Multi-Population Mortality Models and Scenario-Based Projections

Snorre Jallbjørn

Department of Mathematical Sciences
Faculty of Science
University of Copenhagen

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen.

Industrial PhD Thesis by:

Snorre Jallbjørn
Plantagevej 52
DK-2610 Rødovre
snorrejall@gmail.com

Assessment committee:

Associate Professor Bo Markussen (chairperson)
University of Copenhagen

Associate Professor Malene Kallestrup-Lamb
Aarhus University

Adjunct Professor Christian Bressen Pipper
LEO Pharma and University of Southern Denmark

Supervisors:

Affiliate Professor Søren Fiig Jarner
heukno, ATP and University of Copenhagen

Professor Niels Richard Hansen
University of Copenhagen

Professor Mogens Steffensen
University of Copenhagen

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on 27 October 2022. It has been written as part of an Industrial PhD project under File No. 9065-00135B funded jointly by Innovation Fund Denmark (IFD) and the Danish Labour Market Supplementary Pension Fund (ATP).

ISBN: *978-87-7125-062-6.*

Abstract

This thesis covers five topics related to mortality modelling and forecasting with two overarching themes: (coherent) multi-population models and scenario-based projections. In the first part of the thesis, we study desirable model properties from the practitioner's point of view. We present a framework in which multiplicative frailty can be used with stochastic mortality models, and we apply the methodology in a case study of the SAINT mortality model used by the Danish Labour Market Supplementary Pension Fund (ATP). Next, we show that cointegration-based mortality models have more to offer than assuring non-diverging forecasts, and we highlight the limitations of cointegration when applied to models only identifiable under certain identification constraints. Following this, we propose a novel approach to analyzing the global properties of the life expectancy sex gap implied by coherent models, using which we challenge the status of coherence as a universally desirable property. In the second part of the thesis, we formulate a generic causal mortality model in the framework of potential outcomes, which facilitates a discussion of interventions and their direct and indirect effects on forecasts. We show by example the assumptions and data needed to operationalize an empirical analysis. Finally, we introduce the basic ideas for how causal models can be estimated and used to answer interventional queries when data consists of density estimates on marginal distributions observed across multiple populations.

Resumé

Denne afhandling behandler fem emner relateret til dødelighedsmodellering og fremskrivning med to overordnede temaer: (kohærente) multipopulations dødelighedsmodeller og scenariebaserede fremskrivninger. I den første del af afhandlingen studerer vi attraktive modelegenskaber set ud fra et praktisk synspunkt. Vi præsenterer en ramme hvori multiplikativ skrøbelighed (frailty) kan anvendes sammen med stokastiske dødelighedsmodeller, og vi bruger metodikken i et casestudie af SAINT modellen som anvendes af Arbejdsmarkedets Tillægspension (ATP). Dernæst viser vi, at kointegrationsbaserede dødelighedsmodeller har mere at byde på end blot at garantere ikke-divergerende fremskrivninger, og vi belyser begrænsningerne ved en kointegrationsanalyse, når denne anvendes på modeller, som kun er identificerbare under visse identifikationsbetingelser. Derefter foreslår vi en ny tilgang til, hvorledes man kan analysere de globale egenskaber af kønsforskellen i levetid i kohærente modeller, gennem hvilken vi udfordrer status af kohærens som en universel attraktiv modelegenskab. I den anden del af afhandlingen formulerer vi en kausal dødelighedsmodel ved brug af 'potential outcomes' tilgangen, hvilket faciliterer en diskussion af interventioner og deres direkte og indirekte effekter på fremskrivningen. Vi viser gennem et eksempel, hvilke antagelser og data der kræves for at kunne operationalisere en empirisk analyse. Til slut introducerer vi en tilgang til, hvorledes kausale modeller kan estimeres og anvendes til at besvare interventionsspørgsmål, når data består af tæthedsestimater af marginale fordelinger observeret på tværs af populationer.

Preface

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

The work has been carried out under the supervision of Affiliate Professor Søren Fiig Jarner (heukno, ATP, University of Copenhagen), Professor Niels Richard Hansen (University of Copenhagen) and Professor Mogens Steffensen (University of Copenhagen) in the period October 2019 to October 2022 (including 4 weeks of paternity leave). The project was set up as an Industrial PhD project under the Industrial PhD Programme of Innovation Fund Denmark (grant number 9065-00135B) with ATP as the industrial partner.

The thesis consists of an introductory chapter and five manuscripts on different, but related, topics. The manuscripts are written as self-contained, scientific contributions and can be read independently. The reader may find minor notional discrepancies across the chapters, but it is unlikely to cause confusion.

Acknowledgments

First and foremost, I would like to thank Innovation Fund Denmark and ATP for funding the project.

I would like to extend my appreciation and thankfulness to my supervisors and co-authors Affiliate Professor Søren Fiig Jarner and Professor Niels Richard Hansen for three years of wise guidance and fruitful discussions. Søren, your dedicated involvement in the project has been invaluable. Your enthusiastic and positive attitude is contagious, and I have always found myself less stressed and more motivated after our frequent meetings. Thank you for the opportunities you have given me, and for your continued encouragement. I hope that we will have another chance to work together in the future. Niels, I am most grateful for you challenging me to grow as a researcher, and for always making time for me and my questions. You have taught me that writing down and forcing a clear definition of a problem is a simple yet effective way to clarify one's thoughts, and it is a lesson I will carry with me always. I also thank Mogens Steffensen for thoughtful supervision, especially

concerning project management, and for supporting me in both my professional and personal development.

I would like to thank my colleagues at the University of Copenhagen for making my time at the department enjoyable. A special thanks to Christian Furrer for providing me with the LaTeX–template used to create this thesis – it compiled on the first (!) run.

During my studies I had the pleasure of visiting the Department of Biostatistics at the University of Oslo. I would like to express my gratitude to Kjetil Røysland for hosting me on short notice, and to all the people at the department for a rewarding and pleasurable time.

I would also like to thank my colleagues at ATP for taking interest in my research and for supporting me throughout the duration of my studies. To Helene, thank you for encouraging me to take on this project. To Thomas, thank you for always buying me coffee and for always listening.

To all of my friends and family, thank you for your unwavering support and belief in me. Thank you to the strong people at VK Ares who made me forget my hardship, if only for a few moments a week. Thank you to Teis and Jannick for always having my back. Thank you to Sif for not disturbing my sleep (too much). And finally, to Rikke, thank you for your unconditional love and support.

Snorre Jallbjørn
Rødovre, October 2022

For the final version of this thesis, a few typographical errors have been corrected and an ISBN has been added.

Snorre Jallbjørn
Rødovre, January 2023

List of papers

This thesis is comprised of six chapters. Chapter 1 serves as an introductory chapter, providing context for and an overview of the main chapters. Apart from differences in formatting, the five remaining chapters consist of the following self-contained manuscripts:

Chapter 2: Jarner, S. F. and Jallbjørn, S. (2022). The SAINT Model: A Decade Later. *ASTIN Bulletin: The Journal of the IAA*, **52**(2), pp. 483–517. DOI: 10.1017/asb.2021.37

Chapter 3: Jarner, S. F. and Jallbjørn, S. (2020). Pitfalls and merits of cointegration-based mortality models. *Insurance: Mathematics and Economics*, **90**, pp. 80–93. DOI: 10.1016/j.insmatheco.2019.10.005

Chapter 4: Jallbjørn, S. and Jarner, S. F. (2022). Sex Differential Dynamics in Coherent Mortality Models. *Forecasting*, **4**(4), pp. 819–844. DOI: 10.3390/forecast4040045

Chapter 5: Jallbjørn, S., Jarner, S. F., and Hansen, N. R. (2022). Forecasting, Interventions and Selection: The Benefits of a Causal Mortality Model. *Submitted for publication.*

Chapter 6: Jallbjørn, S. and Hansen, N. R. (2022). Aggregated Structural Causal Models. *Working paper.*

Contents

Chapter 1

Introduction

The steady decline of mortality rates observed over the past decades has far-reaching implications for many national economies, putting severe pressure on the sustainability of pension systems and public finances as the share of the population aged 65 and above continues to rise. With many of the consequences of an ageing population still ahead of us, the societal importance of producing accurate and credible mortality forecasts is greater than ever.

This thesis covers five topics related to mortality modelling and forecasting. The main chapters can be divided into two categories. Chapters 2–4 focus primarily on the development and application of multi-population forecasting methods. Chapters 5 and 6 consider causal models and address the opportunities and obstacles of an explanatory approach to mortality forecasting using epidemiological information. The present chapter introduces some important aspects of stochastic mortality modelling and causal inference. The purpose is not to give a complete and exhaustive account on the history of either mortality modelling or causality, but rather to provide the necessary background and context in which this thesis should be read. Building on this introduction, Section 1.4 provides an overview of the main contributions of the thesis and their interconnections.

1.1 The Modern Rise of Life Expectancy

Life expectancy at birth has averaged between 10 and 40 years for most of human history. From the age of hunter-gatherers, through the rise and fall of the Roman empire and until quite recently, wars, food insecurity and infectious disease outbreaks prevented most people from surviving into old ages. However, the epoch of high, volatile death rates suddenly began to end two-hundred years ago. With general living conditions improving in the wake of the industrial revolution, and with the emergence of medical countermeasures for infectious diseases, life expectancy has soared throughout the Western world since the mid- to late-1800s. As a testament to

this fact, Figure 1.1 shows the dramatic increase in life expectancy at birth observed for the Nordic countries[1] over the past two centuries. One can also observe the reduction in volatility that followed the mass production and widespread availability of antibiotics – penicillin in particular – after the Second World War.

As in Jarner et al. (2008), we take a closer look at the age-specific death rates in Figure 1.2 that tell the story behind the large secular increases. Although reductions in mortality across the entire span of ages have affected life expectancy, the rapid gains made between 1800 and 1950 were due mainly to reductions in infant and child mortality. Throughout the 1800s, attrition was very high at the young ages. The probability of a newborn in the Nordics surviving to age 20 was just barely over 50% in 1800. With such a high probability of death, only a select subgroup of very robust individuals made it into adulthood. Indeed, life expectancy conditionally on surviving to age 20 was 57.2 years, a staggering difference of 25.4 years compared to life expectancy at birth. By 1950, the probability of surviving to age 20 had increased to over 95%, almost exhausting the prospects for further gains from the young age groups. Improvements in life expectancy have since been driven mostly by reductions in adult and old-age mortality. Similar analyses confirm this evolution for other developed countries, see e.g. Riley (2001).

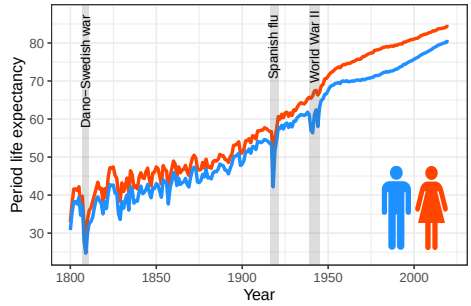

Figure 1.1: Evolution in period life expectancy at birth (i.e, the average number of years one expects to live if one experiences the age-specific death rates prevalent in a particular year) for the Nordic countries.

The story can be told also by breaking down deaths into causes by category. Historically, the share of deaths due to infectious diseases and malnutrition have dominated the rankings, while today the majority of deaths are caused by non-communicable diseases with deaths due to cancers and cardiovascular diseases being the leading ones.[2]

It bears mentioning that the events that have led to the remarkable increases in life expectancy historically are not repeatable – today, the probability of surviving childhood is around 99.5% in most developed countries and infectious diseases are almost eliminated through vaccination and antibiotics. Consequently, scholars have argued that the prospects for further improvements in mortality have diminished,

[1] For the empirical illustrations, we use data from the Human Mortality Database (2022). This dataset covers Denmark (from 1835), Finland (from 1878), Iceland (from 1838), Norway (from 1846) and Sweden (from 1751).

[2] See, for example, `https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death`.

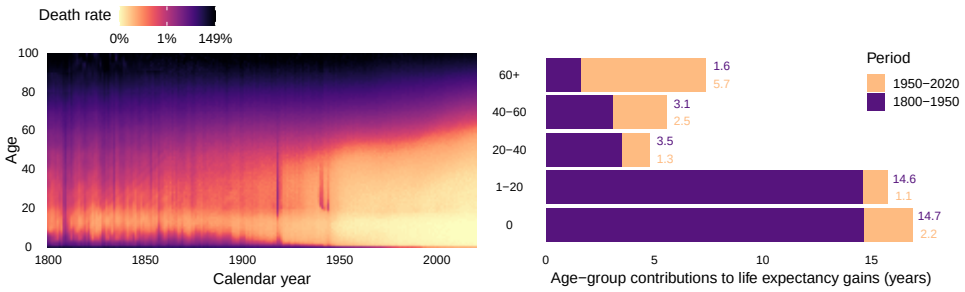

Figure 1.2: Mortality for the Nordic countries. The left panel contains a heat map of the death rates binned by percentiles. The right panel shows contributions to life expectancy at birth decomposed by age groups and periods following the method of Arriaga (1984).

and that a maximum expected lifespan may soon be reached. However, proposed limits have been broken time and time again, and best practice life expectancy[3] is still increasing linearly by roughly two and a half years[4] per decade (Oeppen and Vaupel, 2002; Vaupel et al., 2021). In this light, even though one can imagine many possible futures, both pessimistic and optimistic, it is not entirely unreasonable that improvements in life expectancy will continue to occur at the historical pace.

1.1.1 Societal Response to Increased Longevity

Alongside the continued reductions in old-age mortality over the past decades, financial liabilities associated with a longer lifespan have increased. It is important to recognize that even a slight bias in the mortality expectations can have serious consequences when scaled to, for example, the number of people expected to receive pension benefits:

"*If individuals live three years longer than expected – in line with underestimations in the past – the already large costs of ageing could increase by another 50 percent [of GDP in advanced economies].*" — IMF (2012)

Mitigating the impact of demographic changes is therefore an increasing source of concern and has called for both political and regulatory action.

The pension systems in most developed countries rest on a universal, public pension pillar designed to deal explicitly with poverty alleviation by providing everyone with a minimum level of income. Often, the public pension plans are not pre-funded, and the pension benefits for current retirees are paid by current workers through taxation.

[3]The best practice life expectancy is usually defined as the empirical record average length of life in a national population in a particular year.

[4]With the Nordics experiencing a life expectancy gain of 12.8 years in the period 1950–2020, cf. Figure 1.2, improvements have been lower than the best-practice increase of 17.5 years ($= 7 \cdot 2.5$ years). This slowdown is due to periods of stagnation in mostly Danish mortality caused by unhealthy lifestyles, in particular the consumption of alcohol and tobacco (Juel, 2008; Lindahl-Jacobsen et al., 2016; Kallestrup-Lamb et al., 2020).

While there are many advantages to systems funded on such a pay-as-you-go basis, they are particularly vulnerable to demographic changes. As life expectancy increases, so does public expenditure on pension, threatening to undermine fiscal sustainability.

To improve the long-term sustainability of its pension system, Denmark was the first country to introduce an automatic adjustment mechanism for its state pension age in 2006, creating a one-to-one link between life expectancy and the statuatory retirement age. Since then, six other OECD countries have followed suit with similar links, while about two-thirds of the OECD countries currently employ some kind of automatic adjustment for mandatory components of their pension systems (OECD, 2021). Apart from protecting the financial sustainability of a pension system, an adjustment mechanism also enforces a sense of intergenerational fairness in that the financial costs of living longer are shared between generations.

In the private pension industry, schemes are funded with "real" money in individual or collective accounts. Still, life insurance companies and pension funds face major challenges in relation to increasing lifespans. This risk is often categorized between micro and macro longevity risk (Hári et al., 2008). The former is the risk that a particular individual lives longer than expected, an unsystematic risk that can be diversified away by the company having a sufficiently large portfolio. In contrast, the latter is the risk that all insured on average live longer than expected, a systematic risk that cannot be diversified away by pooling.

