Multiple-Decrement Compositional Forecasting with the Lee-Carter Model

by

Tianyu Guan

B.Sc., Jilin University, 2011

A Project Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2014

APPROVAL

Name: Tianyu Guan

Degree: Master of Science

Title of A Project: Multiple-Decrement Compositional Forecasting with the Lee-Carter Model

Examining Committee: Dr. Tim Swartz, Professor
Chair

Dr. Gary Parker,
Associate Professor, Senior Supervisor

Dr. Cary Chi-Liang Tsai,
Associate Professor, Supervisor

Dr. Michelle Zhou,
Assistant Professor, Internal Examiner

Date Defended: July 10th, 2014

Partial Copyright Licence

Abstract

Changes in cause of death patterns have a great impact on health and social care costs paid by government and insurance companies. Unfortunately an overwhelming majority of methods for mortality projections is based on overall mortality with only very few studies focusing on forecasting cause-specific mortality. In this project, our aim is to forecast cause-specific death density with a coherent model. Since cause-specific death density obeys a unit sum constraint, it can be considered as compositional data. The most popular overall mortality forecasting model, Lee-Carter model, is applied on compositional cause-specific death density. The predicted cause-specific death density is used to calculate life insurance and accidental death rider.

Keywords : Lee-Carter model; compositional data analysis; death density; cause of death; accidental death rider

Acknowledgments

I would like to express my deep gratitude to everyone I have met in Vancouver, because the first time I leave my family and friends, they make Vancouver feel like home. During the past three years, they helped me become stronger, more independent and professional.

I want to thank my supervisor Dr. Gary Parker for his professional guidance and continued support. Without his help and professional insight, this project would not be done. He respected my interest and idea, and was always there to help me. He understood my difficulties and led me to find solutions. I would also like to thank Dr. Cary Tsai and Yi Lu, for teaching me a lot about actuarial science and caring for my career.

Special thanks to Joan Hu and Rachel Altman for their patience in answering my questions about my project. Many thanks to professors Richard Lockhart, Thomas Loughin, Tim Swartz, Jiguo Cao, Derek Bingham, Barbara Sanders and Robin Insley for their support, guidance and precious suggestions.

In addition, I want to thank Charlene Bradbury, Kelly Jay, Sadika Jungic, who provide us with a very warm atmosphere to study. I really appreciate your patience, kindness and constant support in the statistics department.

To all my friends, thank you for your encouragement and support. To my family, thank you for your love and listening.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Government, public and private providers of pension funds and annuities, are exposed to longevity risk. According to Human Mortality Database at http://www.mortality.org, from 1950 to 2010, life expectancy rose from 68 to 79 years in USA, 68 to 81 in Canada and 59 to 83 in Japan. If the longevity risk is underestimated, governments and pension providers will be affected financially in the future. Individuals will also be exposed to financial risks, and living longer may cause them to run out of retirement income and consequently die in poverty or burden their relatives. Among many ways of forecasting mortality, a well known one is the Lee-Carter model (1992). In some cases, the financial risks vary by many factors such as age, sex and causes of death. For instance, the health care costs in the last year of life by cause of death vary a lot for female in Netherlands (Polder et al., 2006). Therefore, disaggregation of death enables us to have a more thorough understanding of the financial risks.

The termination from a given status is called a decrement in actuarial science. The Lee-Carter model is a single decrement model where the status of interest corresponds to an individual being alive and the decrement is the death of that individual. In this context, the single decrement model treats all causes of death as one decrement. When the cause of decrement is also of interest, for example, in a study of health care costs in the last year of life, a multiple decrement model is needed. In biostatistics, multiple decrement models are often referred to as competing risks models.

The aim of this project is to study and forecast financial risks that vary with the cause of death. There are many cause-specific projection models proposed in the literature. Wilmoth (1995) claimed that for "proportional rates of change models" all-cause projection is always more pessimistic than the total mortality projection. The reason is that all-cause mortality tends to be dominated by those causes of death that are decreasing the slowest. Oeppen (2008) stated that "it is clear that the

dependencies or relative balances, between the decrements have not been adequately modeled". In order to solve the problem, in this project we model the cause-specific death density instead of disaggregated mortality. The sum of death densities over causes must add up to 1. We expect that this feature can help obtain coherent forecasts of cause-specific death density that are not as pessimistic as mortality projections disaggregated by cause.

The constraint on cause-specific death density brings the difficulty that standard statistical techniques lose classical interpretation. Compositional Data Analysis (CoDa) introduced by Aitchison (1986) solves the problem by transforming the constrained cause-specific death density to real space where the standard statistical techniques function well. The CoDa equivalent Lee-Carter Model (CoDa LC Model) is a method of producing coherent forecasts of cause-specific death density.

1.2 Lee-Carter Model and its Extensions

Lee and Carter (1992) introduced a famous mortality projection model. Since it was introduced, the model had a variety of applications. The model expresses the log mortality rate matrix as a linear function of a period factor with parameters depending on age. The fit is obtained by Singular Value Decomposition (SVD). The period index is modeled by an ARIMA time series. The forecasts of age-specific rates are based on the forecasts of the period index. The model has several advantages: it combines a parsimonious demographic model with statistical time series methods; forecasting is based on persistent long-term historical trends and patterns; and probabilistic confidence regions are provided for the forecasts (Lee and Carter, 1992).

The Lee-Carter model has many variants and extensions. The variants and extensions improve the Lee-Carter model by using better models for the period index, providing robust estimation, etc. For example, Lee and Miller (2001) proposed a method that is different from the Lee-Carter method mainly in three ways: first, Lee and Miller (2001) only involved data from after 1950 instead of 1900 in the Lee-Carter model; second, the period index is reestimated to match the observed average life expectancy at birth instead of the observed number of deaths; third, the jump-off rates are considered to be the actual rates in the jump-off year. The Booth-Maindonald-Smith variant (BMS model) proposed by Booth et al. (2002), adjusts the period index by fitting the age distribution of death (the Poisson distribution is used to model the death process). The BMS model also introduces a method to choose the most appropriate fitting period (the ending year of fitting period is the latest available data, and the problem reduces to finding the most appropriate starting year) under the assumption of linear period index. Another extension is the Hyndman-Ullah functional data method (Hyndman and Ullah, 2007) with ideas from functional data analysis, nonparametric smoothing and robust statistics (robust estimation allows for temporary shocks such as wars and diseases) combined. The Hyndman-Ullah model assumes that mortality is a smooth function of age and

uses nonparametric smoothing methods for estimation. De Jong and Tickle (2006) proposed the De Jong-Tickle Lee-Carter model, which improves the Lee-Carter model by adding Kalman filtering and multiple principal components. Actually, the Lee-Carter model is a special case of the De Jong-Tickle Lee-Carter model. Booth et al. (2006) compared the short-to-medium-term accuracy of the Lee-Carter model, the Lee-Miller model, the Booth-Maindonald-Smith variant, the Hyndman-Ullah functional data method and the De Jong-Tickle Lee-Carter model.

Booth and Tickle (2008) summarized that mortality projection models mainly contain three factors: age, period and cohort. Obviously, the Lee-Carter model as well as the above variants and extensions are all two-factor models with age and period factors. Some extensions, however, improve the Lee-Carter model by adding a cohort effect. For instance, Renshaw and Haberman (2003) proposed a three-factor model that adds a cohort effect to the Lee-Carter model. In their paper, they provided two iterative procedures to estimate the parameters. Currie (2006) introduced the Age-Period-Cohort model, which is a special case of the Renshaw and Haberman model. Cairns, Blake and Dowd (2006) introduced the CBD model, which fits the logit of mortality rates.

Booth and Tickle (2008) also included a section discussing decomposition by cause of death. Wilmoth (1995) believed that, for "proportional rates of change models" including Lee-Carter, the mortality forecasts will always be higher for the sum of cause-specific forecasts than the overall mortality forecasts. The reason is that "causes of death which are slow to decline come to dominate in the long run" (Booth and Tickle, 2008). They summarized that "although forecasting mortality by separate causes of death has been advocated from a theoretical perspective as a means of gaining accuracy in overall mortality forecasting (e.g. Crimmins, 1981), but subsequent experience has often proved otherwise". In order to solve the above problem, Oeppen (2008) proposed to model the density of death in the life table, which are intrinsically relative since they obey a unit sum constraint for both single-decrement and multiple-decrement life tables. The density of death therefore can be treated as compositional data and analyzed according to *The Statistical Analysis of Compositional Data* written by Aitchison (1986).

1.3 Compositional Data Analysis

Compositional data is defined as random vectors with strictly positive components whose sum is constant. Compositional data is common in a variety of fields, such as geology (compositions of rocks), economy (income/expenditure distribution), chemistry (chemical compositon) and so on. Compositional data analysis is currently a popular topic of research in many fields. The sum constraint on the data makes it hard to perform statistical analysis that is well developed on real space. The problem was once mentioned by Karl Pearson in 1897 and Felix Chayes in 1960's. Aitchison (1986) first proposed theoretically solution in 1980's. In his book "The Statistical Analysis of Compositional Data", he explained in details the simplex sample space for compositional data, the

perturbations (operations of compositional data) and theories based on log-ratios. Aitchison (1986) proposed a way to transform the simplex sample space of compositional data to real space so that the standard statistical methods can be applied. This project uses Aitchison's idea to transform the cause-specific death density into real space and then applies the Lee-Carter structure on the resulting data.

1.4 Outline

The project is organized as follows. Chapter 2 provides a brief introduction to Compositional Data Analysis. We will first introduce the Simplex Sample Space and then briefly describe perturbations, which are operations on compositions. Based on the perturbations, we finally introduce centring and centred log-ratio transformations.

Chapter 3 reviews the Lee-Carter model and presents France (female) mortality projections. Then details about the CoDa LC model are given in Chapter 4 for the single-decrement case and Chapter 5 for the multiple-decrement case. In Chapter 4, France (female) death density is projected. In Chapter 5, the model is applied on Japan (female) cause-specific death density. Chapter 6 discusses whether the CoDa LC model still works when the data is only available for about 10 years. Finally, in Chapter 7, the multiple-decrement CoDa LC model is used to calculate the expectation and variance of a 20-year life insurance with a 20-year accidental death rider using USA data. Chapter 8 is a brief conclusion of this project.

Chapter 2

Compositional Data Analysis (CoDa)

Sometimes for a positive vector **x**, our interest lies on the relative magnitudes x_i/x_j of its parts (proportions) but not on the absolute values. In order to study such proportions, let's consider compositional data, compositional operators and some consequential results. First let's talk about what is compositional data. Any vector **x** with positive elements $x_1, ..., x_D$ representing proportions of some whole is subject to the obvious constraint:

$$x_1 + ... + x_D = 1. \tag{2.1}$$

Compositional data consisting of such vectors play an important role in many disciplines and often display appreciable variability from vector to vector (Aitchison, 1986). Typical examples of compositional data include, mineral compositions of rocks (Geology), chemical composition (Chemistry), portfolio composition (Economics) and so on. For example (Aitchison, 1986), the geochemical compositions of rock (Table 2.1) can be expressed in terms of percentages by weight of ten or more major oxides.

	Percentage compositions of major oxides by weight									
$Type$	SiO_2	TiO_2	Al_2O_3	$TotFe$	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
Permian	60.54	1.32	15.22	6.95	0.21	2.33	3.18	4.81	4.84	0.60
	54.30	1.24	16.67	8.70	0.07	4.24	8.34	3.41	2.52	0.49
	52.17	0.82	20.05	8.38	0.10	2.28	9.29	3.22	2.99	0.69
Post-	55.95	1.26	18.54	7.24	0.28	1.20	3.30	6.14	5.67	0.45
Permian	45.40	1.34	20.14	8.00	0.06	9.29	9.59	3.89	1.38	0.90
	46.59	1.06	15.99	11.20	0.30	10.50	10.45	2.03	1.45	0.43

Table 2.1: Some typical major-oxide compositions of Permian and post-Permian rocks.

If vector **x** is compositional, the vectors **x** and k**x**, with $k > 0$, provide us the same information. It is sometimes difficult to work with the unit-sum constraint, since "it is either ignored or improperly incorporated into the statistical modelling and there results an inadequate or irrelevant analysis with a doubtful or distorted inference." (Aitchison, 1986). Since N-variate data that subjects to a unit sum form an N-1 dimensional sample space or simplex, some well developed statistical methods no longer work on the simplex. As a solution to this problem, Aitchison (1986) suggested to transform the data to the real space by the log-ratio transformation.

2.1 The Simplex Sample Space

Let's define composition first. A composition **x** of D parts is a $D \times 1$ vector with positive components $x_1, ..., x_D$ whose sum is 1. Let $d = D - 1$. Since $x_D = 1 - x_1 - ... - x_{D-1}$, a D-part composition is actually a d-dimensional vector and therefore we are able to introduce some d-dimensional set to represent D-part compositions. This means that a d-part subvector $(x_1, ..., x_d)$ provides complete information of a composition **x**. Another way to determine a composition is to define d ratios:

$$r_i = \frac{x_i}{x_D}, \quad (i = 1, ..., d)\,. \tag{2.2}$$

Then the compositions can be expressed as:

$$\begin{aligned} x_i &= \frac{r_i}{r_1 + ... + r_d + 1}, \quad (i = 1, ..., d)\,, \\ x_D &= \frac{1}{r_1 + ... + r_d + 1}. \end{aligned} \tag{2.3}$$

The sample space for D-part compositional vectors, whose components are proportions of some unit, is the d-dimensional unit simplex:

$$S^d = \{(x_1, ..., x_D) : x_i > 0 \ (i = 1, ..., D), x_1 + ... + x_D = 1\}. \tag{2.4}$$

Back to Table 2.1, we know that the compositions belong to a simplex sample space with $d = 9$. We use R^d and S^d to represent d-dimensional real space and d-dimensional simplex. Then a d-dimensional positive space R^d_+ is defined as:

$$R^d_+ = \{(x_1, ..., x_d) : x_i > 0\}. \tag{2.5}$$

The relationship among S^d, R^d_+, R^d is

$$S^d \subset R^d_+ \subset R^D. \tag{2.6}$$

2.2 Perturbations

For compositional purposes it is convenient to impose on this an algebraic-geometric structure converting S^D into a metric vector space. The fundamental operations of change in the simplex are those of perturbation (compositional addition), inverse perturbation (compositional subtraction), power transformation (compositional multiplication) and inverse power transformation (compositional division). Let **x** be a D-part composition (**x** $\in S^D$) and **y** a D-vector with positive elements (**y** $\in R_+^D$). Then the operation which is termed a perturbation is defined as:

$$\mathbf{x} \oplus \mathbf{y} = C[x_1 y_1, ..., x_D y_D], \tag{2.7}$$

where C is the closure or normalizing operation such that the elements of a positive vector are divided by their sum. The above operation is a one-to-one transformation from S^D to S^D. And since $C\mathbf{y} \in S^d$, we can restrict perturbing vectors to the simplex S^D. For example, **x**=(0.3, 0.5, 0.2) and **y**=(0.6, 0.1, 0.3), then **x** $\oplus$ **y**=$\frac{1}{0.29}$(0.18, 0.05, 0.06)=(0.62, 0.17, 0.21). Ternary diagram (see Figure 2.1) shows the effect of a perturbation (compositional addition). Compositions **x** and **y** can be found as two points, (0.3, 0.5, 0.2) and (0.6, 0.1, 0.3) respectively, and the result of the perturbation **x** $\oplus$ **y** is marked as (0.62, 0.17, 0.21).

We define the inverse perturbation of **x** and **y** as:

$$\mathbf{x} \ominus \mathbf{y} = C[x_1/y_1, ..., x_D/y_D]. \tag{2.8}$$

For example, **x** $\ominus$ **y**=$\frac{1}{6.17}$(0.50, 5.00, 0.67)=(0.08, 0.81, 0.11). Ternary diagram (see Figure 2.2) shows the effect of an inverse perturbation (compositional subtraction). The result of the inverse perturbation **x** $\ominus$ **y** is marked as (0.08, 0.81, 0.11).

Given a D-part composition **x** $\in S^D$ and a real number $a \in R^1$, the power transformed composition (compositional multiplication) is

$$a \otimes \mathbf{x} = C[x_1^a, ..., x_D^a]. \tag{2.9}$$

For example, **x**=(0.3, 0.5, 0.2) and a is -6, -5,..., 5, 6 respectively. The effect of scalar a is shown in Figure 2.3. A distance measure in the simplex space, $\Delta_S : S^D \times S^D \to R_{\geq 0}$, is defined as:

$$\Delta_S(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{D} \left\{ log \frac{x_i}{g(\mathbf{x})} - log \frac{y_i}{g(\mathbf{y})} \right\}^2 \right]^{1/2}, \tag{2.10}$$

where $g(\mathbf{x})$ denotes the geometric mean $(x_1 x_2 \cdots x_D)^{1/D}$. Let **x** $\in S^d$ be a random vector. We define the "centre" $\xi \in S^D$, which minimizes the expectation of $\Delta_S(\mathbf{x}, \xi)$, i.e. $E\{\Delta_S(\mathbf{x}, \xi)\}$, as:

$$\xi = cen(\mathbf{x}) = C(exp(E(log\mathbf{x}))). \tag{2.11}$$

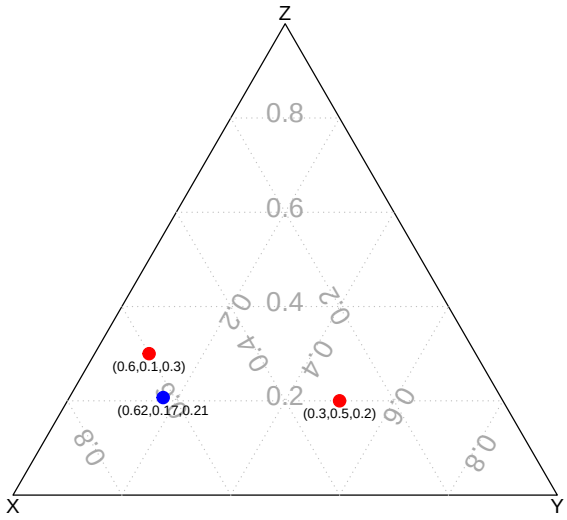

Figure 2.1: Compositional addition.

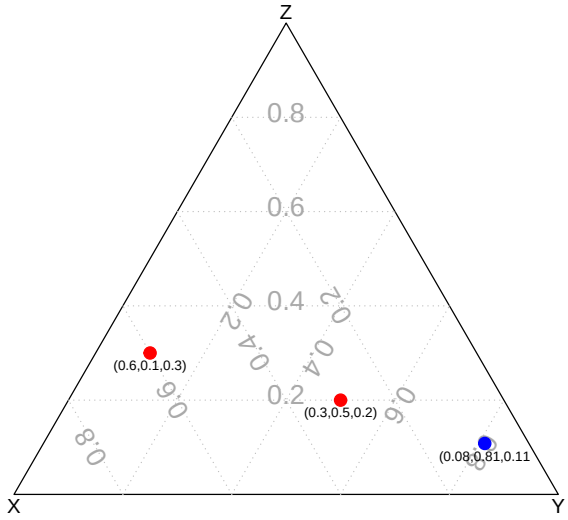

Figure 2.2: Compositional subtraction.

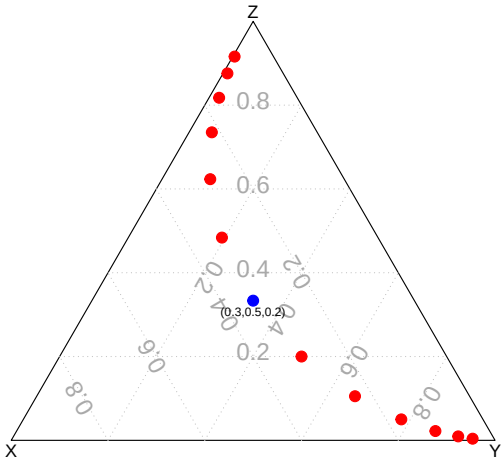

Figure 2.3: Compositional multiplication.

