Modeling issues in longevity and bonus-malus system

Tan, Chong It

2015

Tan, C. I. (2015). Modeling issues in longevity and bonus-malus system. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/63168>

<https://doi.org/10.32657/10356/63168>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**MODELING ISSUES IN LONGEVITY AND
BONUS-MALUS SYSTEM**

CHONG IT TAN

NANYANG BUSINESS SCHOOL

2015

MODELING ISSUES IN LONGEVITY AND BONUS-MALUS SYSTEM

CHONG IT TAN

2015

MODELING ISSUES IN LONGEVITY AND BONUS-MALUS SYSTEM

By

CHONG IT TAN

CHONG IT TAN

Nanyang Business School

A thesis submitted to Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2015

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to my supervisors, Dr. Jackie Li and Dr. Johnny Siu-Hang Li, for their advice and guidance over the years on how to become a good researcher. I would also like to thank Dr. Uditha Balasooriya for inspiring me to pursue actuarial research since my second year in undergraduate studies.

Thanks to other members of my thesis committee – Dr. Shinichi Kamiya, Dr. Keng Leong Yeo and Dr. Jiang Luo – for their valuable participation and suggestions, and to Bowen Yang for many insightful discussions. Special thanks go to Karen Barlaan, Bee Hua Quek and Florence Mui Luang Cher for their indispensable help during my period of studies in Nanyang Business School.

I am also extremely grateful to Aon Benfield Asia Pte. Ltd. for providing me with the financial assistance to complete this thesis. A big thank you to George Attard and John Tze Fei Lian for the wonderful receptions and correspondences.

Finally, I am indebted to my girlfriend, Guek Li Koh, for her care and support, and to my beloved parents, Chye Hock Tan and Siok Hoi Heng, who have always believed in me and encouraged me to pursue my doctoral dream.

Abstract

This thesis explores several modeling issues in longevity and bonus-malus system. On one hand, research in the areas of longevity and mortality are closely related. Mortality is one of the oldest research topics in actuarial science, dates back to year 1825 when Benjamin Gompertz published the law of human mortality. On the other hand, bonus-malus systems are widely used in motor insurance to relate premium amounts to individual past claims experience.

In the study of human longevity, both period and cohort survival curves have their own advantages and shortcomings. We construct a new type of survival curves called hybrid survival curve by combining the strengths of these two types of survival curves. We then propose two stochastic survival models for modeling the evolution of a hybrid survival curve. The estimated time-varying parameters are highly interpretable and their respective trends can be used as an indicator for the rectangularization of survival curve. On top of that, we demonstrate how the time-varying parameters can be extrapolated into the future to obtain projected hybrid survival curves.

In recent years, a collection of stochastic mortality models were developed to forecast the uncertainties in mortality projections. In particular, the time-varying parameters in common stochastic mortality models can be used to construct mortality indexes. We study how existing mortality models can be adapted to satisfy the new-data-invariant property, a property that is required to ensure the resulting mortality indexes are tractable by market participants. We find that the adapted Model M7 (the Cairns-Blake-Dowd model with cohort and quadratic age effects) is the most suitable for constructing mortality indexes. Based on the indexes created from this model, one can write a standardized mortality security called K-forward, which can be used to hedge longevity risk exposures. We also

contribute a hedging method called key K-duration that permits one to calibrate a longevity hedge formed by K-forward contracts. Our numerical illustrations indicate that a K-forward hedge has a potential to outperform a q-forward hedge in terms of the number of hedging instruments required.

Traditionally, when working with cross-sectional data, the motor insurance ratemaking process is separated into two steps – *a priori* ratemaking and *a posteriori* ratemaking. When a bonus-malus system (BMS) with a single set of optimal relativities and a set of simple transition rules is implemented, two inadequacy scenarios are induced because all policyholders are subject to the same *a posteriori* premium relativities independent of their *a priori* characteristics and the same level transitions independent of their current levels occupied. We develop alternatives to alleviate these two inadequacy scenarios under the BMS framework developed for cross-sectional data. In recent times, a panel data structure is available for the ratemaking process. For this kind of data, Boucher and Inoussa (2014) proposed a single-step ratemaking approach to estimate *a priori* premiums and *a posteriori* relativities simultaneously to tackle the first inadequacy scenario. Using our proposed varying transition rules, the second inadequacy scenario can be mitigated under the BMS framework developed for panel data.

Contents

List of Tables	ix
List of Figures	xiii
1 Introduction	1
1.1 Background	1
1.2 Objectives and Outlines of the Thesis	3
2 The Hybrid Survival Curve: Modeling and Interpretations	5
2.1 Introduction	5
2.2 Survival Curves	8
2.2.1 Rectangularization of Survival Curve	8
2.2.2 Existing Types of Survival Curves	11
2.2.3 The Hybrid Survival Curve	14
2.3 Stochastic Models	21
2.3.1 Existing Stochastic Mortality Models	21
2.3.2 A New Class of Stochastic Models	22
2.3.3 The Stochastic Survival Models	24
2.4 Interpreting the Three-Factor Survival Model	27
2.5 Extrapolating the Three Time-Varying Parameters	31
2.6 Conclusion	34
3 Mortality Indexes: Construction, Securitization and Hedging Strategies	37

3.1	Introduction	37
3.2	Constructing Mortality Indexes	41
3.2.1	The New-Data-Invariant Property	43
3.2.2	The Impact of the Proposed Adaptations	45
3.2.3	Finding the Best-Fitting Adapted Model	49
3.2.4	Backtesting the Adapted Models	49
3.2.5	Interpretations of the 3-factor CBD Mortality Indexes	52
3.3	Hedging Longevity Risk with K-forwards	55
3.3.1	Specification of K-forwards	55
3.3.2	Key K-duration	57
3.3.3	Building a Longevity Hedge using Key K-durations	59
3.4	Illustrations of a K-Forward Hedge: Single Cohort	60
3.4.1	The Baseline Results	64
3.4.2	Sampling Risk	66
3.4.3	Sensitivity Tests	68
3.4.3.1	Interest Rate	68
3.4.3.2	Availability of K-forwards	68
3.4.3.3	Age Range	70
3.4.4	Advanced Ages	70
3.5	Generalization to Multiple Cohorts	72
3.5.1	Building K-Forward Hedges	74
3.5.2	Building q-Forward Hedges	75
3.5.3	Hedging Results	78
3.6	Conclusion	80
4	Optimal Design of a Bonus-Malus System	83
4.1	Introduction	83
4.2	Motor Insurance Ratemaking	87
4.2.1	A Priori Ratemaking	87
4.2.2	A Posteriori Ratemaking	88

4.2.2.1	Modeling of Residual Heterogeneity	88
4.2.2.2	Structure of BMS	89
4.2.2.3	Modeling of BMS with Cross-Sectional Data	90
4.2.2.4	Modeling of BMS with Panel Data	92
4.3	Objectives of An Optimal BMS	93
4.3.1	The Three Implications	93
4.3.2	The Two Inadequacy Scenarios	95
4.4	Designing an Optimal BMS with Cross-Sectional Data	96
4.4.1	Determination of Optimal Relativities	97
4.4.2	Design of Transition Rules	101
4.4.3	Varying Transition Rules	103
4.5	Numerical Illustrations: BMS with Cross-Sectional Data	105
4.6	Numerical Illustrations: BMS with Panel Data	109
4.7	Conclusion	113
5	Concluding Remarks and Future Research	115
	Appendix	119
A.	Proof of Identical Maximum Likelihood Estimates for the Two Time- Varying Parameters in Both Stochastic Survival Models	119
	References	123

List of Tables

2.1	The advantages and shortcomings of the period, cohort and hybrid survival curves.	18
3.1	Specifications of the six stochastic mortality models under consideration.	42
3.2	Specifications of the stochastic mortality models that are adapted to satisfy the new-data-invariant property.	45
3.3	The reductions in log-likelihood values due to the proposed adaptations.	47
3.4	Values of the Bayesian Information Criterion (BIC) produced by the five adapted models and Model M5.	48
3.5	The mean absolute percentage error (MAPE) of the forecasted cohort death probabilities over the period of 2000 to 2009 for the cohorts aged 46 to 81 at the beginning of 2000.	51
3.6	The calculated key K-durations and notional amounts of K-forwards. The pension plan and K-forwards are linked to the mortality of English and Welsh males.	64
3.7	The amounts of longevity risk reduction (R) resulting from different combinations of K-forwards and different simulation models. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The VARIMA orders for the time-varying parameters in Models M7* and M5 are (1,1,0) and (5,1,0), respectively. . .	65

3.8	The amounts of risk reduction (R) for pension plans of different initial sizes. The pension plans and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.	68
3.9	The amounts of risk reduction (R) under different interest rate assumptions. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.	69
3.10	The amounts of risk reduction (R) resulting from different spacings between two adjacent key years. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.	69
3.11	The amounts of risk reduction (R) for different age ranges over which the 3-factor CBD mortality indexes are defined. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* (fitted to the corresponding age range) with a VARIMA(1,1,0) process for the time-varying parameters.	70
3.12	The required notional amounts of K-forwards, calibrated by KKD. The pension plan and K-forwards are linked to the mortality of English and Welsh males.	75
3.13	Locations of the key mortality rates of the chosen key cohorts. . . .	76
3.14	The required notional amounts of q-forwards, calibrated by KQD. The pension plan and K-forwards are linked to the mortality of English and Welsh males.	77

3.15	The amounts of risk reduction (R) resulting from the K-forward and q-forward hedges for the hypothetical multi-cohort pension plan. The pension plan and hedging instruments are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.	78
4.1	Transition rules for the simple rules of $-1/+2$ (varying rules of $p = 4$).	106
4.2	Transition rules for the simple rules of $-1/+3$ (varying rules of $p = 3$).	106
4.3	Interaction between a priori and a posteriori ratemakings.	107
4.4	Effectiveness of transition rules.	107
4.5	Optimal relativities for the BMS with cross-sectional data.	108
4.6	Values of maximized log-likelihoods for different combinations of entry level, relativities and transition rules under the Poisson-BMS model.	110
4.7	Optimal relativities for the BMS under the Poisson-BMS model.	111
4.8	Estimated Poisson regression coefficients for the BMS under Poisson-BMS model.	112

List of Figures

2.1	Rectangularization of survival curve.	9
2.2	Three dimensions of the rectangularization of a survival curve. <i>Source:</i> Cheung et al. (2005)	10
2.3	The problem of data incompleteness for cohort survival curves. . . .	13
2.4	The specifications of the death probabilities in the form of age-period cells used to obtain the survival probabilities under the three types of survival curves.	15
2.5	The hybrid survival curves of Swedish males and females in years 1985, 1995 and 2005.	19
2.6	The period, cohort and hybrid survival curves of Swedish males and females.	20
2.7	The logit-transformed annualized survival probabilities and the cor- responding fitted curves using the two-factor survival model and the three-factor survival model for Swedish males and females in years 1985, 1995 and 2005.	24
2.8	The maximum likelihood estimates (MLE) of $\sigma_t^{(i)}$ for $i = 1, 2, 3$ for Swedish males and females.	26
2.9	Values of the Bayesian Information Criterion (BIC) for Swedish males and females.	27
2.10	The hybrid survival curves implied by four different combinations of the time-varying parameters.	30

2.11	Forecasts of $\sigma_t^{(i)}$ for $i = 1, 2, 3$ in years 2010–2030 for Swedish males and females.	33
2.12	The average projections of the hybrid survival curves of in years 2010, 2020 and 2030 for Swedish males and females.	33
3.1	Changes in the logit-transformed mortality curve under different combinations of the 3-factor CBD mortality indexes.	54
3.2	Mortality curves implied by six different combinations of the 3-factor CBD mortality indexes.	54
3.3	Settlement of a K-forward contract at maturity.	56
3.4	The distributions of X and X^* simulated from Model M7*. It is assumed that the underlying population is English and Welsh males and that the pension plan consists of a single pensioner currently aged 65.	66
3.5	The initial demographic structure of the multi-cohort closed pension plan.	73
3.6	The distributions of X and X^* for the hedges that are constructed for both current active members and retirement pensioners in the hypothetical multi-cohort pension plan. The pension plan and hedging instruments are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.	79

Chapter 1

Introduction

1.1 Background

Since the end of World War II, human mortality has been improving consistently across all ages, especially among the elderly. The steady improvements in mortality have led to the rise in human life expectancy that measures the average number of years of remaining lifespan at a given age. For example, United Nations (2011) reported that the global average life expectancy at birth was 48 years during the period of 1950-1955 and increased to 68 years by 2005-2010. Particularly, the average life expectancy of Singaporean has increased from 65.7 years in 1960 to 82.1 years in 2012 according to the World Bank (2014).

In studying the patterns of survival (or mortality) rates, actuaries and demographers analyze survival (or mortality) curves constructed from empirical data. Survival curves allow demographers to investigate several interesting questions regarding the future trends of human longevity. For instance, they could study whether there exists a natural limit to human lifespan or whether there exists a compression of mortality that reduces the variability of age at death. Nonetheless, due to their restricted structures, both period and cohort survival curves are not ideal for analysis purpose in survival modeling.

On the other hand, there is a plethora of research on the modeling of mortality rates because of the important issues in understanding and managing longevity risk. Longevity risk has grown increasingly crucial because pension plans around the world suffer unexpected losses when pensioners live longer than anticipated. When dealing with longevity risk in a pension plan, the traditional method is to sell the liabilities through a pension buy-out, which involves the transfer of all risks (including longevity risk) associated with the liabilities from its balance sheet. Alternatively, the pension plan can choose to execute a pension buy-in by purchasing annuities that are linked to its liabilities to hedge away the longevity risk exposures.

In recent years, we have seen the emergence of a capital market called the life market that allows longevity-linked securities to be traded between counterparties. Building on the idea of using longevity bond (also called survivor bond) to hedge longevity risk (Blake and Burrows, 2001), capital market solutions for longevity risk transfer include longevity bond, longevity swap, mortality catastrophe bond and mortality forward (also known as q-forward). These hedging instruments can be divided into two categories: index-based and customized. For example, in the United Kingdom, a portfolio of index-based q-forward contracts was traded between J.P. Morgan and Lucida PLC in 2008, while ITV executed a customized longevity swap with Credit Suisse in 2011.

Compared to a customized instrument linked to the mortality experience of the hedger's own population, an index-based security that is associated with the mortality experience of a larger pool of lives (such as a national population) is easier to analyze and potentially more liquid. However, the life market has been growing slowly due to the lack of standardized index-based securities. Recently, the Life and Longevity Markets Association (LLMA) was established in London to promote the development of a liquid traded market in longevity- and mortality-related risks.

When insurers are considering the traditional motor insurance ratemaking pro-

cess, two steps are implemented. In the first step, *a priori* risk classification is performed to partition a portfolio of heterogeneous drivers into a number of homogeneous risk classes, in which the drivers that are grouped in the same class would pay the same *a priori* premium. However, the drivers in each risk class would exhibit heterogeneity because there are many hidden risk characteristics that could not be captured in the risk classification. Hence, *a posteriori* ratemaking (also called experience rating) is required in the second step to determine the necessary premium correction or adjustment based on the individual claims experience.

The premium correction can be determined via the greatest accuracy credibility theory first formalized by Bühlmann (1967). An excellent review of the credibility theory can be found in the book by Bühlmann and Gisler (2005). In particular, Dionne and Vanasse (1989) contributed a credibility model to the ratemaking process that integrates both *a priori* and *a posteriori* information on an individual basis. However, this ideal approach is not applicable in practice due to commercial reasons. Instead, insurers develop a bonus-malus system approach to determine the premium adjustment coefficients (formally called relativities). The structure of a bonus-malus system is desirable from a commercial point of view, but its implementation is quite restrictive and thus is subject to a few problematic scenarios.

1.2 Objectives and Outlines of the Thesis

This thesis investigates several modeling issues in longevity and bonus-malus system. In particular, we research on how the weaknesses of the existing approaches or models in these two different topics (related to life and non-life insurance, respectively) can be adapted, and how they may be applied to overcome the great challenges posed to actuaries and demographers.

In Chapter 2, we demonstrate how to construct a new type of survival curves called hybrid survival curve. It possesses the advantages of both period and cohort survival curves. We propose to model the hybrid survival curves using two stochastic survival models, where the estimated time-varying parameters are highly interpretable. Furthermore, we demonstrate how to obtain projected hybrid survival curves by extrapolating the time-varying parameters into the future.

In Chapter 3, we study the construction of mortality indexes using the time-varying parameters in stochastic mortality models. In particular, we investigate how existing mortality models can be adapted to ensure the tractability of indexes by market participants. Based on the mortality indexes created from the chosen mortality model, a standardized mortality-linked security called K-forward can be written to hedge longevity risk. We also contribute a hedging method called key K-duration to calibrate a longevity hedge formed by K-forward contracts.

In Chapter 4, we argue that the restrictive structure of bonus-malus system would cause two inadequacy scenarios associated with a single set of optimal relationships and a set of simple transition rules in its implementation. To mitigate these two problems, we study the optimal design of a bonus-malus system when insurers are working with cross-sectional data. Moreover, the proposal of introducing varying transition rules can also be applied to the existing framework when working with panel data.

We conclude this thesis in Chapter 5 with several suggestions for future research.

Chapter 2

The Hybrid Survival Curve: Modeling and Interpretations

2.1 Introduction

Human life expectancy has been increasing steadily since 1840 and there is no evident sign of slowing down in the near future. In particular, Oeppen and Vaupel (2002) reported that the maximum female life expectancy across a range of countries has risen at an average pace of about three months per year. The rising trend in life expectancy is caused by the mortality improvements across different ages, which have resulted in the so-called rectangularization of survival curve. This concept was coined by Fries (1980), who hypothesized that a decreasing variability of age at death would make a survival curve become more rectangular over time. Since then, a number of indicator measures (see Cheung et al., 2005, for the classifications of measures) were proposed to empirically study the rectangularization of survival curve. There are two existing types of human survival curves – period survival curves and cohort survival curves.

A period survival curve is constructed on the basis of the mortality data in a given period, where the resulting probabilities are called period survival probabili-

ties. However, the period survival probabilities calculated by this approach do not correspond to the actual survival probabilities of real cohorts. In other words, the underlying cohort of a period survival curve is a ‘hypothetical cohort’, so it may not be realistic for certain demographic applications.

In contrast, a cohort survival curve is derived from the mortality data of a specific cohort over time, which implies that the actual survivorship experience of a real cohort can be reflected by the corresponding cohort survival probabilities. However, the construction of a complete cohort survival curve requires the complete cohort mortality data, which would only be available after the death of the last cohort member. As such, cohort survival curve is subject to the problem of data incompleteness due to the lack of timeliness in the cohort mortality data.

The first objective of this chapter is to demonstrate how to construct a new type of survival curves called hybrid survival curve, which possesses the strengths of both period and cohort survival curves. A hybrid survival curve is made up of the survival probabilities experienced by multiple old-age real cohorts (the advantage of cohort survival curves). It is also free from the problem of data incompleteness (the advantage of period survival curves) because it can be constructed before the extinction of a specific cohort. With these advantages, the hybrid survival curve provides an alternative to be used in studying the rectangularization pattern of survival curves.

Specifically, in view of the significant mortality improvements below age 60 that cause the survival probabilities from birth until ages below 60 to be considerably close to 1 and thus relatively unimportant, we choose a fixed starting age of 60 in constructing a hybrid survival curve. Also, we select a fixed ending age of 91 due to the fact that the mortality data beyond age 90 are subject to reliability issues. The collection of actual cohort survival probabilities on a hybrid survival curve resemble the form of cross-sectional survival probabilities as defined by Guillot (2003), though a hybrid survival curve only corresponds to a number of old-age

cohorts instead of a full range of ages. Moreover, our aim here is to study the stochastic modeling of the hybrid survival curve, which is different from Guillot who demonstrated how to calculate the ‘cross-sectional average length of life’ using the cross-sectional survival probabilities from birth.

In recent years, the stochastic modeling of mortality rates that takes into account mortality improvements over time has been analyzed extensively (see, e.g., Cairns et al., 2009, 2011a,b; Dowd et al., 2010a,b) since the pioneering work by Lee and Carter (1992). In particular, a collection of stochastic mortality models were developed to model certain transformed functions of death probabilities or central death rates. For instance, the Lee-Carter model was the first extrapolative model to study the future time trends of logarithmic central death rates in a stochastic setting. However, the modeling of survival probabilities remains largely unexplored, probably because survival probabilities can be expressed in terms of multiplicative products of one minus death probabilities. In this chapter, we examine a different approach and perform the direct modeling of the survival probabilities on the hybrid survival curve.

The second objective of this chapter is to build a new class of stochastic models to model a suitably transformed function of the hybrid survival curve. Following the general structural form of mortality models in Hunt and Blake (2014), we develop two models for the hybrid survival curve. Our proposed models are similar to that of De Jong and Marshall (2007) who performed mortality projection using the Wang transform (see Wang, 2000). The main difference is that we specify a logit-transformation onto the annualized survival probability on the hybrid survival curve as the model response variable instead of using probit-transformed period survival probability.

Our models also resemble the Cairns-Blake-Dowd model by Cairns et al. (2006) except for the different underlying response variables. In more detail, our first proposed model is referred to as two-factor survival model because it consists of

two factors that correspond to the intercept and the slope of the transformed hybrid survival curve. We also consider a simple extension by including a curvature parameter and refer to it as three-factor survival model.

Using the Bayesian Information Criterion (BIC) as our model selection criterion, we find that the three-factor survival model gives lower BIC values for the data sets of Swedish males and females. Besides that, the three time-varying parameters are highly interpretable. Particularly, the trend of each parameter can serve as an indicator for the rectangularization of survival curve. By extrapolating the time-varying parameters with a trivariate random walk with drift, we can readily obtain projected hybrid survival curves.

The remainder of this chapter is organized as follows. Section 2.2 first describes the rectangularization phenomenon and discusses the existing types of survival curves, followed by an illustration of the construction of a hybrid survival curve. Section 2.3 presents the detailed specifications of the two survival models. Section 2.4 describes the interpretations of the time-varying parameters in the three-factor survival model with regards to the rectangularization phenomenon. Section 2.5 demonstrates the extrapolation procedure for the three time-varying parameters in projecting hybrid survival curves. Finally, Section 2.6 concludes this chapter.

2.2 Survival Curves

2.2.1 Rectangularization of Survival Curve

The term ‘rectangularization’ of survival curve was coined by Fries (1980), who hypothesized that the survival curve would become rectangular under declining mortality trends and the fact that human life expectancy at that point in time was believed to have an upper limit of 85 years. This concept is graphically represented in Figure 2.1, where the survival curve tends to shift to the top right corner over

Figure 2.1. Rectangularization of survival curve.

time.

Fries (1980) also conjectured that the distribution of lifetime illness should be narrowly centered at age 85 due to morbidity compression if the exogeneous causes of death were eliminated. As a result, the variability of age at death would decrease over time – a phenomenon known as mortality compression. Strictly speaking, the compression of mortality is different from the compression of morbidity, although Fries (1980) predicted that they would occur simultaneously because the latter is naturally assumed to be having a causal impact on the former.

A number of indicators (for a review, see Cheung et al., 2005) were suggested to study the rectangularization of survival curve and the mortality compression. For example, Wilmoth and Horiuchi (1999) compared 10 highly correlated measures and recommended using the interquartile range (IQR) of the distribution of age at death as the sole indicator of mortality compression. Kannisto (2000) proposed a family of indicators C_α – defined as the shortest age interval covering $\alpha\%$ of the total deaths – and documented the evidences of mortality compression for 22 populations in terms of the declining shortest age intervals. By definition, the value of C_{50} is at most equal to the value of IQR, so the former measure is more effective in detecting the concentration of deaths and can be perceived as an improved

Figure 2.2. Three dimensions of the rectangularization of a survival curve.
Source: Cheung et al. (2005)

measure of IQR.

Furthermore, Cheung et al. (2005) distinguished the rectangularization of a survival curve into three dimensions: horizontalization, verticalization and longevity extension, as depicted in Figure 2.2. Firstly, horizontalization refers to the flattening of the actual survival curve due to the fall of premature deaths and can be measured by the angle β . A lower value of β over time suggests that the survival curve is undergoing horizontalization. Note that premature deaths are defined as the deaths below age $M - 4SD(M+)$ in the left hand tail of the survival curve, where M is the modal age at death normalized to be equal to 1 and $SD(M+)$ is the standard deviation of life durations above M .¹

Secondly, verticalization captures the descending slope of the theoretical survival curve after the elimination of all premature deaths. The degree of verticalization can be measured by the angle θ^* , which is dependent on the concentration of death ages around the modal age M . For this reason, we can interpret the verticalization of survival curve as an equivalent condition to the concentration of death ages (i.e., mortality compression). In this case, a lower value of θ^* indicates verticalization. Moreover, note that the angle θ that corresponds to the slope of the

¹Kannisto (2001) suggested a method to determine the non-normalized value of M in fractions of year to ensure sufficient level of precision.

actual survival curve (taking into account all deaths including premature deaths) is linked to the prolate rectangularity index proposed by Eakin and Witten (1995).

Finally, longevity extension refers to the possible increase of human life durations, as indicated by the changes in the right hand tail of the survival curve $M + 4SD(M+)$. These three measures (β , θ^* and $M + 4SD(M+)$) are free from any arbitrary determination of age and percentile scales (Kannisto, 2001). On the other hand, using the death rates at two suitably chosen ages of 70 and 90, Thatcher et al. (2010) found that the values of M , $SD(M+)$ as well as the life expectancy at the modal age $e(M)$ can be estimated fairly accurately based on the simple logistic model with two parameters (see Thatcher, 1999).

2.2.2 Existing Types of Survival Curves

Here we discuss the advantages and shortcomings of the existing types of survival curves. The conventions below are used throughout the discussion:

- $D_{x,t}$ is observed number of deaths at age x in year t ;
- $E_{x,t}$ is the matching exposures at age x in year t ;
- $m_{x,t}$ is the central death rate at age x in year t ;
- $q_{x,t}$ is the death probability that a person aged x at exact time t will die within a year, which can be written as $q_{x,t} = 1 - e^{-m_{x,t}}$ using the assumption of a constant force of mortality between integer ages²;
- $1 - q_{x,t}$ is the survival probability that a person aged x at exact time t will survive for one year;
- ${}_n s_{x,t}^p$ is the period survival probability that a person aged x at exact time t will survive for n years, based on the one-year death probabilities of multiple

²Another commonly used approximation is $q_{x,t} = \frac{m_{x,t}}{1+0.5m_{x,t}}$, which holds under the assumption that deaths are uniformly distributed (abbreviated as UDD) between integer ages. The proposed approach in this chapter can also be applied using UDD.

cohorts in year t ;

$${}_nS_{x,t}^p = \begin{cases} 1, & n = 0, \\ \prod_{i=0}^{n-1} (1 - q_{x+i,t}), & n \geq 1. \end{cases} \quad (2.1)$$

- ${}_nS_{x,t}^c$ is the cohort survival probability that a person aged x at exact time t will survive for n years until time $t + n$, based on the one-year death probabilities of the specific cohort over time;

$${}_nS_{x,t}^c = \begin{cases} 1, & n = 0, \\ \prod_{i=0}^{n-1} (1 - q_{x+i,t+i}), & n \geq 1. \end{cases} \quad (2.2)$$

A common approach to obtain these probabilities is by using a period life table or a cohort life table. We refer interested readers to Wilmoth and Horiuchi (1999), who analyzed the survival curves of Swedish populations constructed from the life tables on both period and cohort bases. On the other hand, Yue (2012) studied the survival curves of Japan, Sweden and the United States constructed from the raw data in order to avoid the influence of graduation methods on the values in life tables. However, the values obtained from this approach would not be as smooth as the graduated values from the life tables and so do the survival curves. Since our aim is to study the evolution of the hybrid survival curves over the years across the sample period, we calculate the one-year death probabilities and also the resulting survival probabilities from the central death rates defined as $m_{x,t} = \frac{D_{x,t}}{E_{x,t}}$.