Historically, macro longevity risk has been managed by determining insurance premiums on a conservative first-order technical basis; a set of assumptions concerning, for example, mortality, designed to be prudent so that portfolios generate a systematic surplus over time. For this purpose, insurance companies in Denmark used a common first-order basis, i.e., the L66, U74 and G82 risk tables. But while these tables were deemed prudent at the time they were made, the mortality assumptions were quickly overtaken by reality as higher-than-expected improvements continued to materialize. It soon became clear that a new way of managing longevity risk and setting "safe-side" assumptions was needed.

In Denmark, and in many other countries, the solution has been projection-based mortality tables.[5] Since 2011, Danish life insurance companies and pension funds have been required to take future, expected improvements in mortality into account when calculating their reserves. At the same time, the Danish Financial Supervisory Authority introduced a longevity benchmark to be used as the industry standard by all Danish life insurance companies. Companies are allowed to deviate from the benchmark only if their model provides a similar degree of prudence

[5]Also, many European insurers have moved away from guaranteed products and towards products with only conditional, or no guarantees, e.g., unit-link contracts, a shift that has been expedited by the low interest rate environment. For the insurance company, this enables a transfer of both financial and longevity risk to the policyholder.

(Finanstilsynet, 2010). Industry standard models, published and maintained by the Actuarial Profession, are also used in, for example, the UK, the Netherlands and Belgium (Continuous Mortality Investigation, 2016; Antonio et al., 2017).

Within the European Union, insurance regulation was updated and harmonised with the Solvency II regulatory regime, which came into force January 2016. For many insurance companies, Solvency II has become a catalyst of change, incentivizing portfolio-specific stochastic modelling for quantitative risk assessment. The regime dictates market-based valuation of liabilities alongside solvency capital requirements (SCR's) for various risks. The SCR for longevity risk is defined as the capital required to cover all losses due to variation in mortality rates that may occur over a one-year period with at least 99.5% probability. Performing this computation requires a probabilistic model of mortality. Alternatively, the SCR may be calculated using the "Standard Formula", which is meant to reflect a conservative estimate of the above, by evaluating the effects on the liabilities of a permanent, uniform reduction of the mortality rates by 20%. While straightforward to apply, the sheer magnitude of this stress leads to an excessively high solvency requirement for many companies – especially those that update their mortality assumptions annually – pushing them to develop their own internal stochastic models to assess their risks more accurately (Börger, 2010; Jarner and Møller, 2015).

Consequently, stochastic mortality models have come to play a prominent role in both actuarial applications and in policy-making processes for planning appropriate responses to the consequences of population ageing.

1.2 Stochastic Mortality Modelling

Models of human mortality have a long history, dating back to Moivre (1725). The first reasonably accurate model is due to Gompertz (1825), who made the simple observation that mortality tends to increase exponentially as a person ages. Later, Makeham (1867) proposed to add an age-independent component to the equation, the rationale being that not all deaths are due to senescence. The modification resulted in the famous Gompertz-Makeham law of mortality, which has proven surprisingly effective as a model for the age-specific death rates of adults. Although the classical parametric laws can be extended to a dynamic setting, they do not inherently allow for death rates to evolve through time, and have mostly been used as a parsimonious way of summarizing mortality profiles within shorter periods (Booth and Tickle, 2008; Pitacco et al., 2009).

The model of Lee and Carter (1992) marks the beginning of the modern era of mortality modelling. Projection methods prior to the 1990s were primarily focused on point forecasts, and to a large extent based on subjective judgements. In fact, the common method of projection was to simply ask a group of experts, for example

doctors and epidemiologists, to essentially guess the future level of mortality, and then average out over these expectations (Olshansky et al., 2009).[6] The Lee-Carter model signified a move away from expert opinion towards objective, statistical methods, enabling also a quantification of forecast uncertainty.

The overall ambition in stochastic mortality modelling is to specify an accurate, predictive model for the population-level death rate (or transformations thereof). In a continuous-time setting, the death rate is defined for age $x \in \mathbb{R}_+$ and time $t \in \mathbb{R}$ as

$$\mu(t,x) := \lim_{\Delta \to 0^+} \mathsf{P}(x \leq Y_{t-x} < x + \Delta \mid Y_{t-x} \geq x)/\Delta, \tag{1.2.1}$$

where Y_{t-x} is a non-negative random variable describing the lifetime of an individual from the cohort born at time $t - x$. The lifetime, Y, can depend also on, e.g., sex or socio-economic group, but for ease of notation we suppress this dependence.

A mortality model qualifies as being *stochastic* if it can produce a probability distribution for the forecast mortality levels as opposed to a deterministic forecast only. Even though a wide range of behavioural risk factors influence the probability of death, traditional models do not consider population heterogeneity apart from that described by chronological age, calendar time and sex. The rationale is to focus on capturing the secular downward trend in mortality, without having to disentangle the complex relationship between behavioural risks, diseases, genetics and death. Thus, most stochastic mortality models take the form

$$\mu(t,x) = G_\theta(t, x, \{\varepsilon_{s,x}\}_{s \leq t}), \tag{1.2.2}$$

where G_θ is the model parameterized in terms of θ, and the ε's are random variables that capture the stochastic nature of μ. For most applications, the data used to estimate θ are aggregate summaries from vital statistics bureaus on total death counts $\{D(t,x)\}_{t \in \mathcal{T}, x \in \mathcal{X}}$ and total exposure-to-risk-of-death(-estimates) $\{E(t,x)\}_{t \in \mathcal{T}, x \in \mathcal{X}}$, that is, the number of person-years lived at age x during period t, for a given set of ages $\mathcal{X}$ and periods $\mathcal{T}$.

1.2.1 The Lexis Diagram and the Poisson Assumption

A common tool used for representing mortality data is the Lexis diagram (Lexis, 1875; Keiding, 1990). The Lexis diagram is a period-age coordinate system with (calendar) time as abscissa and age as ordinate, see Figure 1.3 for an illustration. Lifetimes of individuals are displayed as line segments of unity slope and deaths as points in the diagram. The two time scales are typically divided into disjoint regions by a partitioning of the coordinate plane into square segments. Let us consider a

[6] For some time, this method was used to generate projections by Statistics Denmark. Following an expert prediction of life expectancy in some target year, a linear interpolation between current life expectancy and target life expectancy was made. To obtain age-specific death rates, a baseline mortality table was scaled to match the life expectancy estimate.

tessellation of the Lexis plane into squares of the form $\Omega_{ij} := [t_{i-1}, t_i) \times [x_{j-1}, x_j)$. The usual unit is one-by-one cells, but other partitions may be used. For modelling purposes, it is common to adopt an assumption of cellwise constant mortality over these period-age cells, that is,

$$\mu(t, x) = \mu_{ij}, \quad (t, x) \in \Omega_{ij}. \tag{1.2.3}$$

If individual life times are observable, the total exposure E_{ij} in Ω_{ij} can be calculated exactly by adding up the length of all life lines intersecting Ω_{ij} (and dividing by $\sqrt{2}$ to convert to person-years). If we denote by D_{ij} the number of deaths that occurred in Ω_{ij}, the observed or central death rate is defined as the number of occurrences divided by the exposures $m_{ij} := D_{ij}/E_{ij}$. This quantity may also be derived as the maximimum likelihood estimate of μ_{ij} under the assumption of cellwise constancy. Suppose we have L i.i.d. observations on $Y_{ij}^{(l)}$, the time individual l lived in Ω_{ij}, and let $\delta_{ij}^{(l)}$ be the indicator of whether individual l died in Ω_{ij} or not. The log-likelihood for μ_{ij} can be written as

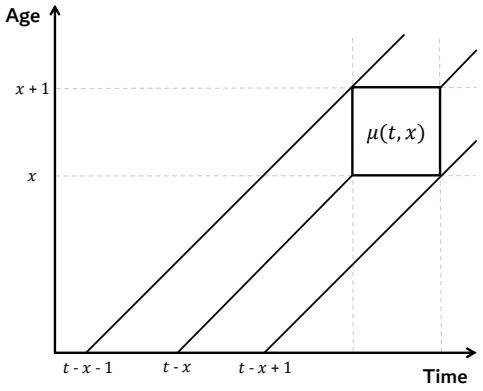

Figure 1.3: Illustration of the Lexis diagram and cellwise constancy.

$$\log \mathcal{L}_{ij} = \log \prod_{l=1}^{L} \mu_{ij}^{\delta_{ij}^{(l)}} e^{-\mu_{ij} Y_{ij}^{(l)}} = D_{ij} \log \mu_{ij} - \mu_{ij} E_{ij}. \tag{1.2.4}$$

Solving the score equation $\frac{\mathrm{d}}{\mathrm{d}\mu_{ij}} \log \mathcal{L}_{ij} = 0$ for μ_{ij} yields the occurrence-exposure rate as desired.

One may notice that inference regarding μ can be drawn from a Poisson likelihood; $\mathcal{L}_{ij}$ is proportional to the likelihood obtained by treating $D_{ij}|E_{ij}$ as Poisson random variable with rate parameter $\mu_{ij}E_{ij}$. This shows that the aggregate quantities available in standard life table data are sufficient statistics for the model. Inference for a given specification of (1.2.2) is therefore often based on the assumption of Poisson distributed deaths, see for example Brouhns et al. (2002) and Currie (2016).

A Short Digression on Period and Cohort Quantities

Life table related quantities such as survival functions and life expectancies are derived from the time- and age-specific mortality rates. However, the interpretation of these quantities is a common cause for confusion, rooted in the distinction between the period and the cohort perspective. Quantities defined in period terms

represent only the mortality experience in a given period. Portrayed on a Lexis diagram, this means that only μ's along the vertical axis enter the computation, and, importantly, that future (expected) mortality improvements are not taken into account. For instance, the survival function $\exp(-\int_0^x \mu(t,y)\,\mathrm{d}y)$ describes the probability of surviving to age x, based on the assumption that one is subject to the mortality rates experienced in period t throughout their entire life. In contrast, the cohort analogue $\exp(-\int_0^x \mu(t-x+y,y)\,\mathrm{d}y)$ uses μ's along a diagonal of the Lexis diagram, and describes the proportion of the cohort born at time $t-x$ that is still alive at time t.

The rationale for using period quantities is that they can be computed given the past experience only, whereas quantities defined in cohort terms either require a very long follow-up or a model for the future, expected rates of death. Since the choice of perspective is often implicit from context, life expectancies in particular are often misinterpreted by the uninitiated.

1.2.2 Forecasting the Secular Trend in Mortality

Extrapolative methods are at the heart of modern mortality forecasting procedures. The basic idea is to exploit that death rates have been steadily declining in a somewhat predictable pattern for a long time. Broadly speaking, if the future is anything like the past, extending these trends ought to give a reasonable estimate of the improvements to come.

The landmark model of Lee and Carter (1992) captures the temporal signal in mortality data with a single time-varying component. Under the Poisson specification, the model takes the form

$$D(t,x)|E(t,x) \sim \mathrm{Pois}(\mu(t,x)E(t,x)), \quad \mu(t,x) = \exp\left(\alpha_x + \beta_x\kappa_t\right), \tag{1.2.5}$$

where the α's and β's are age-specific parameters governing respectively the level of mortality and the rate of improvement in response to the time-varying κ's. The model's fit to historical data is often very accurate because the dependence on age is non-parametric.

While the Lee-Carter specification (1.2.5) itself stands as an important contribution, the pivotal methodological development was the introduction of a secondary model for forecasting. After the parameters in (1.2.5) have been estimated, an appropriate time series model is fitted to the time-varying κ's. Typically, a random walk with drift is chosen

$$\kappa_t = \kappa_{t-1} + \xi + \sigma\omega_t, \quad \omega_t \sim \mathcal{N}(0,1), \tag{1.2.6}$$

although other, more complex, time dynamics could also be used. The (adjusted)

sample mean and variance are used as estimators

$$\hat{\xi} = \frac{1}{T-1}\sum_{t=2}^{T}\Delta\kappa_t = \frac{\kappa_T - \kappa_1}{T-1}, \quad \hat{\sigma}^2 = \frac{1}{T-2}\sum_{t=2}^{T}\left(\Delta\kappa_t - \hat{\xi}\right)^2, \tag{1.2.7}$$

given "observations" $\{\kappa_t\}_{t\in\{1,\ldots,T\}}$ with $\Delta\kappa_t = \kappa_t - \kappa_{t-1}$. The time series model is then used to produce a forecast of the κ's over a desired forecast horizon, and a surface of projected mortality rates is obtained by inserting the forecasted κ's along with the estimated α's and β's into (1.2.5). This idea readily extends to any number of time-varying factors, and, importantly, enables us to assess the probabilistic uncertainty around the point forecasts.

There are several sources of uncertainty one can take into account. Usually, the focus is on the systematic variability, which is the variability that is independent of sample size. For mortality models under the Poisson error structure, we have by the law of total variance that

$$\begin{aligned}\mathsf{Var}[m(t,x)] &= \mathsf{Var}[\mathsf{E}[m(t,x)|E(t,x)]] + \mathsf{E}[\mathsf{Var}[m(t,x)|E(t,x)]] \\ &= \mathsf{Var}[\mu(t,x)] + \mathsf{E}[\mu(t,x)]/E(t,x),\end{aligned} \tag{1.2.8}$$

where the first term on the right-hand-side explains the systematic variability, while the second term explains the unsystematic variability, which tends to zero as the sample size increases.

The systematic variability includes uncertainty from model selection, sampling errors in the parameters and structural uncertainty from the innovation noise from the time series model (Cairns, 2000). Ignoring all sources of uncertainty apart from that of the innovation noise, the uncertainty for the h-step-ahead value of the time-varying parameters translates directly to the uncertainty for μ. To include uncertainty from all of the estimated parameters, bootstrapping techniques can be used (Brouhns et al., 2002, 2005; Koissi et al., 2006). Uncertainty from modelling choices can be assessed through sensitivity analyses or by Bayesian methods.

1.2.3 The Role of Statistics in Mortality Modelling

Advocates of the extrapolative approach to mortality forecasting will emphasize that the models are data-driven and widely based on objective, statistical methods. There are, however, still many subjective choices to be balanced, some of which may turn out to be decisive for the projections. In particular, comparisons between different methods have shown that due to changing patterns of improvements, the choice of data period used to calibrate the model is often more important than the choice of model itself. From this perspective, there is certainly scope for debate about whether the problem of predicting future mortality rates should be treated as any other problem of statistical learning.

Mortality models are special in that they need to capture both past and future trends. When predictive performance is the leading priority, selecting the "best" model is often based on measures of forecasting accuracy. But does it actually make sense to perform model selection through purely statistical means?

From a time series perspective, the estimation window is generally rather short as mortality data observed prior to the 1950s is seldom used.[7] To assess the quality of fit, a hold-out method is typically employed, which leaves around 70 data points for estimation and validation, at best. Accuracy is measured in terms of mean squared error (or some variant thereof), with predictions from the model contrasted against the latest available data for a fixed period between 5 to upwards of 20 years. However, in our view, it is debatable whether (narrowly) besting other models in terms of forecast accuracy on such short- to medium horizons is a strong indicator of a model's ability to predict future trends.

Suppose for instance that the hold-out sample consists of data from the period 2010–2020, during which there have been multiple severe flu seasons and the outbreak of coronavirus disease. The mortality excess caused by these events should arguably not be defining for our view on the long-term trend, but based on this sample, we would lean towards models with fairly modest rates of improvement.