A compositional data set $\mathbf{X} = \{x_{ij}\}$ has D columns and N rows where each row is a composition. The estimate $\hat{\xi}$ of ξ is given by:

$$\hat{\xi} = C[g_1, ..., g_D], \tag{2.12}$$

where g_i=$(x_{1i} \cdots x_{Ni})^{1/N}$ is the geometric mean of the i^{th} column. Then "centring" is defined by calculating the $\hat{\xi}$ composition and using the inverse perturbation operator to subtract it from each row of matrix **X**. In order to open the sum-constrained data of the simplex to the full range of linear models in the real space, Aitchison (1986) defined the centred log-ratio transformation CLR and its inverse CLR^{-1}. The log-ratio transformation, CLR: $S^D \to U^D$, is as follows:

$$\mathbf{z} = CLR(\mathbf{x}) = ln\left[\frac{x_1}{g(\mathbf{x})}, ..., \frac{x_D}{g(\mathbf{x})}\right], \tag{2.13}$$

where U^D is a hyperplane of R^D:

$$U^D = \{(u_1, ..., u_D) : u_1 + ... + u_D = 0\}. \tag{2.14}$$

The inverse transformation $CLR^{-1} : U^D \to S^D$ is

$$\mathbf{x} = CLR^{-1}(\mathbf{z}) = C\left[exp(z_1), ..., exp(z_D)\right]. \tag{2.15}$$

2.3 Rank-r Approximation

There is, for a compositional data **X**, a central result analogous to the singular value decomposition for data sets associated with the sample space R^D, on which much of multivariate statistical methodology is based. Any compositional data matrix **X** can be decomposed in a power-perturbation form as follows:

$$x_n = \hat{\xi} \oplus (u_{n1}s_1 \otimes \beta_1) \oplus ... \oplus (u_{nR}s_R \otimes \beta_R), \tag{2.16}$$

where the u's are power components specific to each composition, s_i's are the "singular values", and the β_j's are orthogonal compositions and R is a readily defined rank of the compositional data set. In practice R is commonly of dimension D-1, the full dimension of the simplex (Aitchison et al., 2003). In a way similar to that for data sets in R^D we may consider an approximation of order $r < R$ to the compositional data set given by:

$$x_n = \hat{\xi} \oplus (u_{n1}s_1 \otimes \beta_1) \oplus ... \oplus (u_{nr}s_r \otimes \beta_r). \tag{2.17}$$

In this project, we will use the above approximation with $r = 1$ or $r = 2$.

Chapter 3

Lee-Carter Model

The Lee-Carter model has been widely used in forecasting mortality since its publication (Lee and Carter, 1992). The authors proposed a model based on a variety of previous work (such as Bozik and Bell (1989) and Lederman (1969)). The Lee-Carter model is a two-factor (age and time) model. More specifically, the Lee-Carter model uses a singular value decomposition (SVD) method to extract the age-specific parameters and a time-varying index. The extracted time-varying index is adjusted by refitting the total observed number of deaths. The Lee-Carter model has some strengths, which are its simplicity, parsimony and robustness in the context of linear trends in an age-specific death rates. While other methods have subsequently been developed (e.g., Brouhns et al., 2002; Renshaw and Haberman, 2003), the Lee-Carter method is often taken as the point of reference (Booth et al., 2006).

3.1 The Model

Let $m_{x,t}$ be the central death rate for age x in year t. The model we will use to fit the matrix of death rates is

$$ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}, \tag{3.1}$$

or

$$m_{x,t} = e^{a_x + b_x k_t + \varepsilon_{x,t}}, \tag{3.2}$$

where a_x's are age-specific constants indicating the average over time of the log mortality, b_x's are age-specific constants indicating which rates decline slowly in response to changes in time-varying index, k_t's are time-varying indices of level of mortality, and $\varepsilon_{x,t}$'s are error terms with mean 0 and variance $\sigma^2_{x,t}$ describing the age-specific historical influences not captured by the model.

Suppose the age-specific constant vectors **a** (with elements a_x's) and **b** (with elements b_x's) and the time-varying index **k** (with elements k_t's) are one solution of (3.1). Then for any constant c,

vectors $\mathbf{a} - \mathbf{b}c$, $\mathbf{b}$ and $\mathbf{k} + c$ are also one solution. Actually we can find that a_x, b_x and k_t are only determined up to a linear transformation. The model is obviously underdetermined. Therefore we need to normalize b_x and k_t by adding two restrictions: $\sum_x b_x = 1$ and $\sum_t k_t = 0$.

The above two restrictions imply that a_x's are simply the averages over time of the log mortality. Assume we have T years of data, then a_x can be expressed as:

$$a_x = \frac{1}{T} \sum_t ln(m_{x,t}). \tag{3.3}$$

3.2 Model Fitting

Now we need to find least squares estimates of the two age-specific constants a_x and b_x as well as the time-varying index k_t. The estimation of parameters of the model cannot be obtained by ordinary regression methods because there are no given regressors. The singular value decomposition (SVD) method can be used to find a least squares solution. SVD is applied to the matrix of the logarithms of the rates after the averages over time of the age-specific rates have been subtracted.

Now we are able to build the matrix for SVD, denoted as $\tilde{\mathbf{m}}$:

$$\tilde{m}_{x,t} = ln(m_{x,t}) - a_x. \tag{3.4}$$

In the matrix form, let $\mathbf{m}$ be the matrix of central death rate with elements $m_{x,t}$ and let vector $\mathbf{a}$ be the row average of $ln(\mathbf{m})$; then we can construct a matrix $\boldsymbol{a}$, which has the same dimension as $ln(\mathbf{m})$ and with every column being vector $\boldsymbol{a}$. The matrix $\tilde{\mathbf{m}}$ is therefore expressed as:

$$\tilde{\mathbf{m}} = ln(\mathbf{m}) - \boldsymbol{a}. \tag{3.5}$$

For $\tilde{\mathbf{m}}$ of size $N \times T$, the SVD is a factorization of the form:

$$\tilde{\mathbf{m}} = U_{N\times N} S_{N\times T} V'_{T\times T}, \tag{3.6}$$

where U and V are orthogonal matrices and S is a diagonal matrix. The columns of U and V are called the left and right singular vectors of $\tilde{\mathbf{m}}$ respectively. The diagonal entries of S are the singular values of $\tilde{\mathbf{m}}$. The first right and left singular vectors and leading singular value of the SVD, after the normalization described above, provide a unique solution. To be specific,

$$b_x = U(x,1) / \sum_x U(x,1) \tag{3.7}$$

and

$$k_t = S(1,1) \times V(t,1) \times \sum_x U(x,1). \tag{3.8}$$

Here k_t is estimated to minimize errors in the logs of death rates rather than the death rates themselves. As a result, we need to take a second step to reestimate k_t, taking the a_x and b_x values

from the first step as given above in (3.3) and (3.7). The reestimation is to adjust the value of k_t, so that given the exposure numbers $e_{x,t}$'s, the implied number of deaths in each year equals the actual number of deaths, that is,

$$\sum_x D_{x,t} = \sum_x e_{x,t} e^{a_x + b_x k_t}, \tag{3.9}$$

where $D_{x,t}$ denotes the actual number of deaths between ages x and $x+1$ in year t. The updated estimates for k_t's are different from the direct SVD estimates. The main reason is that, when fitting the log-transformed rates, the low death rates of the younger ages receive the same weight as the high death rates of the older ages, yet they contribute far less to the total deaths.

3.3 The Fitted Model

3.3.1 The Data

The matrix **D** contains the element $D_{x,t}$, and the matrix **E** contains the element $e_{x,t}$. For both matrices, age is arranged in rows and time in columns. Then we can construct the matrix of central death rates **m**, with elements $m_{x,t}$:

$$\mathbf{m} = \mathbf{D}/\mathbf{E}. \tag{3.10}$$

In many cases, we will need to calculate $q_{x,t}$, the probability of dying in a single year for someone aged x in year t. If we assume the force of mortality, denoted by $\mu_{x,t}$, is constant over each age interval and calendar year, then:

$$q_{x,t} = 1 - e^{-\mu_{x,t}}. \tag{3.11}$$

We will present results for the France 1900 to 2012 mortality experiences for female, with age groups 0, 1-4, 5-9, 10-14,..., 95-99, 100+. The France death and exposure tables can be downloaded from the Human Mortality Database at http://www.mortality.org.

3.3.2 The Estimated Parameters

Fitting a Lee-Carter model to the France mortality data for female, we can obtain the age-specific constants a_x and b_x. The values of a_x and b_x are pictured in Figure 3.1.

3.4 Modeling and Forecasting the Mortality Index, k

The time-varying index k_t values obtained by SVD are shown in the left panel of Figure 3.2. The adjusted k_t's result in the expected number of deaths matching the observed number in each year. The adjusted k_t values are shown in the right panel of the same figure.

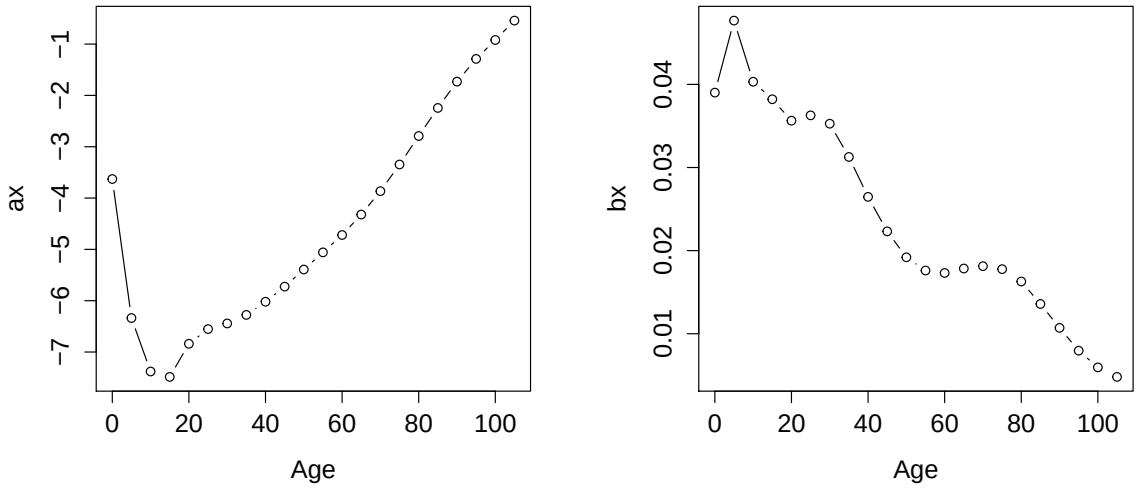

Figure 3.1: France, female, 1900-2012: age-specific constants a_x (left) and b_x (right).

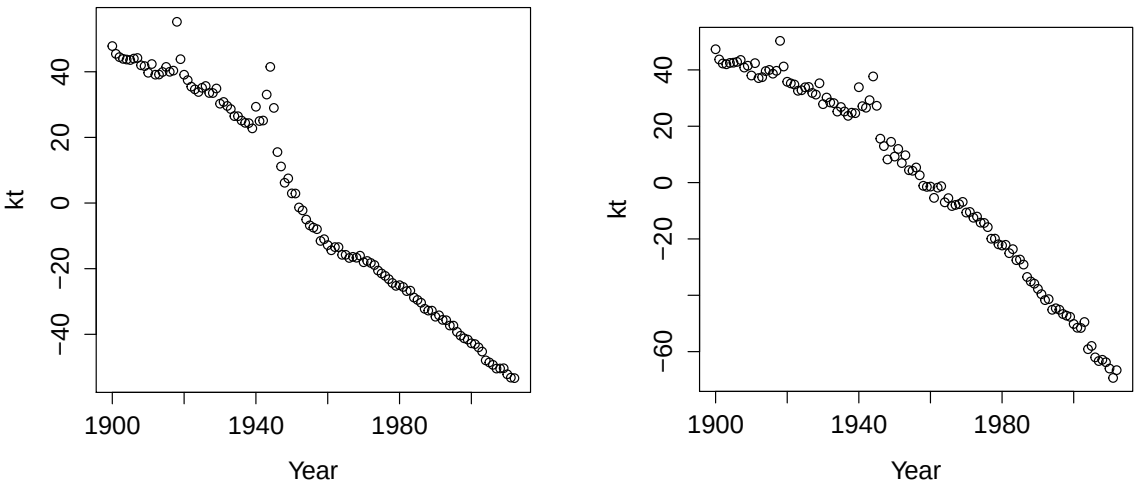

Figure 3.2: France, female, 1900-2012: time-varying index k_t. The k_t values obtained by SVD are on the left, and the adjusted k_t values are on the right.

We have fitted the Lee-Carter model and obtained age-specific constants a_x and b_x. Now we can move on to forecast the mortality index based on our reestimated k_t. We fit k_t to a time series, so the first step is to find an appropriate ARIMA time series model. We choose to use an ARIMA(0,1,0) model for k_t, which is a random walk with drift. From Figure 3.2, we can observe extreme values of k_t in year 1918 and around year 1944. We can have some clues about what caused these extreme values from the documentation provided by HMD. In the documentation, 'Specific Episodes in French Demographic History' have been listed, which includes two world wars, as well as the Spanish flu epidemic during 1918-1919. Therefore, we can explain the extreme value for 1918 by the Spanish flu epidemic and the extreme values around 1944 by World War II during 1939-1945. The next question is how to treat the Spanish influenza epidemic of 1918-1919 and World War II of 1939-1945. Actually the influenza epidemic and World War II are very rare events, so including them in the series might be inappropriate and influence our result. So we deal with the influence of the Spanish influenza epidemic and the World War II by introducing two dummy variables. The model, estimated over 1900-2012, with standard errors in parentheses, is as follows:

$$\begin{aligned} k_t = k_{t-1} &- 1.02 + 9.87flu + 9.42war + \varepsilon_t, \\ &(0.27)\ (2.04) \qquad (2.04) \\ &\varepsilon_t \sim N(0, 8.36), \end{aligned} \tag{3.12}$$

where flu is a dummy variable, which takes a value of 1 in 1918 and 0 elsewhere, and war is a dummy variable, which takes a value of 1 in 1944 and 0 elsewhere. The coefficients of flu and war indicate that the mortality index k_t was 9.87 higher than expected in 1918 and 9.42 higher than expected in 1944. We can see that k_t is drifting downward at an average rate of -1.02 per year. If we do not consider the error terms, the forecasted k_t should be on a downward line with a slope of -1.02. The fitted and the predicted values of k_t appear in Figure 3.3.

Now we obtain the age-specific constants a_x and b_x and the fitted and predicted mortality index k_t. We can use the obtained values to construct the fitted and predicted mortality surface by applying the Lee-Carter model:

$$ln(m_{x,t}) = a_x + b_x k_t. \tag{3.13}$$

As a comparison, Figure 3.4 shows the raw mortality surface (in the logarithm scale), where the height of the surface is $ln(\mathbf{m})$, that is $ln(\mathbf{D}/\mathbf{E})$. Figure 3.5 shows the fitted log mortality by the Lee-Carter model. The mortality surface of the raw data and the predicted mortality surface by the Lee-Carter model is in Figure 3.6.

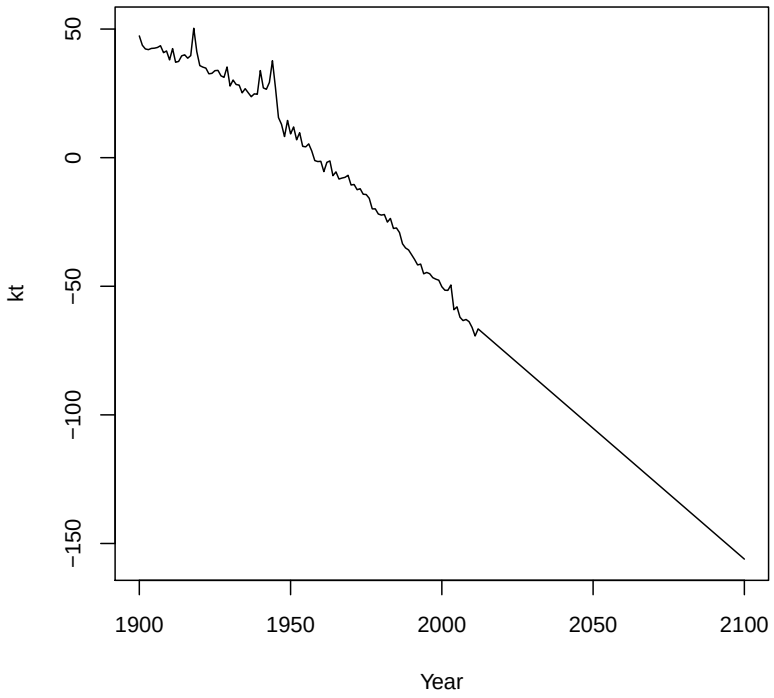

Figure 3.3: France, female: fitted (1900-2012) and forecasted (2013-2100) time-varying index k_t values.

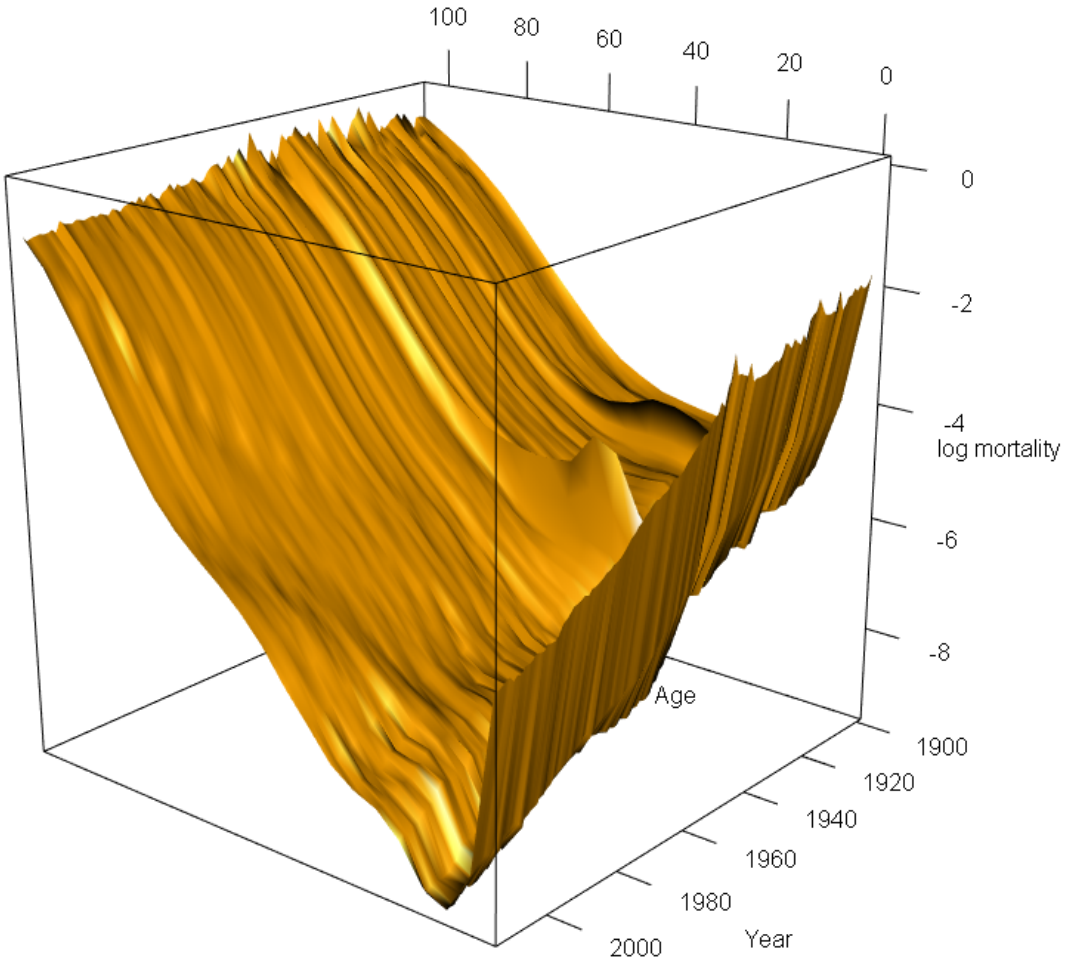

Figure 3.4: France, female, 1900-2012: raw log mortality surface.

