust probabilistic projections. In *Proceedings of the 23rd International conference on machine learning*, pages 33–40.
- Basellini, U. and Camarda, C. G. (2022). Lee-carter cohort mortality forecasts. Paper presented at the 2022 European Population Conference.
- Biffis, E. and Blake, D. (2014). Keeping some skin in the game: How to start a capital market in longevity risk transfers. *North American Actuarial Journal*, 18(1):14–21.
- Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Blake, D. and Burrows, W. (2001). Survivor bonds: Helping to hedge mortality risk. *Journal of Risk and Insurance*, pages 339–348.
- Booth, H., Maindonald, J., and Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population studies*, 56(3):325–336.
- Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243.
- Boyd, M., Porth, B., Porth, L., and Turenne, D. (2019). The impact of spatial interpolation techniques on spatial basis risk for weather insurance: An application to forage crops. *North American Actuarial Journal*, 23(3):412–433.

- Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the X-random case. *International statistical review/revue internationale de Statistique*, pages 291–319.
- Brouhns, N., Denuit, M., and Van Keilegom, I. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, 2005(3):212–224.
- Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and economics*, 31(3):373–393.
- Butt, Z., Haberman, S., and Shang, H.-L. (2014). `ilc`: Lee-Carter Mortality Models Using Iterative Fitting Algorithms. R package version 1.0, URL <http://CRAN.R-project.org/package=ilc>.
- Cairns, A. J. (2000). A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, 27(3):313–330.
- Cairns, A. J. (2013). Robust hedging of longevity risk. *Journal of Risk and Insurance*, 80(3):621–648.
- Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.
- Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2011). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355–367.
- Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). 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.

- Cairns, A. J., Blake, D., Kessler, A., Kessler, M., and Mathur, R. (2024). Covid-19 mortality: The proportionality hypothesis. *European Actuarial Journal*, pages 1–46.
- Cairns, A. J., Blake, D. P., Kessler, A., and Kessler, M. (2020). The impact of COVID-19 on future higher-age mortality. Available at SSRN: <https://ssrn.com/abstract=3606988>.
- Cairns, A. J. and El Boukfaoui, G. (2021). Basis risk in index-based longevity hedges: A guide for longevity hedgers. *North American Actuarial Journal*, 25(sup1):S97–S118.
- Cecinati, F., Wani, O., and Rico-Ramirez, M. A. (2017). Comparing approaches to deal with non-Gaussianity of rainfall data in kriging-based radar-gauge rainfall merging. *Water Resources Research*, 53(11):8999–9018.
- Chan, W.-S. (2002). Stochastic investment modelling: a multiple time-series approach. *British Actuarial Journal*, 8(3):545–591.
- Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297.
- Chen, D., Ou, T., Gong, L., Xu, C.-Y., Li, W., Ho, C.-H., and Qian, W. (2010). Spatial interpolation of daily precipitation in China: 1951–2005. *Advances in Atmospheric Sciences*, 27:1221–1232.
- Chen, H. (2013). A family of mortality jump models applied to U.S. data. *Asia-Pacific Journal of Risk and Insurance*, 8(1):105–121.
- Chen, H. and Cummins, J. D. (2010). Longevity bond premiums: The extreme value approach and risk cubic pricing. *Insurance: Mathematics and Economics*, 46(1):150–161.
- Chen, T., Martin, E., and Montague, G. (2009). Robust probabilistic PCA with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis*, 53(10):3706–3716.
- Clarkson, D. B. (1988). Remark AS R71: A remark on algorithm AS 211. The F-G diagonalization algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37(1):147–151.

- Coelho, E. and Nunes, L. C. (2011). Forecasting mortality in the event of a structural change. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(3):713–736.
- Congressional Research Service (2020). American war and military operations casualties: Lists and statistics. <https://crsreports.congress.gov/product/pdf/RL/RL32492>.
- Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, 2016(4):356–383.
- Davis, R. (1952). On the theory of prediction of nonstationary stochastic processes. *Journal of Applied Physics*, 23(9):1047–1053.
- Deaton, A., Paxson, C., et al. (2001). *Mortality, income, and income inequality over time in Britain and the United States*, volume 8534. National bureau of economic research Cambridge, Mass., USA.
- Delwarde, A., Denuit, M., and Partrat, C. (2007). Negative binomial version of the Lee-Carter model for mortality forecasting. *Applied Stochastic Models in Business and Industry*, 23(5):385–401.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.
- Deng, Y., Brockett, P. L., and MacMinn, R. D. (2012). Longevity/mortality risk modeling and securities pricing. *Journal of Risk and Insurance*, 79(3):697–721.
- Dick, W., Stoppa, A., Anderson, J., Coleman, E., and Rispoli, F. (2011). Weather index-based insurance in agricultural development: A technical guide. *International Fund for Agricultural Development (IFAD)*, 18.