For long-term forecasting, accuracy is to a large degree determined by assumptions about future improvements at the high ages – assumptions that cannot be verified against empirical data. As such, it seems more natural to draw up a list of desiderata against which models can be assessed. A comprehensive list would of course elicit some degree of subjectivity, depending on which features and stylized facts a domain expert deems important. A relatively objective list of basic criteria can be found in Cairns et al. (2008). In addition to this list, one could add some features for the forecasts, for example that improvement rates should vary over time or that mortality for certain populations, e.g. females and males, should evolve in parallel. Expressing these features mathematically in a way that they arise endogenously in the model is challenging, a problem we return to in Chapters 2–4.

Also, one might question whether the role of a mortality model is simply to generate forecasts or if it should also provide insights into the dynamics of mortality, cf. Chapter 3. As we stress in Chapter 2, there are many desirable qualities aside from accuracy that a versatile mortality model should possess. One of those is

[7] As demonstrated in Section 1.1, mortality data can exhibit a number of structural breaks. While it seems reasonable that trends in the (near) future closely resemble trends from the recent past, future trends may be quite different from trends from more distant pasts. For example, a forecast based on improvement patterns observed prior to the 1950s leads to a serious underestimation of actual improvements in old-age mortality. Literature on finding an "optimal" data window agrees with this view, concluding that the window should be chosen as wide as possible, so long as it does not contain periods with radically different improvement patterns, see e.g. Booth et al. (2002). Because mortality models are typically not designed to deal with structural breaks, data observed prior to the 1950s is seldom used.

explainability. From a forecasting perspective, it is important that the model is "anchored in reality" with parameters that can be ascribed some degree of real world meaning – at least if we want to justify extending their trends into the future. From a practical point of view, being able to explain and justify the projected trends is important when political or commercial decisions are based on the model's output.

Nevertheless, mortality models of the form (1.2.2) are not equipped to provide any insights as to what causes the drift in the time-varying parameters. If we want to obtain an improved understanding of the underlying factors that drive the secular trend, epidemiological information has to be included in the model.

1.3 Explanatory Models and Scenario-Based Projections

It is intuitively obvious to even the most casual observer that mortality is the result of various biological processes that are, among other things, influenced by our choice of lifestyle. Asking laypeople, they would probably even expect that mortality projections are made conditionally on the prevalence of various risk factors such as smoking and obesity. In practice, such explanatory models (for forecasting) are still underdeveloped, although they represent a natural next step for obtaining more solidly founded projections.

A model that "explains" mortality requires a functional link between observable risk factors and cause-specific mortality. This comes with several challenges such as unreliability in cause-of-death reporting, a lack of methods to forecast risk behaviour and difficulties with discerning the cause-effect relations between risk factors and disease outcomes. However, with the advent of better data and initiatives such as the Global Burden of Disease Study that make epidemiological information easily accessible, building explanatory models is becoming a realistic objective.

Integrating cause-effect relationships into mortality models is not necessarily done for the purpose of improved best-estimate predictions, but rather to improve our understanding of how mortality develops in the future and the mechanisms responsible. This is achieved by studying the scenarios that arise from eliminating certain causes of death or varying the risk factors we condition on. However, if such interventions are to reflect real-world implementations, the model must be enhanced with a causal interpretation.

1.3.1 The Difference between Probabilistic and Causal Models

A causal model is essentially a probabilistic model equipped with the capability of describing a system when subject to external manipulation. For a given parametrization θ, a probabilistic model $\theta \mapsto \mathbb{P}_\theta$ specifies a single distribution over a system of random variables, whereas a causal model $\theta \mapsto \{\mathbb{P}_\theta^{\mathrm{do}(i)} : i \in \mathbb{I}\}$ specifies an entire family of distributions, one for each intervention. Here, $\mathbb{I}$ denotes the set of possible

interventions, including the observational setting. We use the do-operator of Pearl, e.g. Pearl (2009), to emphasize that $\mathrm{do}(i)$ is something that we actively *do* to the system and not something we passively observe.

The distinction between seeing (i.e., conditioning by observation) and doing (i.e., conditioning by intervention) is cardinal. For example, seeing that someone currently lives in a hospice expressly indicates that death is approaching that person. However, actively relocating someone (healthy) to live in a hospice does not increase that person's risk of death. In this case, the observed association between hospice and death can be explained through a common cause, namely a terminal illness, but this relation does not translate into a causal link between being at a hospice and imminent death.

While this example appears somewhat contrived, the conclusion carries over to more elaborate applications, where the distinction may be less intuitively clear. Suppose we wish to quantify the (causal) effect of smoking, X, on lung cancer, Y. Both variables are affected by place of residence, Z, as people living in poor regions tend to smoke more and are also more exposed to radon, another risk factor for lung cancer. We say that the causal effect from X to Y is confounded by Z, as Z affects both exposure and outcome.

For the sake of exposition, we assume that X, Y, Z are discrete random variables, and are the only ones relevant for determining the effect from X on Y. We can write

$$p(x, y, z) = p(y|x, z)p(x|z)p(z), \tag{1.3.1}$$

where, e.g., $p(x|z)$ is used as short-hand for $\mathsf{P}(X = x|Z = z)$. While the same formulation is possible using the definition of conditional probability, the above relations are formed by recursively considering each variable conditionally on its direct causes, and should be interpreted as autonomous generative mechanisms. That is, we can think of the system as being generated by a program that first draws a z from $\mathbb{P}_Z$, then an x from from $\mathbb{P}_{X|Z=z}$ and finally a y from $\mathbb{P}_{Y|X=x,Z=z}$ (and precisely in this order). The causal interpretation allows us to reason about interventions, since we can replace one or several mechanisms without affecting the remaining ones. Thus, intervening in one part of the system will bring about the "right" consequences "downstream".

In particular, conditioning on $\mathrm{do}(X = x^*)$ corresponds to an atomic intervention that replaces the distribution of X given its cause(s) with a one-point measure. The distinction between conditioning on $\{X = x^*\}$ and conditioning on $\mathrm{do}(X = x^*)$ can now be seen by contrasting the two formulas

$$p(x, y, z|\,\mathrm{do}(x^*)) = \left.\frac{p(y|x, z)p(x|z)p(z)}{p(x|z)}\right|_{x=x^*}, \tag{1.3.2}$$

$$p(x, y, z|x^*) = \left.\frac{p(y|x, z)p(x|z)p(z)}{p(x)}\right|_{x=x^*}. \tag{1.3.3}$$

To infer $p(y|\,\mathrm{do}(x^*))$, the causal effect of smoking on lung cancer, we could perform the experiment $\mathrm{do}(X = x^*)$ in a randomized controlled trial. Randomly assigning people to different values of X breaks its dependence on Z, and renders the effect of interest identifiable. In general, however, randomized controlled trials are expensive and not always feasible. Here, for instance, assigning people to start smoking would not be possible for ethical reasons. A central aspect of causal inference is therefore to state assumptions under which causal quantities become identifiable from observational quantities. In models without unobserved confounding, we can use adjustment formulas, see e.g. Spirtes et al. (2000), Pearl (2009), and Hernán and Robins (2020), but more advanced techniques are needed when the confounders are hidden.

Different frameworks exist for specifying causal models. These include causal graphical models (Spirtes et al., 2000), structural causal models (Pearl, 2009; Peters et al., 2017) and the potential outcomes framework (Imbens and Rubin, 2015; Hernán and Robins, 2020). Depending on context, one framework might be preferable over another in terms of the ease with which assumptions and relationships can be conveyed. In epidemiological and biomedical applications for instance, the potential outcomes frameworks is the most prevalent one. This is the framework we adopt in Chapter 5, while we work in the framework of structural causal models for the purposes of Chapter 6. The frameworks will be introduced when they are needed.

1.4 Overview and Contributions

This thesis has been written as an industrial PhD project in collaboration with the Danish Labour Market Supplementary Pension Fund (ATP), and is therefore thematically motivated by real problems faced by ATP in the context of rising life expectancies. Together with the Danish State Pension, ATP makes up the compulsory part of the Danish pension system. Because ATP provides essentially the entire Danish population with a whole life (nominally guaranteed) annuity, mortality assumptions are a critical component for how ATP is operated. In 2007, ATP developed the SAINT mortality model as part of a new market value annuity product, see Jarner and Kryger (2011). To some extent, Chapters 2–4 build on specific problems encountered during subsequent model revisions. Some of the solutions have been implemented in production at ATP, while others have provided valuable insights into the dynamics of mortality.

In Chapter 2 we survey the major changes made to the SAINT model since Jarner and Kryger (2011) in response to user feedback and regulatory requirements. This is followed by Chapters 3–4 that take a deep dive into some issues related to multi-population forecasting, with a focus on joint modelling of females and males. Generally, women live longer than men, and while this difference varies over time it is believed to persist. However, separate models for the two sexes lead to diverging

forecasts. Chapter 3 considers the use of cointegration techniques for constructing non-diverging mortality projections, while Chapter 4 has life expectancy differentials in multi-population mortality models as its focal point.

An overarching theme in Chapters 2–4 is the strive for a strong intuition and understanding of various model components and how they affect the forecasts produced by the model. Still, it may be difficult for a layperson to appreciate the uncertainty inherent in such estimates, or for a board member to engage in a discussion about varying abstract (model) assumptions.

With the purpose of developing scenario-based mortality projections that can be communicated in a more straightforward, verbal manner, Chapters 5–6 consider an explanatory approach to forecasting. Specifically, Chapter 5 discusses the assumptions and data needed for a scenario-based analysis to be operationalized, while Chapter 6 addresses the issue that, in practice, samples from a joint distribution of risk factors covering the entire population of interest is rarely available. Chapter 6 differs from the remaining chapters in that it is written for a general audience with an interest in causal models.

1.4.1 Stochastic Frailty Models and the SAINT Projection Methodology

In Chapter 2 we go through the major evolutionary steps of the SAINT model since Jarner and Kryger (2011) in response to changing demands arising from practical use and user feedback, and present the SAINT model in its current form used by ATP.

The SAINT model features a frailty component as proposed by Vaupel et al. (1979), that is, a non-negative stochastic quantity Z that acts multiplicatively on the underlying baseline mortality rate, μ_0. Thus, at the level of individuals, mortality reads $\mu(t, x|Z) = Z\mu_0(t, x)$.

Multiplicative frailty is a mathematically tractable way of introducing population heterogeneity and thereby effects of selection. The purpose is twofold: to improve the fit of old-age mortality and to endogenously create changing rates of improvement. The basic idea is that complex (observed) population dynamics may be the result of much simpler (unobserved) individual-level dynamics. For example, frail people have a tendency to die sooner than less frail people and this selection mechanism is one possible explanation for the well-documented old-age mortality plateau (Perks, 1932; Beard, 1959; Vaupel et al., 1979; Kannisto et al., 1994; Thatcher et al., 1998; Thatcher, 1999). In the same way, and perhaps more importantly in a forecasting aspect, lack of historic improvements in mortality among the oldest-old could be attributed to selection (Jarner and Kryger, 2011). This interpretation suggests that improvements among the oldest-old will start to materialize once the frailty composition at these ages begin to change.

The primary theoretical development of the SAINT model since the original version is the generalization of the frailty framework to allow for stochastic rather than deterministic long-term trends. We define a stochastic frailty model as a model for the population-level death rate of the form

$$D(t,x)|E(t,x) \sim \text{Pois}(\mu(t,x)E(t,x)), \quad \mu(t,x) = \mathsf{E}[Z|t,x]\mu_0(t,x),$$

where $\mathsf{E}[Z|t,x]$ is the mean frailty among survivors of the cohort born at time $t-x$. Since mean frailty is given endogenously within the model as a function of μ_0, it is either deterministic or stochastic depending on the nature of μ_0. If μ_0 is deterministic, the model can be estimated and forecasted using standard maximum likelihood and extrapolation techniques. If, on the other hand, μ_0 is stochastic then explicit expressions for $\mathsf{E}[Z|t,x]$ are no longer available. To overcome this issue, we devise a novel method for estimation and forecasting based on a Poisson pseudo-likelihood. The key idea is to replace $\mathsf{E}[Z|t,x]$ with a term that does not depend on the parameters of the baseline mortality model. This leads to a generally applicable procedure by which essentially any (stochastic) mortality model for μ_0 can be combined with frailty. The framework is also extended to allow for frailty-independent mortality components.

Furthermore, we advocate the view that there are a range of other model properties aside from forecasting accuracy that are important from a practitioner's point of view, namely stability, flexibility, explainability and credibility. Based on these properties we motivate additional changes to the SAINT model. We propose to base the long-term mortality trend on a model of the Gompertzian type,

$$\mu_0(t,x;\theta) = \exp\left(\theta_t^1 + \theta_t^2 x + \theta_t^3 (x-75)\mathbb{1}_{\{x\geq 75\}}\right),$$

that, combined with Gamma distributed frailty and a Makeham component, provides an excellent, parsimonious fit of the entire adult age span. Changes made to the time dynamics for the trend are rooted in the findings of Chapter 3.

1.4.2 Cointegration-Based Mortality Models

In Chapter 3 we look at the pitfalls and merits of cointegration-based mortality models. Cointegration-based mortality models were first suggested by Carter and Lee (1992) in a follow-up paper on possible extensions for their original method. By modelling the time-varying mortality indices for multiple populations jointly as a cointegrated process, the indices exhibit a stationary relation that prevents them from diverging, which, in turn, prevents the mortality projections from diverging. The property of non-divergence is also known as coherence; a given forecast of two-population mortality is said to be coherent if the mortality ratios converge to a set of positive, finite age-specific constants (Li and Lee, 2005; Hyndman et al., 2013).

Today, cointegration is widely used as a tool for achieving coherence in multi-population models. This includes the SAINT model that forecasts female and male

time-varying parameters using an error correction model to achieve a common long-run stochastic trend. However, with Chapter 3 we demonstrate that cointegration has more to offer than assuring coherence – it is also a powerful inferential tool for establishing the nature of the long-run dynamics.

We study a p-dimensional vector autoregressive process $\{X_t\}$ with k lags written in error correction form:

$$\Delta X_t := X_t - X_{t-1} = \Pi X_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \Phi D_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma),$$

where Π is the $p \times p$ matrix of autoregression coefficients. We suppose for cointegration that $\operatorname{rank}(\Pi) = r < p$ such that $\Pi = \alpha\beta^\top$ for matrices $\alpha, \beta \in \mathbb{R}^{p \times r}$ of rank r. In the context of mortality modelling, we think of X_t as a stacked vector of time-varying mortality indices. Under some further technical conditions, the Granger representation theorem can be used to decompose X_t into a stochastic and a stationary part, namely

$$X_t = X_0 + \tau(t) + C \sum_{s=1}^{t} \varepsilon_s + Y_t,$$

where $C = I - \alpha(\beta^\top \alpha)^{-1}\beta^\top$, while $\tau(t)$ is a deterministic trend that depends on the initial values, $C \sum_{s=1}^{t} \varepsilon_s$ captures the stochastic trends and $\{Y_t\}$ is a stationary process. As demonstrated in Chapters 2–3, using the Granger representation is key for understanding the model's short- and long run dynamics, an otherwise surprisingly difficult task even in the two-dimensional case.

Cointegration theory offers a statistical framework for identifying and testing stationary relations. The typical procedure for determining the rank of Π is sequential likelihood-ratio testing (Johansen, 1995). However, in a mortality modelling context, the structure is often imposed rather than tested. Some papers do test for cointegration rank, but only as part of a model selection step. Formulating and testing hypotheses on parameters is not part of the analysis. Considering deterministic trends of the form $\Phi D_t = \xi_0 + \xi_1 t$, we look at the model classes and hypothesis tests relevant for mortality modelling.