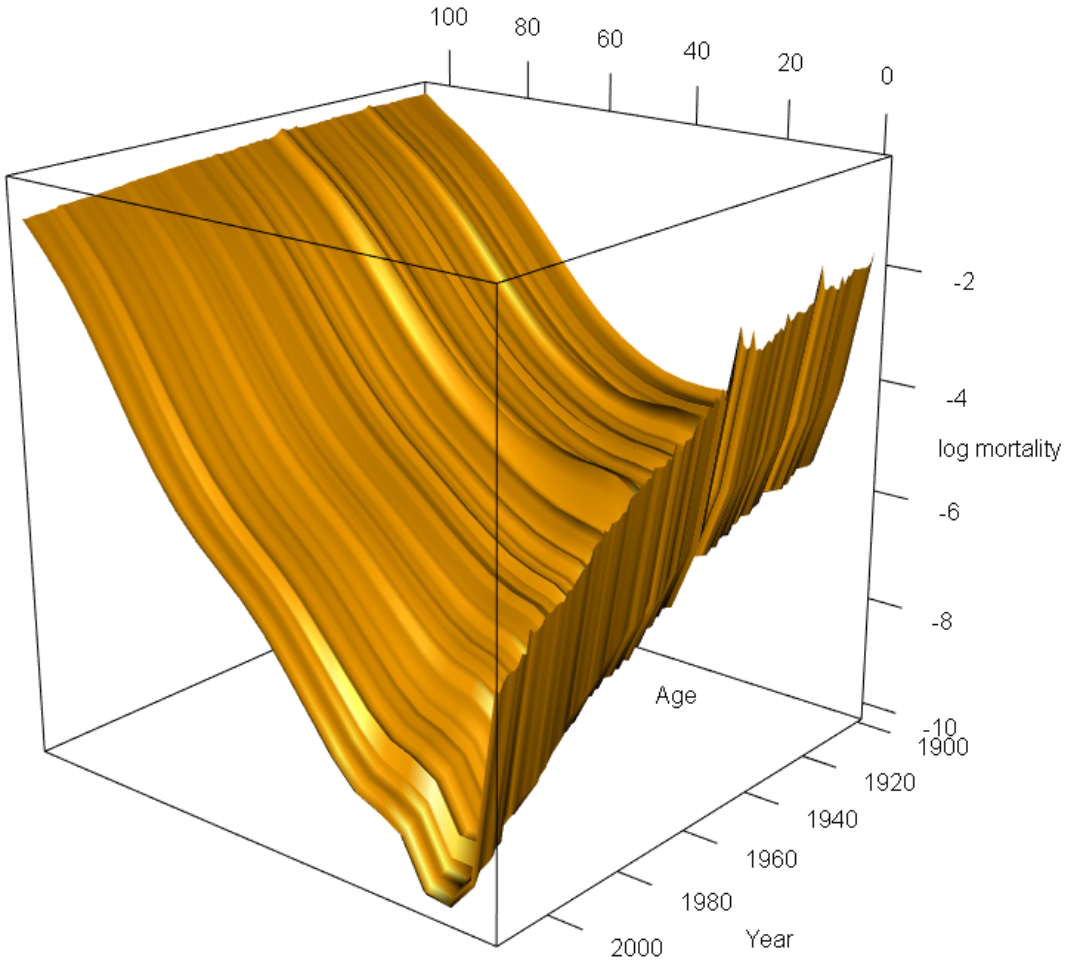

Figure 3.5: France, female, 1900-2012: Lee-Carter fitted log mortality surface.

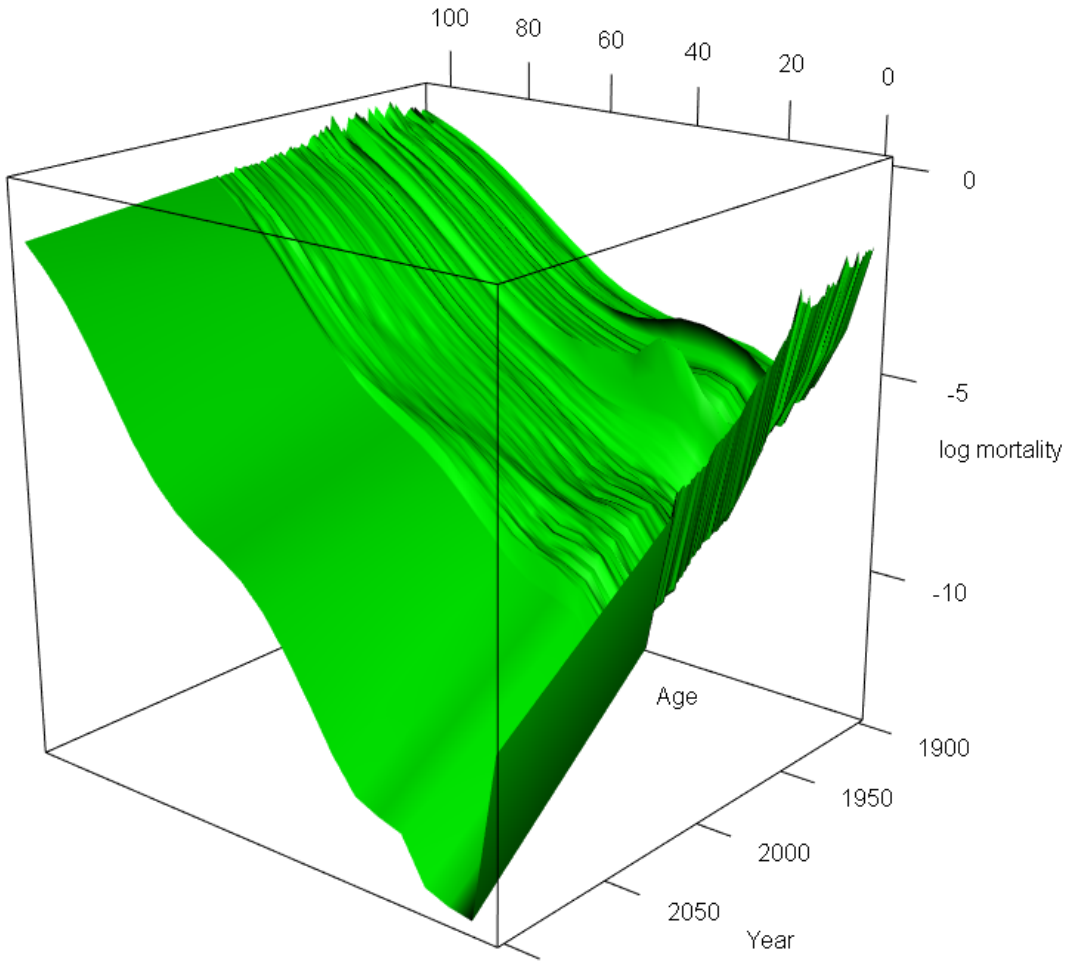

Figure 3.6: France, female: raw mortality surface (1900-2012) and the Lee-Carter predicted mortality surface (2013-2100).

Chapter 4

Single-Decrement CoDa Equivalent Lee-Carter Model

The Lee-Carter model fits and predicts the central death rates. Following the idea of Jim Oeppen's report (Oeppen, 2008), our interest, however, is to model the probability density function of the lifetime random variable, which is also called the density of death. In his report, Oeppen (2008) proposed the idea of modeling the density of death instead of the conventional approach of using log mortality. He mentioned that since the density of death obeys a unit sum constraint (which means the density of death is compositional), the method of compositional data analysis can be used to transform the density of death into the real space where the full range of multivariate statistics can be applied (Oeppen, 2008). He also expressed the structure of Lee-Carter model in the compositional form.

In this chapter, our main purpose is to define the CoDa equivalent Lee-Carter model. We will refer to the CoDa equivalent Lee-Carter model as "CoDa LC model". We will first introduce some basic formulae of the density of death. Next, based on compositional data analysis, we will focus on how to transform the compositional density of death into real space. Then, we will apply the Lee-Carter model on the transformed density of death. Finally, in order to explore the potential of the CoDa LC model, we will present results for the France density of death for female.

4.1 Density of Death

4.1.1 Basic Formula

In demography, ${}_nd_x$ usually denotes the number of deaths between age x and age $x+n$ (Preston et al., 2001). However, in our project, we use ${}_nd_x$ to denote the probability that a newborn dies

between age x and age $x+n$. And we use ${}_nD_x$ to denote the number of deaths between age x and age $x+n$. Let p_x, represents the probability that a life aged x, denoted by (x), survives to at least age $x+1$ and ${}_nq_x$ represents the probability that (x) dies before age $x+n$. Then we can obtain that,

$$ {}_nd_x = {}_xp_0 \cdot {}_nq_x = p_0p_1 \cdots p_{x-1} \cdot {}_nq_x. \tag{4.1}$$

If we add another subscript, year t, to each element of equation (4.1), we get

$$ {}_nd_{x,t} = {}_xp_{0,t} \cdot {}_nq_{x,t}, \tag{4.2}$$

where

$$ {}_xp_{0,t} = p_{0,t} \cdot p_{1,t} \cdots p_{x-1,t}, \tag{4.3}$$

$$ {}_nq_{x,t} = q_{x,t} + p_{x,t} \cdot q_{x+1,t} + \cdots + p_{x,t} \cdots p_{x+n-1,t} \cdot q_{x+n-1,t}. \tag{4.4}$$

Here $p_{x,t}$ denotes the probability that a life aged x in year t survives to at least age $x+1$ and $q_{x,t}$ denotes the probability that a life aged x in year t dies before age $x+1$. Notice that the probabilities on the right hand side of equations (4.3) and (4.4) are all for year t. Generally, one might want to use "cohort" probabilities instead of calendar year ones. Since cohort probabilities require data over a period much longer than what is available here, calendar year data is used as a proxy when applying the CoDa Lee-Carter model in this project.

If ages are grouped into intervals $[x_0, x_1)$, $[x_1, x_2)$, ..., $[x_{n-1}, x_n)$, where x_0 is 0 and $x_n = \omega$ is the limiting age, then we have

$$ {}_{x_{j+1}-x_j}d_{x_j,t} = {}_{x_j}p_{x_0,t} \cdot {}_{x_{j+1}-x_j}q_{x_j,t}, \tag{4.5}$$

where

$$ {}_{x_j}p_{x_0,t} = {}_{x_1}p_{x_0,t} \cdot {}_{x_2-x_1}p_{x_1,t} \cdots {}_{x_j-x_{j-1}}p_{x_1,t}. \tag{4.6}$$

Following the above definition, we have

$$ \sum_{j=0}^{n-1} {}_{x_{j+1}-x_j}d_{x_j,t} = 1. \tag{4.7}$$

The future lifetime of a life aged 0 in year t can be modeled by a continuous random variable, which is denoted by $T_{0,t}$. The ${}_{x_{j+1}-x_j}d_{x_j,t}$'s represent a discretization of the probability density function of $T_{0,t}$. The probabilities of ${}_{x_{j+1}-x_j}d_{x_j,t}$'s will be called the "density of death". According to (4.7), the density of death can be treated as a composition and therefore we can use the method of compositional data analysis to transform it into the real space.

Suppose we have the values of ${}_{x_{j+1}-x_j}q_{x_j,t}$ for all age intervals and years of interest, since ${}_{x_{j+1}-x_j}p_{x_j,t} = 1 - {}_{x_{j+1}-x_j}q_{x_j,t}$, we are able to obtain the values of ${}_{x_{j+1}-x_j}p_{x_j,t}$ for all age intervals and years of interest. Then according to formulae (4.5) and (4.6), we can obtain the values of ${}_{x_{j+1}-x_j}d_{x_j,t}$.

4.1.2 Centred Log Ratio of the Density of Death

In Section 4.1.1, we already derived formulae (4.5) and (4.6) to calculate the density of death. If the age intervals are $[x_0, x_1)$, $[x_1, x_2)$, ..., $[x_{n-1}, x_n)$ and the years of interest are t_1,..., t_N, then similar to the construction of matrix of $ln(\mathbf{m})$, we can construct an $n \times N$ matrix of ${}_{x_{j+1}-x_j}d_{x_j,t}$ with ages in rows and years in columns. In compositional data analysis, it is customary to make every row a composition, so we transpose the density of death matrix. As a result, ages are in columns and years are in rows and according to formula (4.7), each row adds up to one, which is a composition. We use **d** to represent the transposed matrix of density of death and specifically:

$$\mathbf{d} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nn} \end{bmatrix}, \tag{4.8}$$

where the value of $d_{i,j}$ equals ${}_{x_j-x_{j-1}}d_{x_{j-1},t_i}$. In the Lee-Carter model, we apply SVD to the matrix $\tilde{\mathbf{m}}$, which is obtained by subtracting each row's arithmetic average of $ln(\mathbf{m})$ from that row (see formula (3.5)). When using the CoDa LC Model, we will perform a similar operation to matrix **d**, which is called centring in Section 2.2. We will use inverse perturbation (compositional subtraction) defined in (2.8) to subtract the "centre" of **d** from each row of **d**. To be specific, first we find the "centre", which is denoted as g, a $1 \times N$ vector, and equal to:

$$g = C(g_1, g_2, ..., g_n), \tag{4.9}$$

where

$$g_j = (d_{1j}d_{2j} \cdots d_{Nj})^{1/N}, \ for\ j = 1, 2, ..., n. \tag{4.10}$$

Then "centring" means to subtract g from each row of **d**. The row vector, denoted as cen_i, is

$$cen_i = d_i \ominus g = C\left(\frac{d_{i1}}{g_1}, ..., \frac{d_{in}}{g_n}\right), \tag{4.11}$$

where $d_i = (d_{i1}, d_{i2}, ..., d_{in})$ denotes the i^{th} row of **d**.

Let $cen(\mathbf{d})$ be the centred death density matrix, which can be expressed as:

$$cen(\mathbf{d}) = \begin{bmatrix} cen_1 \\ cen_2 \\ \vdots \\ cen_N \end{bmatrix} = \begin{bmatrix} d_1 \ominus g \\ d_2 \ominus g \\ \vdots \\ d_N \ominus g \end{bmatrix}. \tag{4.12}$$

We know that each row of the matrix $cen(\mathbf{d})$ is a composition. In order to transform the compositions into the real space, we use the CLR operator described in Section 2.2 on each row of matrix

$cen\,(\mathbf{d})$. So we have the following formula,

$$CLR\,(cen\,(\mathbf{d})) = \begin{bmatrix} CLR\,(cen_1) \\ CLR\,(cen_2) \\ \vdots \\ CLR\,(cen_N) \end{bmatrix} = \begin{bmatrix} ln\left(\frac{cen_1}{g(cen_1)}\right) \\ ln\left(\frac{cen_2}{g(cen_2)}\right) \\ \vdots \\ ln\left(\frac{cen_N}{g(cen_1)}\right) \end{bmatrix}, \tag{4.13}$$

where $g\,(cen_i)$ is the geometric mean of row vector cen_i. Now we can apply SVD to the matrix $CLR\,(cen\,(\mathbf{d}))$.

4.2 The Model

The SVD of the matrix $CLR\,(cen\,(\mathbf{d}))$ is

$$CLR\,(cen\,(\mathbf{d})) = U_{N\times N} S_{N\times n} V'_{n\times n}. \tag{4.14}$$

Let u_1,..., u_r be the first r left singular vectors, s_1,..., s_r be the first r singular values, and v_1,..., v_r be the first r right singular vectors; then we have the rank-r approximation of the matrix $CLR\,(cen\,(\mathbf{d}))$:

$$\hat{CLR}^r\,(cen\,(\mathbf{d})) = [u_1, ..., u_r] \begin{bmatrix} s_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s_r \end{bmatrix} \begin{bmatrix} v_1' \\ \vdots \\ v_r' \end{bmatrix}. \tag{4.15}$$

Rank-r approximation means that we choose the first r left singular vectors, the first r singular values and the first r right singular vectors to construct the matrix of approximation. We call the left singular vectors period factors and the right singular vectors age factors. We choose the value of r according to the specific datasets. For some datasets, rank-1 approximation is adequate, while for other datasets, rank-2 or even higher ranks might be necessary.

It is worth mentioning that in the Lee-Carter model, a second step (see equation (3.9)) is taken to reestimate the time-varying index. Since we are modeling the distribution of lifetimes (density of death), there seems to be no obvious reason to adjust the values of the period factors to match the average life expectancy or deaths. We do not need to scale the period factors since they automatically sum to zero.

4.3 France Projection

Human Mortality Database at http://www.mortality.org/ provides life tables by country and sex. We use 1×1 life table for France females to explore the CoDa LC model. The France female 1×1 life

table provides data for ages ranging from 0 to 110+, and for years ranging from 1816 to 2012. From the life table, the estimated values for $q_{x,t}$'s are provided; therefore according to Section 4.1.1, based on (4.5) and (4.6), we are able to construct the matrix **d**. Then based on Section 4.1.2, we can obtain the matrix $CLR\,(cen\,(\mathbf{d}))$ to apply SVD. In order to fit the CoDa LC model, we will restrict the data (values of $q_{x,t}$) to ages from 0 to 105+ and years from 1955 to 2005. Rank-2 approximation is chosen and the resulting first age factor is shown in Figure 4.1.

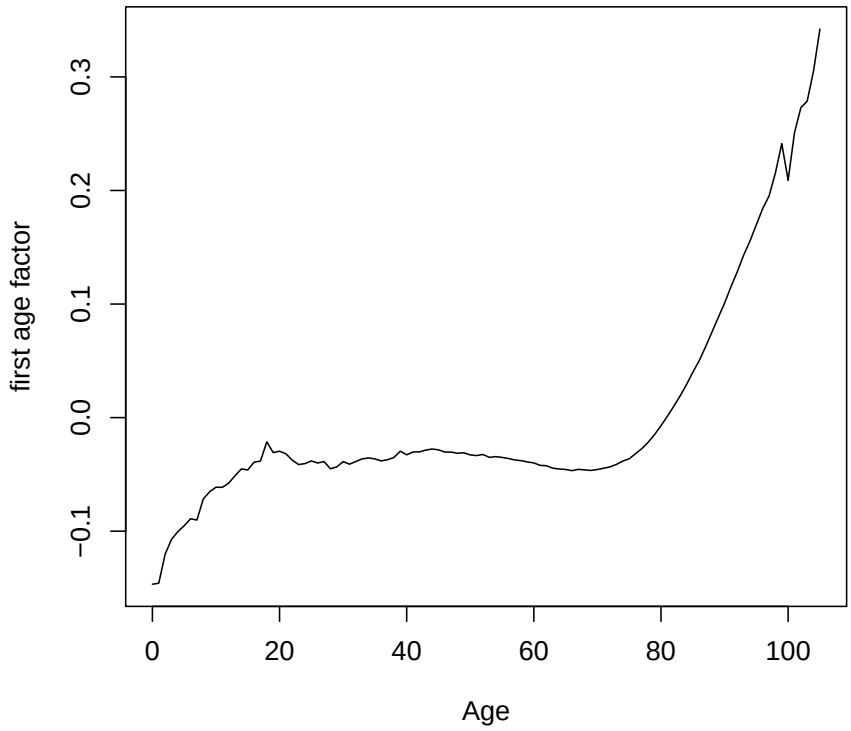

Figure 4.1: France, female, 1955-2005: the first age factor.

We use an ARIMA model to fit the period factor over the period 1955 to 2005 and then predict the period factor for 2006 to 2100. Considering various orders of the ARIMA models, we choose to use an ARIMA(0,1,0) model, which is chosen by AICc (Shumway and Stoffer, 2000). AIC is defined as $AIC = ln\hat{\sigma}_k^2 + \frac{2k}{n}$ with n representing the number of data, k the number of parameters in the fitted model and where $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$. AICc is the corrected form of AIC, which is defined as $AICc = ln\hat{\sigma}_k^2 + \frac{n+k}{n-k-2}$. We use AICc to choose the best ARIMA model, that is the fitted model with

the minimum AICc is selected as the best model. Figure 4.2 shows the fitted and predicted values of the first period factor.