There are two existing types of survival curves. Period survival curves are based on the age-specific death probabilities in a given year, whereas cohort survival curves are derived from the death probabilities of a specific cohort over time. Each period survival curve takes into consideration the complete period mortality data involving multiple birth cohorts. This can be seen from Equation (2.1), where the age-specific death probabilities $q_{x+i,t}$'s for $i = 0, \dots, n - 1$ are used to compute

Figure 2.3. The problem of data incompleteness for cohort survival curves.

the period survival probabilities. However, notice that Equation (2.1) involves the multiplication of the one-year survival probabilities for a group of successive cohorts at a fixed period t . Therefore, except for the case of $n = 1$, each period survival probability is effectively a mix of different cohorts' experience and does not correspond precisely to the actual survival probabilities of any real cohorts. Put differently, the underlying cohort of a period survival curve is a hypothetical cohort that is subject to these one-year survival probabilities obtained in a specific period. As such, period survival curve may not be realistic in practice such as in the projection of life expectancy for real cohorts.

In contrast, each cohort survival curve corresponds to a single real cohort by reflecting its actual survivorship experience over time. The reason is that as expressed in Equation (2.2), the death probabilities $q_{x+i,t+i}$'s for $i = 0, \dots, n-1$ used to calculate the cohort survival probabilities correspond to the same cohort over time. However, each cohort survival curve would suffer from the problem of data incompleteness due to the lack of complete cohort mortality data before the death of the last cohort member, as illustrated in Figure 2.3. This limitation also implies that complete cohort survival curve can only be constructed for extinct cohorts after a long period of time.

From the discussions above, it is clear that each type of survival curves has its own advantages and shortcomings. In particular, cohort survival curves are realistic, while period survival curves are not subject to the problem of data incompleteness. These features of survival curves also explain why most of the previous studies examined period survival curves despite the fact that they are hypothetical in nature, mainly because the complete cohort survival curves of interest are unavailable. Nonetheless, it should be noted that the rectangularization pattern can generally be found in both types of survival curves, where period survival curves would exhibit the pattern of rectangularization over successive time periods and cohort survival curves would move towards a rectangular shape across successive birth cohorts. For further discussions on the relationship between period and cohort measures determined from the respective survival curves, we refer interested readers to Goldstein (2006), Goldstein and Wachter (2006), Wachter (2005) and Wilmoth (2005).

2.2.3 The Hybrid Survival Curve

In this subsection, we demonstrate how to construct a new type of survival curves that possesses the advantages of both period and cohort survival curves. Following our construction approach, the proposed survival curve in each year comprises of the actual survival probabilities of a number of old-age cohorts and is not subject to the problem of data incompleteness. This new type of survival curves is called hybrid survival curve because it builds upon the strengths of the existing types of survival curves.

Specifically, a hybrid survival curve in year t comprises of a collection of ${}_nS_{x_0,t}^h$, for $n = 0, 1, 2, \dots, N$, where each of them is defined as the cohort survival probability for a person with a fixed starting age of x_0 at exact time $t - n + 1$ to survive for n years until time $t + 1$ (i.e., the end of year t). Mathematically, the survival probabilities on a hybrid survival curve can be expressed in terms of the one-year

(a) Period survival curve in year t .

(b) Cohort survival curve for cohort born in year $t - x_0$.

(c) Hybrid survival curve in year t .

Figure 2.4. The specifications of the death probabilities in the form of age-period cells used to obtain the survival probabilities under the three types of survival curves.

death probabilities of the corresponding cohort as follows:

$${}_nS_{x_0,t}^h = \begin{cases} 1, & n = 0, \\ \prod_{i=0}^{n-1} (1 - q_{x_0+i,t-n+1+i}), & n = 1, 2, \dots, N. \end{cases} \quad (2.3)$$

where N is the chosen maximum length of survival (in years). The specifications of the death probabilities used to calculate the survival probabilities in Equations (2.1) – (2.3) are depicted in the age-period cells in Figure 2.4. Specifically, each arrowed line represents a particular point on survival curve, where the coordinate components (x, t) along the way for integer ages x refer to the death probabilities $q_{x,[t]}$'s used in obtaining the survival probabilities. In Figure 2.4(a), we observe that a period survival curve is only based on the one-year death probabilities in a single period. By contrast, cohort and hybrid survival curves are dependent on the one-year death probabilities of the corresponding cohorts over time, as shown by the 45° diagonal lines in Figures 2.4(b) and 2.4(c). The main difference between cohort and hybrid survival curves is that the mortality experience up to year t of multiple cohorts (correspond to $n = 1, 2, \dots, N$ in Equation (2.3)) are considered in constructing the hybrid survival curves of year t , whereas the cohort survival curve is only based on a single cohort over time.

We assume that the starting age x_0 is age 60 because the significant mortality improvements in premature ages (i.e., before aging-related deaths become substantial at older ages) – referred to as the horizontalization of survival curve by Cheung et al. (2005) – make the death probabilities below age 60 relatively unimportant. As a result, the survival probabilities from birth until ages below age 60 are moving closer to 1 over time and do not exhibit an obvious varying pattern. Besides that, the mortality data obtained from the Human Mortality Database beyond age 90 may not be sufficiently reliable because the corresponding population counts are estimated by using the extinct cohort method (see, e.g., Vincent, 1951; Bennett and Horiuchi, 1984). Hence, we assume that length of survival n years from age 60

is at most 31 years, that is, $N = 31$, such that the mortality data beyond age 90 are not required in building a hybrid survival curve. Note that the chosen value of N can be larger to extend to ages beyond 90 if not because of the reliability issues associated with the mortality data of older ages.

By definition, we have ${}_0s_{60,t}^h$ equals to 1. In addition, the survival probabilities with a starting age of 60 can be expressed in terms of the corresponding survival probabilities with a starting age of 0 as follows:

$${}_ns_{60,t}^h = {}_ns_{60,t-n+1}^c = \frac{{}_{60+n}s_{0,t-n-59}^c}{{}_{60}s_{0,t-n-59}^c} \quad n = 1, 2, \dots, 31. \quad (2.4)$$

Therefore, the value of ${}_ns_{60,t}^h$ can serve as a good approximation to ${}_{60+n}s_{0,t-n-59}^c$ if the value of ${}_{60}s_{0,t-n-59}^c$ in the denominator of Equation (2.4) is approaching 1 due to consistent mortality improvements for the ages below 60. Moreover, it should be noted that the second part of the relationship in Equation (2.4) is not applicable for hybrid survival curve, that is, ${}_ns_{60,t}^h \neq \frac{{}_{60+n}s_{0,t}^h}{{}_{60}s_{0,t}^h}$ because the expression of ${}_{60}s_{0,t-n-59}^c$ cannot be rewritten as ${}_{60}s_{0,t}^h$.

As shown in Equation (2.3), in calculating the n -year cohort survival probabilities, the historical mortality data of the underlying cohort over the period of $[t - n + 1, t]$ are needed. With a chosen maximum value of $N = 31$, a data sample period of $[t - 30, t]$ is required to construct the hybrid survival curve in year t . Furthermore, the collection of cohort survival probabilities on a hybrid survival curve resemble the form of cross-sectional survival probabilities for a multitude of cohorts, which are used to compute the cross-sectional average length of life (CAL) as proposed by Guillot (2003), though a full range of ages is needed in that case because those probabilities have a fixed starting age of 0 instead of 60.

Since the hybrid survival curve is made up of the actual survival probabilities of different multiple cohorts, it may not necessarily be strictly decreasing, as opposed to the cases of period survival curve and cohort survival curve that monotonically decline. Nonetheless, the monotonically decreasing property of survival curves is

Period curve	Cohort curve	Hybrid curve
✓ rely on the period mortality data in year t , so the complete survival curve can be constructed with a single year of data	× rely on the cohort mortality data for the cohort born in year $t - x$, so the complete survival curve can only be constructed for extinct cohort	✓ rely on the historical cohort mortality data of multiple cohorts up until year t , so the complete survival curve can be constructed with the data available up to year t
× probabilities of hypothetical cohort	✓ probabilities of real cohort	✓ probabilities of real cohort
✓ strictly decreasing	✓ strictly decreasing	× not strictly decreasing

Table 2.1. The advantages and shortcomings of the period, cohort and hybrid survival curves.

not essential in our study given that our aim is to model the evolution of the hybrid survival curves over time. In other words, we are interested in the evolution of the probabilities ${}_n s_{60,t}^h$ over time, instead of studying the survival probabilities across different values of n on a given hybrid survival curve.

The hybrid survival curve, which possesses the strengths of both existing types of survival curves, provides an alternative to study the rectangularization pattern of survival curves. This is because we can search for its empirical evidence by examining the trends of the survival probabilities exhibited by a multitude of real (not hypothetical) cohorts on a collection of hybrid survival curves over time. Furthermore, despite the limitation that the construction of the hybrid survival curve of year t requires the historical data of 30 years prior to year t (unlike the period survival curve that is only based on the period mortality data in year t), this approach is able to avoid the problem faced by cohort survival curve that can only be constructed for a specific extinct cohort. The advantages and shortcomings of the the three different types of survival curves are summarized in Table 2.1.

Figure 2.5. The hybrid survival curves of Swedish males and females in years 1985, 1995 and 2005.

As an illustration, we consider the gender-specific mortality data of Swedish population with a sample period of years 1940 to 2009 obtained from the Human Mortality Database (2013). With this sample period, the earliest hybrid survival curve for each population is of year 1970. The hybrid survival curves of Swedish males and females in three different years are plotted in Figure 2.5. We observe that the respective hybrid survival curves of both populations shift towards the top right corner over time, providing empirical evidence for the rectangularization phenomenon. Particularly, the hybrid survival curves of Swedish females exhibit a more obvious rectangularization pattern than that of Swedish males. Also, the increment in survival probabilities of Swedish males appear to be more comparable across different ages and symmetrical around age 75 such that the corresponding hybrid survival curves become slightly more vertical over time. On the other hand, the increment in survival probabilities of Swedish females from age 60 to earlier ages are considerably smaller than that of older ages, making the steepness of the hybrid survival curves to remain relatively constant over time.

For graphical comparison purpose, the survival probabilities of the three types

Figure 2.6. The period, cohort and hybrid survival curves of Swedish males and females.

of survival curves (with an age range of 60–91 instead of a full age range) are shown in Figure 2.6. In more detail, the illustrated period and hybrid survival curves are of year 1995, whereas the cohort that was born in year 1935 is chosen for the cohort survival curves, such that this cohort would have attained age 60 at the beginning of year 1995 (identical to the illustrated period and hybrid survival curves). In other words, these three types of survival curves give an identical survival probability of surviving from age 60 to age 61.

Due to the consistent mortality improvements for Swedish populations, the survival probabilities from age 60 to ages 62–91 under the period survival curves (which are only based on the one-year death probabilities in year 1995) are higher than that of under the hybrid survival curves (which are based on the one-year death probabilities prior to year 1995). This finding suggests that using period survival curves may overestimate the survival probabilities of the real cohorts under consideration because these multiple cohorts are subject to the mortality experience of up to year 1995 and not solely dependent on the period mortality data in year 1995.

Similarly, because of the declining mortality rates over time, the cohort survival curves (which are based on the one-year death probabilities after year 1995) yield the highest survival probabilities from age 60 to ages 62–91. Nevertheless, it is evident that cohort survival curves are subject to the problem of data incompleteness (in this chapter, the mortality data is only available up until year 2009) and thus may not be appropriate in demographic applications.

2.3 Stochastic Models

2.3.1 Existing Stochastic Mortality Models

The existing stochastic mortality models can be categorized into two broad classes based on the underlying response variable.³ One of them features the modeling of logarithmic central death rates denoted as $\ln(m_{x,t})$. Specifically, the original Lee-Carter model (Lee and Carter, 1992) was the first extrapolative model developed for the logarithmic central death rates, estimated through the method of singular value decomposition (SVD). Since then, this model has been extended to account for certain important features of mortality rates, to jointly model the mortality rates of multiple populations and to improve the model fitting.

For instance, Renshaw and Haberman (2006) incorporated the cohort effect parameters in the Renshaw-Haberman model, which can be seen as a general case for the Age-Period-Cohort model first studied by Osmond (1985). Li and Lee (2005) suggested the common factor model and its augmented version to cater for the joint mortality modeling of both sexes. Other extensions included the use of higher order SVD terms (Booth et al., 2002) and the incorporation of additional parameters (Renshaw and Haberman, 2003) to enhance the goodness-of-fit. More recently, Li (2013) proposed the Poisson common factor model as an extension of

³Here we refer to models that are extrapolative and able to generate future sample paths. Another class of extrapolative model is the P-splines model by Currie et al. (2004), but it cannot be used to generate sample paths.

the augmented common factor model to allow for additional gender-specific factors.

The logit-transformed death probabilities denoted as $\ln(\frac{q_{x,t}}{1-q_{x,t}})$ is the response variable in another class of mortality models. This type of modeling was first featured in the Cairns-Blake-Dowd (CBD) model by Cairns et al. (2006). In more detail, the original CBD model is a two-factor model, in which its two factors correspond to the intercept and the slope of the logit-transformed $q_{x,t}$ for higher ages above age 60. This model is subsequently extended to incorporate cohort effect and quadratic age effect by Cairns et al. (2009).

The variables of interest in these two classes of mortality models are $m_{x,t}$ and $q_{x,t}$. This may explain why the modeling of period survival probabilities was not considered in previous research, possibly because the modeling of $m_{x,t}$ and $q_{x,t}$ provides a one-to-one correspondence to the modeling of period survival probabilities, which can be expressed in terms of multiplicative products of one minus death probabilities $(1 - q_{x,t})$'s for a fixed period t as shown in Equation (2.1). On the other hand, the modeling of cohort survival probabilities was not addressed mainly because it only keeps track of the survival probabilities of a single cohort over time, which implies that the estimated parameters would only offer little practical use for forecasting purposes. As we are going to illustrate in the later sections, the estimated parameters from the stochastic modeling of hybrid survival curve yield useful interpretations about the rectangularization phenomenon and can be extrapolated into the future to obtain projected hybrid survival curves.

2.3.2 A New Class of Stochastic Models

With the newly developed hybrid survival curve, we are essentially analyzing a collection of cohort survival probabilities from age 60. In this subsection, we aim to develop a new class of stochastic models based on the hybrid survival curve, which can be used to complement the existing mortality models in summarizing the different elements of longevity trends.

In building new stochastic models, a useful reference is Hunt and Blake (2014), who generalized all of the models discussed in Section 2.3.1 except the Renshaw-Haberman model into a general structural form.⁴ In more detail, a link function is needed to transform the observed data into an appropriate form for modeling purpose, and a series of age/period/cohort terms – which could be parametric or non-parametric – are utilized to fit the transformed response variable.

In what follows, we demonstrate that by using a suitably chosen transformation, a hybrid survival curve can be transformed into a form suitable for modeling, in which a series of age functions and time-varying parameters can be utilized to model the transformed hybrid survival curve between ages 61-91 in each year.⁵ Empirically, we find that the appropriate combination of transformations for the survival probabilities are as follows:

- we first annualize the survival probabilities, that is, calculate the $1/n$ -th power of the n -year survival probabilities, for $n = 1, 2, \dots, 31$;
- we then apply the logit transformation onto the annualized survival probabilities,

whereby the resulting transformed survival probabilities appear to be a simple function of age x and can be fitted accordingly. Note that the use of logit transformation ensures that both the annualized and original survival probabilities are bounded between 0 and 1 when they are transformed back from the fitted model response variables.

To illustrate, the transformed hybrid survival curves for Swedish males and females as shown in Figure 2.7 suggest that linear and quadratic functions of age x can be used to fit the curves. Accordingly, we propose to model them by using

⁴Although Hunt and Blake (2014) focused on mortality models, the structural form is still relevant and useful for developing survival models, because both types of models are driven by mortality data.

⁵We treat age 60 as the starting age of a hybrid survival curve, where ${}_0s_{60,t}^h = 1$. However, this data point is not included in the model parameter estimation because it is undefined under the chosen logit transformation.

Figure 2.7. The logit-transformed annualized survival probabilities and the corresponding fitted curves using the two-factor survival model and the three-factor survival model for Swedish males and females in years 1985, 1995 and 2005.

linear and quadratic functions of $(x - \bar{x})$ because the fitted curves by using functions of x and $(x - \bar{x})$ are the same although the estimated parameters are different. This specification choice is also motivated by the fact that the values of time-varying parameters would remain the same when the model is extended to include higher order terms, as shown in Appendix A..

2.3.3 The Stochastic Survival Models

Mathematically, the two survival models are written as follows:

$$\begin{aligned} \ln \left(\frac{n s_{60,t}^h{}^{1/n}}{1 - n s_{60,t}^h{}^{1/n}} \right) &= \sigma_t^{(1)} + \sigma_t^{(2)}(x - \bar{x}) \\ &= \sigma_t^{(1)} + \sigma_t^{(2)}(n - \bar{n}), \end{aligned} \quad (2.5)$$

$$\begin{aligned} \ln \left(\frac{n s_{60,t}^h{}^{1/n}}{1 - n s_{60,t}^h{}^{1/n}} \right) &= \sigma_t^{(1)} + \sigma_t^{(2)}(x - \bar{x}) + \sigma_t^{(3)}((x - \bar{x})^2 - s^2) \\ &= \sigma_t^{(1)} + \sigma_t^{(2)}(n - \bar{n}) + \sigma_t^{(3)}((n - \bar{n})^2 - s^2), \end{aligned} \quad (2.6)$$

where $x = 60 + n$ is the attained age at the end of year t if the person survives, for $n = 1, 2, \dots, 31$ and thus the second model expressions in Equations (2.5) and (2.6) follow; $\sigma_t^{(i)}, i = 1, 2, 3$ are the time-varying parameters; $\bar{x} = 76$ and $s^2 = 80$ are the arithmetic averages of x and $(x - \bar{x})^2$ over the sample age range respectively, or equivalently, $\bar{n} = 16$ and $s^2 = 80$ are the arithmetic averages of n and $(n - \bar{n})^2$.

The model in Equation (2.5) is built similarly as the original CBD model except for the different response variables. In particular, the original CBD model is used to model the logit-transformed death probabilities, whereas our model in Equation (2.5) is built to model the logit-transformed annualized cohort survival probabilities for a given year t . For brevity, we refer to this model as two-factor survival model because it involves two estimated parameters. Similarly, the model specified in Equation (2.6) is referred to as three-factor survival model. In addition, our models are also similar to that of De Jong and Marshall (2007) who proposed the use of probit transformation onto the period survival probabilities and subsequently utilized the Wang transform in the mortality projection.

We perform Gaussian maximum likelihood estimation (MLE) to obtain the time-varying parameters $\sigma_t^{(i)}, i = 1, 2, 3$, which means that Equations (2.5) and (2.6) are assumed to be subject to normally distributed error terms. Note that the estimated time-varying parameters from Gaussian MLE are identical to those obtained from the least squares estimation. This approach is chosen because our response variable is dependent on the mortality experience in the previous n years, which implies that it is not possible to perform the commonly used Poisson MLE (see Brouhns et al., 2002) in the form of $D_{x,t} \sim \text{Poisson}(E_{x,t}m_{x,t})$. Due to the same reason, Poisson MLE was not used in the modeling of mortality improvement rates in Haberman and Renshaw (2012), in which the model response variable is expressed in terms of the mortality rates in two adjacent years.

The fitted curves using both models are also shown in Figure 2.7. Graphically, it can be seen that the three-factor survival model provides a much better fitting than

Figure 2.8. The maximum likelihood estimates (MLE) of $\sigma_t^{(i)}$ for $i = 1, 2, 3$ for Swedish males and females.

that of the two-factor survival model, especially for the survival curves of Swedish females because the logit-transformed hybrid curves exhibit an obvious curvature pattern. This finding indicates that the inclusion of the curvature parameter would provide an improved fitting under the presence of significant curvature effect.

The maximum likelihood estimates of $\sigma_t^{(i)}$ for $i = 1, 2, 3$ for both Swedish populations are depicted in Figure 2.8. In the Appendix A., we prove that the maximum likelihood estimates of $\sigma_t^{(i)}$ for $i = 1, 2$ are identical in both models due to the specification of functions of $(x - \bar{x})$. Apart from that, the two-factor survival model is a special case of the three-factor survival model with $\sigma_t^{(3)} = 0$, so the log-likelihood value of the latter is guaranteed to be larger. To strike a balance between goodness of fit and model parsimony, we make use of the Bayesian Information Criterion (BIC) defined as $\text{BIC}_r = -2\hat{l}_r + v_r \log n_d$, where $\hat{l}_r$ is the maximized log-likelihood value of model r , v_r is the effective number of free parameters and n_d is the number of observations in the data. Using BIC as the model selection metric, we prefer a model with a lower BIC value.

As parameter estimation for each of the survival models is carried out for each

Figure 2.9. Values of the Bayesian Information Criterion (BIC) for Swedish males and females.

year separately, the BIC is also calculated for each year⁶, and the values of BIC are shown in Figure 2.9. It is evident that the BIC values produced by the three-factor survival model are consistently better over time for both populations, indicating that the inclusion of curvature parameter in each year significantly improves the model fitting. Besides that, as we shall discuss in Section 2.4, each of the three parameters is highly interpretable.

2.4 Interpreting the Three-Factor Survival Model

The time-varying parameters of the three-factor survival model can be used to summarize the longevity patterns exhibited by a multitude of cohorts in each year. For each year t , the three time-varying parameters represent the estimated intercept, slope and curvature of the transformed hybrid survival curve, respectively. The im-

⁶If there are also age-specific or cohort-related parameters, we cannot compute the value of BIC for each year.

pact of the change in each of these parameters on the cohort survival probabilities can be illustrated by deriving the following partial derivatives:

$$\frac{\partial {}_n s_{60,t}^h}{\partial \sigma_t^{(1)}} = {}_n s_{60,t}^h \cdot \frac{n}{1 + e^{\sigma_t^{(1)} + \sigma_t^{(2)}(n-\bar{n}) + \sigma_t^{(3)}((n-\bar{n})^2 - s^2)}} > 0, \text{ for } 1 \leq n \leq 31, \quad (2.7)$$

$$\frac{\partial {}_n s_{60,t}^h}{\partial \sigma_t^{(2)}} = {}_n s_{60,t}^h \cdot \frac{n(n-\bar{n})}{1 + e^{\sigma_t^{(1)} + \sigma_t^{(2)}(n-\bar{n}) + \sigma_t^{(3)}((n-\bar{n})^2 - s^2)}} \begin{cases} < 0, \text{ for } 1 \leq n < 16, \\ = 0, \text{ for } n = \bar{n} = 16, \\ > 0, \text{ for } 16 < n \leq 31, \end{cases} \quad (2.8)$$

$$\frac{\partial {}_n s_{60,t}^h}{\partial \sigma_t^{(3)}} = {}_n s_{60,t}^h \cdot \frac{n((n-\bar{n})^2 - s^2)}{1 + e^{\sigma_t^{(1)} + \sigma_t^{(2)}(n-\bar{n}) + \sigma_t^{(3)}((n-\bar{n})^2 - s^2)}} \begin{cases} < 0, \text{ for } 17 < n < 25, \\ > 0, \text{ for } 1 \leq n \leq 7, 25 \leq n \leq 31. \end{cases} \quad (2.9)$$

Apart from that, the trends of these time-varying parameters are useful in detecting the rectangularization and the verticalization of survival curve. In what follows, we first describe the alternative ways of interpreting these two concepts before we associate them with the trends of the time-varying parameters.

The rectangularization of survival curve (of any types, including period, cohort and hybrid) means that given the specific age attained $x = 60 + n$, each of the survival probabilities increases over time. This phenomenon can be graphically represented by an upward shift for the survival probabilities over successive time periods. Therefore, we can make use of the positive change of the area under the survival curve over time as an indication of the rectangularization phenomenon.

On the other hand, the verticalization of survival curve is associated with the increasing magnitude of the steepness of survival curve. The steepening of survival curve can be attributed to the concentration of death ages for period and cohort survival curves, which would be reflected through the declining value of angle θ^* as shown in Figure 2.2. However, due to the different structures of survival curves, the quantities used by Cheung et al. (2005) in measuring the degree of verticalization

are not applicable for hybrid survival curves. To overcome this problem, let us suppose, as an illustrative concept, that the degree of verticalization is approximated by the angle formed between: 1) two specific points on the hybrid survival curve that form an approximated tangential line to the curve, and 2) a vertical line. Note that we do not specify how these two points are chosen, which is in contrast to the exact details specified by Cheung et al. (2005) as illustrated in Figure 2.2. The only requirement is that these two points are located around the centre portion of the hybrid survival curve such that it offers a good approximation to the tangential line.

The main reason for the verticalization of hybrid survival curve is that the mortality improvements for the cohorts of younger ages are faster than those of older ages, in such a way that the increment of survival probabilities for younger ages are larger than that for older ages. Graphically, these movements of survival probabilities lead to a steeper tangential line formed by the two specific points used in the approximation and results in a smaller angle, indicating that the survival curve is subject to verticalization pattern.

With these descriptions in place, in what follows, we discuss the interpretation of each time-varying parameter (keeping other things being equal) towards the phenomenons of rectangularization and verticalization of survival curve.

An upward trend in the value of $\sigma_t^{(1)}$ would result in a higher value of n -year survival probabilities for all n , which can be seen from Equation (2.7). This implies that the hybrid survival curve would shift up for each age x and move towards a rectangular shape, as shown in solid and dashed lines in Figure 2.10. It is easy to see that the area under the curve will grow over time. However, there is no clear-cut interpretation regarding the verticalization of survival curve with respect to the change of $\sigma_t^{(1)}$, largely because the resulting increments of survival probabilities vary with age and do not yield straightforward inferences about the steepness of survival curve.

Figure 2.10. The hybrid survival curves implied by four different combinations of the time-varying parameters.

A larger magnitude of negative value of $\sigma_t^{(2)}$ (i.e., a steeper slope) leads to a more vertical hybrid survival curve, as displayed in solid and dotted lines in Figure 2.10. In more detail, the verticalization is due to the change of survival probabilities in different directions (according to the specific age ranges) with respect to the change of $\sigma_t^{(2)}$, as shown in Equation (2.8). This explanation is aligned with the earlier descriptions about the faster mortality improvements for the cohorts of younger ages than those of older ages. Also, it should be noted that empirically the slope of the logit-transformed curve is negative because the mortality rates at older ages are generally higher, so the logit-transformed annualized survival probabilities are lower for larger value of n (surviving to older ages). Nevertheless, the opposed changes of survival probabilities based on the range of x does not offer simple interpretations about the change of the area under the curve and hence on the rectangularization phenomenon.

A smaller value of $\sigma_t^{(3)}$ leads to a larger area under the curve, as depicted in the dotted and dashed dotted lines in Figure 2.10. The reason for this interpretation is that the directional sensitivity of survival probabilities towards the change of $\sigma_t^{(3)}$ behaves according to the range of x as shown in Equation (2.9) and the magnitude

of changes are larger in the range of $67 < x < 85$. Moreover, it appears that the trend of $\sigma_t^{(3)}$ does not offer clear-cut conclusions regarding the verticalization of survival curve. This is possibly because the implied changes of survival probabilities are relatively smaller due to the smaller magnitude change of $\sigma_t^{(3)}$, such that the steepness of the hybrid survival curve remains unaffected.

Hence, we could interpret an upward trend of $\sigma_t^{(1)}$ and a downward trend of $\sigma_t^{(3)}$ as the indicators for the rectangularization pattern, whereas the verticalization of survival curve can be inferred from the declining value of $\sigma_t^{(2)}$. These interpretations, together with the estimated parameters as shown in Figure 2.8, are aligned with the earlier observations in Figure 2.5:

- The value of $\sigma_t^{(1)}$ ($\sigma_t^{(3)}$) for Swedish females is increasing (decreasing) at a faster pace than those of Swedish males, suggesting that the hybrid survival curves of Swedish females exhibit a more obvious rectangularization pattern over time.
- The value of $\sigma_t^{(2)}$ for Swedish males is subject to a downward trend, indicating that the hybrid survival curves of Swedish males tend to verticalize over time.
- For Swedish females, the upward trend in the value of $\sigma_t^{(2)}$ and the offsetting effect from the increment of $\sigma_t^{(1)}$ (which shifts up survival probabilities for all ages at varying extents) can be used to explain why the steepness of the hybrid survival curves for Swedish females remains relatively constant.