Taking the Lee-Carter model as an example, we point out the limitations of a cointegration analysis when applied to factors that are not fully identifiable. Because the time-varying index in the Lee-Carter model (1.2.5) is only identified up to a linear transformation, we are able to demonstrate that virtually all hypotheses of interest are non-testable without allowing arbitrary identification constraints to influence the analysis. In contrast, when cointegration is applied to identifiable factors a complete analysis is possible, and we show by example the insights that can be obtained in this case.

Cointegrated models yield forecasts of mortality indices that are "marginally" similar to those obtained from applying separate random walk models but produce a

more plausible dependency structure. However, even though the resulting forecasts are well-behaved and empirically justified, they may not be coherent in the strict mathematical sense. This indicates that the definition of coherence is too restrictive – a theme we pursue further in Chapter 4.

1.4.3 Life Expectancy Differentials in Coherent Mortality Models

In Chapter 4 we study the behaviour of the life expectancy differential

$$\mathbb{R} \ni t \mapsto \int_0^\omega \left[\exp\left(- \int_0^y \mu_1(t,z)\,\mathrm{d}z \right) - \exp\left(- \int_0^y \mu_2(t,z)\,\mathrm{d}z \right) \right] \mathrm{d}y.$$

Here, the subscripts on μ identify the mortality experience in two distinct populations, while $\omega \in \mathbb{R}_+$ is some age of truncation.

In the context of two-sex mortality in the Western world, the life expectancy gap widened in favour of women throughout most of the 20th century, but has recently started to close. This development is often attributed to different timings of shifts in unhealthy behaviours. In particular, women adopting health-damaging habits, such as smoking and drinking, is often suggested as an explanation for the reversal of the trend that began during the 1980s. Interestingly, the same sex gap pattern emerged in the SAINT model. That is, in a model with no behavioural effects, and where temporal parameters develop almost linearly over time. This lead us to investigate whether a more fundamental, mathematical reason was the driver behind the dynamics of the gap.

We consider the life expectancy differential implied by two-sex coherent mortality models. Since the introduction of coherence by Li and Lee (2005), a multitude of coherent, multi-population models have been proposed. However, the theoretical properties of the resulting forecasts are rarely studied. On the other hand, theoretical results exist regarding the decomposition and interpretation of observed changes in life expectancy and sex differentials, in particular, those of Glei and Horiuchi (2007) and Cui et al. (2019). In Chapter 4 we review both branches of literature, and use the decomposition results to provide theoretical insights on coherent forecasts.

Technically, we prove that a sufficient condition for the sex gap to be unimodal in strongly coherent mortality models subject to uniform rates of improvement is that

$$\frac{\partial}{\partial x} \log\left(\frac{\mu_m(t, I_m^{-1}(t,x))}{\mu_f(t, I_f^{-1}(t,x))} \right) \leq 0,$$

for all t and x, where $I_g^{-1}(t,z)$ denotes the age, x, at which $z = I_g(t,x) := \int_0^x \mu_g(y,t)\,\mathrm{d}y$ and the $g \in \{f,m\}$ subscript describes whether a quantity relates to females or males. The condition holds for mortality schedules of the same shape, for example when they are log-linear. When the condition holds, we

are in a situation with fixed mortality ratios but where the life expectancy gap both widens and narrows in two distinct epochs. It follows that the life expectancy sex gap should not be used to draw conclusions about mortality "catch up"-effects.

We demonstrate that for Western European levels of male excess mortality (relative to female mortality), the sex gap in the model typically peaks when female life expectancy is between 30 to 50 years. Although only formally proven for a subclass of coherent models, this insight carries over to other more "realistic" models as well, and it explains why coherent two-sex models forecast closing sex gaps for almost all Western European countries and all jump-off years since the 1950s – despite the fact that the actual sex gap was widening until the 1980s. This inadequacy, combined with the findings from Chapter 3, brings us to question the status of coherence as a universally desirable property.

1.4.4 Towards Scenario-Based Projections and Causal Mortality Models

In Chapter 5 we discuss how mortality forecasts are affected by interventions and we show by example the assumptions and data needed for such an analysis to be operationalized. From a communications standpoint, a target audience may be easier to engage when scenarios are formulated in a verbal manner (i.e., as interventions) compared to when scenarios are presented as projections under varying rates of improvement or as quantiles of an excess lifetime distribution.

Generally, we are interested in interventions that target modifiable risk behaviour and interventions that reduce or eliminate certain causes of death. Such queries are inherently causal. To produce scenarios that conform with real-world implementations, we need a model that establishes a functional relationship between risk factors and cause-specific mortality. On its own, discovering these cause-effect relationships and estimating the corresponding dose-response curves is a comprehensive and challenging task. Therefore, realistically, a causal mortality model must be based on existing epidemiological evidence from the literature.

Still, even if we have access to the true causal risk-outcome relationships, there is an additional challenge unique to models that forecast population-level quantities. From a modelling perspective, the joint dynamics of risk prevalence and mortality can be divided into two sub-models describing:

1. The dynamics of the risk factors given survival;

2. The dynamics of survival given the risk factors (as a conditional hazards model).

The usual approach with explanatory mortality models is, however, to treat risk prevalence as an exogenous process, determined independently of mortality. That is,

we first project risk prevalence, and then plug these estimates into the conditional hazards model to obtain a forecast of mortality. This approach is viable for most predictive tasks, but because the dependence between risk prevalence and death only goes one way, feedback effects due to selection are not produced if the system is perturbed (i.e., under interventions).

To facilitate a discussion of interventions and their direct and indirect effects, we formulate a causal mortality model in the framework of potential outcomes. We use techniques from causal mediation theory to decompose the total effect of an intervention into a part directly attributable to the action and a part due to selection. Defining the potential death rate $\mu^{\mathsf{K},\pi}$ as the death rate in a world where causes K are operating and risk prevalence is π, we decompose the total causal effect of cause-elimination on a risk difference scale as

$$\underbrace{\mu^{\mathsf{K},\pi} - \mu^{\mathsf{K}^*,\pi^*}}_{\text{Total effect}} = \underbrace{\mu^{\mathsf{K},\pi} - \mu^{\mathsf{K}^*,\pi}}_{\text{Direct effect}} + \underbrace{\mu^{\mathsf{K}^*,\pi} - \mu^{\mathsf{K}^*,\pi^*}}_{\text{Indirect effect}},$$

where $\mathsf{K}^* \subsetneq \mathsf{K}$ is a reduced set of causes and π^* is the risk prevalence in a world where causes K^* are operating. The direct effect corresponds to removal of causes $\mathsf{K} \setminus \mathsf{K}^*$ while leaving all other death rates unaffected. The indirect effect is the effect that arises due to changes in the post-intervention risk composition over time. If risk prevalence is exogenous to the mortality model, the indirect effect is zero by construction.

We highlight the significance of the indirect effect in both a numerical example and in an application to U.S. mortality and risk data. An interesting direction for future research would be to scale up this application, and explore the effects when a large number of risk factors are included in the model. Increasing the number of risk factors will, of course, also increase the data demands. In particular, a joint distribution of all risk factors will probably not be available. This is an issue we study in Chapter 6.

1.4.5 Aggregated Structural Causal Models

In Chapter 6 we consider a system of variables $X = (X_v : v \in V)$. Rather than having access to a single dataset with i.i.d. samples from the joint distribution $\mathbb{P}$, we assume that we only have density estimates of marginal distributions observed across multiple populations, for example for different countries or over time. Our aim is to formulate a causal model at the level of the observed data, which can be used to answer interventional queries concerning population-level behaviour.

Reasoning about the population of individuals as a whole instead of concentrating on individuals as single entities is natural in a policy intervention context. For instance, we are typically not interested in the reduction of individual i's risk of dying from ischaemic heart disease upon smoking cessation, but rather the reduction

in population-level mortality (i.e., the distributional change) caused by lowering smoking prevalence for the population as a whole. When interventions are formulated as population-based strategies and the inferential target is a distribution, we do not need to recover the full underlying causal structure but only certain aspects of it.

At the level of individuals, we suppose that X is generated according to a structural causal model (SCM) with a hierarchical structure in the errors:

$$X_v := F_v\left(X_{\mathrm{pa}(v)}, \varepsilon_v, \eta_v\right), \quad v \in V.$$

The set $\mathrm{pa}(v) \subseteq V \setminus \{v\}$ contains the indices of the parents (i.e., direct causes) of X_v. The variables $\varepsilon = (\varepsilon_v : v \in V)$ and $\eta = (\eta_v : v \in V)$ are noise variables that have a joint distribution with mutually independent marginals describing, respectively, variation at the level of individuals and variation at the level of populations. The SCM induces interventional distributions by replacing the structural assignment for one or several variables. However, we do not have data at the level of the SCM that can be used to estimate the F_v–map.

Introduce for a subset $I \subseteq V$ the marginal distribution $\mathbb{P}_I(A) = \mathsf{P}(X_I \in A|\eta)$ of X_I conditionally on η. By marginalizing out the individual level variables in the SCM we derive a new structural model for the system, this time at the level of distributions (at which we observe data) with structural equations of the form:

$$\mathbb{P}_I := G_I\left(\mathbb{P}_{\mathrm{pa}(I)}, \eta_I\right), \quad I \in \mathcal{I}.$$

Identifying the aggregated structure may prove sufficient to answer causal questions of interest. That is, if we know the G_I–map we can compute the distribution of X_I under interventions that alter the distribution of its parents. We show that the structural representation of the system in the aggregated SCM is equivalent to the representation in the SCM, in the sense that both models produce the same interventional distribution regardless of the order in which we aggregate and intervene.

Since G_I remains invariant across populations, we can hope that our data samples are sufficiently heterogeneous to reveal it. We outline a regression–based estimation approach in a setting of discrete variables, and discuss how to approach the problem when we have parametric assumptions. Further, we present an algorithm for determining a directed acyclic graph that represents the distributions necessary for formulating the aggregated model as a structural one. We apply the methods in a numerical study of risk factor interventions.

Chapter 2

The SAINT Model: A Decade Later

This chapter contains the manuscript *Jarner and Jallbjørn (2022)*.

Abstract

While many of the prevalent stochastic mortality models provide adequate short- to medium-term forecasts, only few provide biologically plausible descriptions of mortality on longer horizons and are sufficiently stable to be of practical use in smaller populations. Among the very first to address the issue of modelling adult mortality in small populations was the SAINT model, which has been used for pricing, reserving and longevity risk management by the Danish Labour Market Supplementary Pension Fund (ATP) for more than a decade. The lessons learned have broadened our understanding of desirable model properties from the practitioner's point of view and have led to a revision of model components to address accuracy, stability, flexibility, explainability and credibility concerns. This paper serves as an update to the original version published 10 years ago and presents the SAINT model with its modifications and the rationale behind them. The main improvement is the generalization of frailty models from deterministic structures to a flexible class of stochastic models. We show by example how the SAINT framework is used for modelling mortality at ATP and make comparisons to the Lee-Carter model.

Keywords: *SAINT model, Frailty, EM algorithm, Lee-Carter model, Mortality modelling, Multi-population modelling.*

2.1 Introduction

Since the beginning of the 21st century, life annuity providers have faced an upsurge of pensioners to provide for and the need for reliable, long-term mortality projections is perhaps greater than ever. Indeed, the world-wide increases in life expectancy show no signs of slowing down, and populations where mortality rates are already low still experience rates of improvements of the same, or even higher, magnitude than historically. The situation accentuates the importance of powerful predictive models to handle the consequences of an ever older population.

From a practical point of view, there are two partly conflicting aims: (1) producing accurate forecasts and (2) producing forecasts stable under (annual) updates. Accurate forecasting has been the long-standing objective in actuarial and demographic literature and is, broadly speaking, the goal of the academic, while stability is a more recent requirement pertaining to the needs of the practitioner. When applied in practice, the prevailing market value accounting regime dictates that a mortality model should be updated annually to reflect the latest trends in the data. However, many mortality modelling paradigms are very sensitive to the data period used for calibration, and forecasts can therefore vary substantially from year to year. For an annuity provider, large fluctuations or systematic underperformance of a mortality model can lead to significant shifts in liabilities and capital requirements, resulting in huge costs for either the company or the risk collective. Moreover, throughout Europe, mortality models have become an integral part of policy-making as statutory retirement ages are directly linked to gains in life expectancy. For these decisions, stable short-, medium- and long-term forecasts are not just a requirement, but a necessity.

Stability requirements are particularly difficult to meet when forecasting concerns small populations, including, in fact, many countries. Because improvement patterns in these populations exhibit a great deal of variability, simple extrapolations of past trends tend to have poor predictive power over long horizons and projections are prone to dramatic changes following data updates. This holds true for many of the popular projection methodologies such as the model of Lee and Carter (1992). In view of these accuracy and stability concerns, Jarner and Kryger (2011) developed the SAINT model that has been used by the Danish, nationwide pension fund ATP, since 2008.

The SAINT model was designed with the purpose of producing stable, biologically plausible long-term projections for adult mortality. More precisely, projections with smooth, increasing age-profiles and gradually changing rates of improvement over time. With the main application of pricing and reserving for long-term pension liabilities in mind, capturing long-term trends reliably were deemed more important than, for example, short-term fit. However, more than a decade's worth of experience

with the SAINT model in use has broadened our understanding of model requirements from the practitioner's point of view. Even though the overarching structure of the SAINT model has not changed, its components have been revised over the years to address not only accuracy and stability concerns but also the model's flexibility, explainability and credibility. In this paper, we describe how the SAINT model has evolved since Jarner and Kryger (2011) in response to changing demands arising from practical use and user feedback.

In the years following the introduction of the first version of the SAINT model, quantification of longevity risk became a regulatory requirement. As the deterministic trend component used in Jarner and Kryger (2011) was not able to adequately assess this risk, a version of the SAINT model with a stochastic trend was developed. Eventually, this work led to a generally applicable class of stochastic frailty models, which we present in this paper. This methodology constitutes the main theoretical contribution of the paper.

The rest of the paper is organized as follows. Section 2.2 contains a survey of the evolution of the SAINT model over the last decade. In Section 2.3 we discuss how changing rates of improvements can be modelled using frailty. Section 2.4 formalizes the notion of a stochastic frailty model and develops estimation and forecasting procedures. This is followed by Section 2.5 presenting a comprehensive application of the SAINT model to international and Danish data. The findings are discussed in light of comparable results from the Lee-Carter model. Finally, Section 2.6 offers some concluding remarks.

2.2 The Evolution of the SAINT Model

ATP is a funded supplement to the Danish state pension, guaranteeing most of the population a whole-life annuity. In 2008, ATP introduced a new market value (whole-life) annuity. The main characteristic of this annuity is that contributions are converted to pension entitlements on a tariff based on prevailing market rates and an annually updated, best estimate, cohort-specific mortality forecast. Once acquired, pension entitlements are guaranteed for life. The structure gives a high degree of certainty for the members, but it leaves ATP with a substantial longevity risk.

The SAINT model was developed as part of the market value annuity, with the specific aim of producing accurate and stable forecasts in order to manage the longevity risk of ATP. Clearly, accuracy is desirable to avoid long-run deficits due to life expectancy increasing faster than expected (or, in the case of life expectancy increasing slower than expected, pension entitlements being too small). However, stability of forecasts is of equal importance. Each update of the model causes a change in the size of technical provisions, which in turn affects the risk capital allocated to cover longevity risk. Effectively, a stable mortality model is "cheaper"

than a volatile model, because the former frees up risk capital to be used more efficiently elsewhere, for example, to cover a higher market exposure.

Over the course of the last decade, the SAINT model has undergone a number of changes due to changing demands and feedback from its users. Below, we give a brief presentation of the original SAINT model, followed by a survey of the subsequent major changes and their rationale.

2.2.1 The Original SAINT Model

The original SAINT model, as described in Jarner and Kryger (2011), has two core model components:

- A reference population consisting of a large, pooled, international data set;
- A frailty model for modelling increasing rates of improvements over time.