From Figure 4.2, we can see that for fitting years 1955 to 2005, the first period factor has an

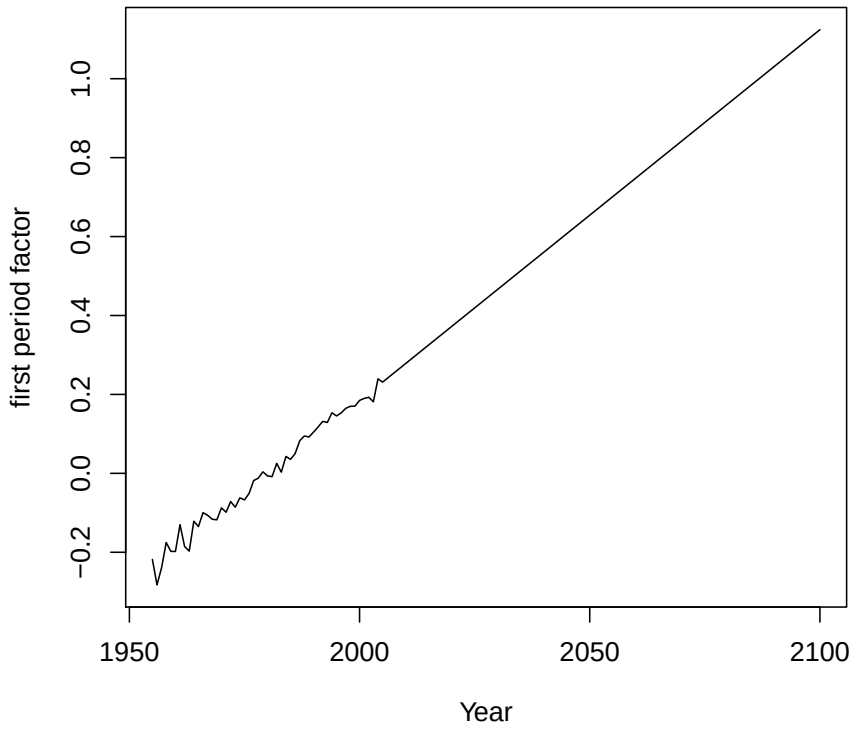

Figure 4.2: France, female: the fitted (1955-2005) and predicted (2006-2100) first period factor.

obvious positive linear trend. By fitting the first age factor from 1955 to 2005 to an ARIMA model, we arrive at the following:

$$\begin{aligned} u_{1,t} &= u_{1,t-1} + 0.0094 + \varepsilon_t, \\ \varepsilon_t &\sim N(0, 0.0006), \end{aligned} \tag{4.16}$$

where $u_{1,t}$ is the element of first left singular vector (first age factor) that corresponds to year t.

In Figure 4.3, we select three years, 1955, 1979 and 2005, to show the centred log ratio data and rank-2 estimates. We can see the estimates are very close to the data.

To end this section, we briefly summarize the CoDa LC model into the following steps:

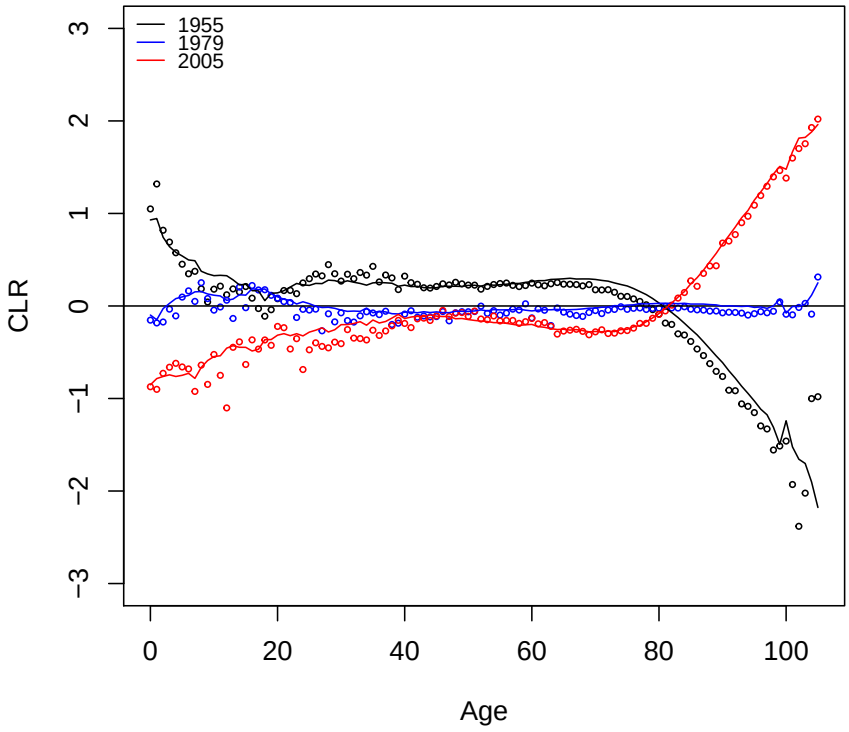

Figure 4.3: France, female: centred log ratio of the centred density of death. Points are data and lines are rank-2 estimates.

1. Construct matrix **d** with time in rows and age in columns. Each row is a composition.
2. Obtain matrix $cen\,(\mathbf{d})$ by centring matrix **d**: calculate the vector of age-specific geometric means (geometric means of columns of **d**) and substract it from each row of the matrix using inverse perturbation operator.
3. Obtain matrix $CLR\,(cen\,(\mathbf{d}))$ by performing CLR operator on each row of $cen\,(\mathbf{d})$ to tranform it into the real space.
4. Apply SVD to $CLR\,(cen\,(\mathbf{d}))$ to obtain age and period factors.
5. Construct the selected low rank approximation to $CLR\,(cen\,(\mathbf{d}))$.

Chapter 5 will extend this single-decrement CoDa LC model to multiple decrements. We will introduce how to obtain the values of density of death from the values of centred log-ratio density of death. We will also introduce how to obtain the values of $q_{x,t}$ from the values of $d_{x,t}$.

Chapter 5

Multiple-Decrement CoDa Equivalent Lee-Carter Model

The CoDa LC model is very easy to extend to the multiple-decrement case. We will follow Jim Oeppen's work (Oeppen, 2008) and explain the model using Japan female multiple-decrement (different causes of death) data.

5.1 Multiple-Decrement Density of Death

Assume there are I causes of death. In Section 4.1.1, we already defined p_x and $p_{x,t}$. Let ${}_nd_x^i$ represent the probability that (0) dies between ages x and $x+n$ from cause i, and ${}_nq_x^i$ denote the probability that (x) dies within the next n years due to death cause i. Similar to formula (4.1), we have, for $i = 1, 2, ..., I$,

$$ {}_nd_x^i = {}_xp_0 \cdot {}_nq_x^i = p_0 p_1 \cdots p_{x-1} \cdot {}_nq_x^i. \tag{5.1}$$

Similar to formula (4.2), if we add another subscript, year t, then we have

$$ {}_nd_{x,t}^i = {}_xp_{0,t} \cdot {}_nq_{x,t}^i. \tag{5.2}$$

From Section 4.1.1, we know that ${}_nD_{x,t}$ denote the number of deaths between ages x and $x+n$ in year t. Now let ${}_nD_{x,t}^i$ denote the number of deaths between ages x and $x+n$ in year t for death cause i. Then intuitively, we can express ${}_nq_{x,t}^i$ as follows:

$$ {}_nq_{x,t}^i = {}_nq_{x,t} \cdot \frac{{}_nD_{x,t}^i}{{}_nD_{x,t}}. \tag{5.3}$$

Therefore the cause-specific density of death can be expressed as:

$$ {}_nd_{x,t}^i = {}_xp_{0,t} \cdot {}_nq_{x,t} \cdot \frac{{}_nD_{x,t}^i}{{}_nD_{x,t}} \tag{5.4}$$

Then if ages are grouped into intervals $[x_0, x_1)$, $[x_1, x_2)$, ... and $[x_{n-1}, x_n)$, where x_0 is 0 and $x_n = \omega$ is the limiting age, then the cause-specific density of death has the following property:

$$\sum_{i=1}^{I} \sum_{j=0}^{n-1} {}_{x_{j+1}-x_j} d^i_{x_j,t} = 1. \tag{5.5}$$

We can use formula (5.4) to construct $\mathbf{d}_{causei}$, the matrix of density of death for cause i, with years in rows and ages in columns. Then we can combine the matrices of density of death for each cause so that we get an $N \times nI$ matrix $\mathbf{d}_{mul}$ which is:

$$\mathbf{d}_{mul} = (\mathbf{d}_{cause1}, \mathbf{d}_{cause2}, ..., \mathbf{d}_{causeI}) . \tag{5.6}$$

The combination ensures that each row of $\mathbf{d}_{mul}$ adds up to 1 and therefore can be treated as a composition.

5.2 Centred Log Ratio of the Density of Death-Multiple Decrements

In Section 4.1.2, we already described how to obtain the centred log ratio of the matrix of single-decrement density of death. Now for $\mathbf{d}_{mul}$ in (5.6), we can repeat exactly what we did to $\mathbf{d}$ in Section 4.1.2 and obtain $CLR\left(cen\left(\mathbf{d}_{mul}\right)\right)$ for the multiple-decrement case.

5.3 The Model

The SVD of the matrix $CLR\left(cen\left(\mathbf{d}_{mul}\right)\right)$ is

$$CLR\left(cen\left(\mathbf{d}_{mul}\right)\right) = U_{N\times N} S_{N\times nI} V'_{nI\times nI} \tag{5.7}$$

where N is the number of years, n is the number of age intervals and I is the number of death causes.

Let u_{mul1},..., u_{mulr} be the first r left singular vectors, s_{mul1},..., s_{mulr} be the first r singular values, and v_{mul1},..., v_{mulr} be the first r right singular vectors; then we have the rank-r approximation of the matrix $CLR\left(cen\left(\mathbf{d}_{mul}\right)\right)$ as:

$$\hat{CLR}^r\left(cen\left(\mathbf{d}_{mul}\right)\right) = [u_{mul1}, ..., u_{mulr}] \begin{bmatrix} s_{mul1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s_{mulr} \end{bmatrix} \begin{bmatrix} v'_{mul1} \\ \vdots \\ v'_{mulr} \end{bmatrix}. \tag{5.8}$$

It is worth mentioning that each right singular vectors includes nI elements, the first n elements are the age factor for death cause 1, the $n+1, ..., 2n$ elements are the age factor for death cause

2,..., and the last n elements, that is the $(I-1)n+1, ..., nI$ elements are the age factor for death cause I. Therefore, for each cause of death, we are able to obtain r age factors of dimension $1 \times n$.

Now we already derived the rank-r approximation for $CLR\left(cen\left(\mathbf{d}_{mul}\right)\right)$, based on which we are able to derive the rank-r approximation for $\mathbf{d}_{mul}$. Referring to (2.13), we can express the rank-r approximation of $\mathbf{d}_{muli}$ as:

$$\hat{\mathbf{d}}^{r}_{muli} = g_{mul} \oplus \left(u_{muli1} s_{mul1} \otimes \beta_{mul1}\right) \oplus ... \oplus \left(u_{mulir} s_{mulr} \otimes \beta_{mulr}\right), \tag{5.9}$$

where

$$\beta_{muli} = C\left(exp\left(v_{i1}\right), exp\left(v_{i2}\right), ..., exp\left(v_{i,nI}\right)\right), \tag{5.10}$$

and g_{mul} is a $1 \times N$ vector with the i^{th} element being the compositional scaled geometric mean of the i^{th} column of matrix $\mathbf{d}_{mul}$, u_{ij} represents the element at the i^{th} row and j^{th} column of matrix U in (5.7), and v_{ij} represents the element at the i^{th} row and j^{th} column of matrix V in (5.7).

In order to indicate whether the fit is good or not, let us first introduce $l^{i}_{x_j}$, which represents the probability that (x_j) will eventually die from cause i:

$$l^{i}_{x_j} = \sum_{k=j}^{n-1} {}_{x_{k+1}-x_k} d^{i}_{x_k}. \tag{5.11}$$

Adding another subscript, year t, to each element of equation (5.11), we get:

$$l^{i}_{x_j,t} = \sum_{k=j}^{n-1} {}_{x_{k+1}-x_k} d^{i}_{x_k,t}. \tag{5.12}$$

Therefore, let j be 0, we will have:

$$l^{i}_{0,t} = \sum_{k=0}^{n-1} {}_{x_{k+1}-x_k} d^{i}_{x_k,t}. \tag{5.13}$$

Then for cause i, we are able to plot $l^{i}_{0,t}$ against year t. Based on the constructed matrix of density of death, we are able to calculate $l^{i}_{0,t}$. Based on the fitted value of density of death, we are able to calculate the estimated $\hat{l}^{i}_{0,t}$. We can plot the $l^{i}_{0,t}$ and $\hat{l}^{i}_{0,t}$ against year t in one figure. If the fitted model is good, we expect the two curves (or dots) to be fairly close.

5.4 Cause-specific Death Rates $q^{i}_{x,t}$

For someone aged x in year t, the probability of dying in year t is $q_{x,t}$. $q_{x,t}$ is used to calculate the expectations and variances of life insurances and annuities. For multiple decrement life table, we are able to solve for the values of $q_{x,t}$ and $q^{i}_{x,t}$ from the density of death. We know that ${}_{x_{j+1}-x_j} d_{x_j,t}$ is equal to

$${}_{x_{j+1}-x_j} d_{x_j,t} = \sum_{i=1}^{I} {}_{x_{j+1}-x_j} d^{i}_{x_j,t}. \tag{5.14}$$

Let j equal to 0, 1,..., n; then we have the following $n-1$ equations based on (4.5) (recall that x_0 is 0):

$$\begin{aligned}
{}_{x_1}d_{0,t} &= {}_0p_{0,t} \cdot {}_{x_1}q_{0,t},\\
{}_{x_2-x_1}d_{x_1,t} &= {}_{x_1}p_{0,t} \cdot {}_{x_2-x_1}q_{x_1,t},\\
{}_{x_3-x_2}d_{x_2,t} &= {}_{x_2}p_{0,t} \cdot {}_{x_3-x_2}q_{x_2,t},\\
&\vdots\\
{}_{x_n-x_{n-1}}d_{x_{n-1},t} &= {}_{x_{n-1}}p_{0,t} \cdot {}_{x_n-x_{n-1}}q_{x_{n-1},t}.
\end{aligned} \tag{5.15}$$

Since ${}_0p_{0,t}$ is 1, from the first equation of (5.15), we can obtain the value of ${}_{x_1}q_{0,t}$, which is equal to ${}_{x_1}d_{0,t}$. Then ${}_{x_1}p_{0,t}$ can be calculated by $1-{}_{x_1}q_{0,t}$. From the second equation of (5.15), we know that ${}_{x_2-x_1}q_{x_1,t} = \frac{{}_{x_2-x_1}d_{x_1,t}}{{}_{x_1}p_{0,t}}$ and ${}_{x_2-x_1}p_{x_1,t}$ is therefore $1-{}_{x_2-x_1}q_{x_1,t}$. Since the value of ${}_{x_2}p_{0,t}$ can be obtained by ${}_{x_2}p_{0,t} = {}_{x_1}p_{0,t} \cdot {}_{x_2-x_1}p_{x_1,t}$, then according to the third equation of (5.15) ${}_{x_3-x_2}q_{x_2,t}$ can be calculated. Therefore the 1^{st} to j^{th} equations of (5.15) will give us the values of ${}_{x_1-x_0}q_{x_0,t}$,..., ${}_{x_j-x_{j-1}}q_{x_{j-1},t}$. We are then able to calculate the cause-specific death rates ${}_{x_{j+1}-x_j}q^i_{x_j,t}$ using

$${}_{x_{j+1}-x_j}d^i_{x_j,t} = {}_{x_j}p_{0,t} \cdot {}_{x_{j+1}-x_j}q^i_{x_j,t}. \tag{5.16}$$

We can apply cause-specific death rates to price some products like accidental death rider, critical illness rider, etc.

5.5 Japan Projection

When constructing a single-decrement matrix **d** we only need the values of q_x, but for the construction of multiple-decrement matrix $\mathbf{d}_{mul}$ (refer to (5.3) and (5.6)) we also need to know the values of $D^i_{x,t}$. The values of $D_{x,t}$ is easily obtained by:

$$D_{x,t} = \sum_{i=1}^{I} D^i_{x,t}. \tag{5.17}$$

In Jim Oeppen's report (Oeppen, 2008), he uses Japan female data to illustrate the multiple-decrement CoDa LC model. The construction of $\mathbf{d}_{mul}$ requires values of q_x and $D^i_{x,t}$. Human Mortality Database provides the life tables for Japan female, and we will use the 5×1 life table. The Japan female 5×1 life table provides data for age groups 0, 1-4, 5-9,..., 105-109 and 110+, and for years from 1947 to 2012. The Berkeley Mortality Database (BMD) provides a table called "Deaths-Causes of death, 1951-1990, (5×1) ". The deaths are grouped into 40 causes and the table provides data for ages 0, 1-4, 5-9,..., 95-99 and 100+, and for years 1951 to 1990. The BMD also provides a documentation named "Data Notes". The 40 causes of death are listed in "Data Notes".

In his report (Oeppen, 2008), Jim Oeppen categorized the 40 causes into 6 groups. Causes 2 to 5, 8 and 9 are categorized as cause “Infectious Disease”; Causes 11 to 21 are categorized as cause “Malignant Neoplasm”; Causes 23 to 25 are categorized as cause “Heart Disease”; Causes 27 to 29 are categorized as cause “Cerebrovascular Disease”; Causes 6, 7, 30 and 31 are categorized as cause “Respiratory Disease”; Causes 1, 10, 22, 26, 32 to 40 are categorized as cause “Miscellaneous Death Cause”. We choose to use age intervals 0, 1-4, 5-9,..., 90-94, 95+. The period we choose to fit the model is 40 years, from 1951 to 1990. We fit the data to multiple-decrement CoDa LC model, the first age factors for 6 different causes are shown in Figure 5.1.

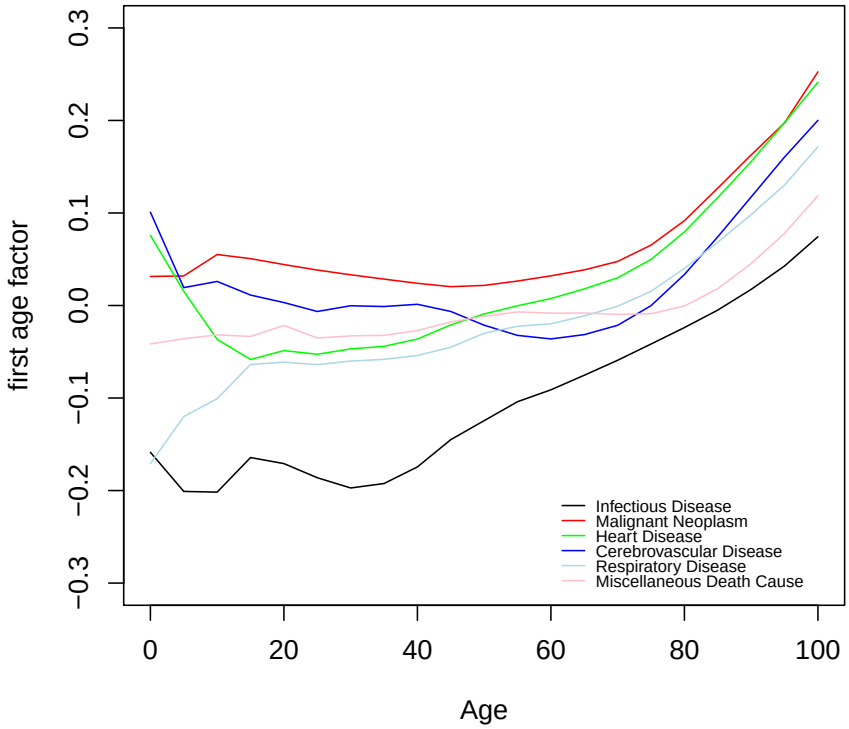

Figure 5.1: Japan, female, 1951-1990: the first cause-specific age factors.

For the period factor, again, we use an ARIMA model to fit it for the period 1951 to 1990 and then predict it for 50 years from 1991 to 2040. Jim Oeppen mentioned in his report (Oeppen, 2008) that he used the AICc criterion (Shumway and Stoffer, 2000) to choose the most appropriate ARIMA model. For both the first and second period factors, ARIMA(0,2,2) is the choice. He also indicated

that a random walk with drift is inadequate. We first fit the first and second period factors to random walks, that is ARIMA(0,1,0), and predict them for years 1991 to 2040. The fitted and predicted first and second period factors are shown in Figure 5.2.

