Five essays in Life Insurance Mathematics

Esben Masotti Kryger

PhD Thesis

Supervisors: Mogens Steffensen, University of Copenhagen
Søren Fiig Jarner, ATP
Michael Preisel, ATP

Submitted: 12 May 2010.

Department of Mathematical Sciences
Faculty of Science
University of Copenhagen

Preface

This thesis has been prepared in partial fulfillment of the requirements for the PhD degree at the Department of Mathematical Sciences under the Faculty of Science at the University of Copenhagen.

The work has been carried out under the supervision of Søren Fiig Jarner, ATP, and Professor Mogens Steffensen, University of Copenhagen, in the period 1 April 2007 to 12 May 2010 at ATP and the Department of Mathematical Sciences at the University of Copenhagen. The project was funded by ATP and the Danish Agency for Science, Technology and Innovation under the Industrial PhD Programme.

The five last chapters in this thesis are written as individual academic papers, and are thus self-contained, and can be read independently. There are minor overlaps in the contents of the papers in Chapters 2 and 3. Likewise there are minor notational overlaps between the different chapters, but it is unlikely to cause confusion.

Acknowledgments

I would like to thank my supervisors and coauthors Søren Fiig Jarner and Mogens Steffensen for their support and advice during the PhD programme. My gratitude extends to my colleagues at ATP and at the Department of Mathematical Sciences, who have shown great interest in my work. I appreciate the discussions with Professor Ragnar Norberg, which initiated fruitful thoughts during my stay at the London School of Economics in the autumn of 2008. Finally, I wish to thank my wife Fulvia and our daughter, Laura, for their ceaseless love, support, and understanding.

Esben Masotti Kryger
Copenhagen, May 2010

Summary

This thesis consists of five papers within the broad area of Life insurance mathematics. There is no unifying topic, but the papers can be divided into three categories as described below.

The first two papers deal with pension scheme design, and are included as Chapters 2 and 3 respectively. Both take the view of an optimiser, who takes intergenerational issues into account in the sense that she is concerned with the benefits of more than one generation in a mutually owned with–profits pension scheme. To this end she picks strategies for investment and bonus allotment. One paper discusses optimal design in the long term, which means that all generations sample the same stationary benefit distribution, in turn implying that there is no issue of systematic intergenerational subsidisation. The other paper meanwhile considers the problem on a shorter horizon. As generations enter at different points in time they do not receive the same benefits. Hence, there is some degree of intergenerational redistribution, which should be taken into account in the design. The distinction between the short– and long–term views gives rise to problems, which are conceptually somewhat different from each other (and have correspondingly different solutions). The two papers are summarised in Section 1.1.

Chapters 4 and 5 contain two papers on the broad topic of mortality. The first of those papers begins by proposing a deterministic model for adult mortality based on frailty theory. Next, the paper suggests a general method for forecasting the (distribution of) mortality of a small population by linking it to a larger, but similar population, for which mortality can be projected robustly – in turn ensuring coherent forecasts. As an illustration the frailty model is estimated for a basket of 19 Western countries, while the coherence method is exemplified by linking Danish mortality development to the aforementioned prognosis for the larger population. The Danish forecast, however, could be based on any mortality projection for the larger population. That is, the two models in the paper are not connected per se. The other paper summarises the historical development of Danish

mortality, and suggests a method to disaggregate the vast improvements in life expectancy that have been observed since data collection started in 1835 into age–specific contributions. This decomposition reveals that increases in life expectancy have historically been carried by improvements among infants and children; gradually moving to the situation today, where possible life expectancy improvements over the next decades can only be brought about by decreasing mortality among ages 60–80. The two papers are summarised in Section 1.2

Finally, Chapter 6 presents a paper, which shows how to extend the Bellmann equation of stochastic control to a new set of problems concerned with choosing the optimal investment strategy in order to maximise some function of terminal wealth. The paper also presents three applications, which are new, and their solutions: 1) A group utility problem for exponential or power utility, where terminal wealth is shared proportionally among group members. This is of central importance to pension scheme designers. 2) Dynamic mean–standard deviation optimisation for a single agent. 3) Endogenous habit formation with quadratic utility for a single agent. The paper is summarised in Section 1.3.

Sammenfatning

Denne afhandling består af fem artikler indenfor det brede felt livsforsikringsmatematik. Afhandlingen har intet forenende tema, men de fem artikler kan inddeles i tre kategorier som beskrevet nedenfor.

De første to artikler omhandler design af pensionssystemer og er inkluderet som afhandlingens Kapitel 2 og Kapitel 3. Begge artikler anlægger synsvinklen af en beslutningstager, der tager intergenerationelle betragtninger med i sine overvejelser – i den forstand at han er interesseret i mere end blot een generations pensionsudbetalinger i et kundeejet pensionsselskab, hvis produkt giver ret til bonus. Med henblik på design af systemet kan han fastlægge investeringsstratgien og reglerne for, hvornår der udloddes bonus. Den første artikel diskuterer optimalt design på langt sigt, hvilket indebærer, at alle generationer sampler den samme, stationære, pensionsudbetalingsfordeling, hvorved systematisk omfordeling mellem generationer ikke bliver et tema. Den anden artikel beskæftiger sig imidlertid med problemet på en kortere horisont. Idet de forskellige generationer kommer ind i systemet på forskellige tidspunkter kan de ikke opnå den samme pensionsudbetalingsfordeling. Der er således et element af systematisk omfordeling mellem generationerne, som designet bør tage højde for. Sondringen mellem det lange og det korte sigt giver anledning til problemer, som konceptuelt er ganske forskellige (og som følgeligt har forskellige løsninger). De to artikler resumeres i Afsnit 1.1.

Kapitlerne 4 og 5 indeholder to artikler om dødelighed. Den første af disse artikler begynder med at foreslå en deterministisk model for voksnes dødelighed baseret på *frailty* teori. Dernæst skitserer artiklen en generel metode til at fremskrive (fordelingen af) dødelighed i en lille befolkning. Dette gøres ved at hægte den lille befolkning til en større, men lignende befolkning, hvis udvikling kan fremskrives robust. Derved opnås koherente fremskrivninger. Som en illustration af modellerne estimeres frailty-modellen for en kurv bestående af 19 vestlige lande, mens metoden til koherente fremskrivninger eksemplificeres ved at betragte Danmark som den lille befolkning og fremskrivningen for de 19 lande som referencepunk-

tet. Fremskrivningen for Danmark kunne dog være baseret på en vilkårlig prognose for udviklingen i den større befolkning. Altså: De to modeller i artiklen er ikke partout forbundne med hinanden. Den anden artikel opsummerer den historiske udvikling i dødeligheden i Danmark og foreslår en metode til at disaggregere de enorme forbedringer i levetiden, der er set siden dataindsamlingen påbegyndtes i 1835, på aldersafhængige bidrag dertil. Denne dekomposition afslører, at øgede levetider historisk set blev frembragt af store fald i babyers og børns dødelighed. Gradvist er udviklingen så gået imod situationen i dag, hvor eventuelle levetidsforbedringer over de næstkommende årtier kun kan frembringes af fald i dødeligheden i aldergrupperne omkring 60–80 år. Disse to artikler resumeres i Afsnit 1.2.

Endelig præsenterer Kapitel 6 en artikel, som demonstrerer, hvordan den klassiske Bellmann–ligning fra stokastisk kontrol kan udvides til en ny familie af problemer, hvor målet er at maksimere en funktion af slutformue ved at kontrollere investeringsstrategien. Artiklen præsenterer også tre nye anvendelser og de tilhørende løsninger: 1) Et gruppenytte–problem for eksponentiel– eller potensnytte, hvori formuen deles proportionalt mellem gruppens medlemmer. Dette er af central betydning for pensionskasser. 2) Dynamisk middelværdi-standardafvigelses optimering for en enkelt investor. 3) Endogen *habit formation* med kvadratisk nytte for en enkelt investor. Artiklen resumeres i Afsnit 1.3.

Contents

1. Introduction

This chapter gives an overview of the contributions of the thesis as well as the perspectives for future research. To this end the division into three separate topics introduced above is maintained. That is, Section 1.1 summarises two papers about pension scheme design, Section 1.2 gives the main points of two mortality–related papers, while Section 1.3 sums up the contents of a paper concerned with optimal portfolio selection.

1.1 Pension scheme design

The paper "Pension fund design under long–term fairness constraints" (Chapter 2) seeks to answer the question "What is the optimal long–term investment strategy for a mutually owned with–profits pension scheme?" The view taken is that of an altruistic board, or a similar authority, wishing to keep the fund solvent in order to ensure future generations' access to the scheme – although this may not be the most desirable outcome for any subset of members, e.g. the present ones may want to dissolve the fund altogether. The solvency condition is equivalent to the existence of a stationary distribution for the funding ratio of the scheme. We note that this notion of fairness is very different from the one typically applied in the literature, which is concerned with arbitrage–free pricing of contingent claims.

The literature on design of fair life insurance products was initiated by a seminal paper by Briys and de Varenne (1994), which has been succeeded by a vast number of extensions, e.g. Grosen and Jørgensen (2002); Ballotta (2005); Bernard *et al.* (2005); Chen and Suchanecki (2007). Their common approach is to price a savings contract consisting of a guaranteed return and an option on terminal non–discretionary bonus. The bonus is paid out if the return on a fixed portfolio is sufficiently high. On the other hand the issuing company may go bankrupt under way (if monitored) or upon expiry. In this literature the purpose is to find the set of fair contracts, which is defined as those with an arbitrage–free net value of zero, by adjusting

the different elements of the contract. Essentially these models consider risk–sharing between two parties, i.e. there are no intergenerational elements. Our paper is inspired by Preisel *et al.* (2010), who modified the aforementioned papers in several crucial ways. The most important manners in which they differ from the mainstream is through the investment strategy, which seeks to protect the solvency (rather than keeping a fixed allocation), via the presence of multiple bonus allotments, and through a study of the stationary properties of the company. Preisel *et al.* (2010) lays out the framework for our analysis, but their model is extended by introducing a members' optimisation criterion, and by discussing how to choose the optimal bonus policy.

In order to comply with the solvency requirement the set of investment strategies is restricted to those in the CPPI–class, that is

$$\pi_t = \alpha \frac{A_t - L_t}{A_t},$$

where π denotes the proportion of total assets, A, allocated to the risky asset in a Black–Scholes market, and L denotes the market value of the scheme's liabilities. The multiplier $\alpha > 0$ then determines the investment strategy, and is set by the board. Bonus is allotted by periodically increasing the guaranteed benefits by a fixed rate for all members, if the funding ratio, $F = A/L$ is above a pre–defined barrier, which is also set by the board, at the end of the period. In that case the bonus given is the quantity that takes the funding ratio back down to the barrier. This rule is inspired by the results in Steffensen (2004). Such a bonus strategy turns out to imply that in the long term it is superior to have chosen a higher barrier, and hence this variable must be set exogenously.

By requiring a certain demography we can derive the stationary distribution of the bonus allotted. The goal is then to maximise the stationary expected power utility of the terminal wealth of a member, who participates in the scheme for several periods. Although the setting is a complete market with deterministic liabilities (between bonus allotments) this completeness does not apply to members' pension savings, which therefore cannot be priced with the usual replicating arguments.

We consider two types of contracts. First, simple ones consisting of only a single contribution, which is converted into a guaranteed benefit

and a compound bonus option, i.e. a payoff that is proportional to

$$\exp\left(\sum_{k=1}^{n} b_k\right),$$

with b_k being the bonus allotments in stationarity. In this case approximate analytical results are derived, and are shown to be accurate. Next we consider a contract to which members contribute in several periods, so that the payoff is proportional to

$$\sum_{j=0}^{n} \exp\left(\sum_{k=j+1}^{n} b_k\right).$$

In this case later bonus is more important because it acts on a larger guaranteed benefit. Then analytical results are out of scope, but simulations indicate that the insights from the approximate results carry over to this case as well – in most cases. Quantitatively, the optimal strategy seems to consist of investing an amount corresponding to the liabilities in a safe asset – and follow *roughly* the same policy as a logarithmic investor in the classic Merton (1969)–world for the remaining assets, the so–called bonus reserve. As in his case, our optimal investment strategy also depends on the coefficient of relative risk aversion, but our dependence is far less pronounced. Therefore it is not nearly as costly to form an investment collective as in Merton (1969). However, our case is very different from his: he values the total return on an initial investment, while we value the payoff from a series of compound options. Consequently, there are other discrepancies: as opposed to the classical case, horizon matters, though only to a mild extent. The reason for this phenomenon is the serial dependence between bonuses, which in turn means that aggressive investment strategies are quite unattractive.

The paper "Fairness vs. efficiency of pension schemes" (Chapter 3) seeks to remedy some of the shortcomings of the previous paper: namely the lack of short–term considerations following from investigating only the stationary properties of the system. In particular, the absence of a method for deriving an optimal bonus barrier. Also, it is possible to analyse systems

that are not as demographically rigid as was required above. In order to work with these extensions a finite horizon is considered.

In our model the benefits received by a generation depend on the initial funding, which is random, and on the investment strategy and bonus barrier, both of which are controlled. Consequently, each generation has its own opinion on optimal design – even if their preferences are equal. We require that the collective, i.e. the generations, must agree on a design there is some loss of efficiency associated with each choice. At the same time some designs will induce less differences between the benefits that different generations receive, than will others. That is, more fairness. In order to analyse the trade–off two independent, hypothetical generations with initial funding ratios κ and $f \in (1, \kappa)$ are considered, with the former being the bonus barrier, and the latter a relatively low funding. We think of the design decision as being taken via a bargaining between the two, not knowing their own identities – in the spirit of Rawls (1971). The fairness measure is defined as the probability that the benefits differ sufficiently little:

$$\mathbb{P}\left(\frac{X\left(s,\kappa,f\right)}{X\left(s,\kappa,\kappa\right)} > 1-\delta\right),$$

with $X(s,\kappa,\cdot)$ denoting the terminal benefits as functions of the investment strategy ($s = \alpha\sigma$), the bonus barrier, and the initial funding level – and with δ being some maximum acceptable redistribution level (as seen from the board's point of view). Each generation has its own efficiency measure defined by the probability of obtaining a certain benchmark, a fraction of its certainty equivalent:

$$\mathbb{P}\left(X(s,\kappa,\cdot) > (1-\beta)\max_{\bar{s},\bar{\kappa}}\ \mathbb{E}\left\{X(\bar{s},\bar{\kappa},\cdot)^{1-\gamma}\right\}^{\frac{1}{1-\gamma}}\right),$$

with γ being a coefficient of relative risk aversion, which turns out to be of minor importance, and with β interpretable as a maximum permitted relative cost of achieving some fairness criterion. The focus on tail probabilities stems from the fact that participation in the system under consideration is assumed compulsory, so that there is no option to walk away from an unattractive scheme.

With those definitions we can maximise fairness subject to an efficiency criterion, or reversely. In the paper the measures are merely plotted for

fixed (β, δ, γ), and it is open which approach is better in which situation. There is, however, another twist to the problem: To add a further trade–off we consider the situation where the lower funding ratio, f, depends on the design, for it is less likely to encounter a future low funding if investments are cautious or if the barrier is high. In this latter setting, however, higher barriers always become desirable, so there is only one design parameter. Conversely, in the case of a fixed f a trade–off between longer period with inequality on one hand, and higher long–term collective benefits on the other hand (both arising from a high barrier) is at the heart of the problem. Unsurprisingly, the results differ substantially – qualitatively and quantitatively – depending on which approach is taken.

As a consequence of the unsatisfying properties of the standard system in which the random funding ratio at entry is crucial two other systems are suggested, both of which turn out to give rise to systems which cost freely provide less intergenerational subsidisation.

A recent paper that is related to ours is Døskeland and Nordahl (2008a). They evaluate the life cycle of a pension scheme and show that early generations subsidise later ones as a consequence of there being no initial bonus reserve, and no third party to sponsor the scheme at start–up, nor to inherit the ultimate bonus reserve. The topic of intergenerational redistribution and risk sharing is more mature in macroeconomics and welfare economics, e.g. Ball and Mankiw (2007); Gollier (2008).

Both papers are implemented with an assumption of annual bonus allotments, but it is straightforward to let the time between possible bonus allotments shrink and analyse the consequences thereof. The latter paper is also suitable to a number of interesting extensions. One obvious idea is the introduction of undiversifiable insurance and market risk. Another possible extension is an analysis of the widespread practice of smoothing bonuses, which would likely result in less redistribution. Finally, other fairness and efficiency measures may be introduced.

1.2 Mortality

The paper "Modelling adult mortality in small populations: The SAINT model" (Chapter 4, written jointly with Søren Fiig Jarner) proposes a stochastic model for adult mortality in small populations. The paper presents two models, which are used interdependently in the paper, but which are not intrinsically connected. The first model is a theory for adult mortality based on frailty theory. More specifically, the mortality intensity at time t and age x of an individual with (unobserved) frailty $z > 0$ is modelled by the (almost) multiplicative structure

$$\mu(t, x; z) = z\mu_s^I(t, x) + \gamma(t),$$

where z comes from a Γ–distribution (although one can readily relax to a more general distributional assumption, cf. Hougaard (1986)). The quantity $\mu_s^I(t, x) > 0$ denotes the senescent mortality of an individual with unit frailty, while γ represents frailty– and age–independent background mortality. Vaupel *et al.* (1979) analysed the connection between population mortality and individual mortality in the same framework, albeit only for a single cohort. Even without specifying further one can derive some nice features about the mortality of a population consisting of individuals with i.i.d. frailties. (Propositions 4.2–4.4). In order to estimate the model, however, further assumptions are required. First, we assume that individual mortality is affected by immediate as well as accumulating factors by specifying

$$\mu_s^I(t, x) = \kappa(t) \exp\left(\int_0^x g(u + t - x, u)du\right).$$

The term κ^{-1} can be thought of as the level of treatment and environmental factors etc. at time t for persons of age x, while g represents a time– and age–dependent force of aging. The motivation for considering this frailty model is two–fold. Firstly, structural models produce credible forecasts of the mortality surface (e.g. monotonous with respect to time and age), and secondly, we allow for improvements in old–age mortality, where none have hitherto been observed – as opposed to purely extrapolative models. The

specification of the frailty model is finalised by putting

$$\begin{aligned} g(t,x) &= g_1 + g_2(t-t_0) + g_3(x-x_0), \\ \kappa(t) &= \exp(\kappa_1 + \kappa_2(t-t_0)), \\ \gamma(t) &= \exp(\gamma_1 + \gamma_2(t-t_0)), \end{aligned}$$

for some x_0, t_0. We can then derive asymptotic properties of annual improvements in mortality as well as age–dependent mortality (Propositions 4.5 and 4.6). As an illustration the model is estimated for a pool consisting of 19 Western countries over 1933–2005 for ages 20–100, and both genders separately. To this end a standard Poisson likelihood is applied, and a discretisation approximation is imposed, since data consists of annual observations, while the model is phrased in continuous time. One of the main findings from this estimation is that women are more heterogeneous than are men, which in turn implies that the difference between the genders' life expectancies is forecasted to widen. Barbi (2003) found the same phenomenon. Another characteristic of the forecast is that the mortality of the oldest old will start to decrease in the first half of this century, although hardly any improvements have been observed historically. We note that the projection is essentially deterministic because all uncertainty at the population level stems from the model parameters, which are extremely precisely determined because of the parsimonious specification.

The second model is based on the observation that realised annual death rates fluctuate quite a lot in small populations, which renders random walk models such as the popular Lee–Carter model (Lee and Carter (1992); Brouhns *et al.* (2002)) and its extensions, e.g. Renshaw and Haberman (2003); Cairns *et al.* (2006), unqualified for forecasting in such populations. The idea is to find a larger population, the mortality development of which is assumed to be roughly similar to that of the smaller population of interest. This has given rise to the name SAINT, an abbreviation of Spread–Adjusted InterNational Trend. The possible convergence in international mortality has been examined by e.g. Wilmoth (1998); Tuljapurkar *et al.* (2000); Wilson (2001); Li and Lee (2005).

The spread between the two populations' mortalities is modelled via the specification of the subpopulation's mortality

$$\bar{\mu}_{\text{sub}}(t,x) = H_{\hat{\theta}}(t,x)\exp(y_t' r_x),$$

where $H_{\hat{\theta}}(t,x)$ is the (possibly stochastic) mortality forecast for the reference population, and where $y_t' = (y_{0,t}, \ldots, y_{n,t})$ and $r_x' = (r_{0,x}, \ldots, r_{n,x})$ for some n define the age– and time–dependent spread, which is thus not restricted to adults. Through the specification of the latter terms one can then control the long–term relationship between the two mortality forecasts.

We illustrate this method by letting $H_{\hat{\theta}}(t,x)$ be the (deterministic) frailty–based projection for the aforementioned group of 19 countries, and by letting Denmark constitute the subpopulation. Notice, however, that one could have used any model for the reference population. Denmark's spread dynamics is then governed by specifying y as a three–dimensional VAR(1)–model, and r a set of regressors of orders 0, 1, and 2. As y turns out to be stationary the ratio between the two forecasts is bounded in probability, and has stationary median 1. Estimation of the spread model proceeds via a Poisson likelihood for each gender, as above. The results indicate that the model's out–of–sample predictions are superior to those of the Lee–Carter model.

Our paper has spurred a research project at Cass Business School, and a paper by Cairns *et al.* (2010). Further, the model is used to set the tariff and calculate the technical provisions in ATP. For this implementation ATP was awarded the 2009 Investment & Pensions silver award for innovation.

The second mortality–themed paper is "The evolution of death rates and life expectancy in Denmark" (Chapter 5, written jointly with Søren Fiig Jarner and Chresten Dengsøe). This paper seeks to decompose the substantial improvement in Danish life expectancy observed since 1835 (when data collection began) into age–specific contributions. To this end the functional derivative of (population) life expectancy, $\bar{e}_0$, with respect to the mortality curve, μ, is calculated:

$$\begin{aligned}\frac{\partial \bar{e}_0\left(\mu(1-\epsilon\delta)\right)}{\partial \epsilon}_{|\epsilon=0} &= \frac{\partial}{\partial \epsilon}\int_0^\infty e^{-\int_0^x \mu(y)(1-\epsilon\delta(y))dy}dx_{|\epsilon=0} \\ &= \int_0^\infty \delta(u)D_\mu(u)du,\end{aligned}$$

where

$$D_\mu(u) = \mu(u)\bar{F}(u)\bar{e}_u.$$

is a kernel. That is, improvements' contributions to changes in life expectancy are the product of three factors: the instantaneous force of mortality, the survival probability, $\bar{F}$, and the *remaining* life expectancy – in total amounting to the unconditional average number of lost years. The life expectancy gains between two mortality tables, μ_s and μ_t, can then be calculated, and under the assumption that the kernel is roughly constant between times s and t, i.e. $D_{\mu_u}(v) \approx \bar{D}(v)$, for some $\bar{D}(v)$, we find the approximation

$$\bar{e}_0(\mu_t) - \bar{e}_0(\mu_s) \approx \int_0^\infty \bar{D}(v) \log \frac{\mu_s(v)}{\mu_t(v)} dv,$$

which in turn can be approximated by discretising the integral. This decomposition thus weighs the improvements in age–specific forces of mortality by the time–averaged age–specific kernel between the two time points in question. The main conclusion from the analysis is that improvements up until about 1950 were driven by improvements in child and infant mortality. Since then, the mortality of those age groups has declined rapidly, so that recent improvements have been primarily caused by declines in the mortality rates of age groups 50 to 80 with a clear trend towards higher ages. Possible future improvements cannot be carried by people under the age of 60, as their survival rate is almost one (once infancy is survived anyway), so it is predicted that the older age groups will drive possible future improvements in Danish longevity.

Although not included in the paper it is easy to see that the method also works for *remaining* life expectancy.

1.3 New portfolio problems

The paper "Some solvable portfolio problems with quadratic and collective objectives" (Chapter 6, written jointly with Mogens Steffensen) extends the class of solvable portfolio problems over terminal wealth (as first studied by Merton (1969)) to those in the form

$$\sup_\pi f\left(t, x, \mathbb{E}_{t,x}\left\{g_1\left(X^\pi(T)\right)\right\}, \ldots, \mathbb{E}_{t,x}\left\{g_n\left(X^\pi(T)\right)\right\}\right),$$

for some regular functions f and $g_1, \dots g_n$ – with π being the relative allocation to risky assets in a Black–Scholes market, and X^π the induced wealth. Three illustrative examples are given briefly. As compared with the classical Bellmann approach (e.g. Björk (2009)) the novelty consists of the presence of t and x as well as several g functions. Some of the ground had been broken, though, by Björk and Murgoci (2008), who considered related, but non–overlapping set of problems.

As examples of the applicability of the technique we solve three portfolio problems that are, to our knowledge, new. The former two problems are not covered by the framework of Björk and Murgoci (2008):

First, group utility with proportional sharing, where the criterion is

$$\sup_{\pi} \sum_{i=1}^{n} u_i^{-1} \left(\mathbb{E}_{t,x} \left\{ u_i \left(\alpha_i X^\pi \left(T \right) \right) \right\} \right),$$

for some utility functions $u_1, \dots u_n$, and some positive $\alpha_1, \dots, \alpha_n$ summing to one, and representing the investors' (constant) stakes in the collective. By imposing a sharing rule that is state–independent we disregard optimal risk sharing and focus on this more realistic way of dividing the wealth of a collective. When a group of heterogeneous individuals with expected exponential utility is considered a fully analytic solution is provided:

$$\pi^* x = n \frac{\Lambda}{\sigma} \frac{e^{-r(T-t)}}{\overline{\xi}(\alpha)},$$

with Λ, σ and r denoting the market price of risk, the volatility, and the interest rate respectively – and with $\overline{\xi}$ a weighed average of the individual coefficients of absolute risk aversion. If all stakes are equal, $\alpha_1 = \dots = \alpha_n = 1/n$, then $\overline{\xi} = \sum_{i=1}^{n} \xi_i / n$, the simple average.

In the, perhaps more interesting, case, where all agents have preferences described by expected power utility, a semi–analytical solution is given

$$\pi^* = \frac{\Lambda}{\sigma} \frac{1}{\gamma \left(t \right)},$$

where $\gamma(t)$ is a (time dependent) weighed average of the individual coefficients of relative risk aversion. A system of coupled ordinary differential equations for the weights forming this average is derived and solved numerically for the particularly simple case with $n = 2$. In both utility settings

the loss of certainty equivalent associated with moving from an individually optimal strategy to one that is decided by the group's preferences can be calculated. This provides agents with an assessment of the costs of joining a collective. The corresponding advantages are not included, but are obvious in practice, and an extension thereto is straightforward. The insights obtained from this analysis are not applied in Chapters 2 and 3.

Secondly, we consider dynamic mean–standard deviation optimisation, where the criterion is

$$\sup_{\pi}\left(\mathbb{E}_{t,x}\left\{X^{\pi}\left(T\right)\right\}-\upsilon\left(\mathbb{V}_{t,x}\left\{X^{\pi}\left(T\right)\right\}\right)^{\frac{1}{2}}\right),$$

for some penalty parameter υ. In some sense this problem is more natural than the one arising from the popular mean–variance criterion (which was recently solved by Basak and Chabakauri (2009b)), because benefit and cost are measured in the same units. However, the optimal allocation to risky assets turns out to be zero. This unusual result is caused by the fact that the penalty term is of magnitude $\sqrt{dt}$, while the mean is of magnitude dt.

Thirdly, endogenous habit formation with quadratic utility is analysed:

$$\inf_{\pi}\left(\mathbb{E}_{t,x}\left\{\frac{1}{2}\left(X^{\pi}\left(T\right)-x\beta\left(t\right)\right)^{2}\right\}\right),$$

where $\beta-1$ can be interpreted as a required return over the remaining horizon. In some situations this dynamic version of quadratic utility is more meaningful than the one typically employed, where the target is fixed initially, and may thus be obtained under way – leading to possibly counterintuitive investment strategies. For the present problem the optimal investment strategy solves the partial differential equation

$$\pi^{*\prime}=k_0(t)+k_1(t)\pi^{*}(t)+k_2\pi^{*}(t)^2+k_3\pi^{*}(t)^3,$$

for some coefficients, which are constant in the likely event that the required rate of return is constant. The optimal strategy that is derived is also compared to the classic, so–called pre–committed, solution. The paper also discusses the distinction between pre–committed and sophisticated investors.

Within our framework one can come up with additional interesting portfolio problems. Also, an extension to other cases as well as to more

involved financial markets, possibly incomplete, is natural. For a different direction one could consider intermediate consumption.

2. Pension fund design under long-term fairness constraints

Background. This chapter is a modified version of Kryger (2010c). The paper was presented at the Nordic Finance Network (NFN) Research Workshop in Bergen, May 2008, and at the 5th Conference in Actuarial Science & Finance on Samos, September 2008. I am indebted to two anonymous referees, Frederik Lundtofte and Kristian Miltersen for useful comments and suggestions.

Abstract. We consider optimal portfolio insurance for a mutually owned with–profits pension scheme. First, intergenerational fairness is imposed by requiring that the pension scheme is driven towards a steady state. Subject to this condition the optimal asset allocation is identified among the class of constant proportion portfolio insurance (henceforth CPPI) strategies by maximising expected power utility of the benefit. For a simple contract approximate analytical results are available and accurate, whereas for a more involved contract Monte Carlo methods must be applied to pick out the best design. The main insights are i) aggressive investment strategies are disastrous for the clients, and ii) the results are far less sensitive to the agent's risk aversion than in the classical case of Merton (1969), and as opposed to his results horizon matters even with constant investment opportunities.

2.1 Introduction

Design of fair pension contracts has received a lot of attention in the academic literature over the last fifteen years. Traditionally the notion of fairness concerns the relationship between disjoint stakeholders, namely (equity) owners and a group of clients. In this line of work the main ref-

erence is Grosen and Jørgensen (2002) (extending Briys and de Varenne (1994); Briys and de Varenne (1997)). These models, however, are defined on a finite horizon and have no intergenerational considerations. Thus, in essence they are single life models in which risk sharing takes place between the customer and the owner. In contrast, we consider a pension scheme on an infinite horizon, in which generations exit and enter – thus allowing transferring wealth between generations through the bonus reserve. Our model scheme is owned solely by its members, i.e. there are no separate equity holders.

The fluctuating bonus reserve will lead to some degree of inequality between different cohorts. To study the long term properties we therefore impose stationarity (of the funding ratio) at the outset, and analyse the company only in its (distant future) invariant condition. This implies that, as seen from today, (distant) future clients are all treated the same. Given this restriction, the best such distribution is identified by maximising expected power utility of the resulting benefit. In turn this yields an optimal strategy for portfolio insurance (to obtain stationarity it is necessary to impose an investment rule that guarantees absence of liquidation).

In designing optimal strategies for managing (distributing and investing) bonus reserves for individual contracts Steffensen (2004) uses the framework of Hindi and Huang (1993) to find the optimal distribution rule, which turns out to give rise to so–called local time payments (loosely speaking the optimal distribution policy consists of giving an infinitesimal amount of bonus whenever a certain barrier is hit, which happens infinitely often), and an optimal asset allocation strategy which, at least in a special case, turns out to be a modified version of the mutual fund separation theorem. Inspired by the results we impose a discrete time version of his bonus distribution rule.

The paper by Grosen and Jørgensen (2000) partly remedies the concerns over single life models, and our approach is very much in the spirit of their work – even if we differ in vital aspects. Also, Preisel *et al.* (2010) abandon the single life approach and their model serves as the blueprint for this paper, albeit slightly modified. Another important contribution taking the clients' point of view and integrating the overall pension scheme dynamics is Døskeland and Nordahl (2008a).

As implied in the preceding paragraphs, we believe that the main short-

coming in the existing literature on pension scheme design is that it lacks disentanglement of the individual contracts from the overall financial status of the company. Such separation is necessary to fully understand the complex dynamics of the entire entity. We offer a new approach to optimising with–profits pension scheme operation that can supplement the existing literature on this topic.

As pointed out by Døskeland and Nordahl (2008a) different, even non-overlapping, generations in a with–profits pension scheme may systematically subsidise each other because the bonus reserve does not belong to a specific subset of the collective, but to the collective as a diffuse whole. One way of partially overcoming this is to price each contract *correctly* by adjusting the terms to reflect the scheme's financial and demographic condition at the time of underwriting. In practice, however, the stipulations are not adapted to fit the economic situation of the scheme, hence it is unattractive to enter funds that are poor (and possibly also funds that are increasing in size because the bonus reserve may be crowded out by new entrants).

We take the view of an altruistic board of a mutually owned with–profits pension scheme seeking to treat future generations equally. This property could be obtained by valuing the bonus option at the time of underwriting, and charging for it. But that approach is not desirable since it will compel the scheme to put the bonus option on the balance sheet. Hence, in the short term, our board accepts that the clients are not treated equally. One reason for giving the board such influence over future generations could be that it represents some external party, say a trade union or a governmental institution, representing the common good rather than the present owners – and possibly subsidising the scheme at its foundation. Alternatively the demand for intergenerational fairness may come from some authority. The point is that in the presence of a positive bonus reserve it may be in the interest of the present clients as a whole to dissolve the company rather than leave anything to generations to come. To avoid that temptation we let some external party design the scheme.

In Section 2.2 the pension scheme, its clients, rules and environment are introduced along with the general contract that is analysed in the paper. Our notion of fairness is defined in Section 2.2.3. Section 2.3 contains "approximate analytical" results for the optimal operation of the scheme,

considering a simple contract, while simulated results are provided in Section 2.4. These are partly intended to assess the accuracy of the analytical approximations of Section 2.3, and partly intended to derive optimal rules for more complex contracts. Further, the speed at which the scheme moves towards stationarity is analysed. Finally, Section 2.5 discusses and summarises the results of the preceding sections, while extensions and limitations are also touched upon.

2.2 Model

The environment in which the pension scheme operates is as follows: Let $(B_t)_{t\geq 0}$ be a Brownian motion on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$ generating the filtration $(\mathcal{F}_t)_{t\geq 0}$ with $\mathcal{F}_t \triangleq \sigma\left(B_s,\ 0 \leq s \leq t\right) \cup \mathcal{N},\ (t \geq 0)$ – i.e. augmented by the null sets.

To the pension scheme, but not necessarily to its individual clients, the financial market is frictionless regarding taxes, divisibility, transaction costs, liquidity and portfolio restrictions. This market consists of a bank account with interest intensity r, and a single risky asset (think of a well-diversified portfolio of risky assets) with volatility σ and market price of risk Λ. Hence, the joint value process is

$$dS_t = \mathrm{diag} S_t \left[(r_t, r_t + \Lambda_t \sigma_t)^\top dt + (0, \sigma_t)^\top dB_t \right],\ (t \geq 0,\ S_0 > (0,0)).$$

We let $r_t = r,\ \Lambda_t = \Lambda > 0$, and $\sigma_t = \sigma > 0$ for all $t \geq 0$, i.e. constant investment opportunities.

The scheme divides its assets, A, between the risky asset and the bank account with a fraction, π, invested in the former. Liabilities, L, representing the progressive reserve, changes by the interest rate – hence between reporting periods indexed by $1, 2, \ldots$

$$\begin{aligned} dA_t &= A_t \left[(r + \pi_t \Lambda \sigma) dt + \pi_t \sigma dB_t \right],\ (t \in i + (0,1),\ i \in \mathbb{N},\ A_0 > 0). \\ dL_t &= L_t r dt,\ (t \in i + (0,1),\ i \in \mathbb{N},\ L_0 > 0). \end{aligned}$$

One could allow for additional, non-marketed noise in the processes A and L, but we impose the simplification that the only source of uncertainty is B, which is hedgeable. Hence, in line with the mainstream, we focus on the savings part of the contract, in particular we disregard mortality.

The pension scheme we consider is a mutual with–profits scheme, i.e. one that is owned by its clients. Also, entry is not voluntary, but governed by, say, legislation, and contributions are fixed. Surrender and free policy options are not available. This ownership structure is non–standard in the literature and hence our model cannot be directly compared to those of Briys and de Varenne (1994); Briys and de Varenne (1997); Grosen and Jørgensen (2000). One could consider "old" and "new" clients as disjoint stakeholders; owners respectively customers, but that approach does not fit our purpose; nor does it reflect actual with–profits pensions systems. The compulsory membership may imply that some clients will enter on unacceptable terms, since it is likely preferable to enter a wealthy scheme.

The *funding ratio* is defined as

$$F_t \triangleq \frac{A_t}{L_t}, \ (t \geq 0).$$

Since avoiding insolvency is an integral part of intergenerational fairness we require throughout that the scheme is always sufficiently liquid – in the special sense that $F > 1 + c$ for some $c > -1$ representing a minimum acceptable funding ratio (from the point of view of the scheme's board, but possibly laid down by some monitoring authority). Hence we denote the surplus assets $A - L(1 + c)$ the *bonus reserve*. We let $c = 0$ – corresponding to liquidation upon insolvency – throughout (except in the sensitivity analysis in Section 2.4.3). Any initial bonus reserve ($F_0 > 1 + c$) could have come from anywhere, e.g. as an inheritance from previous generations or as a subsidy from somewhere else.

In Preisel *et al.* (2010) the asset process is controlled by choosing the fraction of wealth invested in the risky asset by optimising one-period expected power utility of the end-period funding ratio minus one. This criterion gives rise to a CPPI strategy (to be introduced in Section 2.2.1) parameterised by the *manager's* coefficient of relative risk aversion, $\gamma > 0$. We, on the other hand, impose a parameterised investment strategy at the outset and take the clients' point of view as a basis for optimisation. This disparity is a natural consequence of our objective being completely different from theirs. For where their aim is to point out the potential conflict between short- and long viewed stakeholders our ambition is the study of *optimal* design as seen from an altruistic standpoint.

The scheme has a rule of distributing bonus periodically,[1] but only when its funding ratio at the turn of the period exceeds some fixed threshold, $\kappa > 1 + c^+$, and in that case all funds above the threshold are distributed to the clients. We refer to κ as the *bonus barrier*.

At the turn of a reporting period assets and liabilities are A_{i^-} and L_{i^-}. Then new contracts are underwritten with value $\Gamma_i L_{i^-},\ (\Gamma_i \geq 0)$, and converted into liabilities $g_i \Gamma_i L_{i^-}$. The parameter $g_i \in (0, (F_{i^-} + \Gamma_i - 1)/\Gamma_i)$ measures the proportion of contributions which is converted into liabilities at time i.[2] Due to the presence of a bonus reserve this parameter may be less than one. Traditionally, this contribution to the (collective) bonus reserve is not explicit. At the same time contracts mature with market value $\Pi_i L_{i^-},\ (\Pi_i \in [0,1])$.[3] This gives rise to the end year *post bonus* funding ratio

$$\begin{aligned} F_{i^+} &= \frac{A_{i^-} + L_{i^-}(\Gamma_i - \Pi_i)}{L_{i^-}(1 + g_i\Gamma_i - \Pi_i)} \wedge \kappa \\ &= \frac{F_{i^-} + \Gamma_i - \Pi_i}{1 + g_i\Gamma_i - \Pi_i} \wedge \kappa,\ (i \in \mathbb{N}). \end{aligned} \tag{2.1}$$

The bonus, b_i, that is in fact allotted such that

$$L_{i^+} = L_{i^-}\left((1 - \Pi_i)\exp(b_i) + g_i\Gamma_i\right), \quad A_{i^+} = A_{i^-} + L_{i^-}\left(\Gamma_i - \Pi_i \exp(b_i)\right),$$

and (2.1) is satisfied, is

$$b_i = \begin{cases} \log \frac{F_{i^-} - \Gamma_i(\kappa g_i - 1)}{\kappa - \Pi_i(\kappa - 1)}, & F_{i^-} > \kappa - \Pi_i(\kappa - 1) + \Gamma_i(\kappa g_i - 1) \\ 0, & \text{otherwise} \end{cases},\ (i \in \mathbb{N}).$$

We let $g_i = 1$ throughout, and we assume that $\Gamma_i = \Pi_i \max\{F_{i^-}, \kappa, \}/\kappa$, so that net inflow is positive. It is indeed relevant to study schemes that are not as demographically stable as this one, nor as rigid, but for our purpose it makes more sense to consider this case of balanced cash flows. The cash flow restriction implies that new contributions are subsidised by

[1] As an alternative to increasing future benefits the company could pay out a cash dividend.

[2] The proportion $1 - g_i$ is intended to pay for the bonus option.

[3] If $\Gamma_i = 0$ and $\Pi_i = 1$ the scheme closes, and this case is not taken into account below.

the bonus reserve abandoned by benefits paid out, and possibly by the staying members. From the flow assumption we get the simpler relations

$$F_{i^+} = F_{i^-} \wedge \kappa, \text{ and } b_i = \left(\log \frac{F_{i^-}}{\kappa} \right)^+, \ (i \in \mathbb{N}).$$

The described bonus rule is not widespread in the academic world nor in practice, where there are tactical, strategical, distributional (intergenerational), and political reasons for smoothing bonus distribution. Rather it is chosen for technical reasons and because of the results of Steffensen (2004) – and as an approximation to what is in fact practiced.

After settling on a short–term optimisation criterion Preisel *et al.* (2010) investigate the properties of the implied stationary distribution of F. Their aim is to point out the divergence between long- and short viewed stakeholders. Concerning the objectives of the present study their model has a few shortcomings, however. First, their optimisation criterion is inappropriate for our purpose. Second, they do not discuss the choice of bonus barrier. And third, their paper does not study the rate at which the system converges towards stationarity. The aim of this paper is to remedy these weaknesses. We address the first of these reservations by introducing a different, altruistic optimisation criterion in Section 2.3. The second and third points of criticism turn out to be partially interrelated, and we discuss those topics in Sections 2.2.3 and 2.4.4.

2.2.1 Investment strategy

An investor with assets, A_0, and a, possibly random, "floor" on wealth $L_0 < A_0$ is said to follow a CPPI strategy (see e.g. Black and Perold (1992)) with multiplier $\alpha > 0$, if his portfolio is self-financing and his absolute allocation to risky assets at time $t \geq 0$ is $\alpha(A_t - L_t)$. The strategy thus reduces exposure when the cushion, $A - L$, decreases and vice versa. In particular, as the cushion approaches zero, the allocation to risky assets approaches zero. Therefore, if paths are continuous, the strategy ensures that the cushion is always positive.

As mentioned above, the optimisation criterion of Preisel *et al.* (2010) gives rise to a particular CPPI strategy, namely one with multiplier $\Lambda / (\sigma\gamma)$, with γ being the *manager's* coefficient of relative risk aversion. We take a

different approach and impose a general CPPI strategy at the outset. In particular we use the family of parameterised investment strategies

$$\pi_t \triangleq \alpha \frac{A_t - L_t(1+c)}{A_t}, \ (t \geq 0, \ \alpha > 0). \tag{2.2}$$

The motivation for choosing this strategy at the outset is that it implies a zero probability of default in the present model framework.[4] Under condition (2.2) the discrete time funding ratio dynamics between updates has i.i.d. lognormal innovations:

Proposition 2.1.

$$\log\left(\frac{F_{i^-} - (1+c)}{F_{(i-1)^+} - (1+c)}\right) = Z_i, \ (i \geq 1).$$

Here $(Z_i)_{i\in\mathbb{N}}$ *is an i.i.d. sequence with* $Z_1 \sim N(m, s^2)$, $(i \in \mathbb{N})$, $s \triangleq \alpha\sigma$, *and* $m \triangleq s\,(\Lambda - s/2)$.

The main result of Preisel *et al.* (2010) is their Theorem 4.1, which states that $(F_{i^+})_{i\in\mathbb{N}}$ admits a stationary distribution if and only if Z_1 has a strictly positive mean. In our case this translates into the condition $s \in (0, 2\Lambda)$. Note that this requirement is independent of the choice of barrier, κ. s is the volatility of the bonus reserve, and we therefore refer to s as "risk".

2.2.2 The contract

To fulfill the purpose of the paper, we consider a contract spanning a period of length $n \in \mathbb{N}$, which can be taken to represent the typical savings period for a pension scheme client. The contract consists of a set of contributions,

[4] It is not by any means clear that it is optimal for the clients as a whole to impose zero probability of default. Having fairness as our primary concern it does *seem* reasonable, however, to put down this restriction already at the modelling level. An alternative with that property is to implement an Option Based Portfolio Insurance (OPBI) strategy, which – like the CPPI – secures a pre-specified lower boundary on portfolio value at the chosen horizon via a put option on the asset portfolio. For a wholly different approach we may decide to use a "free" strategy not protecting the excess assets, instead allowing for bankruptcy - as does the mainstream in pension scheme design.

$(\xi(t))_{t=0}^{n}$, which are determined at time 0, and a corresponding terminal benefit, $W(n)$. To ease calculations we assume without losing much realism that there are no expenses (administrative costs etc.) associated with the contract. The benefit is

$$W(t) \triangleq \sum_{j=0}^{\lfloor t \rfloor} g(j)\xi(j)e^{r(n-j)}e^{\sum_{k=j+1}^{\lfloor t \rfloor} b_k}, \ (0 \leq t \leq n)\,. \tag{2.3}$$

That is, by the terms of the contract, the contribution at time l is transformed into a claim at expiry of $g(l)\xi(l)\exp(r(n-l))$ – and bonus may be added. Hence, contribution $\xi(l)$ is not awarded bonus until time $l+1$. Since bonus is – in part – intended to reflect the return on the contribution it is natural to refrain from crediting bonus immediately. Typically, either $\xi(t) = 1_{(t=0)}$ or $\xi(t) = 1_{(t\in\{0,1,\ldots,n\})}$, but it is also possible to have, say, an increasing contribution plan reflecting inflation. The contract can obviously be thought of as representing a capital pension. The presence of a bonus reserve, however, is usually linked to an insurance product, and we can think of W as a proxy for the value of a whole life annuity bought at expiry.

Since the guaranteed interest rate equals the market rate, $g = 1$, and there is no default risk the contract is an arbitrage. As argued by Døskeland and Nordahl (2008a) this does not really pose a problem. The fact that there is certain excess return to be earned is a consequence of the intergenerational subsidisation that built up a bonus reserve in the past. In other words, the return on the (random) amount $A_{t^-} - L_{t^-}(1+c)$ that previous generations, or some external party, made available, is handed over to the generation entering the scheme at time t. And the fact that they leave a bonus reserve behind upon exit – possibly more than what they received at entry – does not impair the arbitrage.

For a provocative setting, essentially suggesting no guarantees whatsoever, see Sørensen and Jensen (2001). It seems obvious, however, that this can be a gateway for managerial malpractice. Also, they are partially countered by Døskeland and Nordahl (2008b).

2.2.3 Defining fairness

Our concept of fairness is an intergenerational one. It is based on the wish that future clients who join the scheme at times when it is funded differently will get benefits with the same distribution – as seen from today, regardless of the conditions at the time of underwriting. We therefore say that the scheme is *long–term fair* if F admits a stationary distribution. In our setting this is satisfied for any $\kappa > 1$ and for any $s \in (0, 2\Lambda)$. If no stationary distribution exists the probability of obtaining bonus over any final horizon will tend to zero. It is not obvious that this is in fact unsatisfactory to the clients as a whole, but it is clear that it will favour some generations over others.

2.2.4 Objective

The only control we have at our disposal is the "risk", s, since by stochastic dominance we cannot optimise over κ; for the higher is κ the better off is any client joining, cf. Proposition 2.2 below. In finite time there is a trade–off between waiting for an attractive funding ratio distribution (high κ) on one side, and getting bonus underway while approaching stationarity quickly (low κ) on the other side, cf. Section 2.4.4. Once stationarity is attained no such prioritisation has to be made.

We wish to find the stationary distribution of bonus that satisfies future clients better. We therefore assume that F_{0-} is distributed according to its stationary distribution. To the end of finding the optimal design we maximise expected power utility of the discounted benefit, using a *deterministic*, integrable consumption discounting process $(\nu_t)_{t\geq 0}$, i.e. the maximisation object is

$$\begin{cases} \mathbb{E}\left\{e^{-\int_0^n \nu_t dt}\frac{W(n)^{1-\gamma}}{1-\gamma}\right\}, & \gamma \in [0,\infty)\backslash\{1\}. \\ \mathbb{E}\left\{e^{-\int_0^n \nu_t dt}\log W(n)\right\}, & \gamma = 1 \end{cases}, \tag{2.4}$$

where γ represents the member's coefficient of relative risk aversion. The reason for choosing a utility criterion over a financial one is that pension contracts are usually non–tradeable. The maximisation object we have chosen is standard, but notice that it implicitly considers the contribution as sunk cost (otherwise $X - 1$ is the appropriate object).

2.3 Analytical results

The stationary distribution of F is not explicitly known, and hence we study a different Markov chain for which the invariant distribution can be identified. Following Preisel *et al.* (2010), we study a variant of F with Laplace–distributed innovations instead of normally distributed innovations, i.e. the sequence:

$$\begin{aligned}\tilde{F}_{i^-} &= \left(\tilde{F}_{(i-1)^+} - (1+c)\right)\exp(\tilde{Z}_i) + 1 + c,\ (i \geq 1).\\ \tilde{F}_{i^+} &= \tilde{F}_{i^-} \wedge \kappa,\ (i \in \mathbb{N}).\end{aligned}$$

Now $(\tilde{Z}_i)_{i\in\mathbb{N}}$ an i.i.d. sequence with $\tilde{Z}_1$ Laplace–distributed with location m and scale $\lambda^{-1} \triangleq s/\sqrt{2}$ (picked to match the variance of the true Z_1). The density of $\tilde{Z}_1$ is

$$\frac{\lambda}{2}\exp\left(-\lambda|x-m|\right),\ (x \in \mathbb{R}). \tag{2.5}$$

Similarly, we let

$$\tilde{b}_i = \left(\log\frac{\tilde{F}_{i^-}}{\kappa}\right)^+,\ (i \in \mathbb{N})$$

denote bonus in the new regime. We are not familiar with any continuous time stochastic processes for the financial market bringing about this dynamics, but we use the approximation nevertheless. The assumption may be justified by referring to the fact that this Laplace distribution is also symmetric about its mean and is constructed to have the same variance as the true one. Considering the tail behaviour, however, some differences occur because of the fatter tails in (2.5). The 4^{th} central moment is twice that of the true distribution (corresponding to excess kurtosis of 3), and the higher order even moments differ even more. Notice however that such fatter tails comply with some of the criticism of assuming normally distributed returns.