The rationale for using a reference population, in addition to the target population, is that it is easier to extract a long-term trend from a large dataset, than a small dataset, since idiosyncratic features are typically more pronounced in the latter. The mortality of the target population is subsequently linked to the long-term trend. A similar idea, although differently implemented, was introduced by Li and Lee (2005) in their multi-population extension of the Lee-Carter model. Since Jarner and Kryger (2011), the concept of a reference population has appeared in a number of models and applications, for example Cairns et al. (2011b), Dowd et al. (2011), Börger et al. (2014), Villegas and Haberman (2014), Wan and Bertschi (2015), Hunt and Blake (2017), Villegas et al. (2017), Menzietti et al. (2019), Li and Liu (2019), and Li et al. (2019).

In the notation of the present paper, the SAINT model is of the form:

$$\mu_{\text{target}}(t,x) = \mu_{\text{ref}}(t,x)\exp(y_t^\top r_x), \tag{2.2.1}$$

$$\mu_{\text{ref}}(t,x) = \mathbb{E}[Z|t,x]\mu_0(t,x) + \mu_b(t), \tag{2.2.2}$$

where $\mu_{\text{target}}(t,x)$ and $\mu_{\text{ref}}(t,x)$ are the force of mortality at age x and time t of the target and reference population, respectively. The target mortality is linked to the reference mortality by a set of age-dependent regressors, r_x, with time-dependent coefficients, y_t, termed *spread* parameters. Further, the reference mortality is modelled by the sum of a multiplicative frailty model with baseline mortality μ_0, and age-independent background mortality, μ_b. The term $\mathbb{E}[Z|t,x]$ denotes the (conditional) mean frailty of the given cohort, and we will discuss this in detail later. All mortality rates are gender-specific, although this is not shown explicitly in the notation.

Formally, the SAINT model in use today is still of form (2.2.1)–(2.2.2), but with a different interpretation of the frailty term, and different specifications of time-series dynamics, regressors, and baseline and background mortality. From a methodological point of view, the new frailty model represents by far the greatest of these changes.

2.2.2 From Deterministic to Stochastic Frailty

The SAINT model uses frailty to forecast increasing rates of improvement in age-specific mortality, thereby reducing the risk of underestimating future life expectancy gains. Loosely speaking, changes in baseline mortality affect selection and thereby mean frailty, $\mathbb{E}[Z|t,x]$, which in turn modifies the way baseline mortality affect population-level mortality.

In the original SAINT model, μ_0 and μ_b were assumed to be of a form equivalent to

$$\mu_0(t,x) = \exp(\theta_1 + \theta_2 t + \theta_3 x + \theta_4 tx + \theta_5 x^2), \quad \mu_b(t) = \exp(\theta_6 + \theta_7 t). \tag{2.2.3}$$

This allowed an explicit calculation of $\mathbb{E}[Z|t,x]$ and thereby of μ_{ref}, assuming gamma-distributed frailties, see (14)–(18) in Jarner and Kryger (2011) for details. The resulting model for μ_{ref} had 8 parameters ($\theta_1 - \theta_7$ and a frailty parameter) for each sex, which were estimated by maximizing a standard Poisson likelihood. When forecasting, the parametric form was used to extrapolate μ_{ref} to form a deterministic trend around which μ_{target} would vary. We refer to this as a deterministic frailty model.

Despite its parsimonious structure, the estimated μ_{ref}-surface provided a remarkable fit to the international dataset. However, it soon became clear that the model was not able to fit other datasets equally well. More importantly, assessment of longevity risk was becoming a regulatory requirement, and for this purpose, a deterministic trend model was insufficient.

The current version of the SAINT model features a stochastic frailty model, in which μ_0 and μ_b are stochastic processes. This implies that the frailty term, $\mathbb{E}[Z|t,x]$, also becomes stochastic, and explicit expressions are no longer available. In order to disentangle the dependence between μ_0 and $\mathbb{E}[Z|t,x]$, a pseudo-likelihood approach, reinterpreting the frailty term in terms of observable quantities, had to be devised. This, in turn, led to a generally applicable "fragilization" method by which essentially any frailty distribution can be combined with any baseline and background mortality to form a model, cf. Section 2.4.

2.2.3 Cointegrating Gender Dynamics

In the original SAINT model, male and female mortality were modelled separately, which resulted in a sex differential of just over 10 years in forecasted cohort life

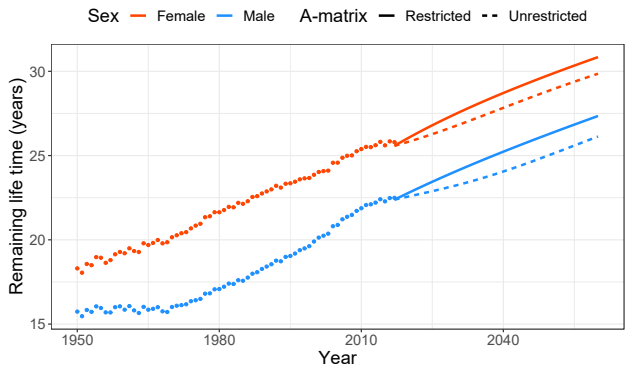

Figure 2.1: Age 60 actual (dots) and forecasted (lines) period life expectancy using the SAINT model with and without restrictions on the A-matrix from Equation (2.5.5).

expectancies at birth. When the first version of the stochastic frailty model was implemented, it was therefore decided to model the new stochastic processes via cointegration to ensure better aligned forecasts. At the time, the baseline and background mortality were modelled as

$$\mu_0(t,x) = \exp(\alpha_t + \beta_t x + 2x^2/10^4), \quad \mu_b(t) = \exp(\zeta_t), \tag{2.2.4}$$

with the processes $\{\alpha_t, \beta_t\}$ governing μ_0 being modelled as in Equation (2.5.5), but with an unrestricted A-matrix. Cointegration reduced the life expectancy difference at birth to about 5 years, and subsequent model development brought it further down to 3.6 years.

In Equation (2.5.5), the B-matrix controls the cointegrating relations, while the A-matrix controls the adjustments to these over time. It later became apparent that an unrestricted A-matrix could lead to complex transitory effects when reestablishing equilibrium relations, as seen on the dashed lines in Figure 2.1. The resulting projections were hard to justify and communicate, and eventually structural zeros were introduced in A to generate more linear projections. Allowing only pairwise dependence between parameters also offered a greater degree of explainability as the forecasting distribution simplified.

The cointegrating relations and the restrictions placed on the A-matrix were imposed, rather than formally tested. Even though formal testing could be done using the comprehensive statistical framework developed by Johansen (1995), it was deemed problematic that the underlying time dynamics could change annually following data updates as this would destabilize projections and potentially damage the model's credibility. Cointegration is therefore used merely as a modelling tool to achieve reasonable projections, rather than to gain insights into the joint behaviour of the time-varying mortality indices, see also Jarner and Jallbjørn (2020) for further discussion on this point.

2.2.4 Improving the Fit

The specification of μ_0 in (2.2.4) was the result of an extensive model search among "simple" models. At the time, it provided a reasonable fit, but eventually failed to adequately capture old-age mortality. Generalizations of linear mortality models typically involve adding a quadratic term to the age effect or introducing cohort components, see for example Cairns et al. (2009). The natural candidate to replace (2.2.4) was therefore the log-quadratic model $\mu_0(t,x) = \exp(\alpha_t + \beta_t x + \kappa_t x^2)$. But despite a clearly superior fit, its parameter estimates turned out exceedingly difficult to forecast.

After further research into the shape of the mortality age profile, the lack of fit was found to be caused by an inflexibility of (2.2.4) at the younger ages, a typical problem when fitting parsimonious models over a large age span. By replacing the fixed quadratic term in (2.2.4) with an excess slope parameter, the low mortality rates of the young were prevented from influencing the trend of the old. The baseline model therefore became

$$\mu_0(t,x) = \exp\left(\widetilde{\alpha}_t + \widetilde{\beta}_t x + \widetilde{\kappa}_t (x-75)\mathbb{1}_{\{x\geq 75\}}\right). \tag{2.2.5}$$

The model's parameters are linearized for forecasting purposes through reparametrization, achieved by setting $\alpha_t = \widetilde{\alpha}_t + 75\widetilde{\beta}_t$, $\beta_t = \widetilde{\beta}_t + \widetilde{\kappa}_t$, and $\kappa_t = -\widetilde{\kappa}_t$ whereby

$$\mu_0(t,x) = \exp\left(\alpha_t + \beta_t(x-75) + \kappa_t (x-75)\mathbb{1}_{\{x<75\}}\right). \tag{2.2.6}$$

Figure 2.2 shows the clear improvement in the model's fit. The fit is particularly impressive at the higher ages and also matches the logistic type behaviour seen in the data for the oldest-old as the frailty component comes into play.

Revisiting the Reference Population and Data Window

In the original paper, Jarner and Kryger (2011) used an aggregate of international data over the years 1933–2005, including, notably, data from the US. Following the first version of the stochastic frailty model, the left cut point of the data window was updated to 1950, while the right cut point was updated annually to match the most recent available data.

Over the years, it became apparent that the slowdown of the improvement rates in the US had begun to manifest itself in the long-term trend. At the time, the US constituted more than 40% of the reference population. In response, an extensive review of the demographic transitions in the Western world was conducted country for country. This work led to a new paradigm for putting together a more homogeneous and balanced data pool, necessitating an exclusion of the US. In the new dataset, two mortality regimes emerged, and, accordingly, the left cut point of the data window was updated to 1970. We note that other countries have also shown recent stagnating

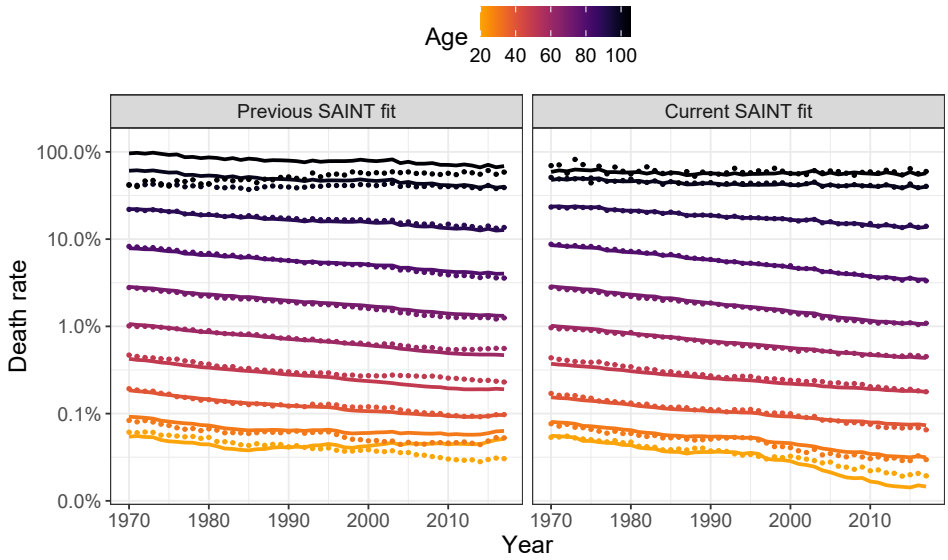

Figure 2.2: Observed (dots) female death rates for select ages with SAINT fits (solid lines) superimposed. The left panel shows the previous version of SAINT with μ_0 as in (2.2.4), estimated on a dataset with the US included and the window of calibration starting in 1950. The right panel shows the current version of SAINT with μ_0 as in (2.2.5), estimated on a dataset with the US excluded and the window of calibration starting in 1970.

rates of improvement, for example the UK, but overall there has been a continued improvement throughout the period, see Figure 2.3.

2.3 Modelling Changing Rates of Improvements

The mortality experience of several countries has shown increasing rates of improvements for older age groups while rates for younger groups have been decelerating, see for example Kannisto et al. (1994), Lee and Miller (2001), Booth et al. (2002), Bongaarts (2005), Li et al. (2013), and Vékás (2019). For long-term projections, in particular, it is important to model these changes to reduce the risk of underestimating future life expectancy gains.

2.3.1 Motivating Example

Although a plethora of models for modelling and forecasting mortality have been proposed in recent years, see e.g. Booth and Tickle (2008) and Janssen (2018) for an overview, the model of Lee and Carter (1992) is still by far the most widely used. Lee and Carter (1992) model the (observed) death rate, m, at time t for age x in a

log-bilinear fashion

$$\log m(t,x) = a_x + b_x k_t + \varepsilon_{t,x}, \tag{2.3.1}$$

where a and b are age-specific parameters and k is a time-varying index in which all temporal trends leading to improvements in mortality are encapsulated. The k-index is typically modelled as a random walk with drift, extrapolating the index linearly from the first to the last data point, resulting in constant rates of improvement in forecasted age-specific mortality.

Figure 2.3 illustrates the problem with assuming constant rates of improvements in forecasts. The figure shows actual and Lee-Carter forecasted period remaining life expectancies at age 60 for Western Europe and Denmark. Except for Western European females, the forecasts vary substantially with the estimation period due to changing rates of improvement. Moreover, the forecasts generally underestimate the future gains in life expectancy because rates of improvement for older age groups tend to increase over time.

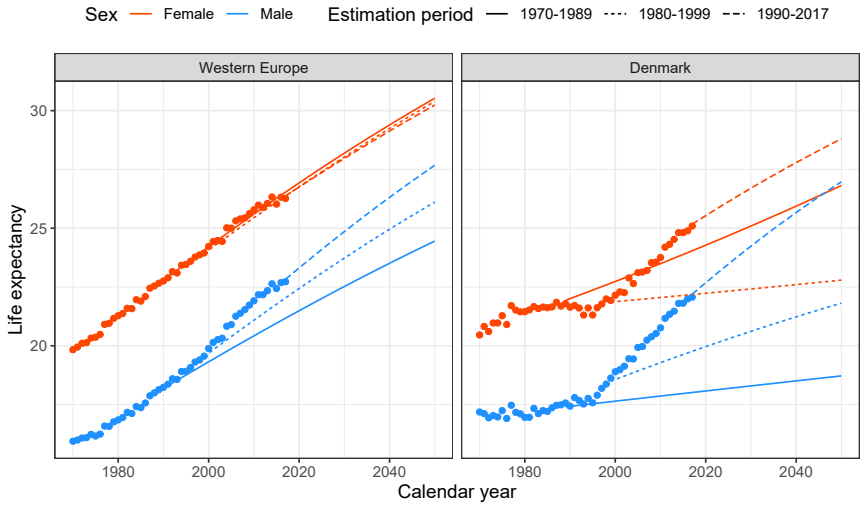

Figure 2.3: Age 60 actual (dots) and forecasted (lines) period life expectancy using a Lee-Carter model based on a rolling estimation window.

This phenomenon is not specific to the Lee-Carter model, but pertains to all models with constant rates of improvements. Fitting these models to shorter, more recent periods of data alleviates the downward bias to some extent, but it does not address the fundamental issue of changing rates. Coherent, multi-population mortality models, for example the model of Li and Lee (2005), can in principle produce changing rates of improvement while the individual populations "lock on" to common rates of improvement. However, the common rates are typically constant over time. Thus, coherence in itself does not guarantee the type of ongoing change

in improvement rates that we advocate. For a more detailed discussion of coherent models and their pros and cons, see Jarner and Jallbjørn (2020) and the references therein.

2.3.2 Frailty Theory

Frailty theory rests on the assumption that cohorts are heterogeneous and that some people are more susceptible to death (frail) than others. The difference in frailty causes selection effects in the population and leads to old cohorts being dominated by low mortality individuals.