- D'Amato, V., Haberman, S., and Russolillo, M. (2012). The stratified sampling bootstrap for measuring the uncertainty in mortality forecasts. *Methodology and Computing in Applied Probability*, 14:135–148.
- Enchev, V., Kleinow, T., and Cairns, A. J. (2017). Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342.
- Erhardt, R. J. and Smith, R. L. (2014). Weather derivative risk measures for extreme events. *North American Actuarial Journal*, 18(3):379–393.
- Fung, M. C., Peters, G. W., and Shevchenko, P. V. (2019). Cohort effects in mortality modelling: a Bayesian state-space approach. *Annals of Actuarial Science*, 13(1):109–144.
- Gillis, N. and Glineur, F. (2011). Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1149–1165.
- Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552.
- Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of hydrology*, 228(1-2):113–129.
- Guo, Y. and Bondell, H. (2023). On robust probabilistic principal component analysis using multivariate  $t$ -distributions. *Communications in Statistics-Theory and Methods*, 52(23):8261–8279.
- Haberman, S. and Renshaw, A. (2009). On age-period-cohort parametric mortality rate projections. *Insurance: Mathematics and Economics*, 45(2):255–270.
- Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.
- Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin: The Journal of the IAA*, 48(2):481–508.

- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.
- Hazell, P., Anderson, J., Balzer, N., Hastrup Clemmensen, A., Hess, U., and Rispoli, F. (2010). The potential for scale and sustainability in weather index insurance for agriculture and rural livelihoods. Technical report, World Food Programme (WFP).
- Hobcraft, J., Menken, J., and Preston, S. (1985). *Age, period, and cohort effects in demography: a review*. Springer.
- Hofstra, N., Haylock, M., New, M., Jones, P., and Frei, C. (2008). Comparison of six methods for the interpolation of daily, European climate data. *Journal of Geophysical Research: Atmospheres*, 113(D21).
- Huber, P. J. and Ronchetti, E. M. (2011). *Robust Statistics*. John Wiley & Sons.
- Hudson, G. and Wackernagel, H. (1994). Mapping temperature using kriging with external drift: theory and an example from Scotland. *International journal of Climatology*, 14(1):77–91.
- Human Mortality Database (2023). Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). Available at <https://www.mortality.org>.
- Hunt, A. and Blake, D. (2021). On the structure and classification of mortality models. *North American Actuarial Journal*, 25(sup1):S215–S234.
- Hunt, A. and Villegas, A. M. (2015). Robustness and convergence in the Lee-Carter model with cohort effects. *Insurance: Mathematics and Economics*, 64:186–202.
- Hyndman, R., Booth, H., Tickle, L., and Maindonald, J. (2014). demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 1.18, URL: <http://CRAN.R-project.org/package=demography>.
- Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

- Ilin, A. and Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000.
- Kibria, B. G. and Joarder, A. H. (2006). A short review of multivariate  $t$ -distribution. *Journal of Statistical research*, 40(1):59–72.
- Kleinow, T. (2015). A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, 63:147–152.
- Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.
- Koissi, M.-C., Shapiro, A. F., and Högnäs, G. (2006). Evaluating and extending the Lee-Carter model for mortality forecasting: Bootstrap confidence interval. *Insurance: Mathematics and Economics*, 38(1):1–20.
- Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.
- Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84(406):460–466.
- Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust statistical modeling using the  $t$  distribution. *Journal of the American Statistical Association*, 84(408):881–896.
- Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4):537–549.
- Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American statistical association*, 87(419):659–671.
- Li, D., Ling, C., Liu, Q., and Peng, L. (2021). Inference for the Lee-Carter model with an AR (2) process. *Methodology and Computing in Applied Probability*, pages 1–29.

- Li, H. and Li, J. S.-H. (2017). Optimizing the lee-carter approach in the presence of structural changes in time and age patterns of mortality improvements. *Demography*, 54:1073–1095.
- Li, J. (2014). A quantitative comparison of simulation strategies for mortality projection. *Annals of Actuarial Science*, 8(2):281–297.
- Li, J. and Heap, A. D. (2008). A review of spatial interpolation methods for environmental scientists. *Geoscience Australia Canberra*.
- Li, J. S.-H. and Chan, W.-S. (2005). Outlier analysis and mortality forecasting: the United Kingdom and Scandinavian countries. *Scandinavian Actuarial Journal*, 2005(3):187–211.
- Li, J. S.-H. and Chan, W.-S. (2007). The Lee-Carter model for forecasting mortality, revisited. *North American Actuarial Journal*, 11(1):68–89.
- Li, J. S.-H., Chan, W.-S., and Cheung, S.-H. (2011). Structural changes in the Lee-Carter mortality indexes: detection and implications. *North American Actuarial Journal*, 15(1):13–31.
- Li, J. S.-H., Hardy, M., and Tan, K. S. (2010). Developing mortality improvement formulas: the Canadian insured lives case study. *North American Actuarial Journal*, 14(4):381–399.
- Li, J. S.-H. and Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, 15(2):177–200.
- Li, J. S.-H., Hardy, M. R., and Tan, K. S. (2009). Uncertainty in mortality forecasting: an extension to the classical Lee-Carter approach. *ASTIN Bulletin: The Journal of the IAA*, 39(1):137–164.
- Li, J. S.-H. and Liu, Y. (2020). The heat wave model for constructing two-dimensional mortality improvement scales with measures of uncertainty. *Insurance: Mathematics and Economics*, 93:1–26.
- Li, J. S.-H., Zhou, K. Q., Zhu, X., Chan, W.-S., and Chan, F. W.-H. (2019). A Bayesian approach to developing a stochastic mortality model for China. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(4):1523–1560.