To study the quantitative properties of bonus we need the funding ratio *prior* to bonus distribution, which has the following stationary distribution function

Proposition 2.2.

$$\mathbb{P}\left(\tilde{F}_{0^-} \leq x\right) = \begin{cases} \frac{\lambda-\rho}{\lambda}\left(\frac{(x-(1+c))^+}{\kappa-(1+c)}\right)^{\rho}, & x \leq \kappa e^m - (1+c)(e^m-1). \\ 1 - \frac{\rho e^{\lambda m}}{\lambda+\rho}\left(\frac{x-(1+c)}{\kappa-(1+c)}\right)^{-\lambda}, & x > \kappa e^m - (1+c)(e^m-1). \end{cases} \tag{2.6}$$

The parameter ρ is the non–zero solution to the non–linear equation $1-(\rho/\lambda)^2 = \exp(-\rho m)$. This implies that $\rho \in (0,\lambda)$.

Expression (2.6) differs from that of Preisel *et al.* (2010) who mix Laplace and Gaussian distributed innovations to derive an approximation to the stationary pre-bonus funding ratio distribution.

The distribution is spread out more the higher is s (and the higher is κ). Since the stationary marginal probability of obtaining bonus is unaffected by κ the bonus increases in κ. The bonus frequency decreases with s, but the *conditional* bonus increases with s. An example of the stationary funding ratio distribution can be seen in Figure 2.1, which demonstrates the points just made.

When we consider $\tilde{F}$ the stationary moments of bonus can be derived. They turn out to be expressed in terms of hypergeometric functions (see Weisstein (2008)), which can be evaluated precisely and quickly.

Proposition 2.3. *All moments of $\tilde{b}_0$ exist, and for $c = 0$*

$$\begin{aligned} \mathbb{E}\left\{\tilde{b}_0\right\} = {} & \frac{\rho(\lambda-\rho)}{\lambda}(\kappa-1)^{-\rho}\left[H_1\left(\rho,(\kappa-1)e^m+1\right) - H_1\left(\rho,\kappa\right)\right] \\ & - \frac{\lambda\rho e^{\lambda m}}{\lambda+\rho}(\kappa-1)^{\lambda}H_1\left(-\lambda,(\kappa-1)e^m+1\right) - \frac{\rho}{\lambda}\log\kappa. \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left\{\tilde{b}_0^2\right\} = {} & \frac{\rho(\lambda-\rho)}{\lambda}(\kappa-1)^{-\rho}\left[H_2\left(\rho,(\kappa-1)e^m+1\right) - H_2\left(\rho,\kappa\right)\right] \\ & - \frac{\lambda\rho e^{\lambda m}}{\lambda+\rho}(\kappa-1)^{\lambda}H_2\left(-\lambda,(\kappa-1)e^m+1\right) - \frac{\rho}{\lambda}\left(\log\kappa\right)^2 \\ & - 2\log\kappa\mathbb{E}\left\{\tilde{b}_0\right\}. \end{aligned}$$

The functions H_1 and H_2 can be found in Section 2.6.

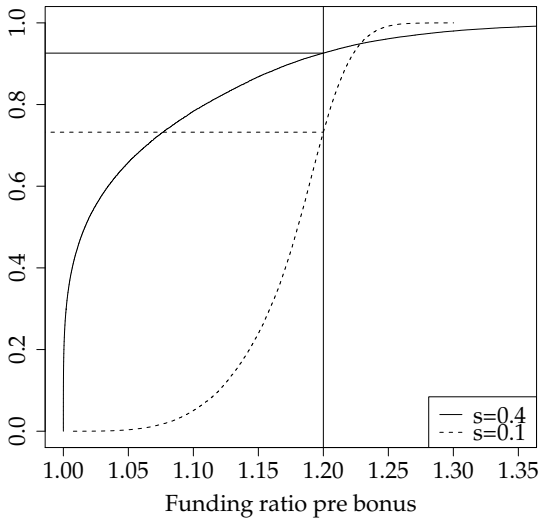

Figure 2.1: Stationary funding ratio distribution. Fixed parameters: $\kappa = 1.2$. Horizontal lines indicate $\mathbb{P}\left(\tilde{F}_{0^-} \leq \kappa\right) = \mathbb{P}\left(\tilde{b}_0 = 0\right)$.

2.3.1 A simple contract

Analytically we can only consider contracts consisting of a single unit of contribution initially transformed into a benefit at time n of

$$\tilde{W}(n) = e^{rn} \exp\left(\sum_{k=1}^{n} \tilde{b}_k\right),$$

i.e. a contract with $\xi(t) = 1_{(t=0)}$ and $g(0) = 1$. For the simple contract maximising (2.4) is equivalent to maximising expected power utility of

$$\tilde{X} \triangleq \exp\left(\sum_{k=1}^{n} \tilde{b}_k\right),$$

i.e. we can disregard ν, r, and the passing of time – and assume that the compounded bonus, X, is received immediately. This is a special feature of the utility functions with constant relative risk aversion used here.

Due to the presence of serially dependent bonuses, $\tilde{X}$ has a complicated distribution that we cannot explicitly derive. Instead we impose two additional approximations to facilitate a semi–analytical solution. First we approximate the serial correlation of $(\tilde{b}_i)_{i\geq 0}$ by analysing the unrestricted underlying random walk with positive drift $(\tilde{Z}_1, \tilde{Z}_1 + \tilde{Z}_2, \dots)$:

Proposition 2.4. *Were* $\left(\tilde{b}_i | \tilde{b}_0 \tilde{b}_i > 0\right)$ *identically distributed for all* $i \geq 0$*, then*

$$\begin{aligned}
\mathbb{E}\left\{\tilde{b}_0 \tilde{b}_1\right\} &= \mathbb{E}\left\{\tilde{b}_0\right\}^2 P_1 \mathbb{P}\left(\tilde{b}_0 > 0\right)^{-1}. \\
\mathbb{E}\left\{\tilde{b}_0 \tilde{b}_2\right\} &= \mathbb{E}\left\{\tilde{b}_0\right\}^2 \left(\frac{P_2}{2} + \frac{P_1^2}{2}\right) \mathbb{P}\left(\tilde{b}_0 > 0\right)^{-1}. \\
\mathbb{E}\left\{\tilde{b}_0 \tilde{b}_3\right\} &= \mathbb{E}\left\{\tilde{b}_0\right\}^2 \left(\frac{P_3}{3} + \frac{P_1^3}{6} + \frac{P_1 P_2}{2}\right) \mathbb{P}\left(\tilde{b}_0 > 0\right)^{-1},
\end{aligned}$$

where $P_j \triangleq \mathbb{P}\left(\tilde{Z}_1 + \cdots + \tilde{Z}_j < 0\right),\ (j \geq 1)$.

Remark 2.5. *To reduce the number of factors in our expressions below we choose to stop at three moments in Proposition 2.4. If desired one could include further moments to improve accuracy, cf. Jarner and Kryger (2009). Also, note that the proposition is valid for the true bonuses as well.*

In order to proceed we need to estimate the higher order serial correlations. To this end assume the following decay:

Approximation 2.6.

$$\begin{aligned}
\mathbb{C}\mathrm{orr}\left\{\tilde{b}_i;\ \tilde{b}_j\right\} &\approx \rho_{0,1} q^{(j-i)-1},\ (j > i \geq 0). \\
\rho_{i,j} &\triangleq \mathbb{C}\mathrm{orr}\left\{b_i;\ b_j\right\},\ (i, j \in \mathbb{N}). \\
q &\triangleq \frac{\rho_{0,3}}{\rho_{0,2}}.
\end{aligned}$$

Remark 2.7. *For the tail behaviour we used* q *– the relationship between step–3 and step–2 covariances. This implies that neither step–2 nor step–3 covariance contribute with our best estimates for them. This simplification does not matter much, though, and can be easily remedied.*

Proposition 2.8. *Under the correlation decay in Approximation 2.6*

$$\begin{aligned}\hat{V}(n) &\triangleq \mathbb{V}\left\{\sum_{i=1}^{n}\tilde{b}_i\right\}/n \\ &\approx \mathbb{V}\left\{\tilde{b}_0\right\}\left(1+\frac{2\rho_{0,1}}{1-q}\left(1-\frac{1-q^n}{n(1-q)}\right)\right),\ (n\geq 1),\end{aligned}$$

which is increasing.

To get an analytical expression for the optimisation problem we also need a distributional approximation for the aggregated bonus. The computationally convenient choice is the normal distribution, which is also asymptotically correct by the central limit theorem. For finite horizons, however, it is flawed by the artificial, negative value space. Nevertheless,

Proposition 2.9. *If Approximation 2.6 held, and if* $\sum_{i=1}^{n}\tilde{b}_i$ *were normally distributed, then maximising expected discounted power utility (with relative risk aversion* $\gamma\in[0,\infty)$*) of the benefit* $\tilde{W}(n)$ *were equivalent to maximising the certainty equivalent bonus*

$$\tilde{b}_{\mathrm{CE}} \triangleq \mathbb{E}\left\{\tilde{b}_0\right\}+\frac{1-\gamma}{2}\hat{V}(n).$$

Remark 2.10. *If* $\gamma=1$ *no approximations are required.*

2.3.2 Optimisation

Proposition 2.9 is now applied to find the optimal s in various cases and analyse these. Throughout this section we keep $\Lambda=1/4$ (having periods of one year in mind). We consider variations in the length of the contract (n), the bonus barrier (κ), and the coefficient of relative risk aversion (γ).

Power utility

The first observation from Table 2.1 presenting the optimal risk is that it does not vary much with κ nor with n (except at high levels of risk aversion). When $\gamma<1$ the increased $\hat{V}(n)$ that results from a longer contract implies higher optimal risk – and oppositely if $\gamma>1$. As can be seen from Proposition 2.9 this is a concavity effect. Similarly, except at

		γ					
$\boldsymbol{\kappa}$	**n**	0	0.5	1	2	5	10
1.1	1	0.277	0.275	0.273	0.268	0.254	0.231
1.1	10	0.281	0.277	0.273	0.264	0.236	0.194
1.1	30	0.283	0.278	0.273	0.261	0.226	0.178
1.1	50	0.284	0.278	0.273	0.261	0.224	0.175
1.2	1	0.276	0.273	0.269	0.262	0.239	0.201
1.2	10	0.283	0.276	0.269	0.254	0.207	0.150
1.2	30	0.287	0.278	0.269	0.250	0.193	0.133
1.2	50	0.287	0.279	0.269	0.249	0.190	0.130
1.3	1	0.276	0.272	0.267	0.258	0.227	0.181
1.3	10	0.284	0.276	0.267	0.247	0.187	0.125
1.3	30	0.289	0.279	0.267	0.241	0.171	0.110
1.3	50	0.290	0.279	0.267	0.240	0.168	0.107

Table 2.1: Optimal risk (s) based on analytical approximation.

the lowest values of γ, an increase in κ will induce lower optimal risk. In this case the reason is that lower barriers are associated with relatively less variable outcomes. As expected the optimal s is decreasing in γ (because $\hat{V}(n)$ increases with s, at least over the relevant range).

It is quite remarkable that even risk neutral clients prefer investment strategies, which are only modestly aggressive, far less than the upper boundary of 2Λ. Apparently the fat right tail of the stationary bonus distribution associated with aggressive investment strategies does not sufficiently compensate the lower marginal probability of obtaining bonus. Across extreme parameterisations $\gamma = 0$, $c = -0.2$, $\kappa = 1.5$, $n = 100$, and $\Lambda \leq 0.4$ one gets an optimal s less than $0.85 \cdot 2\Lambda$. This modest upper limit is astonishing, since it *easily* induces stationarity. In most reasonable cases we are much further from the upper limit as implied in Table 2.1. Also, for reasonable values of γ the optimal s is far above zero. But by taking γ sufficiently high, one can of course get an optimum as close to zero as desired. Most optima are in the range $(0.2, 0.3)$; implying that for a risky

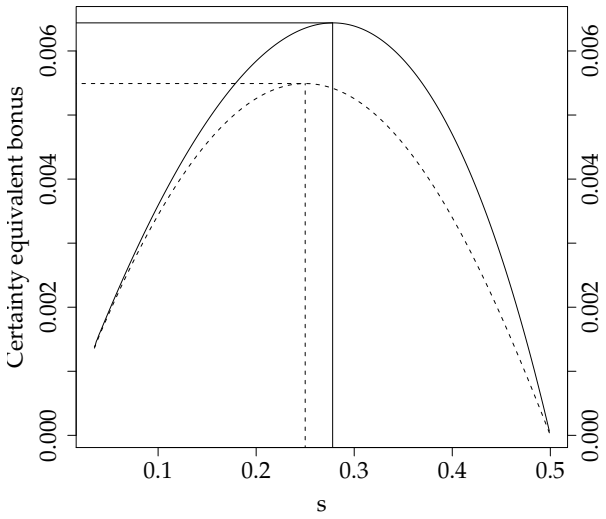

Figure 2.2: Certainty equivalent bonus as a function of s for $\gamma = 0.5$ (full line) and $\gamma = 2$ (dotted line). Fixed parameters: $\kappa = 1.2$, $n = 30$.

asset volatility, σ, of less than 20%, the bonus reserve should be geared since then $\alpha = s/\sigma > 1$.

The classical Merton (1969)–analogue to s is Λ/γ, which would be ∞, 0.5, 0.25, 0.125, 0.05, and 0.025 respectively in the rightmost columns of Table 2.1. Also, in the case of Merton (1969) horizon does not matter, but here n is clearly important (especially at high levels of risk aversion) because bonus is positively serially dependent, so that $\hat{V}(n)$ increases with n. Altogether, except at $\gamma = 1$ the difference is enormous. This should come as no surprise since the problems are very unlike: He considers total return, while our objective is the payoff from a compound option.

Figure 2.2 shows the mapping $s \to \tilde{b}_{\mathrm{CE}}$ for a certain parametrisation, but its appearance is quite representative across a broad range of configurations. The main insight is that it is not very steep around its maximum, implying that it is not overly important to evaluate γ correctly. In the example from Figure 2.2, when $\gamma = 2$, the loss of terminal benefit from

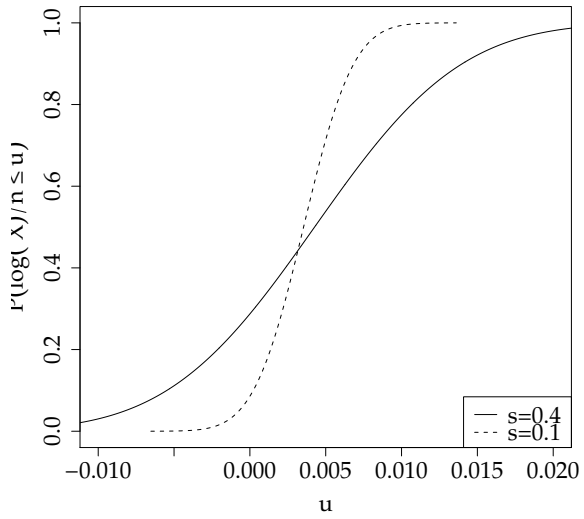

Figure 2.3: Distribution of $\frac{\log \tilde{X}}{n}$. Fixed parameters: $\kappa = 1.2$, $n = 30$.

choosing $s = 0.3$, rather than the optimal $s = 0.25$, is far less than 1% over 30 years.

Seemingly, the choice of barrier, κ, is not so important for low and moderate levels of risk aversion, since the optimal s varies only little with this parameter. As regards the implied investment strategy this is true, of course, but the certainty equivalent bonus differs very markedly across κ (graph not shown). As discussed in Section 2.1 entry to the pension scheme provides an arbitrage and therefore the certainty equivalent bonus is strictly positive.[5]

Figure 2.3 shows the distribution of the average bonus – demonstrating how very adverse (as well as very favourable) outcomes are much more likely as risk increases. Comparison to Figure 2.1 is instructive in clarifying the effect of bonus' serial dependence. For a lengthy discussion of this important concept of being trapped at low funding levels, see Preisel *et al.* (2010).

[5]But when using the normal approximation it needs not be so.

Mean–variance utility

It may be hard to be very specific as to your choice of utility function. To this end, as a pedagogical tool, we show in Figure 2.4 the mean–variance diagram, which provides the set of optimal strategies for any agent with increasing utility of the mean and decreasing utility of the variance of aggregated bonus, and preferences over these two quantities only.

One very useful insight conveyed by Figure 2.4 is the existence of investment strategies inducing stationarity, but which are mean–variance–inefficient. For as can be seen, as s is increased above a certain limit (depending on the parameters) the outcome worsens drastically. This phenomenon is also evident with power utility, cf. Figure 2.1. The reason is the previously mentioned trade–off between frequent and large bonuses, which was illustrated in Figure 2.1, combined with positive serial dependence, which also increases with s.

Limitations and accuracy

In Section 2.4 we provide support for the conclusions above by performing Monte Carlo simulation of the true dynamics. Applying this technique we also find the optimal risk for a contract with several contributions. Further, the convergence of the funding ratio is analysed.

2.4 Numerical results

This section begins by demonstrating the precision of the analytical approximation from above in Section 2.4.1. The simple contract hitherto analysed has a single contribution only, and is thus quite dissimilar to real life contracts. We meet this shortcoming by numerically finding the optimal investment strategy for a contract with several contributions in Section 2.4.2. Section 2.4.3 briefly touches upon the dependence of the results upon the choice of model constants c and Λ. The speed at which the system converges towards stationarity is discussed in Section 2.4.4.

Throughout we impose no approximations, except for using a finite state space rather than all of Ω. Our preferred tool in this section is Monte Carlo simulation, which is described briefly in Section 2.4.5.

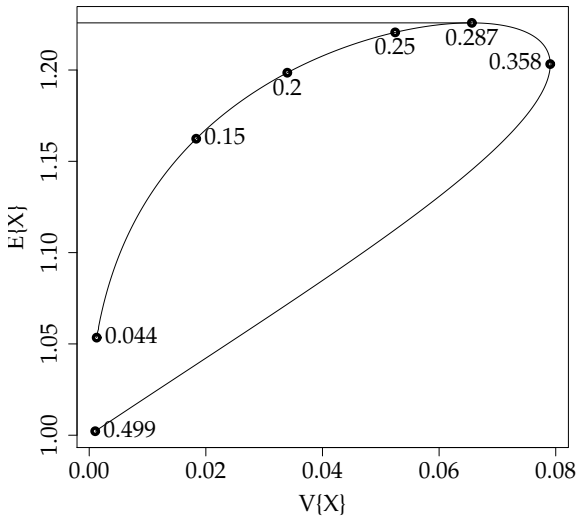

Figure 2.4: Mean–variance diagram for $\tilde{X}$. Fixed parameters: $\kappa = 1.2$, $n = 30$. The labeled points indicate various values of s. Approximation 2.6 was used to calculate the variance.

2.4.1 Comparison to analytical results

This section assesses the overall accuracy of the analytical approximation by simulating the true dynamics using Proposition 2.1. First, the optima are compared, and afterwards, to explain the differences encountered, we compare the benefit distributions.

Power utility optima

First, in Table 2.2 we provide optima comparable to those in Table 2.1. The comparison is quite uplifting; the difference in optimal allocations is less than a few percentage points *except* at high risk aversion, where the previous optima were too low. The main cause for this discrepancy is the fact the value space for X was extended below 1 by matching two moments only (i.e. probability mass was moved quadratically). Such a transformation makes aggressive investment strategies appear artificially

		γ					
$\boldsymbol{\kappa}$	**n**	0	0.5	1	2	5	10
1.1	1	99%	99%	99%	101%	102%	106%
1.1	10	98%	98%	99%	99%	104%	115%
1.1	30	99%	98%	98%	98%	102%	114%
1.1	50	100%	99%	98%	98%	100%	110%
1.2	1	99%	99%	100%	99%	103%	117%
1.2	10	99%	99%	99%	101%	111%	132%
1.2	30	100%	99%	99%	99%	109%	130%
1.2	50	102%	100%	99%	98%	106%	125%
1.3	1	99%	100%	98%	100%	107%	122%
1.3	10	100%	99%	99%	101%	116%	146%
1.3	30	102%	100%	99%	100%	115%	143%
1.3	50	104%	101%	99%	98%	111%	136%

Table 2.2: Optimal risk (s) based on simulation of true process as a percentage of the optima in Table 2.1.

unattractive when the utility function is very concave.

Consequently, the true optima differ far less across γ than did those in Section 2.3, which makes it easier to embrace individuals with different appetites for risk in a common investment policy.

At the outset it is not obvious if short or long contracts are optimised more precisely when applying the analytical approximation. The longer the contract in question the better the normality approximation works. Oppositely, the correlation approximation is worse for long contracts because the correlations decay slower than at rate q. From Table 2.2 we conclude that long–contract–optima are estimated more precisely, i.e. the normality assumption matters more (at least for the levels of risk aversion where the errors are more serious).[6]

The qualitative conclusions from Section 2.3 regarding n, κ, and γ hold true.

[6]In making this conclusion we disregard the case $n = 1$.

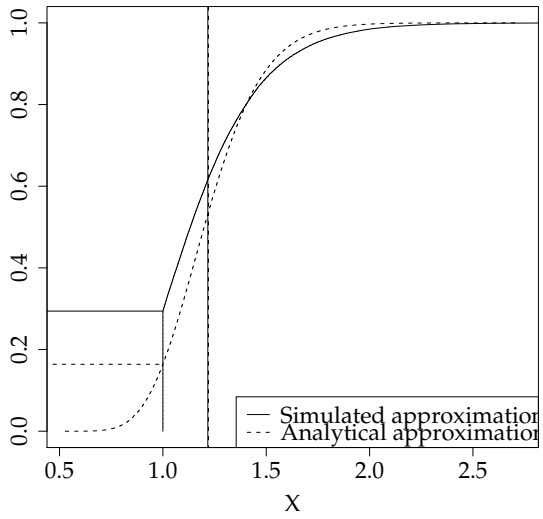

Figure 2.5: Distribution of discounted benefit. Fixed parameters: $s = 0.25$, $\kappa = 1.2$, $n = 30$. Vertical lines show expected discounted benefit (which are equal).

The accuracy of the analytical approximation

In terms of optima we are pleased with the accuracy of the analytical approximation. The approximations were made at a more primitive level, however, and in order to explain the deviations encountered our analysis proceeds at that level. To this end consider Figure 2.5 comparing the true (simulated) distribution of X to that obtained via the analytical approximation. Clearly, the normality approximation in Proposition 2.9 brings about a rather substantial difference between the two by artificially extending the support of X below one. In fact, informal experiments indicate that almost the entire difference between the two distributions in Figure 2.5 stems therefrom, whereas for reasonable horizons, the effect of the approximation regarding serial correlation (Approximation 2.6) is much less. This holds in spite of the latter approximation drastically reducing the true variance of X and thus implying a lighter right tail. Finally,

the Laplace approximation is rather innocuous – it brings about slightly fatter tails, thus partially offsetting the aforementioned error.

2.4.2 A complex contract

In this section we will find the optimal risk for a contract with contributions in every period. This construction implies that later bonuses are more important than earlier ones because they "act" on a larger amount. For simplicity we assume that the contribution vector is

$$\xi \triangleq (1, \exp(\eta + r), \ldots, \exp((\eta + r)n))$$

for some fixed contribution inflation net of interest, η. Then the terminal benefit is

$$W(n) = \exp(nr) \sum_{j=0}^{n} \exp\left(\eta j + \sum_{k=j+1}^{n} b_k\right).$$

We perform the same simulations as above and calculate expected power utility of $W(n)$. For illustrative purposes we use the admittedly high net contribution inflation $\eta = 0.1$, but the qualitative conclusions hold for $\eta = 0$ as well.

The optima are shown, indirectly, in Table 2.3. It turns out that the consequence of increasing net contribution inflation, η, is similar to the effect of reducing n: At low risk aversions the optimal s decreases as η increases, and vice versa. The simple explanation for this is that as η increases more emphasis is put on the last bonus allotted, and thus on the marginal properties of F, while the serial dynamics matters less.

Finally, notice two further points about the optima. First, at low levels of risk aversion the differences between $\eta = -\infty$ (the simple contract) and $\eta = 0.1$ are small. This is because for such agents it is almost exclusively the mean bonus that determines expected utility. Consequently the analytical approximation can be applied with high accuracy to this, somewhat different, problem as well. Second, because serial correlation is downplayed with the approximation, the more risk-averse clients' optima increase substantially – in turn making it even more "feasible" to pool individuals with different attitudes towards risk in a common investment policy. Third, the

		γ					
$\boldsymbol{\kappa}$	**n**	0	0.5	1	2	5	10
1.1	10	99%	100%	101%	102%	106%	110%
1.1	30	99%	100%	101%	103%	107%	112%
1.1	50	98%	100%	101%	103%	107%	113%
1.2	10	99%	100%	102%	104%	112%	121%
1.2	30	98%	100%	102%	106%	116%	127%
1.2	50	97%	100%	103%	107%	118%	129%
1.3	10	98%	100%	103%	107%	118%	131%
1.3	30	96%	100%	103%	110%	125%	141%
1.3	50	95%	100%	104%	112%	128%	145%

Table 2.3: Optimal risk (s) with net contribution inflation $\eta = 0.1$ as a percentage of the optima with $\eta = -\infty$ (the simple contract, cf. Table 2.2). The optima can be backed out using Tables 2.1 and 2.2. Based on simulation of true process.

implied loss of certainty equivalent from choosing a slightly suboptimal s is almost zero because the distribution of X is relatively much narrower when there are several contributions. This property further assists the formation of an investment collective.

2.4.3 Sensitivity analysis

The optimisations above were all performed with a fixed market price of risk, $\Lambda = 1/4$. In practice it is not widely agreed what the magnitude of this quantity might be. Therefore we briefly investigate how the results depend on that vital parameter. Also, pension funds may exist in regimes differing w.r.t. liquidation rules. Hence, we also examine how our results are affected by choosing a non–zero constant for c. As c can probably be observed for every entity this part of the sensitivity analysis does not relate to the insecurity in applying the results; rather to the optima's robustness towards different environments. We have fixed $n = 30$, $\kappa = 1.2$ in this section. For simplicity we perform the sensitivity analysis for the simple

contract only.

As expected the magnitude of the optimal s is very sensitive towards the return/risk-relationship of the financial market. We do not provide the full results, but with $\gamma = 0$ one gets the optima 0.223, 0.287, 0.356 at $\Lambda = 0.2,\ 0.25,\ 0.3$ respectively. Similarly at $\gamma = 10$ the optimal s varies almost as radically, being 0.148, 0.174, 0.199 respectively market prices of risk $\Lambda = 0.2,\ 0.25,\ 0.3$. Clearly, as opposed to the classical rule of Merton (1969), optimal allocation to risky assets is not linear in Λ.

Increasing c corresponds to operating closer to the boundary – all else equal. Therefore the effect of increasing c is similar to that of lowering κ. The numbers confirm this conjecture and are available upon request. Notice, however, that both $\kappa - (1 + c)$ and $(\kappa - (1 + c))/\kappa$ matter, so no direct translation can be made.

2.4.4 Speed towards stationarity

Assuming that it is desirable to obtain fairness via stationarity it is conceivably also attractive to approach such invariance as quickly as possible. For if stationarity is approached too slowly, today's clients will not even approximately sample the same distribution as will future clients. And in that case, stationarity is, more or less, in vain (although there is obviously no connection between a swift approach towards a particular stationary distribution and the desirability thereof).

Casual experiments suggest that for low initial fundings stationarity is approached quicker with high values of s, whereas for high values of F_0 choosing s low results in the faster move towards the invariant distribution. The reason for this difference is the upward drift in F.

The simulations also confirm the presupposition that stationarity is approached faster when $(F_{0^+} - 1) / (\kappa - 1)$ is not too low, nor too high. This ratio, however, does not affect the *limiting* convergence rate, which is determined by s solely. As an example, see Figure 2.6 showing – at a certain parametrisation – how the stationary distribution of the funding ratio is approached rather quickly.

The rate at which a stationary Markov chain moves towards its invariant distribution can in principle be bounded *analytically* (see e.g. Baxendale (2005)), but the bounds turn out useless in our case.

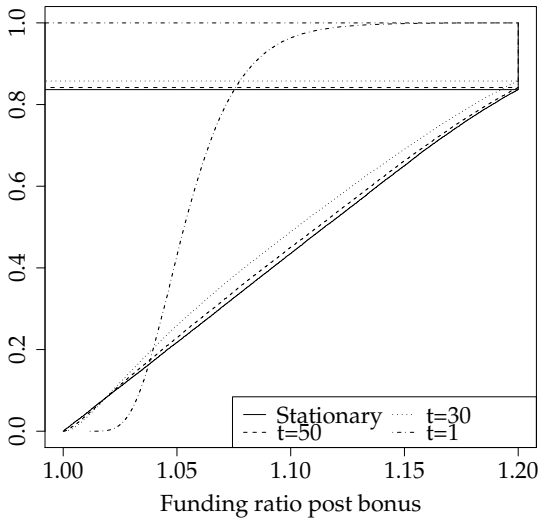

Figure 2.6: Funding ratio distribution at $t \in \{1, 30, 50, \infty\}$. Fixed parameters: $s = 0.25$, $\kappa = 1.2$. Horizontal lines show $\mathbb{P}(F_{t^+} < \kappa | F_{0^-} = 1.05) = \mathbb{P}(b_t = 0 | F_{0^-} = 1.05)$.

2.4.5 Technical description of the Monte Carlo simulation

The simulations were done in the freeware statistical computing package `R`. To evaluate the hypergeometric functions we have used the package `hypergeo`. The stationary distribution of $\tilde{F}$ (from Proposition 2.2) was stratified using a trapezoidal method. To get a proxy for the stationary distribution of F we took 200 Gaussian steps ahead from that stratified distribution.[7] We then simulate F a further n periods ahead to estimate X.

[7]For virtually any choice of (s, κ) 200 steps seems to be more than enough – based on the distance between consecutive (pre-bonus) funding ratio distributions – to get an invariant distribution. In fact the improvement after 20 steps is negligible for most parameter sets and 50 steps seems to be enough for almost all reasonable parameterisations. Starting from a fixed funding ratio, on the other hand, a comparable result requires far more time steps.

Setting the seed manually allows all results to be reproduced. Throughout we did 100,000 trials. The value space for s was approximated by the set of equidistant points $\{0.001, 0.002, \ldots, 2\Lambda - 0.001\}$.

2.5 Concluding remarks

2.5.1 The bonus option

An important issue is whether or not to include the bonus option on the liability side of the balance sheet. From a strictly legal point of view one can absolutely argue in favour of excluding the option, since pension funds are rarely strictly obliged by law to follow a particular bonus policy, even if it has been made public. Grosen and Jørgensen (2000) call this a *counter option* (held by the company). In addition, rules are subject to change for legal, political or strategic reasons. However, if the bonus option is disregarded, the principle of equivalence states that we must have $g = 1$ for the set of pure savings contracts defined by (2.3). Alternatively, one could take a more pragmatic approach to accounting to justify the choice of some $g < 1$ without explicitly regarding the bonus option. It would imply that new entrants contribute explicitly to the bonus reserves, even though they hold no strict statutory claim on it.

If the option *is* included its value depends on whether or not the scheme is open to new members or not. If the scheme is closed the value of the bonus option as a whole equals the bonus reserve, but the total option value may be disaggregated into different cohorts' shares thereof. In a scheme that continues to take in new contributions it is not obvious whether the value of the present members' claim on bonus has a larger or smaller value than the bonus reserve.

2.5.2 Policy implications

To apply the results one must choose a bonus barrier. This choice is a trade–off between obtaining a desirable stationary distribution (high barrier) and attaining fairness quickly (low barrier). After picking appropriate values for contract length, relative risk aversion, and possibly net contribution inflation one only needs to estimate the volatility and market price of risk of a suitable portfolio of risky assets.

The implied optima from cases with very unlike values of γ are not very different. Further, the implied difference in utility between quite different investment strategies is relatively modest, which makes it possible to embrace such different attitudes towards risk in a common investment policy. For such an enterprise a proportion around $\Lambda\sigma^{-1}$ of the bonus reserve invested in risky assets seems to work well for almost all cases. Alternatively, one could set up separate funds for individuals with varying appetite for risk.

2.6 Proofs

Proof of Proposition 2.1. Apply Itô's formula to $F - (1 + c)$. □

Proof of Proposition 2.2. Let $f \in (1 + c, \kappa]$, $x \in (1 + c, \infty)$, and $\nu(f) = (1 + c) + (f - (1 + c))e^m$.

$$\begin{aligned}
\mathbb{P}\left(\tilde{F}_{1^-} \leq x \middle| \tilde{F}_{0^+} = f\right) &= \mathbb{P}\left((f - (1 + c))e^{\tilde{Z}_0} + (1 + c) \leq x\right) \\
&= \mathbb{P}\left(\tilde{Z}_0 \leq \log\left(\frac{x - (1 + c)}{f - (1 + c)}\right)\right) \\
&= \int_{-\infty}^{\log\left(\frac{x-(1+c)}{f-(1+c)}\right)} \frac{\lambda}{2} e^{-\lambda|z-m|} dz \\
&= \begin{cases} \frac{e^{-\lambda m}}{2}\left(\frac{x-(1+c)}{f-(1+c)}\right)^{\lambda}, & x \leq \nu(f) \\ 1 - \frac{e^{\lambda m}}{2}\left(\frac{x-(1+c)}{f-(1+c)}\right)^{-\lambda}, & x > \nu(f) \end{cases}
\end{aligned}$$

Equation (20) of Preisel *et al.* (2010) and integration by parts yields

$$
\begin{aligned}
&\mathbb{P}\left(\tilde{F}_{1^-} \leq x\right) \\
&= \int_{1+c}^{\kappa} \mathbb{P}\left(\tilde{F}_{1^-} \leq x \middle| \tilde{F}_{0^+} = f\right) d\mathbb{P}\left(\tilde{F}_{0^+} \leq f\right) \\
&= \left[\mathbb{P}\left(\tilde{F}_{1^-} \leq x \middle| \tilde{F}_{0^+} = f\right) \mathbb{P}\left(\tilde{F}_{0^+} \leq f\right)\right]_{1+c}^{\kappa} \\
&\qquad - \int_{1+c}^{\kappa} \mathbb{P}\left(\tilde{F}_{0^+} \leq f\right) d\mathbb{P}\left(\tilde{F}_{1^-} \leq x \middle| \tilde{F}_{0^+} = f\right) \\
&= \mathbb{P}\left(\tilde{F}_{1^-} \leq x \middle| \tilde{F}_{0^+} = \kappa\right) + \frac{\lambda - \rho}{2}(\kappa - (1+c))^{-\rho} \\
&\qquad \left[e^{\lambda m}(x-(1+c))^{-\lambda} \int_0^{(\kappa-(1+c))\wedge\left((x-(1+c))e^{-m}\right)} f^{\rho+\lambda-1} df \right. \\
&\qquad \left. + e^{-\lambda m}(x-(1+c))^{\lambda} \int_{(\kappa-(1+c))\wedge((x-(1+c))e^{-m})}^{\kappa-(1+c)} f^{\rho-\lambda-1} df \right] \\
&= \begin{cases} \frac{\lambda-\rho}{\lambda}\left(\frac{x-(1+c)}{\kappa-(1+c)}\right)^{\rho}, & x \leq \nu(\kappa) \\ 1 - \frac{\rho e^{\lambda m}}{\lambda+\rho}\left(\frac{x-(1+c)}{\kappa-(1+c)}\right)^{-\lambda}, & x > \nu(\kappa) \end{cases}
\end{aligned}
$$

□

Proof of Proposition 2.3. To see that any moment of $\tilde{b}_0$ exists let $n \geq 1$, $\epsilon \in (0, (1 \wedge \lambda)/n)$. Then $y^{\epsilon}/\epsilon > \log y,\ (y > 0)$. Hence

$$
\begin{aligned}
\lim_{x\to\infty} \int (x-(1+c))^{-\lambda-1} (\log x)^n \, dx &\leq \epsilon^{-n} \lim_{x\to\infty} \int (x-(1+c))^{-\lambda-1} x^{\epsilon n} dx \\
&\leq \epsilon^{-n} \lim_{x\to\infty} \int \left[(x-(1+c))^{-\lambda-1} \right. \\
&\qquad \left. ((x-(1+c))^{\epsilon n} + (1+c)^{\epsilon n})\right] dx \\
&< \infty \text{ since } \epsilon n - \lambda < 0.
\end{aligned}
$$

To derive the moments define for $q \neq 0$, $j \in \mathbb{N}$, and $x > 1$

$$
H_j(q,x) \triangleq \int (x-1)^{q-1} (\log x)^j \, dx.
$$

Then

$$
\begin{aligned}
H_1(q,x) &= \frac{(x-1)^q}{q}\log x - \frac{x^q}{q^2}F_{2,1}\left(-q,-q,1-q,1/x\right). \\
&\to_{x\to\infty} 0 \text{ for } q\in(-\infty,0). \\
H_2(q,x) &= \frac{2x^q}{q^3}\left[F_{3,2}\left(\left(-q,-q,-q\right),\left(1-q,1-q\right),1/x\right)\right. \\
&\qquad \left. -qF_{2,1}\left(-q,-q,1-q,1/x\right)\log x\right] + \frac{(x-1)^q}{q}\left(\log x\right)^2 \\
&\to_{x\to\infty} 0 \text{ for } q\in(-\infty,0),
\end{aligned}
$$

where $F_{i,j}$ denotes the generalised hypergeometric function, cf. Weisstein (2008).

$$
\begin{aligned}
\mathbb{E}\left\{\tilde{b}_0\right\} &= \int_\kappa^\infty \log(x/\kappa)d\mathbb{P}\left(\tilde{F}_{0^-}\le x\right) \\
&= \frac{\rho(\lambda-\rho)}{\lambda}(\kappa-1)^{-\rho}\int_\kappa^{(\kappa-1)e^m+1}(x-1)^{\rho-1}\log x dx \\
&\quad + \frac{\lambda\rho e^{\lambda m}}{\lambda+\rho}(\kappa-1)^{\lambda}\int_{(\kappa-1)e^m+1}^\infty (x-1)^{-\lambda-1}\log x dx \\
&\quad - \log\kappa\mathbb{P}\left(\tilde{F}_{0^-}\ge\kappa\right) \\
&= \frac{\rho(\lambda-\rho)}{\lambda}(\kappa-1)^{-\rho}\left[H_1\left(\rho,(\kappa-1)e^m+1\right)-H_1\left(\rho,\kappa\right)\right] \\
&\quad - \frac{\lambda\rho e^{\lambda m}}{\lambda+\rho}(\kappa-1)^{\lambda}H_1\left(-\lambda,(\kappa-1)e^m+1\right) - \frac{\rho}{\lambda}\log\kappa.
\end{aligned}
$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left\{\tilde{b}_0^2\right\} &= \int_\kappa^\infty (\log(x/\kappa))^2 \, d\mathbb{P}\left(\tilde{F}_{0^-} \leq x\right) \\
&= \int_\kappa^\infty \left[(\log x)^2 + (\log\kappa)^2 - 2\log\kappa\log x\right] d\mathbb{P}\left(\tilde{F}_{0^-} \leq x\right) \\
&= \int_\kappa^\infty (\log x)^2 \, d\mathbb{P}\left(\tilde{F}_{0^-} \leq x\right) + \frac{\rho}{\lambda}(\log\kappa)^2 \\
&\quad - 2\log\kappa\left[\mathbb{E}\left\{\tilde{b}_0\right\} + \frac{\rho}{\lambda}\log\kappa\right] \\
&= \frac{\rho(\lambda-\rho)}{\lambda}(\kappa-1)^{-\rho}\left[H_2\left(\rho,(\kappa-1)e^m+1\right) - H_2\left(\rho,\kappa\right)\right] \\
&\quad - \frac{\lambda\rho e^{\lambda m}}{\lambda+\rho}(\kappa-1)^{\lambda} H_2\left(-\lambda,(\kappa-1)e^m+1\right) \\
&\quad - \frac{\rho}{\lambda}(\log\kappa)^2 - 2\log\kappa\mathbb{E}\left\{\tilde{b}_0\right\}.
\end{aligned}$$

□

Proof of Proposition 2.4. Let $(\tilde{Z}_i)$ be an i.i.d. sequence. For $i \in \mathbb{N}$ introduce $Y_i \triangleq \log\left(\frac{\kappa-1}{F_{i^+}-1}\right)$ satisfying the recurrence $Y_{i+1} = \left(Y_i + \tilde{Z}_{i+1}\right)^+$. Define

$$\begin{aligned}
S_j &\triangleq \sum_{i=1}^{j} \tilde{Z}_i, \ (j \geq 1). \\
P_j &\triangleq \mathbb{P}\left(S_j \leq 0\right), \ (j \geq 1). \\
\tau_j &\triangleq \mathbb{P}\left(S_1 < 0, \cdots, S_{j-1} < 0, S_j \geq 0\right), \ (j \geq 1). \\
\tau(s) &\triangleq \sum_{j=1}^{\infty} \tau_j s^j, \ (0 \leq s \leq 1).
\end{aligned}$$

P_j gives the probability that the underlying, *unrestricted* random walk, S, is negative j periods ahead. τ_j is the probability that the *unrestricted* random walk stays negative before time j and goes positive (for the first time) at time j. Given we start at full funding, $F_{0^+} = \kappa$ and thus $Y_0 = 0$, it is also the probability that bonus is awarded at time j, but not at

times $1, \ldots, j-1$. $\tau(\cdot)$ is the probability generating function for the (non-delayed) regeneration time of Y with density $(\tau_j)_j$. By differentiating $\tau(\cdot)$ and evaluating at zero we obtain the well-known relation

$$\tau_j j! = \tau^{(j)}(0), \ (j \in \mathbb{N}).$$

Further, by Proposition 1 of Feller (1971), p. 413

$$\tau(s) \triangleq 1 - \exp\left(-\sum_{j=1}^{\infty} \frac{s^j}{j} P_j\right), \ (0 \leq s \leq 1).$$

Differentiation of this expression yields

$$\begin{aligned}
\tau^{(1)}(s) &= (1-\tau(s)) \sum_{j=1}^{\infty} s^{j-1} P_j. \\
\tau^{(2)}(s) &= -\tau^{(1)}(s) \sum_{j=1}^{\infty} s^{j-1} P_j + (1-\tau(s)) \sum_{j=2}^{\infty} (j-1) s^{j-2} P_j. \\
\tau^{(3)}(s) &= -\tau^{(2)}(s) \sum_{j=1}^{\infty} s^{j-1} P_j - 2\tau^{(1)}(s) \sum_{j=2}^{\infty} (j-1) s^{j-2} P_j \\
&\quad + (1-\tau(s)) \sum_{j=3}^{\infty} (j-1)(j-2) s^{j-3} P_j.
\end{aligned}$$

And evaluating at 0 yields

$$\begin{aligned}
\tau_1 &= P_1. \\
\tau_2 &= \frac{-P_1^2 + P_2}{2}. \\
\tau_3 &= \frac{2P_3 + P_1^3 - 3P_1 P_2}{6}.
\end{aligned}$$

Now consider the convolutions

$$F^{*j}(k) \triangleq \mathbb{P}\left(\mathcal{T}_1 + \cdots + \mathcal{T}_j = k\right), \ (k \geq j \geq 1),$$

where $\mathcal{T}_i : \Omega \to \mathbb{N}$ is the i^{th} *non-delayed* regeneration time for Y, $(i \in \mathbb{N}_+)$. The $\mathcal{T}_i$ are i.i.d. according to $(\tau_j)_{j \geq 1}$. Hence $F^{*j}(k)$ is the probability that

the j^{th} regeneration (and thus the j^{th} bonus) occurs at time k. Writing the convolutions in terms of the τ_js gives

$$F^{*1}(k) = \tau_k, \ (k \geq 1).$$
$$F^{*j}(k) = \sum_{i=1}^{k} F^{*(j-1)}(i)\tau_{k-i}, \ (k \geq j > 1).$$

In particular,

$$\begin{aligned}
\sum_{j=1}^{1} F^{*j}(1) &= \tau_1 \\
&= P_1. \\
\sum_{j=1}^{2} F^{*j}(2) &= \tau_2 + \tau_1^2 + 0 \\
&= \frac{P_2}{2} + \frac{P_1^2}{2}. \\
\sum_{j=1}^{3} F^{*j}(3) &= \tau_3 + 2\tau_1\tau_2 + F^{*3}(3) \\
&= \tau_3 + 2\tau_1\tau_2 + \tau_1^3 \\
&= \frac{P_3}{3} + \frac{P_1^3}{6} + \frac{P_1 P_2}{2}
\end{aligned}$$

Writing the "joint bonus probability" in terms of the convolutions yields

$$\begin{aligned}
\mathbb{P}\left(\tilde{b}_0\tilde{b}_i > 0\right)\mathbb{P}\left(\tilde{b}_0 > 0\right)^{-1} &= \mathbb{P}\left(\tilde{F}_{0^+} = \tilde{F}_{i^+} = \kappa\right)\mathbb{P}\left(\tilde{b}_0 > 0\right)^{-1} \\
&= \mathbb{P}\left(\tilde{F}_{i^+} = \kappa \middle| \tilde{F}_{0^+} = \kappa\right) \\
&= \mathbb{P}\left(\mathcal{T}_1 = i\right) + \cdots + \mathbb{P}\left(\mathcal{T}_1 + \cdots + \mathcal{T}_i = i\right) \\
&= \sum_{j=1}^{i} F^{*j}(i), \ (i \geq 1).
\end{aligned}$$

Finally, use the assumption that $\forall i \geq 0 : \left(\tilde{b}_i | \tilde{b}_0 \tilde{b}_i > 0\right)$ are identically distributed. For then, since the random variables $(\tilde{b}_0 | \tilde{b}_0 \tilde{b}_i > 0)$ and $(\tilde{b}_i | \tilde{b}_0 \tilde{b}_i > 0)$ are independent (due to regeneration at $\tilde{b}_0 > 0$), we may calculate

$$\begin{aligned}
\mathbb{E}\left\{\tilde{b}_0 \tilde{b}_i\right\} &= \mathbb{P}\left(\tilde{b}_0 \tilde{b}_i > 0\right) \mathbb{E}\left\{\tilde{b}_0 \tilde{b}_i \middle| \tilde{b}_0 \tilde{b}_i > 0\right\} \\
&= \mathbb{P}\left(\tilde{b}_0 \tilde{b}_i > 0\right) \mathbb{E}\left\{\tilde{b}_0 | \tilde{b}_0 \tilde{b}_i > 0\right\} \mathbb{E}\left\{\tilde{b}_i | \tilde{b}_0 \tilde{b}_i > 0\right\} \\
&= \mathbb{P}\left(\tilde{b}_0 \tilde{b}_i > 0\right) \left(\frac{\mathbb{E}\left\{\tilde{b}_0\right\}}{\mathbb{P}\left(\tilde{b}_0 > 0\right)}\right)^2 \\
&= \sum_{j=1}^{i} F^{*j}(i) \mathbb{E}\left\{\tilde{b}_0\right\}^2 \mathbb{P}\left(\tilde{b}_0 > 0\right)^{-1}, \ (i \geq 1).
\end{aligned}$$

□

Proof of Proposition 2.8. By recursion

$$\mathbb{V}\left\{\sum_{j=1}^{n} \tilde{b}_j\right\} = \mathbb{V}\left\{\tilde{b}_0\right\} \left(n + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{C}\text{orr}\left\{\tilde{b}_i;\ \tilde{b}_j\right\}\right).$$

Use Approximation 2.6 and the familiar property of partial sums of geometric series

$$\sum_{i=0}^{n} q^i = \frac{1 - q^{n+1}}{1 - q}.$$

to derive

$$\begin{aligned}
\sum_{j=i+1}^{n} \mathbb{C}\text{orr}\left\{\tilde{b}_i;\ \tilde{b}_j\right\} &= \rho_{0,1} q^{-i-1} \left(\sum_{j=0}^{n} q^j - \sum_{j=0}^{i} q^j\right) \\
&= \rho_{0,1} q^{-i-1} \frac{1 - q^{n+1} - 1 + q^{i+1}}{1 - q} \\
&= \frac{\rho_{0,1}}{1 - q} \left(1 - q^{n-i}\right).
\end{aligned}$$

Finally, use this expression to calculate

$$\begin{aligned}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{C}\text{orr}\left\{\tilde{b}_i;\ \tilde{b}_j\right\} &= \frac{\rho_{0,1}}{1-q}\sum_{i=1}^{n-1}\left(1-q^{n-i}\right)\\ &= \frac{\rho_{0,1}}{1-q}\left(n-1-q^n\sum_{i=1}^{n-1}q^{-i}\right)\\ &= \frac{\rho_{0,1}}{1-q}\left(n-1-q^n\left(\frac{1-q^{-n}}{1-q^{-1}}-1\right)\right)\\ &= \frac{\rho_{0,1}}{1-q}\left(n-1-\frac{q^n}{1-q^{-1}}\left(q^{-1}-q^{-n}\right)\right)\\ &= \frac{\rho_{0,1}}{1-q}\left(n+\frac{1-q^n}{q-1}\right).\end{aligned}$$

Dividing through by n we get

$$\mathbb{V}\left\{\sum_{j=1}^{n}\tilde{b}_j\right\}/n = \mathbb{V}\left\{\tilde{b}_0\right\}\left(1+2\frac{\rho_{0,1}}{1-q}\left(1+\frac{1-q^n}{n(q-1)}\right)\right).$$

□

Proof of Proposition 2.9. The certainty equivalent of $\tilde{W}$ is

$$\begin{aligned}&(1-\gamma)\mathbb{E}\left\{\left.\frac{\left[e^{-\int_0^n\delta_s ds}\tilde{W}(n)\right]^{1-\gamma}}{1-\gamma}\right|\mathcal{F}_0\right\}^{1/(1-\gamma)}\\ &\qquad= e^{-\int_0^n\delta_s ds}\mathbb{E}\left\{\left.\exp\left(\sum_{k=1}^{n}\left(r+\tilde{b}_k\right)\right)\right|\mathcal{F}_0\right\}\\ &\qquad= e^{-\int_0^n\delta_s ds}\exp\left(nr+\sum_{k=1}^{n}\left(\mathbb{E}\left\{\tilde{b}_0\right\}+\frac{1-\gamma}{2}\hat{V}(n)\right)\right),\end{aligned}$$

demonstrating that the certainty equivalent bonus is $\mathbb{E}\left\{\tilde{b}_0\right\}+\frac{1-\gamma}{2}\hat{V}(n)$.
□

3. Fairness vs. efficiency of pension schemes

Background. The paper in this chapter is a slightly updated version of Kryger (2010a). I thank an anonymous referee for useful comments and suggestions. I am also indebted to David McCarthy, whose ideas on fairness inspired the preliminary reflections in Section 3.3. The paper was presented at the AFIR / LIFE Colloquium in Munich, September 2009.