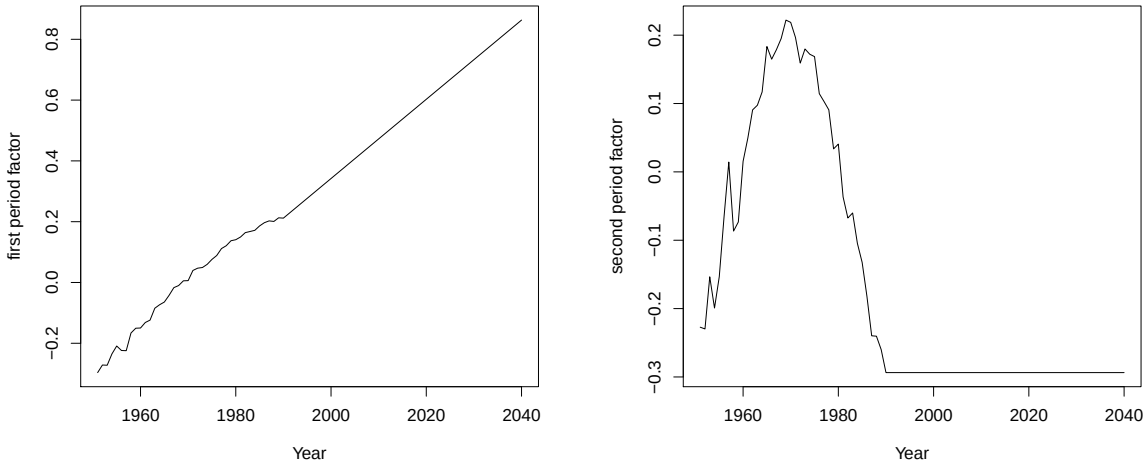

Figure 5.2: Japan, female: the fitted (1951-1990) and predicted (1991-2040) first (left panel) and second (right panel) period factors. ARIMA(0,1,0) is used to fit and predict both the first and second period factors.

Then we fit the first and second period factors to ARIMA(0,2,2) and then predict them for years 1991 to 2040. The fitted and predicted first and second period factors are shown in Figure 5.3.

We also plot the third period factor in Figure 5.4 which shows that the third period factor is basically a random noise. If we fit the 3rd period factor to an ARIMA model, selected by AICc, we will obtain the optimal time series model ARIMA(0,0,0). This means that a rank-3 approximation is unnecessary and consequently a rank-2 approximation is enough. We will also illustrate in the later part why we do not use a rank-1 approximation.

Now according to (5.13), for death cause i, first, we calculate $l^i_{0,t}$ and plot the values of $l^i_{0,t}$ against year as dots. Second, we calculate $\hat{l}^i_{0,t}$ and plot the values of $\hat{l}^i_{0,t}$ against year as curves. We use an ARIMA(0,2,2) model to fit and predict the first and second period factors. Then the plots for rank-1 and rank-2 approximations are shown in Figures 5.5 and 5.6 respectively.

Figure 5.5 shows that rank-1 approximation is inadequate and cannot capture the curvature for each death cause, especially for cause “Cerebrovascular Disease”. So for Japan female data we choose r to be 2. Next, we will illustrate why using random walks for fitting and predicting period factors is also not reasonable. We use an ARIMA(0,1,0) model to fit and predict the first and second

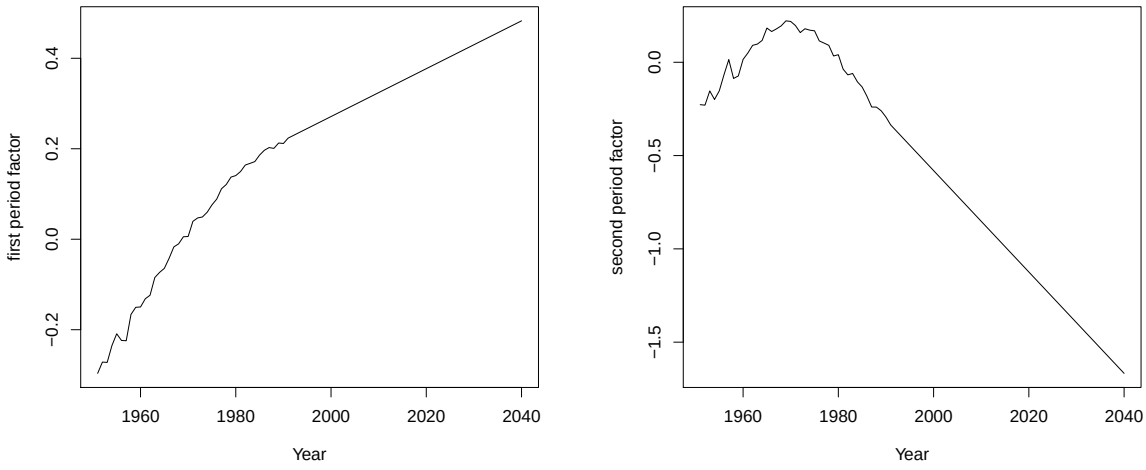

Figure 5.3: Japan, female: the fitted (1951-1990) and predicted (1991-2040) first (left panel) and second (right panel) period factors. ARIMA(0,2,2) is used to fit and predict both the first and second period factors.

period factors. Then the rank-2 approximations based on ARIMA(0,1,0) are shown in Figure 5.7.

From Figure 5.7, we can see that the plots based on an ARIMA(0,1,0) have more sudden angles at the junction parts of the data and the predictions, which means that the ARIMA(0,1,0)-based predictions are more arbitrary and cannot provide us with reliable predictions of the data. To sum up, the rank-2 approximation with first and second period factors fitted to ARIMA(0,2,2) models provides us with fairly reasonable predictions of the data.

Figure 5.4: Japan, female, 1951-1990: the third period factor.

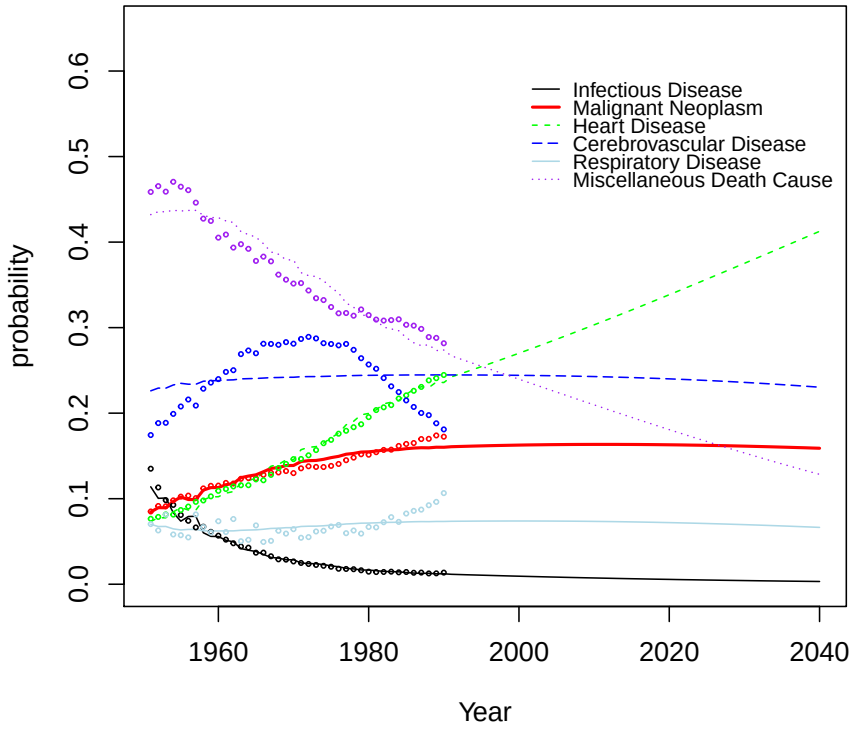

Figure 5.5: Japan, female: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-1 approximation is used. ARIMA(0,2,2) is used to fit and predict the first period factor.

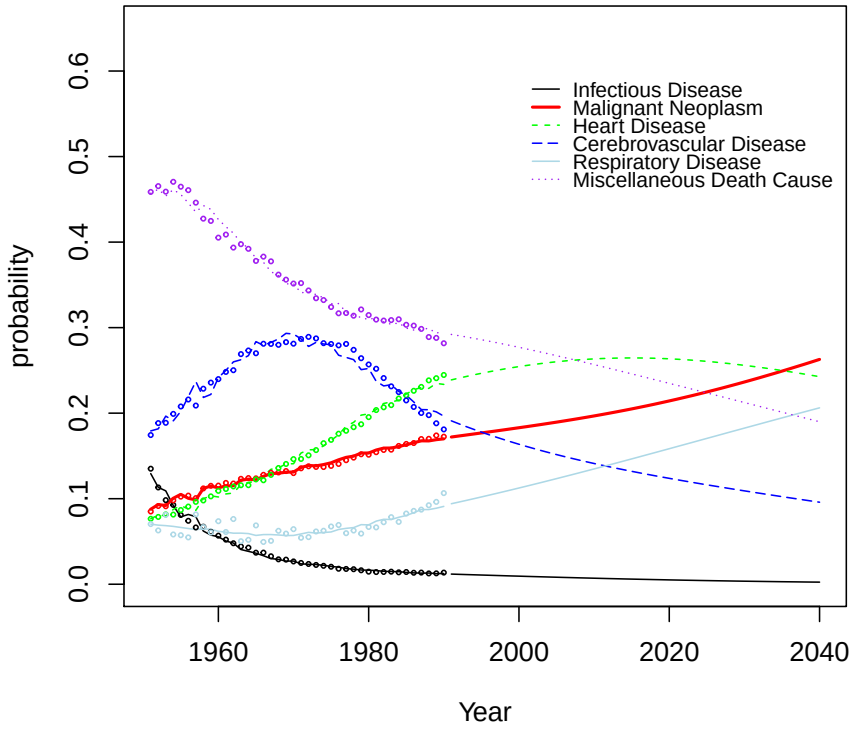

Figure 5.6: Japan, female: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-2 approximation is used. ARIMA(0,2,2) is used to fit and predict the first and second period factors.

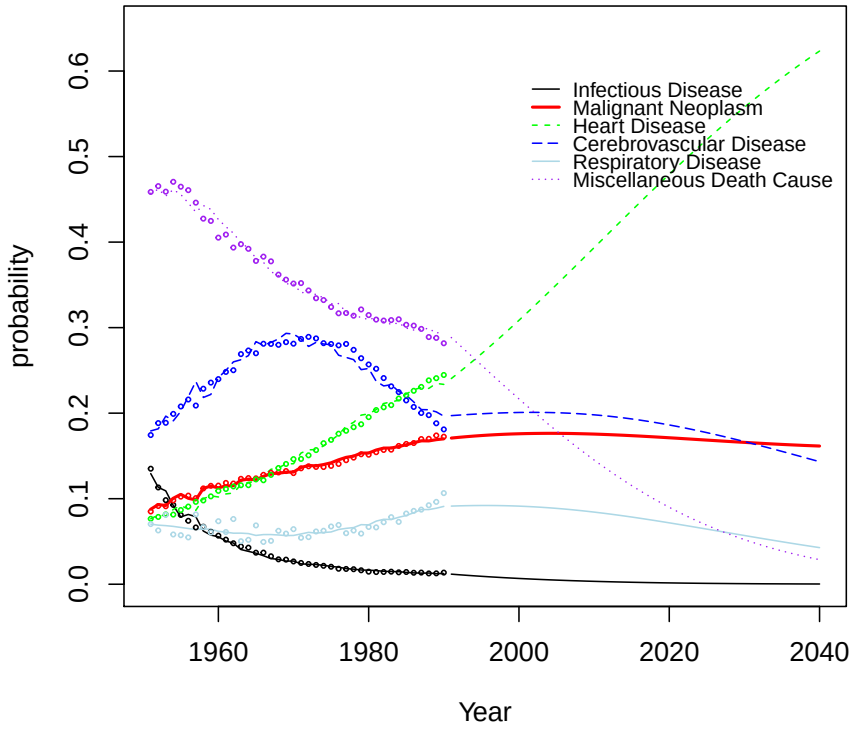

Figure 5.7: Japan, female: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-2 approximation is used. ARIMA(0,1,0) is used to fit and predict the first and second period factors.

Chapter 6

Density of Death Prediction based on Short Observation Period

In Chapter 5, the projection of density of death for Japan female is based on 40 years' observations from year 1951 to 1990. The HMD can usually provide us with life tables for the last 50 years or more. However, the numbers of deaths from different causes might not be available for a very long time period; the reasons might be that there were no mature system of classifications of diseases, like the ICD-10 we use today, during the early and middle of 20^{th} century. Actually ICD-6, published in 1949, was the first to be shaped to become suitable for morbidity reporting. Therefore the systematic records of numbers of deaths from different causes appeared within only the recent decades. In this chapter, we are interested in studying whether a relatively short observation period, say around 10 years, can produce projections based on CoDa LC model that are still reasonable. We will use Japan female and Canada female data to discuss the problem.

6.1 Japan Projection

We will consider three cases for Japan female data. First, Case 1: we use years 1981 to 1990 to fit the model and project the density of death for 50 years from year 1991 to 2040. We use ARIMA(0,2,2) model to fit the first and second period factors. The $l^i_{0,t}$ and $\hat{l}^i_{0,t}$ are plotted in Figure 6.1.

We can see that the fit from 1981 to 1990 is pretty nice and the predictions for the next 20 years from year 1991 to 2010, seem to be reasonable. We can also see that for years 2010 to 2040, the curve for "Malignant Neoplasm" goes up dramatically, which means that by year 2040 almost half of the deaths will be caused by "Malignant Neoplasm". We can compare Figure 6.1 to Figure 5.6. In Figure 5.6, the probability for "Malignant Neoplasm" also increases as years pass by; however, the

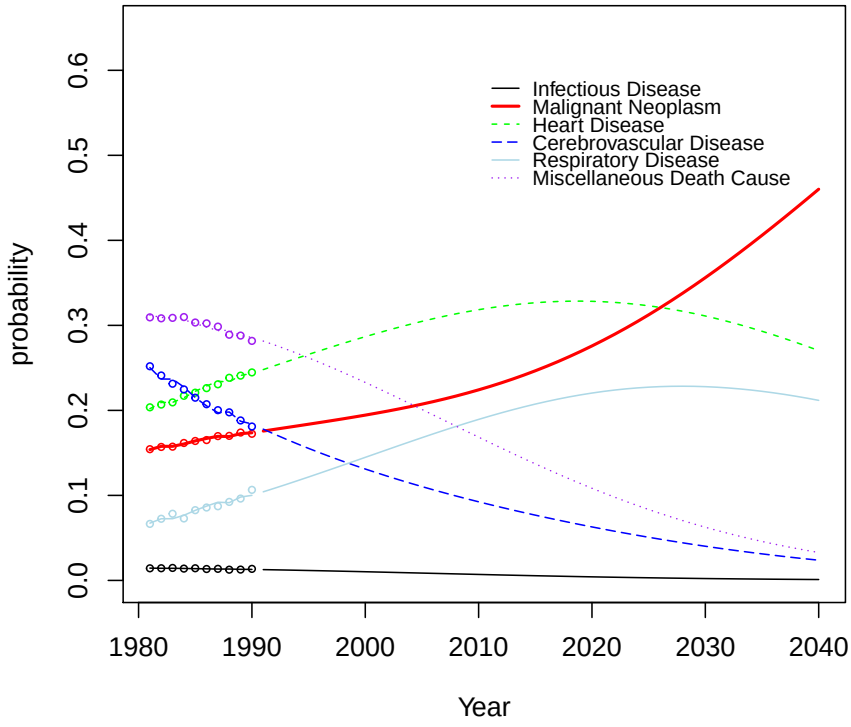

Figure 6.1: Japan, female, Case 1: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-2 approximation is used. ARIMA(0,2,2) is used to fit and predict the first and second period factors.

curve does not behave that dramatically. Since we have no data for year 1991 and after, it is hard for us to comment on the predictions.

Now let's consider Case 2 where we use years 1971 to 1980 to fit the model and project the density of death for 10 years from year 1981 to 1990. Since we have data for years 1981 to 1990, we can comment on the 10-year predictions. The $l^i_{0,t}$ and $\hat{l}^i_{0,t}$ are plotted in Figure 6.2.

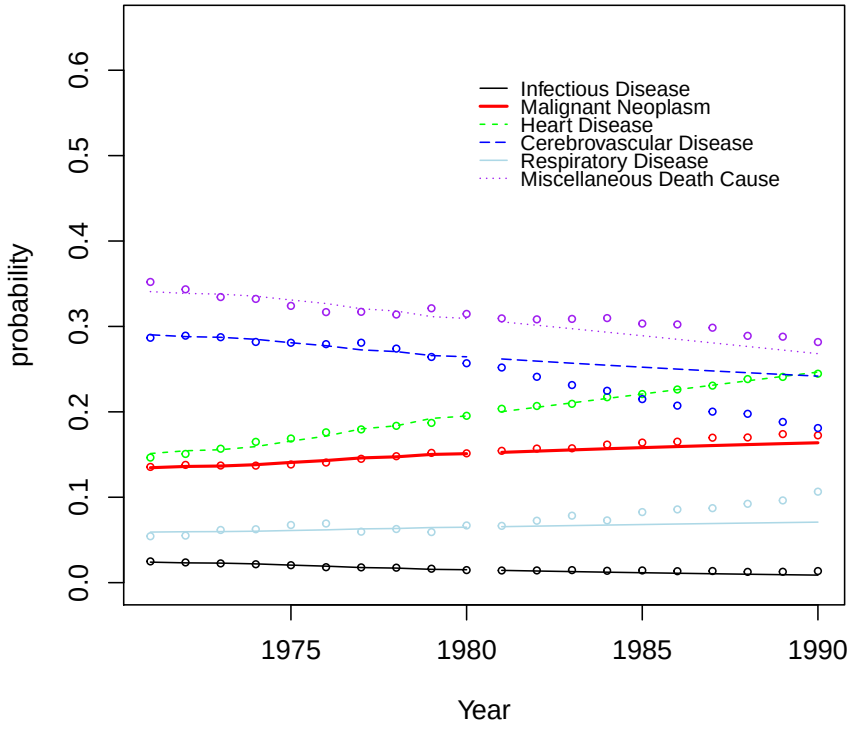

Figure 6.2: Japan, female, Case 2: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-2 approximation is used. ARIMA(0,2,2) is used to fit and predict the first and second period factors.

From Figure 6.2, we can see that for each cause of death, the data for the fitted periods are roughly in a flat line. Based on this dataset, the predictions seem to be 6 flat lines. For cause "Cerebrovascular Disease" the predictions do not capture the downward trend from 1981 to 1990 well. Similarly, for cause "Respiratory Disease" the predictions do not capture the upward trend from 1981 to 1990 at all.

Now as a comparison, in Case 3 we use 30 years from year 1951 to 1980 to fit the model and

project the density of death for the next 10 years from year 1981 to 1990. The plot is in Figure 6.3.