- Li, J. S.-H., Zhou, R., Liu, Y., Graziani, G., Hall, R. D., Haid, J., Peterson, A., and Pinzur, L. (2020). Drivers of mortality dynamics: Identifying age/period/cohort components of historical US mortality improvements. *North American Actuarial Journal*, 24(2):228–250.
- Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42:575–594.
- Lin, T., Wang, C.-W., and Tsai, C. C.-L. (2015). Age-specific copula-AR-GARCH mortality models. *Insurance: Mathematics and Economics*, 61:110–124.
- Liu, C. and Rubin, D. B. (1995). ML estimation of the  $t$  distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, pages 19–39.
- Liu, Q., Ling, C., Li, D., and Peng, L. (2019). Bias-corrected inference for a modified Lee-Carter mortality model. *ASTIN Bulletin: The Journal of the IAA*, 49(2):433–455.
- Liu, Y. and Li, J. S.-H. (2015). The age pattern of transitory mortality jumps and its impact on the pricing of catastrophic mortality bonds. *Insurance: Mathematics and Economics*, 64:135–150.
- Martínez-Cob, A. (1996). Multivariate geostatistical analysis of evapotranspiration and precipitation in mountainous terrain. *Journal of Hydrology*, 174(1-2):19–35.
- Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322.
- Milidonis, A., Lin, Y., and Cox, S. H. (2011). Mortality regimes and pricing. *North American Actuarial Journal*, 15(2):266–289.
- Moral, F. J. (2010). Comparison of different geostatistical approaches to map climate variables: application to precipitation. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 30(4):620–631.

- Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika*, 79(4):747–754.
- Murphy, A. H. (1970). Predicted forage yield based on fall precipitation in California annual grasslands. *Rangeland Ecology & Management/Journal of Range Management Archives*, 23(5):363–365.
- Neves, C., Fernandes, C., and Hoeltgebaum, H. (2017). Five different distributions for the Lee-Carter model of mortality forecasting: A comparison using GAS models. *Insurance: Mathematics and Economics*, 75:48–57.
- Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., and Perla, F. (2019). A deep learning integrated Lee-Carter model. *Risks*, 7(1):33.
- Norton, M. T., Turvey, C., and Osgood, D. (2012). Quantifying spatial basis risk for weather index insurance. *The Journal of Risk Finance*, 14(1):20–34.
- Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International journal of epidemiology*, 14(1):124–129.
- Pan, Q., Porth, L., and Li, H. (2022). Assessing the effectiveness of the actuaries climate index for estimating the impact of extreme weather on crop yield and insurance applications. *Sustainability*, 14(11):6916.
- Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & geosciences*, 30(7):683–691.
- Pedroza, C. (2006). A Bayesian forecasting model: predicting U.S. male mortality. *Biostatistics*, 7(4):530–550.
- Phillips, D. L., Dolph, J., and Marks, D. (1992). A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain. *Agricultural and forest meteorology*, 58(1-2):119–141.
- Quiggin, J., Karagiannis, G., and Stanton, J. (1994). Crop insurance and crop production: an empirical study of moral hazard and adverse selection. *Economics of agricultural crop insurance: theory and evidence*, pages 253–272.

- Rabinowicz, A. and Rosset, S. (2022). Cross-validation for correlated data. *Journal of the American Statistical Association*, 117(538):718–731.
- Renshaw, A. E. and Haberman, S. (2003). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.
- Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and economics*, 38(3):556–570.
- Richman, R. and Wüthrich, M. V. (2021). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science*, 15(2):346–366.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.
- Roznik, M., Brock Porth, C., Porth, L., Boyd, M., and Roznik, K. (2019). Improving agricultural microinsurance by applying universal kriging and generalised additive models for interpolation of mean daily temperature. *The Geneva Papers on risk and Insurance-Issues and practice*, 44:446–480.
- Shope, C. L. and Maharjan, G. R. (2015). Modeling spatiotemporal precipitation: Effects of density, interpolation, and land use distribution. *Advances in Meteorology*, 2015(1):174196.
- Smith, M. (1947). Populational characteristics of American servicemen in World War II. *The Scientific Monthly*, 65(3):246–252.
- Society of Actuaries (2021). Mortality Improvement Scale MP-2021. Available at <https://www.soa.org/resources/experience-studies/2021/mortality-improvement-scale-mp-2021>.

- SriDaran, D., Sherris, M., Villegas, A. M., and Ziveyi, J. (2022). A group regularisation approach for constructing generalised age-period-cohort mortality projection models. *ASTIN Bulletin: The Journal of the IAA*, 52(1):247–289.
- Sun, B., Guo, C., and Cornelis van Kooten, G. (2014). Hedging weather risk for corn production in Northeastern China: The efficiency of weather-indexed insurance. *Agricultural Finance Review*, 74(4):555–572.
- Sun, X., Manton, M., and Ebert, E. E. (2003). *Regional rainfall estimation using double-kriging of raingauge and satellite observations*. Bureau of Meteorology.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622.
- Turvey, C. G. (2001). Weather derivatives for specific event risks in agriculture. *Applied Economic Perspectives and Policy*, 23(2):333–351.
- Villegas, A. M. and Haberman, S. (2014). 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.
- Villegas, A. M., Haberman, S., Kaishev, V. K., and Millosovich, P. (2017). A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin: The Journal of the IAA*, 47(3):631–679.
- Villegas, A. M., Kaishev, V. K., and Millosovich, P. (2018). StMoMo: An R package for Stochastic Mortality Modelling. *Journal of Statistical Software*, 84(3):1–38.
- Wang, C.-W., Huang, H.-C., and Liu, I.-C. (2011). A quantitative comparison of the Lee-Carter model under different types of non-Gaussian innovations. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36:675–696.
- Willets, R. C. (2004). The cohort effect: insights and explanations. *British Actuarial Journal*, 10(4):833–877.