Abstract. The benefits that members of with–profits pension schemes obtain are determined by the scheme design and the controlled funding level at the time of entry. This paper examines efficiency and intergenerational fairness of with–profits pension schemes.

3.1 Introduction

The price of a traded security reacts promptly to changes in the fundamental determinants of its value. As opposed to this, in spite of fair value accounting standards, the price of entering a with–profits pension scheme is typically fixed, regardless of changes to the financial outlooks for participation.

The manifestation of this paradox is that members contribute equally to the collective bonus reserve – even when their prospects for enjoying it are vastly different. In particular, the value of the implicit, compound bonus option that comes with membership depends substantially on the (random, yet controlled) funding ratio at the time of entry. This difference is a source for *systematic* intergenerational redistribution, which may be seen as unfair.

The aim of this study is to discuss and quantify the loss of efficiency

associated with imposing bounds on the pension fund's design (imposed in order to achieve a certain degree of intergenerational fairness). Or reversely put: to analyse the loss of fairness stemming from restrictions imposed for the sake of reaching a specific level of efficiency. This trade–off should be of utmost importance to any regulator or altruistic board. In order to discuss the problem we consider a with–profits pension scheme that does take intergenerational redistribution into account, thereby constraining scheme design.

In a pension context intergenerational redistribution has – to our knowledge – been addressed mainly by Døskeland and Nordahl (2008a). Their model, however, is so vastly different from the one presented below that comparison is futile. They conclude that it is unfavourable to take part in the accumulation phase of a pension fund, and vice versa. One *particular* distinction between Døskeland and Nordahl (2008a) and the model of the present paper is that they consider overlapping generations explicitly whereas we deal with disjoint generations. Overlapping or contemporary generations can easily be studied within this paper's framework, however. Hansen and Miltersen (2002) also briefly discuss redistribution between different generations in the presence of a collective bonus reserve.

There is a rich literature – initiated by Briys and de Varenne (1994) – on the related problem of constructing contracts that are fair between owners and policyholders as a whole. That setup *could* be interpreted as imposing intergenerational fairness, albeit in a rather different way from what we have in mind. Also, none of those papers distinguish between the set of fair contracts (because they value under a unique equivalent martingale measure this would not make sense).

3.1.1 Outline

Section 3.2 introduces the underlying mathematical model. The measurement of fairness and efficiency is discussed in Section 3.3, where some optimisation criteria are subsequently suggested. These criteria are illustrated through Monte Carlo simulation in Section 3.4, while Section 3.5 considers an extension of the model, which reduces redistribution markedly. Finally, Section 3.6 provides a discussion of the preceding modelling and results, and gives concluding remarks.

3.2 Model

We consider a pension fund, which is owned by its present members. The board, which designs the scheme, represents future entrants as well, although these have no formal stakes in the scheme yet. Thus, the board can be seen as a device for solving the coordination problem that arises in any intergenerational enterprise. Such fairness motives are non-standard in the literature, but highly relevant from a practical perspective. Recently, the concept has regained popularity through the book by Akerlof and Shiller (2009).

Rather than starting from scratch the framework of Kryger (2010c) is used, but as opposed to that paper the concern is with the finite time properties of the system. The model is summarised below, and Section 3.5 introduces various extensions that were not dealt with in previous work.

The market values of the scheme's assets and liabilities at time $t \geq 0$ are denoted A_t respectively L_t, while the *funding ratio* is derived as $F \triangleq A/L$. Between reporting periods, indexed by $0, 1, 2, \ldots$, the asset value follows the controlled process

$$A_i > 0,\ dA_{i+t} = A_{i+t}\left((r + \pi_{i+t}\Lambda\sigma)\,dt + \pi_{i+t}\sigma dB_{i+t}\right),\ (i \in \mathbb{N}_0,\ t \in [0,1))\,,$$

where r is the constant risk free interest rate, $\Lambda > 0$ the constant market price of risk, $\sigma > 0$ the constant market volatility, and π the time-varying, controlled proportion of assets allocated to risky assets. B is a one-dimensional standard Brownian motion on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$ driving the financial market, which is frictionless and complete as seen from the scheme's point of view. Individuals are assumed, however, to have limited access to the financial markets. In particular, an individual's guaranteed future benefits cannot be sold or pawned.

In order to avoid insolvency it is required that funding is strictly above one at all times. Hence, we assume that $F_0 > 1$, and that the investment strategy is Constant Proportion Portfolio Insurance (CPPI), that is

$$\pi_t = \alpha\frac{A_t - L_t}{A_t},\ (t \geq 0),$$

where the so–called *multiplier*, $\alpha > 0$, is chosen by the board. Liabilities develop deterministically between reporting times:

$$L_i > 0, \quad L_{i+t} = L_i e^{rt}\ \ (i \in \mathbb{N}_0,\ t \in [0,1))\,.$$

Consequently, the funding ratio process follows the discrete time controlled Markov process

$$F_{0^+} > 1, \quad F_{i^-} = \left(F_{(i-1)^+} - 1\right) \exp(Z_i) + 1, \ (i \in \mathbb{N}), \tag{3.1}$$

where $(Z_i)_{i\in\mathbb{N}}$ is an i.i.d. sequence with $Z_1 \sim N\left(s\Lambda - s^2/2, s^2\right)$, and $s \triangleq \alpha\sigma$ is denoted "risk" as it measures the volatility of the bonus reserve, $A - L$.

At the end of each reporting period, between times i^- and i^+, benefits fall due, new contributions are paid in, and members are awarded bonus. This brings about jumps in asset and liability values, and consequently in the funding ratio. These three types of updates are as follows:

$\Pi_i \in [0, 1]$ denotes the proportion of existing liabilities paid out as benefits (e.g. expiring policies), and similarly, $\Gamma_i \geq 0$ is the amount of new contributions (e.g. new underwritings) relative to existing liabilities.[1] These contributions are converted into liabilities $g_i\Gamma_i L_{i^-}$ for some factor $g_i \in (0, (F_{i^-} + \Gamma_i - 1)/\Gamma_i)$, which is the proportion of the contribution that buys guaranteed benefits. Hence, $1 - g_i$ is the share that – implicitly - buys a compound bonus option. Finally, existing liabilities, L_{i^-}, are increased by a bonus factor $\exp(b_i) \geq 1$, which is determined by using all assets in excess of $\kappa L_{i^-}(1 + g_i\Gamma_i - \Pi_i)$, for a *bonus barrier*, $\kappa > 1$, to enhance guarantees. This barrier is also determined by the board. With the described approach one arrives at the *post bonus* funding ratio

$$\begin{aligned} F_{i^+} &= \frac{A_{i^-} + L_{i^-}(\Gamma_i - \Pi_i)}{L_{i^-}(1 + g_i\Gamma_i - \Pi_i)} \wedge \kappa \\ &= \frac{F_{i^-} + \Gamma_i - \Pi_i}{1 + g_i\Gamma_i - \Pi_i} \wedge \kappa, \ (i \in \mathbb{N}). \end{aligned} \tag{3.2}$$

The bonus, b_i, that is in fact allotted such that

$$L_{i^+} = L_{i^-}\left((1 - \Pi_i)\exp(b_i) + g_i\Gamma_i\right), \quad A_{i^+} = A_{i^-} + L_{i^-}\left(\Gamma_i - \Pi_i \exp(b_i)\right),$$

and (3.2) is satisfied, is

$$b_i = \begin{cases} \log \frac{F_{i^-} - \Gamma_i(\kappa g_i - 1)}{\kappa - \Pi_i(\kappa - 1)}, & F_{i^-} > \kappa - \Pi_i(\kappa - 1) + \Gamma_i(\kappa g_i - 1) \\ 0, & \text{otherwise} \end{cases}, \ (i \in \mathbb{N}). \tag{3.3}$$

[1] If $\Gamma_i = 0$ and $\Pi_i = 1$ the scheme closes, and this case is not taken into account below.

Note that the new contributions do not earn bonus immediately, whereas existing contracts *are* credited. As bonus is, partly, intended to pay for disposable capital, this is only natural.

In this paper we consider contracts, in which members contribute the nominal amount $\xi(t) = \exp((\eta + r)t)$ at time t, for some "net contribution inflation", η. This is converted into a guaranteed benefit at *horizon* time $n \geq t$ of $g_t \xi(t) \exp(r(n-t))$ with present value $g_t \xi(t)$ – plus a compound bonus option. The object of interest is the (to individuals) non-tradeable, discounted terminal benefit

$$\begin{aligned} X &\triangleq e^{-nr} \sum_{j=0}^{n} g_j \xi(j) e^{r(n-j)} e^{\sum_{k=j+1}^{n} b_k} \\ &= \sum_{j=0}^{n} g_j e^{\eta j} e^{\sum_{k=j+1}^{n} b_k}. \end{aligned} \tag{3.4}$$

If necessary, we will equip X with arguments (s, κ, F_{0^+}) representing the "risk", the bonus barrier, and the initial funding ratio respectively. In order to consider intergenerational redistribution we use the rule $g_i = 1$ for all i in the main part of the paper. Section 3.5 explores the consequences of applying other rules.

Actual life insurance contracts give rise to interest rate risk, which is hedgeable in competitive markets, and mortality risk. Also, benefits are typically not received as a lump sum. While none of those factors are considered X can be seen as a proxy for the value of a whole life annuity bought at market terms at time n.

Contributions are compulsory, and there is no free policy option nor any surrender option. This leaves no scope for speculation (via timing of contributions or lapses) against the scheme, i.e. the other members. In Sections 3.2.1 and 3.3 we assume that

$$\Gamma_i = \Pi_i \max\{F_{i^-}, \kappa, \} / \kappa,$$

so that net inflow is positive. This assumption is merely required to get nicer analytical expressions, and the qualitative conclusions are valid in much more flexible scenarios. In the analysis in Sections 3.4 and 3.5 this requirement is not necessary either. Administrative costs, transaction costs, taxes, etc. are disregarded throughout.

From (3.4) we observe that when g is constant it is the release of bonus that determines the outcome. Therefore, the properties of bonus are discussed next.

3.2.1 Properties of bonus

In order to analyse the scheme consider the time until next bonus, as seen from an arbitrary time $i \in \mathbb{N}$,

$$\tau(\theta; s) \triangleq \min\{j \geq 1 :\ b_{i+j} > 0 |\, F_{i^+} = (\kappa - 1)\theta + 1\},\ (\theta \in (0, 1],\ s > 0)\,,$$

where θ measures how far the funding ratio is from the bonus barrier. With this specification the choice of κ does not determine when bonus is awarded, cf. (3.1) and (3.2). The continuous version of τ is the stopping time

$$\tilde{\tau}(\theta, s) \triangleq \inf\left\{t > 0 :\ B_t \geq -\frac{\log\theta}{s} + t\,(s/2 - \Lambda)\right\},\ (\theta \in (0, 1),\ s > 0)\,.$$

Due to discrete time sampling of the funding ratio

$$\tau \geq \lceil \tilde{\tau} \rceil \geq \tilde{\tau},$$

but the approximation error is fairly small, when the barrier is "distant", or the investment strategy is cautious, i.e. θ or s is low (or if time is measured in "small" units).

From e.g. Karatzas and Shreve (2000), p. 197 and Preisel *et al.* (2010) the distribution function of $\tilde{\tau}(\theta, s)$ is

$$\Phi\left(\frac{\log\theta}{s\sqrt{t}} + \sqrt{t}(\Lambda - s/2)\right) + \theta^{\frac{s-2\Lambda}{s}}\,\Phi\left(\frac{\log\theta}{s\sqrt{t}} - \sqrt{t}(\Lambda - s/2)\right),\ (t > 0).$$

For $s \leq 2\Lambda$ this is an inverse Gaussian distribution, but otherwise $\tilde{\tau}$ is defective.

A mere focus on the time until the first bonus allotment certainly has its shortcomings, but it is a nice way of illustrating that cautious strategies (corresponding to low values of s) are unattractive on short horizons – in particular if initial funding is low. This is essentially because of the near–absence of downside risk, i.e. if initial funding is low the best one can

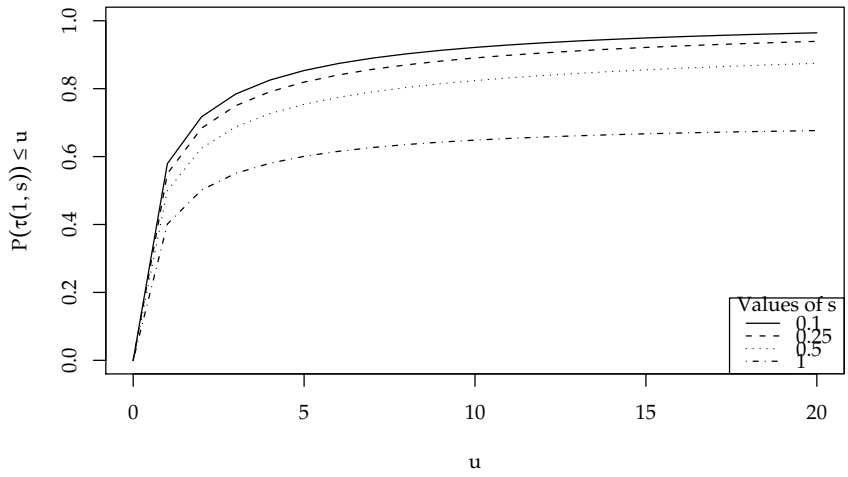

Figure 3.1: The distribution of the non-delayed regeneration time $\tau(1;s)$. Fixed parameters: $\Lambda = 0.25$.

hope for is to get a single bonus. On longer horizons cautious strategies are more attractive, precisely because of downside risk. To realise this one must study $\tau(1,s)$, the time between bonus allotments, which exhibits negative first order stochastic dominance with respect to s, i.e. smaller values of s are preferable. Its distribution can be calculated exactly via the method in Jarner and Kryger (2009), and is shown in Figure 3.1 for four different investment strategies.

A natural supplement to the properties of bonus allotment is the (one-step) *conditional bonus* with distribution

$$\begin{aligned}&\mathbb{P}\left(b_{i+1} \le y \,|F_{i^+} = (\kappa-1)\theta + 1,\ b_{i+1} > 0\right)\\ &\qquad = 1 - \frac{\Phi\left(\Lambda - \frac{s}{2} + s^{-1}\log\theta - s^{-1}\log\frac{\kappa e^y - 1}{\kappa - 1}\right)}{\Phi\left(\Lambda - \frac{s}{2} + s^{-1}\log\theta\right)},\ (y>0)\,,\end{aligned}$$

which follows almost immediately from Preisel *et al.* (2010). The conditional bonus exhibits first order stochastic dominance in s (and κ and θ),

so that higher values are preferable.

Altogether, cautious investment strategies do not give rise to much bonus on short horizons, especially if initial funding is low, whereas on longer horizons the matter is more ambiguous – but with both very cautious and very aggressive strategies inducing only little bonus. As for the barrier – when initial funding is $(\kappa - 1)\theta + 1$, higher barriers are always preferable. The board could, however, encounter a fixed initial funding, and be asked to set a barrier subsequently, which would complicate matters. This is because, for short horizons, more bonus would be given with low barriers, in particular if initial funding is low. But as the horizon increases higher barriers again become more attractive.

For both design parameters one should also have in mind that in case of several contributions ($\eta > -\infty$) the final wealth distribution depends more on later bonuses than on early ones, cf. (3.4). Therefore, the long-run properties are more important than this discussion perhaps suggests.

3.3 Fairness and efficiency

In this section we discuss the measurement of fairness and efficiency. Subsequently, some tangible measures of these two vague notions are introduced.

3.3.1 Preliminary reflections

Since we consider a compulsory scheme the design should satisfy that almost every outcome is acceptable for almost every one. Hence, it is required that – on the vast majority of paths and for almost all types of members – the degree of redistribution between different generations is low, while the outcome must at the same time be satisfactory for almost everyone. We make the precise materialisation of these ideas clear shortly.

To further approach a decision rule we rely on the *original position* of moral philosophy (see Rawls (1971)), which states that any design agreed upon by agents whose identity is unknown to themselves during the bargaining is fair.

It is clear that not all generations can get the same outcome. And as hinted above it is probably not desirable to follow a very cautious investment strategy with the aim of approximating such equality. The ques-

tion is then, how the board should evaluate the inherent redistribution against a possible advantage from investing more aggressively. Although such trade–offs are *acknowledged* in most economic analysis, they are typically disregarded.

In order to pick fairness and efficiency measures we consider a hypothetical bargaining between two members entering the scheme at funding ratios κ and $f \in (1, \kappa]$ respectively, but with the caveat that they must design the system without knowing who enters at which funding. f may depend on the design parameters (s, κ), as will be explained below, but the notation f is used as a shorthand nevertheless, since the meaning will always be clear from the context. The generations represented by the two members are taken to experience identical institutional conditions during their membership periods, but their financial markets are assumed to be governed by independent versions of B, and f is taken to be independent of those. This means that a proper, although not imperative, interpretation of the setup is that the member with funding f enters first, and then after at least n years the other member enters at a time with full funding.

As for the actual measurement, the initial *dogma* of this section and the idea of the original position guides us. Efficiency of the outcome should be associated with the distribution of the terminal benefit, which must overcome some minimum target with a high probability. Fairness ought to be related to the ratio of the terminal benefits, which we want to be close to one with a high probability. Non-overlapping generations are compared, and the ratio of the benefits of two such disjoint generations has a wider distribution than in the case of overlapping generations. Thus, in this respect the discussion in this paper is "worst case" in terms of intergenerational subsidisation.

The two most widely applied measures for evaluating a monetary random variable is measuring its arbitrage-free value or its expected utility. We discard the former approach because of the assumed non-tradeability of the guarantee, and we discard the latter because we prefer to ensure attractive outcomes with a high probability.

3.3.2 Two formulations of the problem

Regarding the choice of initial fundings to compare, it is uncontroversial to use κ as the higher level. As for the lower value one may choose to consider

a constant. In this case it is meaningful to compare different barriers, for there is a trade–off between high and low barriers – as discussed above.

Alternatively, we could let $f = 1 + (\kappa - 1)\theta$ for some $\theta \in (0, 1)$ as in Section 3.2.1, in which case higher barriers are more attractive for both parties, so that it is not possible to optimise over κ. Still, to properly differentiate between candidates for the optimal investment strategy we let θ depend on s, since a high distance from the barrier is more likely with more aggressive strategies. To this end, fix an $\epsilon \in (0, 1)$, and choose $\theta(s)$ such that $\mathbb{P}\left(1 + (\kappa - 1)\theta(s)\right) = \epsilon$ in stationarity. From Preisel *et al.* (2010) we then get

$$\theta(s) = \left(\frac{\epsilon}{1 - s\rho/\sqrt{2}}\right)^{\rho^{-1}}, \ (0 < s < 2\Lambda),$$

where ρ is the unique non-zero solution to

$$1 - \rho^2 s^2/2 = \exp\left(-\rho s(\Lambda - s/2)\right), \ (0 < s < 2\Lambda).$$

For $s \geq 2\Lambda$ no stationary distribution exists, therefore we truncate θ at the value corresponding to the somewhat arbitrary $s = 1.99\Lambda$.

The former setting with a fixed f corresponds to an existing scheme encountering a (low) funding ratio, and wishing to design a fair and efficient scheme going forward. On the other hand, the latter formulation, where f depends on s and κ, covers the case of a new scheme with all the good intentions at the outset, but with an exogenously fixed κ. We refer to the two settings as "case A" and "case B" respectively.

3.3.3 Measuring fairness and efficiency

This section presents our choices for measuring fairness and efficiency. Subsequently the two measures are combined in order to form two different constrained optimisation problems.

Fairness

To measure intergenerational fairness we focus on a threshold for the ratio of the respective terminal benefits:

$$\mathbb{P}\left(\frac{X\left(s,\kappa,f\right)}{X\left(s,\kappa,\kappa\right)}>1-\delta\right), \tag{3.5}$$

where $0 \leq \delta < 1$ measures the maximum permitted redistribution (up to some probability). If the generations were contemporary the ratio in question would be bounded by one. Although studying disjoint generations we use the measure nevertheless, and thus disregard the extent to which the ratio exceeds one.

Efficiency

The quantification of efficiency follows similar lines as above. As previously hinted, a target is needed to calculate efficiency. To this end expected power utility is used as a measure of satisfaction. Since the target will only be used as an auxiliary we prefer this simple approach, because it is easy to communicate, and requires one parameter only. The certainty equivalent of a positive random variable, Y, is then

$$CE\left(\gamma;Y\right) \triangleq \begin{cases} \mathbb{E}\left\{Y^{1-\gamma}\right\}^{\frac{1}{1-\gamma}}, & \gamma \in [0,\infty)\backslash\{1\} \\ \exp\left(\mathbb{E}\left\{\log Y\right\}\right), & \gamma = 1 \end{cases}, \tag{3.6}$$

where $\gamma \geq 0$ denotes the coefficient of relative risk aversion. The suggested efficiency measure for a generation with funding $f < \kappa$ is

$$\mathbb{P}\left(\frac{X(s,\kappa,f)}{\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE\left(\gamma;X(\bar{s},\bar{\kappa},f)\right)}>1-\beta\right). \tag{3.7}$$

For the two generations efficiency is measured by merely adding their respective terminal benefits to a single random variable. Due to linearity of the certainty equivalent, $0 \leq \beta < 1$, can be interpreted as the maximum permitted relative *cost* of obtaining fairness (up to some probability). $\mathcal{S}\times\mathcal{K} \subseteq (0,\infty)\times(1,\infty)$ is a range of considered design variables.

Constrained optimisation

Combining the criteria (3.5) and (3.7) produces two different constrained optimisation problems. First, if we maximise efficiency subject to a fairness side condition we get:

$$\max_{s,\kappa\in\mathcal{S}\times\mathcal{K}} \mathbb{P}\left(\frac{X(s,\kappa,\kappa)+X(s,\kappa,f)}{\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE(\gamma;X(\bar{s},\bar{\kappa},\bar{\kappa})+X(\bar{s},\bar{\kappa},f))}>1-\beta\right) \tag{3.8a}$$

subject to

$$\mathbb{P}\left(\frac{X(s,\kappa,f)}{X(s,\kappa,\kappa)}>1-\delta\right)\geq p, \tag{3.8b}$$

We do not consider values of p for which a Pareto improvement is possible, i.e. for which both parties can be made better off in terms of (3.7). A candidate (s,κ) is thus excluded if:
$\exists(\tilde{s},\tilde{\kappa})\in(\mathcal{S}\times\mathcal{K})\backslash\{(s,\kappa)\}$: (3.8c) and (3.8d) are both satisfied, and at least one of them with strict inequality.

$$\frac{\mathbb{P}(X(s,\kappa,f)>(1-\beta)\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE(\gamma;X(\bar{s},\bar{\kappa},f)))}{\mathbb{P}(X(\tilde{s},\tilde{\kappa},f)>(1-\beta)\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE(\gamma;X(\bar{s},\bar{\kappa},f)))}\leq 1 \tag{3.8c}$$

$$\frac{\mathbb{P}(X(s,\kappa,\kappa)>(1-\beta)\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE(\gamma;X(\bar{s},\bar{\kappa},\bar{\kappa})))}{\mathbb{P}(X(\tilde{s},\tilde{\kappa},\kappa)>(1-\beta)\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE(\gamma;X(\bar{s},\bar{\kappa},\bar{\kappa})))}\leq 1. \tag{3.8d}$$

The reverse constrained optimisation is the one where a fairness criterion is maximised subject to an efficiency threshold condition:

$$\max_{s,\kappa\in\mathcal{S}\times\mathcal{K}} \mathbb{P}\left(\frac{X(s,\kappa,f)}{X(s,\kappa,\kappa)}>1-\delta\right). \tag{3.9a}$$

subject to

$$\mathbb{P}\left(\frac{X(s,\kappa,\kappa)+X(s,\kappa,f)}{\max_{\bar{s},\bar{\kappa}\in\mathcal{S}\times\mathcal{K}}\ CE(\gamma;X(\bar{s},\bar{\kappa},\bar{\kappa})+X(\bar{s},\bar{\kappa},f))}>1-\beta\right)\geq p. \tag{3.9b}$$

In (3.9) the inclusion of (3.8c)-(3.8d) is not imperative, nor meaningful. In (3.8b) and (3.9b) $p\in[0,1]$ is a minimum acceptance probability decided along with β and δ. Before proceeding we warn that for some parameterisations one may end up with probabilities of zero or one, in which case the clever approach is to re-parameterise, unless it was intentional.

Next, we illustrate the suggested criteria through simulations.

3.4 Simulation-based illustrations

In this section we mainly analyse a "mature" fund with equal in- and outflow of $\Pi = \Gamma = 0.02$, net contribution inflation $\eta = 0.02$, a horizon of $n = 50$, and market price of risk, $\Lambda = 0.25$. Following the analysis of the base case each of the main parameters (Π, Γ, η, n, and Λ) are changed, and the derived consequences are briefly discussed. Also, the auxiliary parameters are fixed at $\beta = 0.05$, $\delta = 0.05$, $\gamma = 0.5$, $f = 1.02$ (case A), $\epsilon = 0.05$ (case B), and $\kappa = 1.3$ (case B), but a sensitivity analysis is conducted in Section 3.4.3. Finally, Section 3.4.4 reviews the simulation details.

3.4.1 Base case

Figures 3.2 and 3.3 show the trade–off between fairness and efficiency. In the former graph the bonus barrier and the investment strategy are both to be optimised over (case A), whereas the latter considers a pre-specified bonus barrier (case B). The results are qualitatively in line with the predictions of Section 3.2.1.

In case A, higher barriers yield less fairness (because generations are more different), but more efficiency. Also, the optimal strategies associated with higher barrier are more cautious. If one uses the maximum-efficiency criterion (3.8), a very narrow range of (for this parametrisation) modestly aggressive strategies are non-dominated. The most cautious as well as the most aggressive investment strategies are excluded by Pareto inoptimality, while others are merely dominated. When the maximum-fairness criterion (3.9) is imposed instead, all investment strategies *above* some threshold are candidates for optimum, because redistribution is less for more aggressive strategies (bonus becomes rare).

Conversely, in case B, the heritage to future generations is implicitly considered (through the long-run funding level), which leads to less aggressive strategies being favourable. With a main focus on efficiency, through criterion (3.8), a range of rather cautious investment strategies are non-dominated, as in case A, though these values of s are generally much lower in case B. If fairness is emphasised instead, all strategies *below* a certain threshold are potentially optimal, which highlights the difference between the two cases. In case B, even modestly aggressive investment strategies

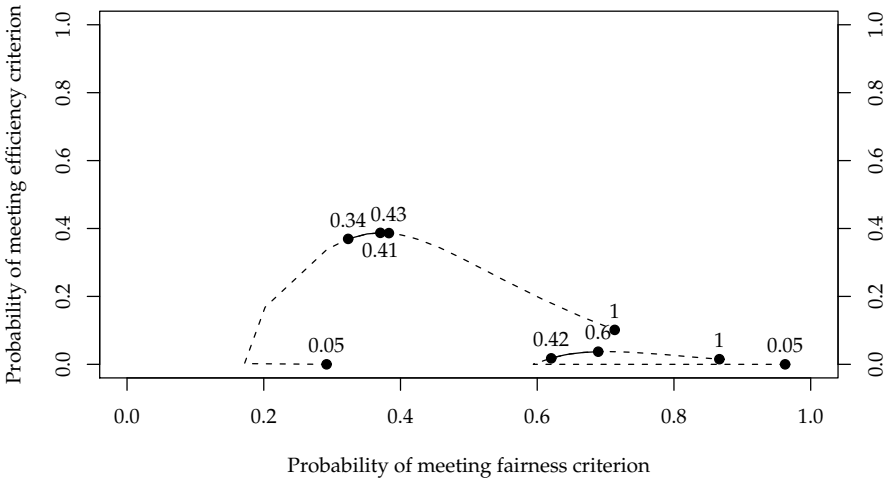

Figure 3.2: Case A: Trade-off between efficiency (3.7) and fairness (3.5) at different values of s indicated in the diagram. The left curve corresponds to $\kappa = 1.5$, and the right curve represents $\kappa = 1.1$. The dashed parts of the graphs correspond to strategies that were discarded due to (3.8c)-(3.8d). Fixed parameters: $g = 1$, $\Gamma = \Pi = 0.02$, $\eta = 0.02$, $n = 50$, $\Lambda = 0.25$, $\gamma = 0.5$, $\beta = \delta = 0.05$, $\mathcal{S} \subseteq (0, 1]$, $\mathcal{K} = \{1.1, 1.5\}$, and $f = 1.02$.

are penalised severely by a higher initial distance from the barrier – in turn making them poor candidates for yielding sufficient efficiency.

Altogether, the results show that it is hard to obtain outcomes that are adequately efficient and fair to a high degree at the same time. One way of overcoming this is to alter the way in which contributions are transformed into benefits, which is the topic for Section 3.5.

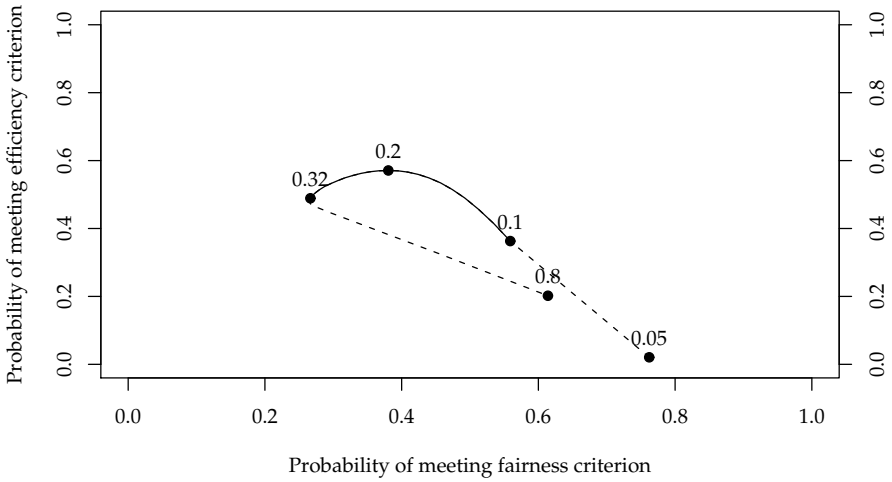

Figure 3.3: Case B: Trade-off between efficiency (3.7) and fairness (3.5) at different values of s indicated in the diagram. The dashed parts of the graph corresponds to strategies that were discarded due to (3.8c)-(3.8d). Fixed parameters: $\kappa = 1.3$, $g = 1$, $\Gamma = \Pi = 0.02$, $\eta = 0.02$, $n = 50$, $\Lambda = 0.25$, $\gamma = 0.5$, $\beta = \delta = 0.05$, $\mathcal{S} \subseteq (0, 1]$, $\mathcal{K} = \{\kappa\}$, and $\epsilon = 0.05$.

3.4.2 Alternative environments

Demography

When the net outflow is positive ($\Pi > \Gamma$) bonuses are higher and more frequent than otherwise. Oppositely, of course, in a fund that is in the process of building up its balance. The aggregate effect on fairness and efficiency is not clear at the outset. The situation with $\Pi = 0.1$, and $\Gamma = 0.02$ is shown in Figure 3.4, which demonstrates that intergenerational subsidisation is slightly less in such a non-accumulating scheme, without harming efficiency. In general, the trade–off governing the design decision is very similar to the base case in Figures 3.2 and 3.3.

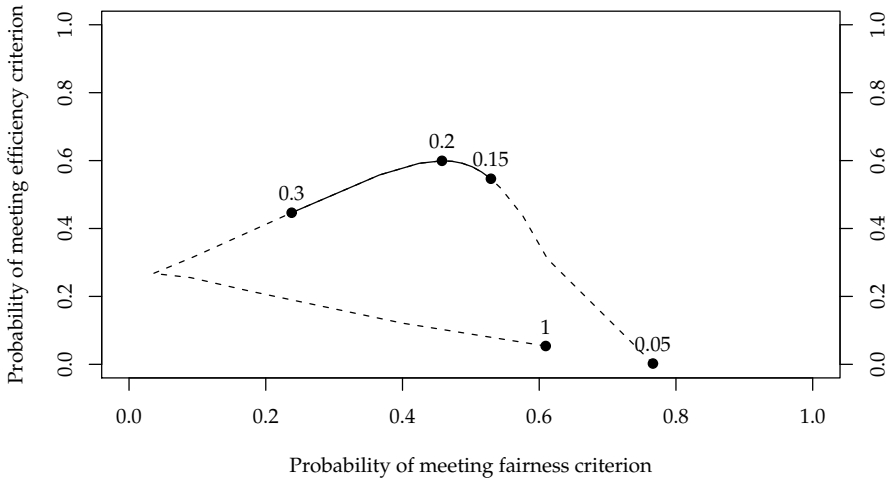

Figure 3.4: Case B: Trade-off between efficiency (3.7) and fairness (3.5) at different values of s indicated in the diagram. The dashed parts of the graph corresponds to strategies that were discarded due to (3.8c)-(3.8d). Fixed parameters: $\kappa = 1.3$, $g = 1$, $\Gamma = 0.02$, $\Pi = 0.1$, $\eta = 0.02$, $n = 50$, $\Lambda = 0.25$, $\gamma = 0.5$, $\beta = \delta = 0.05$, $\mathcal{S} \subseteq (0, 1]$, $\mathcal{K} = \{\kappa\}$, and $\epsilon = 0.05$.

Economy

As net contribution inflation, η, is increased, a higher degree of fairness is obtained, because – as previously mentioned – later bonus matters more, and later bonus is more likely to be the same for both generations. Also, all strategies become more efficient because the terminal distribution is narrower, i.e. less dependent on bonus.

When the market price, Λ, increases, bonuses are larger and more frequent. More aggressive investment strategies are, of course, preferable.

The qualitative *and quantitative* conclusions from the base case are surprisingly insensitive to changes to η and Λ. The only noteworthy effect is that higher contribution inflation implies slightly more cautious investment in case A because of the amplified importance of later bonuses.

Horizon

Extending the horizon, n, makes the system fairer, due to the longer period with identically distributed bonuses, $(n-\tau)^+$. The shapes of the curves in Figures 3.2 and 3.3 are essentially unaltered, however. Only, in case A, slightly more cautious investment strategies are preferred on longer horizons, as the long-run properties become more important. This point emphasises that when sampling from a *fixed* (not affected by the controls) initial distribution (in this case a Dirac distribution), and the terminal conditions do not matter there is a need for someone, be it the board or the regulator, to require long-run stability, for otherwise it is tempting to gradually – or swiftly – exhaust the bonus reserve to the disadvantage of future generations.

3.4.3 Sensitivity

The choice of β, γ, δ, in case A: f, and in case B: ϵ and κ matters little as far as the qualitative conclusion goes.

As for the former two, the reason that the parameters play minor roles is that the benefit distribution is quite narrow (on most trajectories bonus is small compared to contributions), and bounded (far) away from zero. Of course, the higher the values of β and γ, the higher the efficiency. Therefore it is instructive to use a rather low γ-value, since this ensures that the efficiency probabilities are not too high for any agent. The choice of δ affects the *level* of fairness profoundly – but the *shapes* of the curves in Figures 3.2 and 3.3 are unaltered.

In case A higher f implies more fairness and a shift towards slightly more cautious investment. The tail probability, ϵ, was introduced to allow a design-dependent distinction between the two generations, and the extent to which this differentiation is carried out does not affect the results much as long as ϵ is kept reasonably small.

Also, by construction, the choice of κ, in case B, only influences the conditional bonus, and only so to an extent which hardly affects the optimal choice of s.

3.4.4 Simulation details

The simulations were done in the freeware statistical computing package `R`. In all cases 100,000 paths were simulated. The seed was set manually to allow all results to be reproduced, and reused across experiments. It is particularly important to ensure that disjoint generations experience independent market innovations. Oppositely, if contemporary generations were considered, it would be equally important that they sampled the same financial market. Distribution functions are estimated by their sample counterparts.

3.5 Extensions

The previous section showed that although some designs are superior to others it is not possible to obtain high levels of fairness and efficiency at the same time, when the entire contribution is transformed into guaranteed future benefits, i.e. $g = 1$. It is straightforward, however, to design rules that take the differing conditions into account, and hence – partly – overcome systematic intergenerational redistribution. Below we present two such rules.

3.5.1 A solidary rule

When a policy expires its accrued guarantees (including bonus) are paid out, but the free reserve stays in the fund. Therefore, the funding ratio increases as a result of an expiry. This gain is split between existing members and new contributions according to some rule. In the standard case, $g = 1$, new money always benefits from entering in the sense that the value of their guarantee can be approximated by $g_i F_{i^+}$, which is greater than one, but highly dependent on the random timing of entry

Instead, the way in which contributions are transformed into guaranteed future benefits *may* be based on the solidary point of view that $g_i F_{i^+}$ should be the same for all generations, that is

$$g_i \left(\frac{F_{i^-} + \Gamma_i - \Pi_i}{1 + g_i \Gamma_i - \Pi_i} \wedge \kappa \right) = C_i, \tag{3.10}$$

for some positive $C_i < (1 + \Gamma_i - \Pi_i)/\Gamma_i$, which - for the natural choice $C_i = 1$ – is solved by the *solidary rule*

$$g_i = \frac{1 - \Pi_i}{F_{i^-} - \Pi_i} \vee \kappa^{-1}, \ (i \geq 1) \tag{3.11a}$$

$$g_0 = \frac{1}{F_{0^+}} \tag{3.11b}$$

In general C_i could depend on e.g. demographics forecasts and time to expiry.

Because $g < 1$ the funding ratio gets a boost upwards, so that bonuses are larger and more frequent (in return for lower guarantees). As in the case of positive net inflow the combined effect on fairness is ambiguous, depending on the horizon, among others. For the base case it turns out that fairness is enhanced slightly. The fairness measure (3.5) is ill–suited, however, since it focuses exclusively on one–sided deviations. As a matter of fact the generations can be made approximately equally well off when using the solidary rule (in the sense that the density of the ratio between the benefits is much more balanced around 1), which is a major advance over the results obtained with $g = 1$.

3.5.2 An indemnifying rule

As another example we present the *indemnifying rule*

$$\frac{F_{i^-} + \Gamma_i - \Pi_i}{1 + g_i\Gamma_i - \Pi_i} \wedge \kappa = F_{i^-} \wedge \kappa,$$

which gives staying members the same funding ratio regardless of the amount of new entrants and exits. This rule yields

$$g_i\Gamma_i = \Pi_i + \frac{\Gamma_i - \Pi_i}{F_{i^-}}, \ (i \geq 1) \tag{3.12a}$$

$$g_0\Gamma_0 = \Pi_0 + \frac{\Gamma_0 - \Pi_0}{F_{0^-}}, \tag{3.12b}$$

the latter assuming F_{0^-} is known. The indemnifying rule enhances fairness still more the higher the ratio of inflow to outflow, precisely because an accumulating scheme releases relatively little bonus reserve, and thus the

higher the pre-bonus reserve the less the new entrants will receive (in terms of g). The *neutralising* effect of the solidary rule is only achieved if there is no outflow, however – in which case the two rules are almost identical.

3.6 Discussion and conclusion

This section discusses the insights gained from the paper's model. Also, limitations and possible alternatives approaches are described. Policy implications are touched upon, and finally, concluding remarks are given.

One of the main lessons that can be derived from the paper stems from the vast difference between cases A and B. In the former situation both investment strategy and bonus barrier are optimisation variables, and the trade–off is that lower barriers induce less efficient systems that are fairer; and the same effect follows from using very cautious or very aggressive investment strategies. Conversely, in case B higher barriers are always preferable.

If efficiency is the maximisation object and fairness the side condition there is a narrow range of non-dominated strategies in either case (though those ranges are substantially different the cases between). But if fairness is maximised subject to an efficiency constraint the two cases differ more markedly. In case A cautious strategies are dominated, but in case B aggressive strategies are dominated because they imply low initial funding and thus low fairness. These qualitative conclusions are stable under different parameterisations, but are likely sensitive to different *formulations* of the objective.

Another important insight comes from realising how redistribution can be lessened substantially at no cost – by introducing new ways of transforming contributions into guaranteed future benefits.

A third outcome is the ability to exclude certain investment strategies based on dominance arguments.

3.6.1 Limitations and alternatives

Obviously the evaluation of intergenerational fairness is much broader than what can be covered here. For instance, one could argue that by facing identical *rules* generations are treated fairly in that the economic condi-

tions they face are not *explicitly* controlled by other generations. Also, even within the present paradigm, redistribution and efficiency could be measured quite differently.

The alternative rules that were presented do not aim at converting each contribution into benefits in a fair manner. That is, the *value* of the compound bonus option does not equal its implicit *price*, as the former depends on time to maturity and forecasts of demographics etc. Instead, it is assumed that all members have identical contribution plans, so there is no heterogeneity nor any free policy or surrender options.

Default is precluded by construction in the present setting. To overcome this weakness one could allow for default by fixing the portfolio only at the beginning of each period. This would mimic real life investment behaviour more closely than the often employed constant allocation to stocks, while still allowing for bankruptcy. Within the realm of no–default one could change the asset allocation in some non-linear way, while maintaining $\lim_{F_t \to 1^+} \pi(F_t) = 0^+$, e.g. through an Option Based Portfolio Insurance (OBPI) strategy.

Another way of introducing default is to allow for non-marketed shocks to the value of liabilities – interpreted as unanticipated changes in mortality, statute, or the like. Such jumps could occur periodically or at random points in time.

Instead of distributing all excess funding as bonus, some authors argue in favour of smoothing bonus allotment over time precisely with the aim of reducing the effect of random entry time funding levels. This would reduce subsidisation slightly. Another widespread alternative consists of basing the bonus on the past year's financial performance exclusively, which reduces intergenerational redistribution, but enhances solvency problems (if the members do not participate in the downside).

Finally, as previously mentioned one could consider perfectly contemporary generations sampling the same market. Then the interpretation of fairness would be somewhat different, namely related to joining schemes with different initial funding, but which are operated identically. This appears somewhat less interesting from a designer's point of view, but can be very useful in other settings.

3.6.2 Policy implications

A regulator overseeing scheme design, or an altruistic designer should discuss the weighing between short– and long–term objectives as well as the trade–off between fairness and efficiency. In order to use the analysis in the framework laid out here to make an informed decision they must also choose whether case A or B is more appropriate for their purpose.

The most important recommendation stems from noting how fairness can be enhanced greatly at no cost by following the suggestions in Section 3.5.

3.6.3 Concluding remarks

This paper discusses the trade–off between fair and efficient design of with–profits pension schemes. More specifically, strategies for investment and bonus allotment are treated. As in many other cases in social science an important, but often neglected, feature of the problem is the crucial choice of measure for the intangible quantities fairness and efficiency. We have suggested a set of criteria and sketched the characteristics of an optimal design in two situations. First, one where only the present generation is considered, and second the case where the long–term properties (i.e. the heritage to future generations) are implicitly taken into account. It turns out that the optima are very different – quite precisely representing the different approaches. Finally, as a consequence of the somewhat dismal results of that analysis, different ways of converting contributions into guarantees are suggested and shown to yield a substantial improvement.

4. Modelling adult mortality in small populations: The SAINT model

Background. The paper in this chapter is written jointly with Søren Fiig Jarner. It appeared as Jarner and Kryger (2008), and in a more lightweight edition as Kryger (2010b). The paper was presented at a mortality course for the Danish Actuarial Society, December 2008, at a Department Seminar at Dept. of Statistics, London School of Economics, December 2008, and at the 13$^{\text{th}}$ IME Congress in Istanbul, May 2009. It has also been presented on a number of occasions by Søren Fiig Jarner. Since 2007 it has formed part of the syllabus for the annual mortality risk training course by Life & Pensions.

Abstract. The mortality evolution of small populations often exhibits substantial variability and irregular improvement patterns making it hard to identify underlying trends and produce plausible projections. We propose a methodology for robust forecasting based on the existence of a larger reference population sharing the same long–term trend as the population of interest. The reference population is used to estimate the parameters in a frailty model for the underlying intensity surface. A multivariate time series model describing the deviations of the small population mortality from the underlying mortality is then fitted and forecasted. Coherent long–term forecasts are ensured by the underlying frailty model while the size and variability of short- to medium–term deviations are quantified by the time series model. The frailty model is particularly well suited to describe the changing improvement patterns in old age mortality. We apply the method to Danish mortality data with a pooled international data set as reference population.

4.1 Introduction

Mortality projections are of great importance for public financing decisions, health care planning and the pension industry. A large number of forecasts are being produced on a regular basis by government agencies and pension funds for various populations of interest. In many situations the population of interest is quite small, e.g. the population of a small region or the members of a specific pension scheme, and historic data shows substantial variability and irregular patterns. Also, historic data may be available only for a relatively short period of time.

The prevailing methodology for making mortality projections is the method proposed by Lee and Carter (1992). The model describes the evolution in age-specific death rates (ASDRs) by a single time-varying index together with age-specific responses to the index. The structure implies that all ASDRs move up and down together, although not by the same amounts. The method has gained widespread popularity due to its simplicity and ease of interpretation and there has been a wealth of applications, see e.g. Tuljapurkar *et al.* (2000); Booth *et al.* (2006) and references therein. A number of extensions and improvements have been proposed, e.g. Brouhns *et al.* (2002); Lee and Miller (2001); Renshaw and Haberman (2006); Renshaw and Haberman (2003); de Jong and Tickle (2006); Currie *et al.* (2004); Cairns *et al.* (2006), but the original Lee–Carter method still serves as the point of reference.

4.1.1 Small population mortality

The structure of the Lee–Carter model makes it well suited to extrapolate regular patterns with constant improvement rates over time. However, while the mortality experience of large populations often conforms with this pattern the mortality evolution of small populations is generally much more irregular. Lack of fit of the Lee–Carter model for small populations, including the Nordic countries, was reported by Booth *et al.* (2006) in a comparative study. In Denmark, for instance, improvement rates have varied considerably over time within age groups and there has been periods with improvements in some age groups and stagnation or even slight increases in other age groups violating the Lee–Carter assumptions; see Jarner *et al.* (2008); Andreev (2002) for detailed accounts of the evolution

of Danish mortality.

The characteristics of small population mortality makes forecasting based on past trends problematic and very sensitive to the fitting period. Naive extrapolation of historic trends in ASDRs is likely to lead to implausible projections and unrealistic age-profiles, e.g. old age mortality dropping below that of younger ages. Often, however, the population under study can be regarded as a subpopulation of a (larger) reference population obeying a more regular pattern of improvement, e.g. a region within a country, or a small country in a larger geographical area etc. Furthermore, it will often be reasonable to assume that the study and reference populations will share the same long–term trend.

In this paper we propose a methodology for robust small population mortality projection based on the identification of a reference population. The method consists of two steps: First, the reference population is used to estimate a parametric, underlying intensity surface which determines the long-term trend. Second, a multivariate, stationary time series model describing the deviations of the small population mortality from the trend is fitted. Projections are obtained by combining extrapolations of the parametric surface with forecasts of the time series model for deviations. Coherent long–term mortality profiles are guaranteed by the parametric surface while the purpose of the time series is to quantify the short– to medium–term variability in improvement rates of the small population.

It is common to base mortality forecasts on time series models. Typically, parameters describing the evolution of period life tables are estimated assuming either a parametric or non-parametric age-profile. A time series model, often a random-walk with drift, is then fitted and forecasted for each of the parameters. The role of the time series in these methods is to capture both the trend in parameters and their uncertain evolvement around the trend. The structure implies that large short–term variability invariably lead to even larger long–term variability. In contrast, the proposed method by treating deviations from the trend as stationary allows for substantial short–term variability without inflating the long–term uncertainty.

4.1.2 Old age mortality

The modelling of old age mortality presents perhaps the most challenging part of mortality modelling. The historic development in Danish old age mortality, say age 70 and above, shows very modest improvement rates, far below those observed for the younger ages. However, over the last decade or so improvement rates have picked up and currently 70–year–olds experience improvement rates equalling those of younger ages. The picture is the same in most developed countries: old age mortality has historically improved at a slower pace than young and middle age mortality, but recently improvements rates have gradually risen.

The fact that old age improvement rates increase over time cause forecasts based on the Lee–Carter methodology to systematically underpredict the gains in old age mortality, cf. Lee and Miller (2001). As a consequence it has been recommended to use a shorter fitting period over which data conforms better with the assumptions of time-invariant improvement rates, see e.g. Lee and Miller (2001); Tuljapurkar *et al.* (2000); Booth *et al.* (2002). Although this approach clearly forecasts higher improvement rates in old age mortality it seems somewhat ad-hoc and not entirely satisfying. In this paper we take a different route.

Our ambition is to derive a simple model for the population dynamics in which changing mortality patterns naturally arise. This will allow us to characterise how mortality improvements change over time and to make predictions for future improvements in old and oldest-old mortality. Inspired by frailty theory we assume that the population consists of a heterogeneous group of individuals with varying degrees of frailty. Frail individuals have a tendency to die first causing a concentration of robust individuals at high ages. Taken this selection mechanism into account and assuming continued improvements over time a model for the entire intensity surface of the population over time can be derived. We will use the resulting parametric surface to describe the development of the reference population.

The frailty model offers an explanation to the observed lack of improvement in old age mortality. The mortality for a given age group at a given time is influenced by two factors: the current level of health and the average frailty. Over time the level of health improves but so does the average frailty. In effect, as mortality improves the selection mechanism

to reach a given age weakens causing healthier but more frail individuals to become old. In the transition from high to low selection the two effects partly offset each other such that the aggregate mortality for the age group is almost constant. Eventually the health effect will dominate the selection effect and improvements will be seen. We will explore these effects in some detail and derive the asymptotic improvement pattern implied by the model.

For two reasons we focus on adult mortality only in our modelling, i.e. mortality for age 20 and above. First, the nature of infant and child mortality is rather different from adult mortality and more complexity will have to be added to the model to fit the historic evolution adequately. Second, current levels of infant and child mortality are so low that their future course has very little impact on life expectancy and other aggregate measures. Hence, from a forecasting perspective not much is gained by the added complexity. In fact, with current mortality levels already very low up to age 60, say, future life expectancy gains will be driven almost exclusively by the development in old age mortality. However, by modelling the full adult mortality surface we are able to extract information about the general nature of improvement patterns and to predict when improvements will start to occur for age groups where none have been seen historically.