2.5 Extrapolating the Three Time-Varying Parameters

Similar to other stochastic mortality models, the three-factor survival model is an extrapolative model, whereby the time-varying parameters can be extrapolated to obtain future forecasted values. In this section, we demonstrate how extrapolation

and forecasting can be accomplished.

It is common to model the time-varying parameters in a stochastic mortality model with a univariate or multivariate random walk with drift. For example, this approach was used in the original work of Lee and Carter (1992) and Cairns et al. (2006). Following this approach, we utilize a trivariate random walk with drift for the time-varying parameters in the three-factor survival model. Specifically, the time-varying parameters $\sigma_t^{(i)}, i = 1, 2, 3$ are jointly written in the form of a trivariate time series process $\boldsymbol{\sigma}_t = (\sigma_t^{(1)}, \sigma_t^{(2)}, \sigma_t^{(3)})'$ as follows:

$$\boldsymbol{\sigma}_t = \boldsymbol{\sigma}_{t-1} + \boldsymbol{\mu} + \boldsymbol{C}\boldsymbol{Z}_t \quad (2.10)$$

where $\boldsymbol{\mu}$ is a constant 3×1 vector of drift coefficients, $\boldsymbol{C}$ is a constant 3×3 lower triangular matrix and $\boldsymbol{Z}_t$ is a 3×1 vector of standard normal variables. In more detail, note that the innovation terms $\boldsymbol{\epsilon}_t$'s defined as $\boldsymbol{\epsilon}_t = \boldsymbol{\sigma}_t - \boldsymbol{\sigma}_{t-1} - \boldsymbol{\mu}$ are independently and identically distributed trivariate normal vectors with mean zero and constant variance-covariance matrix $\boldsymbol{\Sigma}$, and the matrix $\boldsymbol{C}$ is obtained from the Cholesky decomposition of the matrix $\boldsymbol{\Sigma}$.

The extrapolation of the three-factor survival model is performed as follows. The MLE of $\boldsymbol{\sigma}_t$ for $t = 1970, 1971, \dots, 2009$ in Figure 2.8 are fitted to the process specified in Equation (2.10) to estimate the values of $\boldsymbol{\mu}$ and $\boldsymbol{C}$. We then simulate 5000 realizations of $\boldsymbol{Z}_t$ to obtain the forecasts of $\boldsymbol{\sigma}_t$ in years 2010–2030. For both Swedish males and females, the mean forecasts and the 95% confidence intervals of the time-varying parameters $\sigma_t^{(i)}, i = 1, 2, 3$ are displayed in Figure 2.11. It can be seen that the mean forecasts of the time-varying parameters follow their respective trends in the past due to the specification of random walk in Equation (2.10). Moreover, for each parameter, the forecasting uncertainties in the future periods are captured by their 95% confidence intervals that fan out over time around the underlying trend.

With these forecasted parameters, we can project the hybrid survival curve

Figure 2.11. Forecasts of $\sigma_t^{(i)}$ for $i = 1, 2, 3$ in years 2010–2030 for Swedish males and females.

Figure 2.12. The average projections of the hybrid survival curves of in years 2010, 2020 and 2030 for Swedish males and females.

into the future. Specifically, the average projections of the hybrid survival curve in year t is determined by averaging the survival probabilities ${}_n s_{60,t-n+1}^c$ for each $n = 1, 2, \dots, 31$ obtained from the 5000 simulated scenarios. As an illustration, the projected hybrid survival curves of Swedish males and females are shown graphically in Figure 2.12. We observe that the hybrid survival curves of both populations are expected to rectangularize in the future periods.

Furthermore, it can be seen that the extent of rectangularization is faster for the hybrid survival curves of Swedish females than those of Swedish males, largely due to the faster rising (declining) trends in the forecasted values of $\sigma_t^{(1)}$ ($\sigma_t^{(3)}$) for Swedish females as depicted in Figure 2.11. This finding is aligned with the interpretations presented in Section 2.4 regarding the trends of $\sigma_t^{(1)}$ and $\sigma_t^{(3)}$ to indicate the rectangularization pattern. On the other hand, it appears that the verticalization pattern can be found in the hybrid survival curves for Swedish males. This observation is mainly because of the downward trend exhibited by the extrapolated value of $\sigma_t^{(2)}$ for Swedish males in Figure 2.11, consistent with the inference discussed in Section 2.4 that associates the trend of $\sigma_t^{(2)}$ to verticalization pattern.

2.6 Conclusion

In view of the respective shortcomings of the existing types of survival curves, in this chapter we illustrated the construction of a hybrid survival curve that comprises of the actual cohort survival probabilities and is free from the problem of data incompleteness. Because of these advantages, hybrid survival curve is a potentially better alternative to both period and cohort survival curves in future empirical investigation on the rectangularization phenomenon.

We developed two stochastic models to fit the hybrid survival curve. In particular, the time-varying parameters in the three-factor survival model produce a better fitting to the transformed hybrid survival curve than its two-factor counter-

part. Moreover, each of the three parameters is highly interpretable especially with regards to the rectangularization and the verticalization patterns of survival curves. On top of that, we demonstrated the extrapolation of the time-varying parameters into the future via a trivariate random walk with drift to obtain projected hybrid survival curves.

In extrapolating the time-varying parameters into the future, we employed a simple trivariate random walk with drift process. However, as shown by Chan et al. (2014), this approach does not capture any serial- or cross-correlations that may exist within the time-varying parameters of stochastic mortality models. Instead, they suggested to utilize a general class of Vector Autoregressive Integrated Moving Average (VARIMA) models for these parameters. Note that the specification of time series model would directly affect the resulting forecasted values. For instance, Cairns et al. (2011a) found that the use of different ARIMA models for the cohort effect parameters in certain stochastic mortality models produce ex ante forecasts with different levels of plausibility. In future research, it would be interesting to empirically investigate which of these processes yield projected values that are more plausible and are more appropriate to be used for forecasting purposes.

Chapter 3

Mortality Indexes: Construction, Securitization and Hedging Strategies

3.1 Introduction

Pension plan sponsors can mitigate their longevity risk exposures by trading securities that are linked to future realized mortality. To date, the market for such securities is still in its infancy and has yet to overcome a number of challenges. As Blake et al. (2013) pointed out, one of these challenges is the creation of homogeneous and transparent instruments, which allow the market to concentrate liquidity. A significant step in overcoming this challenge is to develop tractable mortality indexes, upon which standardized mortality-linked securities can be written.

Existing mortality indexes produced by investment banks such as J.P. Morgan, Deutsche Börse and Credit Suisse are constructed without using any model. One disadvantage of this construction approach is that the information content of each index is limited, which means the information reflected by an index is either highly aggregate (e.g., the life expectancy at birth) or specific (e.g., the death probability

at a certain age).⁷ It follows that a large number of such indexes would be needed to effectively hedge longevity risk, which arises from complex and non-parallel shifts in the underlying mortality curve. This problem hinders market development, because liquidity would be diluted across the large spectrum of indexes.

To improve the information content of a mortality index, one may use a model-based construction method, in which mortality indexes are developed from the time-varying parameters in a stochastic mortality model, such as the Lee-Carter model (Lee and Carter, 1992) and the Cairns-Blake-Dowd (CBD) model (Cairns et al., 2006). The model-based approach was first studied by Chan et al. (2014), who argued that the model on which index construction is based must satisfy the new-data-invariant property, which means that when the model is updated with the availability of new mortality data, the mortality indexes (constructed from the model's time-varying parameters) for the previous years would not be affected. This property is crucially important, because once made public, an index value cannot (and should not) be changed. Chan et al. (2014) found that among the six stochastic models documented by Dowd et al. (2010a), the original CBD model (also called Model M5) is the only model that satisfies the new-data-invariant property.

The first objective of this chapter is to further investigate the construction of mortality indexes from stochastic mortality models. We begin with a discussion on how models other than the original CBD model can be adapted to satisfy the new-data-invariant property. We then evaluate the fit of the adapted models to the mortality data from 8 national populations on the basis of the Bayesian Information Criterion (BIC). Among the collection of adapted models, we find that the adapted Model M7 (the Cairns-Blake-Dowd model with cohort and quadratic age effects) yields the best BIC values when fitted to data over the age range of 40-90 for various populations. The model also performs the best in a backtest that compares

⁷The mortality index of Credit Suisse is based on the life expectancy at birth of the US population, while that of J.P. Morgan is based on the death probabilities of four national populations.

projected values with actual values. On top of that, the time-varying parameters in the model are highly interpretable and are able to reflect the varying age pattern of mortality improvement. For these reasons, we propose to use the three time-varying parameters in the adapted Model M7 jointly as mortality indexes. We call these indexes the 3-factor CBD mortality indexes.

Our second objective is to develop standardized securities written on the 3-factor CBD mortality indexes. In particular, we explain how a security called K-forward, proposed as a concept by Chan et al. (2014), can be used to hedge the longevity risk exposures of a pension plan. The structure of a K-forward is identical to that of a q-forward (see, e.g., Coughlan, 2009), except that the reference rates to which the contracts are linked are different. In more detail, the payoff from a q-forward depends on the realized death probability at a reference age in a reference year, whereas that from a K-forward depends on a realized 3-factor CBD mortality index in a reference year. Compared to a q-forward, a K-forward is an even simpler building block, because its reference rate contains only one parameter (the reference year) instead of two.

To ensure that a hedge formed by standardized instruments is effective, there is a need to calibrate it. Following the lines of Cairns (2011), Cairns et al. (2008), Coughlan et al. (2007), Li and Luo (2012), Lin and Tsai (2013), Plat (2009, 2010), Tsai et al. (2010) and Wang et al. (2010), we contribute a measure called key K-duration, which measures a liability's price sensitivity to a specific segment in the time trend of a 3-factor CBD mortality index. The required notional amounts of K-forwards can be determined readily by equating the key K-durations of the portfolio of K-forwards and the liability being hedged. The key K-duration measure is parallel to Li and Luo's (2012) key q-duration, which measures the change in a liability's value due to a small change in a death probability. It also has a close resemblance to Cairns' (2011) approximate deltas (with respect to the time-varying parameters in the original CBD model) and to Plat's (2009) minimum variance hedge ratios that are derived by considering the shifts of the two time-varying

parameters in the original CBD model.

The calculation of key K-durations does not require any simulation, so the execution of the proposed hedging strategy is quick and requires minimal computational effort. Our numerical illustrations indicate that the proposed hedging strategy is effective in reducing a portfolio's longevity risk exposure, even if parameter uncertainty and sampling risk are taken into account. Furthermore, although the mortality indexes are derived from the adapted Model M7, our proposed hedging strategy also works well under scenarios simulated from other stochastic mortality models, suggesting that the success in hedging is largely independent of the simulation model and is likely to be achievable in practice.

The advantage of our proposed hedging strategy is particularly apparent when the portfolio being hedged involves individuals who were born in different years. First, a K-forward hedge is easier to execute in comparison to a q-forward hedge, which requires the hedger to determine the key cohorts in the portfolio. Second, our numerical illustrations suggest that compared to a q-forward hedge, a K-forward hedge giving a comparable hedge effectiveness involves a smaller number of securities. This helps the market to concentrate liquidity, thereby facilitating market development.

The remainder of this chapter is organized as follows. Section 3.2 discusses the construction of mortality indexes using common stochastic mortality models. Section 3.3 specifies a K-forward contract, defines the key K-duration measure and details the proposed hedging strategy. Section 3.4 illustrates the proposed methods with a hypothetical pension plan involving one single cohort and investigates important issues such as sampling risk. Section 3.5 presents the generalization to multiple birth cohorts. Finally, Section 3.6 concludes this chapter.

3.2 Constructing Mortality Indexes

In this section, we revisit the problem of model-based mortality index construction, which was previously studied by Chan et al. (2014). The conventions below are used throughout the discussion:

- $m_{x,t} = \frac{D_{x,t}}{E_{x,t}}$ is the central death rate at age x in year t ;
- $D_{x,t}$ is observed number of deaths at age x in year t ;
- $E_{x,t}$ is the matching exposures at age x in year t ;
- $q_{x,t}$ is the probability that a person aged x at time t will die between time t and $t + 1$;
- $\beta_x^{(i)}, i = 1, 2, 3$, are age-specific parameters;
- $\kappa_t^{(i)}, i = 1, 2, 3$, are time-varying parameters;
- $\gamma_c^{(i)}, i = 3, 4$, where $c = t - x$ denotes year of birth, are cohort-related parameters;
- n_a is the number of ages covered in the sample age range;
- $\bar{x}$ is the mean age over the sample age range;
- $\hat{\sigma}_x^2$ is the mean of $(x - \bar{x})^2$ over the sample age range.

The candidate models under consideration are the six stochastic mortality models discussed by Dowd et al. (2010a).⁸ The specifications of the six models are displayed in Table 3.1. In fitting these models (except Model M5), we need to impose identifiability constraints to stipulate parameter uniqueness. The number of identifiability constraints needed for each model is displayed in parentheses in Table 3.1. We refer interested readers to Cairns et al. (2009) for a deeper discussion on the identifiability constraints.

⁸These six models constitute a subset of the eight stochastic mortality models (labeled as Model M1–M8) examined by Cairns et al. (2009). Model M4 is excluded because of its poor model fitting and its inability to generate future sample paths. Model M8 is not considered because it generates implausible ex ante forecasts, as shown by Cairns et al. (2011b).

Model M1: The Lee-Carter Model (Lee and Carter, 1992)	
$\ln(m_{x,t}) = \beta_x^{(1)} + \beta_x^{(2)} \kappa_t^{(2)}$	(2 constraints)
Model M2: The Renshaw-Haberman Model (Renshaw and Haberman, 2006)	
$\ln(m_{x,t}) = \beta_x^{(1)} + \beta_x^{(2)} \kappa_t^{(2)} + \beta_x^{(3)} \gamma_{t-x}^{(3)}$	(4 constraints)
Model M3: The Age-Period-Cohort Model (Osmond, 1985)	
$\ln(m_{x,t}) = \beta_x^{(1)} + n_a^{-1} \kappa_t^{(2)} + n_a^{-1} \gamma_{t-x}^{(3)}$	(3 constraints)
Model M5: The Original Cairns-Blake-Dowd (CBD) Model (Cairns et al., 2006)	
$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x})$	(No constraint)
Model M6: The CBD Model with a Cohort Effect Term (Cairns et al., 2009)	
$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_{t-x}^{(3)}$	(2 constraints)
Model M7: The CBD Model with Cohort Effect and Quadratic Age Effect Terms (Cairns et al., 2009)	
$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \hat{\sigma}_x^2) + \gamma_{t-x}^{(4)}$	(3 constraints)

Table 3.1. Specifications of the six stochastic mortality models under consideration.

3.2.1 The New-Data-Invariant Property

Chan et al. (2014) argued that the model on which index construction is based must satisfy the new-data-invariant property. In more detail, the time-varying parameters in the model must remain unchanged when the model is fitted to additional years of data. This property is crucial to the tractability of the mortality indexes, because without it historical index values will need to be revised every time when new mortality data becomes available, which is not desirable for the purpose of constructing model-based indexes. Furthermore, if an index value at a particular time point is not fixed, it would be difficult, if not impossible, to define the payoff from an instrument written on the index.

Chan et al. (2014) provided two sufficient conditions for the new-data-invariant property to hold. The first condition is related to the log-likelihood. Assuming $D_{x,t} \sim \text{Poisson}(E_{x,t}m_{x,t})$, the log-likelihood of a stochastic mortality model can be expressed as

$$l = \sum_{t=t_{\text{start}}}^{t_{\text{end}}} \sum_{x=x_0}^{x_1} D_{x,t} \ln(E_{x,t}m_{x,t}) - E_{x,t}m_{x,t} - \ln(D_{x,t}!) = \sum_{t=t_{\text{start}}}^{t_{\text{end}}} \lambda(t), \quad (3.1)$$

where $[x_0, x_1]$ is the sample age range, $[t_{\text{start}}, t_{\text{end}}]$ is the sample period and $\lambda(t)$ represents the contribution to the log-likelihood from data of year t . We require the log-likelihood to be separable, that is, for $t \neq s$, $\lambda(t)$ and $\lambda(s)$ do not contain any common free parameters. This condition implies that we can estimate the mortality indexes in each year independently. The second condition is associated with the identifiability constraints. In particular, we require that the estimation process involves no identifiability constraint, because otherwise when the model is fitted to additional years of data, parameters for the previous years would be changed to conform to the constraint imposed on them. Among the six candidate models we consider, Model M5 is the only model that meets these two conditions.

In what follows, we discuss how the other five models can be adapted to satisfy

the new-data-invariant property. The way we adapt a model depends on whether the model contains age-specific parameters, cohort effect parameters, or both.

To make Model M1 (which contains age-specific parameters) meet the new-data-invariant property, we could keep the age-specific parameters $\beta_x^{(i)}$, $i = 1, 2$, fixed when we update the model. In using this adaptation, the log-likelihood becomes separable and identifiability constraints are no longer needed. We denote the adapted version of Model M1 by Model M1*.

For Models M6 and M7 (which contain cohort effect parameters), we could first estimate the time-varying parameters $\kappa_t^{(i)}$, where $i = 1, 2$ for Model M6 and $i = 1, 2, 3$ for Model M7, and then the cohort effect parameters $\gamma_{t-x}^{(i)}$, where $i = 3$ for Model M6 and $i = 4$ for Model M7, from the residuals. In this way, the estimation of the time-varying parameters would be based on a separable log-likelihood and requires no identifiability constraint. The time-varying parameters in the adapted models thus possess the new-data-invariant property. We denote the adapted versions of Models M6 and M7 by Models M6* and M7*, respectively.

For Models M2 and M3 (which contain both age-specific and cohort effect parameters), we need both of the previously mentioned adaptations. Every time we update the models, we keep the age-specific parameters fixed, estimate the time-varying parameters first and then the cohort effect parameters from the residuals. In this way, the time-varying parameters in the adapted model would meet the new-data-invariant property. We denote the adapted versions of Models M2 and M3 by Models M2* and M3*, respectively.

The five adapted candidate models are specified in Table 3.2, where $\beta_x^{(i)*}$, $i = 1, 2, 3$, and $\gamma_{t-x}^{(i)*}$, $i = 3, 4$, are the parameters that are subject to the proposed adaptations.

Model M1*: The adapted version of model M1	
$\ln(m_{x,t}) = \beta_x^{(1)*} + \beta_x^{(2)*} \kappa_t^{(2)}$	(No constraint)
Model M2*: The adapted version of model M2	
$\ln(m_{x,t}) = \beta_x^{(1)*} + \beta_x^{(2)*} \kappa_t^{(2)} + \beta_x^{(3)*} \gamma_{t-x}^{(3)*}$	(No constraint)
Model M3*: The adapted version of model M3	
$\ln(m_{x,t}) = \beta_x^{(1)*} + n_a^{-1} \kappa_t^{(2)} + n_a^{-1} \gamma_{t-x}^{(3)*}$	(No constraint)
Model M6*: The adapted version of model M6	
$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_{t-x}^{(3)*}$	(No constraint)
Model M7*: The adapted version of model M7	
$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \hat{\sigma}_x^2) + \gamma_{t-x}^{(4)*}$	(No constraint)

Table 3.2. Specifications of the stochastic mortality models that are adapted to satisfy the new-data-invariant property.

3.2.2 The Impact of the Proposed Adaptations

An adapted model yields a reduced goodness-of-fit in comparison to the corresponding original model, because the estimation of an adapted model is conditioned on some pre-estimated parameters. Specifically, the estimation of the time-varying parameters in Models M1*, M2* and M3* is conditioned on the fixed age-specific parameters, which, in practice, can be determined by data over a restricted sample period, say $[t_{\text{start}}, t_{\text{mid}}]$, where $t_{\text{mid}} < t_{\text{end}}$.⁹ On the other hand, the estimation of the cohort effect parameters in Model M2*, M3*, M6* and M7* is conditioned on the estimated time-varying parameters.

To evaluate the impact of the proposed adaptations, we compare the goodness-of-fit to the data over the period of $[t_{\text{mid}} + 1, t_{\text{end}}]$ produced by the original models and the adapted models (whose fixed age-specific parameters, if any, are determined by the data over the period of $[t_{\text{start}}, t_{\text{mid}}]$). The metric we use is $\hat{l}_r - \hat{l}_{r*}$,

⁹Parameter estimation over the restricted sample period of $[t_{\text{start}}, t_{\text{mid}}]$ is still subject to the usual identifiability constraints.

the difference between the maximized log-likelihoods for model r and its adapted version r^* .

We consider gender-specific and unisex mortality data from 8 national populations across different geographical regions, including Australasia (Australia), Asia (Japan), the Nordic region (Norway, Sweden), Western Europe (England and Wales, France) and North America (Canada, United States). All data are obtained from the Human Mortality Database (2013).

For each data set, we use a sample period of 1950 to 2009. We consider a sample age range of 40-90, because models in the CBD family may not fit the accident hump at younger ages well and the data beyond age 90 are subject to reliability problems. With this age range, we have $\bar{x} = 65$ and $\hat{\sigma}_x^2 = \frac{650}{3}$. Furthermore, in fitting the models that are built for $q_{x,t}$ (Models M5, M6, M6*, M7 and M7*), we assume a constant force of mortality between integer ages, which implies that $m_{x,t} = -\ln(1 - q_{x,t})$.

In evaluating the goodness-of-fit, we set t_{mid} to be year 1979, which means the evaluation is based on the fit to the mortality data in the period of 1980 to 2009. The results are tabulated in Table 3.3. From Table 3.3(a), it can be seen that Model M7* consistently gives the smallest reductions in log-likelihood values for all gender-specific populations, indicating that the performances of Models M7* and M7 are similar. One possible explanation is that the three time-varying parameters in Model M7 capture the majority of the historical variations in mortality, so that whether or not the cohort effect parameters are conditionally estimated has little impact on the overall goodness-of-fit. However, other adapted models yield a significantly worse fit in comparison to their original versions. We obtain similar results in Table 3.3(b) when we analyze the unisex mortality data, except that Model M6* yields the smallest reduction in goodness-of-fit for English and Welsh population.

Population	Males					Females				
	M1*	M2*	M3*	M6*	M7*	M1*	M2*	M3*	M6*	M7*
Australia	8088	18309	7815	330	37	4332	6749	4400	2800	66
Canada	58660	24619	7511	161	33	3566	6971	3020	3429	465
England and Wales	101508	19005	11124	1397	160	15627	12850	7505	3811	477
France	42553	56258	11574	15319	4481	20584	33573	15099	51211	1661
Japan	155455	78768	11988	13182	2419	429592	112186	19815	54196	1512
Norway	2689	4063	747	47	4	2259	785	871	1460	67
Sweden	39162	28324	2376	423	25	2396	5323	2859	3418	193
United States	182036	104121	48958	12930	2243	34950	103453	21082	34943	9257

(a) Gender-specific mortality data										
Population	M1*	M2*	M3*	M6*	M7*	M1*	M2*	M3*	M6*	M7*
Australia	25434	42779	10398	1221	45					
Canada	45146	28530	7330	962	468					
England and Wales	129581	35387	11226	360	683					
France	25655	17954	20723	58690	8413					
Japan	686854	189008	22953	56389	5834					
Norway	5607	5516	1115	681	68					
Sweden	12401	29057	4202	2250	140					
United States	169731	320358	45385	29884	10210					

(b) Unisex mortality data										
Population	M1*	M2*	M3*	M6*	M7*	M1*	M2*	M3*	M6*	M7*
Australia	25434	42779	10398	1221	45					
Canada	45146	28530	7330	962	468					
England and Wales	129581	35387	11226	360	683					
France	25655	17954	20723	58690	8413					
Japan	686854	189008	22953	56389	5834					
Norway	5607	5516	1115	681	68					
Sweden	12401	29057	4202	2250	140					
United States	169731	320358	45385	29884	10210					

Table 3.3. The reductions in log-likelihood values due to the proposed adaptations.

Population	Males						Females					
	M1*	M2*	M3*	M5	M6*	M7*	M1*	M2*	M3*	M5	M6*	M7*
Australia	33354	53082	31747	20385	17138	16272	24474	29538	24590	27627	22123	15941
Canada	136685	66715	32133	20052	18401	17344	24078	30818	22455	29275	24758	17469
England and Wales	228698	56788	41008	35996	22608	18796	55213	43998	33794	39992	28235	19211
France	110739	131577	43983	89252	63790	29271	62180	85271	49907	200528	128629	25472
Japan	343090	179063	49214	81324	58414	29708	882156	244212	61078	232404	135556	27762
Norway	19229	22500	15407	13841	13743	13722	17608	15334	15106	17995	16244	13392
Sweden	93526	72312	19992	16841	16000	15221	19541	26013	20757	26919	22110	15434
United States	408199	233034	129313	128091	78291	35659	104268	228801	68072	178509	127756	44833

(a) Gender-specific mortality data						
Population	M1*	M2*	M3*	M5	M6*	M7*
Australia	69358	102991	38208	26870	21192	17344
Canada	110705	75415	32674	25446	22788	19137
England and Wales	291609	90559	43216	45335	25732	20914
France	80709	55950	65345	244553	164331	41429
Japan	1415186	401189	74254	233037	152169	40954
Norway	26029	26391	17250	17238	16316	14785
Sweden	41059	74765	24881	24938	21174	16477
United States	394122	667872	128688	231285	152982	57999

(b) Unisex mortality data

Table 3.4. Values of the Bayesian Information Criterion (BIC) produced by the five adapted models and Model M5.

3.2.3 Finding the Best-Fitting Adapted Model

In this subsection we compare the performances of the five adapted models and Model M5 (which requires no adaptation). The metric we use is the Bayesian Information Criterion (BIC), defined by $\text{BIC}_r = -2\hat{l}_r + v_r \ln n_d$. The BIC indicates a model's goodness-of-fit (reflected by the maximized log-likelihood $\hat{l}_r$), taking into account the number of observations n_d and the effective number of free parameters v_r in the model. We prefer a model with a lower BIC value.

The BIC values produced by the five adapted models and Model M5 are depicted in Table 3.4. We observe that Model M7* gives the best BIC values for all gender-specific and unisex data sets under consideration. This is possibly because, as previously mentioned, the adaptation involved in Model M7* does not have much impact on the model's performance. Another possible explanation is that the quadratic age effect term (which is unique to Model M7*) has a significant explanatory power.

3.2.4 Backtesting the Adapted Models

As remarked by Dowd et al. (2010a), it is quite possible for a model to produce poor ex post out-of-sample forecasts that differ significantly from actual realized outcomes despite the fact that a good in-sample fitting to historical data and plausible ex ante forecasts are provided by the model. Hence, we further analyze the forecasting performances of the five adapted models and Model M5 via the expanding horizon backtests proposed by Dowd et al. (2010a).

The backtesting procedure is summarized as follows:

1. The fixed age-specific parameters of Models M1*, M2* and M3* are estimated using the data over the period of 1950 to 1979.
2. A fixed look-back window of 1980 to 1999 is used to fit the mortality data using the five adapted models and Model M5.

3. For Models M1*, M2* and M3*, the single estimated time-varying parameter is modeled by a univariate random walk with drift. For Models M5, M6* and M7*, the time-varying parameters are modeled by a multivariate random walk with drift.
4. A fixed look-forward window of 2000 to 2009 is used to evaluate the forecasting performance. In our backtesting work, we generate forecasts of cohort death probabilities for the cohorts aged 46 to 81 at the beginning of the look-forward window.¹⁰
5. Simulations are performed to generate forecasts of cohort death probabilities in the 10-year look-forward window.
6. To evaluate the forecasting accuracy, we compute the mean absolute percentage error (MAPE) of the forecasted cohort death probabilities for the cohorts aged 46 to 81 at the beginning of 2000 as follows:

$$\text{MAPE} = \frac{1}{10 \times 36} \sum_{t=1}^{10} \sum_{x=45}^{80} \left| \frac{q_{x+t,1999+t} - \hat{q}_{x+t,1999+t}}{q_{x+t,1999+t}} \right|,$$

where $\hat{q}_{x,t}$ represents the mean forecast of $q_{x,t}$. We prefer a model with a lower MAPE value.