- Wilmoth, J. R. (1990). Variation in vital rates by age, period, and cohort. *Sociological methodology*, pages 295–335.
- Wilmoth, J. R. (1993). Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. Technical report, Department of Demography, University of California, Berkeley.
- Zhou, K. Q. and Li, J. S.-H. (2020). Asymmetry in mortality volatility and its implications on index-based longevity hedging. *Annals of Actuarial Science*, 14(2):278–301.
- Zhou, K. Q. and Li, J. S.-H. (2021). Longevity greys: What do insurers and capital market investors need to know? *North American Actuarial Journal*, 25(sup1):S66–S96.
- Zhou, R., Li, J. S.-H., and Pai, J. (2018). Evaluating effectiveness of rainfall index insurance. *Agricultural Finance Review*, 78(5):611–625.
- Zhou, R., Li, J. S.-H., and Tan, K. S. (2013). Pricing standardized mortality securitizations: A two-population model with transitory jump effects. *Journal of Risk and Insurance*, 80(3):733–774.
- Zhou, R., Wang, Y., Kaufhold, K., Li, J. S.-H., and Tan, K. S. (2014). Modeling period effects in multi-population mortality models: Applications to Solvency II. *North American Actuarial Journal*, 18(1):150–167.

# APPENDICES

# Appendix A

## Relevant Properties of Multivariate Normal Distributions and Multivariate $t$ -Distributions

In this appendix, we present some properties of multivariate normal and  $t$ -distributions. These properties are relevant to the theoretical work presented in Chapter 2 and Appendix B. The properties are presented in their simplified forms that are sufficient for the purpose of this study. Detailed presentations can be found in classical machine learning textbooks such as that authored by Bishop and Nasrabadi (2006).

### A.1 Normal-Normal Hierarchy

We first consider a multivariate normal distribution of which the mean vector is also normally distributed. Let  $\mathbf{y}$  be a  $p$ -dimensional random vector with the following hierarchical structure:

$$\mathbf{y}|z \sim \mathcal{N}(\mathbf{a} + \mathbf{b}z, \Sigma), \quad z \sim \mathcal{N}(0, \tau), \quad (\text{A.1})$$

where  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\Sigma$  and  $\tau$  are fixed parameters, and  $\mathbf{a} + \mathbf{b}z$  is a linear transformation of the latent random variable  $z$ . It can be shown that the resulting marginal distribution of  $\mathbf{y}$

and the posterior distribution  $z|\mathbf{y}$  are both normal:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{a}, \mathbf{b}\mathbf{b}^T\tau + \Sigma), \quad (\text{A.2})$$

$$z|\mathbf{y} \sim \mathcal{N}((\tau^{-1} + \mathbf{b}^T\Sigma^{-1}\mathbf{b})^{-1}\mathbf{b}^T\Sigma^{-1}(\mathbf{y} - \mathbf{a}), (\tau^{-1} + \mathbf{b}^T\Sigma^{-1}\mathbf{b})^{-1}). \quad (\text{A.3})$$

## A.2 Normal-Gamma Hierarchy

Next, we consider a multivariate normal distribution of which the covariance matrix is inversely proportional to a Gamma-distributed random variable. Let  $\mathbf{y}$  be a  $p$ -dimensional random vector with the following hierarchical structure:

$$\mathbf{y}|u \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\Sigma}{u}\right), \quad u \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \quad (\text{A.4})$$

where  $\boldsymbol{\mu}$ ,  $\Sigma$ , and  $\nu$  are fixed parameters. It is noteworthy that  $\text{Gamma}(\nu/2, \nu/2)$  is equivalent to a chi-square distribution with  $\nu$  degrees of freedom. It can be shown that

$$\mathbf{y} \sim t_\nu(\boldsymbol{\mu}, \Sigma), \quad (\text{A.5})$$

with the density function defined in (2.14). This result implies that a multivariate  $t$ -distribution can be interpreted as a infinite mixture of normal multivariate normal distributions. This property forms the foundation for the EM algorithm devised to obtain ML estimates when a multivariate  $t$ -distribution is involved.

Another relevant result is a special case of the so-called normal-gamma conjugacy, which is commonly used in Bayesian inference. This result says that the posterior distribution of  $u$  is also Gamma under the normal likelihood:

$$u|\mathbf{y} \sim \text{Gamma}\left(\frac{\nu + p}{2}, \frac{\nu + (\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})}{2}\right). \quad (\text{A.6})$$

# Appendix B

## Derivation of the EM Algorithm

In this appendix we derive (2.21) to (2.30), the key equations underpinning the EM algorithm presented in Section 2.4.3. The derivations are based on the hierarchical structure of the  $t$ -PPCA model, which is shown in (2.17).