4.1.3 Outline

The rest of the paper is organised as follows: in Section 4.2.1 we present the proposed methodology for small population mortality modelling consisting of a separation of trends and deviations; in Section 4.2.2 we derive a parametric, frailty model for the underlying intensity surface and study some implications and asymptotic properties; and Section 4.2.3 contains a description of the time series model for deviations. In Section 4.3 we give an application to Danish data taken a large international data set as reference population, Section 4.4 contains a study of the fit and forecasting performance of the model, and in Section 4.5 we offer some concluding remarks and indicate further lines of research. All proofs are in Section 4.6.2.

4.2 The model

4.2.1 Methodology

In the following we suggest a methodology for robust forecasting of small population mortality. The evolution of small population mortality is characterised by being more volatile and having less regular improvement patterns compared to what is observed in larger populations. These features make simple projection methodologies very sensitive to the choice of fitting period and lead to very uncertain long–term forecasts. The fundamental idea in the proposed method is to use a large population to estimate the underlying long–term trend and use the small population to estimate the deviations from the trend.

We distinguish between unsystematic and systematic variability. Unsystematic variability refers to the variability associated with the randomness of the time of deaths in a population with a known mortality intensity, while systematic variability refers to the variability of the mortality intensity itself. Since the populations we are interested in are small by assumption we expect noticeable unsystematic variability. For instance, we do not expect crude (i.e. unsmoothed) death rates, constructed from the mortality experience of a single year, to be strictly increasing with age although we believe it to hold for the underlying intensity (at least from some age).

However, even taken the larger unsystematic variability into account it appears that small populations also have a greater *systematic* variability than larger populations. Presumably small populations are more homogeneous and thereby more influenced by specific effects. There are a number of reasons why this might be. Consider for instance the members of an occupational pension scheme. The members have the same or similar education and job, and probably also to some extent similar economic status and life style compared to the population at large. Similarly, the population in a specific country is influenced by common factors, e.g. the health care system and social habits such as smoking. The homogeneity implies that specific changes in e.g. socioeconomic conditions will have a greater impact on the mortality in a small population compared to a large population with greater diversity in background factors.

We will assume that the population under study can be regarded as a

subpopulation of a larger population, e.g. the population of a province is a subpopulation of the national population, and a national population can be regarded a subpopulation of the total population of a larger geographical region, or of similarly developed countries. We will refer to the small population as the *subpopulation* and the large population as the *reference* population. Although both unsystematic and systematic variability will be greater in the subpopulation it is reasonable to assume that the sub- and reference populations will share the same long–term trends in mortality decline, even in the presence of substantial deviations in current mortality levels. The alternative is diverging levels of mortality which in the long run seems highly unlikely for related populations. Wilson (2001); Wilmoth (1998) provide evidence for convergence in global mortality levels due to convergence of social and economic factors.

Data

We will assume that data consists of death counts, $\{D(t,x)\}$, and corresponding exposures, $\{E(t,x)\}$, for a range of years t and ages x. Data is assumed to be available for both the sub- and reference population (distinguished by subscript sub and ref, respectively), but not necessarily for the same ranges of years and ages. Data will typically also be gender specific, but it does not have to be. Since it is of no importance for the formulation of the model we will suppress a potential gender dependence in the notation.

$D(t,x)$ denotes the number of deaths occurring in calender year t among people aged $[x, x+1)$, and $E(t,x)$ denotes the total number of years lived during calender year t by people of age $[x, x+1)$. For readers familiar with the Lexis diagram, $D(t,x)$ counts the number of deaths in the square $[t, t+1) \times [x, x+1)$ of the Lexis diagram and $E(t,x)$ gives the corresponding exposure, i.e. we work with so–called *A-groups*.

Model structure

From the death counts and exposures we can form the (crude) death rates

$$m(t,x) = \frac{D(t,x)}{E(t,x)}, \tag{4.1}$$

which are estimates of the underlying intensity, or force of mortality, $\bar{\mu}(t,x)$.[1]

In order to proceed we assume that we have a family of intensity surfaces $H_\theta(t,x)$ parameterised by θ, and we consider the model where death counts are independent with

$$D_{\text{ref}}(t,x) \sim \text{Poisson}\left(\bar{\mu}_{\text{ref}}(t,x)E_{\text{ref}}(t,x)\right), \tag{4.2}$$

and $\bar{\mu}_{\text{ref}}(t,x) = H_\theta(t,x)$. Based on this model we find the maximum likelihood estimate (MLE) for θ, denoted by $\hat{\theta}$. In principle we could use a Lee–Carter specification of $\bar{\mu}$ in which case the model is the one proposed in Brouhns *et al.* (2002). However, by assumption the evolution of the reference population is smooth which allows us to get a good fit with a more parsimonious specification. In Section 4.2.2 we will develop a specific family of intensity surfaces, which will be shown to provide a very good fit to the reference data in our application. The use of a parametric model also offers insights into the dynamics of improvement rates over time.

The next step is to model the deviations of the subpopulation from the reference population. We will refer to the deviations as the spread and we propose to use a model of the form

$$D_{\text{sub}}(t,x) \sim \text{Poisson}\left(\bar{\mu}_{\text{sub}}(t,x)E_{\text{sub}}(t,x)\right), \tag{4.3}$$

with

$$\bar{\mu}_{\text{sub}}(t,x) = H_{\hat{\theta}}(t,x)\exp(y_t' r_x) \tag{4.4}$$

where $y_t' = (y_{0,t},\ldots,y_{n,t})$ and $r_x' = (r_{0,x},\ldots,r_{n,x})$ for some n. Again, this does in fact allow a Lee–Carter specification of the deviations. However, we will consider the situation in which the coordinates in r are fixed regressors and only the coordinates in y are estimated (by maximum likelihood estimation). Note, that in this case the estimates of y_t only depend on data from period t. In Section 4.2.3 we will propose a specific model with three regressors corresponding to level, slope and curvature of the spread.

The last step is to fit a time-series model for the multivariate time-series y_t. We will use a VAR(1)–model for which standard fitting routines

[1] We use $\bar{\mu}$ to indicate an intensity surface which is constant over calender years and over integer ages. We reserve the use of μ, which will later be used to denote a continuous intensity surface.

exist. If the assumptions behind the modelling approach are fulfilled the time-series should not display drift but rather fluctuate around some level (which may be different from zero). In other words, we expect y_t to be *stationary*.

Forecasts are readily produced by combining trend forecasts with time-series forecasts of the spread. Assuming independence between trend and spread we have, with a slight abuse of notation,[2] the following variance decomposition

$$\mathbb{V}\left\{\log \bar{\mu}_{\text{sub}}(t,x)\right\} = \mathbb{V}\left\{\log H_{\hat{\theta}}(t,x)\right\} + \mathbb{V}\left\{y_t' r_x\right\}. \tag{4.5}$$

We see that there are two sources of (systematic) variability: the trend and the spread. For most specifications of H_θ the variance will increase with the forecasting horizon. The variance of the spread, however, will only increase up to a given level under the assumed stationarity of y_t. The model does not forecast the mortality of the subpopulation to convergence in an absolute sense to that of the reference population, but the spread will be bounded (in probability).

In the following sections we develop a specific model which falls within the framework described above. The model will subsequently be used in an application to Danish mortality taking an international data set as reference population. With this application in mind the model has been dubbed SAINT as an abbreviation for Spread Adjusted InterNational Trend.

4.2.2 Trend modelling

A great number of functional forms have been suggested as models for adult mortality, see e.g. Chapter 2 of Gavrilov and Gavrilova (1991). Classical forms include the ones associated with the name of Gompertz

$$\mu(x) = \alpha \exp(\beta x), \tag{4.6}$$

and Makeham

$$\mu(x) = \alpha \exp(\beta x) + \gamma. \tag{4.7}$$

[2] In equation (4.5) $\hat{\theta}$ is considered as an estimator with a distribution rather than a fixed number.

Both of these capture the approximate exponential increase in intensity observed for adult mortality. The Makeham form also includes an age-independent contribution which can interpreted as a rate of accidents. The additional term, referred to as background mortality, improves the fit at young ages.

Old age mortality, however, is generally overestimated by the assumed exponential increase. Empirical data typically shows decreasing acceleration in mortality at old ages, or even a late-life mortality plateau. A functional form that captures both the (approximate) exponential growth rate seen in adult mortality and the subsequent sub-exponential increase at old ages is the logistic family

$$\mu(x) = \frac{\alpha \exp(\beta x)}{1 + \alpha \exp(\beta x)} + \gamma. \tag{4.8}$$

This form has been proposed as the basis for mortality modelling by several authors, e.g. Cairns *et al.* (2006), Bongaarts (2005), Thatcher (1999), and it has been shown to fit empiric data very well in a number of applications.[3]

Selection and frailty

Various theories have been proposed trying to explain why the increase in the force of mortality slows down at old ages, see e.g. Thatcher (1999) and references therein. In this paper we will focus on frailty theory as it provides a flexible and mathematically tractable framework for modelling mortality.

The theory assumes that the population is heterogeneous with each person having an individual level of susceptibility, or frailty. Frail individuals have a tendency to die earlier than more robust individuals and this selection causes the frailty composition of the cohort to gradually change over time. The continued concentration of robust individuals in effect slows down the mortality of the cohort and causes the cohort intensity to increase less rapidly than the individual intensities. The following example illustrates the idea.

[3]Cairns *et al.* (2006) use the logistic form to describe one-year death probabilities q_x, while we use it as model for the underlying intensity. The two quantities are related by $q_x = 1 - \exp(-\int_{x-1}^{x} \mu(y)dt) \approx \mu(x)$.

Example 4.1 (Gamma-Makeham model)**.** *Assume that the* i^{th} *person of a cohort has his own Makeham intensity:*

$$\mu(x; z_i) = z_i \alpha \exp(\beta x) + \gamma, \tag{4.9}$$

where z_i *is an individual frailty parameter, while* α*,* β *and* γ *are shared by all persons in the cohort. Assume furthermore that* Z *follows a (scaled)* Γ*-distribution with mean* 1 *and variance* Σ^2*. The force of mortality for the cohort then becomes*

$$\mu(x) = \mathbb{E}\{Z|x\}\, \alpha \exp(\beta x) + \gamma = \frac{\alpha \exp(\beta x)}{1 + \Sigma^2 \alpha(\exp(\beta x) - 1)/\beta} + \gamma, \tag{4.10}$$

where $\mathbb{E}\{Z|x\}$ *denotes the conditional mean frailty of the cohort at age* x*. The cohort intensity in this model is of logistic form with an asymptotic value of* $\beta/\Sigma^2 + \gamma$ *as* x *tends to infinity. Hence, although each individual intensity is exponentially increasing the selection mechanism is so strong that the cohort intensity levels off at a finite value.*[4]

The multiplicative frailty model

The ideas of selection and frailty can be generalised to describe the evolution in mortality rates over time for a whole population. In the following we assume the existence of a smooth intensity surface, $\mu(t, x)$, which represents the instantaneous rate of dying for a person aged x at time t, i.e. the probability that the person will die between time t and $t + dt$ is approximately $\mu(t, x)dt$ for small dt.

We start by considering a general, multiplicative frailty model where the mortality intensity for an individual with frailty z has the form

$$\mu(t, x; z) = z\mu_s^I(t, x) + \gamma(t). \tag{4.11}$$

Hence, individual intensities have a senescent (age-dependent) component, $z\mu_s^I(t, x)$, and a background (age-independent) component, $\gamma(t)$. Frailty is assumed to affect the senescent component only and its effect is assumed

[4] In fact, if $\Sigma^2 > \beta/\alpha$ the level of heterogeneity is so large, and the selection effect thereby so strong that the cohort intensity is decreasing with age... For $\Sigma^2 = \beta/\alpha$ the cohort intensity is constant, while for smaller values of Σ^2 the cohort intensity is increasing as expected.

to be multiplicative. Thus z measures the excess (senescent) mortality relative to the mortality of an individual with frailty 1. This multiplicative structure is crucial for the following developments. Vaupel *et al.* (1979) consider the multiplicative frailty model for a single cohort and state results similar to ours.

Proposition 4.2. *Assuming (4.11) the population mortality surface is given by*

$$\mu(t,x) = \mathbb{E}\left\{Z|t,x\right\}\mu_s^I(t,x) + \gamma(t), \tag{4.12}$$

where $\mathbb{E}\left\{Z|t,x\right\}$ *denotes the conditional mean frailty at time* t *for persons of age* x*.*

We will denote the senescent component of the population intensity by $\mu_s(t,x)$, i.e. $\mu_s(t,x) = \mathbb{E}\left\{Z|t,x\right\}\mu_s^I(t,x)$. The result of Proposition 4.2 holds true regardless of the assumed frailty distribution at birth. However, in order to obtain an analytically tractable model we will assume that frailties at birth follow a scaled Γ-distribution with mean 1 and variance Σ^2. Under this assumption the conditional frailty distributions are also Γ-distributed and explicit expressions for the conditional mean and variance can be derived. Let $Z|(t,x)$ denote the conditional frailty distribution at time t for persons of age x.

Proposition 4.3. *Assuming (4.11) and* $Z \sim \Gamma$ *with mean* 1 *and variance* Σ^2 *then* $Z|(t,x) \sim \Gamma$ *with mean and variance given by*

$$\mathbb{E}\left\{Z|t,x\right\} = \left(1 + \Sigma^2 I(t,x)\right)^{-1}, \tag{4.13}$$

$$\mathbb{V}\left\{Z|t,x\right\} = \Sigma^2 \mathbb{E}\left\{Z|t,x\right\}^2, \tag{4.14}$$

where $I(t,x) = \int_0^x \mu_s^I(u+t-x,u)du$*.*

Proposition 4.3 characterises how the frailty composition of a given birth-cohort changes over time. At early ages where the integrated intensity $I(t,x)$ is small the selection is modest and the conditional mean and variance are close to the unconditional values of 1 and Σ^2. As the intensity increases so does $I(t,x)$ and the conditional mean and variance decrease towards 0. Thus, over time the frailty distribution gets more and more concentrated around 0.

The following proposition shows that the conditional mean frailty can also be expressed in terms of the senescent population mortality.

Proposition 4.4. *Under the assumptions of Proposition 4.3*

$$\mathbb{E}\left\{Z|t,x\right\} = \exp\left(-\Sigma^2 H(t,x)\right), \tag{4.15}$$

where $H(t,x) = \int_0^x \mu_s(u+t-x,u)du$.

As an immediate consequence of Proposition 4.4 we have the following inversion formula,

$$\mu_s^I(t,x) = \mu_s(t,x)\exp\left(\Sigma^2 H(t,x)\right), \tag{4.16}$$

which allows us to recover the individual intensities from the population intensity and the level of heterogeneity, Σ^2. The existence of such a formula implies that any population mortality surface can be described by a frailty model with a given level of heterogeneity.

When discussing improvement rates it is most illuminating to focus on the senescent part of the mortality. The background mortality component is primarily included for the purpose of improving the fit among young adults and, relative to the senescent part, its contribution to mortality is negligible for older age groups. Following the notation of Bongaarts (2005) we define the rate of improvement in senescent mortality as

$$\begin{aligned}\rho_s(t,x) &\triangleq -\frac{\partial \log \mu_s(t,x)}{\partial t} \\ &= -\frac{\partial \log \mathbb{E}\left\{Z|t,x\right\}}{\partial t} - \frac{\partial \log \mu_s^I(t,x)}{\partial t}.\end{aligned} \tag{4.17}$$

Generally, healthier conditions and other improvements in individual survival will mean that the contribution from the last term is positive. However, higher survival rates imply less selection and the average frailty of persons of age x will therefore increase to 1, the average frailty at birth, over time. Thus the contribution from the first term is negative. For old age groups with strong selection the changing frailty composition can substantially offset the general improvements but eventually the effect dies out and improvements occur.

To capture the idea that the mortality of an individual is affected by both accumulating and non-accumulating factors we will write the (baseline) individual intensity in the form

$$\mu_s^I(t,x) = \kappa(t,x)\exp\left(\int_0^x g(u+t-x,u)du\right). \tag{4.18}$$

We think of κ as representing the current level of treatment/health at time t for persons of age x, while the accumulating factor g represents the aging process. The idea is that $g(t,x)$ represents the increase in (log) mortality caused by aging at time t for persons of age x. Hence, to get the accumulated effect of aging one needs to integrate from age 0 at the time of birth, $t-x$, to the current age x at time t.

Specification

We will consider the following specialisation of the model given by (4.11) and (4.18):

$$g(t,x) = g_1 + g_2(t-t_0) + g_3(x-x_0), \tag{4.19}$$

$$\kappa(t) = \exp(\kappa_1 + \kappa_2(t-t_0)), \tag{4.20}$$

$$\gamma(t) = \exp(\gamma_1 + \gamma_2(t-t_0)), \tag{4.21}$$

with $x_0 = 60$ and $t_0 = 2000$. The substraction of (year) 2000 in the specification of g, κ and γ and 60 in g is done for interpretability reasons only. Thus $g(2000,60) = g_1$ is the "aging" of a 60-year old in year 2000, g_2 is the additional aging across ages for each calender year, and g_3 is the additional biological aging for each year of age. Similarly, $\kappa(2000) = \kappa_1$ and $\gamma(2000) = \gamma_1$ while κ_2 and γ_2 give the annual rates of change. Notice, that κ depends only on time since the obvious "missing" term, $\kappa_3(x-x_0)$, is already present through g_1. The model has a total of 8 parameters; the 7 parameters appearing in the specification of g, κ and γ together with the variance of the frailty distribution, Σ^2. As we will later demonstrate the model is able to capture the main characteristics of the data very well despite its parsimonious structure.

From a computational perspective it is convenient to think of the intensities as functions of birth year, rather than calender year, and age. By use of Propositions 4.2 and 4.3 we can write μ as

$$\mu(t,x) = \frac{K(t-x,x)}{1+\Sigma^2\int_0^x K(t-x,y)dy} + \gamma(t), \tag{4.22}$$

where $K(b,x) = \mu_s^I(b+x,x)$. This representation highlights the fact that the integral in the denominator relates to a given birth-cohort.

For the model above we have

$$K(b,x) = \kappa(b)\exp\left((g(b,0)+\kappa_2)x + (g_2+g_3)x^2/2\right). \tag{4.23}$$

That is, $K(b,x)$ is log-quadratic in x (for fixed b). When $g_2 + g_3 < 0$ the integral in the denominator can be expressed in terms of the cdf of a normal distribution, while this is not possible when the sum is positive. In either case, it is easy to evaluate the integral numerically.

The model allows us to derive the current and asymptotic improvement patterns in age-specific death rates.

Proposition 4.5. *Assume $\kappa_2 < 0$. If $\kappa_2 + g_2 x < 0$ then*

$$\rho_s(t,x) = -\frac{\partial \log \mathbb{E}\{Z|t,x\}}{\partial t} - (\kappa_2 + g_2 x) \to -(\kappa_2 + g_2 x) \text{ for } t \to \infty. \tag{4.24}$$

The conditions of the proposition imply that all age groups up to age x experience improvements. Note, however, that g_2 may be either positive or negative. Thus the model allows for (asymptotic) improvement rates in senescent mortality to be either increasing or decreasing with age. The presence of the first term means that improvement rates can be substantially lower for a long time before eventually approaching their long–run level. In extreme cases the first term may even dominate the second term, representing general improvements, in which case ASDRs will increase for a period before starting to decrease. Some people have indeed argued that this may happen in old-age mortality. However, we do not find support for increasing ASDRs in the data analysed in this paper.

The model can be viewed as a generalisation of the Gamma-Makeham model of Section 4.2.2. Indeed, that model is obtained by letting all of g, κ and γ be constant (if only g is constant we obtain the model proposed by Vaupel (1999)). However, unlike the Gamma-Makeham model the cohort mortality profiles of our model will generally not have finite asymptotes.

Proposition 4.6. *Assume $\Sigma^2 > 0$. The limit cohort mortality is given by*

$$\lim_{x\to\infty} \mu_s(b+x,x) = \begin{cases} \infty & \text{if } g_2+g_3 > 0, \\ \frac{g(b,0)+\kappa_2}{\Sigma^2} & \text{if} g_2+g_3 = 0 \text{ and } g(b,0)+\kappa_2 > 0, \\ 0 & \text{else.} \end{cases}$$

The cohort mortality profile is the result of two opposite effects: the increase in individual mortality pushes the cohort mortality upwards, while the selection mechanism pushes it downwards. For Γ-distributed frailties and exponential individual intensities the two effects balance each other in such a way that an old-age mortality plateau occurs. This is the case in the Gamma-Makeham model and in the second case of Proposition 4.6. However, when individual intensities increase faster than exponential the individual effect dominates and the cohort mortality converges to infinity (although at a slower pace than the individual intensities). Conversely, for sub-exponential individual intensities the selection effect dominates and the cohort mortality goes to zero. These two situations correspond to the first and third case, respectively, of Proposition 4.6.[5] In the application later in the paper the estimates of g_2 and g_3 are both positive. Hence we find ourselves in the first case. That individual intensities increase faster than exponential was also found by Yashin and Iachine (1997).

Estimation

We next want to estimate the parameters of model (4.19)–(4.21) using the reference data. Since the intensity surface is a continuous function of time and age, while data is aggregated over calender years and one year age groups, we define, for integer values of t and x,

$$\bar{\mu}_{\text{ref}}(t,x) = \frac{1}{4}\left(\mu(t,x) + \mu(t,x+1) + \mu(t+1,x) + \mu(t+1,x+1)\right) \quad (4.25)$$

to represent the average intensity over the square $[t,t+1)\times[x,x+1)$ of the Lexis diagram.[6] We will use the Poisson–model in (4.2) with $\bar{\mu}_{\text{ref}}(t,x)$ given by (4.25) to find the MLE, $\hat{\theta}$, of the parameters

$$\theta = (\Sigma, g_1, g_2, g_3, \kappa_1, \kappa_2, \gamma_1, \gamma_2).$$

[5]The third case of Proposition 4.6 is an extreme case of sub-exponential growth in which the individual intensities are in fact decreasing with age, at least from some age. However, the result holds for any sub-exponential intensity, e.g. polynomial.

[6]There are other possibilities for defining $\bar{\mu}_{\text{ref}}(t,x)$. For instance, $\bar{\mu}_{\text{ref}}(t,x) = \mu(t+1/2, x+1/2)$, or $\bar{\mu}_{\text{ref}}(t,x) = \int_0^1\int_0^1 \mu(t+s,x+u)dsdu$. If the exposure is uniform over the square one may argue in favor of the latter definition, but it is cumbersome to implement and unlikely to yield any substantial differences.

This is achieved by maximising the log–likelihood function

$$\begin{aligned} l(\theta) &= \sum_{t,x} D_{\text{ref}}(t,x)\log(\bar{\mu}_{\text{ref}}(t,x)E_{\text{ref}}(t,x)) - \log(D_{\text{ref}}(t,x)!) \\ &\quad - \bar{\mu}_{\text{ref}}(t,x)E_{\text{ref}}(t,x) \\ &= \sum_{t,x} D_{\text{ref}}(t,x)\log(\bar{\mu}_{\text{ref}}(t,x)) - \bar{\mu}_{\text{ref}}(t,x)E_{\text{ref}}(t,x) + \textit{constant}, \end{aligned}$$

where the last term does not depend on θ and hence need not be included in the maximisation. It is straightforward to implement the log-likelihood function and to maximise it by standard numeric optimisation routines. We have used the `optim` method in the freeware statistical computing package R for our application.

Generally, maximum likelihood estimates are (under certain regularity conditions) asymptotically normally distributed with variance matrix given by the inverse Fisher information[7] evaluated at the true parameter. As an estimate of the variance-covariance matrix we will use

$$\widehat{\text{Cov}}\left\{\hat{\theta}\right\} = -D^2_\theta l(\hat{\theta})^{-1}, \tag{4.26}$$

which can be computed numerically once $\hat{\theta}$ has been obtained. Using the variance estimates and the approximate normality (approximate) 95%-confidence intervals for the parameters can be computed as $\hat{\theta} \pm 1.96\widehat{\mathbb{V}}\left\{\hat{\theta}\right\}$, where $\widehat{\mathbb{V}}\left\{\hat{\theta}\right\}$ denotes the diagonal of $\widehat{\text{Cov}}\left\{\hat{\theta}\right\}$.

Bootstrapping constitutes an alternate approach to assessing the parameter uncertainty which does not rely on asymptotic properties, see e.g. Efron and Tibshirani (1993); Koissi *et al.* (2006). In short, the method consists of simulating a number of new data sets, i.e. new death counts, given the observed exposures and the estimated intensities and calculate the MLE for each data set. The resulting (bootstrap) distribution reflects the uncertainty in the parameter estimates. Although simple in theory the computational burden is in our case substantial as each maximisation takes several minutes.

[7] The Fisher information is defined as minus the expected value of the second derivative of the log-likelihood function, $\mathcal{I}(\theta) = -\mathbb{E}\left\{D^2_\theta l(\theta)|\theta\right\}$.

4.2.3 Spread modelling

The fundamental assumption behind the proposed method for modelling small population mortality is the existence of an underlying (smooth) mortality surface, the trend, around which the small population mortality evolves. In this section we focus on modelling and estimating the deviations of the small population mortality from the trend.

Spread

For given underlying trend, $\bar{\mu}_{\text{ref}}$, we will assume that subpopulation death counts are independent and distributed according to

$$D_{\text{sub}}(t,x) \sim \text{Poisson}\left(\bar{\mu}_{\text{sub}}(t,x)E_{\text{sub}}(t,x)\right), \tag{4.27}$$

where

$$\bar{\mu}_{\text{sub}}(t,x) = \bar{\mu}_{\text{ref}}(t,x)\exp(y_t' r_x) \tag{4.28}$$

with $y_t' = (y_{0,t}, \ldots, y_{n,t})$ and $r_x' = (r_{0,x}, \ldots, r_{n,x})$ for some n. The spread between the mortality of the subpopulation and the trend is modelled by the last term in (4.28). The regressors $r_0, \ldots, r_n$ determine the possible age-profiles of the (log) spread, while $y_0, \ldots, y_n$ describe the evolution over time of the corresponding components of the spread. We will refer to the coordinates of y as the spread parameters.

As opposed to the elaborate trend model we have chosen to use a rather simple log-linear parametrisation of the spread. We do this for two reasons. First, assuming the trend model captures the main features of the mortality surface we expect there to be only little structure left in the spread. Introducing a complex functional form for the spread therefore seems fruitless. Second, a complicated spread model would to some extent counter the idea of the model. The spread is supposed to model only the random, but potentially time-persistent, fluctuations around the underlying mortality evolution.

Regarding the choice of dimensionality, n, we are faced with the usual trade–off. A high number of regressors can fit the spread evolution very precisely, but there is a risk of overfitting thereby impairing forecasting ability. Also, a high number of spread parameters are harder to model and will, typically, increase forecasting uncertainty. A low number of regressors, on the other hand, will fit the spreads less well and can be expected

to capture only the overall shape. However, a low number of spread parameters are easier to model and, generally, provides more robust and less uncertain forecasts.

For a given number of regressors there are essentially two ways to proceed. Either, the regressors are specified directly and only the spread parameters are estimated, or both regressors and spread parameters are estimated simultaneously from the data. We prefer the former method due to ease of interpretability of the spread parameters and presumed better forecasting ability; although we recognise that the latter method provides a better (with-in sample) fit.

Specifically, we propose to parameterise the spread by the three regressors

$$r_{0,x} = 1, \tag{4.29}$$

$$r_{1,x} = (x - 60)/40, \tag{4.30}$$

$$r_{2,x} = (x^2 - 120x + 9160/3)/1000, \tag{4.31}$$

which describe, respectively, the level, slope and curvature of the spread. For ease of interpretability the regressors are chosen orthogonal and they are normalised to (about) unity at ages 20 and 100.[8] The number of regressors reflects a compromise between fit and ease of modelling which appears to work well in our application.

Due to the assumed independence of death counts the MLE of the spread parameters for year t depends only on data for that year. For each year of subpopulation data we obtain the MLE for y_t by maximising the log-likelihood function

$$l(y_t) = \sum_x D_{\text{sub}}(t,x) y_t' r_x - \bar{\mu}_{\text{ref}}(t,x) \exp(y_t' r_x) E_{\text{sub}}(t,x) + \textit{constant},$$

where $\bar{\mu}_{\text{ref}}(t,x)$ is calculated with the maximum likelihood estimates from Section 4.2.2 inserted. Note that the parametric form of the underlying trend ensures that we can calculate $\bar{\mu}_{\text{ref}}(t,x)$ for all x and t. Thus the age

[8] In the application we use mortality data for ages 20 to 100, i.e. 81 one-year age groups. Seen as vectors the three regressors are orthogonal w.r.t. the usual inner product in $\mathbb{R}^{81}$. The regressors are normalised such that $r_{2,20} = -1$ and $r_{2,100} = 1$, while $r_{3,20} = r_{3,100} = 1.053$. If desired we can obtain $r_{3,20} = r_{3,100} = 1$ by changing the normalisation constant from 1000 to 3160/3 in the definition of r_3.

and time windows for which we have data for the sub- and reference population need not coincide, or even overlap. In practice, of course, we expect there to be at least a partial overlap. For example, if the subpopulation is the current and former members of a specific pension scheme, or a specific occupational or ethnic group, we might have only a relatively short history of data, while we might have a considerably longer history of national data which we might want to use as reference data.

Time dynamics

The multivariate series of spread parameters describe the evolution in excess mortality in the subpopulation relative to the reference population. Over time we expect the two populations to experience similar improvements and we therefore believe the spread to be stationary rather than showing systematic drift. We also expect the spread to show time-persistence. If at a given point in time the mortality of the subpopulation is substantially higher or lower than the reference mortality we expect it to stay higher or lower for some time thereafter. Finally, we expect the spread parameters to be dependent rather than independent. The regressors are chosen to have a clear interpretation, but we do not expect, e.g. the level and the slope of the spread to develop independently of each other over time.

The simplest model meeting these requirements is the vector autoregressive (VAR) model which we will adopt as spread parameter model. Specifically we suggest to use the Gaussian VAR(1)–model

$$y_t = Ay_{t-1} + \epsilon_t, \tag{4.32}$$

where A is a three by three matrix of autoregression parameters and the errors ϵ_t are three-dimensional i.i.d. normally distributed variates with mean zero and covariance matrix Ω, i.e. $\epsilon_t \sim N_3(0, \Omega)$.

By not including a mean term in the model we implicitly assume that the spread will converge to zero (in expectation) over time. We believe this is a natural condition to impose for the application to Danish data with an international reference data set considered in this paper. Indeed, it is hard to justify the opposite, that Danish mortality should deviate systematically from international levels indefinitely – even if historic data

were to suggest it. For other applications one may wish to include a mean term in the model and thereby allow for systematic deviations. Similarly, one may wish to consider more general VAR–models with additional lags to capture more complex time-dependence patterns.

The parameters A and Ω of model (4.32) can be estimated by the `ar` routine in R treating the time series of estimated spread parameters, y_t, as observed variables. The routine offers various estimation methods. It would have been in the spirit of this paper to use maximum likelihood estimation, but unfortunately this option is only implemented for univariate time series. Instead we use Yule-Walker estimation which obtains estimates by solving the Yule-Walker equations, cf. e.g. Brockwell and Davis (1991). In our application the estimated A defines a stationary time series, i.e. the modulus of A's eigenvalues is smaller than one, for both men and women. However, in general there is no guarantee for this. As with all statistical analysis where data contradicts modelling assumptions one will then have to propose a more suitable model, e.g. introduce a mean term or additional lags, and reiterate the analysis.

Forecast

Forecasting in the VAR–model (4.32) is based on the conditional distribution of the future values of the spread given the observed values. Assuming year T to be the last year of observation and h to be the forecasting horizon we need to find the conditional distribution of y_{T+h} given the observed values. Due to the Markov property of the VAR(1)–model this distribution depends on the last observed value, y_T, only. Expanding the data generating equation we obtain for $h \geq 1$

$$y_{T+h}|y_T \sim N(m_h, V_h),$$

where m_h and V_h are given by

$$m_h = A^h y_T, \quad V_h = \sum_{i=0}^{h-1} A^i \Omega (A^i)'.$$

From these expressions forecasted values for future spread parameters and corresponding two sided (pointwise) 95%-confidence intervals are easily obtained as

$$\mathrm{CI}_{95\%}(y_{T+h}) = m_h \pm 1.96\sqrt{\mathrm{diag}(V_h)}.$$

Note that due to stationarity the forecasting uncertainty will increase towards a finite limit as the forecasting horizon increases. Thus the deviations of the subpopulation from the reference population are bound (in probability) over time. By use of equation (4.28) we further have

$$\mathrm{CI}_{95\%}\left(\log \bar{\mu}_{\mathrm{sub}}(T+h,x)\right) = \log \bar{\mu}_{\mathrm{ref}}(T+h,x) + m_h' r_x \pm 1.96\sqrt{r_x' V_h r_x}.$$

The stated confidence intervals only reflect the stochasticity of the VAR–model itself without taking the parameter uncertainty into account. They are therefore, in a sense, the "narrowest" possible confidence intervals. Confidence intervals incorporating parameter uncertainty of both the trend and the spread model can be constructed by bootstrap, but we will not pursue that point further. Cairns (2000) also considers model uncertainty, i.e. the uncertainty associated with determining the underlying model, and he discusses how all three types of uncertainty can be assessed coherently in a Bayesian framework.

We have concentrated on assessing the uncertainty of a single ASDR at a future point in time. Since the conditional distribution of the entire future $\{y_{T+1}, \ldots\}$ given y_T is readily available we can also derive simultaneous confidence intervals for any collection of ASDRs by the same method. In principle, it is therefore possible to derive analytic confidence intervals for any functional of the intensity surface. In practice, however, most quantities of interest, e.g. remaining life expectancy, are too complicated to allow analytic derivations. Instead it is necessary to resort to Monte Carlo methods to assess forecasting uncertainty of any but the simplest quantities. Fortunately, this is straightforward to implement. We simply simulate a large number of realisations from the VAR–model (4.32) and calculate the corresponding intensity surface by (4.28). For each surface we calculate the quantity of interest and we thereby obtain (samples from) the forecasting distribution. This is illustrated in Section 4.3.4.

4.3 Application

To demonstrate the model in action we consider the case that gave rise to the name SAINT, namely Denmark as the (sub)population of interest and a basket of developed countries as the reference population. The model is applied to each sex separately.

4.3.1 Data

Data for this study originates from the Human Mortality Database,[9] which offers free access to updated records on death counts and exposure data for a long list of countries. The database is maintained by University of California, Berkeley, United States and Max Planck Institute for Demographics Research, Germany.

We will use both Danish data and a pooled international data set consisting of data for the following 19 countries: USA, Japan, West Germany, UK, France, Italy, Spain, Australia, Canada, Holland, Portugal, Austria, Belgium, Switzerland, Sweden, Norway, Denmark, Finland and Iceland. This set is chosen among the 34 countries represented in the Human Mortality Database because of their similarity to Denmark with respect to past and presumed future mortality. Table 4.6 in Section 4.7 contains a summary of the data.

The subsequent analysis uses data from the years 1933 to 2005 and ages 20 to 100. As far as the time dimension is concerned the cut points are determined by the availability of US data. Concerning the age span the analysis could in principle be based on all ages from 0 to 110, which are all available in the Human Mortality Database. However, since the prime focus is adult mortality and since the mortality pattern at young ages differs markedly from adult mortality all ages below 20 have been excluded. For very high ages the quality of data is poor and sometimes based on disaggregated quantities and for this reason all ages above 100 have also been excluded.[10]

The international data set is constructed as the aggregate of the 19 national data sets. For each year from 1933 to 2005 and each age from 20 to 100 the international death count and international exposure consists of, respectively, the total death count and total exposure of those of the 19 countries for which data exists for that year. Measured in terms of death counts and exposures the international data set is more than 100 times larger than the Danish data set.

[9] See www.mortality.org

[10] In some countries and some years data for ages younger than 100 years is also based on disaggregated quantities, but we suspect this to be of minor importance.

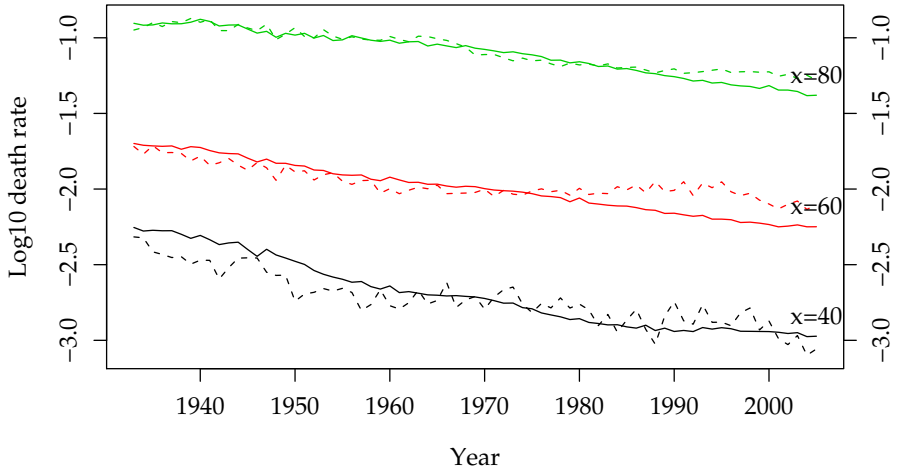

Figure 4.1: Danish (dashed line) and international (solid line) development in female death rates from 1933 to 2005 for selected ages.

4.3.2 Trend

To illustrate the data we have plotted Danish and international female death rates for ages 40, 60 and 80 years in Figure 4.1. Compared to Denmark the international development in death rates has been quite stable with only slowly changing annual rates of improvement. The Danish mortality evolution, on the other hand, shows a much more erratic behavior with large year-to-year variation in improvement rates. The Danish level seems to follow that of the international community in the long run, but there are extended periods with substantial deviations. The most striking of these is the excess mortality of Danish females around age 60 which emerged in 1980, peaked more than a decade later and is still present today although less pronounced. From Figure 4.1 and similar plots it seems reasonable to think of Danish mortality rates as fluctuating around a stable international trend.

We have used the international data set to estimate the trend model de-

Parameter	Women		Men	
	Estimate	95%-CI	Estimate	95%-CI
Σ	$4.2860 \cdot 10^{-1}$	$\pm 1.6 \cdot 10^{-4}$	$2.6243 \cdot 10^{-1}$	$\pm 2.3 \cdot 10^{-4}$
g_1	$9.8965 \cdot 10^{-2}$	$\pm 2.3 \cdot 10^{-6}$	$1.0551 \cdot 10^{-1}$	$\pm 2.2 \cdot 10^{-6}$
g_2	$4.7856 \cdot 10^{-6}$	$\pm 5.1 \cdot 10^{-8}$	$8.3744 \cdot 10^{-5}$	$\pm 3.7 \cdot 10^{-8}$
g_3	$1.3103 \cdot 10^{-3}$	$\pm 1.0 \cdot 10^{-7}$	$5.5903 \cdot 10^{-5}$	$\pm 8.6 \cdot 10^{-8}$
κ_1	$-8.7819 \cdot 10^{0}$	$\pm 1.7 \cdot 10^{-4}$	$-1.0576 \cdot 10^{1}$	$\pm 1.6 \cdot 10^{-4}$
κ_2	$-1.8510 \cdot 10^{-2}$	$\pm 5.8 \cdot 10^{-6}$	$-1.7827 \cdot 10^{-2}$	$\pm 5.0 \cdot 10^{-6}$
γ_1	$-1.1810 \cdot 10^{1}$	$\pm 1.9 \cdot 10^{-3}$	$-7.5222 \cdot 10^{0}$	$\pm 8.4 \cdot 10^{-4}$
γ_2	$-8.9038 \cdot 10^{-2}$	$\pm 3.2 \cdot 10^{-5}$	$-2.5005 \cdot 10^{-2}$	$\pm 2.0 \cdot 10^{-5}$

Table 4.1: Maximum likelihood estimates and 95% confidence intervals for the trend model given in Section 4.3.2. The estimation is based on international mortality data from 1933 to 2005 for ages 20 to 100 years.

scribed in Section 4.2.2. Table 4.1 contains maximum likelihood estimates of the eight parameters and corresponding two sided 95% confidence intervals. The narrow width of the confidence intervals reflects the fact that, relative to the amount of data, we have a very parsimonious model. Using a similar model Barbi (2003) reports standard errors of the same magnitude in an application to Italian data. The small standard errors indicate that the parameters are well determined, but this does not necessarily imply a good fit. The fit of the model can be assessed graphically on Figure 4.2 which shows international female mortality rates together with the estimated trend. Overall, it appears that the model does a remarkably good job at describing the data. There are appreciable deviations only for the very youngest and very highest ages. For now we settle with this informal graphical inspection of goodness-of-fit, but we will return to the issue more formally in Section 4.4.

The model has three parameters to describe three different types of improvement in mortality over time. The parameters g_2 and κ_2 affect the improvement in senescent mortality, while the reduction in background mortality is determined by γ_2. By Proposition 4.5 we know that the limiting rate of improvement in senescent mortality is given by $-(\kappa_2 + g_2 x)$ for

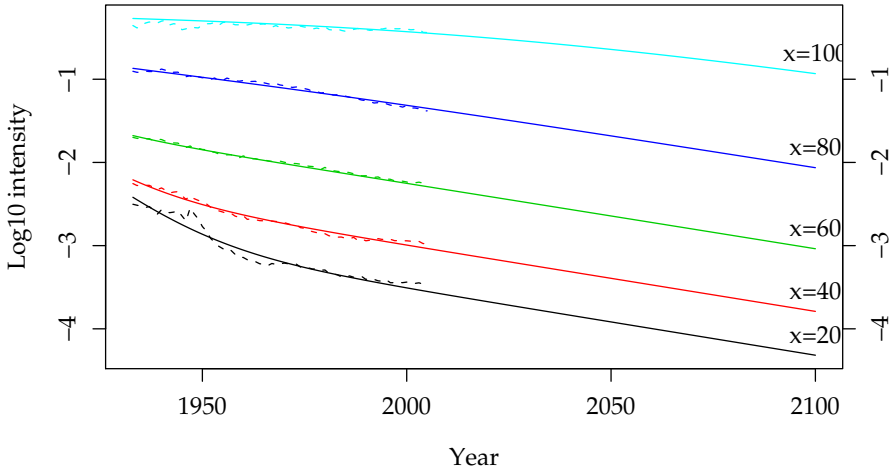

Figure 4.2: Historic development (dashed line) in international female death rate from 1933 to 2005 for the age groups 20, 40, . . . , 100. Model estimate of trend with parameters given in Table 4.1 is superimposed (solid line).

ages x for which this quantity is positive.[11] Since g_2 is positive for both sexes the (limiting) rate of improvement is decreasing in age as expected. Due to frailty we also have that the limiting improvement rate is achieved more slowly for higher ages than for lower ages. This further "steepens" the age-profile of improvement rates and causes it to change shape over time as illustrated in Figure 4.3. Note in particular how the improvement rate for 100–year–olds is projected to double over the next century. The projected increase in improvement rates can also be observed from the curved projection in Figure 4.2 for this age group (the effect is also present for the younger age groups but much less pronounced).

The value of κ_2 is estimated to about -1.8% for both women and men.

[11] With the estimated parameter values this is satisfied up to age 213 for men and 3870 years for women, i.e. for all ages of practical relevance.

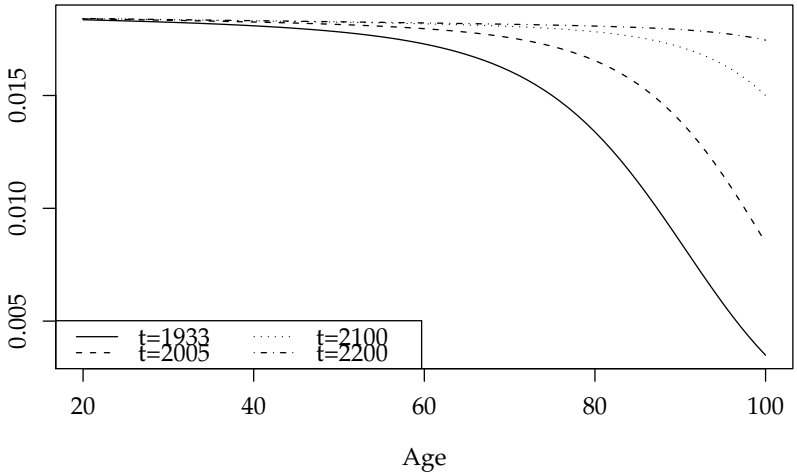

Figure 4.3: Rate of improvement, $\rho_s(t, x)$, in female senescent mortality. Based on parameter values from Table 4.1.

The value of g_2, however, is almost 20 times larger for men than for women. This implies that whereas the limiting age-profile of improvement rates is almost flat for women the age-profile will remain steep for men. Thus 100–year–olds females will eventually experience the same annual rate of improvement as the younger age groups, while old men will continue to have lower improvement rates than younger men.

The improvement rate of background mortality is estimated to about 9% for women and about 2.5% for men. The large value for women is due to the dramatic decrease in mortality among the 20 to 30–year–olds in the beginning of the observation period, cf. Figure 4.2. However, since female background mortality is now negligible compared to senescent mortality further reductions will have virtually no effect on (total) mortality even at young ages. For men, on the other hand, background mortality is much higher (as can be seen by the difference in the values of γ_1) and future reductions will have a substantial effect on young age mortality.

The estimated level of heterogeneity, Σ, is higher for women than for men. Since the Σ parameter controls the "delay" in achieving the asymptotic rate of improvement, i.e. the first term in (4.24) of Proposition 4.5, this implies that the asymptotic rate is approached slower for women than for men. It is not clear why this is the case, but a gender difference of the same magnitude was found by Barbi (2003).

The remaining parameters control the age profile of (senescent) mortality. As expected the parameter for the overall level, κ_1, and the parameter for first-order age dependence, g_1, are both estimated to be smaller for women than for men. Only the second-order age dependence parameter, g_3, is higher, albeit still small, for women than for men. The parameter is positive for both sexes implying that aging is accelerating with age. Thus for all ages of practical relevance female mortality is estimated to be lower than male mortality, but for very advanced ages female mortality will in fact exceed male mortality.

4.3.3 Spread

We will apply the three factor spread model of Section 4.2.3 to describe the Danish fluctuations around the international level. The estimated spread series for women are shown in Figure 4.4, where the excess mortality of Danish women from around 1980 onwards is clearly visible. Note that simultaneously with the increase of the level the curvature has decreased. This means that only women around age 60 experience excess mortality while the mortality of very young and very old Danish women is similar to the international level.

In 2005 the estimated spread parameters were $(0.17, 0.06, -0.16)$ for women and $(-0.01, 0.15, 0.08)$ for men. This in fact implies an excess mortality of more than 25% for Danish women of age 60, but only 6% at age 100. Danish men, on the hand, are very much in line with the international level having an excess mortality of 3% at age 60.

The parameter estimates of the VAR(1)–model, which describes the

dynamics of the spread series, are

$$A = \begin{pmatrix} 0.6861 & -0.1907 & -0.2739 \\ -0.1423 & 0.8724 & -0.1558 \\ -0.2422 & -0.1035 & 0.5179 \end{pmatrix},$$
$$\Omega = 10^{-3} \begin{pmatrix} 2.0449 & -0.7341 & 0.3012 \\ -0.7341 & 2.9779 & -0.7376 \\ 0.3012 & -0.7376 & 1.3278 \end{pmatrix},$$

for women and

$$A = \begin{pmatrix} 0.7885 & -0.1714 & -0.1283 \\ -0.2485 & 0.6387 & 0.0477 \\ -0.0650 & -0.0792 & 0.9130 \end{pmatrix},$$
$$\Omega = 10^{-3} \begin{pmatrix} 1.4840 & -0.3818 & 0.3646 \\ -0.3818 & 3.0056 & -1.1880 \\ 0.3646 & -1.1880 & 3.4158 \end{pmatrix},$$

for men. In both cases the A matrix give rise to stationary series. Note that diagonal and off-diagonal elements of A are of the same magnitude because of the high interdependence between the three spread components. Also the errors are highly correlated.

Figure 4.4 shows the mean forecast for the spread parameters for women with 95% confidence intervals. Due to stationarity all three spread components are forecasted to converge to zero. However, due to the structure of A the convergence is not necessarily monotone. The slope, for instance, starts out positive but is forecasted to become negative before approaching zero.

The width of the confidence intervals reflects the observed variation in the spread over the estimation period. The confidence intervals expand quite rapidly to their stationary values indicating that substantial deviations can build up or disappear in a matter of decades. The confidence intervals do not include parameter uncertainty, but only the uncertainty induced by the error term of the VAR–model. Incorporating parameter, or indeed model, uncertainty will most likely lead to even wider confidence intervals.

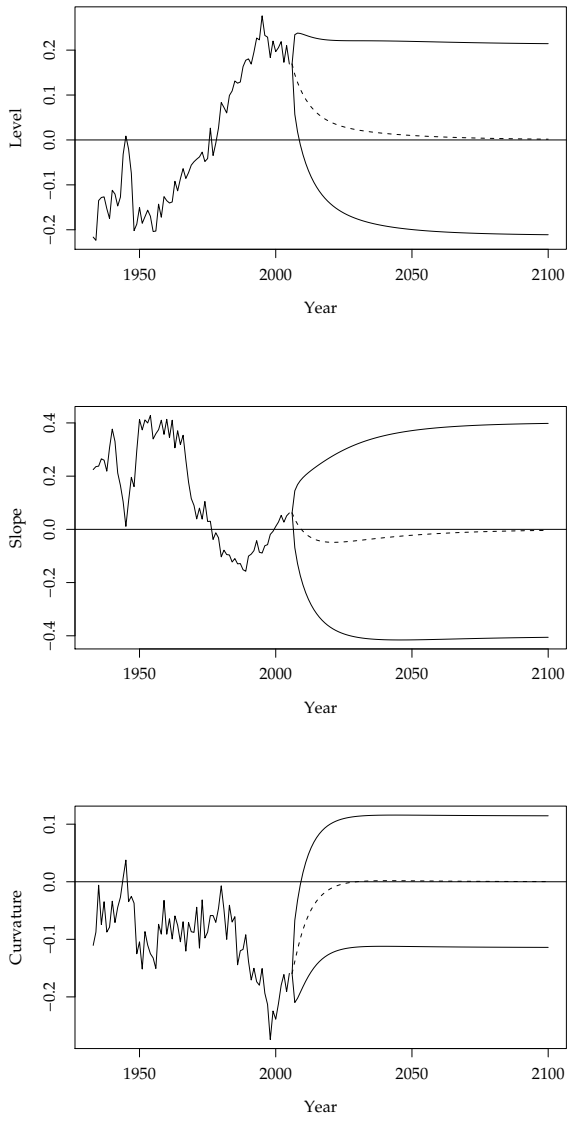

Figure 4.4: Estimated and forecasted spread parameters for Danish women with two sided pointwise 95% confidence intervals.