The backtesting results are shown in Table 3.5. On the basis of the MAPEs, we find that Model M7* produces the most accurate forecasts for 11 out of the 16 gender-specific data sets and 6 out of the 8 unisex data sets under consideration.

Judging from these results, Model M7* is the most suitable model for constructing mortality indexes. Furthermore, as we are going to discuss in the next subsection, the time-varying parameters in Model M7* are easy to interpret and are able to explain the varying age pattern of mortality improvement. For these reasons, we propose to use the three time-varying parameters, $\kappa_t^{(1)}$, $\kappa_t^{(2)}$ and $\kappa_t^{(3)}$, in Model M7* as mortality indexes. We call these indexes the 3-factor CBD mortality

¹⁰We do not need to specify any process to extrapolate the cohort effect parameters, because the years of birth involved in the backtest are covered by the data sample.

(a) Gender-specific mortality data

Population	Males					Females						
	M1*	M2*	M3*	M5	M6*	M7*	M1*	M2*	M3*	M5	M6*	M7*
Australia	10.9%	29.9%	8.4%	6.5%	5.4%	5.4%	12.5%	18.0%	7.2%	11.2%	13.0%	5.8%
Canada	29.6%	28.0%	5.2%	9.6%	8.7%	8.6%	9.8%	14.2%	5.0%	9.3%	9.9%	7.2%
England and Wales	23.8%	17.4%	7.2%	8.7%	7.4%	8.0%	18.5%	15.1%	9.1%	11.2%	12.1%	7.6%
France	13.4%	33.9%	10.8%	11.7%	12.6%	8.2%	12.8%	18.4%	7.9%	19.9%	30.9%	7.8%
Japan	18.8%	12.5%	8.2%	7.7%	9.8%	6.1%	34.4%	29.3%	9.5%	15.4%	23.3%	6.0%
Norway	20.2%	30.2%	13.6%	14.4%	13.8%	12.6%	12.0%	12.0%	14.2%	15.4%	23.9%	9.6%
Sweden	46.9%	45.0%	9.7%	7.7%	7.9%	5.4%	9.4%	13.5%	9.7%	12.6%	18.6%	6.6%
United States	13.4%	20.5%	5.8%	8.9%	6.9%	7.2%	9.9%	35.9%	5.8%	9.2%	10.8%	8.1%

(b) Unisex mortality data

Population	M1*	M2*	M3*	M5	M6*	M7*
Australia	19.9%	18.8%	6.2%	6.8%	6.1%	4.6%
Canada	20.9%	24.0%	4.1%	7.7%	6.8%	7.0%
England and Wales	29.6%	16.1%	7.8%	8.9%	7.3%	7.3%
France	9.5%	13.1%	10.2%	14.5%	17.8%	9.1%
Japan	28.0%	20.6%	8.7%	8.9%	13.4%	6.2%
Norway	25.9%	26.5%	14.5%	12.3%	14.4%	9.2%
Sweden	16.3%	31.8%	9.9%	8.3%	9.7%	4.5%
United States	13.0%	54.0%	5.5%	8.4%	7.5%	7.0%

Table 3.5. The mean absolute percentage error (MAPE) of the forecasted cohort death probabilities over the period of 2000 to 2009 for the cohorts aged 46 to 81 at the beginning of 2000.

indexes.

Note that our recommendation is based on the adapted versions of mortality models in achieving the new-data-invariant property. The original, unrestricted version of models such as Model M2 may perform as good as Model M7, as shown by Cairns et al. (2009) who analyzed data from the United States and England and Wales. Also, we consider the age range of 40-90 throughout our analysis. The use of different age ranges such as those that include the ages below 40 or the ages beyond 90 may lead to a different recommendation with regards to the best adapted model for constructing mortality indexes.

3.2.5 Interpretations of the 3-factor CBD Mortality Indexes

To recap, Model M7* can be written as

$$\ln \left(\frac{q_{x,t}}{1 - q_{x,t}} \right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \hat{\sigma}_x^2) + \gamma_{t-x}^{(4)*}, \quad (3.2)$$

which means that at a given time t , the logit-transformed death probability is a quadratic function of $(x - \bar{x})$, with coefficients being the mortality indexes $\kappa_t^{(1)}$, $\kappa_t^{(2)}$ and $\kappa_t^{(3)}$. This simple structure permits us to interpret the three mortality indexes straightforwardly.

The first index, $\kappa_t^{(1)}$, captures the level of the mortality curve in logit scale. A reduction in $\kappa_t^{(1)}$ signifies an overall mortality improvement across all ages. The impact of a change in this mortality index is illustrated in the top right panel of Figure 3.1 (the solid and dashed lines).

The second index, $\kappa_t^{(2)}$, represents the slope of the logit-transformed $q_{x,t}$ with respect to age x . As demonstrated in the bottom left panel of Figure 3.1 (the solid and dashed lines), an increase in $\kappa_t^{(2)}$ leads to a steeper logit-transformed mortality curve, which implies that mortality at younger ages (below the mean age $\bar{x}$) declines

more rapidly than that at older ages (above the mean age $\bar{x}$).

The third index, $\kappa_t^{(3)}$, measures the curvature of the logit-transformed mortality curve. A higher value of $\kappa_t^{(3)}$ implies an increase in $q_{x,t}$ over the age ranges of 40-50 and 80-90 but a decrease in $q_{x,t}$ over the age range of 51-79. This effect is due to the term $((x - \bar{x})^2 - \hat{\sigma}_x^2)$, which is positive when $51 \leq x \leq 79$ and negative otherwise. The impact of $\kappa_t^{(3)}$ for fixed combinations of $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ is shown in the top left panel of Figure 3.1. The interaction between $\kappa_t^{(3)}$ and the other two indexes is demonstrated in the other panels of Figure 3.1.

Besides being easy to interpret, desirable mortality indexes should be able to capture a wide range of mortality age patterns, so that securities written on the indexes can effectively hedge longevity risk, which fundamentally arises from random non-parallel shifts in the underlying mortality curve. The ability of the 3-factor CBD mortality indexes to represent different age patterns of mortality can be observed from Figure 3.1 (in logit scale) and Figure 3.2 (in the original scale).

Let us illustrate the interpretations of the mortality indexes by considering a portfolio of 3-year term life insurance contracts that are issued to persons aged 40. The present value of this portfolio is higher than expected when:

- the future values of $\kappa_t^{(1)}$ are higher than expected (which implies the overall mortality improves slower than expected),
- the future values of $\kappa_t^{(2)}$ are lower than expected (which implies mortality at younger ages (below age 65) improves slower than expected), and
- the future values of $\kappa_t^{(3)}$ are higher than expected (which implies mortality over the age ranges of 40-50 and 80-90 improves slower than expected).

To further illustrate, we consider a closed pension plan, which makes pension payments to persons aged 65 to 91. The following statements apply to this pension plan:

- The present value is higher than expected if the future values of $\kappa_t^{(1)}$ are

Figure 3.1. Changes in the logit-transformed mortality curve under different combinations of the 3-factor CBD mortality indexes.

Figure 3.2. Mortality curves implied by six different combinations of the 3-factor CBD mortality indexes.

lower than expected (which implies the overall mortality improves faster than expected).

- The present value is higher than expected if the future values of $\kappa_t^{(2)}$ are lower than expected (which implies mortality at older ages (above age 65) improves faster than expected).
- The impact of $\kappa_t^{(3)}$ on the present value is not clear-cut, depending very much on the demographics of the plan. For instance, if the majority of the pensioners are now aged above 80, then lower future values of $\kappa_t^{(3)}$ are likely to yield a higher present value.

3.3 Hedging Longevity Risk with K-forwards

With three mortality indexes capturing different aspects of mortality patterns, we can split longevity risk into three constituent risks, namely “K1 risk”, “K2 risk” and “K3 risk”. K1 risk refers to the uncertainty associated with overall mortality improvements, while K2 and K3 risks reflect the uncertainty surrounding the first and second order age effects, respectively.

Chan et al. (2014) proposed, as a concept, a simple security called K-forward that can be used to hedge each of the three constituent risks. In this section, we detail the mechanism of a K-forward and explain how an effective longevity hedge can be built using a portfolio of K-forward contracts.

3.3.1 Specification of K-forwards

A K-forward contract is a zero-coupon swap that exchanges on the maturity date a fixed amount, determined at time 0 (when the contract is established), for a floating

Figure 3.3. Settlement of a K-forward contract at maturity.

amount that is proportional to a CBD mortality index (the reference rate) for a certain population (the reference population) in some future time (the reference year).

For a K-forward contract with a reference year t^* , the reference rate is $\kappa_{t^*}^{(i)}$, an unknown at time 0. The fixed amount is proportional to the corresponding forward mortality index, denoted as $\tilde{\kappa}_{t^*}^{(i)}$. This forward value is determined in such a way that no payment exchanges hands at time 0. Mathematically, the net payoff to the fixed rate receiver of a K-forward contract can be written as

$$Y \times (\tilde{\kappa}_{t^*}^{(i)} - \kappa_{t^*}^{(i)}), \quad i = 1, 2, 3, \quad (3.3)$$

where Y is the notional amount. The settlement that takes place on the maturity date T^* is illustrated in Figure 3.3. In practice, the maturity date T^* may be slightly later than the reference year t^* due to the time lag in the availability of the mortality index data. For simplicity, we assume that the maturity date T^* is last day of the reference year t^* .

Pension plans and life insurers can hedge their longevity risk exposures with K-forward contracts. As an illustration, we consider a closed pension plan that promises pension payments up to age 91. At present, the majority of the pensioners in the plan are aged 65-79. Since the pension payouts are negatively associated with the future values of $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$, the pension plan sponsor should write K1- and K2-forwards (K-forwards linked to the first two CBD mortality indexes) as a fixed rate receiver to offset the adverse outcome when the future values of $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ turn out to be lower than expected. As the majority of the pensioners are currently younger than age 80, higher than expected values of $\kappa_t^{(3)}$ in the first few

years will result in a larger aggregate payout. After a sufficiently long time period (say, 15 years), the age profile of the pensioners will be concentrated in the age range of 80-90, which means lower than expected values of $\kappa_t^{(3)}$ in later years will lead to more pension payments. To hedge its exposure to K3-risk, the pension plan sponsor should write short-dated K3-forwards as a fixed rate payer and long-dated K3-forwards as a fixed rate receiver.

The mechanism of K-forwards is similar to that of q-forwards, which were transacted in deals such as the one between Lucida PLC and J.P. Morgan in January 2008. Note that both instruments are equally transparent if the same method is used consistently to calculate the 3-factor CBD mortality indexes. The only difference between the two instruments is that the floating and fixed legs of a q-forward are respectively determined by q_{x^*, t^*} , the death probability at the reference age x^* in the reference year t^* , and q_{x^*, t^*}^f , the corresponding forward mortality rate.

Compared to a q-forward, a K-forward is an even simpler security and is more homogeneous, because it is categorized by only one parameter, the reference year. A K-forward longevity hedge may be easier to implement, because it does not require the hedger to make decisions about the reference ages. This advantage, as we are going to demonstrate in Section 3.5, is prominent when the portfolio being hedged covers multiple birth cohorts, involving many possible combinations of reference ages and years. From another viewpoint, K-forwards may be more conducive to the development of liquidity, because the market can focus on contracts written on three indexes only instead of the full range of discrete ages.

3.3.2 Key K-duration

To ensure a K-forward longevity hedge is effective, there is a need to calibrate it. We propose a calibration method called key K-duration (KKD), which is largely parallel to the key q-duration method recently proposed by Li and Luo (2012).

To explain the method of KKD, we need to define the notion of key K-indexes,

the mortality indexes $\kappa_t^{(i)}$, $i = 1, 2, 3$, in future years $t = t_1, t_2, \dots, t_n$ (the key years). It is assumed that K-forwards with reference years equal to the n key years are traded in the market and that no other K-forwards are available.

A KKD measures the change in the value of a liability with respect to a small change in a key K-index. Mathematically, the KKD associated with the i th CBD mortality index and the j th key year can be expressed as

$$KKD_i(P(\boldsymbol{\kappa}), j) = \frac{\partial P(\boldsymbol{\kappa})}{\partial \kappa_{t_j}^{(i)}}, \quad (3.4)$$

where $P(\boldsymbol{\kappa})$ denotes the portfolio value as a function of the vector of the mortality indexes $\boldsymbol{\kappa}$. The idea behind the KKD calibration method is to split the future time line into parts by the n key years. Then the longevity hedge is calibrated by matching the price sensitivities (measured by KKDs) of the pension plan and hedge portfolio over each subperiod.

In practice, it is generally difficult to compute a KKD by analytical means. To solve this problem, we propose a numerical estimation method, which involves the following two working assumptions:

- A shock in the key K-index $\kappa_{t_j}^{(i)}$ by an amount of $\delta^{(i)}(j)$ is accompanied with a level shift in $\kappa_t^{(i)}$ over the period of $t_j \leq t < t_{j+1}$ by the same amount.¹¹ The shock $\delta^{(i)}(j)$ is assumed to have no impact on the trajectory of $\kappa_t^{(i)}$ beyond $t_j \leq t < t_{j+1}$.
- The shock $\delta^{(i)}(j)$ has no impact on $\kappa_t^{(h)}$ for all $h \neq i$ and t .

Admittedly, these working assumptions are rather arbitrary, but they still lead to satisfying hedging results (see Sections 3.4 and 3.5).

Let $\check{\boldsymbol{\kappa}}(\delta^{(i)}(j))$ be the vector of mortality indexes that is subject to the shift $\delta^{(i)}(j)$. The proposed estimation procedure can be summarized as follows:

1. take $\boldsymbol{\kappa}$ as the best estimate of mortality indexes;

¹¹We define t_{n+1} as $+\infty$.

2. assuming the shift $\delta^{(i)}(j)$ is 0.1% of the best estimate, calculate $\check{\kappa}(\delta^{(i)}(j))$;
3. estimate the KKD associated with the i th CBD mortality index and the j th key year as $KKD_i(P(\boldsymbol{\kappa}), j) \approx \frac{P(\check{\kappa}(\delta^{(i)}(j))) - P(\boldsymbol{\kappa})}{\delta^{(i)}(j)}$.

3.3.3 Building a Longevity Hedge using Key K-durations

We now explain how key K-durations can be used to construct a static longevity hedge for a life-contingent liability at time 0. The hedge we build is a value hedge, which aims to reduce the variability of the liability's present value.

We set time 0 to be the beginning of the current year t_0 . Suppose that K1-, K2- and K3-forwards with reference years $t_1, t_2, \dots, t_n$ are available in the market. From the fixed rate receiver's viewpoint, the random present values of the cash flows from the K-forwards with reference year t_j can be expressed as

$$F_j^{(i)}(\boldsymbol{\kappa}) = (1 + r)^{-(T_j - t_0)}(\tilde{\kappa}_{t_j}^{(i)} - \kappa_{t_j}^{(i)}), \quad i = 1, 2, 3, \quad (3.5)$$

where r is the interest rate at which the cash flows are discounted.¹² It can be seen that

$$KKD_i(F_j^{(i)}(\boldsymbol{\kappa}), j) = -(1 + r)^{-(T_j - t_0)}, \quad (3.6a)$$

$$KKD_i(F_j^{(i)}(\boldsymbol{\kappa}), h) = 0, \quad j \neq h, \quad (3.6b)$$

$$KKD_h(F_j^{(i)}(\boldsymbol{\kappa}), j) = 0, \quad i \neq h. \quad (3.6c)$$

Equations (3.6b) and (3.6c) arise from the two working assumptions in Section 3.3.2.

We let $w^{(1)}(j)$, $w^{(2)}(j)$ and $w^{(3)}(j)$ be the notional amounts of the K1-, K2- and K3-forwards with reference year t_j , respectively. The present value of hedge

¹²The settlement date is T_j , which is generally later than the reference year t_j . As previously mentioned, to simplify our calculations, we assume that the settlement date T_j is the last day of the reference year t_j .

portfolio at time 0 is

$$H(\boldsymbol{\kappa}) = \sum_{h=1}^n \sum_{g=1}^3 w^{(g)}(h) F_h^{(g)}(\boldsymbol{\kappa}). \quad (3.7)$$

Let $L(\boldsymbol{\kappa})$ be the random present value of liability at time 0. To achieve effective hedging, the KKDs of hedge portfolio must match the KKDs of liability. This is equivalent to solving the following set of conditions for $i = 1, 2, 3$ and $j = 1, 2, \dots, n$:

$$\begin{aligned} KKD_i(L(\boldsymbol{\kappa}), j) &= KKD_i(H(\boldsymbol{\kappa}), j) \\ &= KKD_i\left(\sum_{h=1}^n \sum_{g=1}^3 w^{(g)}(h) F_h^{(g)}(\boldsymbol{\kappa}), j\right) \\ &= w^{(i)}(j) KKD_i(F_j^{(i)}(\boldsymbol{\kappa}), j). \end{aligned} \quad (3.8)$$

The simplification, which follows from Equations (3.6b) and (3.6c), allows us to determine the required notional amount of each K-forward separately without solving a system of equations. Furthermore, by using Equation (3.6a), we obtain

$$w^{(i)}(j) = -(1+r)^{T_j-t_0} KKD_i(L(\boldsymbol{\kappa}), j), \quad (3.9)$$

where $KKD_i(L(\boldsymbol{\kappa}), j)$ can be numerically estimated by the algorithm described in Section 3.3.2. A longevity hedge can be constructed readily by writing K-forwards with the notional amounts specified by Equation (3.9).

3.4 Illustrations of a K-Forward Hedge: Single Cohort

In this section, we illustrate a K-forward hedge for a hypothetical pension plan involving only a single birth cohort. The following assumptions about the pension plan and the hedging instruments are made:

- Assumptions about the hypothetical pension plan:

1. The pension plan contains one pensioner. It pays the pensioner \$1 at the beginning of each year starting from age 65 until the pensioner dies or attains age 91.
 2. There is no basis risk, that is, the pensioner's mortality experience is the same as the experience of the population to which the K-forwards are linked.
 3. All cash flows are discounted at an interest rate of 3%.
- Assumptions about the hedging instruments:
 1. The 3-factor CBD mortality indexes are derived from Model M7* that is fitted to data over the age range of 40-90.
 2. The current year t_0 (i.e. time 0) is taken as the beginning of the year immediately after the end of the data sample period (see Table ??). For instance, in the illustration that is based on data from English and Welsh population, time 0 is set to the beginning of year 2010.
 3. K-forwards with reference years 2015, 2020, 2025 and 2030 are available in the market.
 4. The forward mortality indexes, $\tilde{\kappa}_{t_j}^{(i)}$, are identical to the best estimates of the corresponding mortality indexes. This assumption, which implies a zero risk premium, affects the cost of the hedge but not the performance of hedge.
 5. The best estimates of the future mortality indexes required in Assumption 4 and the calculation of KKD values are obtained by extrapolating the time-varying parameters in Model M7* (fitted to the age range of 40-90 and the entire data sample period) through a Vector Autoregressive Integrated Moving Average (VARIMA) process.¹³

¹³The modeling of the CBD time-varying parameters with VARIMA processes is described by Chan et al. (2014). We also refer interested readers to Tiao and Box (1981) and Wei (2006) for further details about VARIMA modeling.

We evaluate the effectiveness of a longevity hedge by the amount of longevity risk reduction, R , defined as

$$R = 1 - \frac{\sigma^2(X^*)}{\sigma^2(X)}, \quad (3.10)$$

where $\sigma^2(.)$ is the variance function,

$$X = L(\boldsymbol{\kappa}) - \mathbb{E}(L(\boldsymbol{\kappa})) \quad (3.11)$$

is the random present value of unexpected cash flows (the random present value of future cash flows minus the expected present value of future cash flows) when the plan is not hedged, and

$$X^* = L(\boldsymbol{\kappa}) - \mathbb{E}(L(\boldsymbol{\kappa})) - H(\boldsymbol{\kappa}) + \mathbb{E}(H(\boldsymbol{\kappa})) \quad (3.12)$$

is the random present value of unexpected cash flows when a hedge is in place. A longevity hedge is considered to be effective if the variability of X^* is significantly smaller than that of X , so a higher value of R implies a better hedging performance.

Model M7* plays two roles in the longevity hedge. First, the mortality indexes to which the instruments are linked are derived from Model M7*. Second, in computing KKDs, the best estimates of the future mortality indexes are derived from a Model M7* that is fitted to historical data. So, a natural question to ask is whether the longevity hedge still works if future mortality does not follow Model M7*. To address this concern, we evaluate hedge effectiveness on the basis of mortality scenarios simulated from five different stochastic mortality models:

- Model M7*: the adapted version of model M7
- Model M5: the original CBD model
- Model M3: the original Age-Period-Cohort model
- Model M2: the original Renshaw-Haberman model

- Model MRW: the multivariate random walk model¹⁴ (Bell, 1997)

If the hedge is effective regardless of what simulation model is used, then a high hedge effectiveness is likely to be achievable in reality. The simulation procedure is summarized as follows:

1. Specify the stochastic processes for the time-varying and cohort effect parameters in the simulation models.
 - For Model M7*, the three time-varying parameters are modeled by a VARIMA process, whereas the cohort effect parameter is modeled by an AR(1) process.
 - For Model M5, the two time-varying parameters are modeled by a VARIMA process.
 - For Models M2 and M3, the time-varying parameter is modeled by a random walk with drift, whereas the cohort effect parameter is modeled by an ARIMA(1,1,0) process.¹⁵
 - For Model MRW, the evolution of the log death rates at different ages is modeled by a multivariate random walk with drift.
2. For each model, simulate 5,000 realizations of future death probabilities and transform them to realizations of the 3-factor CBD mortality indexes.¹⁶ Parameter uncertainty is incorporated into the simulations via the parametric bootstrap approach (see Brouhns et al., 2005).
3. For each realization, calculate the values of X and X^* . Then for each simulation model, compute the variances of X and X^* and finally the amount of longevity risk reduction R .

¹⁴It models the log death rates at different ages directly with a multivariate random walk with drift. No time-varying or age-specific parameters are involved.

¹⁵The ARIMA orders used for the cohort effect parameters in Models M7*, M2 and M3 are taken from the paper by Cairns et al. (2011b).

¹⁶The step is necessary, because we need the realized 3-factor CBD mortality indexes to calculate the simulated payoffs from the K-forwards. This step is not needed when the simulation model is Model M7* or when the hedging instruments used are q-forwards.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
Key year	2015	2020	2025	2030
$KKD_i(F_j^{(i)}(\kappa), j)$	-0.8375	-0.7224	-0.6232	-0.5375
$KKD_1(L(\kappa), j)$	-0.8373	-0.7540	-0.5825	-0.3155
$KKD_2(L(\kappa), j)$	-5.8541	-9.0194	-9.8330	-6.8644
$KKD_3(L(\kappa), j)$	139.2352	54.5539	-40.2955	-80.3121
$w^{(1)}(j)$	0.9997	1.0437	0.9347	0.5869
$w^{(2)}(j)$	6.9901	12.4850	15.7791	12.7698
$w^{(3)}(j)$	-166.2541	-75.5153	64.6624	151.2644

Table 3.6. The calculated key K-durations and notional amounts of K-forwards. The pension plan and K-forwards are linked to the mortality of English and Welsh males.

3.4.1 The Baseline Results

In this subsection, we present the baseline results for English and Welsh male population. The calculated KKDs and notional amounts are tabulated in Table 3.6. Note that a positive notional amount means the pension plan participates as a fixed rate receiver, whereas a negative notional amount means the pension plan participates as a fixed rate payer. It can be seen that the signs of the calculated notional amounts agree with the arguments presented in Section 3.3.1.

The longevity risk reductions calculated from different simulation models are displayed in Table 3.7, and the density functions for X and X^* simulated from Model M7* are shown in Figure 3.4. Using a portfolio of K1-, K2- and K3-forwards, the amounts of longevity risk reduction obtained range from 94.0% to 96.0%. The results depend minimally on the simulation model used, indicating that the hedge effectiveness is achievable even if future mortality does not follow Model M7*. The results also indicate that a high hedge effectiveness can be achieved by using K-forwards associated with four key years only.

Table 3.7 also shows the amounts of longevity risk reduction resulting from different subsets of the available K-forwards. It can be seen that K1 risk is the most crucial whereas K3 risk has the smallest impact. In particular, when the simulation model is Model M5, the use of K3-forwards seems to have no marginal

Simulation K1 only model	K2 only	K3 only	K1&K2	K1&K3	K2&K3	K1&K2&K3
M7*	68.8%	43.5%	0.5%	93.7%	69.4%	94.7%
M5	89.5%	3.3%	0.0%	96.0%	89.5%	96.0%
M3	95.6%	4.6%	0.2%	95.7%	95.5%	95.6%
M2	93.7%	16.1%	0.0%	95.1%	93.7%	95.2%
MRW	77.2%	39.3%	0.6%	93.5%	77.9%	94.0%

Table 3.7. The amounts of longevity risk reduction (R) resulting from different combinations of K-forwards and different simulation models. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The VARIMA orders for the time-varying parameters in Models M7* and M5 are (1,1,0) and (5,1,0), respectively.

benefit. This result can be explained using the following equation:

$$X_{\text{with K3-forward}}^* = X_{\text{without K3-forward}}^* - \sum_{j=1}^4 w^{(3)}(j)(1+r)^{-(T_j-t_0)}(\mathbb{E}(\kappa_{t_j}^{(3)}) - \kappa_{t_j}^{(3)}). \quad (3.13)$$

Because of the nature of Model M5, which assumes that future mortality is driven entirely by $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$, the values of $\kappa_t^{(3)}$ simulated from Model M5 are practically zero. It follows that the variability in the second term of the right-hand-side of Equation (3.13) is close to zero. Therefore, $X_{\text{with K3-forward}}^*$ and $X_{\text{without K3-forward}}^*$ have almost the same variability, which implies that the impact of the K3-forwards on the hedge effectiveness is negligible.

In general, the third CBD mortality index may take values that are significantly different from zero.¹⁷ In this case, adding K3-forwards to the hedge portfolio may improve hedge effectiveness. For example, when Model M7* is assumed, adding K3-forwards to a portfolio of K1- and K2-forwards increases the value of R from 93.7% to 94.7%. We also found that the benefit of adding K3-forwards is more apparent for some other populations such as U.S. males.

In principle, given a collection of simulated mortality scenarios, we can gradually adjust the notional amounts of the K-forwards in the hedge portfolio to compute the highest attainable value of R . The value of R obtained by matching

¹⁷For all populations under consideration, Model M7* results in better BIC values compared to Model M6 (see Table 3.4). This supports the existence of a quadratic age effect, which implies that the third CBD mortality index generally takes non-zero values.

Figure 3.4. The distributions of X and X^* simulated from Model M7*. It is assumed that the underlying population is English and Welsh males and that the pension plan consists of a single pensioner currently aged 65.

KKDs is generally less than its highest attainable value, because KKD is only an approximate measurement of the sensitivity to the underlying mortality indexes. We found that the maximum achievable R values for a portfolio of K1-, K2- and K3-forwards are 97.3% (for Model M7*), 99.1% (for Model M5), 96.6% (for Model M3), 95.8% (for Model M2) and 96.4% (for Model MRW). These values are very close to those obtained by matching KKD, providing a strong case for the KKD method. Note that the computation of the highest attainable R value requires heavy computational resources, because it involves simulations and a maximization of a multi-variable function. By contrast, as discussed in Section 3.3.3, the method of KKD is rather easy to execute.

3.4.2 Sampling Risk

Sampling risk, or small-sample risk, refers to the risk that the actual mortality experience of a population turns out to be different from the ‘true’ underlying mortality rates. In this subsection, we study the impact of sampling risk on hedge effectiveness. Again, we use the mortality data from English and Welsh males

population to illustrate.