### B.1 The Complete Log-Likelihood

We first derive the complete log-likelihood, an elaborated version of (2.20). Using the probability density functions of the distributions involved in the  $t$ -PPCA hierarchical structure specified in (2.17), we obtain the following:

$$\begin{aligned} f(\mathbf{y}_t|z_t, u_t) &= \left( (2\pi)^p \cdot \left| \frac{\sigma^2 \mathbf{I}}{u_t} \right| \right)^{-1/2} \\ &\times \exp \left( -\frac{1}{2} (\mathbf{y}_t - \mathbf{a} - \mathbf{b}z_t)^T \cdot \left( \frac{\sigma^2 \mathbf{I}}{u_t} \right)^{-1} \cdot (\mathbf{y}_t - \mathbf{a} - \mathbf{b}z_t) \right), \end{aligned} \quad (\text{B.1})$$

$$f(z_t|u_t) = \left( \frac{2\pi}{u_t} \right)^{-1/2} \cdot \exp \left( -\frac{z_t^2}{2u_t^{-1}} \right), \quad (\text{B.2})$$

$$f(u_t) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u_t^{(\nu/2)-1} \cdot \exp \left( -\frac{\nu}{2} u_t \right). \quad (\text{B.3})$$

Substituting the expressions above into (2.20), we obtain the complete log-likelihood as follows:

$$\begin{aligned}
L_c &= \sum_{t=1}^n \log[f(\mathbf{y}_t, z_t, u_t)] = \sum_{t=1}^n \log[f(\mathbf{y}_t|z_t, u_t)f(z_t|u_t)f(u_t)] \\
&= - \sum_{t=1}^n \left[ \frac{1}{2} \log \left| \frac{\sigma^2 \mathbf{I}}{u_t} \right| + \frac{1}{2} (\mathbf{y}_t - \mathbf{a} - \mathbf{b}z_t)^T \cdot \frac{u_t \mathbf{I}}{\sigma^2} \cdot (\mathbf{y}_t - \mathbf{a} - \mathbf{b}z_t) \right. \\
&\quad \left. + \frac{1}{2} \log \left( \frac{1}{u_t} \right) + \frac{1}{2} u_t z_t^2 - \frac{\nu}{2} \left( \log \frac{\nu}{2} + \log u_t - u_t \right) + \log \Gamma \left( \frac{\nu}{2} \right) \right] + \text{constant}. \\
&= - \sum_{t=1}^n \left[ \frac{p}{2} \log \sigma^2 + \frac{u_t}{2\sigma^2} (\mathbf{y}_t - \mathbf{a})^T (\mathbf{y}_t - \mathbf{a}) - \frac{1}{\sigma^2} (u_t z_t) \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}) \right. \\
&\quad \left. + \frac{1}{2\sigma^2} \mathbf{b}^T \mathbf{b} u_t z_t^2 - \frac{\nu}{2} \left( \log \frac{\nu}{2} + \log u_t - u_t \right) + \log \Gamma \left( \frac{\nu}{2} \right) \right] + \text{constant}. \tag{B.4}
\end{aligned}$$

Note that the parameters to be estimated are  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$ , so any terms that do not involve the parameters can be treated as constants.

Then, taking conditional expectation on both sides of  $L_c$  with respect to  $\mathbf{y}_t$ 's, we obtain the expression of  $\langle L_c \rangle$  that is shown in (2.21).

## B.2 Expectations in the E-Step

Next, we derive the posterior expectations, (2.22) to (2.26), involved in the E-step of the EM algorithm.

### B.2.1 $\langle u_t \rangle$ (2.22)

First, from (2.17), we have  $\mathbf{y}_t|u_t \sim \mathcal{N}(\mathbf{a}, (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})/u_t)$ , as shown in (2.18), and  $u_t \sim \text{Ga}(\nu/2, \nu/2)$ . By the conjugacy between normal likelihood and gamma prior, we obtain

$$u_t | \mathbf{y}_t \sim \text{Ga} \left( \frac{p + \nu}{2}, \frac{(\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{y}_t - \mathbf{a}) + \nu}{2} \right), \tag{B.5}$$

which immediately implies (2.22) holds:

$$\langle u_t \rangle = \mathbb{E}[u_t | \mathbf{y}_t] = \frac{\nu + p}{\nu + (\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{y}_t - \mathbf{a})}. \quad (\text{B.6})$$

### B.2.2 $\langle \log u_t \rangle$ (2.23)

Following (B.5), we can derive  $\langle \log u_t \rangle$ , that is,  $\mathbb{E}[\log u_t | \mathbf{y}_t]$ . Consider a generic random variable  $Y = \log X$ , where  $X \sim \text{Ga}(\alpha, \beta)$ . Its first moment can be easily derived through the moment generating function as follows:

$$\begin{aligned} M_Y(t) &= \mathbb{E}[e^{Yt}] = \mathbb{E}[X^t] = \frac{\Gamma(\alpha + t)}{\Gamma(\alpha)} \cdot \beta^{-t} \\ \implies \mathbb{E}[Y] &= M'_Y(t)|_{t=0} = \frac{1}{\Gamma(\alpha)} [\Gamma'(\alpha + t)\beta^{-t} - \Gamma(\alpha + t)\beta^{-t} \log \beta] \Big|_{t=0} \\ &= \psi(\alpha) - \log \beta, \end{aligned} \quad (\text{B.7})$$

where  $\psi(\cdot) = \Gamma(\cdot)/\Gamma'(\cdot)$  is the digamma function. Substituting  $\alpha$  and  $\beta$  by the corresponding parameters in (B.5), we get

$$\langle \log u_t \rangle = \mathbb{E}[\log u_t | \mathbf{y}_t] = \psi\left(\frac{\nu + p}{2}\right) - \log\left(\frac{\nu + (\mathbf{y}_t - \mathbf{a})^T (\mathbf{b}\mathbf{b}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{y}_t - \mathbf{a})}{2}\right). \quad (\text{B.8})$$