4.3.4 Forecasting life expectancy

Life expectancies are an intuitively appealing way to summarise a mortality surface. Letting $\bar{\mu}$ denote an intensity surface which is constant over Lexis squares, cf. Section 4.2.1, and letting t and x be integers we get the following approximation for the *cohort* mean remaining life time of individuals born at $t - x$ (at time t)

$$\begin{aligned}\bar{e}^{\mathrm{c}}(t,x) &= \mathbb{E}\left\{T - x \middle| T \geq x\right\} \\ &= \int_0^\infty \exp\left(-\int_0^s \bar{\mu}(t+v, x+v)dv\right) ds \\ &\approx \sum_{j=0}^{M} \exp\left(-\sum_{i=0}^{j-1} \bar{\mu}(t+i, x+i)\right) \frac{1 - \exp\left(-\bar{\mu}(t+j, x+j)\right)}{\bar{\mu}(t+j, x+j)}\end{aligned} \tag{4.33}$$

for some large M. Similarly, the *period* mean remaining life time is

$$\bar{e}^{\mathrm{p}}(t,x) \approx \sum_{j=0}^{M} \exp\left(-\sum_{i=0}^{j-1} \bar{\mu}(t, x+i)\right) \frac{1 - \exp\left(-\bar{\mu}(t, x+j)\right)}{\bar{\mu}(t, x+j)}. \tag{4.34}$$

The cohort life expectancy is calculated from the ASDRs of a specific cohort, i.e. along a diagonal of the Lexis diagram, while the period life expectancy is calculated from the ASDRs at a given point in time, i.e. along a vertical line of the Lexis diagram. The cohort life expectancy represents the actual life expectancy of a cohort taking the future evolution of ASDRs into account. The period life expectancy, on the other hand, is the life expectancy assuming no future changes in ASDRs. For this reason the cohort life expectancy is (substantially) higher than the corresponding period life expectancy.

In Table 4.2 we have shown selected cohort and period life expectancies for women based on point estimates of the intensities (obtained from the mean forecast of the spread). Period life expectancies based on observed death rates for 2005 are also shown. We note that they correspond very well with the model estimates indicating a good fit of the model in the jump–off year.

The table contains period life expectancy forecasts up to year 2105 while cohort life expectancies are forecasted only up to year 2025. In principle, we can calculate cohort life expectancies for 2105 also. However,

	International				Denmark			
	Age				Age			
Year	20	60	70	80	20	60	70	80
2005	71.67	27.37	17.88	10.17	71.67	27.05	17.41	9.66
2025	74.96	30.27	20.36	11.98	74.96	30.26	20.34	11.96

	International				Denmark			
	Age				Age			
Year	20	60	70	80	20	60	70	80
2005	62.92	24.85	16.54	9.64	60.89	23.10	15.11	8.68
	(62.84)	(25.12)	(16.83)	(9.70)	(60.89)	(23.11)	(15.19)	(8.84)
2025	65.98	27.43	18.75	11.28	65.91	27.36	18.69	11.24
2045	68.94	30.02	21.04	13.08	68.94	30.02	21.04	13.08
2105	77.27	37.70	28.15	19.13	77.27	37.70	28.15	19.13

Table 4.2: Upper panel: Cohort remaining life expectancy in years for women. The numbers are based on model forecasts and calculated using (4.33) with $M = 120$. Lower panel: Period remaining life expectancy in years for women. The numbers are based on model forecasts and calculated using (4.34) with $M = 120$. The period remaining life expectancy based on observed death rates for year 2005 (with $M = 110$) is shown in brackets.

assuming a maximal age of 120 years this would require that we project ASDRs to year 2205. No matter how good a model, we cannot give credence to quantities based on projections 200 years into the future and we have therefore chosen to omit the numbers.

The current excess mortality for Danish women can be seen as lower period life expectancies in 2005. The Danish cohort life expectancies are also lower than the international levels but the differences are smaller due to future convergence of Danish rates to the international trend. After 20 years the differences between Danish and international life expectancies have virtually disappeared.

The forecasting uncertainty of complicated functionals such as life expectancy can be assessed by Monte Carlo methods as described in Section 4.2.3. As an illustration of this approach we show in Figure 4.5 the forecasting distribution for the cohort life expectancy of a 60-year-old Danish women in 2005, $\bar{e}^{\mathrm{c}}(2005, 60)$, based on 100,000 simulations. The em-

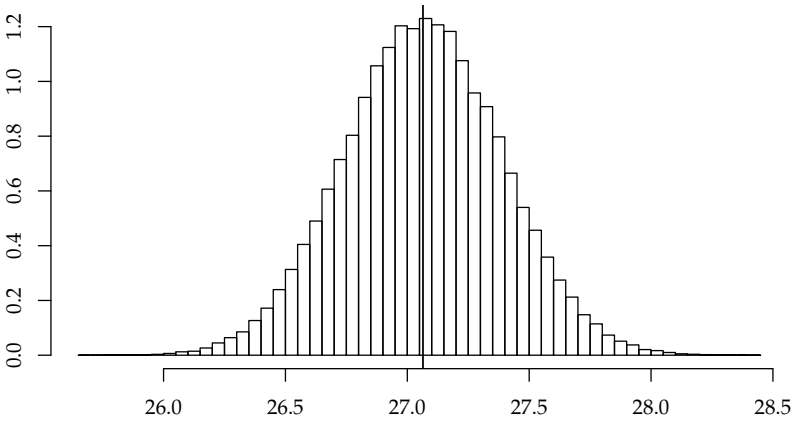

Figure 4.5: Forecasting distribution of the cohort remaining life expectancy $\bar{e}^{\mathrm{c}}(2005, 60)$ for Danish women. The mean and median (vertical line) are both 27.06, and the standard deviation is 0.32. Based on 100,000 simulations.

pirical mean is 27.06 which is very close to the point estimate of 27.05 in Table 4.2. Note that this need not necessarily be true in general for non-linear functionals. Confidence intervals can be obtained using either the empirical standard deviation of 0.32 and a normal approximation, or the percentiles of interest can be calculated directly from the sample.

4.4 Goodness-of-fit

Evaluation of a statistical model's performance is of uttermost importance, and hence we devote this section to investigating the fit of the SAINT model. To this end we evaluate the model's fit within-sample as well as out-of-sample. Having applications in mind we emphasise the latter. Our benchmark is the widespread "Poisson version" of the Lee–Carter (LC) model, cf. Brouhns *et al.* (2002). This model assumes that

$$D(t,x) \sim \text{Poisson}\left(\bar{\mu}(t,x)E(t,x)\right) \text{ with } \bar{\mu}(t,x) = \exp(\alpha_x + \beta_x \kappa_t), \quad (4.35)$$

where the parameters are subject to the constraints $\sum_t \kappa_t = 0$ and $\sum_x \beta_x = 1$ to ensure identifiability.

To assess the performance we measure the cellwise errors in death counts, i.e. the deviation from the expectation in the hypothesised Poisson distribution. We then sum these, their absolute values or their squares. For comparison and interpretation we normalise by the total number of deaths in the two former cases. Thus we consider

$$\begin{aligned} G_1 =& \sum_{t,x} \left(D(t,x) - \bar{\mu}(t,x)E(t,x)\right) \Big/ \sum_{t,x} D(t,x) \\ =& 1 - \sum_{t,x} \bar{\mu}(t,x)E(t,x) \Big/ \sum_{t,x} D(t,x) \,, \end{aligned} \tag{4.36}$$

$$G_2 = \sum_{t,x} \left|D(t,x) - \bar{\mu}(t,x)E(t,x)\right| \Big/ \sum_{t,x} D(t,x) \,, \tag{4.37}$$

$$G_3 = \sum_{t,x} \left(D(t,x) - \bar{\mu}(t,x)E(t,x)\right)^2 , \tag{4.38}$$

where $\bar{\mu}$ is either fitted or forecasted. The forecasted values for the SAINT model (LC model) are based on the mean forecast of the spread (κ-index).

Notice that G_1 is the weighted (by actual deaths) average of $1 - \bar{\mu}(t,x)/m(t,x) \approx \log\left(m(t,x)/\bar{\mu}(t,x)\right)$. The weighted average of the latter therefore being a comparable measure of fit. In comparing some versions of the LC method Booth *et al.* (2006) calculated the *unweighed* averages of $\log\left(m(t,x)/\bar{\mu}(t,x)\right)$ and its absolute value. Modulo a log approximation this corresponds to weighing by $\left(\bar{\mu}(t,x)E(t,x)\right)^{-1}$ in (4.36), which seems unreasonable considering the likelihood. Apart from these suggestions, most mortality models we have encountered base their evaluation of fit on graphical inspection.

G_2 and G_3 evaluate the fit in two different ways and we will use both measures. As a supplement we also use G_1 to measure overall bias. Alternatively, one could evaluate how well key figures such as annuity values and remaining life expectancy match, but we do not pursue that here.

	Women		Men	
Estimation period	SAINT	Lee–Carter	SAINT	Lee–Carter
1933–1950	5.08	5.38	5.30	5.91
1933–1970	5.02	5.55	5.39	6.09
1933–1990	6.00	6.55	5.00	5.78
1933–2005	6.25	6.31	4.89	5.96

Table 4.3: Within-sample error measured by G_2 (as percentages). Danish data.

4.4.1 Within-sample performance

Within-sample both models perform well with an absolute relative error of about 6%. Table 4.3 reveals that there is not much to choose between the two, but it is encouraging that the more sparsely parameterised SAINT model fits at least as good over any subperiod considered. We choose to use G_2 because it is comparable across sexes and periods, but G_3 gives the same conclusion. Recently, Dowd *et al.* (2008b) came up with some statistically testable suggestions to measure goodness–of–fit for mortality models, which we shall not go into – partly because we are somewhat skeptical of the underlying independence and dispersion assumptions.

4.4.2 Out-of-sample performance

The papers we are aware of are to a large extent silent regarding the out-of-sample performance of their respective models. Two exceptions are Booth *et al.* (2006) as mentioned above, and Dowd *et al.* (2008a), whose ideas are appealing. For sake of brevity we shall not perform any of their four suggested inspections in depth here.

In essence we believe that a good mortality forecast must fulfill two equally important criteria. First, the model should provide accurate forecasts over short and long horizons. And secondly, using different input data the mortality forecasts ought to be as little sensitive towards the choice of estimation period as possible, i.e. robust.

	Forecast period length					
	10		15		20	
Estimation period	SAINT	LC	SAINT	LC	SAINT	LC
1933–1950	0.69	0.43	0.93	0.69	1.34	1.31
1933–1970	1.48	2.32	2.39	4.27	3.50	6.96
	Forecast period length					
	25		30		35	
Estimation period	SAINT	LC	SAINT	LC	SAINT	LC
1933–1950	2.43	2.81	4.26	5.34	6.44	8.58
1933–1970	4.82	9.70	6.22	13.0	7.27	15.7

Table 4.4: Out-of-sample error for different estimation periods and different forecast periods measured by G_3 (normalised by 10^6). Danish women.

Accuracy

To evaluate the accuracy we will calculate G_3 for forecast horizons ranging from 10 to 35 years for two different, but overlapping, estimation periods. The results are displayed in Table 4.4 from which we conclude (albeit based on limited evidence and overlapping estimation periods) that the LC method predicts slightly more accurately over short forecast periods, whereas on long horizons the SAINT model's performance is superior. Further analysis has indicated that the tipping point lies between about five and 15 years' forecast.

At the cost of potentially even worse long run forecasts it has been suggested to improve the short run accuracy by calibrating the Lee–Carter model to the latest observed death rates. This would likely reinforce the difference between the two models.

Very short term forecasts (not shown) are quite accurate in both cases—because of short term smoothness of data and the relatively dense parametrisation. All conclusions above apply to men as well.

For the same estimation and forecast periods we have calculated G_1 as a measure of bias. The results are shown in Table 4.5. Note that negative values imply upward bias of death rates, i.e. projected death rates

	Forecast period length					
Estimation	10		15		20	
period	SAINT	LC	SAINT	LC	SAINT	LC
1933-1950	-7.97	-1.64	-6.84	-0.62	-6.36	-0.33
1933-1970	-4.86	-3.28	-3.64	-2.35	-2.10	-1.21
	Forecast period length					
Estimation	25		30		35	
period	SAINT	LC	SAINT	LC	SAINT	LC
1933-1950	-6.82	-1.04	-6.97	-1.51	-6.49	-1.39
1933-1970	+0.33	+0.80	+1.83	+1.93	+2.78	+2.53

Table 4.5: Out-of-sample error for different estimation periods and different forecast periods measured by G_1 (as percentages). Danish women.

are higher than realised death rates. For the short estimation period the SAINT model has a higher bias than LC, while the bias of the two models is essentially the same for the long estimation period.

Examining the contributions of G_1 more thoroughly (numbers not shown) reveals that there is a seemingly systematical pattern in the errors of the SAINT model with most of the upward bias concentrated at ages below 40. Death rates for ages 40–60 are in fact slightly downward biased, while death rates for ages above 60 are essentially unbiased. Due to its structure the LC model suffers no such systematic age bias.

Dowd *et al.* (2008a) point out that most mortality forecasts are upward biased. We find the same, but of course this is not an intrinsic feature of the models.

Robustness

We will check robustness in two ways—by examining the stability of forecasts towards the inclusion of additional years in each end of the data window. This serves two distinct purposes. Adding extra years in the "left" side of the interval we examine the sensitivity towards the otherwise arbitrary choice of left end point of the input data. On the other hand

adding years in the "right" side allows analysis of the desired feature that forecasts do not change substantially when the model is calibrated using new data. Of course these two tests are closely linked. For the sake of brevity we will provide graphical indications only but obviously a version of the G_3 measure, or something similar, should be considered as well to assess how close forecasts based on different data are.

For the former analysis see Figure 4.6. This graphical inspection clearly indicates that the SAINT model is more "backward robust". The particular evidence is based on two scenarios only, but in fact the conclusion applies to other ages and periods and to Danish men as well.

Finally, we consider the stability towards including new data. This is essentially no different from the preceding analysis, and the conclusion is repeated from above. Figure 4.7 compares mortality intensity forecasts at two key ages and suggests that the SAINT model is slightly more "forward robust". This conclusion is also representative across sexes, ages, and estimation periods.

We do not believe in the existence of an intrinsically optimal length for the sample period. Hence, we do not investigate this. Instead we have faith in the underlying model and use as much data as possible whenever it is deemed being of an acceptable quality.

As a closing remark we note that any full evaluation of the out-of-sample performance should take the entire fitted and forecasted distribution into account, cf. Dowd *et al.* (2008a). At first glance our model seems to provide reasonably wide distributions on both short and long horizons, thus nicely accompanying the reasonable forecasts. Presently we do not, however, elaborate further on this.

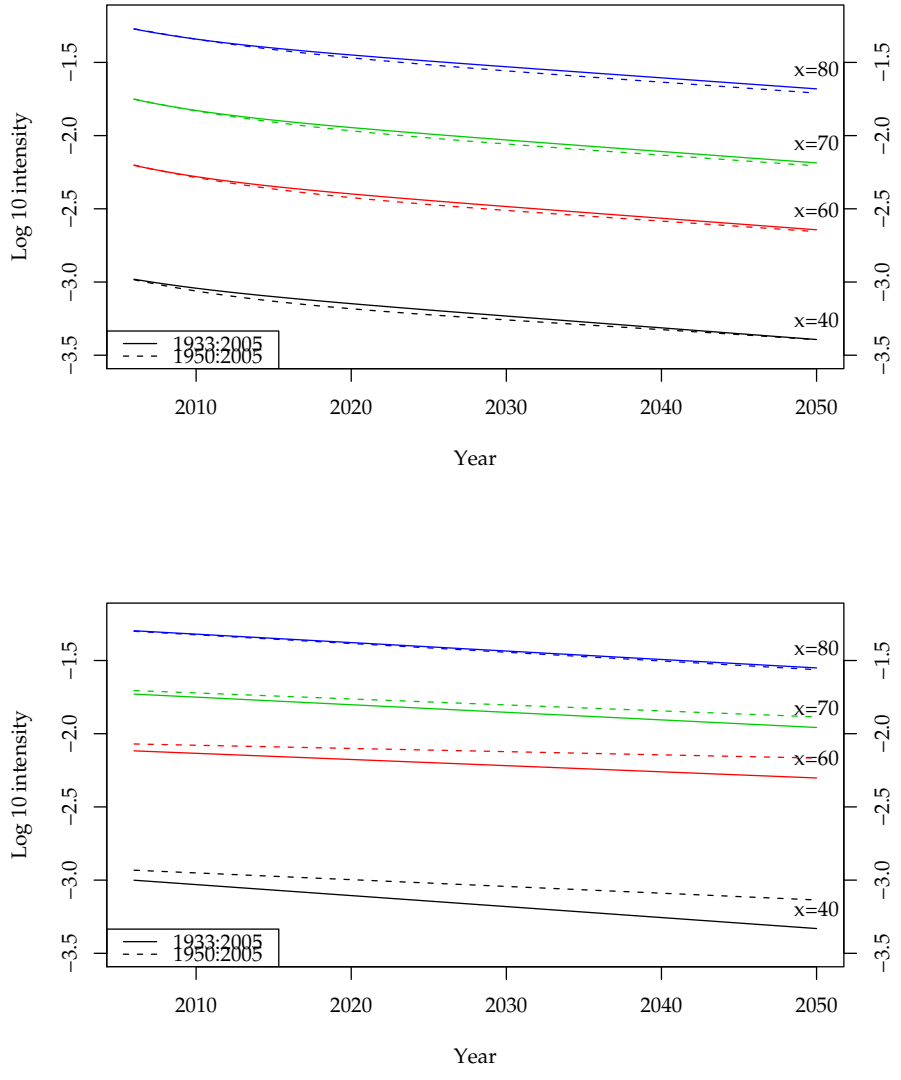

Figure 4.6: Mortality intensity forecasts based on two different estimation periods. Upper panel: SAINT model. Lower panel: Lee–Carter model. Danish women.

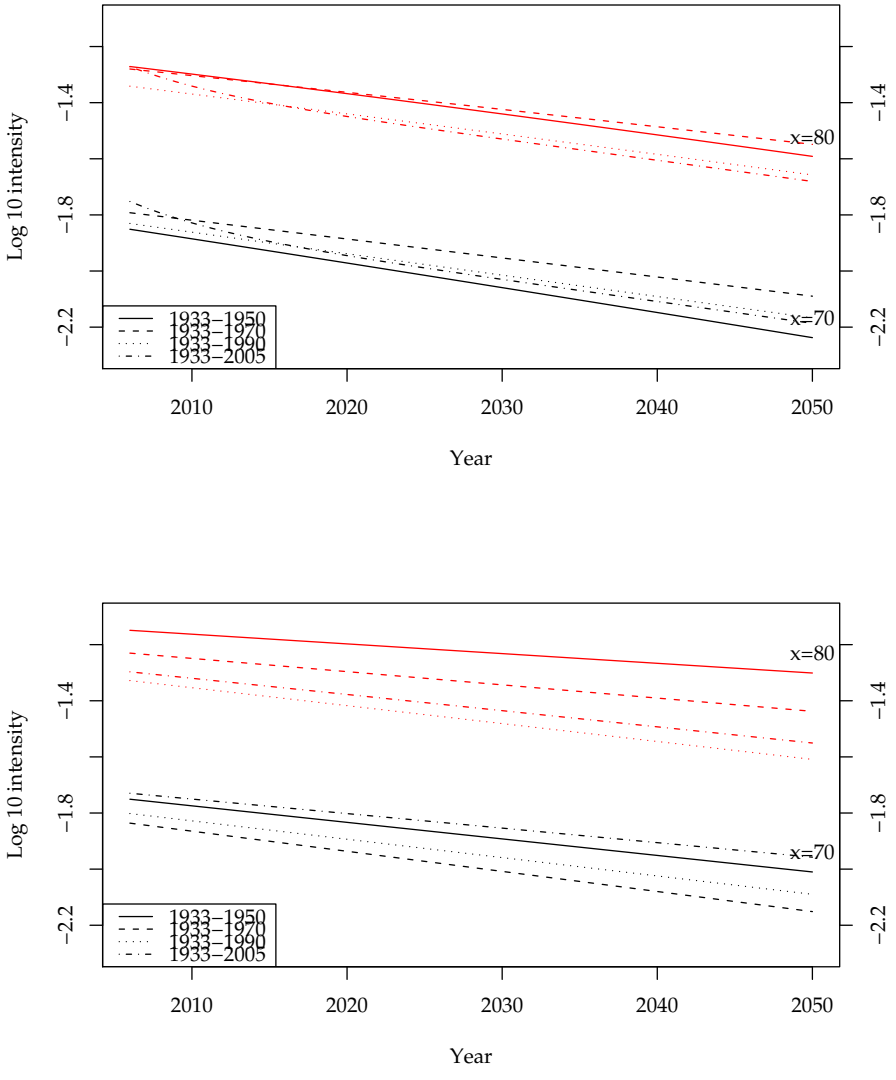

Figure 4.7: Mortality intensity forecasts based on four different estimation periods. Upper panel: SAINT model. Lower panel: Lee–Carter model. Danish women.

4.5 Concluding remarks

The mainstream in mortality modelling builds on linear time series of unobserved underlying random processes. This typically works very well when the population in question is sufficiently large that realised death rates are smooth, in particular over relatively short forecast horizons. Over longer horizons, and for small populations on the other hand the performance of these models is less convincing, and the estimates may be very sensitive to the choice of input data. We therefore developed a two-step approach—modelling first the mortality of a larger reference population, then the mortality spread between the two populations.

We have left the choice of reference population a subjective one. The reference population should be related to the population of interest as we have to believe that the two populations share the same long term trend. Observing this, we recommend to choose it as large as possible for best identification of the trend. The analysis obviously depends on the choice of reference population, but in this respect the choice of reference population is no different from the choice of estimation window, or indeed the choice of model.

In the presented model we have focused on forecasting a single population. However, the methodology can easily be extended to produce coherent, i.e. non-diverging, forecasts for a group of related populations by using the group as reference population and treat each population as a subpopulation of the group. Similarly, we could consider men and women as subpopulations of the same population rather than estimate separate models for each gender as done in the application. Using common and individual components to produce coherent forecasts for related populations in the Lee–Carter framework has been suggested by Li and Lee (2005).

We have used a two-stage estimation routine in which we first estimate the trend parameters and then estimate the spread parameters with the trend kept fixed. This approach can be justified when the reference population is substantially larger than the subpopulation, as in case of Danish and international data. For applications in which the reference and subpopulation are of comparable size one might consider to estimate the trend and the spread jointly. It is straightforward to write down the likelihood function so in principle this is possible, but it is numerically involved due to the large number of parameters involved.

The trend component of the SAINT model imposes structure on how mortality can evolve over time and across ages. The parametric form provides insight into the improvement patterns and it guarantees biologically plausible forecasts. Compared to the Lee–Carter model the structure of the SAINT model lead to more precise long–term forecasts at the price of higher bias. The higher bias was primarily for young ages which is not surprising as the focus of our modelling has been on old age mortality. The bias at young ages could undoubtedly be reduced by more careful modelling of these age groups if so desired.

We have concentrated on the uncertainty generated by the stochastic term of the spread model, and only briefly mentioned the possibility of including parameter uncertainty. In the case of the trend, however, the latter source of uncertainty would be negligible since the trend parameters are so well-determined. Another possibility which might better reflect the uncertainty of the trend would be to introduce stochastic terms in the trend model also, thus treating both the trend and the spread as stochastic processes. This however is not straightforward since we wish to preserve the overall structure of the trend. We believe that constructing confidence intervals which properly reflect the intrinsic uncertainty of mortality projections is an important topic which calls for more attention than so far received.

4.6 Background and proofs

4.6.1 Background

Recall that the density of the Γ-distribution with shape parameter $\lambda > 0$ and scale parameter $\beta > 0$ is given by

$$f(z) = \frac{\beta^{-\lambda}}{\Gamma(\lambda)} z^{\lambda-1} \exp\left(-z/\beta\right), \ (z \geq 0). \tag{4.39}$$

This distribution has mean $\beta\lambda$ and variance $\beta^2\lambda$. Letting $\lambda = \beta^{-1} = \Sigma^{-2}$ we obtain a Γ-distribution with mean 1 and variance Σ^2.

The survival function, $\bar{F}(t, x)$, denotes the proportion of the cohort born at time $t - x$ still alive at time t (at age x). Similarly, the individual survival function, $\bar{F}(t, x; z)$, denotes the probability that a person with

frailty z born at time $t-x$ is still alive at time t. If f denotes the density of the frailty distribution at birth we have

$$\bar{F}(t,x) = \int_0^\infty \bar{F}(t,x;z) f(z) dz, \tag{4.40}$$

while the conditional frailty density at time t for persons of age x is given by

$$f(z|t,x) = \frac{f(z)\bar{F}(t,x;z)}{\bar{F}(t,x)}. \tag{4.41}$$

The survival function can be expressed in terms of the force of mortality as

$$\bar{F}(t,x) = \exp\left(-\int_0^x \mu(u+t-x,u)du\right), \tag{4.42}$$

and, conversely,

$$\mu(t,x) = \left[-\frac{d}{d\delta}\log \bar{F}(t+\delta, x+\delta)\right]_{|\delta=0}. \tag{4.43}$$

The same relationships hold for $\bar{F}(t,x;z)$ and $\mu(t,x;z)$.

4.6.2 Proofs

Proof of Example 4.1. The first equality in (4.10) follows from Proposition 4.2 with $\mu_s^I(t,x) = \alpha\exp(\beta x)$ and $\gamma(t) = \gamma$, and the second equality follows from Proposition 4.3. □

Proof of Proposition 4.2. By (4.43), (4.40), (4.41) and (4.11) the population mortality satisfies

$$\begin{aligned}
\mu(t,x) &= \frac{\left[-\frac{d}{d\delta}\bar{F}(t+\delta,x+\delta)\right]_{|\delta=0}}{\bar{F}(t,x)} \\
&= \frac{\int_0^\infty f(z)\left[-\frac{d}{d\delta}\bar{F}(t+\delta,x+\delta;z)\right]_{|\delta=0} dz}{\bar{F}(t,x)} \\
&= \frac{\int_0^\infty f(z)\bar{F}(t,x;z)\mu(t,x;z)dz}{\bar{F}(t,x)} \\
&= \int_0^\infty f(z|t,x)\left(z\mu_s^I(t,x)+\gamma(t)\right)dz \\
&= \mathbb{E}\left\{Z|t,x\right\}\mu_s^I(t,x) + \gamma(t).
\end{aligned}$$

□

Proof of Proposition 4.3. Using (4.11), the individual survival function can be written

$$\bar{F}(t,x;z) = \exp\left(-zI(t,x) - G(t,x)\right), \tag{4.44}$$

where $I(t,x) = \int_0^x \mu_s^I(u+t-x,u)du$ and $G(t,x) = \int_0^x \gamma(t-x+u)du$. Inserting (4.44) in (4.40) with f given by (4.39) for $\lambda = \beta^{-1} = \Sigma^{-2}$ we get

$$\begin{aligned}\bar{F}(t,x) &= \frac{\lambda^\lambda}{\Gamma(\lambda)} \int_0^\infty \exp\left(-z\left[\lambda + I(t,x)\right]\right) z^{\lambda-1} dz \exp\left(-G(t,x)\right) \\ &= \left(\frac{1}{1+\Sigma^2 I(t,x)}\right)^{1/\Sigma^2} \exp\left(-G(t,x)\right). \end{aligned} \tag{4.45}$$

Finally, inserting (4.44) and (4.45) in (4.41) with f as above we obtain

$$f(z|t,x) = \frac{\left(\Sigma^{-2} + I(t,x)\right)^{\Sigma^{-2}}}{\Gamma\left(\Sigma^{-2}\right)} z^{\Sigma^{-2}-1} e^{-z\left(\Sigma^{-2}+I(t,x)\right)},$$

which we recognise as a Γ-density with $\lambda = \Sigma^{-2}$ and $\beta^{-1} = \Sigma^{-2} + I(t,x)$.
□

Proof of Proposition 4.4. First note that in the notation of Proposition 4.3 we have

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right) I(t,x) = \mu_s^I(t,x),$$

and thereby

$$\begin{aligned}\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right) \log \mathbb{E}\left\{Z|t,x\right\} &= -\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right) \log\left(1+\Sigma^2 I(t,x)\right) \\ &= -\frac{\Sigma^2 \mu_s^I(t,x)}{1+\Sigma^2 I(t,x)} \\ &= -\Sigma^2 \mu_s^I(t,x) \mathbb{E}\left\{Z|t,x\right\} \\ &= -\Sigma^2 \mu_s(t,x),\end{aligned}$$

where the last equality follows from Proposition 4.2. Using $\mathbb{E}\{Z|t-x,0\} = 1$ we then get

$$\begin{aligned}\log \mathbb{E}\{Z|t,x\} &= \int_0^x \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right) \log \mathbb{E}\{Z|t-x+u,u\}\, du \\ &= -\Sigma^2 \int_0^x \mu_s(t-x+u,u)du.\end{aligned}$$

□

Proof of Proposition 4.5. The equality in (4.24) follows from (4.17) and the specification of μ_s^I. To show convergence we first use Proposition 4.3 to write

$$\frac{\partial \log \mathbb{E}\{Z|t,x\}}{\partial t} = -\frac{\Sigma^2 \frac{\partial}{\partial t} I(t,x)}{1+I(t,x)}, \tag{4.46}$$

where $I(t,x) = \int_0^x K(t-x,y)dy$. By (4.23) and dominated convergence we have

$$I(t,x) = \int_0^x \exp\left((\kappa_2 + g_2 y)t\right) K(-x,y)dy \to 0 \text{ for } t \to \infty,$$

since $\kappa_2 + g_2 y < 0$ for all $0 \leq y \leq x$ by assumption. Similarly,

$$\frac{\partial}{\partial t} I(t,x) = \int_0^x \exp\left((\kappa_2 + g_2 y)t\right)(\kappa_2 + g_2 y) K(-x,y)dy \to 0 \text{ for } t \to \infty,$$

and we conclude that (4.46) also converges to zero. □

Proof of Proposition 4.6. First case follows by applying l'Hôpital's rule to (4.22), second case is the Gamma-Makeham model, and third case is trivial. □

4.7 Data summary

		Male		Female	
Country	Period	Deaths	Exposure	Deaths	Exposure
Australia	1933-2004	3656554	298774817	3050957	304038560
Austria	1947-2005	2362681	146948386	2544281	172938922
Belgium	1933-2005	3959202	237960436	3655934	254715105
Canada	1933-2005	5818067	501315204	4711059	510104624
Denmark	1933-2005	1736235	118573885	1651576	124503893
England & Wales	1933-2005	18790451	1155893702	18906239	1301265668
Finland	1933-2005	1649803	105805700	1493332	118540078
France	1933-2005	19554792	1197266420	18601741	1332188094
Iceland	1933-2005	50595	4531626	46223	4597037
Italy	1933-2004	17840263	1231244196	16616547	1360392086
Japan	1947-2005	22944373	2115001379	19852118	2274227259
Netherlands	1933-2005	3785909	304422103	3442716	316942482
Norway	1933-2005	1326283	94013584	1240955	98579501
Portugal	1940-2005	2859456	189334582	2769074	216359565
Spain	1933-2005	10815395	785643871	9940870	862709017
Sweden	1933-2005	2969115	202088902	2792398	209883920
Switzerland	1933-2005	1949244	145118601	1888688	159060151
USA	1933-2005	67513137	4756496475	58192508	5127612213
West Germany	1956-2005	16193960	1081937335	17307798	1227848814

Table 4.6: Summary of deaths and exposures for ages 20–100 over the period 1933–2005 for the countries included in the international data set. Source: The Human Mortality Database (www.mortality.org).

5. The evolution of death rates and life expectancy in Denmark

Background. The paper in this chapter is written jointly with Søren Fiig Jarner and Chresten Dengsøe. It was initiated by invitation from the Scandinavian Actuarial Journal, and appeared as Jarner *et al.* (2008) as part of a special longevity issue of the journal. The paper has been presented on various occasions, including the mortality risk course arranged by the Danish Actuarial Society in December 2008 held in Copenhagen.

Abstract. From 1835 till today Denmark has experienced an increase in life expectancy at birth of about 40 years for both sexes. Over the course of the last 170 years life expectancy at birth has increased from 40 years to 80 years for women, and from 36 years to 76 years for men, and it continues to rise.

Using a new methodology we show that about half of the total historic increase can be attributed to the sharp decline in infant and young age death rates up to 1950. Life expectancy gains from 1950 till today can, on the other hand, primarily be attributed to improvements in the age–specific death rates for the age group from 50 to 80, although there is also a noticeable contribution from the further decline in infant mortality over this period. With age–specific death rates up to age 60 now at a very low absolute level substantial future life expectancy improvements must necessarily arise from improvements in age–specific death rates for ages 60 and above. Using the developed methodology we quantify the impact of further reductions in age–specific mortality.

Despite being one of countries with the highest life expectancy at the beginning of the 20^{th} century and despite the spectacular historic increase in life expectancy since then Denmark, in fact, is lacking behind compared to many other countries, notably the other Nordic countries. The main reason being an alarming excess mortality for cause–specific death rates related to ischaemic heart diseases and, in particular, a number of cancer diseases. Age–specific death rates continue to improve in most countries, and a likely scenario is that Denmark in the future will experience improvement rates at the international level or perhaps even higher as a result of a catch–up effect.

5.1 Introduction

Death is, if not defeated, then on retreat and has been so for a very long time indeed. The most striking feature of the evolution of death rates in Denmark, and many other countries, is the constant improvement over at least the last two hundred years for which we have reliable data. Moreover, there is no sign of mortality rates levelling out or improvement rates even slowing down in any near future.

The continuing mortality improvements have far–reaching consequences for pension funds and the future financing of public health care and state pension system. Being tax–paid the challenges for public financing are aggravated by low fertility rates causing the ratio between old people and young people to increase over time. In Denmark this development has not gone unnoticed and the heavy implications for the future financing of the so–called Danish welfare system are analysed in Danish Welfare Commission (2004), which also provides recommendations for easing the future financing burden. The analysis rests, among other things, on a mortality forecast based on the Lee–Carter methodology which predicts a very modest increase in life expectancy of about 4 years for both sexes over the next 50 years, Haldrup (2004).

One of the recommendations of the Danish Welfare Commission has recently been implemented. The state pension retirement age will in the future follow the development in life expectancy. The current retirement age of 65 will be in effect till 2023 and then increase with six months each year to 67 in 2027. From then on future life expectancy gains will be reflected in a similar increase in the retirement age.

The majority of Danish pension plans consists of funded systems. These pension funds are facing two distinct problems in relation to falling mortality rates. The first problem concerns providing adequate reserves for annuity contracts already entered. In most contracts currently in effect the terms, including mortality assumptions, cannot be changed over the course of the contract. Many of these contracts are based on the, at the time, conservative technical basis G82 which has an assumed life expectancy of about 77 years for women and about 73 years for men. As reality has overtaken these assumptions a funding problem has arisen.

The second problem concerns assessing the future mortality pattern in order to base new contracts on more robust mortality assumptions. From

the point of view of the pension industry this problem can to some extent be handled through securitisation. There has been some academic progress in this area, see e.g. Dahl *et al.* (2008); Cairns *et al.* (2005); Lin and Cox (2005), but the market for longevity bonds and related products is still in its infancy. Another popular approach is to avoid the problem altogether by changing the products into "mortality free" savings products, or into annuities in which the mortality assumptions can be changed. However, the basic problem of assessing the future length of the retirement period still persists.

Despite the overall picture of constantly declining mortality rates there has been periods with no improvements or even slight increases in mortality for certain age groups. One such period lasted from around 1980 to 1995 during which life expectancy rose by only about 1 year. This period with almost stagnation in life expectancy was also observed to some extent in other countries but it was more pronounced in Denmark. In fact, the slower Danish pace of improvement called for political action and in 1992 the Life Expectancy Commission was formed by the Danish Ministry of Health. In a series of reports they documented Danish excess mortality for a number of heart and cancer diseases, see e.g. Life Expectancy Commission (1998). For related work see the very comprehensive, descriptive analysis of the evolution of Danish mortality in Andreev (2002) and the discussion in Hansen *et al.* (2006).

The purpose of the present paper is to provide an overview of the evolution of age–specific death rates and to explore the link between improvements in age–specific death rates and life expectancy gains. Life expectancy (at birth) is a much used statistic used to summarise a given life table in a succinct way. Normally, this quantity is calculated for a given period life table, i.e. the mortality pattern of a population in a given calender year, and used in this way it conveys information about the general level of mortality in the population at that instant in time. It does not, however, correspond to the expected life time of a newborn, except in the hypothetical situation with no future improvements in death rates.

We develop a new sensitivity measure which relates improvements in age–specific death rates to life expectancy gains (the measure corresponds to the functional derivative of life expectancy with respect to the mortality curve). Using this tool we can quantify the historic and future contribu-

tions to life expectancy gains by the different age groups and quantify statements like "life expectancy gains used to be caused by falling infant mortality, but are now due to improvements in old–age mortality".

We will also contrast the evolution of Danish mortality with the international development and compare cause–specific mortality rates for selected countries.

5.2 Data and notation

Data used in the paper originate from the Human Mortality Database[1] (HMD), which offers free access to updated records on death counts and exposure data for a long list of countries. The database is maintained by University of California, Berkeley, United States and Max Planck Institute for Demographic Research, Germany.

The data consists of gender specific death counts, $D(t, x)$, and the corresponding exposures, $E(t, x)$, for a range of years t and ages x. More precisely, $D(t, x)$ counts the number of deaths occurring in calender year t among people aged x last birthday, and $E(t, x)$ gives the total number of years lived during calender year t by people of age x. For readers familiar with the Lexis diagram, $D(t, x)$ counts the number of deaths in the square $[t, t + 1) \times [x, x + 1)$ of the Lexis diagram and $E(t, x)$ is the corresponding exposure.

For Denmark we are fortunate to have data from 1835 onwards making the Danish series one of the longest data series available in HMD. Recent population data for Denmark is of a very high quality being based on the Central Population Register (CPR). The register was introduced in 1968 and used for the first time in the 1976 census. Before that time censuses were held every five years with varying levels of detail in the recording of ages. The pre–1976 data is therefore based on interpolation in both the time and age dimension and also extrapolation at high ages, since the maximal age recorded has varied over time. The interested reader is referred to Andreev (2002) and Wilmoth *et al.* (2005) for a detailed account of the structure of the underlying data and the methods used to create the HMD data series.

[1] See www.mortality.org

From the death counts and exposures we form the (crude) death rates

$$m(t,x) = \frac{D(t,x)}{E(t,x)} \quad \text{for } t = 1835,\ldots,2006,\ x = 0,\ldots,99, \tag{5.1}$$

which will form the basis of our analysis. Death counts and exposures are available also for ages $100,\ldots,109,110+$, where the last age is an open–ended interval covering age 110 and above. However, as these data are very noisy and, for the early part of the series, constructed from age 100+ data we choose to ignore these also for more recent years. Hence, we let "$m(t,x) = \infty$" for $x \geq 100$ and all t, meaning that the highest attainable age is 100 years. Life expectancy at birth is only marginally influenced by this assumption even in recent years, but the life expectancy at very old ages will be slightly underestimated.

5.2.1 Force of mortality

The original data is aggregated over calender years and age groups of one year. However, it turns out to be convenient to work with a continuous formulation, in particular, when discussing the sensitivity of life expectancy to changes in mortality.

For an individual with (continuous) life time T the survival function is defined as $\bar{F}(x) = \mathbb{P}(T > x)$, i.e. the probability that the person will live longer than x years. The *force of mortality*, also called the *intensity* or the *hazard*, is defined as

$$\mu(x) = -\frac{d}{dx}\log \bar{F}(x) = \frac{f(x)}{\bar{F}(x)}, \tag{5.2}$$

where f denotes the density of the life time distribution. The force of mortality can be interpreted as the instantaneous death rate immediately after age x given survival to age x.

The survival function and the conditional survival functions given survival to age y can be expressed in terms of the intensity as

$$\bar{F}(x|y) = \mathbb{P}(T > y + x | T > y) = e^{-\int_y^{y+x}\mu(u)du},$$

and the expected remaining life time for an y–year–old can be expressed as

$$\bar{e}_y = \mathbb{E}\left\{T - y | T > y\right\} = \int_0^\infty \bar{F}(x|y)dx = \int_0^\infty e^{-\int_y^{y+x}\mu(u)du}dx.$$

Of particular interest to us is the quantity $\bar{e}_0$, life expectancy at birth.

We shall be using these formulas when calculating survival probabilities and expected life times for period life tables, $\{m(t,y)\}_{y=0,\ldots,100}$, and in order to do so we define an intensity function by $\mu_t(x) = m(t,y)$ for $x \in [y, y+1)$, and base the calculations on μ_t. Thus when speaking of the life expectancy at year t we mean the quantity $\bar{e}_0(\mu_t)$, and similarly for other summary statistics related to year t.

All reported life expectancies in the paper are calculated using this procedure based on period life tables calculated from HMD data truncated at age 100. Our life expectancies might therefore deviate slightly from numbers published by the various national bureaus of statistics, e.g. Statistics Denmark.

5.3 Death rates

By and large death rates have been constantly improving in the historic period considered, see Figure 5.1. However, looking at the evolution of age–specific death rates in more detail reveals a more subtle structure with great variability in both the pace of improvements over time and the age groups affected. In the following we will give an account of the age–specific improvements and we will relate these to life expectancy gains using the decomposition technique developed in Section 5.4.

5.3.1 Infant and child mortality

In 1835 infant mortality was about 19% for girls and 24% for boys; thus only four out of five babies survived their first year of living. Young age mortality was also very high and life expectancy at birth at that time was only about 40 years for women and 36 years for men. During the late 19th century infant mortality declined somewhat but it still remained very high. At the turn of the century infant mortality had fallen to 13% for girls and 16% for boys. At the same time life expectancy had risen to 53 years for women and 50 years for men. However, this life expectancy gain can be contributed mainly to the decline in child mortality (between age 1 and 10) which dropped by 70% from 1835 to 1900.

Figure 5.2 shows that the reduction in child mortality contributed with

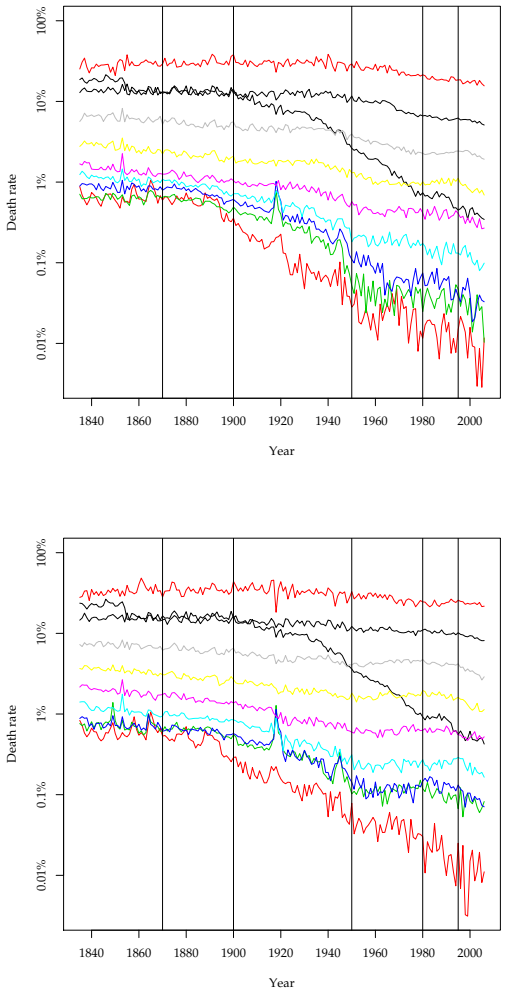

Figure 5.1: Development in Danish female (top) and male (bottom) age–specific death rates from 1835 till 2006 for ages $0, 10, \ldots, 90$. The line starting out as the second highest and crossing several others is the death rate for 0–year–olds. The other lines represent ages 10 to 90 in increasing order.

about twice as much as the reduction in infant mortality to the life expectancy gain over the period. The reader is referred to Section 5.4 for a thorough description of how the age–specific life expectancy contributions in Figure 5.2 are calculated.

The turn of the century marked the beginning of uninterrupted improvements in infant mortality. From 1900 onwards infant mortality has been declining with about 3% each year. In 1950 infant mortality had dropped to 3% for girls and 4% for boys. During the first half of the 20th century life expectancy rose to 72 years for women and 69 years for men; about one third of the increase can be attributed to the reduction in infant mortality. The improvements in infant mortality continued after 1950 and in 2006 infant mortality had been further reduced to 0.3% for girls and 0.4% for boys. The life expectancy in 2006 was 80 years for women and 76 years for men. However, only a smaller part of the life expectancy increase after 1950 was due to reductions in infant mortality.

The historic development in infant mortality has been truly remarkable. Coming from a level in 1835 corresponding to the death rate of an 80–year–old the level has been dramatically reduced in both absolute and relative terms. In 2006 infant mortality was at the same level as the mortality of an 50–year–old. Infant mortality seems to continue to fall, however, being now at a low absolute level future reductions will not have a great impact on life expectancy.

The evolution of child mortality is very similar to the development in infant mortality with respect to the timing and size of improvement rates. This can be seen from Figure 5.1 where the curves for infant and 10–year–old's mortality have developed almost in parallel. However, the improvements in child mortality started already around 1870, some 30 years before infant mortality began to decrease.

Child mortality in 1835 ranged from 6% for 1–year–old boys and girls to 1% for 10–year–old boys and girls. By 1900 these death rates had been reduced to about 2% for 1–year–olds and 0.3% for 10–year–olds. About half of the life expectancy gain over this period can be attributed to this reduction. Apart from the fall in infant mortality from 1900 to 1950 the reduction in child mortality in the late 19th century is the single most important contribution to the increase in life expectancy from 1835 to 2006.

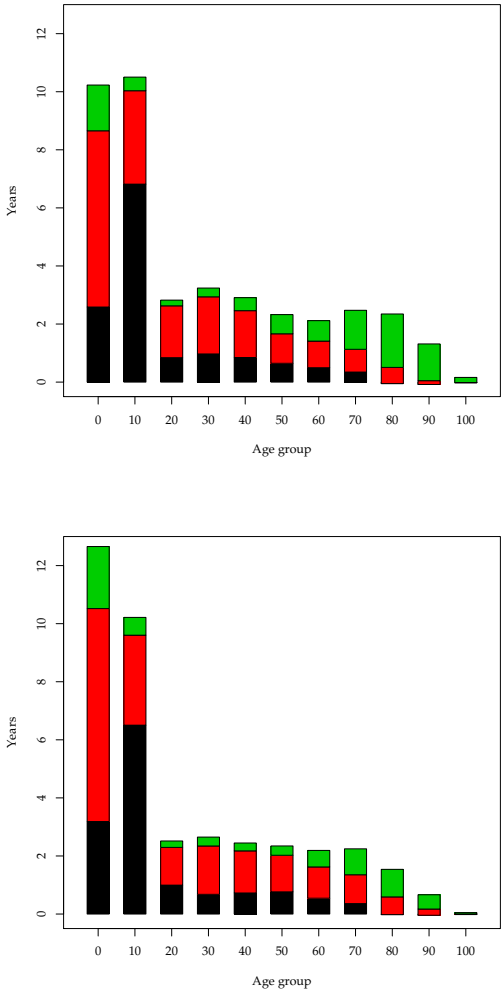

Figure 5.2: Decomposition of the life expectancy gain from 1836 to 2006 for Danish females (top) and males (bottom). The plot shows the life expectancy gain attributable to age groups 0, 1–10, 11–20, . . . , 91–100 over the three periods 1835–1900 (bottom boxes), 1900–1950 (middle boxes) and 1950–2006 (top boxes); see Section 5.4.2 for details.

From 1900 onwards child mortality steadily declined to below 0.3% in 1950 for both boys and girls and further down to below 0.04% in 2006. The fall from 1900 to 1950 had an appreciable effect on life expectancy of about 3 years, while the impact in terms of life expectancy gains of the fall in the second half of the 20th century has been minimal. As for infant mortality child mortality is now at a very low level and further reductions will have only marginal effects on life expectancy.

In combination more than half of the 40 years gain in life expectancy since 1835 can be attributed to improvements in infant and child mortality.

5.3.2 Young and middle age mortality

The evolution of young (ages 11 to 30) and middle age (ages 31 to 60) mortality exhibits considerably variation between the sexes and across different time periods and ages. Whereas the evolution of infant and child mortality shows a fairly regular pattern with a steady decrease of the same magnitude for both sexes since the late 19th century, the improvement in young and middle age mortality is characterised by periods with high rates of improvement and periods with virtually no improvements.

Generally, the improvement rate is decreasing with age. Even for the periods with high rates of improvement in young and middle age mortality these rates are below those observed for infant and child mortality. This also holds true for old age (above age 60) mortality which has had the lowest rates of improvement of all age groups; a point to which we shall return below.

We have informally, i.e. not based on any objective measure apart from our judgement, divided the historic period into six subperiods (indicated by the vertical lines in Figure 5.1) in which the age–specific rates of improvement appear approximately constant. Since the death rates are plotted on a logarithmic scale this corresponds to the death rates approximately following straight lines with age–specific slopes within each subperiod.