We retain all assumptions specified in Section 3.4, except that we now assume the pension plan has a cohort of pensioners aged 65 at time 0 instead of only one single pensioner. We let $l(65)$ be the initial number of pensioners, and $l(x)$ be the number of pensioners who survive to age x , where $65 < x \leq 91$. To incorporate sampling risk, we treat the cohort of pensioners as a random survivorship group. Specifically, we model $l(x)$ for $65 < x \leq 91$ by a binomial process:

$$l(x+1) \sim \text{Binomial}(l(x), 1 - q_{x,1945+x}), \quad x = 65, 66, \dots, 90. \quad (3.14)$$

The procedure for simulating liability cash flows is adapted as follows:

1. Simulate a future mortality curve from the assumed mortality model.
2. Given the simulated mortality curve, simulate the number of survivors $l(x)$, for $x = 66, 67, \dots, 91$, from the binomial process.
3. Calculate the liability cash flows based on the simulated values of $l(x)$.
4. Repeat the steps above to obtain 5,000 scenarios of liability cash flows.

The incorporation of sampling risk does not affect the method by which the KKDs are calculated. Therefore, in this illustration, the values of $KKD_i(F_j^{(i)}(\boldsymbol{\kappa}), j)$ are the same as those in Table 3.6, whereas the values of $KKD_i(L(\boldsymbol{\kappa}), j)$ are equal to $l(65)$ multiplied by the corresponding values in Table 3.6 (because the initial number of pensioners is $l(65)$ instead of one). It follows that the required notional amounts are simply $l(65)$ multiplied by the notional amounts shown in Table 3.6.

The amounts of longevity risk reduction for different values of $l(65)$ are shown in Table 3.8. The row with $l(65) = \infty$ shows the values of R when there is no sampling risk. It can be seen that the hedge effectiveness drops as $l(65)$ decreases. When $l(65) = 1000$, the value of R resulting from the hedge built from all available K-forwards is only 63.7%. This finding suggests that a small pension plan should consider alternative de-risking solutions, such as pension buy-ins (see Coughlan et

$l(65)$	K1 only	K2 only	K3 only	K1&K2	K1&K3	K2&K3	K1&K2&K3
∞	68.8%	43.5%	0.5%	93.7%	69.4%	44.4%	94.7%
10000	65.6%	41.3%	0.6%	89.3%	66.3%	42.3%	90.3%
5000	63.0%	39.8%	0.5%	85.8%	63.6%	40.6%	86.8%
3000	59.6%	37.3%	0.6%	80.8%	60.2%	38.2%	81.9%
1000	46.5%	29.0%	0.4%	63.0%	46.9%	29.6%	63.7%

Table 3.8. The amounts of risk reduction (R) for pension plans of different initial sizes. The pension plans and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.

al., 2013), instead of using standardized K-forwards.

3.4.3 Sensitivity Tests

In this subsection, we sensitivity test several assumptions used in estimating hedge effectiveness for the baseline case of one pensioner in Section 3.4.1.

3.4.3.1 Interest Rate

In the previous illustrations, an interest rate of 3% per annum was used. We now examine the relationship between the hedging results and the assumed interest rate. The results are summarized in Table 3.9. A higher interest rate produces a lower standard deviation of X , because the more distant and so more uncertain cash flows are discounted by a larger extent. For the same reason, the required notional amounts are lower in magnitude when the assumed interest rate is higher. Nevertheless, the amounts of risk reduction under different interest rate assumptions are similar, indicating that the effectiveness of a K-forward hedge is not very sensitive to the interest rate assumption.

3.4.3.2 Availability of K-forwards

In the KKD framework, the key years are the reference years for which K-forwards are available. It is assumed in the baseline calculations that the separation between

Interest Rate	$\sigma(X)$	K1 only	K2 only	K3 only	K1&K2	K1&K3	K2&K3	K1&K2&K3
1%	0.325	68.7%	45.0%	0.3%	94.3%	69.0%	45.8%	95.2%
2%	0.269	68.7%	44.3%	0.4%	94.0%	69.2%	45.1%	95.0%
3%	0.223	68.8%	43.5%	0.5%	93.7%	69.4%	44.4%	94.7%
4%	0.187	68.8%	42.8%	0.6%	93.3%	69.6%	43.7%	94.4%
5%	0.157	68.7%	41.9%	0.8%	92.8%	69.8%	42.9%	94.1%

Table 3.9. The amounts of risk reduction (R) under different interest rate assumptions. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.

Separation between two adjacent key years (total number of key years)	K1 only	K2 only	K3 only	K1&K2	K1&K3	K2&K3	K1&K2&K3
8 (3)	53.3%	32.5%	0.5%	77.5%	53.9%	33.2%	78.3%
7 (3)	59.1%	36.6%	0.4%	84.5%	59.6%	37.3%	85.3%
6 (4)	62.8%	39.2%	0.5%	88.5%	63.5%	40.0%	89.4%
5 (4)	68.8%	43.5%	0.5%	93.7%	69.4%	44.4%	94.7%
5 (3)	67.6%	42.4%	0.3%	91.5%	68.2%	43.0%	92.3%
5 (2)	61.1%	36.7%	0.1%	81.4%	61.5%	36.8%	81.7%

Table 3.10. The amounts of risk reduction (R) resulting from different spacings between two adjacent key years. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.

two adjacent key (reference) years is 5 years. Here we examine how the hedging results may change when different separations are assumed. Throughout the analysis, the earliest key year is fixed to 2015. The hedging results are tabulated in Table 3.10. It can be seen that no matter 3 or 4 key years are used, a smaller separation produces a (slightly) higher amount of risk reduction. This result may possibly be explained by the fact that closer key years (from 2015) yield better hedging results for the earlier random cash flows but worse hedging results for the later ones, which are discounted more heavily and hence less important to the overall hedging performance. It can also be seen that, as expected, the amount of risk reduction increases with the number of key years (i.e., the number of instruments used).

Age range	$\bar{x}$	K1 only	K2 only	K3 only	K1&K2	K1&K3	K2&K3	K1&K2&K3
40-90	65	68.8%	43.5%	0.5%	93.7%	69.4%	44.4%	94.7%
50-90	70	79.8%	22.3%	3.8%	91.1%	81.1%	27.8%	94.1%
60-90	75	87.9%	7.6%	5.4%	90.2%	89.9%	14.4%	93.5%

Table 3.11. The amounts of risk reduction (R) for different age ranges over which the 3-factor CBD mortality indexes are defined. The pension plan and K-forwards are linked to the mortality of English and Welsh males. The simulation model is Model M7* (fitted to the corresponding age range) with a VARIMA(1,1,0) process for the time-varying parameters.

3.4.3.3 Age Range

In our baseline calculations, we use an age range of 40-90 to define the 3-factor CBD mortality indexes. We now study how the age range over which the indexes are defined may affect the hedging results. In Table 3.11 we compare the baseline hedging results with the hedging results that are based on two alternative age ranges, 50-90 and 60-90. It can be observed that the choice of age range does not have a significant impact on the amount of risk reduction produced by the hedge with all three types of K-forwards. However, the R values for other combinations of K-forwards vary significantly with the age range used. In particular, the performances of the K1-forward only and K3-forward only hedges improve with the mean age in the age range, but the reverse is true for the K2-forward only hedge. This pattern suggests that if the 3-factor CBD mortality indexes are defined over an older age range, the first and third CBD indexes would capture a larger proportion of the overall longevity risk, while the second CBD index would capture less.

3.4.4 Advanced Ages

Here, we examine how a K-forward hedge may perform if we extend the coverage of the hypothetical pension plan to age 101. While the age coverage of the pension plan is extended, the definition of the 3-factor CBD mortality indexes remains unchanged; that is, the indexes are still derived by fitting Model M7* to data over the age range of 40-90. We use the following procedure to calculate the amount of

risk reduction produced by the K-forward hedge for the extended pension plan.

1. Calibrate the K-forward hedge as follows:
 - (a) Compute the best estimates of future mortality indexes by extrapolating from a VARIMA process that is fitted to the historical index values.
 - (b) Calculate the best estimates of the death probabilities over the age range of 65-101 using Equation (3.2) with $\bar{x} = 65$ and $\hat{\sigma}_x^2 = \frac{650}{3}$.
 - (c) On the basis of the values from Steps (a) and (b), calculate $P(\boldsymbol{\kappa})$, $P(\check{\boldsymbol{\kappa}}(\delta^{(i)}(j)))$ and $KKD_i(P(\boldsymbol{\kappa}), j)$ for $i = 1, 2, 3$ and $j = 1, \dots, n$.
 - (d) Using the KKD values computed in Step (c), compute the notional amounts of the K-forwards in the hedge portfolio.
2. Simulate 5000 realizations of future death probabilities from a stochastic mortality model that is fitted to historical data over the age range of 40-100. As before, the simulation models considered are Models M7*, M5, M3, M2 and MRW.
3. Calculate the simulated present values of pension payments from age 65 to 101.
4. Transform the simulated future death probabilities over the age range of 40-90 to realizations of the 3-factor CBD mortality indexes, which are then used to determine the simulated payoffs from the K-forwards.
5. Compute X , X^* and finally the amount of risk reduction R .

We use the same key years despite the coverage of this pension plan is 10 years longer. With this setup, the amounts of longevity risk reduction produced by a calibrated portfolio of K1-, K2- and K3-forwards are 93.9% (for Model M7*), 93.4% (for Model M5), 94.3% (for Model M3), 95.1% (for Model M2) and 93.9% (for Model MRW). These findings suggest that the KKD strategy still produces a satisfactory hedge effectiveness when the pension plan is extended to include more

advanced ages.

3.5 Generalization to Multiple Cohorts

In real life, individuals in a pension plan are generally born in different years. For a multi-cohort pension plan, longevity risk is associated with a two-dimensional mortality surface, composed by a collection of mortality curves for various birth cohorts. The illustrations in the previous section may be viewed as a special case, whereby the longevity risk exposure arises from a particular diagonal over the mortality surface.

In this section, we use a more realistic hypothetical pension plan to illustrate how the proposed method can be applied to situations involving multiple birth cohorts. We retain all assumptions stated in Section 3.4, except those about the pensionable age and the plan's demographics. Here, we assume that individuals in the plan are currently aged 50 to 85, following the age distribution displayed in Figure 3.5. The pensionable age is 60, so the individuals who are currently aged 50 to 59 are now active members (who will receive pension payments when they reach age 60) while those who are currently aged 60 or above are now retirement pensioners (who are currently receiving pension payments). It is further assumed that the plan is closed to new entrants and that all active members remain in their jobs until retirement.

We consider two hedging scenarios:

1. A hedge for the current retirement pensioners only

This hedge aims to reduce the variability of the present values of the pension payments to the current retirement pensioners in the next 31 years.

2. A hedge for both current active members and retirement pensioners

This hedge aims to reduce the variability of the present values of the pen-

Figure 3.5. The initial demographic structure of the multi-cohort closed pension plan.

sion payments to the current retirement pensioners and the deferred pension payments to the current active members in the next 31 years.

As the market has started to see static longevity hedges for active members, the results for the second hedging scenario may be of many practitioners' interest.¹⁸

Another objective of this section is to compare the K-forward longevity hedges with the corresponding longevity hedges that are composed of q-forwards, a standardized instrument that has been promoted by the Life and Longevity Markets Association (LLMA). We calibrate q-forward hedges for the aforementioned two hedging scenarios using the method of key q-duration (KQD), proposed by Li and Luo (2012).

As before, all results in this section are computed using mortality data from English and Welsh males population.

¹⁸In January 2011, the Pall (UK) pension fund executed a £70 million deal with J.P. Morgan to hedge the longevity risk associated with its active and deferred members (see Blake et al., 2014).

3.5.1 Building K-Forward Hedges

In this illustration, we assume that K-forwards with reference years 2015, 2020, 2025, 2030 and 2035 are available. An additional reference year (2035) is considered, because the hedging horizon here is longer than that for the single-cohort pension plan, which makes at most 26 years of payments (from age 65 to 91). The KKD with respect to each index and reference year can be estimated by using the algorithm described in Section 3.3.2. Given the KKD values, the required notional amounts (shown in Table 3.12) can be computed accordingly.

It can be observed that the signs of the notional amounts are completely in line with the arguments presented in Section 3.3.1. It is interesting to note that the pension plan's positions in the K3-forward with reference year 2025 are different in the two hedging scenarios. For the hedge involving current retirement pensioners only, the plan's position in that K3-forward should be a fixed rate receiver, because in 2025 the individuals associated with the hedge will attain ages 75-90, such that the total pension payments at that time will be negatively related to the index value to which that K3-forward is linked.

It is also interesting to compare the required notional amounts of K1- and K2-forwards in the two hedging scenarios. Since more risk is being hedged, the notional amounts of K1- and K2-forwards in the larger hedge covering both current active members and retirement pensioners are generally higher. The only exception is the K2-forward with reference year 2015. This is because from 2015 to 2019, all individuals associated with the smaller hedge (covering the current retirement pensioners only) will be at least 65 years old, but some individuals associated with the larger hedge will still be younger than 65. As explained in Section 3.2.5, pension payments below and above age $\bar{x} = 65$ respond to the second CBD mortality index in opposite directions. The offsetting effects thus reduce the amount of K2 risk that the larger hedge has to mitigate.

Reference year t_j	Type	Notional amount	
		Retirement pensioners only	Active members and retirement pensioners
2015	K1	11821	16887
	K2	110873	100389
	K3	−1174133	−2220565
2020	K1	10736	17719
	K2	137888	151606
	K3	−357008	−1775855
2025	K1	8495	16116
	K2	138762	191452
	K3	518404	−696304
2030	K1	5058	12385
	K2	99205	185632
	K3	874180	373324
2035	K1	1588	6730
	K2	35931	121384
	K3	469823	822578

Table 3.12. The required notional amounts of K-forwards, calibrated by KKD. The pension plan and K-forwards are linked to the mortality of English and Welsh males.

3.5.2 Building q-Forward Hedges

We now use the KQD method to develop the corresponding q-forward hedges. The first step in the KQD method is to identify the key cohorts in the pension portfolio. The pension plan involves 36 birth cohorts in total, with the oldest born in year 1925 and the youngest in year 1960. In the smaller hedge for the current retirement pensioners only, we consider five key cohorts, which were born in 1930, 1935, 1940, 1945 and 1950. In the larger hedge that covers additionally the current active members, we use two extra key cohorts, which were born in 1955 and 1960.

The next step is to identify the key mortality rates in each key cohort. Following Li and Luo (2012), the key mortality rates are chosen in such a way that they are no more than five ages apart (see Table 3.13). It is assumed that the q-forwards linked to the chosen key mortality rates are available in the market.

We then use the q-forwards that are linked to the key mortality rates to build

Key cohort k	Year of birth c_k	Number of key mortality rates n_k	Location of key mortality rates
1	1930	1	Age 85
2	1935	2	Ages 80, 85
3	1940	3	Ages 75, 80, 85
4	1945	4	Ages 70, 75, 80, 85
5	1950	5	Ages 65, 70, 75, 80, 85
6	1955	5	Ages 65, 70, 75, 80, 85
7	1960	4	Ages 65, 70, 75, 80

Table 3.13. Locations of the key mortality rates of the chosen key cohorts.

longevity hedges for the two hedging scenarios. The reference rate of the q -forward that corresponds to the l th key mortality rate of the k th key cohort (which was born in year c_k) is $q_{x_{l,k}, t_{l,k}}$, where $x_{l,k}$ and $t_{l,k} = c_k + x_{l,k}$ are the q -forward's reference age and reference year, respectively. For instance, the reference rates of the five q -forwards associated with the key cohort that was born in 1950 are $q_{65,2015}$, $q_{70,2020}$, $q_{75,2025}$, $q_{80,2030}$ and $q_{85,2035}$. Note that the smaller hedge consists of 15 q -forwards, while the larger hedge contains 9 more q -forwards, as two extra key cohorts are introduced.

The notional amounts are determined by matching the KQD values of the pension plan and the portfolio of q -forwards. The KQD values are estimated using the procedure detailed in Section 5 of Li and Luo (2012). The required notional amounts of q -forwards for each hedging scenario are tabulated in Table 3.14. As a consequence of how the KQD values are estimated, the required notional amounts of the 10 q -forwards corresponding to the first four key cohorts are identical in both hedging scenarios. Also, as expected, the pension plan acts as a fixed rate receiver in all q -forwards, because the pension payments are negatively associated with the realized mortality rates.

Key cohort k	Reference age $x_{l,k}$	Reference year $t_{l,k}$	Notional amount	
			Retirement pensioners only	Active members and retirement pensioners
1	85	2015	35107	35107
2	80	2015	75556	75556
	85	2020	19215	19215
3	75	2015	210216	210216
	80	2020	64683	64683
	85	2025	36228	36228
4	70	2015	418885	418885
	75	2020	162233	162233
	80	2025	102002	102002
	85	2030	58076	58076
5	65	2015	283558	500145
	70	2020	159431	255112
	75	2025	115337	183150
	80	2030	73061	113965
	85	2035	41998	60440
6	65	2020		564347
	70	2025		242804
	75	2030		163699
	80	2035		85929
	85	2040		23716
7	65	2025		341518
	70	2030		138863
	75	2035		83317
	80	2040		30373

Table 3.14. The required notional amounts of q-forwards, calibrated by KQD. The pension plan and K-forwards are linked to the mortality of English and Welsh males.

Hedging scenarios	K1-, K2- and K3-forwards	q-forwards
Retirement pensioners only	94.5%	91.6%
Active members and retirement pensioners	96.0%	95.5%

Table 3.15. The amounts of risk reduction (R) resulting from the K-forward and q-forward hedges for the hypothetical multi-cohort pension plan. The pension plan and hedging instruments are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.

3.5.3 Hedging Results

The hedging results are reported in Table 3.15 and illustrated graphically in Figure 3.6. For the hedging scenario involving the current retirement pensioners only, the amounts of risk reduction produced by the K-forward hedge and the q-forward hedge are 94.5% and 91.6%, respectively. For the other scenario that includes additionally the current active members, the amounts of risk reduction resulting from the K-forward hedge and the q-forward hedge are 96.0% and 95.5%, respectively. The simulation results indicate that both types of instruments can yield highly satisfactory hedging results, even if the portfolio being hedged covers multiple birth cohorts. They also indicate that a K-forward hedge may yield a higher amount of risk reduction compared to a q-forward hedge.

More importantly, what distinguishes the two instruments is the number of securities needed to achieve a given level of hedge effectiveness. For the hedging scenario involving both current retirement pensioners and active members, although both hedges lead to similar R values, the q-forward hedge (calibrated by the KQD method) requires 9 more securities in comparison to the K-forward hedge (calibrated by the KKD method). It follows that when the number of cohorts covered is large, a K-forward hedge is more compact and thus more preferred.

Technically speaking, the KQD measure relies heavily on certain assumptions about how a shift in a key mortality rate affects the mortality rates in the neighbouring cohorts. Li and Luo (2012) showed empirically that these assumptions are

Figure 3.6. The distributions of X and X^* for the hedges that are constructed for both current active members and retirement pensioners in the hypothetical multi-cohort pension plan. The pension plan and hedging instruments are linked to the mortality of English and Welsh males. The simulation model is Model M7* with a VARIMA(1,1,0) process for the time-varying parameters.

reasonably accurate when the separation between two key cohorts is small, but are less likely to hold if the key cohorts are farther apart. Therefore, in implementing the KQD method, additional key cohorts and hence q-forward contracts have to be introduced when the portfolio being hedged covers more cohorts of individuals. Otherwise, the estimated KQD values may not accurately reflect the portfolio's sensitivity to changes in the underlying mortality curve, thereby leading to a less satisfactory hedging result.

By contrast, the KKD value for a particular CBD mortality index depends only on time, having no relationship with the individuals' years of birth. It follows that there is no need to alter the choices of the key K-indexes when more cohorts of individuals are included in the portfolio being hedged. In our illustration, if the pension plan covers even more birth cohorts (say those who are aged 40-49 at time 0), the number of key K-indexes and hence the number of K-forwards will still be 15. To hedge the risk associated with these additional cohorts, all the hedger needs to do is to recalculate the notional amounts.

Moreover, the satisfactory hedging results as shown above suggest that the amount of residual risks due to the use of key K-index (as opposed to the K-index in each year) for a K-forward hedge and key mortality rates (as opposed to the mortality rates of each cohort) for a q-forward hedge are similar to each other. In other words, the use of a K-forward hedge that is categorized by only reference year does not produce a larger residual risk as compared to a q-forward hedge that is linked to both reference age and reference year.

The benefit of K-forwards to market development has now become clear. With a finite number of K-forwards, the hedging needs of pension plans with different demographic structures can be met. However, the collection of q-forwards that are applicable to one pension plan may not be suitable and/or sufficient for another pension plan that has a different age profile. Hence, relative to q-forwards, K-forwards are easier to attract demand from hedgers and are potentially more liquid.

3.6 Conclusion

In this chapter, we studied the construction of mortality indexes using stochastic mortality models. We recommend using the three time-varying parameters in Model M7* (i.e., Model M7 that is adapted to satisfy the new-data-invariant property) to construct mortality indexes. One basis of our recommendation is that among the collection of six appropriately adapted models, Model M7* provides the best fitting when applied to data over the age range of 40 to 90 for various populations. Model M7* also performs the best in a backtest that compares projected values with actual values. Another basis is that the three time-varying parameters, which we call the 3-factor CBD mortality indexes, are easy to interpret and are able to reflect the varying age pattern of mortality.

We also investigated the potential applications of the 3-factor CBD mortality indexes to securitization. In particular, we further developed a standardized

security called K-forward, the payoff of which is linked to a 3-factor CBD mortality index. We also explained how and why a portfolio of K-forward contracts can be used to hedge the longevity risk associated with a life-contingent portfolio. In terms of structure, the proposed K-forwards are simpler than the existing q-forwards, because each K-forward contract is characterized by one single parameter (the reference year) only.

The proposed KKD measure permits one to calibrate a longevity hedge that is formed by K-forward contracts. From the results of the simulation studies we conducted, the following conclusions can be drawn. First, a calibrated K-forward hedge can substantially reduce the variability in the values of a pension portfolio, even if parameter uncertainty and sampling risk are taken into account. Second, the effectiveness of a K-forward hedge does not vary significantly with the assumed simulation model, indicating that the success in hedging is likely to be achievable in real life. Finally, compared to a q-forward hedge, a K-forward hedge giving a comparable hedge effectiveness requires a smaller number of securities, thereby helping the market to concentrate liquidity.

One important issue that is not addressed is the pricing of K-forwards. Our illustrations are based on the assumption that the K-forwards are costless, so they provide no information about how a K-forward hedge may cost the hedger. It is warranted in future research to study how a K-forward contract may be priced by no-arbitrage approaches (see, e.g., Cairns et al., 2006; Denuit et al., 2007a; Dowd et al., 2006; Li et al., 2011) or economic methods (see, e.g., Zhou et al., 2011, 2013a,b). Given that K-forwards have a very simple structure, it may be possible to derive close-form pricing formulas for them.

Chapter 4

Optimal Design of a Bonus-Malus System

4.1 Introduction

When insurers are working with cross-sectional data, the motor insurance ratemaking is performed in two separate steps. In the first step, *a priori* ratemaking aims to determine the amount of *a priori* premium through the use of risk classification variables. In the second step, *a posteriori* ratemaking is imposed as a correction mechanism using the claims experience information. In practice, a bonus-malus system (BMS) is commonly used in *a posteriori* ratemaking due to commercial reasons. It consists of the following three building blocks:

- BMS levels: The number of BMS levels is specified in advance by the insurer. Every new driver enters into the BMS at a pre-specified starting level.
- transition rules: After each year, the driver moves up or down these levels according to a set of transition rules, depending on the claims experience in that year.
- relativities: Each level is associated with a premium adjustment coefficient

called relativity, which is to be multiplied with the base premium to obtain the premium amount.

Generally, insurers consider the frequency of claims reported (see, e.g., Lemaire, 1995; Tremblay, 1992; Walhin and Paris, 1999) as the relevant claims experience information, but a few studies (see, e.g., Frangos and Vrontos, 2001; Mahmoudvand and Hassani, 2009; Pinquet, 1997; Tzougas et al., 2014) took into account both the frequency and severity of claims.

Since the occupancy of BMS level is a function of the claims experience, a driver who has a lower *a priori* expected claim frequency is expected to gravitate towards the lower BMS levels. Conversely, a driver with a higher *a priori* expected claim frequency tends to occupy the higher BMS levels. Using such an argument, Taylor (1997) pointed out that the drivers occupying the BMS level with a lower average *a priori* expected claim frequency tend to pay a lower *a priori* premium. Hence, he further commented that the justifiable BMS relativities need to consider the differentiation of underlying claim frequencies by experience, but only to the extent that this differentiation has not been recognized within the *a priori* premiums.

The above remarks have three important implications. Firstly, they tell us that the main purpose of a BMS is to deal with the heterogeneity within each risk class but not the heterogeneity between different risk classes. Secondly, the information of *a priori* expected claim frequency should be incorporated into the determination of premium relativities. Finally, the integration of both *a priori* and *a posteriori* ratemakings should produce a BMS in which the average *a priori* expected claim frequency appears to be as indistinguishable as possible across different BMS levels once the steady state has been reached. Put it differently, the average *a priori* expected claim frequencies of BMS levels should exhibit as little variation as possible to ensure that the implementation of BMS does not further reinforce the *a priori* risk classification. This last implication can be used to alleviate the clustering of policyholders in the BMS levels according to their *a priori* risk characteristics, but

surprisingly it is often overlooked in the design of BMS.

In general, the set of transition rules is easy to understand except for a few special bonus rules such as in the former Belgian BMS (see Pitrebois et al., 2003a). Specifically, the same number of level transitions is applied to all drivers regardless of their current levels occupied and *a priori* characteristics, as long as the drivers make the same number of claims in the current year. For instance, a common set of transition rules is $-1/+2$ (i.e. one level bonus for no claims; two levels malus for each claim). Given the number of BMS levels and the specified transition rules, the optimal relativity for each BMS level was derived by Norberg (1976) through the maximization of asymptotic predictive accuracy, more formally known as Norberg's criterion.

When the general BMS above – with a single set of optimal relativities and a set of simple transition rules – is implemented, it leads to two inadequacy scenarios. The first inadequacy scenario was pointed out by Pitrebois et al. (2003b) who argued that unfairness is induced because of the same *a posteriori* correction in terms of relativity being applied to the policyholders independent of their *a priori* characteristics, hence overpenalizing the drivers who are *a priori* bad risks.¹⁹ Similarly, in this chapter we argue that the second inadequacy scenario arises because of the same *a posteriori* correction in terms of level transition being applied to the policyholders independent of their current levels occupied, hence overpenalizing the drivers staying in the higher BMS levels.

One proposed solution to tackle the first inadequacy scenario is a single-step ratemaking approach developed by Boucher and Inoussa (2014) when a panel data structure is available for insurers. Specifically, both the *a priori* premiums and *a posteriori* relativities are estimated simultaneously via statistical log-likelihood approach using what they referred to as the bonus-malus system for panel data (BMS-panel model). The BMS-panel models are feasible because the structure of

¹⁹Specifically, they argued that the bad drivers that are required to pay a larger *a priori* premium should be rewarded more and penalized less than the good drivers.

panel data enables insurers to obtain premiums based on empirical considerations.

However, the single-step ratemaking approach in Boucher and Inoussa is not applicable for the BMS framework developed for cross-sectional data. In this chapter, we fill in this gap by proposing a new objective function in the determination of optimal relativities. In more detail, our objective function, as a generalization of the Norberg's criterion, aims to minimize the expected squared difference between the true premium and the actual premium (in absolute terms). By minimizing the new objective function, the first inadequacy scenario could be alleviated because on average the actual premiums payable by policyholders are of the smallest difference from their corresponding true premiums. A financial equilibrium constraint is added to the optimization such that the set of optimal relativities has an average of 100%. This constraint was first considered by Coene and Doray (1996) who devised a financially-balanced BMS. The solution can then be derived analytically by using the Lagrangian method. Moreover, we argue that the last implication should be given more consideration in designing an optimal BMS and address this issue by introducing a new metric called effectiveness of transition rules.

On top of that, we introduce the concept of varying transition rules, which can alleviate the second inadequacy scenario and may improve effectiveness. As an illustrative example, we propose to specify the transition rules as a functional form of the current level occupied instead of a rigid setting. The numerical results indicate that the proposed varying transition rules can lead to a higher level of effectiveness. We also apply the concept of varying transition rules onto the BMS-panel model, whereby numerical results suggest that the varying transition rules may perform better than their corresponding simple rules based on the log-likelihood values.