### B.2.3 $\langle z_t \rangle$ (2.24)

As presented in (2.17), we have  $\mathbf{y}_t | z_t, u_t \sim \mathcal{N}(\mathbf{a} + \mathbf{b}z_t, \sigma^2 \mathbf{I}/u_t)$  and  $z_t | u_t \sim \mathcal{N}(0, 1/u_t)$ . To derive (2.24), we apply result (A.3) for the conditioning of multivariate normal distributions, and we can obtain

$$f(z_t | \mathbf{y}_t, u_t) \propto f(\mathbf{y}_t | z_t, u_t) f(z_t | u_t) \sim \mathcal{N}\left((\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}), \frac{\sigma^2 (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1}}{u_t}\right). \quad (\text{B.9})$$

Next, since  $u_t \sim \text{Ga}(\nu/2, \nu/2)$ , applying the scale mixture Gaussian representation of multivariate  $t$ -distributions, as shown in (A.4) and (A.5), we obtain the following result:

$$z_t | \mathbf{y}_t \sim t_\nu\left((\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}), \sigma^2 (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1}\right), \quad (\text{B.10})$$

which immediately implies that (2.24) holds:

$$\langle z_t \rangle = \mathbb{E}[z_t | \mathbf{y}_t] = (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}). \quad (\text{B.11})$$

#### B.2.4 $\langle u_t z_t \rangle$ (2.25)

By the law of total expectation, we have

$$\langle u_t z_t \rangle = \mathbb{E}[u_t z_t | \mathbf{y}_t] = \mathbb{E}[\mathbb{E}[u_t z_t | u_t, \mathbf{y}_t] | \mathbf{y}_t] = \mathbb{E}[u_t | \mathbf{y}_t] \cdot \mathbb{E}[z_t | u_t, \mathbf{y}_t] = \langle u_t \rangle \langle z_t \rangle. \quad (\text{B.12})$$

The second last step is implied by the observation from (B.9) that  $\mathbb{E}[z_t | u_t, \mathbf{y}_t]$  depends only on  $\mathbf{y}_t$ , but not on  $z_t$  and  $u_t$ . The last step uses  $\mathbb{E}[z_t | u_t, \mathbf{y}_t] = \mathbb{E}[z_t | \mathbf{y}_t]$ , which can be noticed from (B.9) and (B.10).

#### B.2.5 $\langle u_t z_t^2 \rangle$ (2.26)

Similarly, by the law of total expectation:

$$\begin{aligned} \langle u_t z_t^2 \rangle &= \mathbb{E}[\mathbb{E}[u_t z_t^2 | u_t, \mathbf{y}_t] | \mathbf{y}_t] \\ &= \mathbb{E}[u_t ((\mathbb{E}[z_t | u_t, \mathbf{y}_t])^2 + \text{Var}(z_t | u_t, \mathbf{y}_t)) | \mathbf{y}_t] \\ &= \mathbb{E}[u_t \cdot \text{Var}(z_t | u_t, \mathbf{y}_t) | \mathbf{y}_t] + \mathbb{E}[u_t | \mathbf{y}_t] \cdot (\mathbb{E}[z_t | u_t, \mathbf{y}_t])^2 \\ &= \mathbb{E}[u_t \cdot \text{Var}(z_t | u_t, \mathbf{y}_t) | \mathbf{y}_t] + \langle u_t \rangle \langle z_t \rangle^2 \\ &= \mathbb{E} \left[ u_t \cdot \frac{\sigma^2 (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1}}{u_t} \middle| \mathbf{y}_t \right] + \langle u_t \rangle \langle z_t \rangle^2 \\ &= \sigma^2 (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} + \langle u_t \rangle \langle z_t \rangle^2. \end{aligned} \quad (\text{B.13})$$

Similar to the arguments above, the third step uses the observation that  $\mathbb{E}[z_t | u_t, \mathbf{y}_t]$  depends on  $\mathbf{y}_t$  only, and the fourth step again draws on the fact that  $\mathbb{E}[z_t | u_t, \mathbf{y}_t] = \mathbb{E}[z_t | \mathbf{y}_t]$ . The fifth step follows from the fact that  $\text{Var}(z_t | u_t, \mathbf{y}_t) = \sigma^2 (\mathbf{b}^T \mathbf{b} + \sigma^2)^{-1} / u_t$ , as shown in (B.9).

### B.3 Updating Formulas in the M-Step

Finally, we derive the updating formulas, (2.27) to (2.30), in the M-step. These formulas are obtained by setting the first-order partial derivatives of  $\langle L_c \rangle$  (2.21) with respect to  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\sigma^2$  and  $\nu$  to zero.

$$\frac{\partial \langle L_c \rangle}{\partial \mathbf{a}} = - \sum_{t=t_1}^{t_n} \left[ (-2) \cdot \frac{\langle u_t \rangle}{2\sigma^2} (\mathbf{y}_t - \mathbf{a}) + \frac{1}{\sigma^2} \langle u_t z_t \rangle \mathbf{b} \right] = 0 \quad (\text{B.14})$$

$$\implies \mathbf{a} = \frac{\sum_{t=t_1}^{t_n} \langle u_t \rangle \mathbf{y}_t - \mathbf{b} \langle u_t z_t \rangle}{\sum_{t=t_1}^{t_n} \langle u_t \rangle} = \frac{\sum_{t=t_1}^{t_n} \langle u_t \rangle (\mathbf{y}_t - \mathbf{b} \langle z_t \rangle)}{\sum_{t=t_1}^{t_n} \langle u_t \rangle}. \quad (\text{B.15})$$