We mention in passing that the age–specific log–linear structure within each subperiod is the assumption underlying the popular Lee–Carter model, Lee and Carter (1992). It is quite clear, however, that in case of Danish mortality estimates and forecasts based on this methodology will be very sensitive to the chosen data window as the improvement rates vary considerably over time, cf. Hansen *et al.* (2006) for a quantification of this

	Females					Males				
Period	0	10	30	60	99	0	10	30	60	99
1835–1870	1.0	1.1	0.5	0.6	-0.1	1.0	1.3	0.7	0.8	0.1
1870–1900	0.2	2.6	1.1	0.8	-0.2	0.2	2.4	0.5	0.4	-0.2
1900–1950	3.1	4.4	4.0	1.8	0.5	2.9	3.8	3.0	1.8	0.7
1950–1980	4.1	3.3	1.4	0.7	1.5	4.5	2.9	0.2	-0.1	0.2
1980–1995	3.3	2.0	2.2	0.8	0.1	3.3	3.2	1.1	1.0	0.4
1995–2006	2.4	3.9	5.3	3.4	1.7	2.5	6.5	3.3	2.2	2.2
1835–2006	2.3	2.9	2.1	1.2	0.5	2.3	2.8	1.5	1.1	0.4

Table 5.1: Median annual rates of improvements (in percentages) for selected periods and age groups: 0, 1–10, 11–30, 31–60, 61–99.

effect.

For each age x and each subperiod the constant, annual rate of improvement has been calculated, i.e. the number r_x such that $m(t,x) = m(s,x)(1-r_x)^{t-s}$ for the period from s to t. Within each subperiod and within each of the age groups 0, 1–10, 11–30, 31–60, 61–99 the median rate of improvement across ages was then calculated. The results are shown in Table 5.1. The last row contains the median rate of improvement within each age group for the whole period. We chose to use the median rather than the mean to get more robust results less sensitive to outlying rates of improvement at specific ages. As for infant mortality the rates of reduction in young and middle age mortality were relatively modest during the 19th century. The annual rate of improvement was somewhat below 1% for most age groups. Still, from 1835 to 1900 even this low rate of improvement gave rise to mortality rates being reduced by about 50%. However, the combined impact of this reduction for the ages 11 to 60 on the gain in life expectancy over this period was only 4 years or 2 years less than the effect of the decrease in child mortality alone, cf. Figure 5.2. The reason the effect was not more pronounced is that the level of infant and child mortality was still very high in 1900. More than 20% died before the age of 20 and these people did not benefit from the improved mortality rates at higher ages.

The first half of the 20th century was a period with unprecedented mortality improvements across ages. Infant and child mortality fell sharply and young age mortality also fell rapidly at a pace of 3% or more a year, with female mortality declining slightly more than male mortality. From 1900 to 1950 mortality rates were reduced by 80% or more. Also middle age mortality fell noticeably at a rate of just below 2% a year. The effect on life expectancy over this period was an impressive increase of 18 years for women and 19 years for men; life expectancy rose from 53.4 years to 71.5 years for women and from 50.1 years to 69.1 years for men. Of this gain about 7 years can be attributed to the reduction in young and middle age mortality.

The period from 1900 to 1950 saw the largest general increase in mortality, but it also witnessed some of the greatest disasters in human history. The influenza pandemic known as the Spanish Flu in the aftermath of the World War I killed somewhere between 20 and 40 million worldwide in the years 1918–1919. The effect of the influenza was particularly devastating because of its high morbidity for the ages 20 to 40, while influenza normally is most deadly for children and elderly people. The impact of the influenza is clearly visible on Figure 5.1 (the cholera epidemic in 1853 can also be identified on the plot, although, much less prominent).

The other tragic event of the period was World War II. The effect of the war can be seen as spikes in the death rates for ages 20 to 40 around 1945. Both these events caused a period with high excess mortality in young age groups, but death rates quickly fell back to previous levels from where they continued to decline. Therefore neither of the events affect the mortality improvements when measuring from 1900 to 1950.

After the strong decline in mortality for both sexes up to 1950 came a period of about 30 years in which improvement rates for young and middle age mortality diverged for females and males. Female mortality continued to decline although at a slower pace, while male mortality ceased to improve. In fact, from 1950 to 1980 male death rates slightly *increased* for ages 50 to 75.

The life expectancy gap between women and men was 3 years in 1950. A difference which had been roughly constant since 1835. As a consequence of the stagnation in male mortality improvements and the continuing female improvements this gap widened to 6 years from 1950 to 1980; over

the course of the 30 years life expectancy rose from 71.5 years to 77.2 years for women and from 69.1 years to 71.2 years for men.

Like Denmark many other developed countries also experienced a deceleration in life expectancy gains from 1950 to 1980. In Denmark, however, the slow increase was followed by almost stagnation in life expectancy, particularly for women, from 1980 to 1995, while in most other countries life expectancy improvements began to pick up again around 1980. Danish women had a life expectancy improvement of only 0.6 years over these 15 years, while male life expectancy rose by 1.6 years. The life expectancy gap between women and men thereby narrowed by 1 year to about 5 years.

Looking at Figure 5.1 and Table 5.1 we see that improvements in infant, child and young age mortality in fact persisted throughout the period, but the effect on life expectancy was hardly noticeable since death rates at these ages were already very low.

From 1995 to the time of writing Denmark has again experienced high rates of improvements, and for the first time in history death rates are improving simultaneously for all age groups including the oldest ages. The improvements, particularly in middle and old age mortality, have caused a life expectancy increase of 2.7 years for women and 3.2 years for men. Over the last decade the life expectancy gap between women and men has thus been further reduced by half a year. The life expectancy in 2006 was 80.5 years for women and 75.9 for men.

The current rate of improvement in death rates and life expectancy is historically high, only exceeded by the improvements observed from 1900 to 1950. Since 1995 there has been an average, annual increase in life expectancy of 0.24 years for women and 0.29 years for men. In comparison, there has been an average, annual increase over the whole period of about 0.23 years for both sexes. This average, of course, includes the spectacular period from 1900 to 1950 during which there was an average, annual increase of about 0.37 years for both sexes. Thus, apart from this period the average, annual increase has been considerable lower and in this perspective the current level of improvement is indeed very high.

5.3.3 Old age mortality

As mortality rates decline and life times increase the perception of "old age" also gradually changes. In 1835 with less than 40% of newborns reaching

the age of 60, cf. Figure 5.3, this was certainly considered a very old age. Nowadays, with about 90% reaching age 60 a person of this age is no longer "old", but merely an adult. However, for the sake of this paper we take old age to mean ages 61 and above as a compromise reflecting the historic period as a whole.

Through the 19th century death rates for 70–year–olds did not improve, while for even older age groups they actually increased. We should, of course, treat this finding with caution as we must consider data from this early period less reliable and to some extent prone to age misreporting, in particular, at high ages.

During the first half of the 20th century old age mortality started to improve but at a rate of less than 1% a year for males and even lower for females. These rates were much lower than those observed for the younger ages. Of the historic increase in life expectancy of about 19 years from 1900 to 1950, less than 2 years can be attributed to improvements in old age mortality, cf. Figure 5.2.

The pattern of no improvement in male mortality and continued improvement in female mortality seen from 1950 to 1980 at ages below 60 is also seen in old age mortality to an even wider degree. In this period female death rates for age 70 and 80 declined with almost 2% a year and with 1% for age 90. At the same time, male death rates for the old ages decreased only marginally or, in the case, of the 70–year–olds increased slightly.

As for the younger ages the period from 1980 to 1995 saw hardly any improvements in old age mortality. However, since 1995 old age death rates have been steadily declining and the current rate of about 2% a year is the highest in history. Of the life expectancy gain of about 13 years for both sexes from 1950 to 2006 reductions in old age mortality has contributed with about 4.5 years for women and about 2.5 years for men.

From the start of the data period till today old age mortality has in general improved at a slower pace than the younger age groups and the improvements have occurred later in time. Also the effect on life expectancy has been moderate compared to, in particular, the reductions in infant and child mortality. However, with death rates now at quite low levels up to age 60 future life expectancy gains will come almost exclusively from improvements in old age mortality.

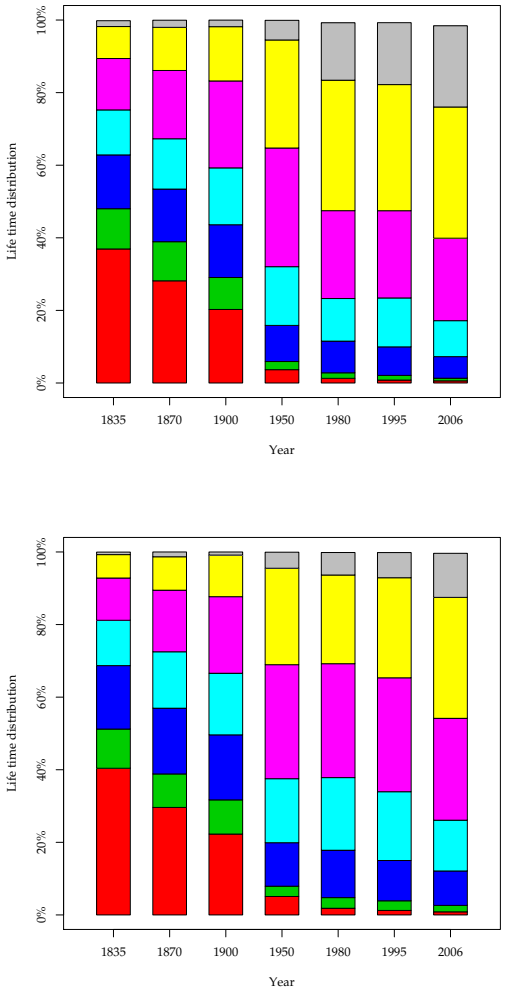

Figure 5.3: Life time distribution of Danish females (top) and males (bottom) at selected years. Each column shows the probability of dying before the age of 20, 40, 60, 70, 80, 90 and 100 based on the period life table for that year.

5.3.4 Life time distributions

Another perspective on the aggregated effect of the evolution of death rates can be obtained from looking at the life time distributions implied by period life tables. As for life expectancy calculations life time distributions based on period life tables do not represent the actual distribution of the life length of newborns (unless no further improvements occur). Still the distributions give a useful snapshot of the current mortality pattern of the population.

Figure 5.3 depicts life time distributions at selected years corresponding to the endpoints of the subperiods identified above. These years represent structural breaks in (close to) constant improvement regimes. For each year the plot shows the probability of dying before the age of 20, 40, 60, 70, 80, 90 and 100, i.e. the cumulative distribution function of life length evaluated at these ages.

The combined effect of declining death rates for infants and children is clearly visible. In 1835 the probability of dying before the age of 20 was about 40%. Over the next 115 years this probability decreased rapidly to below 2% in 1950, and further down to less than 1% in 2006. In fact, in 2006 the probability for females to die before age 40 was less than 1.5%, and about 2.5% for males.

Over the period all age–specific death rates have improved. However, the rate of improvement has not been uniform across ages. Generally, the younger ages have had the largest rates of improvements, the older ages the lowest, and the oldest–old almost no improvements at all. The effect can be seen on Figure 5.1 where the death rates are much more spread out in 2006 than they were in 1835.

The life time distributions provide information about the aggregate effect of improvements in age–specific death rates up to a given age. For instance, the probability of reaching age 80 in 1835 was only about 10% for women and 7% for men[2]. In 2006 these numbers had increased to 60% and 46%, respectively. The probability of reaching the age of 90 has also increased considerably over the period although far less, while the probability of becoming a centenarian has only improved from 0.2% to 1.6% for women and from 0.03% to 0.4% for men.

[2]Figure 5.3 gives the probability of *dying* before a given age and the numbers are thus obtained as 100% minus the values in the figure.

Historically male death rates have been about 10% higher than female death rates (plots not shown). In the period from 1950 to 1980 where male death rates stagnated while female death rates continued to decline this gap widened to an excess male mortality of about 70%. Currently, male death rates are about 50% higher than female death rates for most age groups, with the exception of young male mortality (ages 20 to 40) which is more than twice as high as female mortality. This gender difference give rise to substantially lower probabilities for males to attain high ages as seen on Figure 5.3.

Improvements in age–specific death rates have caused people to die at still higher ages. However, the probability of attaining a very high age of 100, say, has not improved by much. This phenomenon is sometimes referred to as rectangularisation. The term refers to period survival functions looking increasingly "rectangular" staying close to 1 up to high ages and then dropping to 0 over a short age span at very high ages. In 2006 about 60% of female deaths occurred between ages 80 and 100, a span of only 20 years, while in 1835 the same percentage of deaths was spread out between ages 25 and 100.

The nature of oldest–old mortality and how it will develop in the future is the object of an interesting, but somewhat speculative, debate. Some argue that there is a biological highest age for the human body. This would imply that medical and other advances can improve death rates only up to a certain age and people will tend to die in a still more narrow age span just below the highest possible age. Others argue that no such upper limit exists and that mortality of the oldest–old will indeed improve in the future and still higher ages will be attained. The interested reader is referred to Thatcher (1999); Rose and Mueller (2000); Yashin and Iachine (1997) and references therein.

One should keep in mind that the development of oldest–old mortality is largely of academic interest. The practical and economic implications of even a drastic improvement among the oldest–old will be limited as this group is quite small and will continue to be so for a long time.

5.4 Life expectancy

Life expectancy at birth, or simply life expectancy, is the usual way of summarising the age–specific death rates of a population at a given point in time. It is also the measure of choice when describing the effect of improvements in death rates over time and when comparing "the state of mortality" in different countries; in the present paper we make use of it for both purposes.

We stress again that despite the intuitive appeal of the name, the life expectancy calculated from a period life table does not represent the expected life time of a newborn. The latter quantity, the so–called cohort life expectancy, is generally, substantially higher since newborns will typically experience age–specific death rates lower than the current level due to future improvements. However, for a cohort still alive the calculation of its life expectancy must necessarily be partly subjective and based on a specific model for the as yet unknown future death rates; only for extinct cohorts can life expectancies be calculated from observed death rates only.

The main advantage of period life expectancies over cohort life expectancies is that they are objective summaries of observed death rates and for this reason we focus on the former in the present descriptive study, although, cohort life expectancies are arguably of more interest in some situations.

The life expectancy at birth depends on all age–specific death rates from age 0 to the highest attainable age which, in this paper, is set to age 100. Similarly, one can calculate the remaining life time given survival to a given age. These quantities depend on the age–specific death rates from the conditioning age onwards and they provide information about the tail of the life time distribution.

Figure 5.4 shows the total expected life time for females and males given survival to a given age for selected period life tables. The height of the first box indicates the life expectancy at birth, the combined height of the first and second box indicates the expected *total* life time given survival to age 20, and so on for ages 40, 60, 70, 80 and 90. Thus the height of the second box represents the increase in total life expectancy when surviving from age 0 to age 20, the height of the third box represents the increase when surviving from age 20 to age 40, and so on for the higher ages. Note that the expected *remaining* life time at a given age

can be obtained from the graph as the expected total life time minus the conditioning age. Two features of Figure 5.4 stand out. The first is the sharp increase in life expectancy at birth from 1835 to 1950 (and the more moderate increase hereafter). The second is the absence of substantial improvements in expected total life time given survival to high ages.

As already noted the increase in life expectancy at birth from 1835 to 1950 was primarily due to a marked reduction in infant and child mortality over the period. The very high level of infant and child mortality in the 19th century can indirectly be seen on Figure 5.4. In 1835 male life expectancy at birth was 36 years, while the expected total life length given survival to age 20 was 59 years. Thus surviving the first 20 years gave you an expected gain in total life time of 23 years... A discrepancy of that size implies that the chance of surviving to age 20 must have been rather small. In fact, as can be seen from Figure 5.3 about 40% of newborn males died before age 20. Hence the high level of infant and child mortality manifests itself as large total life expectancy gains from surviving the first 20 years, i.e. in the large size of the second box.

After 1950 we see a very different pattern in which life expectancy at birth is almost the same as the expected total life length given survival to age 20 and 40. The additional expected total life time given survival to age 60 and 70 is also quite small. The new pattern that has arisen is caused by all age–specific death rates up to age 40, say, now being at a very low level. The almost collapse of expected total life time given survival up to age 60 means that future gains in life expectancy at birth will be mirrored almost one to one in gains in expected life time of 60–year–olds.

The second striking feature of Figure 5.4 is that the remaining expected life time for 90–year–olds has been almost constant at a level of about 4 years throughout the period (of course the picture is slightly exaggerated by our setting the death rate at age 100 to 1). Gains in expected remaining life time for 80–year–olds have also been modest. It has increased from 5 years in 1835 to 7 in 2006 for males and from 6 years to 9 years for females. Even the expected remaining life time for 60–year–olds has increased with only 7 years for males and 8 years females over the period from 1835 to 2006. However, the increase since 1995 has been about 2 years. Overall, the aggregate effect of improvements in old age mortality have been fairly modest and the gains in terms of expected remaining life time have

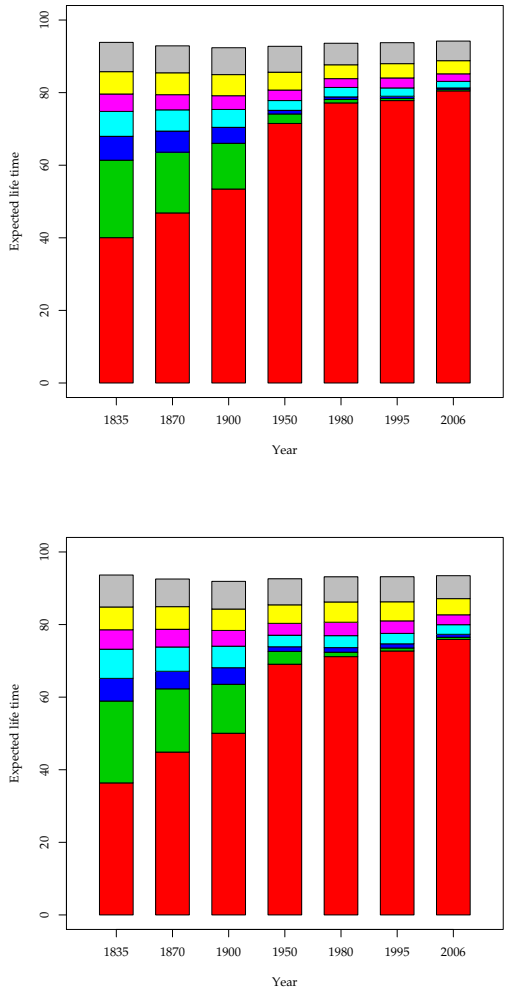

Figure 5.4: Expected total life times for Danish females (top) and males (bottom) at selected years. Each column shows the expected total life time at birth (first box) and given survival to age 20, 40, 60, 70, 80 and 90 based on the period life table for that year.

occurred quite recently.

The purpose of the rest of this section on life expectancy is to develop a sensitivity measure which relates improvements in age–specific death rates to increases in life expectancy and to use this to decompose historic life expectancy gains into age–specific contributions. A tool which we have already used in Section 5.3.

5.4.1 Sensitivity measure

Life expectancy is a complicated function of the entire intensity curve and it is not easy a priori to say how it will respond to changes in (part of) the curve. Apart from infant and child mortality the force of mortality is increasing with age and reaches very high levels at old ages. This, however, does not imply that improvements in old age mortality will have an appreciable effect on life expectancy since only few people will benefit from the improvements.

For any age x we can express life expectancy at birth as a term related to those dying before age x and a term related to those surviving to age x. Using the notation introduced in Section 5.2 we have

$$\bar{e}_0 = \mathbb{E}\left\{T|T \leq x\right\}\mathbb{P}(T \leq x) + \mathbb{E}\left\{T|T > x\right\}\mathbb{P}(T > x).$$

Changing the intensity for ages higher than x will affect the expected total life time given survival to age x, $\mathbb{E}\left\{T|T > x\right\}$, but the effect on life expectancy at birth, $\bar{e}_0$, will be dampened by the probability of surviving to age x, $\mathbb{P}(T > x)$.

Note that the probability of surviving to, or dying before, a given age depends only on the part of the intensity curve *before* that age, while the expected remaining life time, or the expected total life time, given survival to a given age depends only on the part of the intensity curve *after* that age. In that sense the life time distribution and the expected total life times, Figures 5.3 and 5.4, are dual representations of the same information.

The formula above is valid for one age at a time and can be used to derive a sensitivity measure for changes above a given age. However, to understand the simultaneous impact of changes to the entire curve we will form the (functional) derivative which measures the rate with which life expectancy will change when changing the intensity curve in a given

direction. We will consider age–specific *relative improvements* of rate δ of the intensity curve μ and thus calculate the following derivative

$$\begin{aligned}\frac{\partial \bar{e}_0\left(\mu(1-\epsilon\delta)\right)}{\partial \epsilon}_{|\epsilon=0} &= \frac{\partial}{\partial \epsilon}\int_0^\infty e^{-\int_0^x \mu(y)(1-\epsilon\delta(y))dy}dx_{|\epsilon=0} \\ &= \int_0^\infty \int_0^x \mu(u)\delta(u)e^{-\int_0^x \mu(y)dy}dudx \\ &= \int_0^\infty \int_u^\infty \mu(u)\delta(u)e^{-\int_0^x \mu(y)dy}dxdu \\ &= \int_0^\infty \delta(u)D_\mu(u)du,\end{aligned}$$

where the kernel is given by

$$D_\mu(u) = \mu(u)\int_u^\infty e^{-\int_0^x \mu(y)dy}dx = \mu(u)\bar{F}(u)\bar{e}_u = f(u)\bar{e}_u. \tag{5.3}$$

The kernel measures the (marginal) effect of a relative improvement of the intensity at a given age, u, and is equal to the fraction of people dying at that age, measured by the density $f(u)$, times the expected remaining life time $\bar{e}_u$. The result is very intuitive but could hardly have been anticipated in advance. Note that since working in a continuous framework one has to integrate over all the age–specific improvements using the kernel as a weight function to get the aggregate effect, and that improvements at one age only has no effect.

For all years from 1835 to 2006 we have calculated the kernel using formula (5.3) and subsequently computed the average over the periods 1835–1899, 1900–1949 and 1950–2006. The result is shown in Figure 5.5. In the 19th century the sensitivity of life expectancy to improvements in age–specific mortality was almost monotone decreasing in age, apart from the hump for young males. This was to be expected due to the high level of infant and child mortality and the low fraction of people reaching high ages. In the first half of the 20th century with infant and child mortality much reduced the highest sensitivity is now to be found for ages 50 to 80, although life expectancy gains from improvements in young age mortality is still substantial. The kernel has a value of about 0.13 for females between age 20 and 40, which means that a simultaneous improvement of 10%, say, of the intensity for this age group would increase life expectancy at birth

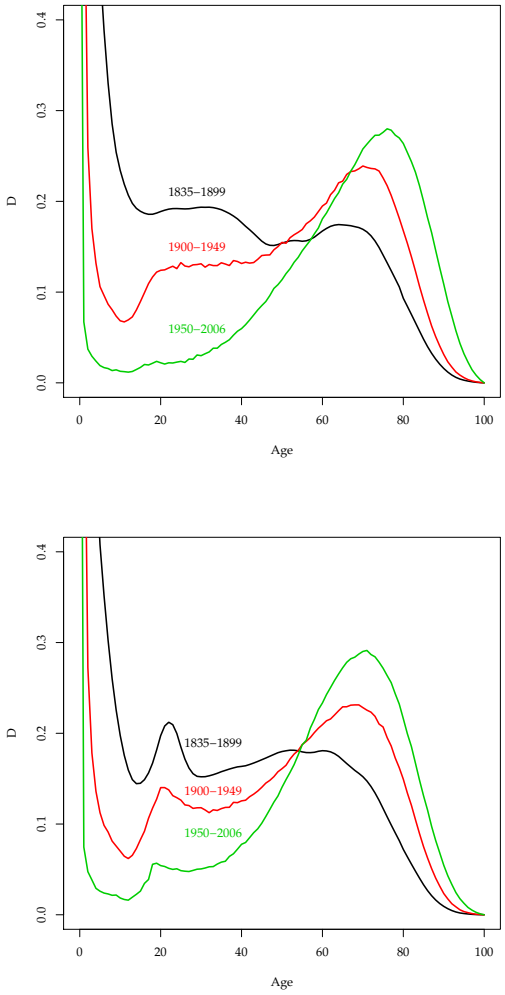

Figure 5.5: Sensitivity of expected life time at birth to improvements in age–specific death rates for Danish females (top) and males (bottom). The plot shows the average kernel, D, over the periods 1835–1899, 1900–1949 and 1950–2006; see Section 5.4.1 for the definition and interpretation of D. The value at age 0 for the three periods is 6.9, 4.6 and 0.8 for females and 8.0, 5.7 and 1.1 for males.

with approximately 0.26 ($= 0.13 \cdot 10\% \cdot 20$) years. The same reduction for the age group from 60 to 80 would have an effect almost twice as high.

Apart from the peak at age 0 the sensitivity curve for the last period is shifted further towards the high ages and has a well–defined hump around age 75 for women and age 70 for men. As opposed to the previous periods improvements in the mortality at very high ages will now have an appreciable effect on life expectancy at birth. For females the sensitivity at age 85 is about the same as at age 50 and the same relative improvement in the two age–specific death rates will therefore have the same impact on life expectancy at birth.

5.4.2 Decomposing life expectancy gains

The decline in age–specific mortality rates over time causes life expectancy to rise and it is illuminating to study how the different age groups have contributed to the increase. Below we suggest a way to decompose life expectancy gains, a tool which we have already used repeatedly throughout Section 5.3.

Generally all age–specific death rates will be different when comparing two life tables and a first attempt of decomposing the life expectancy difference could be to decide on an order, e.g. from the lowest age to the highest age, in which to change the death rates from one table to the other and assign the change in life expectancy after each change to that age. The result of this procedure, however, will depend on the chosen order and one would arrive at a different result when changing the updating scheme.

Having access to only two period life tables at time s and time t, say, there is no unique way to obtain the desired decomposition. However, imagine the idealised situation in which we could observe each of the age–specific intensities at any time u between s and t, $\mu_u(y)$, and assume that the transition from $\mu_s(y)$ to $\mu_t(y)$ is smooth. Under these assumption we can calculate the time derivative of life expectancy using a similar line of

reasoning as in the previous section

$$\begin{aligned}\frac{\partial}{\partial u}\bar{e}_0(\mu_u) &= \int_0^\infty \frac{\partial}{\partial u} e^{-\int_0^x \mu_u(y)dy} dx \\ &= -\int_0^\infty \int_0^x \mu_u'(v) e^{-\int_0^x \mu_u(y)dy} dv dx \\ &= -\int_0^\infty \int_v^\infty \mu_u'(v) e^{-\int_0^x \mu_u(y)dy} dx dv \\ &= -\int_0^\infty \frac{\mu_u'(v)}{\mu_u(v)} D_{\mu_u}(v) dv,\end{aligned}$$

where the kernel, $D_{\mu_u}(v)$, is defined in (5.3). The change in life expectancy from time s to time t, $\bar{e}_0(\mu_t) - \bar{e}_0(\mu_s)$, can then be written as

$$\begin{aligned}\bar{e}_0(\mu_t) - \bar{e}_0(\mu_s) &= \int_s^t \frac{\partial}{\partial u}\bar{e}_0(\mu_u) du \\ &= -\int_s^t \int_0^\infty \frac{\mu_u'(v)}{\mu_u(v)} D_{\mu_u}(v) dv du \\ &= -\int_0^\infty \int_s^t \frac{\partial}{\partial u} \left\{\log \mu_u(v)\right\} D_{\mu_u}(v) du dv.\end{aligned}$$

The inner integral in the last formula can be interpreted as (the density of) the age–specific contribution to the life expectancy gain over the period related to age u, which is precisely what we are after. Note, however, that to calculate this expression we would have to make an assumption about the value of $\mu_u(v)$ for non–integer values of u, e.g. linear or exponential between the observed values at the neighbouring integers.

In order to obtain a simpler formula we will instead make the assumption that the kernel, $D_{\mu_u}(v)$, exhibits only a weak dependence on time, i.e. $D_{\mu_u}(v) \approx \bar{D}(v)$, for some $\bar{D}(v)$. Under this assumption we get the relation

$$\bar{e}_0(\mu_t) - \bar{e}_0(\mu_s) \approx \int_0^\infty \bar{D}(v) \log \frac{\mu_s(v)}{\mu_t(v)} dv, \tag{5.4}$$

which expresses the increase in life expectancy in terms of improvements in the age–specific log mortality rates and the previously introduced kernel. Note that only the intensities at time s and t are needed under this assumption.

In Figure 5.2 we have used formula (5.4) to decompose the life expectancy gains for each of the periods 1835–1900, 1900–1950 and 1950–2006 into contributions from different age groups. For each of three periods we have used the average kernel over the period, shown in Figure 5.5, as $\bar{D}$ and for each age v between 0 and 100 we have then calculated its contribution to the life expectancy gain as $\bar{D}(v)\log[\mu_s(v)/\mu_t(v)]$. To make the age–specific contributions sum to the observed life expectancy gain over the period we have scaled them by a common factor. Finally, we have grouped the contributions into the age groups 0, 1–10, 11-20, ..., 91–100, and stacked the contributions for each period on top of each other.

The plot shows three distinct improvement patterns. In the first period life expectancy gains were driven mainly by reductions in child, and to some extent infant, mortality. Over the second period a wide range of age groups contributed to the increase in life expectancy, with the reduction in infant mortality being the single most important. While in the most recent period life expectancy gains have come mainly from improvements in high age mortality but also to some degree from further reductions in infant mortality.

5.5 Denmark and the World

Mortality improvements is by no means an isolated Danish phenomenon. The evolution of Danish mortality which has resulted in an increase of 40 years in life expectancy from 1835 to 2006 is certainly remarkable, but many developed countries have seen even larger improvements over that period. In fact, the evolution in large parts of the developed world has surpassed the Danish evolution to such an extent that over the course of the last century Denmark has moved from being a top–ranking country with respect to life expectancy to currently being in the bottom half. This development has caused concern not least because other Nordic countries, e.g. Sweden and Norway, have been able to maintain their position as world–leading countries with respect to life expectancy.

All countries have had their own unique mortality history depending on national characteristics. Therefore, to put the Danish development into an international perspective without going into country–specific details we have constructed a single, international death rate based on data for 18

Country	1900	1950	1980	1995	2004
Denmark	53.4 (3)	71.5 (3)	77.2 (6)	77.8 (7)	79.8 (7)
Sweden	53.6 (2)	72.4 (2)	78.8 (2)	80.8 (4)	82.6 (3)
Norway	55.1 (1)	73.2 (1)	79.1 (1)	81.4 (3)	82.3 (4)
France	46.7 (5)	69.2 (6)	78.4 (4)	81.9 (2)	83.8 (2)
UK	48.1 (4)	71.3 (4)	76.7 (7)	79.4 (5)	80.7 (5)
US		71.0 (5)	77.4 (5)	79.1 (6)	80.2 (6)
Japan		60.9 (7)	78.7 (3)	82.8 (1)	85.4 (1)

Table 5.2: Female life expectancy at birth for selected countries and years (rank in parenthesis). For UK last year available is 2003.

developed countries[3].

Data is extracted from the Human Mortality Database and for each country it consists of gender specific death counts and exposures in the format described in Section 5.2. From these we have constructed an international death rate as the ratio between the total death count and the total exposure in all of the countries for which data exists for the given year. To ensure that most of the larger countries are represented each year we consider only the time period from 1900 to 2004.

Tables 5.2–5.3 show life expectancies for selected countries and years. And the plots in Figure 5.6 compare Danish and international death rates for adult ages.

At the start of the 20^{th} century Danish death rates were about 15% lower than the international level for both females and males. The Danish life expectancy was comparable to the life expectancy in Sweden and Norway, and substantially higher than in France and UK.

In the first half of the 20th century mortality rates declined substan-

[3]The 18 countries with the time range of available data in parenthesis are: Australia (1921–2004), Austria (1947–2005), Belgium (1841–2005), Canada (1921–2004), UK, Civilian Population (1841–2003), Finland (1878–2005), France, Civilian Population (1899–2005), West Germany (1956–2004), Iceland (1838–2005), Italy (1872–2003), Japan (1947–2005), Netherlands (1850–2004), Norway (1846–2006), Portugal (1940–2005), Spain (1908–2005), Sweden (1751–2006), Switzerland (1876–2005), US (1933–2004).

Country	1900	1950	1980	1995	2004
Denmark	50.1 (3)	69.1 (3)	71.2 (4)	72.7 (6)	75.1 (7)
Sweden	50.7 (2)	69.8 (2)	72.8 (2)	74.8 (3)	78.3 (2)
Norway	51.7 (1)	69.9 (1)	72.3 (3)	76.2 (2)	77.5 (3)
France	43.0 (5)	63.4 (6)	70.2 (6)	73.8 (5)	76.7 (4)
UK	44.1 (4)	66.5 (4)	70.7 (5)	74.2 (4)	76.4 (5)
US		65.4 (5)	70.0 (7)	72.7 (7)	75.2 (6)
Japan		57.5 (7)	73.4 (1)	76.5 (1)	78.7 (1)

Table 5.3: Male life expectancy at birth for selected countries and years (rank in parenthesis). For UK last year available is 2003.

tially in many developed countries. The Danish life expectancy rose by almost 19 years for both females and males and Denmark was, at that time, still one of the countries with highest life expectancy in the world. However, life expectancy rose by even more in France and UK so although Denmark remained its relative position the gap had already narrowed, particulary for women.

From 1950 to 1980 Danish female mortality continued to improve although at a lower pace than in the previous period. However, improvements in international death rates did not slow down as much and as a result Denmark was no longer a leading country with respect to life expectancy at the end of the period. Of the seven countries listed in Table 5.2 only UK women had a lower life expectancy than Danish women in 1980.

In the same period Danish males experienced an almost stagnation in death rates and life expectancy. International death rates, on the other hand, continued to decline and at the end of period they had essentially caught up with the Danish level. Danish male life expectancy in 1980 was still in the high end, but the gap to the other countries had shrunk substantially.

When discussing life expectancy evolution Japan stands out. In 1950 the Japanese life expectancy trailed many other developed countries by almost 10 years for both males and females. However, over a period of only 30 years they came to have one of the highest life expectancies in the

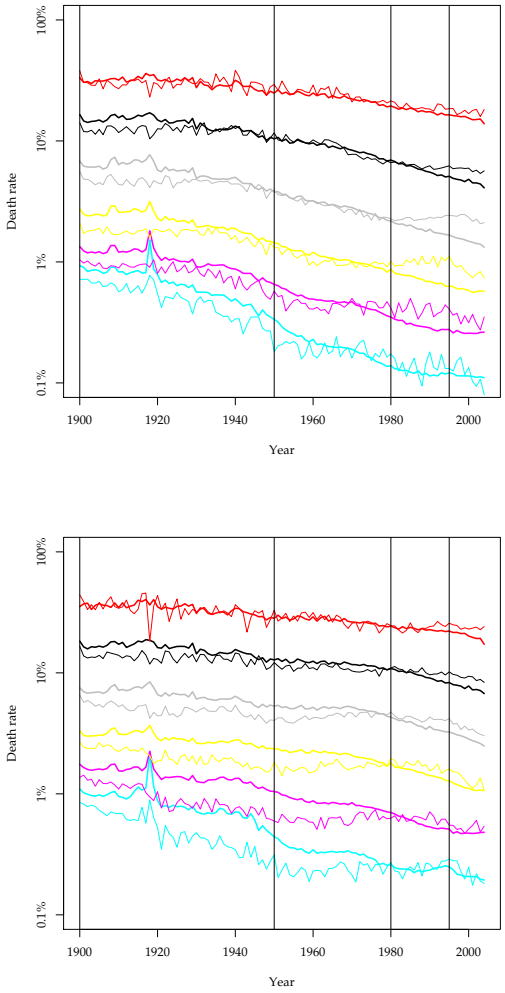

Figure 5.6: Development in Danish and international female (top) and male (bottom) age–specific death rates from 1900 to 2004 for ages 40, 50, . . . , 90. Danish rates are shown with thin lines and international rates with thick lines. Top lines represent 90–year–olds and the other ages follow below in decreasing order.

world. From 1950 to 1980 life expectancy rose with 16 years for men and 18 years for women corresponding to an annual increase of over half a year. In comparison, the Danish life expectancy increased with only 2 years for men and 6 years for women.

The following 15 years from 1980 to 1995 saw a stagnation in Danish female mortality. The international level continued to fall over the period and a large gap between the Danish and the international level was established. At the end of the period the excess mortality in Denmark for women aged 50 to 70 was more than 50% compared to the international level. In 1995 Danish female life expectancy was the lowest of the seven countries in Table 5.2 more than one year behind US as the second lowest.

Danish males was also overtaken by the international development from 1980 to 1995, but the excess Danish male mortality at the end of the period was far less than for females. Still, in 1995 Danish male life expectancy was at almost the same level as that of US and together they ranked lowest of the seven countries in Table 5.3.

In Japan life expectancy continued to rise at a high rate from 1980 to 1995, and at the end of the period they ranked number one for both females and males. Hence, it took Japan less than half a century to move from the bottom of the list to the top. Truly a remarkable achievement. During the same period Denmark fell from the top to the bottom.

Since 1995 Denmark has again experienced substantial life expectancy gains. From 1995 to 2004 life expectancy rose with about 2 years for women and about 2 and a half years for men. However, most other developed countries had similar gains and the ranking of the seven countries in Tables 5.2–5.3 in 1995 and in 2004 is therefore almost identical.

In 2004 Denmark was still the lowest ranking country, while Japan, France, Sweden and Norway constituted the top half. The gap between Denmark and the two other Nordic countries was between 2 and 3 years. Compared to Japan, however, the life expectancy gap was about 2 and a half years for males and about 5 and a half years for females.

Throughout the period life expectancy gains for women have been largest in Japan. Japanese women were already the longest living in 1995 but even so Japan experienced the largest increase from 1995 to 2004 among the seven countries in Table 5.2. This is fascinating in its own right but it also shows that large increases can occur in countries where

life expectancy is already high. Hence, further, substantial increases in life expectancy can be expected for both Sweden and Norway. In Denmark we can hope for a future reduction in excess mortality, particularly for women, with the effect of even higher future increases in life expectancy as Denmark catches up.

5.5.1 The big why

There is no agreed upon explanation as to why the mortality progress in Denmark from 1950 to 1995 was so much slower than in most other developed countries. It is a fact, though, that this period which had the largest economic progress in history and during which the so–called Welfare State was founded with a large public health sector saw relatively modest mortality improvements.

Many explanations have been put forward trying to explain this paradox including increased female labour market participation rates in the 1960s; increased consumption of tobacco and alcohol; changes in diet and other life style changes; lack of physical exercise and increased obesity; and less efficient screening programs and treatments, particularly for cancer. None of these explanations is entirely satisfying on its own, however, since most developed countries have had a development similar to Denmark in many of these areas.

Some light can be shed on the Danish excess mortality by looking at cause–specific mortality rates. For the seven countries previously compared Table 5.4 shows age–standardised death rates for all causes, cancer and circulatory diseases. The latter two being the major death causes in the developed world. The reported death rates are computed as a weighted average of age–specific death rates using (the age composition of) the same standard population as weights. This is done to take account of differences in age structure of the populations being compared. Data is extracted from OECD Health Data 2007 which contains statistics on health and health care systems in OECD countries from 1960 to 2006[4]. However, for reasons of comparison we have chosen to use data from 2001 as this is the latest year for which death rates for all of the selected countries are available (the restricting country being Denmark). The excess mortality of Danish

[4]See www.sourceoecd.org

Country	Females			Males		
	All	M.n.	C.d.	All	M.n.	C.d.
Denmark	593 (7)	186 (7)	184 (5)	871 (7)	245 (6)	308 (6)
Sweden	460 (3)	134 (3)	182 (4)	705 (2)	180 (1)	303 (5)
Norway	481 (4)	140 (4)	171 (3)	783 (4)	208 (3)	299 (4)
France	411 (2)	116 (2)	116 (2)	774 (3)	252 (7)	203 (2)
UK	536 (5)	155 (6)	194 (6)	804 (5)	225 (5)	314 (7)
US	554 (6)	142 (5)	198 (7)	826 (6)	207 (2)	297 (3)
Japan	329 (1)	102 (1)	103 (1)	638 (1)	215 (4)	174 (1)

Table 5.4: Standardised death rates per 100,000 population based on data for 2001 (rank in parenthesis). "All" is all causes, "M.n." stands for malignant neoplasms (cancer), and "C.d." abbreviates circulatory diseases.

women is indeed alarming and much of it can be attributed to excess mortality related to cancer (malignant neoplasms). Compared to Sweden and Norway the Danish women have a cancer related excess mortality of more than 30%, while the level of mortality related to heart diseases (circulatory diseases) is comparable in the three countries. Even when comparing with UK, which has the second highest cancer related death rate, the Danish women have an excess mortality of 20%. Once again the Japanese (and the French) women stand out by having cause–specific death rates much lower than the women in any other country.

The picture for males is less clear. First of all, the variation between the seven populations is smaller for men than for women. For both males and females Japan has the lowest all–cause death rate and Denmark the highest. However, whereas Danish women have an all–cause excess mortality of 80% compared with Japan the Danish men have an all–cause excess mortality of "only" 37%. The two cause–specific death rates for Danish males are both high, but neither of them stands out. For cancer there is an excess mortality compared to Sweden and Norway, but the Danish level is comparable to that of France and UK. For heart diseases, of which ischaemic heart diseases count the most (numbers not shown), the countries can be divided in two groups: Japan and France in the top and the five

other countries at a comparable level at the bottom.

Looking at the pattern of cause–specific death rates across countries it seems reasonable to conclude that diet and smoking habits must be at least part of the explanation of the observed differences between countries. Undoubtedly, numerous other factors are also important and the question of why the Danish excess mortality is so high must still be considered largely unresolved.

5.6 Concluding remarks

The story of mortality evolution is one of continued improvements. Whether age–specific death rates can decline indefinitely and life expectancy continue to rise or whether there exists an unsurmountable biological barrier for human life spans is a question of philosophical nature. However, it seems almost certain that we will witness appreciable mortality improvements in the foreseeable future, in particular for countries like Denmark which is lacking behind other developed countries. Here a reduction in excess mortality will in itself give rise to substantial life expectancy gains.

Going back in history one can point at a number of factors behind the observed reductions in death rates: improvements in nutrition and sanitary conditions, higher standards of living, better housing and working conditions, public health measures, better hygiene in hospitals, medical advances etc. Over the course of history these changes have led to death causes changing from infectious diseases such as tuberculosis, diphtheria and cholera to degenerative diseases such as cancer and heart diseases.

Detailed knowledge of causality is valuable for understanding the past but it is of limited value when trying to predict the future. The mechanism governing death rates is too complex and the impact of future medical inventions, economic development, demographic changes etc. cannot possibly be foreseen. Consequently, most mortality projections are based on purely statistical models extrapolating past trends.

The most widely used model for mortality projections is still the one proposed by Lee and Carter (1992) although numerous extensions and other model types have been proposed since then, see e.g. Brouhns *et al.* (2002); Lee and Miller (2001); Renshaw and Haberman (2006); de Jong and Tickle (2006); Currie *et al.* (2004); Cairns *et al.* (2006). For recent

comparisons of selected models see Cairns *et al.* (2007) and Booth *et al.* (2006).

All these models provide more or less structured projections of age–specific death rates and their main strength is their ability to extrapolate regular improvement patterns. Changes in improvement rates are considered as structural breaks and data before the last structural break is often disregarded. Large populations, like the US, do indeed show regular patterns with near constant annual rates of improvement over long periods and this approach, although hardly optimal, is feasible. However, for small regions, like Denmark, the mortality evolution has been much more erratic with many periods with very different improvement patterns and basing a possibly long–term projection on the last regular period is neither robust nor trustworthy. Despite the wealth of models in existence we feel there is a need for developing a methodology which can make convincing projections from volatile mortality rates making proper use of all available data. In a forthcoming paper we propose a new model for small region mortality projections.

6. Some solvable portfolio problems with quadratic and collective objectives

Background. The paper in this chapter is written jointly with Mogens Steffensen. The chapter is a slightly updated version of Kryger and Steffensen (2010).

Abstract. We present a verification result for a general class of portfolio problems, where the standard dynamic programming principle does not hold. Explicit solutions to a series of cases are provided. They include dynamic mean–standard deviation, endogenous habit formation for quadratic utility, and group utility. The latter is defined by adding up the certainty equivalents of the group members, and the problem is solved for exponential and power utility.

6.1 Introduction

For decades the class of Hamilton–Jacobi–Bellmann–solvable dynamic asset allocation problems over terminal wealth, $X(T)$, has been limited to those in the form

$$\sup_{\pi} \mathbb{E}_{t,x} \left\{ F\left(X(T)\right) \right\},$$

for some function F, with π being an allocation control. Björk and Murgoci (2008) extended this class to those in the form

$$\sup_{\pi} \left[\mathbb{E}_{t,x} \left\{ F\left(t, x, X(T)\right) \right\} + G\left(t, x, \mathbb{E}_{t,x} \left\{ X(T) \right\}\right) \right], \tag{6.1}$$

for some function G, which allowed them to calculate the optimal *time consistent* investment strategy for a mean–variance investor. This result

was first published by Basak and Chabakauri (2009b) in a quite general incomplete market framework. Basak and Chabakauri (2009a) also use their methodology to calculate the optimal time consistent strategy for a variance–minimising investor holding a non–tradeable asset.

The novelty of Björk and Murgoci (2008) is, apart from working in a general Markovian financial market, the dependence on (t, x) in their F as well as the mere presence of a G that is not affine in the conditional expectation of terminal wealth. Furthermore they allow for consumption, skipped in (6.1). The dependence on (t, x) and the non-affine G rule out the use of the classical dynamic programming–technique based on iterated expectations, and, consequently, they refer to such problems as *time inconsistent*. Equivalently, the definition of time inconsistent solutions in Basak and Chabakauri (2009b) is "policies, from which the investor has [an] incentive to deviate".

The aim of the present paper is to study the class of problems in the form

$$\sup_{\pi} f\left(t, x, \mathbb{E}_{t,x}\left\{g_1\left(X(T)\right)\right\}, \ldots, \mathbb{E}_{t,x}\left\{g_n\left(X(T)\right)\right\}\right),$$

for some integer n, and where f is allowed to be non-affine in the g-functions. Our main application is a group utility problem, where a group of investors seek to maximise a specific notion of group utility, where investors share terminal wealth equally. Whereas utility maximisation for a single investor may be considered a classic problem it is not clear how to formalise the preferences of a group of heterogeneous agents. We suggest to maximise the sum of certainty equivalents, and thereby form the objective

$$\sum_{i=1}^{n} u_i^{-1}\left(\mathbb{E}_{t,x}\left\{u_i\left(\alpha_i X^{\pi}\left(T\right)\right)\right\}\right),$$

for individual utility functions $u_1, \ldots, u_n$. Note that due to monotonicity of u_1 this problem is equivalent to the standard problem in the single investor case. The positive constant α_i indicate the proportion of total wealth that agent i is entitled to.

A different problem of interest that is contained in our general objective is mean–standard deviation optimisation.

Both problems call on our general objective, and are not special cases of Björk and Murgoci (2008). On the other hand, due to the presence of

t, x in their F–function they can deal with problems that we cannot treat.

The concept of time–inconsistency was first treated formally by Strotz (1956), who considered a so–called "Cake–Eating Problem" (i.e. one of allocating an endowment between different points in time). He showed that the optimal solution is time consistent only for exponential discounting. Strotz (1956) described three different types of agents, and Pollak (1968) contributed further to the understanding and naming of them: 1) the pre–committed agent does not revise his initially decided strategy even if that makes his strategy time–inconsistent; 2) the naive agent revises his strategy without taking future revisions into account even if that makes his strategy time–inconsistent; 3) the sophisticated agent takes possible future revisions into account, thereby making his strategy time consistent. Which type is more relevant depends on the entire framework of the decision in question. Here, we focus on the pre–committed and sophisticated agents and pay no attention to the naive agent. Strotz (1956) suggests that, although (in some sense) optimal, it may be difficult to pre–commit.

In recent years the concept of non-exponential (e.g. hyperbolic) discounting has received a lot of attention as a prime example of a time inconsistent problem. Solano and Navas (2010) give an overview over which strategies the three different types of agents should use.

A proposition, which characterises the solution (in a Black–Scholes market) to our class of problems is provided in Section 6.2, while Sections 6.3 and 6.4 present applications, some of which are - to our knowledge - new. Finally, Section 6.5 wraps up the findings and provides an outlook on further work within this area.

6.2 The main result

We consider a Black–Scholes market consisting of a bank account with interest intensity r, and a stock with dynamics given by

$$dS(t) = (r + \Lambda\sigma) S(t)\, dt + \sigma S(t)\, dB(t), S(0) > 0,$$

with market price of risk Λ and volatility $\sigma > 0$. B is a standard Brownian motion on an abstract probability space $(\Omega, \mathbb{F}, \mathbb{P})$ equipped with a filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ satisfying the usual conditions; and with each $\mathcal{F}_t \supseteq \sigma\{B(s),\ 0 \leq s \leq t\}$.

We consider an investor, who places the proportion $\pi(t)$ of his wealth in the stock at time t. Denoting by $X^\pi(t)$ his wealth at time t given the investment strategy π, the dynamics of his wealth becomes

$$\begin{aligned} dX^\pi(t) &= (r + \pi(t)\Lambda\sigma) X^\pi(t)\, dt + \pi(t)\sigma X^\pi(t)\, dB(t), \\ X^\pi(0) &= x_0 > 0, \end{aligned} \tag{6.2}$$

where x_0 is the initial wealth. The strategy is self-financing in the sense that we disregard consumption and injection of capital.

Before introducing the objectives we introduce two conditional expectations

$$\begin{aligned} y^\pi(t,x) &= \mathbb{E}_{t,x}\left\{g\left(X^\pi(T)\right)\right\}, \\ z^\pi(t,x) &= \mathbb{E}_{t,x}\left\{h\left(X^\pi(T)\right)\right\}, \end{aligned}$$

for functions g and h. The subscript t,x denotes conditioning on the event $X^\pi(t) = x$.

The objective of the investor is to find

$$V(t,x) = \sup_\pi V^\pi(t,x) = \sup_\pi f\left(t,x,y^\pi(t,x), z^\pi(t,x)\right), \tag{6.3}$$

for a given regular function $f \in C^{1,2,2,2}$, and to find the corresponding optimal investment strategy, π^*.

As opposed to Björk and Murgoci (2008) we only treat problems over terminal wealth. Also, we restrict ourselves to a (one–dimensional) Black–Scholes market.