The remainder of this chapter is organized as follows. Section 4.2 reviews the motor insurance ratemaking technique for both cross-sectional and panel data structures. Section 4.3 discusses the three implications and describes the two inadequacy scenarios as noted above. Section 4.4 proposes the new objective function for

the determination of optimal relativities and introduces the effectiveness of transition rules as a metric in the BMS framework developed for cross-section data. The rationales of using a set of varying transition rules are presented, followed by the specification of the proposed rules. Numerical illustrations for the BMS when working with cross-sectional data and panel data are provided in Section 4.5 and Section 4.6, respectively. Finally, Section 4.7 concludes this chapter.

4.2 Motor Insurance Ratemaking

The motor insurance ratemaking process consists of two steps called *a priori* ratemaking and *a posteriori* ratemaking. Specifically, the *a priori* ratemaking classifies the policyholders into different tariff risk classes using the observable risk classification variables under the framework of generalized linear models (see McCullagh and Nelder, 1989), such that the policyholders in each tariff class pay the same *a priori* premium. The residual heterogeneity that still remains is tackled by an *a posteriori* ratemaking mechanism, which uses past claims experience information to determine the premium corrections. The two common *a posteriori* ratemaking mechanisms are credibility premium and BMS. In this chapter we focus on the design of BMS, which can be perceived as the commercial version of the credibility premium. We refer readers to the seminal paper by Dionne and Vanasse (1989) for discussions on the credibility premium.

4.2.1 A Priori Ratemaking

Let us consider a portfolio with n policies. Denote d_{it} as the risk exposures of policyholder i in year t with number of claims reported Y_{it} and a vector of *a priori* variables $\mathbf{X}_{it}^T = (X_{it1}, X_{it2}, \dots, X_{itq})$, for $i = 1, 2, \dots, n$. The Poisson regression technique is commonly used to model Y_{it} , but other count distributions (see, e.g., Boucher et al., 2007) can also be used. Mathematically, the Poisson regression

model for Y_{it} can be written as

$$Y_{it} \sim \text{Poisson} \left(d_{it} \exp(\beta_0 + \sum_{m=1}^q \beta_m x_{itm}) \right), i = 1, 2, \dots, n, \quad (4.1)$$

where β_m 's are the regression coefficients. If we denote $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_q$ as the estimated regression coefficients, then we have

$$\lambda_{it} = d_{it} \exp(\hat{\beta}_0 + \sum_{m=1}^q \hat{\beta}_m x_{itm}), i = 1, 2, \dots, n \quad (4.2)$$

as the predicted *a priori* expected claim frequency for policyholder i in year t , whereby the policyholders with the same predicted *a priori* expected claim frequency are grouped into the same tariff class.

4.2.2 A Posteriori Ratemaking

4.2.2.1 Modeling of Residual Heterogeneity

In practice, insurers use an extensive list of *a priori* variables to obtain more accurate λ_{it} 's. However, the residual heterogeneity would remain in the risk classes, simply because there are many unobservable variables that have not been considered. The modeling of residual heterogeneity has been studied extensively (see, e.g., Lemaire, 1995; Denuit et al., 2007b). For example, by assuming a conditional distribution for the number of claims reported Y_{it} , the residual heterogeneity is modeled by adding a random heterogeneity term into the conditional distribution.

Specifically, let Θ_i be the random effect that captures the residual heterogeneity of policyholder i . Given $\Theta_i = \theta$, following from the Poisson distribution assumption for Y_{it} in Section 4.2.1 implies that

$$\Pr[Y_{it} = k | \Theta_i = \theta] = \exp(-\lambda_{it}\theta) \frac{(\lambda_{it}\theta)^k}{k!}, k = 0, 1, 2, \dots \quad (4.3)$$

The most commonly used distribution for Θ_i is the Gamma (a, a) distribution with the following density function:

$$f(\theta) = \frac{1}{\Gamma(a)} a^a \theta^{a-1} \exp(-a\theta), \theta \in \mathbb{R}^+, \quad (4.4)$$

where the use of Poisson-Gamma mixture leads to a Negative Binomial distribution for Y_{it} . With these specifications, we obtain $\mathbb{E}[\Theta_i] = 1$ and $\mathbb{E}[Y_{it}] = \mathbb{E}[\mathbb{E}[Y_{it}|\Theta_i]] = \lambda_{it}$. Note that other specifications of mixed Poisson models can also be used. We refer readers to Boucher et al. (2007) and Yip and Yau (2005) for further details. Moreover, time-varying random effect Θ_{it} can be used to account for the impact of BMS in reducing moral hazard. This modeling choice can be attributed to the fact that individual efforts in preventing accidents (i.e., counteracting moral hazard) are unobserved. For instance, the empirical studies conducted by Pinquet et al. (2001) and Bolancé et al. (2003) support the existence of time-dependent random effects.

4.2.2.2 Structure of BMS

As mentioned, the structure of BMS consists of BMS levels, transition rules and relativities. The descriptions for BMS levels and transition rules are detailed as follows:

- There are j BMS levels, numbered from 1 to j .
- A specified starting level is assigned to a new driver.
- Each claim-free year is rewarded by a bonus transition (i.e. the driver moves down the BMS level), whereas each reported claim is penalized by a malus transition (i.e. the driver goes up the BMS level).
- Since the number of claims in the current year and the current level occupied suffice to determine the level occupied in the next year, the BMS can be represented by a Markov Chain: the future level in year $t + 1$ depends on the

present information (the current level occupied and the number of claims in year t) but not on the past history.

- Generally, the number of transition levels imposed is easy to understand because it is applied to all policyholders independent of their *a priori* expected claim frequencies and current levels occupied.
- Nevertheless, any forms of transition rules can be used in determining the occupancy of BMS level. For instance, the old Belgian (see Pitrebois et al., 2003a) and the old Portuguese (see Centeno and Silva, 2002) systems are both dependent on the number of consecutive claim-free years. In this case, the states of BMS should be redefined in such a way that the Markovian property continues to hold.²⁰

Given the number of BMS levels and the transition rules, the optimal relativity for each level ℓ , denoted as r_ℓ can be determined through the optimization of certain objective function. For example, under the assumption of conditional Poisson distribution in modeling the number of claims, the policyholders who occupy level ℓ are required to pay an amount of premium equal to $r_\ell\%$ of the *a priori* premium, but this multiplicative form may not hold true for other count distributions. In the following two subsections, we present the modeling techniques of BMS when insurers are working with cross-sectional data and panel data, respectively.

4.2.2.3 Modeling of BMS with Cross-Sectional Data

When insurers are working with cross-sectional data, the ratemaking process is done in two separate but coherent steps, that is, BMS is used as an *a posteriori* ratemaking mechanism after the *a priori* ratemaking classification. For this kind of data, each of the subscript t in the previous notations is dropped. Besides that, the previous mathematical descriptions correspond to policyholder i , but the insurer

²⁰Fictitious levels have to be introduced to establish the Markovian property. For instance, Lemaire (1995) was the first to propose splitting certain levels into sublevels to accommodate the special bonus rule.

is considering a portfolio of policyholders rather than an individual in designing a BMS.

Suppose that a policyholder is picked at random from the portfolio. By incorporating the random effect Θ , we can express the actual expected claim frequency for this policyholder as $\Lambda\Theta$, where Λ is defined as the unknown *a priori* expected claim frequency. Note that since the random effect Θ captures the residual heterogeneity, the random variables Λ and Θ may reasonably be assumed to be mutually independent.²¹ We also assume that the policyholders are subdivided into h risk classes, where w_g is denoted as the proportion of policyholders belonging to the g -th risk class whose expected claim frequency is λ_g , i.e. $\Pr[\Lambda = \lambda_g] = w_g$.

Let us further denote the transition probability of moving from level ℓ_1 to level ℓ_2 for a policyholder with expected claim frequency $\lambda\theta$ as $p_{\ell_1\ell_2}(\lambda\theta)$, for $\ell_1, \ell_2 = 1, 2, \dots, j$. The one-step transition matrix can be written as $\mathbf{P}(\lambda\theta) = \{p_{\ell_1\ell_2}(\lambda\theta)\}$. Moreover, it can be shown that $\mathbf{P}^{(m)}(\lambda\theta) = \mathbf{P}^m(\lambda\theta)$, that is, the m -step transition matrix $\mathbf{P}^{(m)}(\lambda\theta)$ is the m -th power of the one-step transition matrix $\mathbf{P}(\lambda\theta)$. Finally, we denote the stationary distribution by $\boldsymbol{\pi}(\lambda\theta) = (\pi_1(\lambda\theta), \pi_2(\lambda\theta), \dots, \pi_j(\lambda\theta))^T$, where $\pi_\ell(\lambda\theta)$ is the stationary probability for a policyholder with expected claim frequency $\lambda\theta$ to be in level ℓ . The stationary distribution $\boldsymbol{\pi}(\lambda\theta)$ can be obtained by solving the following two conditions:

$$\begin{cases} \boldsymbol{\pi}(\lambda\theta)^T = \boldsymbol{\pi}(\lambda\theta)^T \mathbf{P}(\lambda\theta) \\ \boldsymbol{\pi}(\lambda\theta)^T \mathbf{1} = 1, \end{cases} \quad (4.5)$$

where $\mathbf{1}$ is the column vector of 1's. These modeling details are needed in Section 4 when we discuss the determination of optimal relativities and the design of transition rules.

²¹Using panel data, Boucher and Denuit (2006) showed that *a priori* covariates (and hence Λ) and Θ are correlated. However, we think that this independence assumption is reasonably appropriate (see Denuit et al., 2007b) under the context of BMS modeling with cross-sectional data.

The stationary distribution is required because in this chapter we choose to determine the set of optimal relativities on the basis of the asymptotic Norberg's criterion. This approach is different from the transient criterion proposed by Borgan et al. (1981), who argued that the age of policy should be taken into consideration especially when a relatively longer period is needed for a BMS to reach its steady state.

4.2.2.4 Modeling of BMS with Panel Data

When a panel data structure is available, the insurer can make use of the empirical experience data in designing a BMS. By showing that a two-step approach is no longer coherent in a panel data setting, Boucher and Inoussa (2014) developed a single-step approach to combine the estimation of *a priori* premiums and *a posteriori* relativities simultaneously using what they called the BMS-panel model.

Mathematically, this modeling approach involves maximizing the log-likelihood of the following joint distribution for all policyholders:

$$\begin{aligned} \Pr(Y_{i1}, \dots, Y_{is} | \mathbf{X}_{i1}, \dots, \mathbf{X}_{is}) &= \sum_{\ell(1)=1}^j \Pr(Y_{i1}, \dots, Y_{is} | \mathbf{X}_{i1}, \dots, \mathbf{X}_{is}, L(1) = \ell(1)) \Pr(L(1) = \ell(1)) \\ &= \sum_{\ell(1)=1}^j \left(\prod_{t=1}^s \Pr(Y_{it} | \mathbf{X}_{it}, L(t) = \ell(t)) \right) \Pr(L(1) = \ell(1)), \end{aligned} \tag{4.6}$$

where s is the number of years of available panel data and $L(t)$ for $t = 1, \dots, s$ is the BMS level occupied by the policyholder in year t .

Under this setup, the random heterogeneity term Θ_i is no longer required in the expression of conditional distribution for Y_{it} , instead it will be replaced by the relativity of level $\ell(t)$ denoted as $r_{\ell(t)}$. For instance, if we assume that $Y_{it} | L(t) = \ell(t), \mathbf{X}_{it}$ is Poisson distributed, then the resulting model is called Poisson-BMS-panel model (abbreviated as Poisson-BMS) with the following probability mass

function:

$$\Pr(Y_{it} = k | \mathbf{X}_{it}, L(t) = \ell(t)) = \exp(-\lambda_{it} r_{\ell(t)}) \frac{(\lambda_{it} r_{\ell(t)})^k}{k!}. \quad (4.7)$$

Hence, the use of panel data enables the insurer to estimate both Poisson regression coefficients $\hat{\beta}_m$'s and relativities r_ℓ 's in a single step by numerically maximizing the resulting log-likelihood. Moreover, the joint estimation is based on empirical considerations rather than relying on the stationary probabilities that are required in the modeling of BMS with cross-sectional data.

4.3 Objectives of An Optimal BMS

In this section, we discuss the objectives of an optimal BMS by examining the three implications from the remarks made by Taylor (1997) and describing the problems of existing BMS which lead to two inadequacy scenarios. In particular, we aim to pinpoint the important issues that should be put into consideration in designing a BMS.

4.3.1 The Three Implications

Due to the clustering of policyholders in the BMS levels according to their *a priori* risks, Taylor (1997) pointed out that a BMS should rely on the claims experience information in its design, but only restricted to the part of the information that does not reflect the observable *a priori* characteristics. In our opinion, these remarks carry the following three implications:

- implication 1: The main purpose of a BMS is to deal with the heterogeneity within each risk class but not the heterogeneity between different risk classes. This implication is due to the fact that heterogeneity remains no matter how extensive the use of *a priori* classification variables is. This implication requires the modeling of residual heterogeneity as discussed in Section 4.2.2.1.

It is also important to note that a BMS should not attempt to tackle the heterogeneity between different risk classes, which implies that neither different transition rules nor different set of optimal relativities should be imposed on the policyholders in different risk classes.

- implication 2: The information of *a priori* expected claim frequency should be incorporated into the determination of premium relativity.

Following from implication 1, the heterogeneity between different risk classes should be treated as given and be embedded into the computation of optimal relativities. This approach ensures the role of BMS as an *a posteriori* correction mechanism. Taylor (1997) considered this implication by performing simulations, followed by Pitrebois et al. (2003b) who derived an analytical formula for the optimal relativity.

- implication 3: The integration of both *a priori* and *a posteriori* ratemakings should produce a BMS in which the average *a priori* expected claim frequency appears to be as indistinguishable as possible across different BMS levels once the steady state (only applicable for the BMS with cross-sectional data) has been reached.

This implication ensures that the implementation of BMS does not further reinforce the *a priori* risk classification. It can be used to alleviate the clustering of policyholders in the BMS levels according to their *a priori* risk characteristics, which is exactly the main argument used by Taylor (1997) in his remarks. However, this implication has not been analyzed in previous studies.

There are two explanations for the clustering phenomenon of policyholders in certain BMS levels according to their *a priori* risk classifications. Firstly, it is due to the probabilistic inferences of different predicted λ_g 's for the policyholders who belong to the different tariff classes. Other things being equal, the policyholders with a higher (lower) λ_g are expected to occupy a higher (lower) BMS level. It is

therefore essential for a BMS to incorporate the information of *a priori* ratemaking into the design of BMS, particularly in determining the set of optimal relativities.

Secondly, it may be attributed to the fact that there is no differentiation of level transitions according to the *a priori* characteristics and the current level occupied. While it is indeed not necessary to enforce different transition rules according to the *a priori* characteristics based on implication 1, this explanation provides us an objective that the transition rules should be specified in such a way that implication 3 can be achieved. In addition, this objective motivates us to construct a measure of interest as a basis to compare different transition rules of BMS.

4.3.2 The Two Inadequacy Scenarios

In the implementation of BMS, the following two inadequacy scenarios arise because of the single set of optimal relativities and the simple but inflexible transition rules, respectively:

- inadequacy scenario 1: The unfairness is induced because of the identical optimal relativities applied to all the policyholders regardless of their *a priori* expected claim frequencies, as pointed out by Pitrebois et al. (2003b). As a result, given the same relativity of a level, the drivers with a higher λ_g and so a higher *a priori* premium are penalized more.
- inadequacy scenario 2: The unfairness is induced because of the identical level transitions applied to all the policyholders independent of their current levels occupied, using the similar argument as above. Since the drivers staying in higher BMS levels tend to be of a higher *a priori* risk (due to clustering), they are penalized more if they are subject to the same transition rules.

Note that if we are using credibility premium, these two inadequacy scenarios would not arise. However, as a commercial version of credibility premium, a BMS is restricted in its implementation. For instance, the set of optimal relativities is

fixed for all policyholders, while the set of transition rules has to be introduced as part of the BMS.²²

With regards to the first inadequacy scenario, Pitrebois et al. (2003b) proposed to design the optimal relativities for urban and rural drivers separately. Under this proposal, the unfairness could be reduced by imposing a set of lower optimal relativities to urban drivers who are required to pay a higher *a priori* premium. This proposed solution is equivalent to the approach of separating a single portfolio of heterogeneous drivers into a number of more homogeneous portfolios and subsequently computing the corresponding set of optimal relativities for each portfolio. In practice, however, insurers prefer to design a single set of optimal relativities for the entire portfolio of drivers due to commercial reasons. Moreover, the introduction of multiple sets of optimal relativities means that the heterogeneity between different risk classes are dealt with separately by the insurer, contradicting implication 1.

4.4 Designing an Optimal BMS with Cross-Sectional Data

The issues discussed in the previous section are crucial towards the determination of optimal relativities and the specification of transition rules for the BMS developed for cross-sectional data, which are the main subjects in this section. Specifically, we take into account the first two implications and the first inadequacy scenario in proposing a new objective function to determine the optimal relativities. Furthermore, we propose a metric called effectiveness of transition rules by considering the third implication.

On the other hand, Boucher and Inoussa (2014) argued that the first inade-

²²In computing credibility premium, there are no BMS levels, transition rules and optimal relativities. The premium is derived on an individual basis as a function of past claims experience and *a priori* expected claim frequency.

quacy scenario is mainly due to the separation of *a priori* ratemaking and *a posteriori* ratemaking. Therefore, they proposed the single-step ratemaking approach as detailed in Section 4.2.2.4 to tackle the first inadequacy scenario.

4.4.1 Determination of Optimal Relativities

Pitrebois et al. (2003b) derived an analytical formula for optimal relativity by minimizing the expected squared difference between the ‘true’ relative premium Θ and the relative premium r_L applicable to the policyholders after the steady state has been reached. Mathematically, it can be written as

$$\min \mathbb{E}[(\Theta - r_L)^2]. \quad (4.8)$$

This approach takes into account the *a priori* expected claim frequency (implication 2) in the computation but the objective function itself does not consider the obtained *a priori* premium λ_g ’s. By ignoring the amount of *a priori* premium in the objective function, the resulting optimal relativity may not minimize the expected squared difference between the ‘true’ premium $\Lambda\Theta$ and the actual premium Λr_L paid by the policyholders staying in the BMS level L .

In practice, the insurer is concerned about the collected premium in absolute terms rather than relative terms, which suggests that the amount of *a priori* premium should be incorporated into the objective function. In fact, Norberg (1976) considered the expected squared difference between the true premium and the actual premium (in absolute rather than relative terms), in which the policyholders are not separated into different risk classes and the minimization function can be further simplified as follows:

$$\min \mathbb{E}[(\bar{\lambda}\Theta - \bar{\lambda}r_L)^2] \equiv \min \mathbb{E}[(\Theta - r_L)^2], \quad (4.9)$$

where $\bar{\lambda}$ is the constant expected claim frequency for all policyholders in the ab-

sence of *a priori* risk classification. Therefore, we propose the minimization of the following objective function in the determination of optimal relativities:

$$\begin{aligned}
\mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2] &= \sum_{\ell=1}^j \mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2 | L = \ell] \Pr[L = \ell] \\
&= \sum_{\ell=1}^j \mathbb{E}[\mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2 | L = \ell, \Lambda] | L = \ell] \Pr[L = \ell] \\
&= \sum_{\ell=1}^j \sum_{g=1}^h \mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2 | L = \ell, \Lambda = \lambda_g] \Pr[\Lambda = \lambda_g | L = \ell] \Pr[L = \ell] \\
&= \sum_{\ell=1}^j \sum_{g=1}^h \int_0^{+\infty} (\lambda_g \theta - \lambda_g r_\ell)^2 \pi_\ell(\lambda_g \theta) w_g f(\theta) d\theta \\
&= \sum_{g=1}^h w_g \int_0^{+\infty} \sum_{\ell=1}^j (\lambda_g \theta - \lambda_g r_\ell)^2 \pi_\ell(\lambda_g \theta) f(\theta) d\theta.
\end{aligned} \tag{4.10}$$

Specifically, the proposed objective function in Equation (4.10) aims to minimize the expected squared difference between the true premium and the actual premium, so its solution has already taken into account the differences in *a priori* premiums, which is a major issue that leads to the first inadequacy scenario. Although the problem of overpenalization of higher *a priori* risk drivers still exists given a single set of optimal relativities²³, the solution obtained can alleviate the unfairness because the *a priori* premiums are embedded directly into the objective function in Equation (4.10).

In addition, it is important that the optimal relativity has an average of 100%, so that the bonuses and maluses exactly offset each other and result in a financial equilibrium condition. As such, the financial equilibrium constraint is added to the minimization of Equation (4.10) such that it becomes the following constrained

²³Given a single set of optimal relativities, the drivers with a higher *a priori* risk are still penalized more than the drivers with a lower *a priori* risk.

optimization:²⁴

$$\min \mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2], \quad \text{subject to } \mathbb{E}[r_L] = 1. \quad (4.11)$$

In this case, we can make use of the Lagrangian method to solve Equation (4.11). Specifically, we define the Lagrangian as

$$\begin{aligned} \mathcal{L}(\mathbf{r}, \alpha) &= \mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2] + \alpha(\mathbb{E}[r_L] - 1) \\ &= \sum_{\ell=1}^j \mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2 | L = \ell] \Pr[L = \ell] + \alpha\left(\sum_{\ell=1}^j r_\ell \Pr[L = \ell] - 1\right), \end{aligned} \quad (4.12)$$

where $\mathbf{r} = (r_1, r_2, \dots, r_\ell)^T$. We can derive the following first order conditions:

$$\Pr[L = \ell](2\mathbb{E}[\Lambda^2\Theta - \Lambda^2 r_L | L = \ell] - \alpha) = 0, \quad \text{for } \ell = 1, 2, \dots, j, \quad (4.13)$$

$$1 - \sum_{\ell=1}^j r_\ell \Pr[L = \ell] = 0. \quad (4.14)$$

The solution set for α and $r_\ell, \ell = 1, 2, \dots, j$ is given as:

$$\alpha = \frac{\left(\sum_{\ell=1}^j \frac{\mathbb{E}[\Lambda^2\Theta | L = \ell] \Pr[L = \ell]}{\mathbb{E}[\Lambda^2 | L = \ell]}\right) - 1}{\sum_{\ell=1}^j \frac{\Pr[L = \ell]}{2\mathbb{E}[\Lambda^2 | L = \ell]}}, \quad (4.15)$$

$$r_\ell = \frac{\mathbb{E}[\Lambda^2\Theta | L = \ell]}{\mathbb{E}[\Lambda^2 | L = \ell]} - \frac{\alpha}{2\mathbb{E}[\Lambda^2 | L = \ell]}, \quad (4.16)$$

²⁴In the derivation by Pitrebois et al. (2003b), the solution for optimal relativity is $r_\ell = \mathbb{E}[\Theta | L = \ell]$, so it is guaranteed that $\mathbb{E}[r_L] = \mathbb{E}[\mathbb{E}[\Theta | L]] = \mathbb{E}[\Theta] = 1$.

$$\text{where} \quad \Pr [L = \ell] = \sum_{g=1}^h w_g \int_0^{+\infty} \pi_\ell(\lambda_g \theta) f(\theta) d\theta, \quad (4.17)$$

$$\mathbb{E}[\Lambda^2 \Theta | L = \ell] = \frac{\sum_{g=1}^h w_g \int_0^{+\infty} \lambda_g^2 \theta \pi_\ell(\lambda_g \theta) f(\theta) d\theta}{\sum_{g=1}^h w_g \int_0^{+\infty} \pi_\ell(\lambda_g \theta) f(\theta) d\theta}, \quad (4.18)$$

$$\mathbb{E}[\Lambda^2 | L = \ell] = \frac{\sum_{g=1}^h w_g \int_0^{+\infty} \lambda_g^2 \pi_\ell(\lambda_g \theta) f(\theta) d\theta}{\sum_{g=1}^h w_g \int_0^{+\infty} \pi_\ell(\lambda_g \theta) f(\theta) d\theta}. \quad (4.19)$$

Note that if we relax the financial equilibrium constraint, we have $\alpha^{\text{unconstrained}} = 0$ and

$$r_\ell^{\text{unconstrained}} = \frac{\mathbb{E}[\Lambda^2 \Theta | L = \ell]}{\mathbb{E}[\Lambda^2 | L = \ell]}. \quad (4.20)$$

However, the premium relativity from the unconstrained optimization may have an average of less than 100%, which is undesirable for the insurer in the long run. In addition, if the insurer does not enforce any *a priori* ratemaking (i.e. all the λ_g 's are assumed to be $\bar{\lambda}$), we have Equations (4.15) and (4.16) respectively reduced to

$$\begin{aligned} \alpha^{\text{without a priori}} &= \frac{\left(\sum_{\ell=1}^j \frac{\bar{\lambda}^2 \mathbb{E}[\Theta | L = \ell] \Pr [L = \ell]}{\bar{\lambda}^2 \mathbb{E}[1 | L = \ell]} \right) - 1}{\sum_{\ell=1}^j \frac{\Pr [L = \ell]}{2 \bar{\lambda}^2 \mathbb{E}[1 | L = \ell]}} \\ &= \frac{\mathbb{E}[\Theta] - 1}{\frac{1}{2 \bar{\lambda}^2}} = 0 \end{aligned} \quad (4.21)$$

and

$$\begin{aligned} r_\ell^{\text{without a priori}} &= \frac{\bar{\lambda}^2 \mathbb{E}[\Theta | L = \ell]}{\bar{\lambda}^2 \mathbb{E}[1 | L = \ell]} \\ &= \mathbb{E}[\Theta | L = \ell] \\ &= \frac{\int_0^{+\infty} \theta \pi_\ell(\bar{\lambda} \theta) f(\theta) d\theta}{\int_0^{+\infty} \pi_\ell(\bar{\lambda} \theta) f(\theta) d\theta}, \end{aligned} \quad (4.22)$$

which have been derived by Norberg (1976) and Pitrebois et al. (2003b).

4.4.2 Design of Transition Rules

In this subsection, we present a new measure called effectiveness of transition rules to quantify implication 3. It can be used to compare different transition rules for the BMS with cross-sectional data. As before, let us assume that the policyholders are partitioned into h risk classes in the *a priori* classification, where $w_g = \Pr[\Lambda = \lambda_g]$ is denoted as the proportion of policyholders in the g -th risk class. In this case, we have

$$\mathbb{E}[\Lambda] = \sum_{g=1}^h \lambda_g w_g, \quad (4.23)$$

$$\mathbb{V}[\Lambda] = \sum_{g=1}^h (\lambda_g - \mathbb{E}[\Lambda])^2 w_g. \quad (4.24)$$

Pitrebois et al. (2003b) proposed the following expression to quantify the interaction of *a priori* and *a posteriori* ratemakings

$$\mathbb{E}[\Lambda|L = \ell] = \frac{\sum_{g=1}^h w_g \int_0^{+\infty} \lambda_g \pi_\ell(\lambda_g \theta) f(\theta) d\theta}{\sum_{g=1}^h w_g \int_0^{+\infty} \pi_\ell(\lambda_g \theta) f(\theta) d\theta}. \quad (4.25)$$

In more detail, this measure of interaction can be interpreted as the average *a priori* expected claim frequency in the BMS level $L = \ell$. From implication 3, we understand that a good BMS should exhibit small variations of $\mathbb{E}[\Lambda|L = \ell]$ across different BMS levels. Hence, we decompose $\mathbb{V}[\Lambda]$ as a summation of two components:

$$\begin{aligned} \mathbb{V}[\Lambda] &= \mathbb{E}[\mathbb{V}[\Lambda|L]] + \mathbb{V}[\mathbb{E}[\Lambda|L]] \\ &= \sum_{\ell=1}^j \left(\sum_{g=1}^h (\lambda_g - \mathbb{E}[\Lambda|L = \ell])^2 w_g \int_0^{+\infty} \pi_\ell(\lambda_g \theta) f(\theta) d\theta \right) \\ &\quad + \sum_{\ell=1}^j \left(\mathbb{E}[\Lambda|L = \ell] - \mathbb{E}[\Lambda] \right)^2 \Pr[L = \ell], \end{aligned} \quad (4.26)$$

where the first component $\mathbb{E}[\mathbb{V}[\Lambda|L]]$ is the expected value of the conditional variances and the second component $\mathbb{V}[\mathbb{E}[\Lambda|L]]$ is the variance of the conditional means.