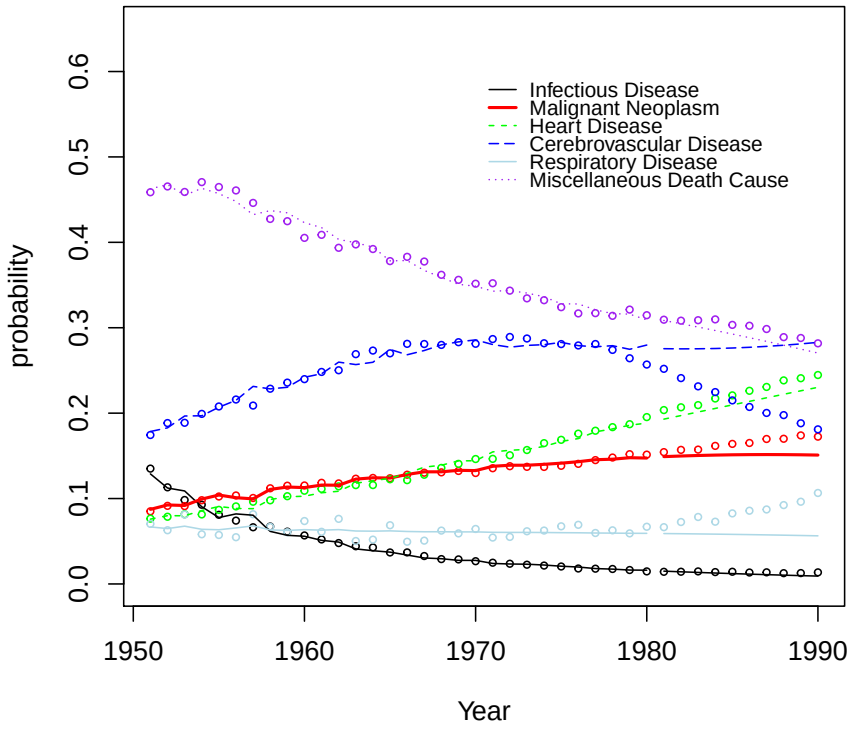

Figure 6.3: Japan, female, Case 3: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-2 approximation is used. ARIMA(0,2,2) is used to fit and predict the first and second period factors.

From Figure 6.3, we can see that the predictions are not very good especially for causes "Cerebrovascular Disease" and "Respiratory Disease". Comparing Figure 6.2 and Figure 6.3, we can see that although the fitted period for Case 2 (10 years) is less than that for Case 3 (30 years), the 10-year predictions, in general, are closer to the data points for Case 2 than Case 3. Therefore, whether the predictions based on the short period's data are good or not depends on the chosen time periods. Taking cause "Cerebrovascular Disease" as an example, we can see if we use data before 1980 to fit the model and predict into the future, it is really hard to capture the trends, as a result, more information from hospitals, clinics, research labs, etc. needs to be collected and various investigations need to be done to make better predictions.

6.2 Canada Projection

In this section, we will use Canada female data to examine projections based on short observation periods. Human Mortality Database provides the life tables for Canada female, and we will use the 5×1 life table. The Canada female 5×1 life table provides data for age groups 0, 1-4, 5-9,..., 105-109 and 110+, and for years from 1921 to 2009. Statistics Canada provides CANSIM Tables. CANSIM is Statistics Canada's key socioeconomic database. Updated daily, CANSIM provides fast and easy access to a large range of the latest statistics available in Canada. The CANSIM database is very user-friendly. For example, we use Table 102-0561: Leading causes of death, total population, by age group and sex, Canada. Under the title of the table, we can find a button "Add/Remove data". After we click on the "Add/Remove data" button, a page with 7 steps appears. By selecting the specific items from each step, we can create our customized CANSIM table. We choose to use number of deaths for Canada, female, all causes of death from year 2001 to 2010 with age intervals 1-14, 15-19,..., 85-89 and 90+. Therefore the data from HMD is adjusted to the same age intervals (1-14, 15-19,..., 85-89 and 90+) and period (2001-2010). The causes of death are categorized into 6 groups, which are "Infectious Disease", "Malignant Neoplasm", "Nervous system Disease", "Respiratory Disease", "Accident" and "Miscellaneous Death Cause".

After we fit our data to multiple-decrement CoDa LC model, we find that the first and second period factors hardly have a clear trend. As a result, the best choice of ARIMA models for the first and second period factors by AICc is ARIMA(0,0,0). Therefore the rank-1 or rank-2 predictions are basically equal to the values of the most recent data. The $l^i_{0,t}$ and $\hat{l}^i_{0,t}$ are plotted in Figure 6.4.

We can see that for Canada female data, the multiple-decrement CoDa LC model predictions are questionable. From Figure 6.4, we can see that except for the cause "Nervous system Disease", the points for the fitted periods for each cause of death are basically on a flat line. Due to the unit sum constraint of cause-specific density of death, we can conclude that the downward trend of cause "Nervous system Disease" is counterbalanced by several minor upward trends of other causes. As long as the overall density of death has a relatively clear trend, the single-decrement CoDa LC model is probably able to describe it. It seems that the multiple-decrement CoDa LC model, however, does not work when there is no obvious trends for the density of death for most of the causes of death (even if the overall density of death has a clear downward trend).

In Chapter 7, we will use our model to fit a short period's USA data (12 years), and then predict the density of death for the next 50 years. After that, we will apply the predicted values to calculate the expectation and variance of a term life insurance with a term accidental death rider.

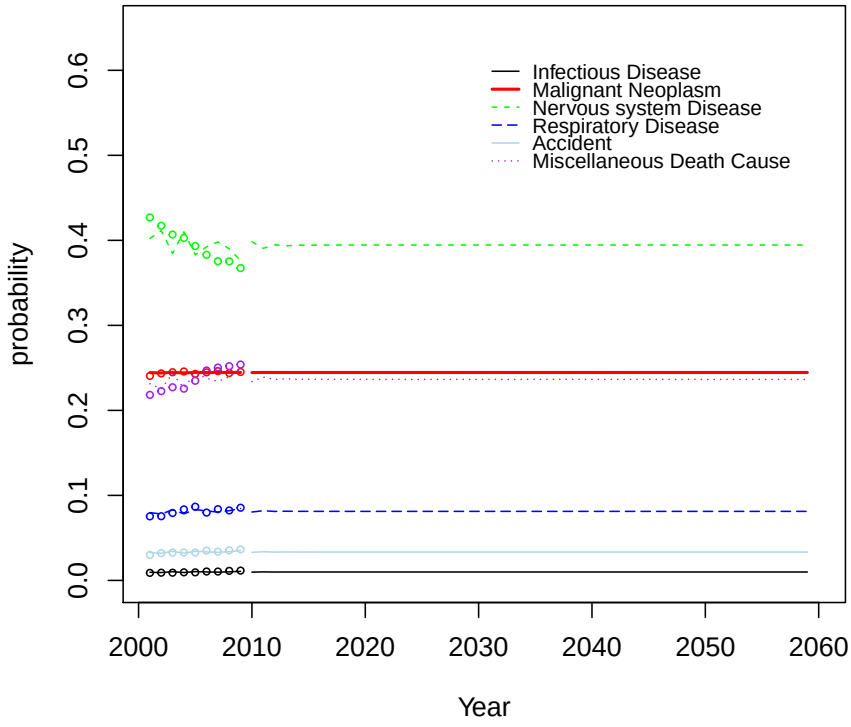

Figure 6.4: Canada, female: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates and predictions based on rank-2 approximation. ARIMA models for the first and second period factors are chosen by AICc.

Chapter 7

Pricing Life Insurance with a Rider

The Lee-Carter model is able to project the mortality rates and one can apply these projections to price life insurances and life annuities. By using a multiple-decrement CoDa LC model, we are able to project the cause-specific death rates and apply them to price more complicated products, such as an accidental death rider. In this chapter, we will price, for different ages and years, a 20-year life insurance with an accidental death rider and calculate the associated variances.

7.1 Life Insurance and Accidental Death Rider

Consider an n-year term insurance that pays a death benefit of $\$1$ at the end of the year of death if within n years. Let K_x be the curtate future lifetime random variable for someone aged x; then the present value random variable for the benefit for someone aged x is

$$Z = \begin{cases} v^{K_x+1} & if \;\; K_x \leq n-1 \\ 0 & if \;\; K_x \geq n \end{cases} . \tag{7.1}$$

The probability function of K_x is

$$P\left(K_x = k\right) = {}_kp_x \cdot q_{x+k}, \tag{7.2}$$

where ${}_kp_x$ represents the probability that someone aged x lives at age $x+k$ and q_{x+k} represents the probability that someone aged $x+k$ dies between ages $x+k$ and $x+k+1$. We will use ${}_k|q_x$ to represent $P\left(K_x = k\right)$. The EPV of the benefit, $E\left(Z\right)$, is therefore

$$E\left(Z\right) = \sum_{k=0}^{n-1} v^{k+1} \cdot {}_k\,|q_x. \tag{7.3}$$

We are not only interested in the first moment, that is the expectation of Z, we are also interested in the second moment of Z, which is helpful to calculate the variance,

$$E\left(Z^2\right) = \sum_{k=0}^{n-1} v^{2(k+1)} \cdot_k |q_x. \tag{7.4}$$

Now let's consider an n-year term accidental death rider, which pays a benefit of $\$1$ at the end of the year of death if within n years due to an accidental cause. The present value random variable for the accidental death benefit for someone aged x is

$$Y = \begin{cases} v^{K_x+1} & if \;\; K_x \leq n-1 \; and \; accidental \; death \\ 0 & if \;\; K_x \geq n \; or \; non-accidental \; death \end{cases}. \tag{7.5}$$

Let ${}_k|q_x^{(ac)}$ be the probability that someone aged x will die between ages $x+k$ and $x+k+1$ due to an accidental cause. The EPV of the benefit, $E\left(Y\right)$, is therefore

$$E\left(Y\right) = \sum_{k=0}^{n-1} v^{k+1} \cdot_k |q_x^{(ac)}. \tag{7.6}$$

Note that

$${}_k|q_x^{(ac)} = {}_kp_x \cdot q_{x+k}^{(ac)}, \tag{7.7}$$

where $q_{x+k}^{(ac)}$ represents the probability that someone aged $x+k$ will die in the next year due to accident. Then the second moment of Y is

$$E\left(Y^2\right) = \sum_{k=0}^{n-1} v^{2(k+1)} \cdot_k |q_x^{(ac)}. \tag{7.8}$$

We are interested in an n-year life insurance with an accidental death rider. The present value random variable is therefore Z+Y, denoted as W:

$$W = \begin{cases} 2v^{K_x+1} & if \;\; K_x \leq n-1 \; and \; accidental \; death \\ v^{K_x+1} & if \;\; K_x \leq n-1 \; and \; non-accidental \; death \\ 0 & if \;\; K_x \geq n \end{cases}. \tag{7.9}$$

Let ${}_k|q_x^{(nac)}$ be the probability that someone aged x will die between ages $x+k$ and $x+k+1$ due to a non-accidental cause, then ${}_k|q_x^{(nac)} = {}_k\, p_x \cdot q_{x+k}^{(nac)}$. Therefore the first and second moments of W are

$$E\left(W\right) = 2\sum_{k=0}^{n-1} v^{k+1} \cdot_k |q_x^{(ac)} + \sum_{k=0}^{n-1} v^{k+1} \cdot_k |q_x^{(nac)} \tag{7.10}$$

and

$$E\left(W^2\right) = 4\sum_{k=0}^{n-1} v^{2(k+1)} \cdot_k |q_x^{(ac)} + \sum_{k=0}^{n-1} v^{2(k+1)} \cdot_k |q_x^{(nac)}. \tag{7.11}$$

The variance of W is

$$V(W) = E\left(W^2\right) - \left(E(W)\right)^2. \tag{7.12}$$

Now say that we want to calculate the expectation and variance of a life insurance with an accidental death rider for someone aged x in year t. Let $W_{x,t}$ represent the present value random variable; then the expectation of $W_{x,t}$ can be expressed as:

$$\begin{aligned} E(W_{x,t}) = 2\sum_{k=0}^{n-1} v^{k+1} \cdot p_{x,t} \cdot p_{x+1,t+1} \cdots p_{x+k-1,t+k-1} \cdot q^{(ac)}_{x+k,t+k} \\ + \sum_{k=0}^{n-1} v^{k+1} \cdot p_{x,t} \cdot p_{x+1,t+1} \cdots p_{x+k-1,t+k-1} \cdot q^{(nac)}_{x+k,t+k}. \end{aligned} \tag{7.13}$$

The second moment of $W_{x.t}$ can be expressed as:

$$\begin{aligned} E\left(W^2_{x,t}\right) = 4\sum_{k=0}^{n-1} v^{2(k+1)} \cdot p_{x,t} \cdot p_{x+1,t+1} \cdots p_{x+k-1,t+k-1} \cdot q^{(ac)}_{x+k,t+k} \\ + \sum_{k=0}^{n-1} v^{2(k+1)} \cdot p_{x,t} \cdot p_{x+1,t+1} \cdots p_{x+k-1,t+k-1} \cdot q^{(nac)}_{x+k,t+k}. \end{aligned} \tag{7.14}$$

The variance of $W_{x,t}$ is therefore

$$V(W_{x,t}) = E\left(W^2_{x,t}\right) - \left(E(W_{x,t})\right)^2. \tag{7.15}$$

7.2 Numerical Illustrations-USA Data

7.2.1 Cause-Specific Density of Death

In this section, we will calculate for various ages (every age from 25 to 60) and years (calendar year 2000, 2010, 2020, 2030 and 2040), the expectations and variances of a 20-year life insurance with an accidental death rider. We will apply the model described in Chapter 5 to project the USA cause-specific density of death and then to calculate the expectations and variances. We collect our data from the Human Mortality Database and Centers for Disease Control and Prevention (CDC). Human Mortality Database provides the life tables for USA, for both sexes. The USA life tables provide data for years as early as 1933 up to 2010. The Centers for Disease Control and Prevention is the national public health institute of the United States with a main goal of protecting public health and safety through the control and prevention of disease, injury and disability. CDC provides a variety of useful reports, such as National Vital Statistics Reports, National Health Statistics Reports and so on. The National Vital Statistics Reports include one report called Deaths: Leading Causes for XXXX (a calendar year), from which we can obtain the cause-specific numbers of deaths for USA in year XXXX. We use 12 reports: Deaths: Leading Causes for 1999, 2000,... and 2010. The

reports can be downloaded from http://www.cdc.gov/nchs/products/nvsr.htm. In each report, Table 10 provides us with the numbers of deaths from 113 selected causes by age for United States for that particular year. In Table 10, the ages are grouped as “Under 1 year”, “1-4 years”, “5-14 years”, “15-24 years”,...,“75-84 years” and “85 years and over”, and we are able to calculate the death rates ($q_{x,t}$) in such age groups based on HMD 1×1 or 5×1 life tables. We categorize the 113 causes into 9 major groups (with reference to International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) Version for 2010). The 9 groups are listed in Table 7.1 and how we categorize the selected 113 causes into the 9 causes is shown in Table 7.2.

Key	Categorization
Cause 1	Infectious Disease
Cause 2	Malignant Neoplasm
Cause 3	Diabetes
Cause 4	Cardiovascular Disease
Cause 5	Respiratory Disease
Cause 6	Genitourinary Disease
Cause 7	Accident
Cause 8	Digestive Disease
Cause 9	Miscellaneous Death Cause

Table 7.1: 9 categorizations of death causes

We are now able to construct the matrix of $\mathbf{d}_{mul}$, the cause-specific density of death matrix for USA, use the multiple-decrement CoDa LC model to fit the density of death for years 1999 to 2010 and then make projection for years 2011 to 2060. The age groups are 0-15, 15-24, 25-34, ..., 75-84 and 85+. The first age factors for 9 different causes are shown in different colors in Figure 7.1.

The first and second period factors are fitted for years 1999 to 2010 and predicted for years 2011 to 2060, by an ARIMA model. We use the AICc criterion (Shumway and Stoffer, 2000) to choose the most appropriate ARIMA model for the first period factor. It turns out that a random walk, that is an ARIMA(0,1,0), is the best choice. For the second period factor, the optimal time series chosen by AICc is an ARIMA(1,0,0). But since the coefficient is around 0.6 and the most recent second period factor in year 2010 is 0.22, the period factors for years 2011 and after are smaller and smaller, which means that when we make projections about future densities of deaths, the second period factor will have insignificant effects. Finally, we choose a rank-1 approximation and use a random walk model for the first period factor. The first period factor and ARIMA(0,1,0) predictions are shown in Figure 7.2 (left panel) and the second period factor and ARIMA(1,0,0) predictions are shown in Figure 7.2 (right panel).

According to (5.13), we can calculate $l^i_{0,t}$ and its estimate $\hat{l}^i_{0,t}$, and then plot them against years. The plots are shown in Figure 7.3.

Cause of death	Key
Salmonella infections (A01-A02)	Cause 1
Shigellosis and amebiasis (A03,A06)	Cause 1
Certain other intestinal infections (A04,A07-A09)	Cause 1
Respiratory tuberculosis (A16)	Cause 1
Other tuberculosis (A17-A19)	Cause 1
Whooping cough (A37)	Cause 1
Scarlet fever and erysipelas (A38,A46)	Cause 1
Meningococcal infection (A39)	Cause 1
Septicemia (A40-A41)	Cause 1
Syphilis (A50-A53)	Cause 1
Acute poliomyelitis (A80)	Cause 1
Arthropod-borne viral encephalitis (A83-A84,A85.2)	Cause 1
Measles (B05)	Cause 1
Viral hepatitis (B15-B19)	Cause 1
Human immunodeficiency virus (HIV) disease (B20-B24)	Cause 1
Malaria (B50-B54)	Cause 1
Other and unspecified infectious and parasitic diseases and their sequelae (A00,A05,A20-A36, A42-A44,A48-A49,A54-A79,A81-A82,A85.0-A85.1, A85.8,A86-B04,B06-B09,B25-B49,B55-B99)	Cause 1
Malignant neoplasms of lip, oral cavity and pharynx (C00-C14)	Cause 2
Malignant neoplasm of esophagus (C15)	Cause 2
Malignant neoplasm of stomach (C16)	Cause 2
Malignant neoplasms of colon, rectum and anus (C18-C21)	Cause 2
Malignant neoplasms of liver and intrahepatic bile ducts (C22)	Cause 2
Malignant neoplasm of pancreas (C25)	Cause 2
Malignant neoplasm of larynx (C32)	Cause 2
Malignant neoplasms of trachea, bronchus and lung (C33-C34)	Cause 2
Malignant melanoma of skin (C43)	Cause 2
Malignant neoplasm of breast (C50)	Cause 2
Malignant neoplasm of cervix uteri (C53)	Cause 2
Malignant neoplasms of corpus uteri and uterus, part unspecified (C54-C55)	Cause 2
Malignant neoplasm of ovary (C56)	Cause 2
Malignant neoplasm of prostate (C61)	Cause 2
Malignant neoplasms of kidney and renal pelvis (C64-C65)	Cause 2
Malignant neoplasm of bladder (C67)	Cause 2
Malignant neoplasms of meninges, brain and other parts of central nervous system (C70-C72)	Cause 2
Hodgkin's disease (C81)	Cause 2
Non-Hodgkin's lymphoma (C82-C85)	Cause 2
Leukemia (C91-C95)	Cause 2
Multiple myeloma and immunoproliferative neoplasms (C88,C90)	Cause 2
Other and unspecified malignant neoplasms of lymphoid, hematopoietic and related tissue (C96)	Cause 2

Table 7.2: Categorization for the selected 113 death causes

Cause of death	Key
All other and unspecified malignant neoplasms (C17,C23-C24,C26-C31,C37-C41,C44-C49,C51-C52,C57-C60,C62-C63,C66,C68-C69,C73-C80,C97)	Cause 2
In situ neoplasms, benign neoplasms and neoplasms of uncertain or unknown behavior (D00-D48)	Cause 9
Anemias (D50-D64)	Cause 9
Diabetes mellitus (E10-E14)	Cause 3
Malnutrition (E40-E46)	Cause 9
Other nutritional deficiencies (E50-E64)	Cause 9
Meningitis (G00,G03)	Cause 9
Parkinsons disease (G20-G21)	Cause 9
Alzheimers disease (G30)	Cause 9
Acute rheumatic fever and chronic rheumatic heart diseases (I00-I09)	Cause 4
Hypertensive heart disease (I11)	Cause 4
Hypertensive heart and renal disease (I13)	Cause 4
Acute myocardial infarction (I21-I22)	Cause 4
Other acute ischemic heart diseases (I24)	Cause 4
Atherosclerotic cardiovascular disease, so described (I25.0)	Cause 4
All other forms of chronic ischemic heart disease (I20,I25.1I25.9)	Cause 4
Acute and subacute endocarditis (I33)	Cause 4
Diseases of pericardium and acute myocarditis (I30-I31,I40)	Cause 4
Heart failure (I50)	Cause 4
All other forms of heart disease (I26-I28,I34-I38,I42-I49,I51)	Cause 4
Essential hypertension and hypertensive renal disease (I10,I12,I15)	Cause 4
Cerebrovascular diseases (I60-I69)	Cause 4
Atherosclerosis (I70)	Cause 4
Aortic aneurysm and dissection (I71)	Cause 4
Other diseases of arteries, arterioles and capillaries (I72-I78)	Cause 4
Other disorders of circulatory system (I80-I99)	Cause 9
Influenza (J10-J11)	Cause 5
Pneumonia (J12-J18)	Cause 5
Acute bronchitis and bronchiolitis (J20-J21)	Cause 5
Unspecified acute lower respiratory infections (J22)	Cause 5
Bronchitis, chronic and unspecified (J40-J42)	Cause 5
Emphysema (J43)	Cause 5
Asthma (J45-J46)	Cause 5
Other chronic lower respiratory diseases (J44,J47)	Cause 5
Pneumoconioses and chemical effects (J60-J66,J68)	Cause 5
Pneumonitis due to solids and liquids (J69)	Cause 5
Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98)	Cause 5
Peptic ulcer (K25-K28)	Cause 8
Diseases of appendix (K35-K38)	Cause 8
Hernia (K40-K46)	Cause 8
Alcoholic liver disease (K70)	Cause 8
Other chronic liver disease and cirrhosis (K73-K74)	Cause 8
Cholelithiasis and other disorders of gallbladder (K80-K82)	Cause 8

Table 7.2: Categorization for the selected 113 death causes-Con.