The portfolio problem presented in (6.3) is, in general, not a classical portfolio problem. If f does not depend on (t,x) and is affine in $(y^\pi(t,x), z^\pi(t,x))$, the problem can be written in a classical way,

$$V(t,x) \propto \sup_\pi \mathbb{E}_{t,x}\left\{\hat{F}\left(X^\pi(T)\right)\right\} \tag{6.4}$$

for some $\hat{F}$. The problem (6.3) is, at first glance, just a mathematical abstract generalisation of the problem (6.4). However, as we argue below, there are examples of this generalisation that make good economic sense. Truly, there are also examples of (6.3) that make no economic sense. But this is not an argument against solving (6.3) in its generality, as long as we

have some interesting and useful applications in mind. Here we present a list of five such examples, which we solve for in Sections 6.3 and 6.4. The first four examples are based on the specifications $g(x) = x$ and $h(x) = x^2$. In the fifth example, g and h are utility functions, and y and z are also modified – and so–called group utility is maximised.

1. Mean–variance optimisation with pre–commitment

   This is a classical quadratic utility optimisation problem corresponding to
   $$f(t, x, y, z) = ay + bz + c \tag{6.5}$$
   When studying this example in detail in Section 6.3.4, we explain how this choice of f can deal with both mean–variance utility maximisation and variance minimisation under minimum return constraints. Essentially, we do not need the generalisation (6.3) for this problem. This relates to the fact that f does not depend on (t, x) and is affine in (y, z), so the problem is on the form (6.4).

2. Mean–variance optimisation without pre–commitment
   $$f(t, x, y, z) = y - \frac{\upsilon(t, x)}{2}\left(z - y^2\right)$$
   If υ does not depend on (t, x), then f does not depend on (t, x). But the non–affinity in y makes the problem non–standard. For υ constant, this is the problem treated by Basak and Chabakauri (2009b) in an incomplete market framework. It is studied as a special (the simplest) case by Björk and Murgoci (2008). The case of $\upsilon(x) = \upsilon/x$ is investigated by Björk *et al.* (2009).

3. Mean–standard deviation optimisation
   $$f(t, x, y, z) = y - \upsilon\left(z - y^2\right)^{\frac{1}{2}}$$
   Due to the non–affinity of f in (y, z), this case is not covered by Björk and Murgoci (2008)).

4. Quadratic utility with endogenous habit formation
   $$f(t, x, y, z) = -\left(\frac{1}{2}z + \frac{1}{2}x^2\beta^2(t) - x\beta(t)\, y\right),$$

for some time–dependent required return, β. We provide the full solution to this problem although the case can also be solved using Björk and Murgoci (2008).

5. Collective of heterogeneous investors

$$\begin{aligned} f(t,x,y,z) &= g^{-1}(y) + h^{-1}(z) \\ y^{\pi}(t,x) &= \mathbb{E}_{t,x}\left\{g\left(\alpha X^{\pi}(T)\right)\right\} \\ z^{\pi}(t,x) &= \mathbb{E}_{t,x}\left\{h\left((1-\alpha)X^{\pi}(T)\right)\right\}, \end{aligned}$$

where g and h are utility functions. e.g. for power utility

$$\begin{aligned} g(x) &= x^{1-\gamma_1}/(1-\gamma_1), \\ h(x) &= x^{1-\gamma_2}/(1-\gamma_2), \\ f(t,x,y,z) &= ((1-\gamma_1)y)^{(1-\gamma_1)^{-1}} + ((1-\gamma_2)z)^{(1-\gamma_2)^{-1}} \\ &= \alpha\mathbb{E}_{t,x}\left\{X^{\pi}(T)^{1-\gamma_1}\right\}^{\frac{1}{1-\gamma_1}} \\ &\quad + (1-\alpha)\mathbb{E}_{t,x}\left\{X^{\pi}(T)^{1-\gamma_2}\right\}^{\frac{1}{1-\gamma_2}}. \end{aligned}$$

The functions g and h form the utility of terminal wealth, whereas the function f adds up the so–called certainty equivalents of the two investors. Whereas it may make no economic sense to add up the indirect utility from each investor (e.g. that would add up currency unit in different power), it makes good economic sense to add up certainty equivalents (at least that would add up linear currency units). However, the transition into certainty equivalents before adding up makes the problem non-standard due to the non-linearity of g^{-1} and h^{-1}. In order to extend to more than two agents, f needs more arguments, of course. To our knowledge this problem is new.

One can come up with several other interesting examples, but these are the ones we study in the present paper.

The result that facilitates the solution of this new class of problems is the following proposition.

Proposition 6.1. *Let $f : [0,T] \times \mathbb{R}^3 \to \mathbb{R}$ be a function from $C^{1,2,2,2}$. Let g and h be real functions. The set of* admissible *strategies are those,*

for which the stochastic integrals in (6.49) and (6.54) are martingales, and for which the partial differential equations (6.46)-(6.47) and (6.50)-(6.51) have solutions. Note that admissibility depends on the choice of g, h.

Define $V(t,x) = \sup_\pi f(t,x,y^\pi(t,x), z^\pi(t,x))$ *with the supremum taken over all admissible strategies, and with*

$$y^\pi(t,x) = \mathbb{E}_{t,x}\{g(X^\pi(T))\},$$
$$z^\pi(t,x) = \mathbb{E}_{t,x}\{h(X^\pi(T))\}.$$

If there exist functions F, G, H *such that*

$$F_t - f_t = \inf_\pi \left[-(r + \Lambda\sigma\pi) x (F_x - f_x) - \frac{1}{2}(\sigma\pi)^2 x^2 (F_{xx} - U) \right], \quad (6.6)$$
$$F(T,x) = f(T,x,g(x),h(x)), \quad (6.7)$$
$$G_t = -(r + \Lambda\sigma\pi^*) x G_x - \frac{1}{2}(\sigma\pi^*)^2 x^2 G_{xx}, \quad (6.8)$$
$$G(T,x) = g(x),$$
$$H_t = -(r + \Lambda\sigma\pi^*) x H_x - \frac{1}{2}(\sigma\pi^*)^2 x^2 H_{xx}, \quad (6.9)$$
$$H(T,x) = h(x),$$

where

$$U(f,y,z) = f_{xx} + 2f_{xy}y_x + 2f_{xz}z_x + f_{yy}y_x^2 + 2f_{yz}y_x z_x + f_{zz}z_x^2, \quad (6.10)$$

and

$$\pi^* = \arg\inf_\pi \left[-(r + \Lambda\sigma\pi) x (F_x - f_x) - \frac{1}{2}(\sigma\pi)^2 x^2 (F_{xx} - U) + f_t \right]$$

Then

$$V(t,x) = F(t,x), y^{\pi^*}(t,x) = G(t,x), z^{\pi^*}(t,x) = H(t,x),$$

and the optimal investment strategy is given by π^*.

We find the optimising investment strategy in terms of the value function by differentiating with respect to π inside the square brackets of (6.6) and get

$$\pi^* = -\frac{\Lambda}{\sigma x} \frac{F_x - f_x}{F_{xx} - U(f,y,z)} \quad (6.11)$$

(provided $U > F_{xx}$). Feeding this control process back into the Bellman–like equation we get the following system of partial differential equations (PDEs) that we need to solve:

$$F_t = -rx\left(F_x - f_x\right) + \frac{1}{2}\frac{\Lambda^2\left(F_x - f_x\right)^2}{F_{xx} - U\left(f, G, H\right)} + f_t, \tag{6.12}$$

$$G_t = -\left(rx - \frac{\Lambda^2(F_x - f_x)}{F_{xx} - U\left(f, G, H\right)}\right)G_x - \frac{1}{2}\left(\frac{\Lambda(F_x - f_x)}{F_{xx} - U\left(f, G, H\right)}\right)^2 G_{xx}, \tag{6.13}$$

$$H_t = -\left(rx - \frac{\Lambda^2(F_x - f_x)}{F_{xx} - U\left(f, G, H\right)}\right)H_x - \frac{1}{2}\left(\frac{\Lambda(F_x - f_x)}{F_{xx} - U\left(f, G, H\right)}\right)^2 H_{xx}, \tag{6.14}$$

with boundary conditions

$$\begin{aligned} F\left(T, x\right) &= f\left(T, x, g\left(x\right), h\left(x\right)\right), \\ G\left(T, x\right) &= g\left(x\right), \\ H\left(T, x\right) &= h\left(x\right). \end{aligned}$$

We also present the system in terms of π^*, since this is sometimes convenient to work with:

$$F_t = -rx\left(F_x - f_x\right) - \frac{1}{2}\Lambda\sigma\pi^*\left(F_x - f_x\right)x + f_t, \tag{6.15}$$

$$G_t = -\left(r + \Lambda\sigma\pi^*\right)xG_x - \frac{1}{2}\left(\sigma\pi^*\right)^2 x^2 G_{xx}, \tag{6.16}$$

$$H_t = -\left(r + \Lambda\sigma\pi^*\right)xH_x - \frac{1}{2}\left(\sigma\pi^*\right)^2 x^2 H_{xx}, \tag{6.17}$$

(with unchanged boundary conditions).

Remark 6.2. *The proposition can easily be extended to cover more than two transformations of terminal wealth:*

$$V(t, x) = \sup_{\pi} f\left(t, x, y_1^{\pi}(t, x), \ldots, y_n^{\pi}(t, x)\right),$$

in which case

$$U(f, y_1, \ldots, y_n) = \frac{\partial^2 f}{\partial x^2} + 2\sum_{i=1}^{n}\frac{\partial^2 f}{\partial x y_i}\frac{\partial y_i}{\partial x} + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 f}{\partial y_i y_j}\frac{\partial y_i}{\partial x}\frac{\partial y_j}{\partial x},$$

for $y_i^{\pi} = \mathbb{E}_{t,x}\left\{g_i\left(X^{\pi}\left(T\right)\right)\right\}$.

Remark 6.3. *The standard case can be formalised by*

$$f\left(t,x,y^{\pi}\left(t,x\right),z^{\pi}\left(t,x\right)\right)=y^{\pi}\left(t,x\right)$$

Then the result collapses into a standard Bellman equation. This is seen by realising that

$$f_t = f_x = U = 0. \tag{6.18}$$

In this case $F = G$ and the differential equation for G (and also for H of course, since f does not depend on z) is redundant in the proposition.

Also the proof collapses into a standard proof for the Bellman equation.

In the next two sections we solve the five problems listed above and variations thereof.

6.3 Quadratic objectives

This section analyses the first four problems from the list in Section 6.2, albeit in a different order.

6.3.1 Mean–variance without pre–commitment

In this section we consider the optimisation problem

$$V\left(t,x\right)=\sup_{\pi}\left(\mathbb{E}_{t,x}\left\{X^{\pi}\left(T\right)\right\}-\frac{\upsilon(t,x)}{2}\mathbb{V}_{t,x}\left\{X^{\pi}\left(T\right)\right\}\right).$$

When $\upsilon(t,x)=\upsilon$ the solution to this problem was found by Basak and Chabakauri (2009b) in a relatively general incomplete market. Björk and Murgoci (2008) also give the solution as *the* example of their rather general method.

For constant υ the function f is given by

$$\begin{aligned} f\left(t,x,y,z\right) &= y-\frac{\upsilon}{2}\left(z-y^{2}\right), \qquad (6.19)\\ f_y &= 1+\upsilon y, f_{yy}=\upsilon, f_z=-\frac{\upsilon}{2},\\ f_t &= f_x = f_{xx} = f_{zz} = f_{xy} = f_{xz} = f_{yz} = 0. \end{aligned}$$

From (6.10) we can now derive

$$U = \upsilon G_x^2.$$

Plugging $f_t = f_x = 0$ and U into (6.11) and (6.12) we get the following optimal investment candidate and PDE that we need to solve together with (6.16),

$$\pi^* = -\frac{\Lambda}{\sigma x}\frac{F_x}{F_{xx} - \upsilon G_x^2} \tag{6.20}$$

$$F_t = -rxF_x + \frac{1}{2}\Lambda^2\frac{F_x^2}{F_{xx} - \upsilon G_x^2}, \tag{6.21}$$

$$F(T, x) = x.$$

In this particular case the PDE for F involves G but not H and therefore we do not need to pay attention to the PDE for H. After having derived the solution to (6.21), this is plugged into (6.20) to form the optimal investment strategy as a function of (t, x). Plugging this strategy into (6.9) results in a PDE characterising H. However we do not need this characterisation in order to find the optimal investment and the value function.

We now search for a solution in the form

$$F(t, x) = p(t)\, x + q(t), G(t, x) = a(t)\, x + b(t).$$

Note that, for such a solution we can immediately derive from (6.19) that

$$H(t, x) = (ax + b)^2 + \frac{2}{\upsilon}(ax + b - px - q).$$

The partial derivatives are

$$F_t = p'(t)\, x + q'(t), F_x = p(t), F_{xx} = 0,$$
$$G_t = a'(t)\, x + b'(t), G_x = a(t), G_{xx} = 0,$$

such that the optimal investment candidate (6.20) becomes

$$\pi^* = \frac{\Lambda}{\upsilon\sigma x}\frac{p(t)}{a^2(t)}.$$

Plugging this strategy and the partial derivatives into (6.21) and (6.16) gives the system

$$p'(t)\,x + q'(t) = -rxp(t) - \frac{1}{2}\Lambda^2 \frac{p(t)^2}{\upsilon a(t)^2},$$
$$a'(t)\,x + b'(t) = -rxa(t) - \Lambda^2 \frac{p(t)}{\upsilon a(t)}.$$

Collecting terms with and without x gives

$$p'(t) = -rp(t), p(T) = 1,$$
$$q'(t) = -\frac{1}{2}\Lambda^2 \frac{p(t)^2}{\upsilon a(t)^2}, q(T) = 0,$$
$$a'(t) = -ra(t), a(T) = 1,$$
$$b'(t) = -\frac{\Lambda^2}{\upsilon}\frac{p(t)}{a(t)}, b(T) = 0,$$

with solutions

$$p(t) = e^{r(T-t)},$$
$$a(t) = e^{r(T-t)}.$$

Also,

$$q(t) = b(t)/2 = \frac{\Lambda^2}{2\upsilon}(T-t).$$

The optimal investment strategy finally becomes

$$\pi^*(t,x)\,x = \frac{\Lambda}{\upsilon\sigma}e^{-r(T-t)}.$$

This verifies the result of Basak and Chabakauri (2009b) and Björk and Murgoci (2008).

The optimal strategy consists of putting a nominally increasing dollar amount in the risky asset - on most paths corresponding to a decreasing relative allocation.

A constant υ is not an obvious model choice in that this penalty parameter must necessarily be estimated from the time–0 distribution of

terminal wealth, which in turn depends on time to maturity (and thus calendar time) as well as present wealth. Therefore it could also be updated dynamically as the terminal wealth distribution changes as a result of market dynamics (and deterministically changing time to maturity). That is, υ could depend on x, and possibly on t. In the case treated above the agent pre–commits to υ but not to the target in the quadratic deviation forming the variance, cf. Section 6.3.4. Section 6.3.4 describes the classical case with pre–commitment to both quantities. Within the framework of the present section Björk *et al.* (2009) found a solution for the special case $\upsilon(x) = \upsilon/x$, where the investor does not pre–commit to any of the two.

6.3.2 Mean–standard deviation

Inspired by the discussion in the preceding section it is natural to modify the problem, seemingly slightly, to penalise with standard deviation instead of variance. In single–period models it is well–known that mean–variance and mean–standard deviation are equivalent – in the sense that the set of risk aversions maps into the same set of controls. As it turns out, this equivalence does not carry over to the dynamic model.

The optimisation problem considered in this section is thus

$$\sup_{\pi} \left(\mathbb{E}_{t,x} \left\{ X(T) \right\} - \upsilon \left(\mathbb{V}_{t,x} \left\{ X(T) \right\} \right)^{\frac{1}{2}} \right).$$

To our knowledge this problem has not been studied before, but our extension of Björk and Murgoci (2008) makes it open to investigation.

The problem corresponds to the function f given by

$$f(t,x,y,z) = y - \upsilon \left(z - y^2 \right)^{\frac{1}{2}}, \tag{6.22}$$

with

$$\begin{aligned}
f_t &= f_x = f_{xx} = f_{xy} = f_{xz} = 0,\\
f_y &= 1 + y\upsilon\left(z - y^2\right)^{-\frac{1}{2}},\\
f_{yy} &= \upsilon\left(z - y^2\right)^{-\frac{1}{2}} + y^2\upsilon\left(z - y^2\right)^{-\frac{3}{2}},\\
&= \upsilon z(z - y^2)^{-\frac{3}{2}},\\
f_z &= -\frac{1}{2}\upsilon\left(z - y^2\right)^{-\frac{1}{2}},\\
f_{zz} &= \frac{1}{4}\upsilon\left(z - y^2\right)^{-\frac{3}{2}},\\
f_{yz} &= -\frac{1}{2}y\upsilon\left(z - y^2\right)^{-\frac{3}{2}}.
\end{aligned}$$

From (6.10) we can now derive

$$\begin{aligned}
U &= \frac{1}{4}\upsilon\left(H - G^2\right)^{-\frac{3}{2}} H_x^2 + \left(\upsilon\left(H - G^2\right)^{-\frac{1}{2}} + G^2\upsilon\left(H - G^2\right)^{-\frac{3}{2}}\right) G_x^2\\
&- G\upsilon\left(H - G^2\right)^{-\frac{3}{2}} G_x H_x\\
&= \upsilon\left(H - G^2\right)^{-\frac{1}{2}}\left(\frac{1}{4}\left(H - G^2\right)^{-1}\left(H_x - 2GG_x\right)^2 + G_x^2\right)\\
&= \upsilon(H - G^2)^{-\frac{3}{2}}\left(HG_x^2 - GG_xH_x + H_x^2/4\right).
\end{aligned}$$

Plugging $f_t = f_x = 0$ and U into (6.11) and (6.12) we get the following optimal investment candidate and PDE that we need to solve together with (6.16) and (6.17),

$$\pi^* = -\frac{\Lambda}{\sigma x}\frac{F_x}{F_{xx} - \upsilon\left(H - G^2\right)^{-\frac{1}{2}}\left(\frac{1}{4}\left(H - G^2\right)^{-1}\left(H_x - 2GG_x\right)^2 + G_x^2\right)}, \tag{6.23}$$

$$F_t = -rxF_x + \frac{\Lambda^2 F_x^2/2}{F_{xx} - \upsilon\left(H - G^2\right)^{-\frac{1}{2}}\left(\frac{1}{4}\left(H - G^2\right)^{-1}\left(H_x - 2GG_x\right)^2 + G_x^2\right)},$$

and $F(T, x) = x$. We now search for a solution in the form

$$F(t, x) = p(t)\, x, G(t, x) = a(t)\, x, H(t, x) = c(t)\, x^2,$$

with $c \geq a^2$. We know immediately from (6.22) that the following relation must hold

$$p(t) = a(t) - \upsilon \left(c(t) - a^2(t)\right)^{\frac{1}{2}}.$$

The partial derivatives are

$$\begin{aligned}
F_t &= p'(t)\, x, F_x = p(t), F_{xx} = 0, \\
G_t &= a'(t)\, x, G_x = a(t), G_{xx} = 0, \\
H_t &= c'(t)\, x^2, H_x = 2c(t)\, x, H_{xx} = 2c(t),
\end{aligned}$$

such that the function U and optimal investment candidate (6.23) become

$$\begin{aligned}
U(t,x) &= \frac{\upsilon \left(c(t) - a^2(t)\right)^{-\frac{1}{2}} c(t)}{x}, \\
\pi^* &= \frac{\Lambda}{\upsilon\sigma} \frac{p(t)}{\left(c(t) - a^2(t)\right)^{-\frac{1}{2}} c(t)}.
\end{aligned}$$

Plugging this strategy and the partial derivatives into (6.15), (6.16), and (6.17) (and (6.12), (6.13), and (6.14) for the boundary conditions) gives the system

$$\begin{aligned}
p' &= -\left(r + \frac{1}{2}\Lambda\sigma\pi^*\right) p,\ p(T) = 1, \\
a' &= -\left(r + \Lambda\sigma\pi^*\right) a,\ a(T) = 1, \\
c' &= -\left(2\left(r + \Lambda\sigma\pi^*\right) + \left(\sigma\pi^*\right)^2\right) c, c(T) = 1.
\end{aligned}$$

Surprisingly, the solution is $\pi^* = 0$ via $c = a^2$. Note that for this solution, actually U is infinite. However, since $\pi^* U$ is finite, the solution is valid. For this solution,

$$\begin{aligned}
p' &= -rp,\ p(T) = 1, \\
a' &= -ra,\ a(T) = 1, \\
c' &= -2rc,\ (T) = 1,
\end{aligned}$$

such that

$$\begin{aligned}
p = a &= e^{r(T-t)}, \\
c &= e^{2r(T-t)}.
\end{aligned}$$

This can also be seen from deriving a differential equation for π which gives a DE in the form

$$\begin{aligned}(\pi^*)' &= k_1(t)\pi^* + k_2(\pi^*)^2 + k_3(\pi^*)^3\\ \pi^*(T) &= 0,\end{aligned}$$

with solution

$$\pi^* = 0.$$

This of course makes the case less interesting, although even this is an important insight.

The intuition behind this result is as follows: When the magnitude of the deviations from the mean is smaller than unity, standard deviation punishes these deviations more than does variance. Over an infinitesimal time interval, dt, standard deviation is of order $\sqrt{dt}$, which means that the punishment is so hard that any risk taking is unattractive.

6.3.3 Endogenous habit formation

In this section we consider the optimisation problem

$$\inf_{\pi}\left(\mathbb{E}_{t,x}\left\{\frac{1}{2}\left(X^{\pi}(T) - x\beta(t)\right)^2\right\}\right).$$

This setup is relevant when investors have a time dependent return target, β. To our knowledge the result is new.

The problem corresponds to the function f given by

$$\begin{aligned}f(t,x,y,z) &= -\frac{1}{2}z - \frac{1}{2}x^2\beta^2 + x\beta y \qquad (6.24)\\ f_t &= -x^2\beta\beta' + xy\beta',\\ f_x &= -x\beta^2 + \beta y, f_{xx} = -\beta^2,\\ f_y &= x\beta, f_{xy} = \beta,\\ f_{yz} &= f_{yy} = f_{xz} = f_{zz} = 0.\end{aligned}$$

From (6.10) we can now derive

$$U = 2\beta G_x - \beta^2.$$

Plugging U into (6.11) and (6.12) we get the following optimal investment candidate and PDE that we need to solve together with (6.16) and (6.17),

$$\pi^* = -\frac{\Lambda}{\sigma x}\frac{F_x + x\beta^2 - \beta G}{F_{xx} + \beta^2 - 2\beta G_x}, \tag{6.25}$$

$$F_t = -rx\left(F_x + \beta^2 x - \beta G\right) + \frac{\Lambda^2}{2}\frac{\left(F_x + \beta^2 x - \beta G\right)^2}{F_{xx} + \beta^2 - 2\beta G_x} - x^2\beta\beta' + x\beta' G,$$

and $F(T,x) = -\frac{x^2}{2}\left(1 - \beta(T)\right)^2$. We now search for a solution in the form

$$F(t,x) = \frac{1}{2}p(t)\,x^2, G(t,x) = a(t)\,x, H(t,x) = c(t)\,x^2,$$

with $p < 2a\beta - \beta^2$, and $a(T) = c(T) = 1$. We know immediately from (6.24) that the following relation must hold

$$p(t) = 2\beta(t)\,a(t) - c(t) - \beta^2(t), \tag{6.26}$$

for $c > 0$. The partial derivatives are

$$\begin{aligned}
F_t &= \frac{1}{2}p'(t)\,x^2, F_x = p(t)\,x, F_{xx} = p(t),\\
G_t &= a'(t)\,x, G_x = a(t), G_{xx} = 0,\\
H_t &= c'(t)\,x^2, H_x = 2c(t)\,x, H_{xx} = 2c(t),
\end{aligned}$$

such that the function U and optimal investment candidate (6.25) becomes, using (6.17),

$$\begin{aligned}
U(t,x) &= 2\beta(t)\,a(t) - \beta^2(t),\\
\pi^*(t) &= -\frac{\Lambda}{\sigma}\frac{p(t) - \beta^2(t) + \beta(t)\,a(t)}{p(t) - \beta^2(t) + 2\beta(t)\,a(t)}\\
&= \frac{\Lambda}{\sigma}\frac{\beta(t)\,a(t) - c(t)}{c(t)},
\end{aligned}$$

where, in the last equation we use (6.26).

Plugging this strategy and the partial derivatives into (6.15), (6.16), and (6.17) gives the system

$$\begin{aligned}
\frac{1}{2}p' &= -\left(r + \Lambda\sigma\pi^*/2\right)(a\beta - c) + \beta'(a - \beta),\\
a' &= -\left(r + \Lambda\sigma\pi^*\right)a,\\
c' &= -\left(2\left(r + \Lambda\sigma\pi^*\right) + \left(\sigma\pi^*\right)^2\right)c.
\end{aligned}$$

We can derive the following ODE for π. This is important because then we do not have to calculate a and c in order to derive π^*.

$$\begin{aligned}
\pi^{*\prime} &= \frac{\Lambda}{\sigma}\frac{c\left(\beta' a+\beta a'-c'\right)-c'\left(\beta a-c\right)}{c^2} \\
&= \frac{\Lambda}{\sigma}\frac{\beta' a+\beta a'-\frac{c'}{c}\beta a}{c} \\
&= \frac{\Lambda}{\sigma}\left(\frac{\beta'}{\beta}+r+\Lambda\sigma\pi^*+\left(\sigma\pi^*\right)^2\right)\frac{\beta a}{c} \\
&= \frac{\Lambda}{\sigma}\left(\frac{\beta'}{\beta}+r+\Lambda\sigma\pi^*+\left(\sigma\pi^*\right)^2\right)\left(\pi^*\left(t\right)\frac{\sigma}{\Lambda}+1\right) \\
&= \left(\frac{\beta'}{\beta}+r+\Lambda\sigma\pi^*+\left(\sigma\pi^*\right)^2\right)\left(\pi^*\left(t\right)+\frac{\Lambda}{\sigma}\right) \\
&= k_0(t)+k_1(t)\pi^*(t)+k_2\pi^*(t)^2+k_3\pi^*(t)^3,
\end{aligned}$$

with $k_0(t)=(\beta'/\beta+r)\Lambda/\sigma$, $k_1(t)=\beta'/\beta+r+\Lambda^2$, $k_2=2\Lambda\sigma$, and $k_3=\sigma^2$. The boundary condition is $\pi^*(T)=\Lambda\left(\beta(T)-1\right)/\sigma$.

Because of the terms $\left(\beta'/\beta+r\right)\Lambda/\sigma$ the solution is not zero, although $\pi^*\left(T\right)=0$ for $\beta\left(T\right)=1$, which is the more meaningful value for $\beta(T)$. The quantity $-\beta'/\beta$ represents the the target rate of return of the investor. Therefore it is reasonable to let $-\beta'/\beta$ be a constant larger than r. If $-\beta'/\beta=r$, then the optimal strategy *is* zero, precisely because this target can be obtained via a full allocation to the bank account.

An example of the optimal strategy can be seen in Figure 6.1, which assumes a required rate of return of r+2%. For comparison, the optimal control in the corresponding pre–commitment case (formalised by (6.27) below with $\beta=x_0\exp\left((0.02+r)T\right)$) is (for $T=50$, $\Lambda=\sigma$) *initially* $\pi^*(0,x_0)=\left(e^{0.02T}-1\right)\Lambda/\sigma\approx 172\%$, but (otherwise) path–dependent. e.g. the optimal allocation tends to zero if performance is good, and vice versa – in contrast to the case presented here.

6.3.4 Mean–variance optimisation with pre–commitment

In this section we consider the optimisation problem formalised by

$$V\left(t,x\right)=\sup_{\pi}\mathbb{E}_{t,x}\left\{-\frac{1}{2}\left(X^{\pi}\left(T\right)-\beta\right)^2\right\} \tag{6.27}$$

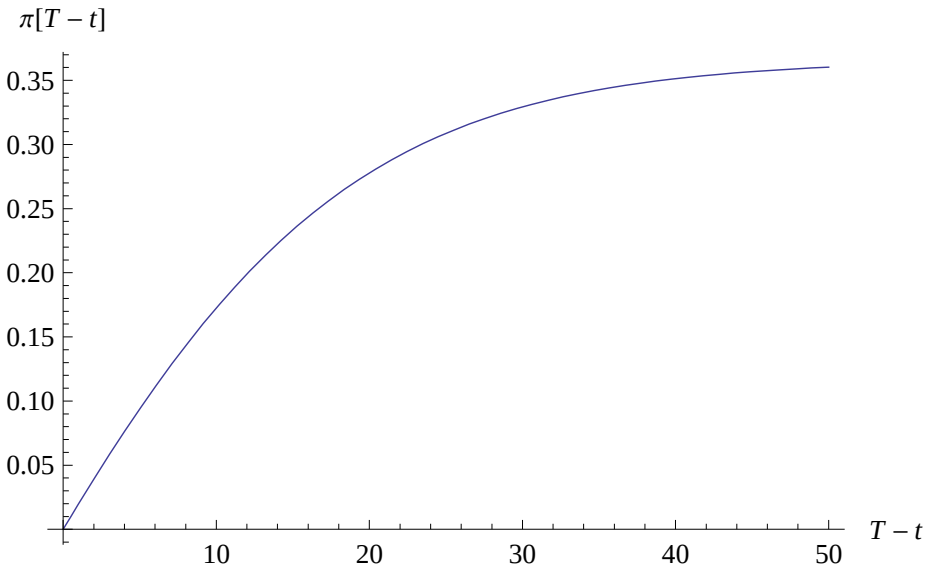

Figure 6.1: Optimal allocation to risky assets for an investor with endogenous habit formation and quadratic utility with required return $-\beta'/\beta = r + 0.02$. The market is $(\Lambda, \sigma) = (0.2, 0.2)$

for a constant β.

We start out by explaining how this problem is the 'first step' in solving a variation of mean–variance utility optimisation, namely 'with pre–commitment'. Consider the problem

$$V(0, x_0) = \sup_{\pi} \left(\mathbb{E}\left\{X^{\pi}(T)\right\} - \frac{\upsilon}{2} \mathbb{V}\left\{X^{\pi}(T)\right\} \right). \tag{6.28}$$

The term pre–commitment refers to the target given implicitly by considering the variance as the quadratic deviation from the target $\mathbb{E}\left\{X^{\pi}(T)\right\}$. One possibility is to actually update this target with (t, x) on the construction of $V(t, x)$. In Section 6.3.1 we updated the target to $\mathbb{E}_{t,x}\left\{X^{\pi}(T)\right\}$ in order to formalise the mean–variance optimisation problem *without* pre–commitment. An alternative is to refrain from updating the target at all. Therefore we say that we pre–commit ourselves to the target $\mathbb{E}_{0,x_0}\left\{X^{\pi}(T)\right\}$ determined at time 0, and we speak of the problem 'with pre–commitment'. This is what we study in this section.

First we write the value function of the problem (6.28) with pre–commitment, i.e. without updating the target

$$V(t,x) = \sup_{\pi} \mathbb{E}_{t,x}\left\{X^{\pi}(T) - \frac{\upsilon}{2}\left(X^{\pi}(T) - \mathbb{E}_{0,x_0}\left\{X^{\pi}(T)\right\}\right)^2\right\}. \quad (6.29)$$

This can be rewritten as

$$\begin{aligned} V(t,x) &= \sup_{\pi,K:\mathbb{E}_{0,x_0}\left\{X^{\pi(K)}(T)\right\}=K} \mathbb{E}_{t,x}\left\{X^{\pi}(T) - \frac{\upsilon}{2}\left(X^{\pi}(T) - K\right)^2\right\} \\ &= \sup_{\pi,K:\mathbb{E}_{0,x_0}\left\{X^{\pi(K)}(T)\right\}=K} \mathbb{E}_{t,x}\left\{-\frac{\upsilon}{2}X^{\pi}(T)^2 + (1+\upsilon K)X^{\pi}(T)\right. \\ &\quad \left. -\frac{\upsilon}{2}K^2\right\}. \end{aligned} \quad (6.30)$$

The optimisation over π and K can be decomposed in two steps: One solves the optimisation problem for a general K and finds the optimal strategy $\pi^*(K)$. Then one calculates $\mathbb{E}_{0,x_0}\left\{X^{\pi^*(K)}(T)\right\}$ and determines the optimal K^* as the solution to the nonlinear equation $\mathbb{E}_{0,x_0}\left\{X^{\pi^*(K^*)}(T)\right\} = K^*$. The solution (π^*, K^*) solves the problem formalised by (6.29).

Rewriting

$$\begin{aligned} V(t,x) &= \upsilon \sup_{\pi,K:\mathbb{E}_{0,x_0}\left\{X^{\pi(K)}(T)\right\}=K} \mathbb{E}_{t,x}\left\{-\frac{1}{2}\left(X^{\pi}(T) - \left(\frac{1}{\upsilon}+K\right)\right)^2\right\} \quad (6.31) \\ &\quad + \frac{1}{2\upsilon} + K \\ &= \upsilon \sup_{\pi,\beta:\mathbb{E}_{0,x_0}\left\{X^{\pi(\beta)}(T)\right\}=\beta-\frac{1}{\upsilon}} \mathbb{E}_{t,x}\left\{-\frac{1}{2}\left(X^{\pi}(T)-\beta\right)^2\right\} + \beta - \frac{1}{2\upsilon}, \end{aligned}$$

gives us that solving (6.27) is the first step of solving (6.29). The second step is to solve

$$\mathbb{E}_{0,x_0}\left\{X^{\pi(\beta)}(T)\right\} = \beta - \frac{1}{\upsilon} \quad (6.32)$$

for β and plug the solution β^* back into π^*. The problem (6.27) corre-

sponds to the function f given by

$$\begin{aligned} f(t,x,y,z) &= -\left(\frac{1}{2}z + \frac{1}{2}\beta^2 - \beta y\right), \\ f_y &= \beta, f_z = -\frac{1}{2}, \\ f_t &= f_x = f_{xx} = f_{yy} = f_{zz} = f_{xy} = f_{xz} = f_{yz} = 0. \end{aligned}$$

Since all the double derivatives of f are zero we get from (6.10) that $U = 0$. Plugging $f_t = f_x = U = 0$ into (6.11) and (6.12) we get the following optimal investment candidate (and corresponding PDE that we need to solve),

$$\pi^* = -\frac{\Lambda}{\sigma x}\frac{F_x}{F_{xx}}, \tag{6.33}$$

$$F_t = -rxF_x + \frac{1}{2}\Lambda^2\frac{F_x^2}{F_{xx}}, \tag{6.34}$$

$$F(T,x) = -\frac{1}{2}(x-\beta)^2.$$

In this particular case the PDE for F does not involve G and H and therefore we do not need to pay attention to the PDEs for G and H. After having derived the solution to (6.34), this is plugged into (6.33) to form the optimal investment strategy as a function of (t,x). Plugging this strategy into (6.16) and (6.17) results in PDEs characterising G and H. However, we do not need these characterisations in order to find the optimal investment and the value function.

We now search for a solution in the form

$$F(t,x) = \frac{1}{2}p(t)(x-q(t))^2$$

The partial derivatives are

$$\begin{aligned} F_t &= \frac{1}{2}p'(t)(x-q(t))^2 - q'(t)p(t)(x-q(t)), \\ F_x &= p(t)(x-q(t)), \\ F_{xx} &= p(t), \end{aligned}$$

such that the optimal investment candidate (6.33) becomes

$$\pi^* x = \frac{\Lambda}{\sigma}\left(q\left(t\right) - x\right). \tag{6.35}$$

Plugging the partial derivatives into (6.34) gives

$$\begin{aligned}
&\frac{1}{2}p'\left(t\right)\left(x - q\left(t\right)\right)^2 - q'\left(t\right)p\left(t\right)\left(x - q\left(t\right)\right) \\
&= -p\left(t\right)\left(x - q\left(t\right)\right)^2 r - q\left(t\right)p\left(t\right)\left(x - q\left(t\right)\right)r + \frac{1}{2}\Lambda^2 p\left(t\right)\left(x - q\left(t\right)\right)^2.
\end{aligned}$$

Collecting terms with $\left(x - q\left(t\right)\right)^2$ and $\left(x - q\left(t\right)\right)$ gives

$$\begin{aligned}
p'\left(t\right) &= \left(-2r + \Lambda^2\right)p\left(t\right), p\left(T\right) = -1, \\
q'\left(t\right) &= rq\left(t\right), q\left(T\right) = \beta.
\end{aligned}$$

This system has the solutions

$$\begin{aligned}
p\left(t\right) &= -e^{\left(2r - \Lambda^2\right)(T-t)}, \\
q\left(t\right) &= e^{-r(T-t)}\beta.
\end{aligned}$$

The full solution can be found by plugging the control (6.35) into (6.16) and (6.17) and guessing a linear and quadratic solution in x to G and H. One finds that

$$\begin{aligned}
G(t,x) &= \beta\left(1 - e^{-\Lambda^2(T-t)}\right) + xe^{\left(r - \Lambda^2\right)(T-t)}, \\
H(t,x) &= \beta^2\left(1 - e^{-\Lambda^2(T-t)}\right) + x^2 e^{\left(2r - \Lambda^2\right)(T-t)}.
\end{aligned}$$

This is the full solution to the problem without any further specification of β. If we want to solve the mean variance optimisation problem with pre–commitment, what remains is to determine β in accordance with (6.32),

$$\begin{aligned}
G\left(0, x_0\right) = \beta - \frac{1}{\upsilon} &\Leftrightarrow \\
\beta = \frac{1}{\upsilon}e^{\Lambda^2 T} + x_0 e^{rT} &\Rightarrow \\
q\left(t\right) = e^{rt}\left(x_0 + \frac{1}{\upsilon}e^{\left(\Lambda^2 - r\right)T}\right).
\end{aligned}$$

With this representation of q we can now express the optimal wealth and the optimal strategy in terms of the diffusion B. First we note that $q - X^{\pi}$ follows a geometric Brownian motion,

$$\begin{aligned} d\left(q\left(t\right) - X^{\pi^*}\left(t\right)\right) &= \left(r - \Lambda^2\right)\left(q\left(t\right) - X^{\pi^*}\left(t\right)\right)dt \\ &\quad - \Lambda\left(q\left(t\right) - X^{\pi^*}\left(t\right)\right)dB\left(t\right). \end{aligned}$$

The solution is

$$\begin{aligned} q\left(t\right) - X^{\pi^*}\left(t\right) &= \left(q\left(0\right) - x_0\right)e^{\left(r - \Lambda^2 - \frac{1}{2}\Lambda^2\right)t - \Lambda B(t)} \\ &= \frac{1}{\upsilon}e^{\left(\Lambda^2 - r\right)(T-t)}e^{-\frac{1}{2}\Lambda^2 t - \Lambda B(t)}, \end{aligned}$$

such that

$$\begin{aligned} X^{\pi^*}\left(t\right) &= q\left(t\right) - \frac{1}{\upsilon}e^{\left(\Lambda^2 - r\right)(T-t)}e^{-\frac{1}{2}\Lambda^2 t - \Lambda B(t)} \\ &= x_0 e^{rt} + \frac{1}{\upsilon}\left(e^{\Lambda^2 T}e^{-r(T-t)} - e^{\left(\Lambda^2 - r\right)(T-t)}e^{-\frac{1}{2}\Lambda^2 t - \Lambda B(t)}\right). \end{aligned}$$

Specifically,

$$X^{\pi^*}\left(T\right) = x_0 e^{rT} + \frac{1}{\upsilon}\left(e^{\Lambda^2 T} - e^{-\frac{1}{2}\Lambda^2 T - \Lambda B(T)}\right). \tag{6.36}$$

In continuation,

$$\begin{aligned} \pi^* X^{\pi^*}\left(t\right) &= \frac{\Lambda}{\sigma}\left(q\left(t\right) - X^{\pi^*}\left(t\right)\right) \\ &= \frac{\Lambda}{\sigma}\left(\frac{1}{\upsilon}e^{\left(\Lambda^2 - r\right)(T-t)}e^{-\frac{1}{2}\Lambda^2 t - \Lambda B(t)}\right). \end{aligned} \tag{6.37}$$

The exponential terms in (6.36) and (6.37) containing the Brownian motion are recognised as as the state price density process times e^{rt}, we can of course express the optimal terminal wealth and the optimal strategy in terms of this process instead. Then the solution in (6.37) is recognised as the classical solution, see e.g. Basak and Chabakauri (2009b), their formulas (37) and (38). The state price representation comes out directly when using the martingale method. In our solution the optimal wealth

process can be rewritten only after recognising the connection between these processes.

The presence of the Brownian motion, or equivalently x_0, in $\pi^* X^{\pi^*}(t)$ shows that the solution is time–inconsistent.

It is easily verified from (6.35) that both the optimal proportion and the optimal amount invested in stocks is decreasing in wealth. This is a well–known feature and one of the main arguments for not being convinced about the objective concerning practical applications. Actually, this problematic feature is one of the reasons for hunting for alternatives like we did in the preceding sections.

We conclude by a remark on the mean–variance optimisation problem formalised by

$$\inf_{\pi:\mathbb{E}\{X^\pi(T)\}\geq K} \mathbb{V}\left\{X^\pi(T)\right\}.$$

We argue that this problem is equivalent to the problem studied above. By rewriting the problem in terms of a Lagrange multiplier,

$$\begin{aligned} V(0,x_0) &= \inf_{\pi,\lambda:\mathbb{E}\{X^\pi(T)\}=K} \mathbb{E}\left\{\left(X^\pi(T) - \mathbb{E}\left\{X^\pi(T)\right\}\right)^2 - \lambda X^\pi(T)\right\} \\ &= \inf_{\pi,\lambda:\mathbb{E}\{X^\pi(T)\}=K} \mathbb{E}\left\{\left(X^\pi(T) - K\right)^2 - \lambda X^\pi(T)\right\} \\ &= \inf_{\pi,\lambda:\mathbb{E}\{X^\pi(T)\}=K} \mathbb{E}\left\{X^\pi(T)^2 - (2K+\lambda) X^\pi(T) + K^2\right\}. \end{aligned}$$

Now we form the value function with pre–commitment,

$$V(t,x) = \inf_{\pi,\lambda:\mathbb{E}\{X^\pi(T)\}=K} \mathbb{E}_{t,x}\left\{X^\pi(T)^2 - (2K+\lambda) X^\pi(T) + K^2\right\},$$

where the term pre–commitment refers to the fact that the target K equals $\mathbb{E}\left\{X^\pi(T)\right\}$ rather than $\mathbb{E}_{t,x}\left\{X^\pi(T)\right\}$. This problem is essentially equivalent to the problem formalised by

$$V(t,x) = \sup_{\pi,\lambda:\mathbb{E}\{X^\pi(T)\}=K} \mathbb{E}_{t,x}\left\{-\frac{1}{2}\left(X^\pi(T) - (K+\lambda/2)\right)^2\right\}. \qquad (6.38)$$

But this problem is equivalent to the problem (6.31). In (6.31) the parameter υ is fixed and the target K is subject to the constraint. In (6.38) the target K is fixed and the parameter λ is subject to the constraint. So, the optimal portfolios arising from different values of υ in (6.31) correspond to the optimal portfolios arising from different values of K in (6.38).

6.4 Collective objectives

In this section we apply Proposition 6.1 to a new set of problems that arise for a collective of heterogenous investors. We think of a group of n investors who, despite their different attitudes towards risk, invest in the same mutual fund. The task is to form an optimal investment strategy for this mutual fund. Such a study is e.g. relevant for compulsory pension schemes.

For simplicity we assume that all investors participate in the fund over the same period. Also, they share the same beliefs about the financial market. At the end of the optimisation horizon the terminal wealth $X^{\pi}(T)$ is distributed such that investor i receives $\alpha_i X^{\pi}(T)$. The constant $\alpha_i \in (0,1)$ represent his relative stake in the collective. Agents may be entitled to unequal proportions (e.g. due to different contributions). Thus, the risk *sharing* is fixed and is not subject to optimisation. However, the aggregate wealth $X^{\pi}(T)$ is not fixed and is subject to optimisation via the investment strategy π. It is important to understand that we are considering the problem of optimal investment for a group of investors which is marginal to the total number of investors in the economy. Therefore, there is no equilibrium theory or asset price formation taking place here. Equilibrium asset prices are given and this marginal group of investors with heterogenous risk aversions plays the investment game together for one reason or another (e.g. in order to save on transaction costs (widely defined) or because they are forced to).

The question is now, what is the objective of the group. A first naive idea is to add up the indirect utility from each investor to achieve the value function.

$$\sup_{\pi} \sum_{i=1}^{n} \mathbb{E}_{t,x} \left\{ u_i \left(\alpha_i X^{\pi}(T) \right) \right\} = \sup_{\pi} \mathbb{E}_{t,x} \left\{ \sum_{i=1}^{n} u_i \left(\alpha_i X^{\pi}(T) \right) \right\}. \tag{6.39}$$

This problem can in principle be solved via standard techniques, but it suffers from serious drawbacks: There is no economic point in adding up different utility functions. For each investor, the utility function expresses his preferences, but it is merely ordinal. Thus, since the utility functions are not comparable, they tell nothing about preferences across the group of investors. A simple check of economic reasonability is the unit of the terms

in the sum. For heterogenous investors we are adding up different functions of the currency unit, and this is also a warning that the formulation (6.39) is completely useless.

The idea that we will introduce here is to align each investor's indirect utility before summation by calculating his certainty equivalent. Thus, we propose instead the formalisation

$$\sup_{\pi} \sum_{i=1}^{n} u_i^{-1} \left(\mathbb{E}_{t,x} \left\{ u_i \left(\alpha_i X^{\pi} \left(T \right) \right) \right\} \right). \tag{6.40}$$

This makes economic sense: At time t we are adding up certain time t-amounts which are definitely comparable. From a mathematical point of view, though, the problem (6.40) seems more challenging, due to the non-linearity of the utility functions, but our Proposition 6.1 is able to cope with that.

We re–emphasise that the proportional division of terminal wealth is pre-imposed, so it is not possible to increase group utility by assigning all wealth to the more risk-tolerant agent. There may exist more optimal risk sharing rules - especially should one know more about the agents' endowments. Still, the simple rule that we have outlined is highly relevant from a practical perspective.

6.4.1 A collective of exponential utility investors

For exponential utility with coefficients of absolute risk aversion $\xi_i > 0$, $n = 2$, and $\alpha_1 = \alpha_2 = 1/2$, the problem (6.40) is

$$\sup_{\pi} \left(\frac{-1}{\xi_1} \log \mathbb{E} \left\{ e^{-\xi_1 X^{\pi}(T)/2} \right\} + \frac{-1}{\xi_2} \log \mathbb{E} \left\{ e^{-\xi_2 X^{\pi}(T)/2} \right\} \right).$$

This corresponds to the function f given by

$$\begin{aligned} f\left(y,z\right) &= -\frac{\log y}{\xi_1} - \frac{\log z}{\xi_2}, \\ f_t &= f_x = f_{xx} = f_{xy} = f_{xz} = f_{yz} = 0, \\ f_y &= -\frac{1}{\xi_1 y}, f_{yy} = \frac{1}{\xi_1 y^2}, \\ f_z &= -\frac{1}{\xi_2 z}, f_{zz} = \frac{1}{\xi_2 z^2}. \end{aligned}$$

From (6.10) we can now derive

$$U = \frac{1}{\xi_1}\left(\frac{G_x}{G}\right)^2 + \frac{1}{\xi_2}\left(\frac{H_x}{H}\right)^2.$$

Plugging $f_t = f_x = 0$ and U into (6.11) we get the following optimal candidate,

$$\pi^* = -\frac{\Lambda}{\sigma x}\frac{F_x}{F_{xx} - \frac{1}{\xi_1}\left(\frac{G_x}{G}\right)^2 - \frac{1}{\xi_2}\left(\frac{H_x}{H}\right)^2}. \tag{6.41}$$

With this specification of π^* we now search for a solution to (6.15), (6.16), and (6.17) in the form

$$F(t,x) = p(t)\,x + q(t), G(t,x) = e^{g_1(t)x+g_2(t)}, H(t,x) = e^{h_1(t)x+h_2(t)}.$$

The partial derivatives are

$$\begin{aligned}
F_t &= p'(t)\,x + q'(t), F_x = p(t), F_{xx} = 0,\\
G_t &= e^{g_1(t)x+g_2(t)}\left(g_1'(t)\,x + g_2'(t)\right),\\
G_x &= e^{g_1(t)x+g_2(t)}g_1(t), G_{xx} = e^{g_1(t)x+g_2(t)}g_1^2(t),\\
H_t &= e^{h_1(t)x+h_2(t)}\left(h_1'(t)\,x + h_2'(t)\right),\\
H_x &= e^{h_1(t)x+h_2(t)}h_1(t), H_{xx} = e^{h_1(t)x+h_2(t)}h_1^2(t),
\end{aligned}$$

such that the optimal investment candidate (6.41) becomes

$$\pi^* x = \frac{\Lambda}{\sigma}\frac{p(t)}{\frac{1}{\xi_1}g_1^2(t) + \frac{1}{\xi_2}h_1^2(t)}. \tag{6.42}$$

Plugging this strategy into (6.15), (6.16), and (6.17), leads to ordinary differential equations (ODEs) for p, g_1, and h_1, with terminal conditions and solutions

$$\begin{aligned}
p'(t) &= -rp(t); p(T) = 1: \ p(t) = e^{r(T-t)},\\
g_1'(t) &= -rg_1(t); g_1(T) = -\frac{\xi_1}{2}: \ g_1(t) = -\frac{\xi_1}{2}e^{r(T-t)},\\
h_1'(t) &= -rh_1(t); h_1(T) = -\frac{\xi_2}{2}: \ h_1(t) = -\frac{\xi_2}{2}e^{r(T-t)}.
\end{aligned}$$

Plugging these into (6.42) yields

$$\pi^* x = 2\frac{\Lambda}{\sigma}\frac{e^{-r(T-t)}}{\bar{\xi}},$$

with $\bar{\xi} = \sum_{i=1}^{n} \xi_i/n$ defining the average risk aversion.

This strategy may be compared to the classical solution for a single investor with risk aversion ξ who invests optimally the amount

$$\frac{\Lambda}{\sigma}\frac{e^{-r(T-t)}}{\xi}.$$

We see that the collective of investors calculates an average absolute risk aversion coefficient $\bar{\xi}$, and then invests two times the amount that such an average investor would, i.e. one (time) for each participant. Notice that this strategy is not the simple average of individually optimal strategies.