Given the value of $\mathbb{V}[\Lambda]$, implication 3 means that the variations should be mainly driven by the first component, which is equivalent to say that the policyholders staying in each BMS level should have large variations in the *a priori* expected claim frequency. In other words, a BMS should provide sufficient flexibility such that the policyholders from any *a priori* risk classes would occupy any BMS levels in the steady state. A larger value of the first component then implies a smaller value of the second component, which indicates that the BMS does not further reinforce the *a priori* risk classification.

It can be seen that the value of the first component is dependent on the number of risk classes h , the values of *a priori* expected claim frequency λ_g 's, the number of BMS levels j and the specification of transition rules, whereas the value of $\mathbb{V}[\Lambda]$ depends only on the first two factors. Since the number of BMS levels is usually fixed in advance, we propose to use

$$\tau_{\text{rule}} = \frac{\mathbb{E}[\mathbb{V}[\Lambda|L]]}{\mathbb{V}[\Lambda]} \quad (4.27)$$

as a measure of the effectiveness of transition rules. Other things being equal, a larger value of τ_{rule} is preferred because of the arguments of implication 3. The insurer may consider using this metric to compare different candidates of transition rules. Alternatively, the insurer could specify a minimum threshold value for τ_{rule} such that the set of transition rules to be used in the BMS should satisfy the criterion.

4.4.3 Varying Transition Rules

With a set of simple transition rules, both the single-step and two-step approaches are not able to tackle the second inadequacy scenario. To overcome this problem, we argue that a set of varying transition rules may help reduce the overpenalization on the drivers staying in the higher BMS levels. The logic behind is to introduce a set of varying transition rules based on the current levels occupied to correct the tendency of those policyholders occupying the higher BMS levels (who tend to have a higher λ_g) to continue staying in the higher BMS levels. As a result, varying transition rules may lead to a fulfilment of implication 3. In more detail, it can be achieved by imposing a lower malus transition and a higher bonus transition on those policyholders occupying the higher BMS levels instead of a fixed number of level transitions for all policyholders. Conversely, the varying transition rules could also impose a higher malus transition and a lower bonus transition on the policyholders occupying the lower BMS levels (who tend to have a lower λ_g).

Let us denote $t_{\ell,k}$ as the effective level transition²⁵ imposed on the policyholders occupying level ℓ in the BMS and making k claims in the current year, for $\ell = 1, 2, \dots, j$ and $k \geq 0$. The varying transition rules should satisfy the following structure:

- malus transition: a non-negative effective malus transition²⁶, and a smaller effective malus transition for the drivers occupying the higher BMS levels

$$\begin{aligned} t_{\ell,k} &\geq 0, & \text{for } k \geq 1, \\ t_{\ell_2,k} &\leq t_{\ell_1,k}, & \text{for } k \geq 1, \ell_2 > \ell_1. \end{aligned} \tag{a}$$

- bonus transition: a non-positive effective bonus transition, and a larger ef-

²⁵Since the BMS consists of a finite number of levels, the actual level transition for the policyholders staying in the top or bottom levels may not be the same as the level transition specified by the transition rules. Since it is the effective level transition that will be considered in the transition matrix, we choose to work with the effective level transition.

²⁶The strict inequality holds true for certain BMS levels but not all BMS levels, due to the fact that the BMS consists of a finite number of levels. In the same vein, all the inequalities presented here are non-strict inequalities.

fective bonus transition (in magnitude) for the drivers occupying the higher BMS levels,

$$\begin{aligned} t_{\ell,0} &\leq 0, \\ |t_{\ell_2,0}| &\geq |t_{\ell_1,0}|, \text{ for } \ell_2 > \ell_1. \end{aligned} \tag{b}$$

As an example, we propose the following set of transition rules

$$\begin{aligned} t_{\ell,0} &= \begin{cases} 0, & \text{for } \ell = 1, \\ -1, & \text{for } 2 \leq \ell \leq \lceil \frac{j}{2} \rceil + 1, \\ -2, & \text{for } \ell > \lceil \frac{j}{2} \rceil + 1, \end{cases} \\ t_{\ell,k} &= \begin{cases} \min \left[j - \ell, \max \left[k, \lceil \frac{j-\ell}{p} \times k \rceil \right] \right], & \text{for } k \geq 1, \ell < j, \\ 0, & \text{for } k \geq 1, \ell = j, \end{cases} \end{aligned} \tag{4.28}$$

where $\lceil x \rceil$ is the ceiling function of x , which is used because the level transition has to be an integer. If the policyholders are currently staying in level 1, no effective bonus transition can be given. We propose using at most two levels of bonus transition due to practical consideration. The cut-off level of $\lceil j/2 \rceil + 1$ is chosen to ensure that more levels are subject to one level bonus transition, so that the transition rules will not be too soft.²⁷ For the effective malus transition, the insurer is given the freedom to specify p , which is the smallest number of claims required for the drivers to move from level ℓ to level j . The maximum function in $t_{\ell,k}$ ensures that there is at least one level of malus transition for each claim reported, while the minimum function in $t_{\ell,k}$ limits the number of malus transitions to $j - \ell$. Besides that, there is no effective malus transition for the policyholders who are staying in the highest level j .

With these specifications, the varying level transitions imposed in the BMS are both sufficiently flexible and commercial. The varying rules can be implemented in practice as long as the clear specification of transition rules are provided, such

²⁷The numbers of BMS levels that are subject to 0, 1 and 2 levels of bonus transitions are 1, $\lceil \frac{j}{2} \rceil$ and $j - 1 - \lceil \frac{j}{2} \rceil$, respectively, where $\lceil \frac{j}{2} \rceil > j - 1 - \lceil \frac{j}{2} \rceil$ for all j .

as the special bonus rules introduced in the old Belgian BMS (see Pitrebois et al., 2003a). Note that the proposed functional form is not unique, so any appropriate functional forms of transition rules that satisfy conditions (a) and (b) can be used. In other words, the specification of varying transition rules can be of any design that fulfill the two conditions in (a) and (b) instead of the rules in the form of -1 or $-2/+k$ as discussed above. A more important consideration is to ensure that the effectiveness of the specified varying rules is higher than the corresponding simple rules.

4.5 Numerical Illustrations: BMS with Cross-Sectional Data

In this section, we use the cross-sectional claims data set as studied by Pitrebois et al. (2004) for illustration purposes, where the policyholders are partitioned into 24 risk classes through the use of risk classification variables. The predicted $\lambda_g, g = 1, 2, \dots, 24$ and the corresponding w_g 's are listed in their work. The estimated parameter a of the Gamma distribution for Θ is 1.2401. With these details, we obtain $\mathbb{E}[\Lambda] = 0.1462$ and $\mathbb{V}[\Lambda] = 0.002625$.

We consider a BMS that consists of $j = 9$ levels, similar to the BMS as discussed by Taylor (1997). Moreover, we analyze four sets of transition rules, two of which are simple while the other two of which are varying. In particular, we compare each set of simple rules with a corresponding set of varying rules. For instance, the simple rules of $-1/+2$ is compared against the varying rules of $p = 4$ according to Equation (4.28).²⁸ Similarly, the simple rules of $-1/+3$ is compared against the varying rules of $p = 3$.

The effective level transitions for these rules are shown in Tables 4.1 and 4.2.

²⁸The basis of comparison is the smallest number of claims required to move to the highest level j . For example, the rules of $-1/+2$ require $p = 4$ claims for the drivers occupying level 1 to move to level 9.

Starting level	Level occupied if				
	0	1	2	3	≥ 4
	claims are reported				
9	8 (7)	9 (9)	9 (9)	9 (9)	9 (9)
8	7 (6)	9 (9)	9 (9)	9 (9)	9 (9)
7	6 (5)	9 (8)	9 (9)	9 (9)	9 (9)
6	5 (5)	8 (7)	9 (8)	9 (9)	9 (9)
5	4 (4)	7 (6)	9 (7)	9 (8)	9 (9)
4	3 (3)	6 (6)	8 (7)	9 (8)	9 (9)
3	2 (2)	5 (5)	7 (6)	9 (8)	9 (9)
2	1 (1)	4 (4)	6 (6)	8 (8)	9 (9)
1	1 (1)	3 (3)	5 (5)	7 (7)	9 (9)

Table 4.1. Transition rules for the simple rules of $-1/+2$ (varying rules of $p = 4$).

Starting level	Level occupied if			
	0	1	2	≥ 3
	claims are reported			
9	8 (7)	9 (9)	9 (9)	9 (9)
8	7 (6)	9 (9)	9 (9)	9 (9)
7	6 (5)	9 (8)	9 (9)	9 (9)
6	5 (5)	9 (7)	9 (8)	9 (9)
5	4 (4)	8 (7)	9 (8)	9 (9)
4	3 (3)	7 (6)	9 (8)	9 (9)
3	2 (2)	6 (5)	9 (7)	9 (9)
2	1 (1)	5 (5)	8 (7)	9 (9)
1	1 (1)	4 (4)	7 (7)	9 (9)

Table 4.2. Transition rules for the simple rules of $-1/+3$ (varying rules of $p = 3$).

For each pair of comparable rules, it is evident that the varying rules are more flexible than the simple rules, as shown by the smaller effective malus transitions and the larger effective bonus transitions for the drivers occupying the higher BMS levels. Also, it can be seen that the rules of $-1/+2$ and $p = 4$ are softer than the rules of $-1/+3$ and $p = 3$.

For each set of transition rules, the values of $\Pr[L = \ell]$ and $\mathbb{E}[\Lambda|L = \ell]$ are tabulated in Table 4.3. In comparison with the corresponding simple rules, the proportion of drivers staying in the lower BMS levels (in this case, $\ell \leq 5$) under the varying rules are higher, while the reverse is true for the higher BMS levels. In addition, the drivers tend to occupy the three lowest BMS levels under the softer rules of $-1/+2$ and $p = 4$ as compared to the more severe rules of $-1/+3$ and

level ℓ	$-1/+2$		Varying $p = 4$		$-1/+3$		Varying $p = 3$	
	$\Pr[L = \ell]$	$\mathbb{E}[\Lambda L = \ell]$	$\Pr[L = \ell]$	$\mathbb{E}[\Lambda L = \ell]$	$\Pr[L = \ell]$	$\mathbb{E}[\Lambda L = \ell]$	$\Pr[L = \ell]$	$\mathbb{E}[\Lambda L = \ell]$
9	2.87%	19.26%	0.70%	21.35%	5.06%	18.02%	1.19%	20.10%
8	2.35%	18.27%	0.54%	19.23%	4.09%	17.21%	1.12%	18.07%
7	2.31%	17.37%	1.26%	18.73%	4.17%	16.36%	2.83%	17.38%
6	2.31%	16.78%	2.34%	17.34%	3.94%	15.90%	2.14%	16.59%
5	3.20%	16.01%	4.24%	16.70%	3.72%	15.58%	5.51%	16.28%
4	3.37%	15.68%	4.17%	16.17%	7.29%	14.83%	9.18%	15.25%
3	8.09%	14.90%	8.97%	15.16%	6.20%	14.70%	7.57%	15.04%
2	6.84%	14.77%	7.44%	14.97%	5.33%	14.59%	6.34%	14.87%
1	68.66%	13.97%	70.34%	14.03%	60.21%	13.87%	64.13%	13.97%

Table 4.3. Interaction between a priori and a posteriori ratemakings.

Transition rules	τ_{rule}
$-1/+2$	93.87%
varying $p = 4$	94.74%
$-1/+3$	94.56%
varying $p = 3$	95.18%

Table 4.4. Effectiveness of transition rules.

$p = 3$. We also find that the value of $\mathbb{E}[\Lambda|L = \ell]$ increases with the BMS level ℓ for all four sets of transition rules, providing supporting evidence for the remarks made by Taylor (1997). Particularly, the values of $\mathbb{E}[\Lambda|L = \ell]$ under the rules of $-1/+3$ are the smallest for every BMS level. This is largely due to its severe malus transitions, whereby it would result in a larger proportion of drivers staying in the higher BMS levels, hence the average *a priori* expected claim frequency would be smaller.

More importantly, we should jointly consider the values of $\Pr[L = \ell]$ and $\mathbb{E}[\Lambda|L = \ell]$ to derive meaningful interpretations regarding the complementary role of *a posteriori* ratemaking. We examine the effectiveness of transition rules in alleviating the clustering problem of policyholders by considering the values of τ_{rule} shown in Table 4.4. We observe that both the varying transition rules are more effective than the corresponding simple rules. This can be explained by the fact that the varying level transitions imposed are able to produce a higher variation of λ_g 's in each BMS level and thus resulting in a higher value of $\mathbb{E}[\mathbb{V}[\Lambda|L]]$. On the contrary, the simple rules may lack sufficient flexibility to achieve implication 3.

level ℓ	-1/+2		Varying $p = 4$		-1/+3		Varying $p = 3$	
	r_ℓ	$r_\ell^{\text{unconstrained}}$	r_ℓ	$r_\ell^{\text{unconstrained}}$	r_ℓ	$r_\ell^{\text{unconstrained}}$	r_ℓ	$r_\ell^{\text{unconstrained}}$
9	266.57%	261.07%	420.41%	419.11%	233.67%	227.73%	293.04%	288.85%
8	235.37%	229.27%	317.11%	315.62%	205.54%	199.03%	230.71%	225.56%
7	207.95%	201.22%	307.54%	305.95%	176.56%	169.34%	211.92%	206.36%
6	188.51%	181.29%	234.82%	233.11%	159.35%	151.70%	183.85%	177.74%
5	162.42%	154.47%	215.07%	213.20%	146.58%	138.59%	173.05%	166.69%
4	150.04%	141.74%	186.85%	184.91%	115.41%	106.55%	133.57%	126.30%
3	118.70%	109.49%	127.08%	125.08%	109.31%	100.29%	124.15%	116.67%
2	112.85%	103.46%	118.27%	116.22%	103.93%	94.75%	116.18%	108.51%
1	72.92%	62.35%	69.43%	67.10%	66.34%	56.11%	70.86%	62.08%
$\mathbb{E}[r_L]$	100%	90.20%	100%	97.79%	100%	90.68%	100%	91.90%

Table 4.5. Optimal relativities for the BMS with cross-sectional data.

In practice, more sets of transition rules should be compared especially when the number of BMS levels j is much higher, and those rules that have a τ_{rule} value larger than the pre-specified threshold can be used in the BMS. However, it should be noted that the value of τ_{rule} would be different if the *a priori* risk classification is carried out in a different way.²⁹

Next, we perform the calculation of optimal relativities for the BMS under each set of transition rules. The results are displayed in Table 4.5, where the resulting value of $\mathbb{E}[r_L]$ is also shown for each set of optimal relativities. Note that r_ℓ and $r_\ell^{\text{unconstrained}}$ correspond to the determination of optimal relativities with and without the financial equilibrium constraint, respectively. For each set of transition rules, the values of r_ℓ are higher than $r_\ell^{\text{unconstrained}}$. This observation is supported by the finding that the values of $\mathbb{E}[r_L]$ for unconstrained optimal relativities are all below 100%. In other words, when the financial equilibrium constraint is imposed, higher optimal relativities are required such that $\mathbb{E}[r_L] = 100\%$ is obtained.

The set of optimal relativities under the transition rules of -1/+3 are considerably lower for every level ℓ , mainly because the policyholders are more likely to occupy the higher BMS levels when the claims are more heavily penalized and hence a lower set of optimal relativities are required to achieve financial equilibrium condition. In contrast, the varying transition rules with $p = 4$ have the highest

²⁹For instance, if the policyholders are partitioned into more or fewer risk classes, the value of τ_{rule} would change accordingly and may not be above 90% as shown in our illustrations.

optimal relativities for all BMS levels except the lowest level $\ell = 1$. One possible explanation is that the probability of drivers occupying the higher BMS levels are the smallest under this set of transition rules, thus yielding a larger optimal relativities for these BMS levels.

4.6 Numerical Illustrations: BMS with Panel Data

For illustration purposes in this section, we make use of the simulated panel data set provided by De Jong and Heller (2008). The data set contains the claim counts of 40000 policies for a consecutive period of three years. The risk classification variables are driver's and vehicle's value, each with six categories, resulting in a total of $6 \times 6 = 36$ risk classes.

As before, we consider a BMS of $j = 9$ levels and the four sets of transition rules as shown in Tables 4.1 and 4.2. We further assume that each driver is a new driver in the BMS. In other words, we set $L(1)$ equals to the specified entry level, which is to be varied in selecting between different structures of BMS. We consider two forms of relativities, namely general relativities r_ℓ^{general} and linear relativities $r_\ell^{\text{linear}} = \alpha + \delta(\ell - 1)$ (see Gilde and Sundt, 1989). The linear form of relativities is preferred in practice because the relativities for two adjacent levels are separated by a constant amount δ , thus avoiding numerical values of general relativities that may exhibit irregular patterns. Moreover, we consider a Poisson-BMS model in maximizing the log-likelihood of the joint distribution expressed in Equation (4.6).

The values of maximized log-likelihood obtained through numerical procedures are tabulated in Table 4.6 for different possible combinations of entry level, form of relativities and transition rules. Firstly, as expected, imposing the restricted form of linear relativities yields a lower log-likelihood value than the case of general

Entry level $L(1)$	$r_{\ell}^{\text{general}}$				r_{ℓ}^{linear}			
	$-1/+2$	Varying $p = 4$	$-1/+3$	Varying $p = 3$	$-1/+2$	Varying $p = 4$	$-1/+3$	Varying $p = 3$
1	-68733	-68319	-69423	-68900	-69918	-69847	-71014	-70903
2	-67822	-67569	-68519	-68038	-70213	-70199	-70838	-70766
3	-68453	-67572	-69497	-68060	-72778	-72954	-72910	-72869

Table 4.6. Values of maximized log-likelihoods for different combinations of entry level, relativities and transition rules under the Poisson-BMS model.

relativities, whereby the difference can be interpreted as the tradeoff for having desirable linear relativities in the implementation of BMS. Secondly, the two sets of varying rules give higher log-likelihood values than their corresponding simple rules for different combinations of entry level and form of relativities, except for the comparison between the varying rules of $p = 4$ and the simple rules of $-1/+2$ under the BMS structure of $L(1) = 3$ and linear relativities. This finding suggests another advantage of varying transition rules over simple transition rules in terms of the empirical log-likelihood. Finally, the most optimal BMS structure for both forms of relativities have the same varying rules of $p = 4$ with different entry levels ($L(1) = 2$ for general relativities; $L(1) = 1$ for linear relativities).

In addition, to compare the different specifications of simple and varying transition rules, we identify the following optimal structures of BMS for different combinations of relativities and transition rules based on the value of log-likelihood:

- general relativities and simple rules: $L(1) = 2, -1/+2$;
- general relativities and varying rules: $L(1) = 2, \text{varying } p = 4$;
- linear relativities and simple rules: $L(1) = 1, -1/+2$;
- linear relativities and varying rules: $L(1) = 1, \text{varying } p = 4$.

The values of optimal relativities obtained from the single-step Poisson-BMS model for these optimal structures are shown in Table 4.7.

For the case of general relativities $r_{\ell}^{\text{general}}$, we observe different magnitude of increase associated with the level of BMS, indicating the drawback of irregular

level ℓ	r_ℓ^{general}		r_ℓ^{linear}	
	$L(1) = 2, -1/+2$	$L(1) = 2, \text{varying } p = 4$	$L(1) = 1, -1/+2$	$L(1) = 1, \text{varying } p = 4$
9	261.16%	230.98%	733.48%	591.83%
8	91.01%	97.00%	647.66%	522.45%
7	101.56%	59.72%	561.84%	453.07%
6	64.75%	52.44%	476.01%	383.69%
5	49.26%	39.23%	390.19%	314.31%
4	32.52%	25.96%	304.36%	244.93%
3	23.35%	18.62%	218.54%	175.56%
2	13.99%	11.15%	132.72%	106.18%
1	7.22%	5.77%	46.89%	36.80%

Table 4.7. Optimal relativities for the BMS under the Poisson-BMS model.

patterns in the numerical values obtained in terms of abrupt rise or fall from one level to another. Furthermore, the values of r_ℓ^{general} for $\ell = 1, \dots, 9$ obtained under the varying rules of $p = 4$ are strictly increasing, while the values of r_ℓ^{general} under the simple rules of $-1/+2$ are not.³⁰ The results suggest that imposing varying rules may offer sufficient flexibility with regards to the determination of optimal relativities. In contrast, the linear form of relativities ensures a steady increase in the values of r_ℓ^{linear} .

On the other hand, it is easy to see that the values of r_ℓ^{linear} are significantly higher than that of r_ℓ^{general} . This is mainly due to the technique of single-step estimation for both the *a priori* premiums and *a posteriori* relativities using a panel data structure, where the estimation provides a set of numerical values consisting of $\hat{\beta}_m$, for $m = 1, \dots, q$ and r_ℓ^{general} (or r_ℓ^{linear}), for $\ell = 1, \dots, j$ that maximizes the resulting log-likelihood. In other words, we should jointly consider the values of $\hat{\beta}_m$ and r_ℓ^{general} (or r_ℓ^{linear}) in comparing between different structures of BMS.

The estimated Poisson regression coefficients $\hat{\beta}_m$'s are displayed in Table 4.8. Note that $\hat{\beta}_0$ represents the estimated intercept and $\hat{\beta}_m$, for $m = 1, \dots, 5$ (for $m = 6, \dots, 10$) correspond to the categorical variable of driver's age (vehicle's value). Other things being equal for each BMS structure, the drivers in the youngest age category are charged a higher *a priori* premium because of the negative values

³⁰It should be noted that this finding could not be generalized because the values are obtained from numerical procedures.

$\hat{\beta}_m$	r_ℓ^{general}		r_ℓ^{linear}	
	$L(1) = 2, -1/+2$	$L(1) = 2, \text{varying } p = 4$	$L(1) = 1, -1/+2$	$L(1) = 1, \text{varying } p = 4$
$\hat{\beta}_0$	0.5692	0.7915	-1.0209	-0.7806
$\hat{\beta}_1$	-0.1080	-0.1048	-0.1104	-0.1101
$\hat{\beta}_2$	-0.1749	-0.1699	-0.1773	-0.1762
$\hat{\beta}_3$	-0.2669	-0.2585	-0.2816	-0.2809
$\hat{\beta}_4$	-0.2055	-0.2063	-0.2150	-0.2151
$\hat{\beta}_5$	-0.1559	-0.1565	-0.1578	-0.1574
$\hat{\beta}_6$	0.1039	0.1065	0.1217	0.1226
$\hat{\beta}_7$	0.0760	0.0924	0.1120	0.1191
$\hat{\beta}_8$	-0.3361	-0.3278	-0.4976	-0.5043
$\hat{\beta}_9$	-0.2526	-0.2807	-0.1878	-0.1798
$\hat{\beta}_{10}$	-0.8710	-0.8548	-0.9834	-0.9515

Table 4.8. Estimated Poisson regression coefficients for the BMS under Poisson-BMS model.

of $\hat{\beta}_m$, for $m = 1, \dots, 5$. On the other hand, in comparison to the lowest value category, the drivers with a vehicle in the top three highest value categories (correspond to $\hat{\beta}_8, \hat{\beta}_9$ and $\hat{\beta}_{10}$) are required to pay a smaller premium, while the impact is reversed for the drivers with a vehicle in the two intermediate value categories (correspond to $\hat{\beta}_6$ and $\hat{\beta}_7$).

Moreover, we find that the values of $\hat{\beta}_0$ differ significantly depending on the form of relativities. For example, the value of $\hat{\beta}_0 = 0.7915$ for the BMS structure of $L(1) = 2$ and the varying rules of $p = 4$ under general relativities implies a multiplicative constant of $\exp(0.7915) = 2.2067$ onto the *a priori* premiums. In contrast, the multiplicative constant is only $\exp(-1.0209) = 0.3603$ for the case of $L(1) = 1$ and the simple rules of $-1/+2$ under linear relativities. The results indicate that the differences of $\hat{\beta}_0$ obtained under these BMS structures can be used to explain the significantly different values of r_ℓ^{linear} as compared to r_ℓ^{general} . Nevertheless, it can be seen that the other regression coefficients are numerically similar under different BMS structures, except for some notable differences in the values of $\hat{\beta}_8, \hat{\beta}_9$ and $\hat{\beta}_{10}$ depending on the form of relativities.

4.7 Conclusion

In this chapter, we argued that there are two inadequacy scenarios in a BMS that consists of a single set of optimal relativities and a set of simple transition rules. The underlying reason of the first inadequacy scenario is the separation of the *a priori* expected claim frequency $\lambda_{i,t}$ (or λ_i for cross-sectional data) from the relativity r_ℓ in the estimation of *a posteriori* relativities. For this reason, in determining the set of optimal relativities, Boucher and Inoussa (2014) developed the single-step ratemaking approach that is applicable for panel data, while we proposed a new objective function that builds upon the asymptotic Norberg's criterion to be applied to cross-sectional data.

In addition, considering the third implication from the remarks by Taylor (1997), we introduced a metric to compare different specifications of transition rules. Due to the flexibility of varying transition rules, we argued that varying rules may be more effective than their corresponding simple rules in tackling the second inadequacy scenario. By analyzing two kinds of BMS frameworks depending on the data structure, we provided numerical illustrations to support the proposal of implementing varying transition rules in designing a BMS.

For the design of BMS with cross-sectional data, the set of optimal relativities is determined by using the stationary distribution on the basis of the asymptotic Norberg's criterion. In practice, however, policyholders may take a long period of time to reach the steady state, hence the premium relativities obtained may not be adequate in the short run. Under these circumstances, insurers could replace the stationary distribution $\boldsymbol{\pi}(\lambda\theta)$ with the scaled column sum of m -th step transition matrix $\frac{\sum_{\ell_1=1}^j p_{\ell_1\ell_2}(\lambda\theta)}{\sum_{\ell_2=1}^j \sum_{\ell_1=1}^j p_{\ell_1\ell_2}(\lambda\theta)}$, which represents the probability of moving to level ℓ_2 after m years. In more detail, insurer may re-optimize the design of BMS every year with the arrival of new data, simply by substituting the value of $\boldsymbol{\pi}(\lambda\theta)$ in the optimization procedure, such that the other steps as discussed in 4.4 can be performed in the same way to obtain the set of optimal relativities.

As mentioned, the value of $\mathbb{V}[\Lambda]$ is dependent on the number of risk classes h and the values of *a priori* expected claim frequency λ_g 's. Suppose the insurer categorizes the policyholders into a number of very different risk classes given a set of more useful *a priori* variables such that the variability of Λ is much larger. In this case, the appropriate transition rules with an effectiveness larger than the pre-specified threshold may be different from the case of having a smaller $\mathbb{V}[\Lambda]$.³¹ It is warranted in future research to investigate how the varying extent of *a priori* risk classification would affect the corresponding choice of sufficiently effective transition rules.

³¹For instance, the use of a smaller set or less useful information content carried by the classification variables may lead to a smaller value of $\mathbb{V}[\Lambda]$.

Chapter 5

Concluding Remarks and Future Research

In this thesis, we first considered the issues associated with the structures of period and cohort survival curves in Chapter 2. We demonstrated the construction of hybrid survival curves through a hybrid approach that retain the strengths of both period and cohort survival curves. For this reason, the hybrid survival curve may potentially be more suitable in conducting empirical investigation on the rectangularization phenomenon. Besides that, we presented two stochastic survival models to model the evolution of a hybrid survival curve.