Cause of death	Key
Acute and rapidly progressive nephritic and nephrotic syndrome (N00-N01,N04)	Cause 6
Chronic glomerulonephritis, nephritis and nephropathy not specified as acute or chronic, and renal sclerosis unspecified (N02-N03,N05-N07,N26)	Cause 6
Renal failure (N17-N19)	Cause 6
Other disorders of kidney (N25,N27)	Cause 6
Infections of kidney (N10-N12,N13.6,N15.1)	Cause 6
Hyperplasia of prostate (N40)	Cause 6
Inflammatory diseases of female pelvic organs (N70-N76)	Cause 6
Pregnancy with abortive outcome (O00-O07)	Cause 9
Other complications of pregnancy, childbirth and the puerperium (O10-O99)	Cause 9
Certain conditions originating in the perinatal period (P00-P96)	Cause 9
Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99)	Cause 9
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)	Cause 9
All other diseases (residual)	Cause 9
Motor-vehicle accidents (V02-V04,V09.0,V09.2,V12-V14, V19.0-V19.2,V19.4-V19.6,V20-V79,V80.3-V80.5,V81.0-V81.1, V82.0-V82.1,V83-V86,V87.0-V87.8,V88.0-V88.8,V89.0,V89.2)	Cause 7
Other land transport accidents (V01,V05-V06,V09.1,V09.3-V09.9, V10-V11,V15-V18,V19.3,V19.8-V19.9,V80.0-V80.2,V80.6-V80.9, V81.2-V81.9,V82.2-V82.9,V87.9,V88.9,V89.1,V89.3, V89.9)	Cause 7
Water, air and space, and other and unspecified transport accidents and their sequelae (V90-V99,Y85)	Cause 7
Falls (W00-W19)	Cause 7
Accidental discharge of firearms (W32-W34)	Cause 7
Accidental drowning and submersion (W65-W74)	Cause 7
Accidental exposure to smoke, fire and flames (X00-X09)	Cause 7
Accidental poisoning and exposure to noxious substances (X40-X49)	Cause 7
Other and unspecified nontransport accidents and their sequelae (W20-W31,W35-W64, W75-W99,X10-X39, X50-X59,Y86)	Cause 7
Intentional self-harm (suicide) by discharge of firearms (X72-X74)	Cause 9
Intentional self-harm (suicide) by other and unspecified means and their sequelae (*U03,X60-X71,X75-X84,Y87.0)	Cause 9
Assault (homicide) by discharge of firearms (*U01.4,X93-X95)	Cause 9
Assault (homicide) by other and unspecified means and their sequelae (*U01.0-*U01.3,*U01.5-*U01.9,*U02,X85-X92,X96-Y09,Y87.1)	Cause 9
Legal intervention (Y35,Y89.0)	Cause 9
Discharge of firearms, undetermined intent (Y22-Y24)	Cause 9
Other and unspecified events of undetermined intent and their sequelae (Y10-Y21,Y25-Y34,Y87.2,Y89.9)	Cause 9
Operations of war and their sequelae (Y36,Y89.1)	Cause 9
Complications of medical and surgical care (Y40-Y84,Y88)	Cause 9

Table 7.2: Categorization for the selected 113 death causes-Con.

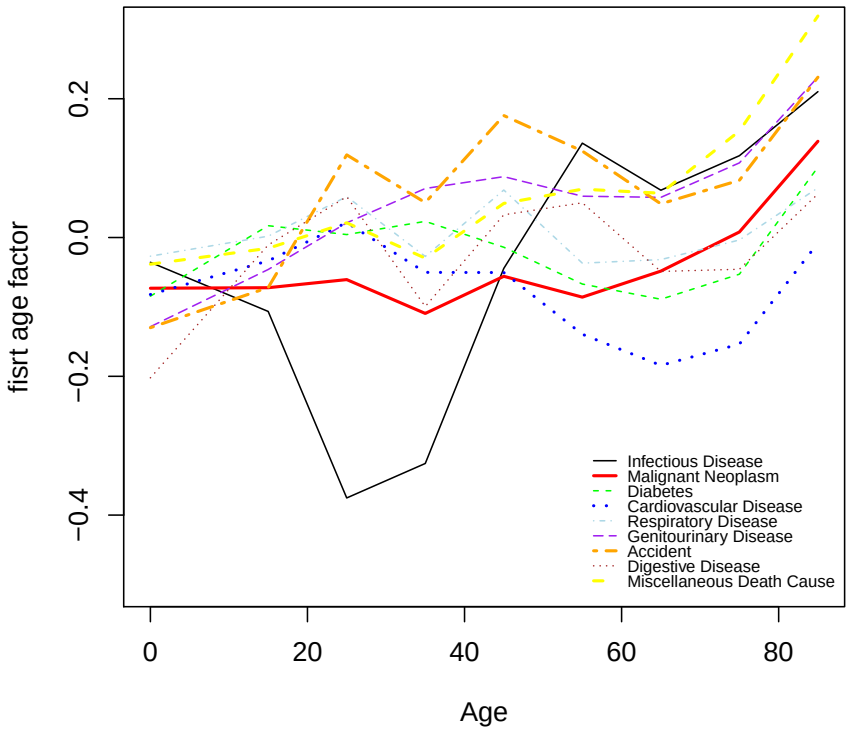

Figure 7.1: USA, both sexes, 1999-2010: the first cause-specific age factors.

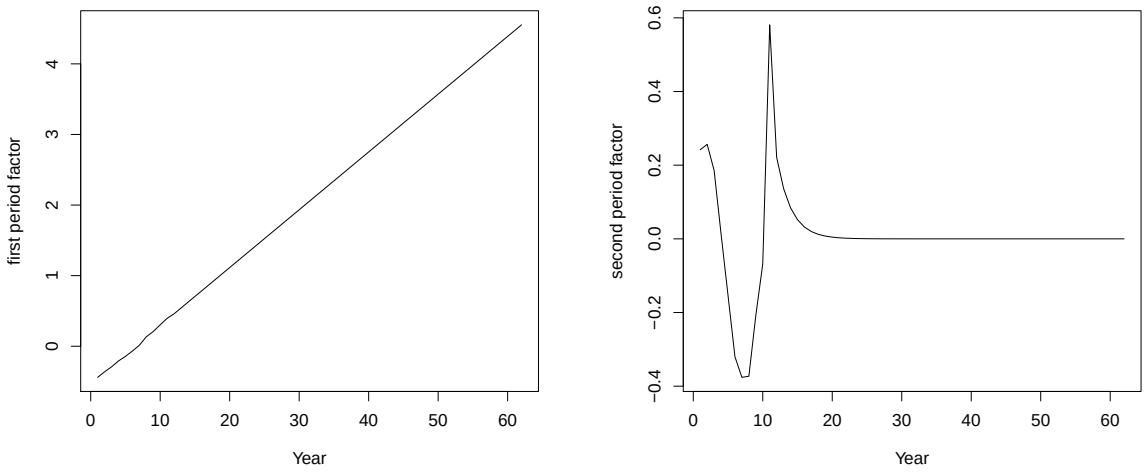

Figure 7.2: USA, both sexes: the fitted (1999-2010) and predicted (2011-2060) first (left panel) and second (right panel) period factors. ARIMA(0,1,0) is used to fit and predict the first period factor. ARIMA(1,0,0) is used to fit and predict the second period factor.

From Figure 7.3, we can see that the rank-1 approximation fits the data (shown as dots in the figure) very well and the predictions are reasonable. According to our result, the probability of death by major cardiovascular diseases decreases from 0.42 in 1999 to 0.04 in 2060. This means that in the next 50 years, there should be some very effective treatments developed for cardiovascular diseases. I believe, however, that the probability of death caused by major cardiovascular diseases will decrease to some level, and then at some point will be more stable and stop decreasing. Anyway, according to our model, it seems that for all the causes of death the $l^i_{0,t}$ decreases or is stable, except for the category "Miscellaneous Death Cause" where $l^i_{0,t}$ increases a lot and contributes over 75% of all the deaths in 2060.

Based on Section 5.4, we are able to obtain the values of ${}_{x_{j+1}-x_j}q^i_{x_j,t}$, which are important for calculating the expectations and variances of insurance contracts and riders.

7.2.2 Expectations and Variances

The death rates we obtained are ${}_{15}q^i_{0,t}$, ${}_{10}q^i_{15,t}$, ${}_{10}q^i_{25,t}$, ..., ${}_{\omega}q^i_{85,t}$. To calculate the expectation and variance of a 20-year life insurance with a 20-year accidental death rider, we need to know the values of $q^7_{0,t}$, $q^7_{1,t}$,..., etc. (note that the accidental death is Cause 7). Within each age intervals (15- , 15-24, 25-34,..., 75-84 and 85+), we use a uniform distribution of death (UDD) assumption. The

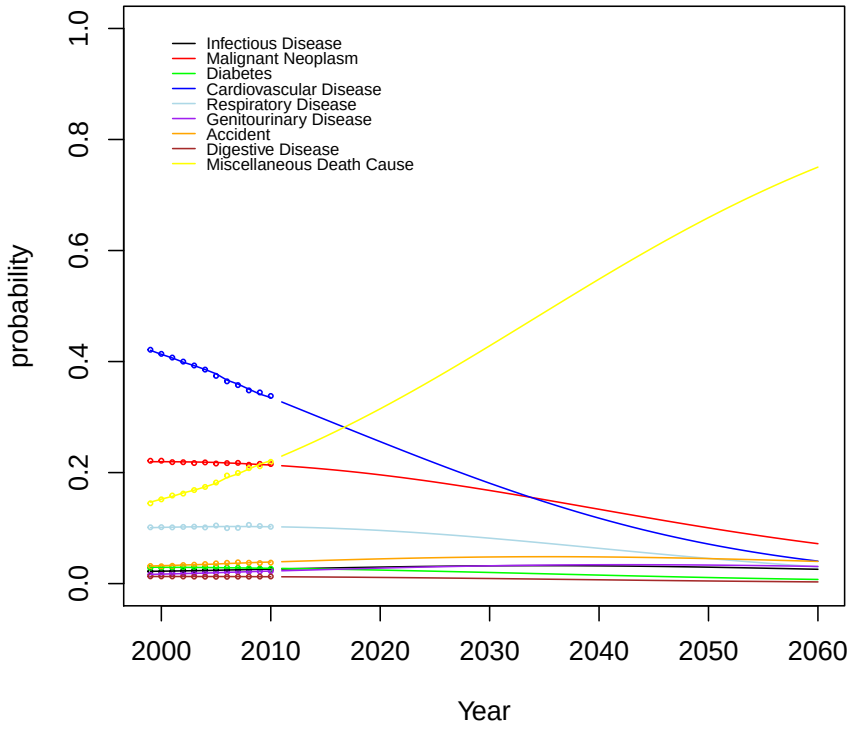

Figure 7.3: USA, both sexes: probability that a newborn will eventually die from a specific cause. Points represent data and lines represent estimates. Rank-1 approximation is used. ARIMA(0,1,0) is used to fit and predict the first period factor.

details about UDD assumption can be found from Bowers et al. (1997). We plot of the surface of $q_{x,t}$, $q_{x,t}^{(ac)}$ and $q_{x,t}^{(nac)}$ for ages 25 to 80 and years 1999 to 2060 in Figures 7.4, 7.5 and 7.6 respectively.

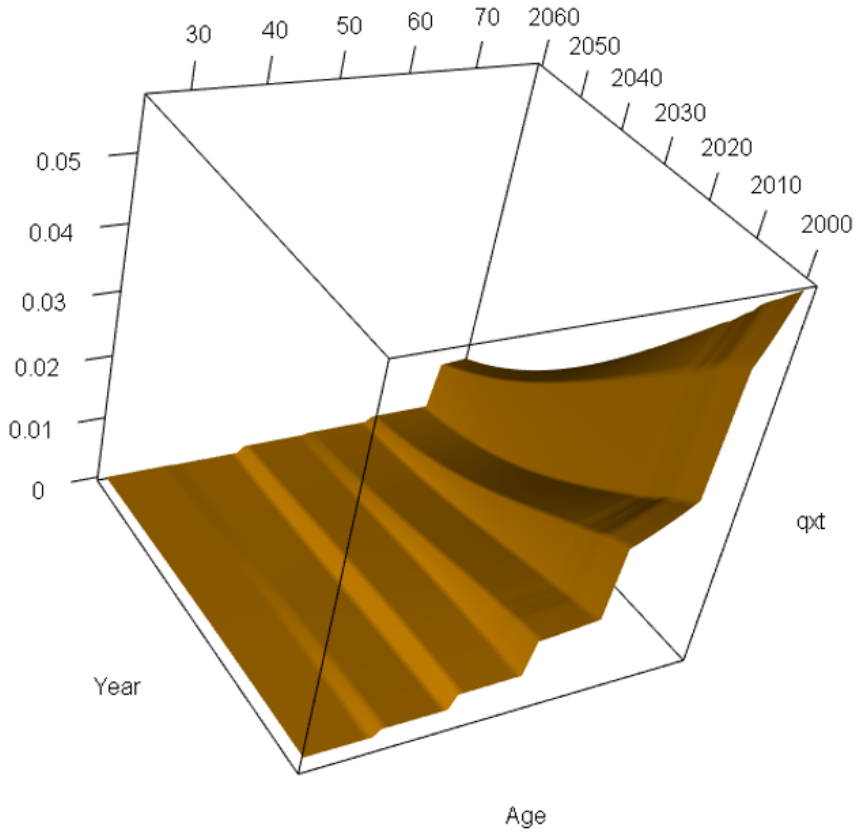

Figure 7.4: USA, both sexes: $q_{x,t}$ for ages 25 to 79 and for years 1999 to 2060.

From Figure 7.4 we can see that for a given year t, the death rates $q_{x,t}$ increases as age increases. For a given age x, the death rates $q_{x,t}$ decreases as a function of t, especially for older ages. According to our model, $q_{79,t}$ decreases by 83% from 0.06 in year 1999 to 0.01 in year 2006. The surface of $q_{x,t}^{(nac)}$ is very similar to that of $q_{x,t}$, since the accidental death rates are relatively small compared to the overall death rates. Now let's take a look at the surface of $q_{x,t}^{(ac)}$ in Figure 7.5; the shape is a little bumpy. For different ages the peaks appear in different calendar years. And for years 2030 and after, it seems that the peaks appear around ages 45 and 55 (in order to see this, we also include Figure 7.7, which is only angle-different from Figure 7.5). As a result when we are calculating the expectations, for years 2030 and beyond, the weights of those accidental deaths at ages around 45 and 55 are heavier than those at elder ages around 80, which means that after

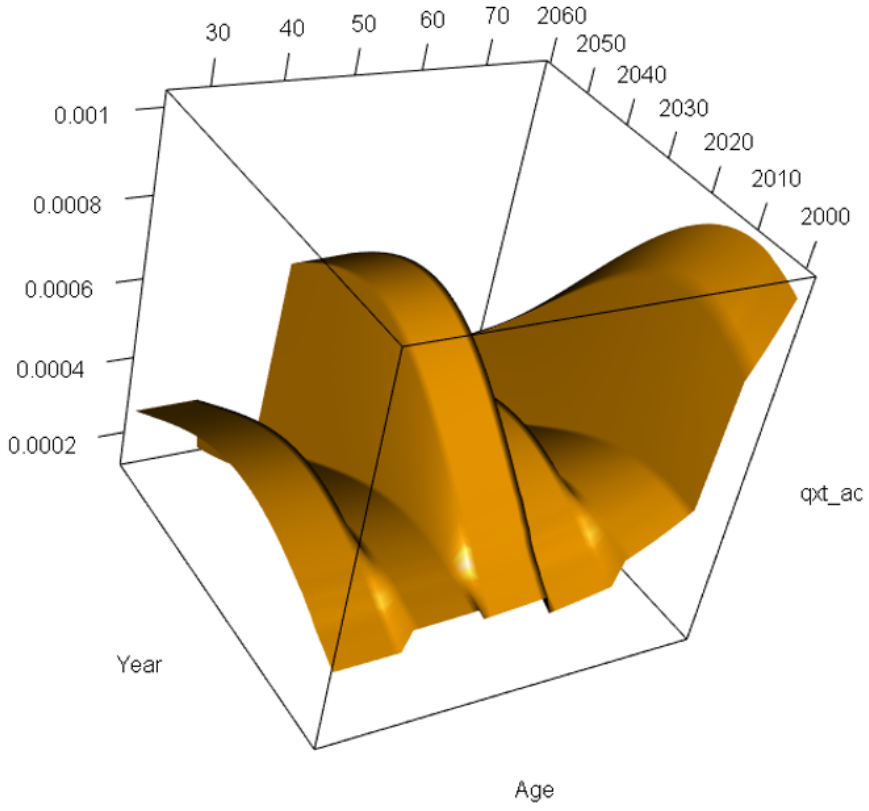

Figure 7.5: USA, both sexes: $q_{x,t}^{(ac)}$ for ages 25 to 79 and for years 1999 to 2060.

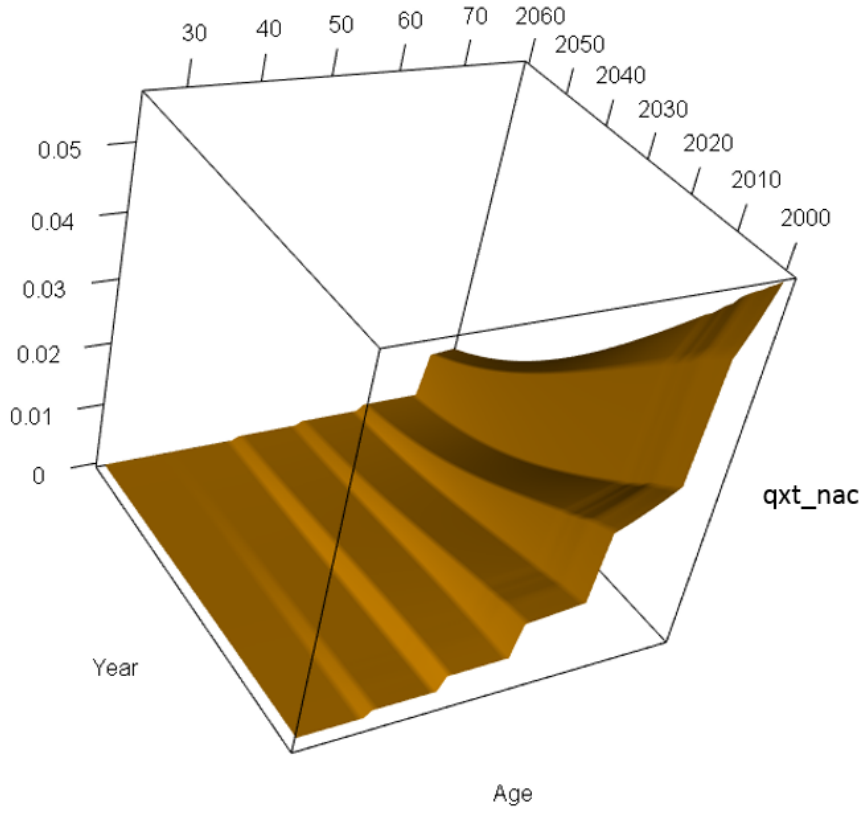

Figure 7.6: USA, both sexes: $q_{x,t}^{(nac)}$ for ages 25 to 79 and for years 1999 to 2060.

year 2030 the expectations at older ages will not increase as much as they will until 2030.