For the full solution we also solve the ODEs for q, g_2, and h_2, and get

$$\begin{aligned} q(t) &= \frac{n\Lambda^2}{2\bar{\xi}}(T-t), \\ g_2(t) &= -\frac{\Lambda^2(T-t)\xi_1}{2\bar{\xi}^2}\left(2\bar{\xi} - \xi_1\right) \\ h_2(t) &= -\frac{\Lambda^2(T-t)\xi_2}{2\bar{\xi}^2}\left(2\bar{\xi} - \xi_2\right). \end{aligned}$$

The group–optimal discounted certainty equivalent is thus

$$\begin{aligned} e^{-r(T-t)}F(t,x) &= n\left(\frac{x}{n} + e^{-r(T-t)}\frac{\Lambda^2}{2\bar{\xi}}(T-t)\right) \\ &= e^{-r(T-t)}\sum_{i=1}^{n}\frac{-1}{\xi_i}\left[\frac{-\xi_i}{n}e^{r(T-t)}x - \frac{\Lambda^2(T-t)\xi_i}{2\bar{\xi}^2}\left(2\bar{\xi} - \xi_i\right)\right] \\ &= \sum_{i=1}^{n}\left[\frac{x}{n} + e^{-r(T-t)}\frac{\Lambda^2(T-t)}{2\bar{\xi}^2}\left(2\bar{\xi} - \xi_i\right)\right], \end{aligned}$$

with the individual terms in the sum corresponding to the i^{th} individual's certainty equivalent. On the other hand, if each individual invests on his own, he obtains the comparable optimal discounted certainty equivalent

$$\frac{x}{n} + e^{-r(T-t)}\frac{\Lambda^2}{2\xi_i}(T-t),$$

such that his relative loss (of discounted certainty equivalent after subtraction of x/n) from entering the collective is

$$1 - \frac{e^{-r(T-t)}\xi_i^{-1}\frac{\Lambda^2(T-t)\xi_i}{2\bar{\xi}^2}\left(2\bar{\xi}-\xi_i\right)}{e^{-r(T-t)}\frac{\Lambda^2}{2\xi_i}(T-t)} = \left(1-\frac{\xi_i}{\bar{\xi}}\right)^2,$$

which could be compared to the estimated gains from economies of scale. These losses are – unsurprisingly – independent of initial wealth. In the present case of two investors they both lose the same proportion, but the formulae actually hold when there are more agents. Then some can be hit substantially harder than others (and some may not suffer at all, of course).

The results above can easily be extended to the case of n investors with coefficients $\xi_1, \ldots \xi_n$, who put up proportions $\alpha_1, \ldots, \alpha_n$ (with all $\alpha_i > 0$, and $\sum_{i=1}^n \alpha_i = 1$). The optimal strategy is

$$\begin{aligned}\pi^* x &= n\frac{\Lambda}{\sigma}\frac{e^{-r(T-t)}}{\sum_{i=1}^n n\alpha_i^2\xi_i} \\ &= n\frac{\Lambda}{\sigma}\frac{e^{-r(T-t)}}{\bar{\xi}}\frac{\sum_{i=1}^n \xi_i}{\sum_{i=1}^n (n\alpha_i)^2\xi_i},\end{aligned}$$

where the former expression shows that a "representative" agent has risk aversion $\sum_{i=1}^n n\alpha_i^2\xi_i$, and the latter expression demonstrates how the standard solution is to be multiplied by a correction factor, which depends on how different the stakes are.

As in the mean–variance case (Section 6.3.1) it is relevant to let the coefficients of absolute risk aversion depend on t, x. However, our methodology cannot cope with this setting. For a single investor the example is a special case of Björk and Murgoci (2008).

6.4.2 A collective of power utility investors

As another, and perhaps more interesting, example consider a collective of power utility investors with strictly positive coefficients of relative risk aversion γ_i. For illustration we consider a small collective with $n = 2$, and

with equal proportions $\alpha_1 = \alpha_2 = 1/2$. Then the problem (6.40) is[1]

$$\sup_{\pi} \frac{1}{2}\left[\left(\mathbb{E}_{t,x}\left\{\left(X^{\pi}(T)\right)^{1-\gamma_1}\right\}\right)^{(1-\gamma_1)^{-1}} + \left(\mathbb{E}_{t,x}\left\{\left(X^{\pi}(T)\right)^{1-\gamma_2}\right\}\right)^{(1-\gamma_2)^{-1}}\right].$$

This corresponds to the function f given by

$$\begin{aligned} f(y,z) &= \frac{1}{2}\left(y^{(1-\gamma_1)^{-1}} + z^{(1-\gamma_2)^{-1}}\right), \qquad (6.43)\\ f_t &= f_x = f_{xx} = f_{xy} = f_{xz} = f_{yz} = 0,\\ f_y &= \frac{1}{2}(1-\gamma_1)^{-1} y^{\frac{\gamma_1}{1-\gamma_1}}, f_{yy} = \frac{1}{2}\frac{\gamma_1}{(1-\gamma_1)^2} y^{\frac{2\gamma_1-1}{1-\gamma_1}},\\ f_z &= \frac{1}{2}(1-\gamma_2)^{-1} z^{\frac{\gamma_2}{1-\gamma_2}}, f_{zz} = \frac{1}{2}\frac{\gamma_2}{(1-\gamma_2)^2} z^{\frac{2\gamma_2-1}{1-\gamma_2}}. \end{aligned}$$

From (6.10) we can now derive

$$U = \frac{1}{2}\frac{\gamma_1}{(1-\gamma_1)^2} G^{(1-\gamma_1)^{-1}}\left(\frac{G_x}{G}\right)^2 + \frac{1}{2}\frac{\gamma_2}{(1-\gamma_2)^2} H^{(1-\gamma_2)^{-1}}\left(\frac{H_x}{H}\right)^2.$$

Plugging $f_t = f_x = 0$ and U into (6.11) we get the following optimal candidate,

$$\pi^* = -\frac{\Lambda}{\sigma x}\frac{F_x}{F_{xx} - \frac{1}{2}\left(\frac{\gamma_1}{(1-\gamma_1)^2} G^{(1-\gamma_1)^{-1}}\left(\frac{G_x}{G}\right)^2 + \frac{\gamma_2}{(1-\gamma_2)^2} H^{(1-\gamma_2)^{-1}}\left(\frac{H_x}{H}\right)^2\right)}. \qquad (6.44)$$

With this specification of π^* we now search for a solution to (6.15), (6.16), and (6.17) in the form

$$F(t,x) = p(t)\,x, G(t,x) = a^{1-\gamma_1}(t)\,x^{1-\gamma_1}, H(t,x) = c^{1-\gamma_2}(t)\,x^{1-\gamma_2},$$

The partial derivatives are

$$\begin{aligned} F_t &= p'(t)\,x, F_x = p(t), F_{xx} = 0,\\ G_t &= (1-\gamma_1)a^{-\gamma_1}(t)\,a'(t)\,x^{1-\gamma_1},\\ G_x &= (1-\gamma_1)a^{1-\gamma_1}(t)\,x^{-\gamma_1}, G_{xx} = -\gamma_1(1-\gamma_1)a^{1-\gamma_1}(t)\,x^{-\gamma_1-1},\\ H_t &= (1-\gamma_2)c^{-\gamma_2}(t)\,c'(t)\,x^{1-\gamma_2},\\ H_x &= (1-\gamma_2)c^{1-\gamma_2}(t)\,x^{-\gamma_2}, H_{xx} = -\gamma_2(1-\gamma_2)c^{1-\gamma_2}(t)\,x^{-\gamma_2-1}, \end{aligned}$$

[1]Allowing for $\gamma = 1$ (logarithmic utility) is notationally cumbersome, so we do not treat it formally.

and $p = (a + c)/2$, such that the optimal candidate (6.44) becomes

$$\begin{aligned}\pi^* &= \frac{\Lambda}{\sigma}\frac{a\,(t) + c\,(t)}{a\,(t)\,\gamma_1 + c\,(t)\,\gamma_2} \\ &= \frac{\Lambda}{\sigma}\frac{1}{\gamma\,(t)},\end{aligned}$$

with $\gamma\,(t)$ defined as a weighted average of the underlying coefficients of relative risk aversion – with time–dependent weights,

$$\gamma\,(t) = \frac{a\,(t)\,\gamma_1 + c\,(t)\,\gamma_2}{a\,(t) + c\,(t)}.$$

This formulation means that (in contrary to the exponential case) there can never be an agent, who can be taken to be representative for the collective over the entire period.

Plugging this strategy into (6.15), (6.16), and (6.17), leads to a system of ODEs for p, a, and c, with terminal conditions. The differential equations for a and c can be solved isolated from p and are sufficient for determination of π. We find the following representation in terms of γ,

$$\begin{aligned}a'\,(t) &= -\left(r + \frac{\Lambda^2}{\gamma(t)} - \frac{\Lambda^2}{2}\frac{\gamma_1}{\gamma(t)^2}\right) a(t); a\,(T) = 1, \\ c'\,(t) &= -\left(r + \frac{\Lambda^2}{\gamma(t)} - \frac{\Lambda^2}{2}\frac{\gamma_2}{\gamma(t)^2}\right) c(t); c\,(T) = 1.\end{aligned}$$

We have no explicit solution to the two-dimensional system of ODEs. We can however characterise the solution a bit further by calculating an ODE for the quantity $w = a/(a + c)$, which is the weight on agent 1's coefficient of relative risk aversion in the formation of the group's ditto:

$$w' = w(1 - w)(\gamma_1 - \gamma_2)\frac{\Lambda^2}{2}\left[\gamma_2 + w(\gamma_1 - \gamma_2)\right]^{-2},\ w(T) = 1/2,$$

with the property that for $\gamma_1 < \gamma_2$, w is a decreasing function of time (and thus an increasing function of time to expiry) so that the more risk tolerant agent has the larger weight. For $w > \gamma_2/(\gamma_1 + \gamma_2)$, the weight on agent 1's individually optimal strategy is larger than a half. This is equivalent

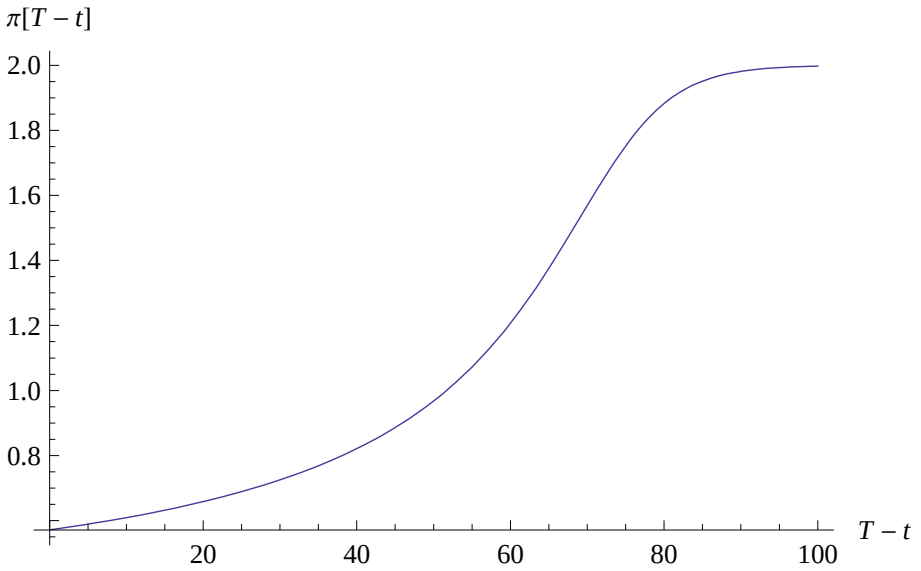

Figure 6.2: Group–optimal allocation to risky assets for a collective formed by two investors with $\gamma_1 = 0.5$, $\gamma_2 = 3$. The market is $(\Lambda, \sigma) = (0.2, 0.2)$. The two would prefer 200% respectively 33.3% allocated to risky assets.

to local concavity of w which will occur for sufficiently long time horizons, because $w \in (0, 1)$.

For n investors n differential equations can be reduced to $n - 1$ using this technique, but the advantage is not nearly as obvious.

When participating in the group, the discounted certainty equivalent of individual 1 is $xa(t) \exp(-r(T-t))$, while as an individual he would be indifferent between participating in the lottery and receiving the amount $x \exp\left(\Lambda^2(T-t)/(2\gamma_1)\right)$. Depending on the measurement of loss one or the other investor is worst off.

An illustration of the development over time of the optimal strategy for the group can be seen in Figure 6.2, whereas Table 6.1 shows the corresponding certainty equivalents, and contrasts them to those of the individuals forming the collective. The figure reveals that the collective's optimal strategy changes rather slowly over time – except during a relatively short transition phase.

	individual	**group**
Agent 1	1.492	1.223
Agent 2	1.069	1.028

Table 6.1: Discounted optimal certainty equivalents (normalised by wealth) with ten years to expiry for agents with relative risk aversions $\gamma_1 = 0.5$, $\gamma_2 = 3$. The second row gives the certainty equivalents obtainable by individuals, and the third row shows the corresponding figures, when the group formed by the two decides the optimal strategy. The market is $(\Lambda, \sigma) = (0.2, 0.2)$.

If agents put up different proportions, say $\alpha \in (0,1)$, and $1-\alpha$, the only modification is that the appropriate average becomes

$$\gamma(t) = \frac{\alpha a(t)\gamma_1 + (1-\alpha)c(t)\gamma_2}{\alpha a(t) + (1-\alpha)c(t)}.$$

6.4.3 A collective of mean–variance utility investors without pre–commitment

We paid a lot of attention to the mean–variance utility investor in the previous section. Let us see what happens if we apply our certainty equivalent approach to a group of heterogenous mean–variance utility investors. This becomes particularly simple in the case without pre–commitment, since the utility inversion just becomes the identity function. We therefore study the problem

$$\begin{aligned}&\sup_{\pi} \sum_{i=1}^{n} \left(\mathbb{E}_{t,x}\left\{ \frac{X^{\pi}(T)}{n} \right\} - \frac{\upsilon_i}{2} \mathbb{V}_{t,x}\left\{ \frac{X^{\pi}(T)}{n} \right\} \right) \\ &= \sup_{\pi} \left(\mathbb{E}_{t,x}\left\{ X^{\pi}(T) \right\} - \frac{\bar{\upsilon}/n}{2} \mathbb{V}_{t,x}\left\{ X^{\pi}(T) \right\} \right),\end{aligned}$$

with $\bar{\upsilon} = \sum_{i=1}^{n} \upsilon_i / n$ defining the average risk aversion. But this problem is equivalent to the problem of a single investor with wealth x and risk aversion $\bar{\upsilon}/n$. The optimal investment strategy then becomes

$$\pi^* x = n \frac{\Lambda}{\sigma} \frac{e^{-r(T-t)}}{\bar{\upsilon}}.$$

This should be compared with the solution for a single investor with risk aversion υ, who invests optimally the amount

$$\pi^* x = \frac{\Lambda}{\sigma} \frac{e^{-r(T-t)}}{\upsilon}.$$

As was the case for exponential utility collectives, we find that the group-optimal amount invested in stocks is found by using the average risk aversion $\bar{\upsilon}$, and then investing this amount for each of the n participants.

Since mean–variance is not a real utility function there need not be a loss associated with joining a group. An individual investor, i, has an optimal discounted certainty equivalent of

$$\frac{x}{n} + e^{-r(T-t)} \frac{\Lambda^2 (T-t)}{\upsilon_i},$$

while as a group member his corresponding "indifference amount" is

$$\frac{x}{n} + e^{-r(T-t)} \frac{\Lambda^2 (T-t)}{\bar{\upsilon}},$$

so that he incurs a loss (again, in certainty–equivalent terms) by joining the group iff $\upsilon_i < \bar{\upsilon}$, i.e if he is less cautious than the group as a whole.

6.4.4 A collective of mean–variance utility investors with pre–commitment

We can also consider the mean–variance utility with pre–commitment *for a collective*. Here it becomes important in which order we implement the different arguments. Does each investors realise that the utility inversion of a mean–variance utility is the identity function *before* he decides to pre–commit himself to his time 0-target? Or does he pre–commit to his time 0-target for thereafter to realise that the utility inversion is no longer just the identity function?

If we implement the identity utility inversion, we get the problem

$$\begin{aligned}
&\sup_{\pi} \sum_{i=1}^{n} \left(\mathbb{E}_{t,x} \left\{ \frac{X^{\pi}(T)}{n} \right\} - \frac{\upsilon_i}{2} \mathbb{E}_{t,x} \left\{ \frac{X^{\pi}(T)}{n} - \mathbb{E}_{0,x_0} \left\{ \frac{X^{\pi}(T)}{n} \right\} \right\}^2 \right) \\
&= \sup_{\pi:\mathbb{E}_{0,x_0}\{X^{\pi}(T)\}=K} \sum_{i=1}^{n} \left(\mathbb{E}_{t,x} \left\{ \frac{X^{\pi}(T)}{n} \right\} - \frac{\upsilon_i}{2} \mathbb{E}_{t,x} \left\{ \frac{X^{\pi}(T)}{n} - \frac{K}{n} \right\}^2 \right) \\
&= \sup_{\pi:\mathbb{E}_{0,x_0}\{X^{\pi}(T)\}=K} \left(\mathbb{E}_{t,x} \left\{ X^{\pi}(T) \right\} - \frac{\bar{\upsilon}}{2n} \mathbb{E}_{t,x} \left\{ X^{\pi}(T) - K \right\}^2 \right) \\
&= \sup_{\pi,h:\mathbb{E}_{0,x_0}\left\{X^{\pi(h)}(T)\right\}=h-\frac{n}{\bar{\upsilon}}} \mathbb{E}_{t,x} \left\{ -\frac{1}{2} \left(X^{\pi}(T) - h \right)^2 \right\}
\end{aligned}$$

with $\bar{\upsilon} = \sum_{i=1}^{n} \upsilon_i / n$ defining the average risk aversion. But this problem is equivalent to the problem of a single investor with wealth x and risk aversion $\bar{\upsilon}/n$. The optimal investment strategy then becomes

$$\pi^* x = \frac{\Lambda}{\sigma} \left(q(t) - x \right)$$

with

$$q(t) = e^{rt} \left(x_0 + \frac{n}{\bar{\upsilon}} e^{\left(\Lambda^2 - r\right)T} \right).$$

This can be compared to the solution for a single investor with risk aversion υ_i and initial wealth x_0/n, who invests optimally the amount

$$\pi^* x/n = \Lambda \left(q_i(t) - x/n \right) / \sigma$$

with

$$q_i(t) = \exp(rt) \left(x_0/n + \exp\left(\left(\Lambda^2 - r \right) T \right) / \upsilon_i \right).$$

We see that the collective of investors calculates an average target process $\bar{q}$ based on the average aversion $\bar{\upsilon}$,

$$\bar{q}(t) = \exp(rt) \left(x_0/n + \exp\left(\left(\Lambda^2 - r \right) T \right) / \bar{\upsilon} \right),$$

and then invests n times this amount, that is

$$\pi^* x = n \frac{\Lambda}{\sigma} \left(\bar{q}(t) - x/n \right) = \frac{\Lambda}{\sigma} \left(q(t) - x \right).$$

The alternative is to start with the pre–commitment such that objective of investor i, before starting the investment collective, is

$$V_i(t,x) = \sup_{\pi,\beta_i:\mathbb{E}\left\{X^{\pi(\beta_i)}(T)\right\}=\beta_i-\frac{1}{\upsilon_i}} \mathbb{E}_{t,x}\left\{-\frac{1}{2}\left(X^{\pi}(T)-\beta_i\right)^2\right\}.$$

Now the utility inversion is no longer the identity function, and dealing with the case turns out to be surprisingly difficult. First we have to assume that $X^{\pi}/n \leq \beta_i$ a.s. for all i. Then the collective of investors faces the problem

$$V(t,x) = \sup_{\pi,\beta_i:\mathbb{E}\left\{\frac{X^{\pi(\beta_i)}(T)}{n}\right\}=\beta_i-\frac{1}{\upsilon_i}} \sum_{i=1}^{n}\left(\beta_i + \sqrt{\mathbb{E}_{t,x}\left\{\left(\frac{X^{\pi}(T)}{n}-\beta_i\right)^2\right\}}\right),$$

which seems intractable.

6.5 Concluding remarks

Björk and Murgoci (2008) point out that to any non-standard problem within their set of study corresponds a standard problem. Here we argue that this also hold in our case. Rearranging the terms of (6.6) yields

$$\begin{aligned} F_t = \inf_{\pi}\Big[&-\left(r+\Lambda\sigma\pi\right)xF_x - \frac{1}{2}\left(\sigma\pi\right)^2x^2F_{xx} \\ &+f_t + \left(r+\Lambda\sigma\pi\right)xf_x + \frac{1}{2}\left(\sigma\pi\right)^2x^2U\Big]. \end{aligned}$$

One can recognise this as the standard Hamilton–Jacobi–Bellmann (HJB) equation to the problem

$$\sup_{\pi}\mathbb{E}_{t,x}\left\{\begin{array}{l} -\int_t^T\left(f_s + \left(r+\Lambda\sigma\pi\right)X(s)f_x + \frac{1}{2}\left(\sigma\pi X(s)\right)^2U(f,y,z)\right)ds \\ +f\left(T,X(T),g\left(X(T)\right),h\left(X(T)\right)\right) \end{array}\right\} \tag{6.45}$$

with appropriate arguments. Björk and Murgoci (2008) calculated specifically the equivalent standard problem for the mean–variance case and their

result can be recognised in (6.45). Formalising the "extra" terms in (6.6) as "utility of consumption" is straightforward here (although nothing is actually consumed), and probably in cases much more involved than ours likewise. However, it is of only marginal interest since we do not know any examples where the standard problem induced by a non-standard problem has a meaningful economic interpretation in its own respect.

In this paper we have concentrated on the pure investment problem. In Björk and Murgoci (2008), consumption is also taken into account. Their preferences over consumption contribute to the inconsistency only via dependence on wealth (like endogenous habit formation). More generally, inconsistency could also arise from taking a non-linear function of the expected utility of consumption. This is a natural subject for further research.

We have completely accepted the non-standard problem as meaningful. The game theoretical foundations for interpretation of the non-standard problem are taken as given in our presentation and we refer to Björk and Murgoci (2008) for theoretical considerations in this regard. Once having accepted the problem as meaningful we are allowed to attack it directly in continuous time with more or less standard control theoretical techniques. Therefore the generalised HJB equation and the examples of its solution stand out as the primary contribution of our paper.

6.6 Proof

Proof of Proposition 6.1. Consider an arbitrary admissible strategy π, and let $U^{\pi} = U\left(f, Y^{\pi}, Z^{\pi}\right)$.

1. First we argue that if there exists a function $Y^{\pi}\left(t,x\right)$ such that

$$Y_t^{\pi} = -\left(r + \Lambda\sigma\pi\right) x Y_x^{\pi} - \frac{1}{2}\left(\sigma\pi\right)^2 x^2 Y_{xx}^{\pi}, \tag{6.46}$$

$$Y^{\pi}\left(T,x\right) = g\left(x\right), \tag{6.47}$$

then

$$Y^{\pi}\left(t,x\right) = y^{\pi}\left(t,x\right). \tag{6.48}$$

Namely,

$$\begin{aligned} Y^{\pi}(t, X(t)) &= -\int_t^T dY^{\pi}(s, X^{\pi}(s)) + Y^{\pi}(T, X^{\pi}(T)) \\ &= -\int_t^T \left(\begin{array}{l} Y_s^{\pi}(s, X^{\pi}(s))\, ds \\ +Y_x^{\pi}(s, X^{\pi}(s)) \left(\begin{array}{l} (r + \Lambda\sigma\pi(s)) X^{\pi}(s)\, ds \\ +\sigma\pi(s) X^{\pi}(s)\, dB(s) \end{array} \right) \\ +\frac{1}{2} Y_{xx}^{\pi}(s, X^{\pi}(s)) (\sigma\pi(s))^2 X^{\pi}(s)^2\, ds \end{array} \right) \\ &\quad + Y^{\pi}(T, X^{\pi}(T)). \end{aligned}$$

Inserting (6.46) and (6.47) gives

$$Y^{\pi}(t, X(t)) = -\int_t^T \sigma\pi(s) X^{\pi}(s)\, dB(s) + g(X^{\pi}(T)). \tag{6.49}$$

Now, taking conditional expectation on both sides gives

$$Y^{\pi}(t, x) = \mathbb{E}_{t,x}\{g(X^{\pi}(T))\} = y^{\pi}(t, x).$$

From similar arguments (replace y and Y by z and Z) we get that if there exists a function $Z^{\pi}(t, x)$ such that

$$Z_t^{\pi} = -(r + \Lambda\sigma\pi) x Z_x^{\pi} - \frac{1}{2}(\sigma\pi)^2 x^2 Z_{xx}^{\pi}, \tag{6.50}$$

$$Z^{\pi}(T, x) = h(x), \tag{6.51}$$

then

$$Z^{\pi}(t, x) = z^{\pi}(t, x). \tag{6.52}$$

2. Second we obtain an expression for

$$f(t, X^{\pi}(t), y^{\pi}(t, X^{\pi}(t)), z^{\pi}(t, X^{\pi}(t))).$$

From (6.48) and (6.52) we have that this equals

$$f(t, X^{\pi}(t), Y^{\pi}(t, X^{\pi}(t)), Z^{\pi}(t, X^{\pi}(t))).$$

Since f is sufficiently differentiable, then by Itô's formula

$$
\begin{aligned}
&f\left(t, X^{\pi}(t), y^{\pi}\left(t, X^{\pi}(t)\right), z^{\pi}\left(t, X^{\pi}(t)\right)\right) \\
&= -\int_{t}^{T} df\left(s, X^{\pi}(s), Y^{\pi}\left(s, X^{\pi}(s)\right), Z^{\pi}\left(s, X^{\pi}(s)\right)\right) \\
&\quad + f\left(T, X^{\pi}(T), Y^{\pi}\left(T, X^{\pi}(T)\right), Z^{\pi}\left(T, X^{\pi}(T)\right)\right) \\
&= -\int_{t}^{T} \left(\begin{array}{l} \left(f_{s}+f_{y} Y_{s}^{\pi}+f_{z} Z_{s}^{\pi}\right) ds \\ +\left(f_{x}+f_{y} Y_{x}^{\pi}+f_{z} Z_{x}^{\pi}\right) dX^{\pi}(s) \\ +\frac{1}{2}\left(\begin{array}{l} f_{xx}+2 f_{xy} Y_{x}^{\pi}+2 f_{xz} Z_{x}^{\pi} \\ +f_{y} Y_{xx}^{\pi}+f_{z} Z_{xx}^{\pi}+f_{yy}\left(Y_{x}^{\pi}\right)^{2} \\ +2 f_{yz} Y_{x}^{\pi} Z_{x}^{\pi}+f_{zz}\left(Z_{x}^{\pi}\right)^{2} \end{array} \right)\left(\sigma \pi(s)\right)^{2} X^{\pi}(s)^{2} ds \end{array} \right) \\
&\quad + f\left(T, X^{\pi}(T), Y^{\pi}\left(T, X^{\pi}(T)\right), Z^{\pi}\left(T, X^{\pi}(T)\right)\right),
\end{aligned}
$$

where we have skipped some arguments under the integral. Inserting (6.46), (6.47), (6.50), (6.51) and (6.2) we have that

$$
\begin{aligned}
&f\left(t, X^{\pi}(t), y^{\pi}\left(t, X^{\pi}(t)\right), z^{\pi}\left(t, X^{\pi}(t)\right)\right) \\
&= -\int_{t}^{T} \left(\begin{array}{l} \left(\begin{array}{l} f_{s}+f_{y}\left(-\left(r+\Lambda \sigma \pi(s)\right) x Y_{x}^{\pi}-\frac{1}{2}\left(\sigma \pi(s)\right)^{2} x^{2} Y_{xx}^{\pi}\right) \\ +f_{z}\left(-\left(r+\Lambda \sigma \pi(s)\right) x Z_{x}^{\pi}-\frac{1}{2}\left(\sigma \pi(s)\right)^{2} x^{2} Z_{xx}^{\pi}\right) \end{array} \right) ds \\ +\left(f_{x}+f_{y} Y_{x}^{\pi}+f_{z} Z_{x}^{\pi}\right)\left(\begin{array}{l} \left(r+\Lambda \sigma \pi(s)\right) X^{\pi}(s) dt \\ +\sigma \pi(s) X^{\pi}(s) dB(s) \end{array} \right) \\ +\frac{1}{2}\left(\begin{array}{l} f_{xx}+2 f_{xy} Y_{x}^{\pi}+2 f_{xz} Z_{x}^{\pi} \\ +f_{y} Y_{xx}^{\pi}+f_{z} Z_{xx}^{\pi}+f_{yy}\left(Y_{x}^{\pi}\right)^{2} \\ +2 f_{yz} Y_{x}^{\pi} Z_{x}^{\pi}+f_{zz}\left(Z_{x}^{\pi}\right)^{2} \end{array} \right)\left(\sigma \pi(s)\right)^{2} X^{\pi}(s)^{2} ds \end{array} \right) \\
&\quad + f\left(T, X^{\pi}(T), g\left(X^{\pi}(T)\right), h\left(X^{\pi}(T)\right)\right).
\end{aligned}
$$

Abbreviating and inserting (6.10) we get

$$
\begin{aligned}
&f\left(t, X^{\pi}(t), y^{\pi}\left(t, X^{\pi}(t)\right), z^{\pi}\left(t, X^{\pi}(t)\right)\right) \\
&= -\int_{t}^{T} \left(\begin{array}{l} f_{s} ds+f_{x}\left(r+\Lambda \sigma \pi(s)\right) X^{\pi}(s) ds \\ +\left(f_{x}+f_{y} Y_{x}^{\pi}+f_{z} Z_{x}^{\pi}\right) \sigma \pi(s) X^{\pi}(s) dB(s) \\ +\frac{1}{2} U^{\pi}\left(\sigma \pi(s)\right)^{2} X^{\pi}(s)^{2} ds \end{array} \right) \qquad (6.53) \\
&\quad + f\left(T, X^{\pi}(T), g\left(X^{\pi}(T)\right), h\left(X^{\pi}(T)\right)\right).
\end{aligned}
$$

3. Third, we establish on the basis of (6.53) that

$$F(t,x) \geq \sup_{\pi} f\left(t,x,y^{\pi}(t,x),z^{\pi}(t,x)\right).$$

An Itô calculation on F gives that

$$\begin{aligned} F\left(t,X^{\pi}(t)\right) &= -\int_{t}^{T} dF\left(s,X^{\pi}(s)\right) + F\left(T,X^{\pi}(T)\right) \\ &= -\int_{t}^{T}\left(F_s ds + F_x dX^{\pi}(s) + \frac{1}{2}F_{xx}\left(\sigma\pi(s)\right)^2 X^{\pi}(s)^2 ds\right) \\ &\quad + F\left(T,X^{\pi}(T)\right). \end{aligned}$$

Inserting (6.6) which for an arbitrary strategy π means that

$$F_t \leq f_t - (r+\Lambda\sigma\pi)\, x\left(F_x - f_x\right) - \frac{1}{2}\left(\sigma\pi\right)^2 x^2\left(F_{xx} - U^{\pi}\right),$$

with $x = X^{\pi}(s)$, and inserting (6.7) and (6.2) we get that

$$\begin{aligned} F\left(t,X^{\pi}(t)\right) \geq\ & f\left(T,X^{\pi}(T),g\left(X^{\pi}(T)\right),h\left(X^{\pi}(T)\right)\right) \\ & - \int_{t}^{T}\left(\begin{array}{l} \left(\begin{array}{l} f_s - (r+\Lambda\sigma\pi(s))\, X^{\pi}(s)\left(F_x - f_x\right) \\ -\frac{1}{2}\left(\sigma\pi(s)\right)^2 X^{\pi}(s)^2\left(F_{xx} - U^{\pi}\right) \end{array}\right) ds \\ +F_x\left(r+\Lambda\sigma\pi(s)\right) X^{\pi}(s)\, ds \\ +F_x\sigma\pi(s)\, X^{\pi}(s)\, dB(s) \\ +\frac{1}{2}F_{xx}\left(\sigma\pi(s)\right)^2 X^{\pi}(s)^2 ds \end{array}\right). \end{aligned}$$

Abbreviation gives

$$\begin{aligned} F\left(t,X^{\pi}(t)\right) \geq -\int_{t}^{T}\left(\begin{array}{l} \left(\begin{array}{l} f_s + f_x\left(r+\pi(s)\Lambda\sigma\right) X^{\pi}(s) \\ +\frac{1}{2}\left(\sigma\pi(s)\right)^2 X^{\pi}(s)^2 U^{\pi} \end{array}\right) ds \\ +F_x\sigma\pi(s)\, X^{\pi}(s)\, dB(s) \end{array}\right) \\ + f\left(T,X^{\pi}(T),g\left(X^{\pi}(T)\right),h\left(X^{\pi}(T)\right)\right). \end{aligned}$$

Inserting (6.53) we get that

$$\begin{aligned} F\left(t,X^{\pi}(t)\right) \geq\ & f\left(t,X^{\pi}(t),y^{\pi}\left(t,X^{\pi}(t)\right),z^{\pi}\left(t,X^{\pi}(t)\right)\right) \\ & + \int_{t}^{T}\left(f_x + f_y Y_x^{\pi} + f_z Z_x^{\pi} - F_x\right)\sigma\pi(s)\, X^{\pi}(s)\, dB(s). \end{aligned} \tag{6.54}$$

Now, due to sufficient integrability, taking conditional expectation on both sides and thereafter supremum over π on both sides finally gives

$$F(t,x) \geq \sup_{\pi} f\left(t,x,y^{\pi}(t,x),z^{\pi}(t,x)\right). \tag{6.55}$$

Consider the specific strategy π^*, and let $U^{\pi^*} = U\left(f, Y^{\pi^*}, Z^{\pi^*}\right)$.

1. First, since $G(t,x) = Y^{\pi^*}(t,x)$ and $H(t,x) = Z^{\pi^*}(t,x)$ we have from (6.48) and (6.52) that

$$\begin{aligned} G(t,x) &= y^{\pi^*}(t,x), \\ H(t,x) &= z^{\pi^*}(t,x). \end{aligned}$$

2. Second, also for this specific strategy we have that

$$\begin{aligned} &f\left(t, X^{\pi^*}(t), y^{\pi^*}\left(t, X^{\pi^*}(t)\right), z^{\pi^*}\left(t, X^{\pi^*}(t)\right)\right) \\ &= -\int_t^T \left(\begin{array}{l} f_s ds + f_x\left(r + \Lambda\sigma\pi^*(s)\right) X^{\pi^*}(s)\, ds \\ + \left(f_x + f_y Y_x^{\pi^*} + f_z Z_x^{\pi^*}\right) \sigma\pi^*(s)\, X^{\pi^*}(s)\, dB(s) \\ + \frac{1}{2} U^{\pi^*}\left(\sigma\pi^*(s)\right)^2 X^{\pi^*}(s)^2\, ds \end{array} \right) \\ &\quad + f\left(T, X^{\pi^*}(T), g\left(X^{\pi^*}(T)\right), h\left(X^{\pi^*}(T)\right)\right). \end{aligned} \tag{6.56}$$

3. Third, we establish on the basis of (6.56) that

$$F(t,x) \geq \sup_{\pi} f\left(t,x,y^{\pi}(t,x),z^{\pi}(t,x)\right).$$

An Itô calculation on F gives that

$$\begin{aligned} F\left(t, X^{\pi^*}(t)\right) &= -\int_t^T dF\left(s, X^{\pi^*}(s)\right) + F\left(T, X^{\pi^*}(T)\right) \\ &= -\int_t^T \left(F_s ds + F_x dX^{\pi^*}(s) \right. \\ &\qquad \left. + \frac{1}{2} F_{xx}\left(\sigma\pi^*(s)\right)^2 X^{\pi^*}(s)^2\, ds \right) \\ &\quad + F\left(T, X^{\pi^*}(T)\right). \end{aligned}$$

Inserting (6.6) which for the strategy π^* means that

$$F_t = f_t - (r + \Lambda\sigma\pi^*)\, x\, (F_x - f_x) - \frac{1}{2}\,(\sigma\pi^*)^2\, x^2 \left(F_{xx} - U^{\pi^*}\right),$$

with $x = X^{\pi^*}(s)$, and inserting (6.7) and (6.2) with the strategy π^* we get that

$$\begin{aligned} F\left(t, X^{\pi^*}(t)\right) &= f\left(T, X^{\pi^*}(T), g\left(X^{\pi^*}(T)\right), h\left(X^{\pi^*}(T)\right)\right) \\ &\quad - \int_t^T \left(\begin{array}{l} \left(\begin{array}{l} f_s - (r + \Lambda\sigma\pi^*(s))\, X^{\pi^*}(s)\, (F_x - f_x) \\ -\frac{1}{2}\,(\sigma\pi^*(s))^2\, X^{\pi^*}(s)^2 \left(F_{xx} - U^{\pi^*}\right) \end{array} \right) ds \\ +F_x \left(\begin{array}{l} (r + \Lambda\sigma\pi^*(s))\, X^{\pi^*}(s)\, ds \\ +\sigma\pi^*(s)\, X^{\pi^*}(s)\, dB(s) \end{array} \right) \\ +\frac{1}{2} F_{xx}\,(\sigma\pi^*(s))^2\, X^{\pi^*}(s)^2\, ds \end{array} \right) \end{aligned}$$

Abbreviation gives

$$\begin{aligned} F\left(t, X^{\pi^*}(t)\right) &= f\left(T, X^{\pi^*}(T), g\left(X^{\pi^*}(T)\right), h\left(X^{\pi^*}(T)\right)\right) \\ &\quad - \int_t^T \left(\begin{array}{l} \left(\begin{array}{l} f_s + f_x\,(r + \Lambda\sigma\pi^*(s))\, X^{\pi^*}(s) \\ +\frac{1}{2}\,(\sigma\pi^*(s))^2\, X^{\pi^*}(s)^2\, U^{\pi^*} \end{array} \right) ds \\ +F_x \sigma\pi^*(s)\, X^{\pi^*}(s)\, dB(s) \end{array} \right). \end{aligned}$$

Inserting (6.56) we get that

$$\begin{aligned} F\left(t, X^{\pi^*}(t)\right) &= f\left(t, X^{\pi^*}(t), y^{\pi^*}\left(t, X^{\pi^*}(t)\right), z^{\pi^*}\left(t, X^{\pi^*}(t)\right)\right) \\ &\quad + \int_t^T \left(f_x + f_y Y_x^{\pi^*} + f_z Z_x^{\pi^*} - F_x\right) \sigma\pi^*(s)\, X^{\pi^*}(s)\, dB(s). \end{aligned}$$

Now, assuming sufficient integrability, taking conditional expectation on both sides finally gives

$$\begin{aligned} F(t,x) &= f\left(t, x, y^{\pi^*}(t,x), z^{\pi^*}(t,x)\right) \\ &\leq \sup_{\pi} f\left(t, x, y^{\pi}(t,x), z^{\pi}(t,x)\right). \end{aligned} \tag{6.57}$$

(6.55) together with (6.57) gives that

$$F(t,x) = \sup_{\pi} f\left(t,x,y^{\pi}(t,x), z^{\pi}(t,x)\right).$$

From the arguments above we learn that this supremum is obtained by the strategy π^*.

□

Bibliography

Akerlof, G. A. and Shiller, R. J. (2009). *Animal Spirits*. Princeton University Press.

Andreev, K. F. (2002). *Evolution of the Danish Population from 1835 to 2000*. Monographs on Population Aging, 9. Odense University Press.

Ball, L. and Mankiw, N. G. (2007). Intergenerational Risk Sharing in the Spirit of Arrow, Debreu, and Rawls, with Applications to Social Security Design. *Journal of Political Economy 115 (4)*, 523–547.

Ballotta, L. (2005). A Lévy process-based framework for the fair valuation of participating life insurance contracts. *Insurance: Mathematics and Economics 37*, 173–196.

Barbi, E. (2003). Assessing the rate of ageing of the human population. Max Planck Institute for Demographic Research Working Paper 2003-008.

Basak, S. and Chabakauri, G. (2009a). Dynamic Hedging in Incomplete Markets: A Simple Solution. Available at `http://ssrn.com/abstract=1297182`.

Basak, S. and Chabakauri, G. (2009b). Dynamic Mean-Variance Asset Allocation. Available at `http://ssrn.com/abstract=965926`, forthcoming in Review of Financial Studies.

Baxendale, P. H. (2005). Renewal Theory and Computable Convergence Rates for Geometrically Ergodic Markov Chains. *Annals of Applied Probability 15 (1A)*, 700–738.

Bernard, C., Courtois, O. L., and Quittard-Pinon, F. (2005). Market value of life insurance contracts under stochastic interest rates and default risk. *Insurance: Mathematics and Economics 36*, 499–516.

Björk, T. (2009). *Arbitrage Theory in Continuous Time*. Oxford University Press, third edition.

Björk, T. and Murgoci, A. (2008). A General Theory of Markovian Time Inconsistent Stochastic Control Problems. Available at

`http://econtent.essec.fr/mediabanks/ESSEC-PDF/Enseignement%20et%20Recherche/Enseignement/Departement/seminaire/Finance/2008-2009/Tomas_Bjork-Seminaire.pdf`

Björk, T., Murgoci, A., and Zhou, X. (2009). Time Inconsistent Control and Mean Variance Portfolios with State Dependent Risk Aversion. Available at `www.math.kth.se/pdefinance/2009/presentations/Bjork.pdf`.

Black, F. and Perold, A. F. (1992). Theory of constant proportion portfolio insurance. *Journal of Economic Dynamics and Control 16*, 403–426.

Bongaarts, J. (2005). Long-range trends in adult mortality: Models and projection methods. *Demography 42*, 23–49.

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

Booth, H., Maindonald, J., and Smith, L. (2002). Applying Lee-Carter under conditions of varying mortality decline. *Population Studies 56*, 325–336.

Briys, E. and de Varenne, F. (1994). Life Insurance in a Contingent Claim Framework: Pricing and Regulatory Implicatons. *The Geneva Papers on Risk and Insurance Theory 19*, 53–72.

Briys, E. and de Varenne, F. (1997). On the risk of insurance liabilities: debunking some common pitfalls. *Journal of Risk and Insurance 64*, 673–695.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods, Second Edition*. Springer-Verlag.

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.

Cairns, A. J., Blake, D., and Dowd, K. (2005). Pricing death: Frameworks for the valuation and securitization of mortality risk. *ASTIN Bulletin 36*, 79–120.

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

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2007). A quantitative comparison of stochastic mortality models using data from England & Wales and the United States. Available at `www.mortalityrisk.org`.

Cairns, A. J. G. (2000). A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics 27*, 313–330.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2010). Bayesian Stochastic Mortality Modelling for Two Populations. Pensions Institute Discussion Paper PI-1001.

Chen, A. and Suchanecki, M. (2007). Default risk, bankruptcy procedures and the market value of life insurance liabilities. *Insurance: Mathematics and Economics 40*, 231–255.

Currie, I. D., Durban, M., and Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling 4*, 279–298.

Dahl, M., Melchior, M., and Møller, T. (2008). On systematic mortality risk and risk-minimization with survivor swaps. *Scandinavian Actuarial Journal 108*, 114–146.

Danish Welfare Commission (2004). *Fremtidens velfærd kommer ikke af sig selv*. Book in Danish available at www.velfaerd.dk.

de Jong, P. and Tickle, L. (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies 13*, 1–18.

Døskeland, T. M. and Nordahl, H. A. (2008a). Intergenerational Effects of Guaranteed Pension Contracts. *The Geneva Risk and Insurance Review 33*, 19–46.

Døskeland, T. M. and Nordahl, H. A. (2008b). Optimal pension insurance design. *Journal of Banking & Finance 32*, 382–392.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2008a). Backtesting Stochastic Mortality Models: An Ex-Post Evaluation of Multi-Period-Ahead Density Forecasts. Pensions Institute Discussion Paper PI-0803.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2008b). Evaluating the Goodness of Fit of Stochastic Mortality Models. Pensions Institute Discussion Paper PI-0802.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. II, second edition*. Wiley.

Gavrilov, L. A. and Gavrilova, N. S. (1991). *The Biology of Life Span: A Quantitative Approach, ed. V. P. Skulachev*. Harwood Academic Publishers.

Gollier, C. (2008). Intergenerational risk-sharing and risk-taking of a pension fund. *Journal of Public Economics 92*, 1463–1485.

Grosen, A. and Jørgensen, P. L. (2000). Fair valuation of life insurance liabilities: The impact of interest rate guarantees, surrender options, and bonus policies. *Insurance: Mathematics and Economics 26*, 37–57.

Grosen, A. and Jørgensen, P. L. (2002). Life Insurance Liabilities at Market Value: An Analysis of Insolvency Risk, Bonus Policy, and Regulatory

Intervention Rules in a Barrier Option Framework. *Journal of Risk & Insurance 69 (1)*, 63–91.

Haldrup, N. (2004). Estimation af middellevetider for mænd og kvinder i Danmark 2002-2100 baseret på Lee-Carter metoden. Available (in Danish) at Danish Welfare Commission: `www.velfaerd.dk`.

Hansen, M. and Miltersen, K. R. (2002). Minimum Rate of Return Guarantees: The Danish Case. *Scandinavian Actuarial Journal*, 280–318.

Hansen, M. F., Pedersen, L. H., and Stephensen, P. (2006). Forventet levetid for forskellige aldersgrupper. Available (in Danish) at DREAM: `www.dreammodel.dk`.

Hindi, A. and Huang, C. (1993). Optimal consumption and portfolio rules with durability and local substitution. *Econometrica 6 (1)*, 85–121.

Hougaard, P. (1986). Survival Models for Heterogeneous Populations Derived from Stable Distributions. *Biometrika 73 (2)*, 387–396.

Jarner, S. F. and Kryger, E. M. (2008). Modelling adult mortality in small populations: The SAINT model. Pensions Institute Discussion Paper PI-0902.

Jarner, S. F. and Kryger, E. M. (2009). Exact hitting time and ladder height distributions of random walks. Non-published manuscript available on request.

Jarner, S. F., Kryger, E. M., and Dengsøe, C. (2008). The evolution of death rates and life expectancy in Denmark. *Scandinavian Actuarial Journal 108*, 147–173.

Karatzas, I. and Shreve, S. E. (2000). *Brownian Motion and Stochastic Calculus, second edition*. Springer.

Koissi, M.-C., Shapiro, A. F., and Högnäs, G. (2006). Evaluating and extending the Lee-Carter model for mortality forecasting: Bootstrap confidence intervals. *Insurance: Mathematics and Economics 38*, 1–20.

Kryger, E. M. (2010a). Fairness vs. efficiency of pension schemes. *Forthcoming in European Actuarial Journal*.

Kryger, E. M. (2010b). Mortality Forecasting for Small Populations: The SAINT Framework. *AENorm 17*, 6–9.

Kryger, E. M. (2010c). Optimal Pension Fund Design Under Long-Term Fairness Constraints. *Forthcoming in Geneva Risk and Insurance Review*.

Kryger, E. M. and Steffensen, M. (2010). Some solvable portfolio problems with quadratic and collective objectives. Available at `http://ssrn.com/abstract=1577265`.

Lee, R. D. and Carter, L. R. (1992). Modeling and Forecasting of U.S. Mortality. *Journal of the American Statistical Association 87*, 659–675.

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

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography 42*, 575–594.

Life Expectancy Commission (1998). *Danskernes dødelighed i 1990'erne. 1. delrapport fra Middellevetidsudvalget*. Danish Ministry of Health.

Lin, Y. and Cox, S. H. (2005). Securitization of mortality risks in life annuities. *Journal of Risk and Insurance 72*, 227–252.

Merton, R. C. (1969). Lifetime Portfolio Selection Under Uncertainty: The Continuous-Time Case. *Review of Economics and Statistics 51*, 247–257.

Pollak, R. A. (1968). Consistent Planning. *The Review of Economic Studies 35 (2)*, 201–208.

Preisel, M., Jarner, S. F., and Eliasen, R. (2010). Investing for Retirement Through a With-Profits Pension Scheme: A Client's Perspective. *Scandinavian Actuarial Journal*, 15–35.

Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

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.

Rose, M. R. and Mueller, L. D. (2000). Ageing and immortality. *Philosophical Transactions of the Royal Society of London. Series B 355*, 1657–1662.

Solano, J. M. and Navas, J. (2010). Consumption and portfolio rules for time-inconsistent investors. *European Journal of Operational Research 201 (3)*, 860–872.

Sørensen, C. and Jensen, B. A. (2001). Paying for minimum interest rate guarantees : who should compensate who? *European financial management 7 (2)*, 183–211.

Steffensen, M. (2004). On Merton's Problem for Life Insurers. *ASTIN Bulletin 34 (1)*, 5–25.

Strotz, R. H. (1956). Myopia and Inconsistency in Dynamic Utility Maximization. *The Review of Economic Studies 23 (3)*, 165–180.

Thatcher, A. R. (1999). The Long-Term Pattern of Adult Mortality and the Highest Attained Age. *Journal of the Royal Statistical Society. Series A 162*, 5–43.

Tuljapurkar, S., Li, N., and Boe, C. (2000). A universal pattern of mortality decline in the G7 countries. *Nature 405*, 789–792.

Vaupel, J. W. (1999). Discussion on "The Long-Term Pattern of Adult Mortality and the Highest Attained Age" by A. R. Thatcher. *Journal of the Royal Statistical Society. Series A 162*, 31–32.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*, 439–454.

Weisstein, E. W. (2008). *Hypergeometric Function. From MathWorld – A Wolfram Web Resource.* Available at `http://mathworld.wolfram.com/HypergeometricFunction.html`.

Wilmoth, J. R. (1998). Is the pace of Japanese mortality decline converging toward international trends? *Population and Development Review 24*, 593–600.

Wilmoth, J. R., Andreev, K., Jdanov, D., and Glei, D. A. (2005). Methods protocol for the Human Mortality Database. Available at `www.mortality.org`.

Wilson, C. (2001). On the scale of global demographic convergence 1950-2000. *Population and Development Review 27*, 155–171.

Yashin, A. I. and Iachine, I. A. (1997). How frailty models can be used for evaluating longevity limits: Taking advantage of an interdisciplinary approach. *Demography 34*, 31–48.