Frailty theory is well-established in biostatistics and survival analysis, and several monographs are devoted to the topic, for example Duchateau and Janssen (2008), Wienke (2010), and Hougaard (2012). In demographic and actuarial science frailty models are also known as heterogeneity models. They have been used in mortality modelling to fit the logistic form of old-age mortality, see e.g. Wang and Brown (1998), Thatcher (1999), Butt and Haberman (2004), Olivieri (2006), Cairns et al. (2006), Spreeuw et al. (2013), and Li and Liu (2019), and to allow for overdispersion in mortality data, cf. Li et al. (2009). The SAINT model, however, employs frailty theory with the dual purpose of fitting old-age mortality and generating changing rates of improvement.

Below, we present a flexible class of continuous time models spanning multiple birth cohorts, with additive frailty and non-frailty components. With additive models, we can distinguish between "selective" mortality influenced by frailty and "background" mortality not affected by frailty, for example accidents. Following Vaupel et al. (1979), individual frailty is a non-negative stochastic quantity Z that acts multiplicatively on an underlying baseline mortality rate. We assume that frailty is assigned at birth (according to some distribution) and remains constant throughout an individual's life span. In this context, we can interpret frailty as individual (congenital) genetic differences. Conditionally on frailty being Z, mortality at age x at time t takes the form

$$\mu(t,x|Z) = Z\mu_0(t,x) + \mu_b(t,x), \tag{2.3.2}$$

where μ_0 is the baseline rate describing age-period effects influenced by individual frailty and μ_b is background mortality common to all individuals regardless of their respective frailties.

Equation (2.3.2) describes the mortality rate of an individual, but this quantity is not observable in population-level data. In fact, we only observe an aggregate of the death rates. We can derive an explicit expression for this aggregate, namely the population-level rate, by writing up the survival function

$$S(t,x) = e^{-\int_0^x \mu(t-x+u,u)\,\mathrm{d}u} = \mathbb{E}\left[e^{-\int_0^x \mu(t-x+u,u|Z)\,\mathrm{d}u}\right] \tag{2.3.3}$$

and differentiating $-\log S(t,x)$ to get

$$\mu(t,x) = \mathbb{E}[Z|t,x]\mu_0(t,x) + \mu_b(t,x). \tag{2.3.4}$$

Here, $\mathbb{E}[Z|t,x]$ is the mean frailty among the survivors of the cohort born at time $t-x$. As a matter of convention, we assume without loss of generality that mean frailty is one at birth. It is useful to introduce the Laplace transform $\mathcal{L}(s) = \mathbb{E}[\exp(-sZ)]$ of the common frailty distribution in which case

$$\mathbb{E}[Z|t,x] = \frac{-\mathcal{L}'(\mathcal{M}_0(t,x))}{\mathcal{L}(\mathcal{M}_0(t,x))}, \tag{2.3.5}$$

where $\mathcal{M}_0(t,x) = \int_0^x \mu_0(t-x+u,u)\,\mathrm{d}u$ is the cumulated baseline rate.

So far, the expressions above relating mean frailty to the baseline rate are standard in survival analysis. For later use, we establish an additional relationship between mean frailty and the cumulated cohort rate adjusted for background mortality, namely

$$\mathcal{M}(t,x) = \int_0^x (\mu(t-x+u,u) - \mu_b(t-x+u,u))\,\mathrm{d}u, \tag{2.3.6}$$

via its survival function

$$\exp(-\mathcal{M}(t,x)) = S(t,x)e^{\int_0^x \mu_b(t-x+u,u)\,\mathrm{d}u} = \mathcal{L}(\mathcal{M}_0(t,x)). \tag{2.3.7}$$

Introducing the function $\nu(\cdot) = -\log\mathcal{L}(\cdot)$, we have $\mathcal{M}(t,x) = \nu(\mathcal{M}_0(t,x))$ and $\mathcal{M}_0(t,x) = \nu^{-1}(\mathcal{M}(t,x))$ which gives us

$$\mathbb{E}[Z|t,x] = \nu'(\mathcal{M}_0(t,x)) = \nu'(\nu^{-1}(\mathcal{M}(t,x))), \tag{2.3.8}$$

upon insertion into (2.3.5).

The relation between mean frailty and mortality described by Equation (2.3.8) will be central to our estimation approach. Whereas $\mathcal{M}_0$ is given solely in terms of the baseline rate, $\mathcal{M}$ can be estimated using empirical death rates. Substituting $\mathcal{M}$ by such an estimate disentangles the frailty distribution from the baseline rate which greatly simplifies the estimation procedure whenever parametric structures have been imposed on μ_0 and μ_b. We return to the specifics in Section 2.4.

To apply frailty theory in practice, we must identify suitable choices of frailty distributions. A brief account of appropriate distributions is given in Appendix 2.A. Essentially, useful distributions are the ones with an explicit Laplace transform. The most commonly used distribution is the Gamma distribution which has a tractable Laplace transform, see Example 2.4.1, along with other desirable properties.

2.3.3 Frailty leads to Changing Rates of Improvements

To clarify how the inclusion of frailty leads to changing rates of improvements, we define the rate of improvement in selective mortality as

$$\rho_s(t,x) = -\frac{\partial}{\partial t}\log\mathbb{E}[Z|t,x] - \frac{\partial}{\partial t}\log\mu_0(t,x), \tag{2.3.9}$$

where $\rho_0(t,x) = -\frac{\partial}{\partial t}\log\mu_0(t,x)$ is the rate of improvement in baseline mortality.

Suppose that we model the period effect so that baseline mortality is decreasing over time, that is, $\mu_0(t,x) \to 0$ as $t \to \infty$ for fixed age x. Then, the cumulated baseline will also be decreasing over time, that is $\mathcal{M}_0(t,x) \to 0$ as $t \to \infty$, while the mean frailty in successive cohorts will be *increasing* over time due to less and less selection of frail individuals, so $\mathbb{E}[Z|t,x] \to 1$ as $t \to \infty$. From (2.3.9), we see that improvements in cohort mortality are smaller than, but gradually increasing to, the rate of improvement at the individual level.

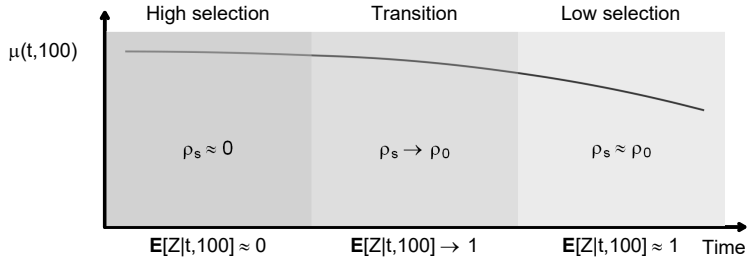

Figure 2.4: Illustration of population level mortality at age 100 over time. When selection is high, observed mortality rates do not improve much even though rates are assumed to be decreasing at the individual level. This is due to improvements in baseline mortality being (partially) offset by increases in mean frailty. As mean frailty eventually approaches one, observed improvements and improvements in the underlying mortality are approximately equal.

This line of thinking offers an explanation to the small improvement rates observed in old-age mortality, and it suggests that we might expect to see higher rates of improvements in the future. At old ages where death rates – and thereby selection – are high, the change in mean frailty over time can substantially offset improvements in baseline mortality. This makes improvements in observed mortality close to zero, cf. Equation (2.3.9), a behaviour illustrated in Figure 2.4. As improvements in baseline mortality continue to occur at the individual level, the selection mechanism gradually weakens and improvements in observed mortality get closer to the underlying improvement rates. The pattern of gradually changing improvement rates of old-age mortality resembles what is seen in the data. We will return to this point in Section 2.5.

The Difference between Frailty and the Traditional Cohort Perspective

While conditional mean frailty $\mathbb{E}[Z|t,x]$ may be regarded as a cohort component in the sense that it focuses on the evolution of a cohort through time, the notion is qualitatively different from the traditional cohort perspective in mortality modelling. As an illustrative example, consider the cohort extended Lee-Carter model of Renshaw

and Haberman (2006), that is

$$\log \mu(t,x) = a_x + b_x k_t + c_{t-x}. \tag{2.3.10}$$

The cohort component $\exp(c_{t-x})$ assumes the role of a dummy variable and offsets mortality by the same factor throughout the life span of individuals born at time $t-x$. Although the multiple $\mathbb{E}[Z|t,x]\mu_0(t,x)$ does in principle contain the structure (2.3.10) with $\mu_0(t,x) = \exp(a_x + b_x k_t)$ being the Lee-Carter model, $\mathbb{E}[Z|t,x]$ is not constant over time except in degenerate cases. Indeed, $\mathbb{E}[Z|t,x]$ progressively decreases for a given cohort as selection intensifies and should therefore not be regarded as a cohort component in the traditional sense.

2.3.4 Alternative Ways of addressing Changing Improvement Rates

With a growing body of empirical evidence against the assumption of constant age-specific rates of improvements, several other approaches have been developed to address the issue. One suggestion involves finding an "optimal" calibration period for which model assumptions are not violated. This usually involves fitting (Lee-Carter) models to shorter and more recent periods of data, see for example Lee and Miller (2001), Tuljapurkar et al. (2000), Booth et al. (2002), and Li and Li (2017). For projections over longer horizons the problem persists; future increases in especially old-age mortality beyond what has been seen historically cannot be captured by making an informed decision about the period of calibration.

Within the Lee-Carter framework, another solution is to extend the original model by replacing the time-invariant age-response with a time-varying version, so instead of Equation (2.3.1) we would have

$$\log m(t,x) = a_x + B_{t,x} k_t + \varepsilon_{t,x}. \tag{2.3.11}$$

A freely varying $B_{t,x}$ introduces far too many parameters, and some restrictions have to be imposed. Li et al. (2013) suggest that $B_{t,x}$ describe the "rotating" pattern of mortality improvements, namely that improvement rates are declining for the young but increasing for the old. Consequently, Equation (2.3.11) is sometimes referred to as the rotated Lee-Carter model. Li et al. (2013) achieve rotation by letting the b_x's from the original model converge (smoothly) to some assumed long-term target B_x. The approach has since been adopted and adapted by a number of authors, for example Vékás (2019) and Gao and Shi (2021).

Another idea is to model and project improvement rates directly as opposed to projecting death rates, see for example Haberman and Renshaw (2012). Various forms of projections built from age-specific improvement rates applied to reference mortality tables are also becoming popular among actuarial practitioners, see for example Jarner and Møller (2015) for a detailed account of the longevity benchmark

employed by the Danish financial supervisory authority or the model used by the Continuous Mortality Investigation (2016).

The alternative approaches mentioned above each have their own merits as ways of addressing changing rates of improvements. Frailty, however, has the unique advantage that it can be introduced into any existing mortality model to forecast improvement rates higher than those observed historically, while preserving both the original model structure and the underlying driver of the system.

2.4 Stochastic Frailty Models

In the following, we detail how frailty can be used with any stochastic mortality model and we show how to estimate and forecast these models. The extension from deterministic to stochastic frailty is a fundamental point in practical applications where the ability to describe uncertainty of forecasts is essential for managing longevity risk.

2.4.1 Data and Terminology

Data are assumed to be on the form of death counts, $D(t,x)$, and corresponding exposures, $E(t,x)$, over time-age cells of the form $[t,t+1)\times[x,x+1)$ for a range of calendar years $t\in\{t_{\min},\dots,t_{\max}\}=\mathcal{T}\subseteq\mathbb{N}$ and ages $x\in\{x_{\min},\dots,x_{\max}\}=\mathcal{X}\subseteq\mathbb{N}_0$. From the death counts and risk exposures, we define the observed (empirical) death rate as the ratio

$$m(t,x)=D(t,x)/E(t,x). \tag{2.4.1}$$

The empirical death rate is a nonparametric estimate of the underlying cohort rate, $\mu(t,x)$, which for modelling purposes is also assumed constant over $[t,t+1)\times[x,x+1)$.

2.4.2 Modelling Framework

For ease of presentation, we will consider stochastic models for baseline and background mortality of the form

$$\mu_0(t,x)=F(\theta_t,\eta_x), \tag{2.4.2}$$
$$\mu_b(t,x)=G(\zeta_t,\xi_x), \tag{2.4.3}$$

where F and G are functional forms taking parameters in the age-period dimension as input. All quantities may be multidimensional, and further generalizations to include cohort effects and general dependence structures are possible if so desired. We assume that parameters are to be estimated from data, but they can also be fixed or empty.

We define a (generalized) stochastic frailty model as a model of the form

$$D(t,x) \sim \text{Poisson}(\mu(t,x)E(t,x)), \tag{2.4.4}$$

$$\mu(t,x) = \mathbb{E}[Z|t,x]\mu_0(t,x) + \mu_b(t,x), \tag{2.4.5}$$

with μ_0 and μ_b given by (2.4.2)–(2.4.3) while $\mathbb{E}[Z|t,x]$ denotes conditional mean frailty as in Section 2.3.2. Note that $\mathbb{E}[Z|t,x]$ is now a stochastic quantity since it depends on μ_0. The frailty distribution at birth is the same for all cohorts, and it is assumed to belong to a family indexed by ψ with Laplace transform $\mathcal{L}_\psi$ available in explicit form. The parameters of the model are thus $(\psi, \theta, \eta, \zeta, \xi)$ where all components can be vectors.

Based on (2.3.8) we can write

$$\mu(t,x) = \nu'_\psi(\mathcal{M}_0(t,x))F(\theta_t, \eta_x) + G(\zeta_t, \xi_x) \tag{2.4.6}$$

with $\nu_\psi(\cdot) = -\log \mathcal{L}_\psi(\cdot)$ and $\mathcal{M}_0(t,x) = \sum_{u=0}^{x-1} F(\theta_{u+t-x}, \eta_u)$. Inserting the above into (2.4.4), all parameters can be estimated jointly from the resulting likelihood. This likelihood is, however, rather intractable with frailty and remaining parameters occurring in a complex mix. Consequently, estimation has to be handled on a case-by-case basis depending on the choice of frailty distribution and mortality model. Below we propose an alternative, generally applicable pseudo-likelihood approach which greatly simplifies the estimation task.

2.4.3 Pseudo-Likelihood Function

We seek to replace the problematic term $\mathbb{E}[Z|t,x]$ with an estimate that does not depend on baseline parameters. From (2.3.8), we have that $\mathbb{E}[Z|t,x] = \nu'_\psi(\nu_\psi^{-1}(\mathcal{M}(t,x)))$, where $\mathcal{M}$ is the cumulated frailty-dependent part of mortality. At first sight this does not seem to help much since $\mathcal{M}$ is even more complicated than $\mathcal{M}_0$. However, in contrast to $\mathcal{M}_0$ we can obtain an estimate of $\mathcal{M}$ which does not involve the baseline parameters. With a slight abuse of notation, suppressing the dependency on the background parameters, we estimate $\mathcal{M}$ by

$$\widetilde{\mathcal{M}}(t,x) = \sum_{u=0}^{x-1} \widetilde{m}(t-x+u, u), \tag{2.4.7}$$

where

$$\widetilde{m}(t,x) = \begin{cases} m(t_{\min}, x) - G(\zeta_{t_{\min}}, \xi_x) & \text{for } x_{\min} \le x \le x_{\max} \text{ and } t < t_{\min}, \\ m(t,x) - G(\zeta_t, \xi_x) & \text{for } x_{\min} \le x \le x_{\max} \text{ and } t_{\min} \le t \le t_{\max}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4.8}$$

is an extension of the empirical death rates (with background mortality subtracted). The extension is required because the summation in (2.4.7) falls partly outside the

data window; we need to know the death rates from birth to the present or maximum age for all cohorts entering the estimation. The gray area of Figure 2.5 illustrates the "missing" death rates. The extension implies that selection prior to $t_{\min}$ has happened according to initial rates (rather than actual rates) and that all cohorts have mean frailty one at age $x_{\min}$ (rather than at birth).