For  $\mathbf{b}$ , we have

$$\frac{\partial \langle L_c \rangle}{\partial \mathbf{b}} = - \sum_{t=t_1}^{t_n} \left[ -\frac{1}{\sigma^2} \langle u_t z_t \rangle (\mathbf{y}_t - \mathbf{a}) + 2 \cdot \frac{1}{2\sigma^2} \langle u_t z_t^2 \rangle \mathbf{b} \right] = 0 \quad (\text{B.16})$$

$$\implies \mathbf{b} = \left[ \sum_{t=t_1}^{t_n} \langle u_t z_t^2 \rangle \right]^{-1} \left[ \sum_{t=t_1}^{t_n} (\mathbf{y}_t - \mathbf{a}) \langle u_t z_t \rangle \right]. \quad (\text{B.17})$$

For  $\sigma^2$ , we have

$$\frac{\partial \langle L_c \rangle}{\partial \sigma^2} = - \sum_{t=t_1}^{t_n} \left[ \frac{p}{2\sigma^2} - \frac{\langle u_t \rangle}{2\sigma^4} (\mathbf{y}_t - \mathbf{a})^T (\mathbf{y}_t - \mathbf{a}) + \frac{1}{\sigma^4} \langle u_t z_t \rangle \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}) - \frac{1}{2\sigma^4} \mathbf{b}^T \mathbf{b} \langle u_t z_t^2 \rangle \right] = 0 \quad (\text{B.18})$$

$$\implies \sigma^2 = \frac{1}{np} \sum_{t=t_1}^{t_n} \left[ \langle u_t \rangle (\mathbf{y}_t - \mathbf{a})^T (\mathbf{y}_t - \mathbf{a}) - 2 \langle u_t z_t \rangle \mathbf{b}^T (\mathbf{y}_t - \mathbf{a}) + \mathbf{b}^T \mathbf{b} \langle u_t z_t^2 \rangle \right]. \quad (\text{B.19})$$

For  $\nu$ , we have

$$\frac{\partial \langle L_c \rangle}{\partial \nu} = - \sum_{t=t_1}^{t_n} \left[ -\frac{1}{2} \left( \log \frac{\nu}{2} + \langle \log u_t \rangle - \langle u_t \rangle \right) - \frac{1}{2} + \frac{1}{2} \frac{\Gamma'(\nu/2)}{\Gamma(\nu/2)} \right] = 0 \quad (\text{B.20})$$

$$\implies 1 + \log \frac{\nu}{2} - \psi \left( \frac{\nu}{2} \right) + \frac{1}{n} \sum_{t=t_1}^{t_n} (\langle \log u_t \rangle - \langle u_t \rangle) = 0. \quad (\text{B.21})$$

# Appendix C

## Theoretical Properties of the Iterative SVD Algorithm

In this appendix, we provide more technical details about the iterative SVD algorithm described in Algorithm 3.

Recall that the objective optimization problem is to update the age-cohort parameters  $(\mathbf{c}, \boldsymbol{\gamma})$  by minimizing the target loss function:

$$L(\mathbf{c}, \boldsymbol{\gamma}) = \sum_{x,s \in \mathcal{O}} (z_{x,s} - c_x \gamma_s)^2, \quad (\text{C.1})$$

where  $s = t - x$  is the cohort and  $\mathcal{O}$  is the set of index of observed values. For sake of the notational clarity, we use  $\hat{z}_{x,s}(\boldsymbol{\theta}) := c_x \gamma_s$  to denote the estimator of the observation  $z_{x,s}$ , as a function of the model parameters  $\boldsymbol{\theta} = (\mathbf{c}, \boldsymbol{\gamma})$ . Thus, we can rewrite the optimization problem as finding  $\boldsymbol{\theta}$  to minimize the  $L^2$  loss function of the observed data:

$$L_{obs}(\boldsymbol{\theta}) = \sum_{x,s \in \mathcal{O}} (z_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2. \quad (\text{C.2})$$

The iterative SVD algorithm, rather than directly solving (C.2), iteratively performs imputation and SVD on the complete data matrix  $\mathbf{Z}_c$ . Let us write the  $L^2$  loss function of the complete data and missing data as:

$$L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) = L_{obs}(\boldsymbol{\theta}) + L_{mis}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) \quad (\text{C.3})$$

$$L_{mis}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) = \sum_{x,s \notin \mathcal{O}} (\tilde{z}_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2, \quad (\text{C.4})$$

where  $\tilde{z}_{x,s}$  is the imputed missing value. The iterative SVD algorithm minimizes the total  $L^2$  error (C.3) as a function  $L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}})$  with respect to both the model parameter  $\boldsymbol{\theta}$  and the set of imputed values  $\tilde{\mathbf{z}} = \{\tilde{z}_{x,s} | x, s \notin \mathcal{O}\}$ .