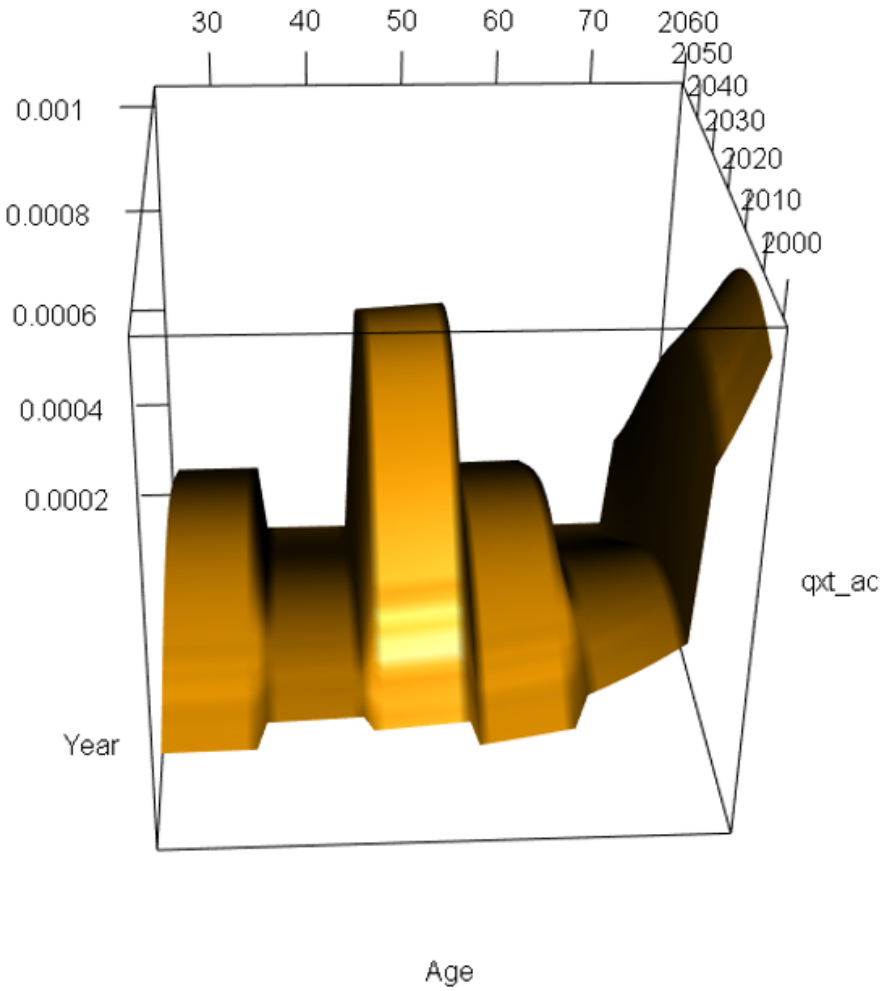

Figure 7.7: USA, both sexes, $q_{x,t}^{(ac)}$ for ages 25 to 79 and for years 1999 to 2060.

We calculate the moments for ages 25 to 60 and for years 2000, 2010, 2020, 2030 and 2040 with an interest rate of 5%. The expectations and variances can be found in Figures 7.8 and 7.9 respectively.

From Figure 7.8, we can observe a more gentle slope for the expectations between ages 45 and 55 for years 2030 and 2040. In Figure 7.9, we can see the same special age periods from 45 to 55.

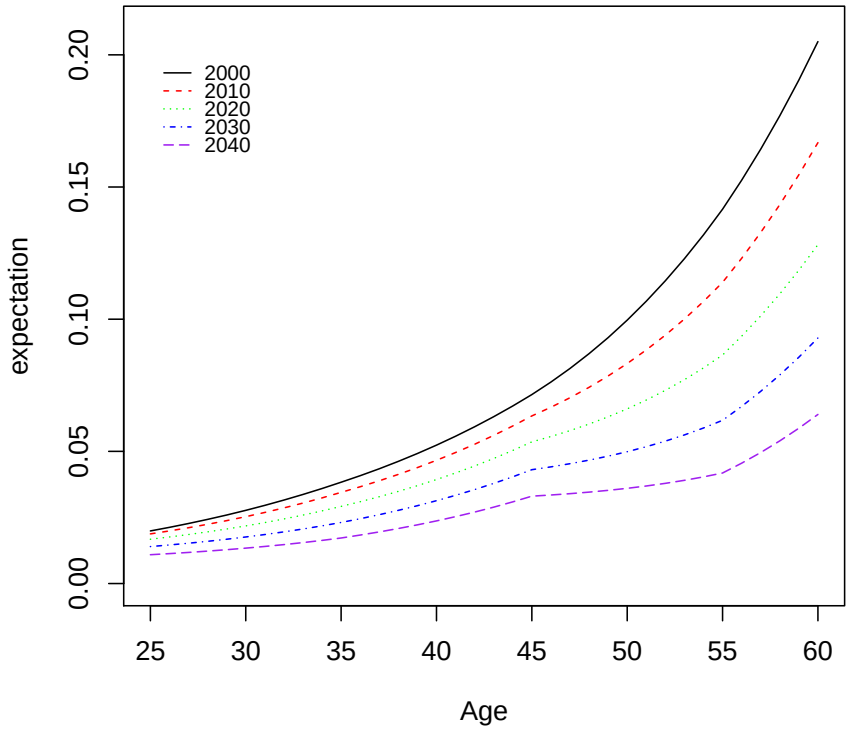

Figure 7.8: USA, both sexes: expected present value for 20-year life insurance with 20-year accidental death rider with interest rate of 5%.

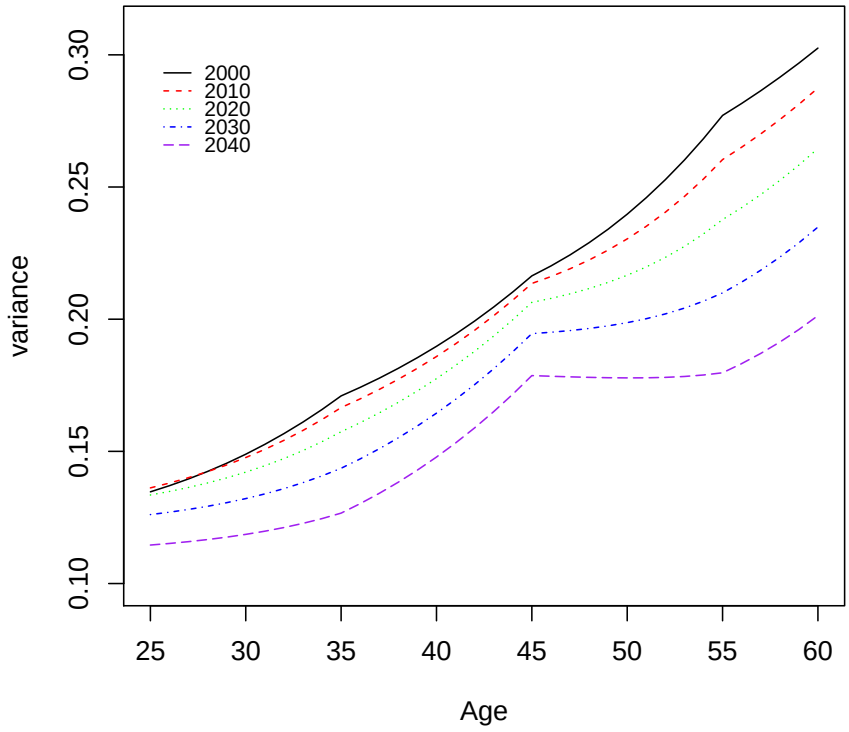

Figure 7.9: USA, both sexes: variance of present value for 20-year life insurance with 20-year accidental death rider with interest rate of 5%.

Chapter 8

Conclusion

In this project, we use Compositional Data Analysis to solve the problem of coherent forecasting of multiple-decrement life tables. Instead of modeling the mortality rates, we are interested in the cause-specific density of death, which can be treated as compositional data. We extend the Lee-Carter model so that the Lee-Carter structure can be applied on compositional data. Following the work of Jim Oeppen (Oeppen, 2008), we apply the model on several multiple-decrement datasets, including Japan (female), Canada (female) and USA (both sexes). Like the Lee-Carter model, the CoDa LC model is easy to understand and simple to implement.

One of the Lee-Carter model's advantages is its simplicity. The CoDa LC model inherits this advantage. At the same time, the CoDa LC model extends the Lee-Carter model from single decrement to multiple decrements. The Lee-Carter model only uses the first left and right singular vectors and the leading value of the SVD to obtain the unique solution of **b** and **k**. And then the LC model takes a second step to reestimate **k**. The CoDa LC model, on the other hand, has more flexibility that enables us to choose the rank of the approximation. Another advantage of the CoDa LC model is that there is no necessity of reestimating the period factors and therefore saves the step of reestimation.

After fitting the model, we are able to do predictions. Very similar to the Lee-Carter model, we use an ARIMA model to fit the period factors and make predictions. But we use AICc to choose the optimal ARIMA model. If we use rank-r approximation, then we need to use AICc to choose the optimal ARIMA model r times, each time for each of the 1st, 2nd,..., and r^{th} period factors. After we project the future density of death, we are able to do some applications such as insurance pricing.

When density of death is only available for around 10 years, whether our model works really depends on what periods we choose to fit. If the period we choose is only 10 years, it is very likely that the model does not work since 10 years' cause-specific death rates might not reflect the downward curvatures as single decrement death rates for every cause of death, and therefore our predictions might not be enough regarding the future trends. While for the USA (both sexes)

data, the fits and predictions seem to be reasonable. For the Lee-Carter model, since there is no disaggregation of death, the downward trends of mortality rates are usually very obvious.

The disaggregation of death is very beneficial. The CoDa LC model enables us to predict the cause-specific density of death. The causes of death includes “Accidents” and therefore we are able to price an n-year term insurance with an n-year accidental death rider. The death rates reach a peak around ages 45 to 55 for years 2030 and after. This characteristic affects the shape of the curves of the expectations and variances for the combination of insurance and rider issued at various ages and different calendar years.

There is still much room for future work. One possible direction would be to consider the problem how many categories one should disaggregate the causes of death into. If we disaggregate the causes of death into too many categories, the multiple-decrement CoDa LC model might not work. Another possible direction would be to obtain the confidence intervals, which will be important for pricing. Applying the LC model variants or extensions might also be another interesting topic.

Bibliography

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall Ltd, London New York.

[2] Aitchison, J. (1986). *A Concise Guide to Compositional Data Analysis.*

[3] Aitchison, J., Mateu-Figueras, G. and Ng, K.W. (2003). Characterization of distributional forms for compositional data and associated distributional tests. *Mathematical Geology*, 35: 667-680.

[4] Berkeley Mortality Database (2014). University of California, Berkeley, USA, http://www.demog.berkeley.edu/ bmd.

[5] Booth, H., Maindonald J. and Smith L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56: 325-336.

[6] Booth, H., Hyndman, R.J., Tickle, L. and de Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, 15: 289-310.

[7] Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: a review of methods. *Annals of Actuarial Science*, 3: 3-43.

[8] Bowers, N., Gerber, H. U., Hickman, J. C., Jones, D. A. and Nesbitt, C. J. (1997). *Actuarial Mathematics, Second Edition*. The Society of Actuaries, Schaumburg, Illinois.

[9] Bozik, J. and Bell, W. (1989). Time series modeling for the principal components approach to forecasting age-specific fertility, paper presented at the 1989 meetings of the Population Association of America, Baltimore.

[10] 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: 373-393.

[11] Cairns, A. J. G., Blake, D. and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *The Journal of Risk and Insurance*, 73: 687-718.

[12] Cairns, A. J. G., Blake, D. and Dowd, K. (2008). Modelling and management of mortality risk: a review. *Scandinavian Actuarial Journal*, 2/3: 79-113.

[13] Cairns, A. J. G., 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: 1-35.

[14] Centers for Disease Control and Prevention, http://www.cdc.gov.

[15] Chayes, F. (1960). On correlation between variables of constant sum. *J. Geophys. Res.*, 65: 4185-4193.

[16] Crimmins, E.M. (1981). The changing pattern of American mortality decline, 1947-1977, and its implication for the future. *Population and Development Review*, 7: 229-254.

[17] Currie, I. D. (2006). Smoothing and forecasting mortality rates with P-splines. Paper given at the Institute of Actuaries, June 2006.

[18] De Jong, P. and Tickle, L. (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies*, 13: 1-18.

[19] Dickson, D.C.M., Hardy, M.R. and Waters, H.R. (2013). *Actuarial Mathematics for Life Contingent Risks, Second Edition*. Cambridge University Press.

[20] Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics and Data Analysis*, 51: 4942-4956.

[21] Lederman, S. (1969). Nouvelles tables-type de mortalité. *Travaux et Documents*, n.53, Paris: Institut National d'études démographiques.

[22] Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87: 659671.

[23] Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38: 537-549.

[24] 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.

[25] Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: A test using Japanese cause of death data (Technical report).

[26] Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc.*, 60: 489-498.

[27] Polder, J.J., Barendregt, J.J. and van Oers, H. (2006). Health care costs in the last year of life - the Dutch experience. *Social Science and Medicine*, 63: 1720-1731.

[28] Preston, S.H., Heuveline, P. and Guillot, M. (2001). *Demography: measuring and modeling population processes*. Blackwell Publishers, Malden, MA.

[29] Renshaw, A. E. and Haberman, S. (2003). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33: 255-272.

[30] Shumway, R.H and Stoffer, D.S. (2000). *Time series analysis and its applications*. Springer texts in statistics, New York.

[31] Wilmoth, J.R. (1995). Are Mortality Projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies*, 5: 293-319.

[32] Wilmoth, J.R., Shkolnikov, V. and Barbieri, M. (2014). Human Mortality Database. University of California at Berkeley, Max Planck Institute for Demographic Research and INED, Paris, http://www.mortality.org.

Appendix A

HMD: Life Table

The values of $q_{x,t}$'s used in our numerical illustrations were obtained from the life tables provided by Human Mortality Database at http://www.mortality.org. An excerpt of the 5×1 life table for USA, both sexes, is shown below.

Year	Age	mx	qx	ax	lx	dx	Lx	Tx	ex
1933	0	0.06167	0.05883	0.22	100,000	5,883	95,405	6,087,969	60.88
1933	1-4	0.00484	0.01911	1.41	94,117	1,799	371,812	5,992,564	63.67
1933	5-9	0.00164	0.00815	2.31	92,318	752	459,563	5,620,752	60.88
1933	10-14	0.00135	0.00674	2.61	91,565	617	456,355	5,161,189	56.37
1933	15-19	0.00229	0.01137	2.70	90,948	1,034	452,360	4,704,834	51.73
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1933	100-104	0.42799	0.89546	1.75	169	152	354	388	2.29
1933	105-109	0.52376	0.93829	1.58	18	17	32	34	1.89
1933	110+	0.60163	1.00000	1.66	1	1	2	2	1.66
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2010	0	0.00619	0.00615	0.07	100,000	615	99,425	7,883,243	78.83
2010	1-4	0.00027	0.00107	1.58	99,385	106	397,281	7,783,818	78.32
2010	5-9	0.00011	0.00057	2.44	99,278	57	496,246	7,386,537	74.40
2010	10-14	0.00014	0.00071	2.81	99,221	71	495,952	6,890,291	69.44
2010	15-19	0.00049	0.00246	2.99	99,151	244	495,263	6,394,339	64.49
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2010	100-104	0.41857	0.89552	1.81	2,167	1,941	4,636	5,031	2.32
2010	105-109	0.56925	0.95520	1.52	226	216	380	395	1.74
2010	110+	0.69050	1.00000	1.45	10	10	15	15	1.45

Table A.1: The United States of America, life table (period 5×1), last modified: 24-Jun-2013.

Appendix B

BMD: Cause-Specific Number of Death

The cause-specific numbers of deaths for Japan females used in Chapter 5 were obtained from a table available in the Berkeley Mortality Database, called “Deaths-Causes of death, 1951-1990, (5×1)”. (See http://www.demog.berkeley.edu/ bmd). An excerpt of the table is shown below. Each cause of death is represented by a unique integer. The correspondence between the integers and the names of the causes can be found in a document called “Data Notes”.

Year	Cause	Age	Female	Male	Total
1951	Total	Total	406,458	432,540	838,998
1951	Total	0	56,005	66,864	122,869
1951	Total	1-4	38,694	40,936	79,630
⋮	⋮	⋮	⋮	⋮	⋮
1951	Total	100+	33	10	43
1951	Total	Unknown	9	23	32
1951	1	Total	7,355	5,852	13,207
1951	1	0	86	109	195
1951	1	1-4	241	248	489
⋮	⋮	⋮	⋮	⋮	⋮
1951	1	100+	1	0	1
1951	1	Unknown	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮
1951	40	Total	78,803	85,919	164,722
1951	40	0	27,483	33,936	61,419
1951	40	1-4	5,434	6,188	11,622
⋮	⋮	⋮	⋮	⋮	⋮
1951	40	100+	2	0	2
1951	40	Unknown	3	4	7
⋮	⋮	⋮	⋮	⋮	⋮

Table B.1: Japan, female, cause-specific numbers of death, 1951-1990 (5×1).

Appendix C

CANSIM Table 102-0561

The cause-specific numbers of deaths for Canadian females used in Chapter 6 were provided by CANSIM Table 102-0561. The CANSIM website is http://www5.statcan.gc.ca/cansim. An excerpt of the table is shown below.

Death age	**Causes of death (ICD-10)**	2001	2002	...	2010
⋮	⋮	⋮	⋮	⋮	⋮
1-14	All causes	385	383	...	285
1-14	Salmonella infections [A01-A02]	0	0	...	0
1-14	Shigellosis and amoebiasis [A03, A06]	0	0	...	0
1-14	Tuberculosis [A16-A19]	0	0	...	0
1-14	Whooping cough [A37]	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮
15-19	All causes	305	305	...	274
15-19	Salmonella infections [A01-A02]	0	0	...	0
15-19	Shigellosis and amoebiasis [A03, A06]	0	0	...	0
15-19	Tuberculosis [A16-A19]	0	0	...	0
15-19	Whooping cough [A37]	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮
90+	All causes	20,615	21,518	...	28,415
90+	Salmonella infections [A01-A02]	0	2	...	1
90+	Shigellosis and amoebiasis [A03, A06]	1	0	...	0
90+	Tuberculosis [A16-A19]	6	7	...	1
90+	Whooping cough [A37]	0	0	...	0

Table C.1: Cause-specific numbers of death by age: Canada, female, 2001-2010.

Appendix D

National Vital Statistics Report

The cause-specific numbers of deaths for USA both sexes used in Chapter 7 were provided by Table 10 of the National Vital Statistics Reports. The National Vital Statistics Reports can be searched at Centers for Disease Control and Prevention at http://www.cdc.gov. An excerpt of the table for calendar year 2007 is shown below.

Cause of death (ICD-10)	**All ages**	< 1	1-4	5-14	...	85+
All causes	2,423,712	29,138	4,703	6,147	...	201
Salmonella infections (A01-A02)	30	2	2	-	...	9
Shigellosis and amebiasis (A03,A06)	4	-	-	2	...	-
Certain other intestinal infections (A04,A07-A09)	6,758	11	14	5	...	2,864
Tuberculosis (A16-A19)	554	1	-	1	...	90
Respiratory tuberculosis (A16)	424	2	-	1	...	74
Other tuberculosis (A17-A19)	130	-	-	-	...	16
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table D.1: Number of deaths from 113 selected causes by age: United States, 2007.