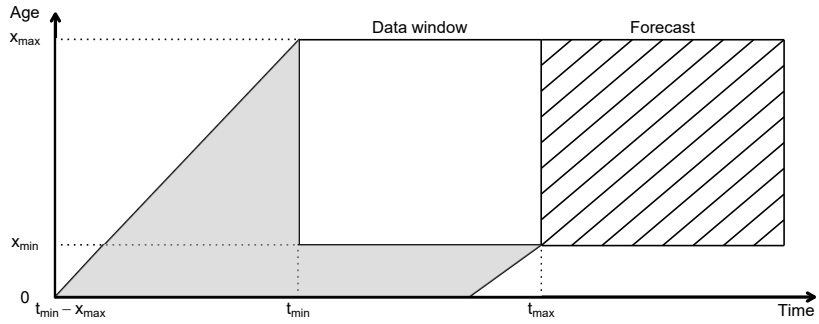

Figure 2.5: Data is available for years between $t_{\min}$ and $t_{\max}$ and for ages between $x_{\min}$ and $x_{\max}$. The grey area below and to the left of the data window illustrates the part of the trajectories needed for calculation of $\widetilde{\mathcal{M}}$ that falls outside the data window. The cross hatched area to the right illustrates the years and ages for which we wish to forecast mortality.

We propose to base estimation of (2.4.4)–(2.4.5) on a likelihood function in which the term $\mathbb{E}[Z|t,x]$ is replaced by $\nu'_\psi(\nu_\psi^{-1}(\widetilde{\mathcal{M}}(t,x)))$. The resulting approximate likelihood function is referred to as a pseudo-likelihood function, cf. Besag (1975), and corresponds to estimating the modified model

$$D(t,x) \sim \text{Poisson}(\mu(t,x)E(t,x)), \tag{2.4.9}$$

$$\mu(t,x) = \nu'_\psi(\nu_\psi^{-1}(\widetilde{\mathcal{M}}(t,x)))F(\theta_t,\eta_x) + G(\zeta_t,\xi_x). \tag{2.4.10}$$

Importantly, the cohort rate μ is separable in frailty and baseline parameters. The model is therefore considerably easier to handle than (2.4.4)–(2.4.5) in the sense that joint estimation can be based on marginal estimation procedures for the baseline and background intensities, μ_0 and μ_b. Equation (2.4.10) might still look daunting, but it simplifies in specific cases.

Example 2.4.1. *When Z follows a Gamma distribution with mean one and variance σ^2 the Laplace transform and conditional mean frailty are given by*

$$\mathcal{L}(s) = \left(1+\sigma^2 s\right)^{-1/\sigma^2}, \tag{2.4.11}$$

$$\mathbb{E}[Z|t,x] = \left(1+\sigma^2\mathcal{M}_0(t,x)\right)^{-1} = \exp\left(-\sigma^2\mathcal{M}(t,x)\right), \tag{2.4.12}$$

whereby Equation (2.4.10) *reads*

$$\mu(t,x) = \exp\left(-\sigma^2\widetilde{\mathcal{M}}(t,x)\right)F(\theta_t,\eta_x) + G(\zeta_t,\xi_x). \tag{2.4.13}$$

2.4.4 Maximum Likelihood Estimation

Maximum likelihood estimates of the model (2.4.9)–(2.4.10) can be obtained by optimizing the profile (pseudo) log-likelihood function,

$$\ell(\psi) = \log L(\psi, \hat{\theta}(\psi), \hat{\eta}(\psi), \hat{\zeta}(\psi), \hat{\xi}(\psi)), \tag{2.4.14}$$

where L is the likelihood resulting from (2.4.9)–(2.4.10) and $\hat{\theta}(\psi), \hat{\eta}(\psi), \hat{\zeta}(\psi)$ and $\hat{\xi}(\psi)$ denote the maximum likelihood estimates for fixed value of ψ. Since the frailty family is typically of low dimension, the profile log-likelihood function can usually be optimized reliably by general purpose optimization routines. This is particularly simple when $\mu_b(t, x) = 0$ for all $t \in \mathcal{T}$ and $x \in \mathcal{X}$.

In the general setting with non-zero background mortality, the model describes a situation of competing risks. The interpretation is that individuals of age x at time t are susceptible to death from two different, mutually exclusive sources with intensities μ_s and μ_b. This structure is natural to consider in many contexts, but the likelihood function is complicated and direct estimation of the parameters may prove difficult even for fixed ψ.

Assuming that we have separate routines available for estimating the baseline and background mortality models, we can exploit the additive structure of (2.4.9)–(2.4.10) using the EM algorithm of Dempster et al. (1977). It can be shown that the likelihood is increased in each step of the EM algorithm and thus converges to a local maximum, although it may do so rather slowly. It is, however, a great advantage that we can use the same top-level procedure to estimate virtually any combination of mortality models, especially when estimation routines for the underlying models are readily available. It is also straightforward to extend the EM algorithm to more than two competing risks.

Alternatively, if computational efficiency is essential, we can carry out estimation by Newton-Raphson sweeps over frailty and mortality model parameters, but this requires a substantial amount of tailor-made code. We find that optimizing the log-likelihood using the EM algorithm is both flexible, easy to implement and in our experience sufficiently fast and robust to be of practical use. The algorithm is detailed in Appendix 2.B.

2.4.5 Forecasting

Suppose that we have estimates $(\hat{\psi}, \hat{\theta}, \hat{\eta}, \hat{\zeta}, \hat{\xi})$ available. Following the usual approach in stochastic mortality modelling, death rates are to be projected using a time-series model for the time-varying parameters $\{\theta_t, \zeta_t\}_{t \in \mathcal{T}}$. Typically a (multi-dimensional) random walk with drift is used, see for example Lee and Carter (1992) or Cairns et al. (2006), but models with more complex structure can also be applied. Let an overbar denote projected parameters and assume that these are available for

$t \in \{t_{\max}+1, \ldots, t_{\max}+h\}$ given a forecasting horizon $h \in \mathbb{N}_+$. The forecast region is illustrated as the cross-hatched box in Figure 2.5.

Baseline and background mortality rates are readily projected by inserting $\bar{\theta}_t$ and $\hat{\eta}_x$ into (2.4.2) and $\bar{\zeta}_t$ and $\hat{\xi}_x$ into (2.4.3). Forecasting mean frailty is slightly more involved. We notice that while it was convenient to specify mean frailty in terms of $\mathcal{M}$ for estimation purposes, it is practical to express it in terms of $\mathcal{M}_0$ when forecasting, because $\mathcal{M}_0$ can be computed directly from F throughout the forecast region. Mortality is thus projected via

$$\mu(t,x) = \nu'_{\hat{\psi}}(\widetilde{\mathcal{M}}_0(t,x))F(\bar{\theta}_t, \hat{\eta}_x) + G(\bar{\zeta}_t, \hat{\xi}_x), \tag{2.4.15}$$

where $\widetilde{\mathcal{M}}_0(t,x)$ in the forecast region is given by the recursion

$$\widetilde{\mathcal{M}}_0(t,x) = \begin{cases} \widetilde{\mathcal{M}}_0(t_{\max}, x-1) + F(\hat{\theta}_{t_{\max}}, \hat{\eta}_{x-1}) & \text{for } x_{\min} < x \text{ and } t_{\max}+1 = t, \\ \widetilde{\mathcal{M}}_0(t-1, x-1) + F(\bar{\theta}_{t_{\max}}, \hat{\eta}_{x-1}) & \text{for } x_{\min} < x \text{ and } t_{\max}+1 < t, \\ 0 & \text{for } x_{\min} = x. \end{cases} \tag{2.4.16}$$

We underline that G does not enter $\widetilde{\mathcal{M}}_0$ in the forecast, whereas in the data window $\widetilde{\mathcal{M}}_0$ is defined by the transformation $\widetilde{\mathcal{M}}_0(t,x) = \nu^{-1}_{\hat{\psi}}(\widetilde{\mathcal{M}}(t,x))$ to ensure consistency with the estimated model.

Example 2.4.2. *Continuing Example 2.4.1 where frailty is Gamma distributed with mean one and estimated variance $\hat{\sigma}^2$, we get using* (2.4.12) *in conjunction with* (2.4.15) *that mortality should be forecasted via*

$$\mu(t,x) = \Big(1 + \hat{\sigma}^2 \widetilde{\mathcal{M}}_0(t,x)\Big)^{-1} F(\bar{\theta}_t, \hat{\eta}_x) + G(\bar{\zeta}_t, \hat{\xi}_x), \tag{2.4.17}$$

with $\widetilde{\mathcal{M}}_0$ given by (2.4.16) *in the forecast region. In the data window, $\widetilde{\mathcal{M}}_0(t,x)$ can be expressed as $\widetilde{\mathcal{M}}_0(t,x) = [\exp(\hat{\sigma}^2 \widetilde{\mathcal{M}}(t,x)) - 1]/\hat{\sigma}^2$ using* (2.4.12)*.*

2.5 An Application to International Mortality

We consider the implementation of the stochastic frailty model currently used at ATP and make comparisons to the usual Lee-Carter methodology. The application is based on mortality data retrieved from the Human Mortality Database (2021). To allow the reader to reproduce the results, we use Danish data to model the spread, rather than proprietary ATP data. Sections 2.5.1–2.5.3 cover reference population mortality, while Section 2.5.4 discusses spread modelling for target population mortality.

2.5.1 An International Reference Trend

The reference mortality trend, denoted μ_{ref}, belongs to the class of stochastic frailty models (2.4.2)–(2.4.3), and is gender-specific although this is not shown explicitly

in the notation. The model assumes Gamma distributed frailty with mean one and variance σ^2 and takes the functional form

$$\mu_{\text{ref}}(t,x) = \exp\left(-\sigma^2\widetilde{\mathcal{M}}(t,x)\right)\mu_0(t,x) + \mu_b(t), \tag{2.5.1}$$

$$\mu_0(t,x) = \exp\left(\alpha_t + \beta_t(x-75) + \kappa_t(x-75)\mathbb{1}_{\{x<75\}}\right), \tag{2.5.2}$$

$$\mu_b(t) = \exp\left(\zeta_t\right). \tag{2.5.3}$$

In (2.5.1), the variance of the frailty distribution expresses the amount of heterogeneity in the population, but since any estimate depends on the choice of μ_0, the quantity can only be interpreted in a model-specific context. On the other hand, its influence on the mortality curve is clear. If death rates increase with age, the function $x \mapsto \exp(-\sigma^2\widetilde{\mathcal{M}}(t,x))$ decreases from one towards zero and describes how μ_0 is "dragged" down by the frailty component. If σ^2 is close to zero, then mean frailty is close to one for all ages. As the variance grows, the decline in mean frailty steepens. This drags down the old-age part of the mortality curve and eventually so much that the rates fall into a decline.

A Lee–Carter Baseline?

Instead of the Gompertzian model applied in (2.5.2), one could use a Lee-Carter model for μ_0, see Jarner (2014) for such an application. However, we find that the parametric structure in (2.5.2) is favourable in terms of stability and preserving the overall structure of the data and, in particular, smooth and increasing age-profiles, which is not guaranteed in long-term Lee-Carter forecasts.

Moreover, the parameters of the Lee-Carter model are not identifiable without additional constraints which precludes the use of more flexible time dynamics such as error-correction models, a limitation the parametric (identifiable) structure does not have, see for example Hunt and Blake (2018) and Jarner and Jallbjørn (2020) for a detailed discussion.

Lastly, unlike the parametric form (2.5.2) that readily expands in the age dimension, that is the model extends to ages not part of the estimation, the Lee-Carter model applies only to the age span used in the calibration. This is a pronounced problem at the highest ages where data are often sparse. To obtain reliable and stable rates in both model and forecasts, one typically employs a Kannisto extension (Kannisto et al., 1994), or a similar procedure, for example the methods discussed in Pitacco et al. (2009), for the oldest-old. Irrespective of the extrapolation procedure, the coupling of two separate models adds an additional layer of complexity and defeats part of the purpose for introducing frailty, namely to capture the logistic type old-age mortality behaviour seen in the data.

Selecting a suitable Reference Population

To establish an international reference population, we have to decide on a suitable list of countries to use. While all countries appear to follow the same long-term trend, mortality improvements occur at different times for individual countries and variation in improvement rates differ as well. Ideally, the reference trend should consist of countries that reflect the prevalent mortality regime, but their historical development ought to be comparable as well. That is, the countries chosen ought to have undergone the same stages of demographic transition at roughly the same time.

It proves quite difficult to find a set of rules for selecting countries satisfying these broad criteria. Kjærgaard et al. (2016) propose an out-of-sample selection criterion as a way of constructing an "optimal" set of countries. Others have established hard inclusion criteria based on various socio-demographic-indicators, for example the Dutch actuarial society who base their official projections on a peer group of all European countries with a per capita GDP above the European average (Antonio et al., 2017).

We remain agnostic with regard to specific rules. Having populations entering or leaving the data pool following (annual) data updates is almost certainly bound to cause problems in terms of model stability. Our advice is to choose a wide range of countries with the intention of sticking with them in the long term. With an outset in the countries available from the Human Mortality Database and excluding countries only if they clearly violate the criteria above, we are left with primarily the western part of Europe. In particular, the SAINT model is currently based on pooled data from the following 18 countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany (West), Iceland, Italy, Luxembourg, Netherlands, Norway, Scotland, Spain, Sweden, Switzerland and UK (England and Wales).

2.5.2 Time Dynamics

The SAINT model (2.5.2)–(2.5.3) has four time-varying parameters that need to be forecasted (for each sex). The Makeham component ζ and the excess slope κ are nuisance parameters included in the model to add sufficient flexibility to fit the historical data and to enhance interpretability of the level α and the slope β. Even though $\{\zeta_t\}_{t\in\mathcal{T}}$ and $\{\kappa_t\}_{t\in\mathcal{T}}$ appear non-stationary, see Figure 2.6, modelling their trending behaviour has little effect on mortality projections. Striking a compromise between model complexity and performance, we forecast these parameters using a random walk without drift, so for any horizon $h \in \mathbb{N}_0$ we have

$$\kappa_{t_{\max}+h}|\kappa_{t_{\max}} \sim \mathcal{N}\left(\kappa_{t_{\max}}, h\sigma_\kappa^2\right) \quad \text{and} \quad \zeta_{t_{\max}+h}|\zeta_{t_{\max}} \sim \mathcal{N}\left(\zeta_{t_{\max}}, h\sigma_\zeta^2\right), \tag{2.5.4}$$

where $\sigma_\kappa, \sigma_\zeta \in \mathbb{R}_+$.

Projecting $\{\alpha_t, \beta_t\}$ is a more delicate task. In particular, some thought should go into the joint behaviour of the gender-specific forecasts. It is a well established fact that women live longer than men and while this gender gap varies over time it is believed to persist. Since forecasting even closely related populations independently will lead to diverging forecasts, cf. Tuljapurkar et al. (2000), we must deal with the problem through joint modelling in order not to have undesirable scenarios such as projecting men to live longer than women.

An Error-Correction Model

To ensure that female and male parameters "stay together", not just in a median forecast but also for every stochastic realization, we need parameters to cointegrate, that is a given linear combination of them should be stationary. This is achieved by forecasting from the error-correction model

$$\Delta Y_t = AB^\top Y_{t-1} + C + \omega_t, \tag{2.5.5}$$

where $\omega_t \overset{\text{iid}}{\sim} \mathcal{N}_4(0, \Omega)$, $\Delta Y_t = Y_t - Y_{t-1}$, and

$$Y_t = \begin{pmatrix} \alpha_t^f \\ \alpha_t^m \\ \beta_t^f \\ \beta_t^m \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & 0 \\ a_{21} & 0 \\ 0 & a_{32} \\ 0 & a_{42} \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}, \tag{2.5.6}$$

with superscript f and m denoting female and male parameter values, respectively. Equation (2.5.6) contains two critical assumptions:

1. Structural zero's have been placed in the A-matrix making parameter pairs weakly exogenous, so that the relation between the α's will not affect the relation between the β's and vice versa.