## C.1 Convergence of the Iterative SVD Algorithm

We first show that the iterative SVD algorithm always converges by showing that the algorithm can be represented as an alternating minimization procedure:

1. For a fixed  $\tilde{\mathbf{z}}$ , let  $\boldsymbol{\theta}^*$  be the minimizer of  $L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}})$  with respect to  $\boldsymbol{\theta}$ :

$$\begin{aligned} \boldsymbol{\theta}^* &= \arg \min_{\boldsymbol{\theta}} L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) \\ &= \arg \min_{\boldsymbol{\theta}} \left[ \sum_{x,s \in \mathcal{O}} (z_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2 + \sum_{x,s \notin \mathcal{O}} (\tilde{z}_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2 \right] \\ &= \arg \min_{\boldsymbol{\theta}} \left[ \sum_{x,s} (z'_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2 \right], \end{aligned} \quad (\text{C.5})$$

where  $z'_{x,s}$  denotes the data with imputed missing values, which equals  $z_{x,s}$  for  $x, s \in \mathcal{O}$  and equals  $\tilde{z}_{x,s}$  for  $x, s \notin \mathcal{O}$ . Since  $\hat{z}_{x,s}(\boldsymbol{\theta}) := c_x \gamma_s$ , the minimizer  $\boldsymbol{\theta}^*$  can be found by performing PCA to the approximate complete matrix  $\mathbf{Z}_c$ , as described in Step 2 of Algorithm 3.

2. For a fixed  $\boldsymbol{\theta}$ , let  $\tilde{\mathbf{z}}^*$  be the minimizer of  $L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}})$  with respect to  $\tilde{\mathbf{z}}$ :

$$\begin{aligned} \tilde{\mathbf{z}}^* &= \arg \min_{\tilde{\mathbf{z}}} L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) \\ &= \arg \min_{\tilde{\mathbf{z}}} [L_{mis}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) + \text{constant}] \\ &= \arg \min_{\tilde{\mathbf{z}}} \left[ \sum_{x,s \notin \mathcal{O}} (\tilde{z}_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2 + \text{constant} \right]. \end{aligned} \quad (\text{C.6})$$

It is straightforward to see that the minima is achieved when  $\tilde{z}_{x,s} = \hat{z}_{x,s}(\boldsymbol{\theta})$  and so  $\tilde{\mathbf{z}}^* = \{\tilde{z}_{x,s}(\boldsymbol{\theta}) | x, s \notin \mathcal{O}\}$ . This solution is exactly imputing the missing values by the PCA reconstruction  $\hat{z}_{x,s}(\boldsymbol{\theta})$  with parameters  $\boldsymbol{\theta}$ , as described in Step 3 of Algorithm 3.

## C.2 The Iterative SVD Algorithm Minimizes the Target Loss Function

We next show that the iterative SVD algorithm minimizes the target loss function (C.2). More precisely, the minimizer  $\boldsymbol{\theta}^*$  obtained by iteratively minimizing the total  $L^2$  error  $L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}})$  is equivalent to the one obtained by directly minimizing the  $L^2$  error  $L_{obs}(\boldsymbol{\theta})$  of the observed data.

Following (C.6),

$$\tilde{\mathbf{z}}^* = \arg \min_{\tilde{\mathbf{z}}} L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}}) = \arg \min_{\tilde{\mathbf{z}}} \left[ \sum_{x,s \notin \mathcal{O}} (\tilde{z}_{x,s} - \hat{z}_{x,s}(\boldsymbol{\theta}))^2 \right], \quad (\text{C.7})$$

which immediately implies that

$$\left. \frac{\partial L_{tot}}{\partial \tilde{\mathbf{z}}} \right|_{\tilde{\mathbf{z}}=\tilde{\mathbf{z}}^*} = 0 \quad (\text{C.8})$$

and

$$L_{mis}(\boldsymbol{\theta}, \tilde{\mathbf{z}}^*) = \sum_{x,s \notin \mathcal{O}} (\hat{z}_{x,s}(\boldsymbol{\theta}) - \hat{z}_{x,s}(\boldsymbol{\theta}))^2 = 0, \quad (\text{C.9})$$

Therefore, we can obtain

$$L_{tot}(\boldsymbol{\theta}, \tilde{\mathbf{z}}^*) = L_{obs}(\boldsymbol{\theta}) + L_{mis}(\boldsymbol{\theta}, \tilde{\mathbf{z}}^*) = L_{obs}(\boldsymbol{\theta}). \quad (\text{C.10})$$

and consequentially,

$$\frac{dL_{obs}}{d\boldsymbol{\theta}} = \left. \frac{dL_{tot}}{d\boldsymbol{\theta}} \right|_{\tilde{\mathbf{z}}=\tilde{\mathbf{z}}^*} = \left. \frac{\partial L_{tot}}{\partial \boldsymbol{\theta}} \right|_{\tilde{\mathbf{z}}=\tilde{\mathbf{z}}^*} + \underbrace{\left. \frac{\partial L_{tot}}{\partial \tilde{\mathbf{z}}} \right|_{\tilde{\mathbf{z}}=\tilde{\mathbf{z}}^*}}_{=0 \text{ from (C.8)}} \cdot \frac{\partial \tilde{\mathbf{z}}}{\partial \boldsymbol{\theta}} = \left. \frac{\partial L_{tot}}{\partial \boldsymbol{\theta}} \right|_{\tilde{\mathbf{z}}=\tilde{\mathbf{z}}^*} \quad (\text{C.11})$$

which suggests that the minimizer  $\boldsymbol{\theta}^*$  of  $L_{tot}(\boldsymbol{\theta}, \mathbf{z}_{imp})$  coincides with the minimizer of  $L_{obs}(\boldsymbol{\theta})$ , and thus shows that the iterative SVD algorithm indeed implicitly solves our objective optimization problem (3.22).