We do not further examine the hybrid survival curves via the existing indicator measures as reviewed by Cheung et al. (2005). This provides an avenue for further research to study whether the use of the hybrid survival curve could yield different conclusions regarding the rectangularization phenomenon on the basis of those indicator measures. In particular, since the structure of hybrid survival curve is different from that of period and cohort survival curves, modifications or adaptations may be needed in defining certain existing indicator measures. Alternatively, new measures that are specific to hybrid survival curve could be developed in the future. Besides that, the model expressions in Equations (2.5) and (2.6) should

also be applicable to period and cohort survival curves. In future research, it is warranted to investigate the practicality and usefulness of extending the proposed survival models to those two types of survival curves.

In recent years, the joint modeling of mortality rates for multiple populations has become an important issue, partly because of the biological consistency between populations that are closely related to each other. For example, gender specific populations of the same country would be subject to similar change in mortality rates and exhibit an appropriate level of biological reasonableness due to their exposures to a common set of mortality drivers. In future research, it is warranted to consider the joint modeling of hybrid survival curves for handling multiple populations. For instance, the three-factor survival model could be extended to accommodate multiple populations in the same vein as the augmented common factor model in Li and Lee (2005). Besides that, the time series modeling of the time-varying parameters for multiple populations may be further examined by taking into consideration the correlation of parameters obtained from different populations.

To overcome the challenge of concentrating liquidity in the capital market for longevity risk trading, in Chapter 3 we studied the construction of mortality indexes using the time-varying parameters obtained from adapted mortality models. We recommended using the 3-factor CBD mortality indexes estimated from Model M7* to construct mortality indexes. In addition, we developed the key K-duration method to calibrate a longevity hedge formed by K-forward contracts, in which the payoff of each K-forward is linked to a 3-factor CBD mortality index.

The proposed KKD method may be improved in future research in a number of ways. For instance, similar to the recent contributions by Lin and Tsai (2013), Tsai et al. (2010) and Wang et al. (2010), a convexity component can be added to improve the approximation of a portfolio's sensitivity to the 3-factor CBD mortality indexes. Also, following the footsteps of Cairns (2011), it would be interesting to

extend the KKD hedging framework from static to dynamic, permitting hedgers to periodically adjust their K-forward portfolios. Periodic adjustments allow a K-forward portfolio to reflect the changes in KKD values over time due to, for example, fluctuations in expected interest and mortality rates, thereby possibly leading to even better hedging results.

A major concerning issue in using a standardized K-forward hedge is population basis risk, which arises from the differences between the mortality experience of the hedger's population and the population to which the K-forwards are linked. Recently, a number of multi-population stochastic mortality models have been developed by researchers including Cairns et al. (2011a), Dowd et al. (2011), Jarner and Kryger (2011), Li and Lee (2005), Yang and Wang (2013) and Zhou et al. (2013b, 2014) to address the issue of population basis risk. In addition, Coughlan et al. (2011) and Li and Hardy (2011) proposed frameworks for quantifying population basis risk, while the Institute and Faculty of Actuaries (IFoA) and the Life and Longevity Markets Association (LLMA) founded the Longevity Basis Risk Working Group (LBRWG) to develop a market-friendly method for assessing population basis risk underlying longevity transactions. Further research is required to find out how population basis risk can be incorporated in the KKD hedging strategy and how it affects the effectiveness of a K-forward hedge.

In Chapter 4, we analyzed the two frameworks of BMS modeling in mitigating the two inadequacy scenarios. With respect to the modeling of BMS when insurers are working with cross-sectional data, we proposed a new objective function in the determination of optimal relativities and also introduced the effectiveness of transition rules to compare different specifications of transition rules. Furthermore, we suggested to implement a set of varying transition rules in a BMS because it provides sufficient flexibility in alleviating the second inadequacy scenario, as supported by the numerical illustrations applied to both BMS frameworks.

Throughout our discussion, we considered a quadratic loss function in deter-

mining the optimal relativities for the BMS developed for cross-sectional data. An avenue for further research is to incorporate different types of loss functions such as exponential loss function in the determination of optimal relativities (see, e.g., Denuit and Dhaene, 2001; Denuit et al., 2007b) via the proposed objective function in Equation (4.10). Moreover, apart from the general form of relativities that we considered in Section 4.4, we could also impose the linear form of relativities into the optimization of the proposed objective function. Under these different combinations of assumed loss functions and form of relativities, it may be possible to derive analytical formulae for the optimal relativities that also satisfy the financial equilibrium condition.

Furthermore, we assume that the number of BMS levels j is specified in advance. It is possible to relax this pre-specification by adding j as a parameter in the optimization procedure. Specifically, the objective of the optimization procedure is to select both the number of levels j and the set of optimal relativities r_L that minimize the value of $\mathbb{E}[(\Lambda\Theta - \Lambda r_L)^2]$. In future research, it would be interesting to examine the marginal improvement of introducing the number of levels j as an extra parameter into the optimal design of BMS.

Appendix

A. Proof of Identical Maximum Likelihood Estimates for the Two Time-Varying Parameters in Both Stochastic Survival Models

Here we prove that the maximum likelihood estimates of $\sigma_t^{(i)}$, for $i = 1, 2$ are identical in both survival models due to the specification of functions of $(x - \bar{x})$.

For brevity, we denote the response variable in Equations (2.5) and (2.6) as $V_{n,t}$ and the realized value as $v_{n,t}$. We can then express Equations (2.5) and (2.6) as follows:

$$V_{N \times w} = U_{N \times 2}^{2\text{-factor}} \Sigma_{2 \times w}^{2\text{-factor}} + \epsilon_{N \times w}^{2\text{-factor}},$$

$$V_{N \times w} = U_{N \times 3}^{3\text{-factor}} \Sigma_{3 \times w}^{3\text{-factor}} + \epsilon_{N \times w}^{3\text{-factor}},$$

where w is the number of hybrid survival curves constructed in the sample period, N is the maximum length of survival under consideration (throughout Section 2, we assume $N = 31$, but it can be generalized),

$$\begin{aligned}
\mathbf{V}_{N \times w} &= \begin{pmatrix} v_{1,t_1} & v_{1,t_2} & \cdots & v_{1,t_w} \\ v_{2,t_1} & v_{2,t_2} & \cdots & v_{2,t_w} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N,t_1} & v_{N,t_2} & \cdots & v_{N,t_w} \end{pmatrix}, \\
\mathbf{U}_{N \times 2}^{2\text{-factor}} &= \begin{pmatrix} 1 & (x_1 - \bar{x}) \\ 1 & (x_2 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_N - \bar{x}) \end{pmatrix}, & \mathbf{\Sigma}_{2 \times w}^{2\text{-factor}} &= \begin{pmatrix} \sigma_{t_1}^{(1)} & \sigma_{t_2}^{(1)} & \cdots & \sigma_{t_w}^{(1)} \\ \sigma_{t_1}^{(2)} & \sigma_{t_2}^{(2)} & \cdots & \sigma_{t_w}^{(2)} \end{pmatrix}, \\
\mathbf{U}_{N \times 3}^{3\text{-factor}} &= \begin{pmatrix} 1 & (x_1 - \bar{x}) & ((x_1 - \bar{x})^2 - s^2) \\ 1 & (x_2 - \bar{x}) & ((x_2 - \bar{x})^2 - s^2) \\ \vdots & \vdots & \vdots \\ 1 & (x_N - \bar{x}) & ((x_N - \bar{x})^2 - s^2) \end{pmatrix}, & \mathbf{\Sigma}_{3 \times w}^{3\text{-factor}} &= \begin{pmatrix} \sigma_{t_1}^{(1)} & \sigma_{t_2}^{(1)} & \cdots & \sigma_{t_w}^{(1)} \\ \sigma_{t_1}^{(2)} & \sigma_{t_2}^{(2)} & \cdots & \sigma_{t_w}^{(2)} \\ \sigma_{t_1}^{(3)} & \sigma_{t_2}^{(3)} & \cdots & \sigma_{t_w}^{(3)} \end{pmatrix},
\end{aligned}$$

and $\boldsymbol{\epsilon}_{N \times w}^{2\text{-factor}}$ and $\boldsymbol{\epsilon}_{N \times w}^{3\text{-factor}}$ are the Gaussian error terms. In this case, the maximum likelihood estimates of $\mathbf{\Sigma}^{2\text{-factor}}$ and $\mathbf{\Sigma}^{3\text{-factor}}$ are identical to those obtained from the least squares estimation, so we have

$$\begin{aligned}
\mathbf{\Sigma}^{2\text{-factor}} &= \left(\mathbf{U}^{2\text{-factor}'} \mathbf{U}^{2\text{-factor}} \right)^{-1} \left(\mathbf{U}^{2\text{-factor}'} \mathbf{V} \right) \\
&= \begin{pmatrix} N & \sum_{i=1}^N (x_i - \bar{x}) \\ \sum_{i=1}^N (x_i - \bar{x}) & \sum_{i=1}^N (x_i - \bar{x})^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^N v_{i,t_1} & \cdots & \sum_{i=1}^N v_{i,t_w} \\ \sum_{i=1}^N v_{i,t_1} (x_i - \bar{x}) & \cdots & \sum_{i=1}^N v_{i,t_w} (x_i - \bar{x}) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{N} & 0 \\ 0 & \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N v_{i,t_1} & \cdots & \sum_{i=1}^N v_{i,t_w} \\ \sum_{i=1}^N v_{i,t_1} (x_i - \bar{x}) & \cdots & \sum_{i=1}^N v_{i,t_w} (x_i - \bar{x}) \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sum_{i=1}^N v_{i,t_1}}{N} & \cdots & \frac{\sum_{i=1}^N v_{i,t_w}}{N} \\ \frac{\sum_{i=1}^N v_{i,t_1} (x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} & \cdots & \frac{\sum_{i=1}^N v_{i,t_w} (x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \end{pmatrix},
\end{aligned}$$

because $\sum_{i=1}^N (x_i - \bar{x}) = 0$. Similarly, we obtain

$$\begin{aligned}
\mathbf{\Sigma}^{\text{3-factor}} &= \left(\mathbf{U}^{\text{3-factor}'} \mathbf{U}^{\text{3-factor}} \right)^{-1} \left(\mathbf{U}^{\text{3-factor}'} \mathbf{V} \right) \\
&= \begin{pmatrix} N & \sum_{i=1}^N (x_i - \bar{x}) & \sum_{i=1}^N ((x_i - \bar{x})^2 - s^2) \\ \sum_{i=1}^N (x_i - \bar{x}) & \sum_{i=1}^N (x_i - \bar{x})^2 & \sum_{i=1}^N ((x_i - \bar{x})^3 - s^2(x_i - \bar{x})) \\ \sum_{i=1}^N ((x_i - \bar{x})^2 - s^2) & \sum_{i=1}^N ((x_i - \bar{x})^3 - s^2(x_i - \bar{x})) & \sum_{i=1}^N ((x_i - \bar{x})^2 - s^2)^2 \end{pmatrix}^{-1} \\
&\quad \begin{pmatrix} \sum_{i=1}^N v_{i,t_1} & \cdots & \sum_{i=1}^N v_{i,t_w} \\ \sum_{i=1}^N v_{i,t_1} (x_i - \bar{x}) & \cdots & \sum_{i=1}^N v_{i,t_w} (x_i - \bar{x}) \\ \sum_{i=1}^N v_{i,t_1} ((x_i - \bar{x})^2 - s^2) & \cdots & \sum_{i=1}^N v_{i,t_w} ((x_i - \bar{x})^2 - s^2) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{N} & 0 & 0 \\ 0 & \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2} & 0 \\ 0 & 0 & \frac{1}{\sum_{i=1}^N ((x_i - \bar{x})^2 - s^2)^2} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N v_{i,t_1} & \cdots & \sum_{i=1}^N v_{i,t_w} \\ \sum_{i=1}^N v_{i,t_1} (x_i - \bar{x}) & \cdots & \sum_{i=1}^N v_{i,t_w} (x_i - \bar{x}) \\ \sum_{i=1}^N v_{i,t_1} ((x_i - \bar{x})^2 - s^2) & \cdots & \sum_{i=1}^N v_{i,t_w} ((x_i - \bar{x})^2 - s^2) \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sum_{i=1}^N v_{i,t_1}}{N} & \cdots & \frac{\sum_{i=1}^N v_{i,t_w}}{N} \\ \frac{\sum_{i=1}^N v_{i,t_1} (x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} & \cdots & \frac{\sum_{i=1}^N v_{i,t_w} (x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \\ \frac{\sum_{i=1}^N v_{i,t_1} ((x_i - \bar{x})^2 - s^2)}{\sum_{i=1}^N ((x_i - \bar{x})^2 - s^2)^2} & \cdots & \frac{\sum_{i=1}^N v_{i,t_w} ((x_i - \bar{x})^2 - s^2)}{\sum_{i=1}^N ((x_i - \bar{x})^2 - s^2)^2} \end{pmatrix},
\end{aligned}$$

because $\sum_{i=1}^N ((x_i - \bar{x})^2 - s^2) = 0$ and

$$\begin{aligned}
\sum_{i=1}^N ((x_i - \bar{x})^3 - s^2(x_i - \bar{x})) &= \sum_{i=1}^N (x_i - \bar{x})^3 - s^2 \sum_{i=1}^N (x_i - \bar{x}) \\
&= \sum_{i=1}^N (x_i - \bar{x})^3 \\
&= 0.
\end{aligned}$$

Note that the last equality is due to the strictly consecutive integer values of x_i .

In more detail, this implies that the value of $\bar{x}$ equals to the median value of the integers x_i 's, that is, $\bar{x} = x_1 + \frac{N-1}{2}$, such that the values of $(x_i - \bar{x})$'s are

$$\begin{aligned}
&\{(x_1 - \bar{x}), (x_1 + 1 - \bar{x}), \dots, -1, 0, 1, \dots, (x_1 + N - 2 - \bar{x}), (x_1 + N - 1 - \bar{x})\} \quad (\text{for odd } N), \\
&\{(x_1 - \bar{x}), (x_1 + 1 - \bar{x}), \dots, -0.5, 0.5, \dots, (x_1 + N - 2 - \bar{x}), (x_1 + N - 1 - \bar{x})\} \quad (\text{for even } N),
\end{aligned}$$

and hence the summation of any odd functions of these values is zero.

References

- Bell, W. (1997). Comparing and assessing time series methods for forecasting fertility and mortality rates. *Journal of Official Statistics*, **13**, 279-303.
- Bennett, N. and Horiuchi, S. (1984). Mortality estimation from registered deaths in less developed countries. *Demography*, **21**, 217-233.
- Blake, D. and Burrows, W. (2001). Survivor bonds: Helping to hedge mortality risk. *Journal of Risk and Insurance*, **68**, 339-348.
- Blake, D., Cairns, A., Coughlan, G., Dowd, K. and MacMinn, R. (2013). The new life market. *Journal of Risk and Insurance*, **80**, 501-557.
- Blake, D., MacMinn, R. Li, J.S.-H. and Hardy, M.R. (2014). Longevity risk and capital markets: The 2012-13 update. *North American Actuarial Journal*, **18**, 1-13.
- Bolancé, C., Guillén, M., and Pinquet, J. (2003). Time-varying credibility for frequency risk models: Estimation and tests for autoregressive specification on the random effect. *Insurance: Mathematics and Economics*, **33**, 273-282.
- Booth, H., Maindonald, J. and Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, **56**, 325-336.
- Borgan, Ø., Hoem, J.M. and Norberg, R. (1981). A nonasymptotic criterion for the evaluation of automobile bonus systems. *Scandinavian Actuarial Journal*, **3**, 165-178.

- Boucher, J.-P. and Denuit, M. (2006). Fixed versus random effects in Poisson regression models for claim counts: Case study in motor insurance. *ASTIN Bulletin*, **36**, 285-301.
- Boucher, J.-P., Denuit, M. and Guillén, M. (2007). Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, **11**, 110-131.
- Boucher, J.-P. and Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin*, **44**, 587-612.
- Box, G.E.P., and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd edition. San Francisco: Holden-Day Press.
- 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.
- Brouhns, N., Denuit, M. and Keilegom, I.V. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, **3**, 212-224.
- Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin*, **4**, 199-207.
- Bühlmann, H. and Gisler, A. (2005). *A Course in Credibility Theory and Its Applications*. Berlin: Springer-Verlag.
- Cairns, A.J.G. (2011). Modelling and management of longevity risk: Approximations to survival functions and dynamic hedging. *Insurance: Mathematics and Economics*, **49**, 438-453.
- Cairns, A.J.G., Blake, D. and Dowd, K. (2006). A two-factor model for stochastic mortality with parametric uncertainty: Theory and calibration. *Journal of Risk and Insurance*, **73**, 687-718.

- 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.
- 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.
- Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D. and Khalaf-Allah, M. (2011a). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, **41**, 29-55.
- Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D. and Khalaf-Allah, M. (2011b). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, **48**, 355-367.
- Centeno, M.L. and Andrade e Silva, J.M. (2002). Optimal bonus scales under path-dependent bonus rules. *Scandinavian Actuarial Journal*, **2**, 129-136.
- Chan, W.S., Li, J.S.-H. and Li, J. (2014). The CBD mortality indexes: Modeling and applications. *North American Actuarial Journal*, **18**, 38-58.
- Cheung, S.L.K., Robine, J.-M., Tu, E.J. and Caselli, G. (2005). Three dimensions of the survival curve: horizontalization, verticalization, and longevity extension. *Demography*, **42**, 243-258.
- Coene, G. and Doray, L.G. (1996). A financially balanced bonus-malus system. *ASTIN Bulletin*, **26**, 107-116.
- Coughlan, G.D. (2009). Longevity risk transfer: Indices and capital market solutions. In Barrieu, P.M. and Albertini, L. (eds.) *The Handbook of Insurance Linked Securities*, 261-282. London: Wiley.
- Coughlan, G., Blake, D., MacMinn, R. and Cairns, A.J.G. (2013). Longevity Risk and Hedging Solutions. In Dionne, G. (ed.) *Handbook of Insurance*, 2nd edition, 997-1036. New York: Springer.

- Coughlan, G.D., Epstein, D., Ong, A., Sinha, A., Hevia-Portocarrero, J., Gingrich, E., Khalaf-Allah, M. and Joseph, P. (2007). *LifeMetrics: A toolkit for measuring and managing longevity and mortality risks*. London: J.P. Morgan.
- Coughlan, G.D., Khalaf-Allah, M., Ye, Y., Kumar, S., Cairns, A.J.G., Blake, D. and Dowd, K. (2011). Longevity hedging 101: A framework for longevity basis risk analysis and hedge effectiveness. *North American Actuarial Journal*, **15**, 114-146.
- Currie, I.D. (2006). Smoothing and forecasting mortality rates with P-splines. Paper presented at the Institute of Actuaries, June 2006. Available at: <http://www.macs.hw.ac.uk/~iain/research/talks/Mortality.pdf>.
- Currie, I. D., Durban, M. and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279-298.
- De Jong, P. and Heller, G.Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press.
- De Jong, P. and Marshall, C. (2007). Mortality projection based on the Wang transform. *ASTIN Bulletin*, **37**, 149-161.
- Denuit, M. and Dhaene, J. (2001). Bonus-malus scales using exponential loss functions. *German Actuarial Bulletin*, **25**, 13-27.
- Denuit, M., Devolder, P. and Goderniaux, A.-C. (2007). Securitization of longevity risk: Pricing survivor bonds with Wang transform in the Lee-Carter framework. *Journal of Risk and Insurance*, **74**, 87-113.
- Denuit, M., Maréchal, X., Pitrebois, S. and Walhin, J.F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Chichester: John Wiley & Sons.
- Dionne, G. and Vanasse, C. (1989). A generalization of actuarial automobile insurance rating models: the Negative Binomial distribution with a regression component. *ASTIN Bulletin*, **19**, 199-212.

- Dowd, K., Blake, D., Cairns, A.J.G. and Dawson, P. (2006). Survivor swaps. *Journal of Risk and Insurance*, **73**, 1-17.
- Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D. and Khalaf-Allah, M. (2010a). Backtesting stochastic mortality models: An ex-post evaluation of multi-period-ahead density forecasts. *North American Actuarial Journal*, **14**, 281-298.
- Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D. and Khalaf-Allah, M. (2010b). Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics*, **47**, 255-265.
- Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D. and Khalaf-Allah, M. (2011). A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, **15**, 334-356.
- Eakin, T. and Witten, M. (1995). How square is the survival curve of a given species? *Experimental Gerontology*, **30**, 33-64.
- Frangos, N. and Vrontos, S. (2001). Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *ASTIN Bulletin*, **31**, 1-22.
- Fries, J.F. (1980). Aging, natural death, and the compression of morbidity. *New England Journal of Medicine*, **303**, 130-135.
- Gilde, V. and Sundt, B. (1989). On bonus systems with credibility scales. *Scandinavian Actuarial Journal*, **1**, 13-22.
- Goldstein, J.R. (2006). Found in translation? A cohort perspective on tempo-adjusted life expectancy. *Demographic Research*, **14**, 71-84.
- Goldstein, J. R. and Wachter, K.W. (2005). Relationships between period and cohort life expectancy: Gaps and lags. *Population Studies*, **60**, 257-269.
- Guillot, M. (2003). The cross-sectional average length of life (CAL): A cross-

- sectional mortality measure that reflects the experience of cohorts. *Population Studies*, **57**, 41-54.
- Haberman, S. and Renshaw, A.E. (2012). Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and Economics*, **50**, 309-333.
- Human Mortality Database. University of California, Berkeley (USA), and Max-Planck Institute of Demographic Research (Germany). Available at: <http://www.mortality.org> or <http://www.humanmortality.de>. (data downloaded on 24 February 2013).
- Hunt, A. and Blake, D. (2014). A general procedure for constructing mortality models. *North American Actuarial Journal*, **18**, 116-138.
- Jarner, S.F. and Kryger, E.M. (2011). Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin*, **41**, 377-418.
- Kannisto, V. (2000). Measuring the compression of mortality. *Demographic Research*, **3**. Available at: <http://www.demographic-research.org/Volumes/Vol13/6>.
- Kannisto, V. (2001). Mode and dispersion of the length of life. *Population: An English Selection*, **13**, 159-171.
- Lee, R.D. and Carter, L. (1992). Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association*, **87**, 659-671.
- Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*. Boston: Kluwer Academic Publishers.
- Li, J. (2013). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population Studies*, **67**, 111-126.
- Li, J.S.-H. and Hardy, M.R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, **15**, 177-200.

- Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of population: An extension of the Lee-Carter method. *Demography*, **42**, 575-594.
- Li, J.S.-H. and Luo, A. (2012). Key q-duration: A framework for hedging longevity risk. *ASTIN Bulletin*, **42**, 413-452.
- Li, J.S.-H., Ng, A.C.Y. and Chan, W.S. (2011). On the calibration of mortality forward curves. *Journal of Futures Markets*, **31**, 941-970.
- Lin, T. and Tsai, C.C.-L. (2013). On the mortality/longevity risk hedging with mortality immunization. *Insurance: Mathematics and Economics*, **53**, 580-596.
- Mahmoudvand, R. and Hassani, H. (2009). Generalized bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *ASTIN Bulletin*, **39**, 307-315.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.
- Norberg, R. (1976). A credibility theory for automobile bonus system. *Scandinavian Actuarial Journal*, **2**, 92-107.
- Oeppen, J. and Vaupel, J.W. (2002). Broken limits to life expectancy. *Science*, **296**, 1029-1031.
- Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology*, **14**, 124-129.
- Pinquet, J. (1997). Allowance for costs of claims in bonus-malus systems. *ASTIN Bulletin*, **27**, 33-57.
- Pinquet, J., Guillén, M. and Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin*, **31**, 337-348.
- Pitrebois, S., Denuit, M. and Walhin, J.F. (2003a). Fitting the Belgian bonus-malus system. *Belgian Actuarial Bulletin*, **3**, 58-62.
- Pitrebois, S., Denuit, M. and Walhin, J.F. (2003b). Setting a bonus-malus scale

- in the presence of other rating factors: Taylor's work revisited. *ASTIN Bulletin*, **33**, 419-436.
- Pitrebois, S., Denuit, M. and Walhin, J.F. (2004). Bonus-malus scales in segmented tariffs: Gilde & Sundt's work revisited. *Australian Actuarial Journal*, **10**, 107-125.
- Plat, R. (2009). Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, **45**, 123-132.
- Plat, R. (2010). One-year Value-at-Risk for longevity and mortality. *Insurance: Mathematics and Economics*, **49**, 462-470.
- Renshaw, A.E. and Haberman, S. (2003). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, **33**, 255-272.
- Renshaw, A.E. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**, 556-570.
- Taylor, G. (1997). Setting a bonus-malus scale in the presence of other rating factors. *ASTIN Bulletin*, **27**, 319-327.
- Thatcher, A.R. (1999). The long-term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society (A)*, **162**, 5-43.
- Thatcher, A.R., Cheung, S.L.K., Robine, J.-M. and Horiuchi, S. (2010). The compression of deaths above the mode. *Demographic Research*, **22**, 502-538.
- Tiao, G.C. and Box, G.E.P. (1981). Modelling multiple time series with applications. *Journal of the American Statistical Association*, **76**, 802-816.
- Tremblay, L. (1992). Using the Poisson Inverse Gaussian in bonus-malus systems. *ASTIN Bulletin*, **22**, 97-106.
- Tsai, J.T., Wang, J.L. and Tzeng, L.Y. (2010). On the optimal product mix in

- life insurance companies using conditional value at risk. *Insurance: Mathematics and Economics*, **46**, 235-241.
- Tzougas, G., Vrontos, S. and Frangos, N. (2014). Optimal bonus-malus systems using finite mixture models. *ASTIN Bulletin*, **44**, 417-444.
- United Nations, Department of Economic and Social Affairs, Population Division. (2011). *World Population Prospects: The 2010 Revision, Volume II: Demographic Profiles*. ST/ESA/SER.A/317.
- Vincent, P. (1951). La Mortalité des Vieillards. *Population*, **6**, 181-204.
- Walhin, J.F. and Paris, J. (1999). Using mixed Poisson distributions in connection with bonus-malus systems. *ASTIN Bulletin*, **29**, 81-99.
- Wachter, K.W. (2005). Tempo and its tribulations. *Demographic Research*, **13**, 201-222.
- Wang, S.S. (2000). A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, **67**, 15-36.
- Wang, J.L., Huang, H.C., Yang, S.S. and Tsai, J.T. (2010). An optimal product mix for hedging longevity risk in life insurance companies: the immunization theory approach. *Journal of Risk and Insurance*, **77**, 473-497.
- Wei, W.W.S. (2006). *Time Series Analysis: Univariate and Multivariate Method*, 2nd edition. New York: Addison Wesley/Pearson.
- Wilmoth, J.R. (2005). On the relationship between period and cohort mortality. *Demographic Research*, **13**, 231-280.
- Wilmoth, J.R. and Horiuchi, S. (1999). Rectangularization revisited: Variability of age at death within human populations. *Demography*, **36**, 475-495.
- World Bank. (2014). *World Development Indicators 2014*. Retrieved from: <http://data.worldbank.org/indicator/SP.DYN.LE00.IN>.
- Yang, S.S. and Wang, C.W. (2013). Pricing and securitization of multi-country

- longevity risk with mortality dependence. *Insurance: Mathematics and Economics*, **52**, 157-169.
- Yip, K.C.H. and Yau, K.K.W. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, **36**, 153-163.
- Yue, C.J. (2012). Mortality compression and longevity risk. *North American Actuarial Journal*, **16**, 434-448.
- Zhou, R., Li, J.S.-H. and Tan, K.S., (2011). Economic pricing of mortality-linked securities in the presence of population basis risk. *Geneva Paper of Risk and Insurance: Issues and Practice*, **36**, 544-566.
- Zhou, R., Li, J.S.-H. and Tan, K.S. (2013a). Economic pricing of mortality-linked securities: A Tâtonnement approach. *Journal of Risk and Insurance*, doi: 10.1111/j.1539-6975.2013.12008.x.
- Zhou, R., Li, J.S.-H. and Tan, K.S. (2013b). Pricing standardized mortality securitizations: a two-population model with transitory jump effects. *Journal of Risk and Insurance*, **80**, 733-774.
- Zhou, R., Wang, Y., Kaufhold, K., Li, J.S.-H. and Tan, K.S. (2014). Modeling period effects in multi-population mortality models: Applications to Solvency II. *North American Actuarial Journal*, **18**, 150-167.