2. The B-matrix imposes two cointegrating relations, one between each parameter pair, with coefficients of unity so that the model corrects any disequilibrium that may arise in the difference between the parameters.

The error-correction behaviour of (2.5.5) can be made more precise by decomposing the drift term. For a $p \times q$ matrix X of full rank, we say that another matrix $X_\perp$ of full rank and dimension $p \times (p-q)$ such that $X^\top X_\perp = 0$ is its orthogonal complement. With $M = A(B^\top A)^{-1}B^\top$ it can be shown that $I - M = B_\perp(A_\perp^\top B_\perp)^{-1}A_\perp^\top$, cf. Chapter 3 of Johansen (1995). Using this identity we can rewrite (2.5.5) so that

$$\Delta Y_t = A(B^\top Y_{t-1} - C_D) + C_\Delta + \omega_t, \tag{2.5.7}$$

where $C_D = -(B^\top A)^{-1}B^\top C = \left(C_{D_\alpha}, C_{D_\beta}\right)^\top$ is the parameter difference in stationarity with $C_{D_\alpha} = \frac{c_1 - c_2}{a_{21} - a_{11}}$ and $C_{D_\beta} = \frac{c_3 - c_4}{a_{42} - a_{32}}$, while $C_\Delta = (I - M)C =$

$(C_{\Delta_\alpha}, C_{\Delta_\alpha}, C_{\Delta_\beta}, C_{\Delta_\beta})^\top$ is the common drift with $C_{\Delta_\alpha} = \frac{a_{21}c_1 - a_{11}c_2}{a_{21} - a_{11}}$ and $C_{\Delta_\beta} = \frac{a_{42}c_3 - a_{32}c_4}{a_{42} - a_{32}}$. From Equation (2.5.7), it is clear Y_t is updated in response to the disequilibrium error $B^\top Y_{t-1} - C_D$ with a force depending on the magnitude of the a's.

An explicit representation of the (median) forecast for a horizon $h \in \mathbb{N}_0$ can be discerned via the Granger representation theorem, see e.g. Jarner and Jallbjørn (2020). We have

$$\mathbb{E}[Y_{t_{\max}+h} | Y_{t_{\max}}] = Y_{t_{\max}} + C_\Delta h - A(B^\top A)^{-1} \begin{pmatrix} 1 - \lambda_\alpha^h & 0 \\ 0 & 1 - \lambda_\beta^h \end{pmatrix} (B^\top Y_{t_{\max}} - C_D), \tag{2.5.8}$$

where $\lambda_\alpha = 1 + a_{11} - a_{21}$ and $\lambda_\beta = 1 + a_{32} - a_{42}$ are eigenvalues of $I + AB^\top$. Equation (2.5.8) highlights the asymptotic random walk behaviour, with the initial disequilibrium error decaying exponentially to zero provided that $\lambda_\alpha, \lambda_\beta < 1$.

Writing out the error-correction behaviour helps us identify means of ensuring stable and robust forecasts. With this aim in mind, adjusting C_D and C_Δ to equal desired long-term values might be preferable compared to unconstrained estimation. Equation (2.5.8) details how the target equilibrium may severely affect projections. It shows that the error-correction model can bring about a number of undesirable features in the median forecast, like short- to medium-term increases in mortality for one of the genders during the restoration of the equilibrium. It is difficult to justify such behaviours in best estimate projections. To avoid them, we assume that $B^\top Y_t$ is distributed according to its stationary distribution by equating C_D to the difference between jump-off values, namely $\widehat{C}_{D_\theta} = \widehat{\theta}^f_{t_{\max}} - \widehat{\theta}^m_{t_{\max}}$ where θ is either α or β. This makes the median forecast of the error-correction model coincide with the median forecast of a random walk model, since the ultimate term in (2.5.8) vanishes. The error-correction interpretation still applies, however, when considering a stochastic realization of the process. Moreover, looking at the evolution of mortality through the years, for example Figure 2.3, it is clear that female mortality has developed more steadily than mortality for the males. We therefore use the empirical average of the female parameters to model the common slope, that is $\widehat{C}_{\Delta_\theta} = \frac{1}{t_{\max} - t_{\min} + 1} \sum_t \Delta \widehat{\theta}^f_t$ where θ is either α or β.

2.5.3 Model Fit and Forecast

Since international mortality develops more steadily than country-specific mortality, we are able to use a relatively wide window of data for model calibration. The endpoints should be chosen such that we do not introduce structural breaks if data for some countries are missing. Balancing these concerns, we apply the model to international mortality data, ages 20–100 and calendar years 1970–2017 with 2017 being the last year where data exist for all the countries considered at the time of

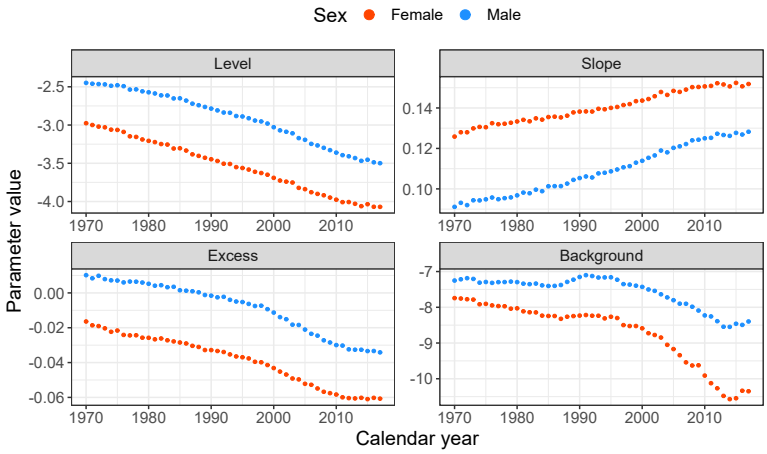

Figure 2.6: Estimated parameters for baseline (2.5.2) and background (2.5.3) mortality in the SAINT model.

writing. The model is estimated separately for females and males, and follows the EM algorithm described in Appendix 2.B.

The estimated parameters of the SAINT model (2.5.2)–(2.5.3) are shown in Figure 2.6. To justify the use of the error-correction model (2.5.5), the series $\{\alpha_t\}_{t\in\mathcal{T}}$ and $\{\beta_t\}_{t\in\mathcal{T}}$ must be integrated of order 1 for both genders, i.e. the stochastic part of these processes must be non-stationary. We conclude from unit root tests (test results not shown) that this is indeed the case. Further, we take the observed stable difference between parameters as (weak) evidence that they engage in an equilibrium correcting relationship.

In mortality forecasting applications, the choice of jump-off point is a key consideration to match the start of the projection with recently observed data (Lee and Miller, 2001). Because the SAINT model fits the empirical death rates very well, cf. Figure 2.2, no jump-off correction is needed and we use the model values in the jump-off year to determine the jump-off rates. Moreover, the Poisson assumption (2.4.4) ensures that the model approximates the total number of deaths in the data, see also the discussion in Brouhns et al. (2002).

Comparing SAINT and Lee–Carter Forecasts

To put forecasting into perspective, the SAINT projections are compared to projections generated by a Lee-Carter model. The Lee-Carter model is, in the spirit of this paper, estimated under the Poisson assumption as in Brouhns et al. (2002) and we use a random walk with drift for forecasting as is customary. Extrapolating the time-varying index linearly implies that age-specific death rates decay exponentially at a *constant* rate. We shall see that this assumption causes forecasts based on the

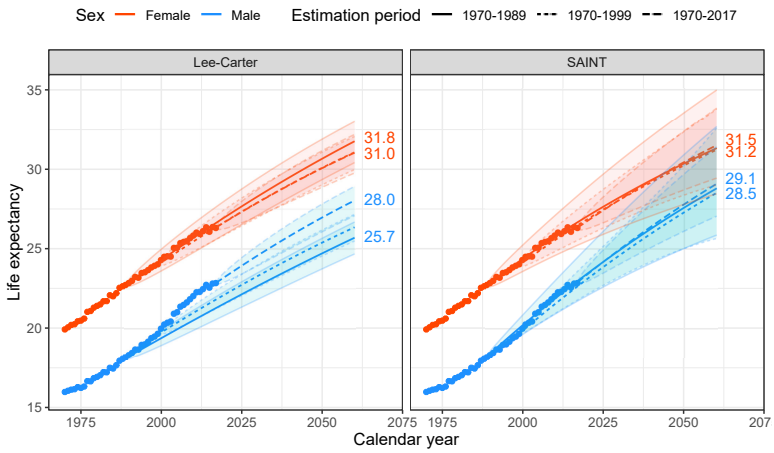

Figure 2.7: The panels show actual (dots) and forecasted (lines) remaining life expectancies for age 60 females and males with pointwise 95% confidence bands based on expanding estimation windows. Even though a rolling fixed-length data window is a more common back test approach in the literature, an expanding data window corresponds to how a mortality model is typically updated in practice.

Table 2.1: Empirical mean (and standard deviation in parentheses) of projected period life expectancies for a 60-year-old based on 10,000 simulations for select years.

Model	Female			Male		
	2020	2040	2060	2020	2040	2060
Calibration period: 1970–1989						
Lee-Carter	27.01 (0.54)	29.50 (0.62)	31.74 (0.65)	21.63 (0.38)	23.73 (0.46)	25.71 (0.51)
SAINT	27.06 (0.81)	29.38 (1.15)	31.40 (1.47)	23.46 (0.76)	26.35 (1.14)	28.88 (1.51)
Calibration period: 1970–1999						
Lee-Carter	26.73 (0.41)	29.04 (0.49)	31.04 (0.52)	22.18 (0.31)	24.39 (0.40)	26.36 (0.43)
SAINT	26.86 (0.53)	29.22 (0.78)	31.26 (1.07)	23.12 (0.53)	26.04 (0.90)	28.65 (1.49)
Calibration period: 1970–2017						
Lee-Carter	26.72 (0.22)	29.02 (0.52)	31.01 (0.61)	23.28 (0.16)	25.83 (0.40)	28.04 (0.47)
SAINT	26.75 (0.25)	29.31 (0.71)	31.51 (1.00)	23.39 (0.25)	26.44 (0.71)	29.09 (1.04)

Lee-Carter methodology to have a tendency of underestimating the actual gains in old-age mortality. The purpose of the comparison is not to show superiority of SAINT over other models, but to illustrate the beneficial effects of cointegration and frailty.

Forecasts are compared by considering period life expectancies. While cohort life expectancies taking future improvements of mortality into account are generally of more interest, the period life expectancy summarizes the level of mortality at a given time and is better suited for illustrative purposes as it can be compared with the actual experience. Recently, Arnold et al. (2019) added perspective on the stability of period versus cohort tables, arguing that the former might be preferable for practitioners looking to minimize capital adjustments following life tables updates.

Historical and forecasted life expectancies of a 60-year-old are shown in Figure 2.7.

The figure shows that both models produce highly similar median forecasts for female life expectancy and that predictions are quite robust to the choice of the data window used to calibrate the model. This similarity can be attributed to the stable rates of improvements observed since the 1970s. For the males, however, the improvement rates considered during the period of estimation are considerably lower than the actual rates in the periods that follow. The Lee-Carter model fails to capture this development, and the median forecasts are a good way from agreeing with the actual experience. Since improvement rates are increasing over time, predictions from the Lee-Carter model grow increasingly optimistic as we include more recent data.

The SAINT model lends its strength in the stable female improvement rates by the coupling of the genders described in Section 2.5.2. This leads to the forecasts being decidedly more robust to the choice of estimation period, and the resulting almost linear median life expectancy projections resemble the actual experience better than the scattered Lee-Carter forecasts. In fact, it is quite remarkable how well even the first SAINT forecast based on 1970-1989 predicts present day male life expectancy.

To contrast the forecasting uncertainty of the two methods, Table 2.1 reports summary statistics for the projected life time distributions based on 10,000 simulations. Both the table and the figure reveal that there is a major difference in the forecasting uncertainty between the two methods, even when point estimates are similar. It is evident that the uncertainty is greatest in the SAINT model, while it is, at least with hindsight, worryingly low for the Lee-Carter model. Specifically for the males using the two earliest calibrations, we get no warning that the projections might be well off target; the models do not capture the final years of observed data, let alone the 1970–2017 model's median projection, within their 95%-confidence bands.

Figure 2.8 compares the average improvement rates over a long horizon for select high ages. For the females, improvement rates are similar in both models over the short term; the height of the light-red bars align with the dashed lines. On longer horizons, however, the SAINT model gives rise to increasing rates of improvements endogenously within the model as the frailty composition changes over time. The Lee-Carter model and many of its descendants cannot predict rates of improvements higher than those observed historically and are therefore likely to be understating future increases in old-age mortality, even for large and stable datasets. In fact, the likeness between forecasts seen in Figure 2.7 is deceptive; the cohort life expectancy for a newborn female in 2017 is 94.81 years in the SAINT model, but nearly 2 years less in the Lee-Carter prediction at 92.96 years. For the males, the overall level of the improvement rates is higher in the SAINT forecast as a result of the genders being tied together compared to the Lee-Carter predictions where this is not so.

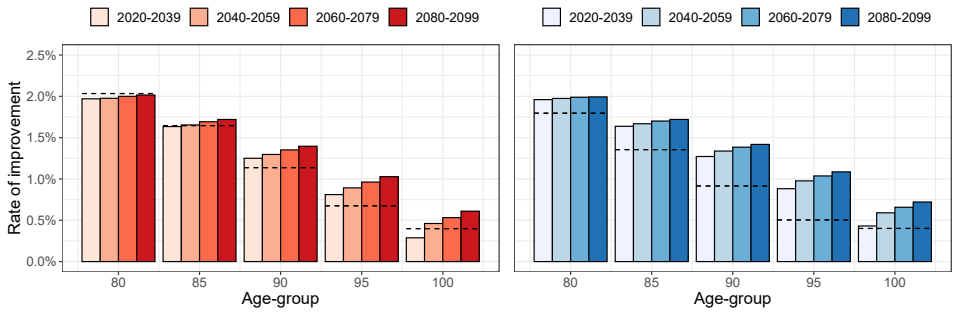

Figure 2.8: The panels show predicted average rates of improvement for select ages using the SAINT model estimated on the full data period 1970-2017. Female rates are shown in the left panel and male rates in the right. The corresponding improvement rates for the Lee-Carter model are superimposed as the dashed lines.

2.5.4 Spread Modelling

Given an underlying model for reference population mortality, we model mortality in the target population as deviations from the trend. As the name suggests, the trend should capture the main features of the mortality evolution and the spread should therefore be a model flexible enough to adequately fit observed mortality in the target population but introduce as little complexity as possible.

Since Plat (2009), common practice is to take a regression approach using a linear structure

$$D_{\text{target}}(t,x) \sim \text{Poisson}(\mu_{\text{target}}(t,x)E_{\text{target}}(t,x)), \tag{2.5.9}$$

$$\mu_{\text{target}}(t,x) = \mu_{\text{ref}}(t,x)\exp\left(y_t^\top r_x\right), \tag{2.5.10}$$

with parameters $y_t = (y_{1,t},\dots,y_{p,t})^\top \in \mathbb{R}^p$ describing the evolution of the spread over time and age-dependent regressors $r_x = (r_{1,x},\dots,r_{p,x})^\top \in \mathbb{R}^p$. Any standard GLM-routine can be used to fit the model by specifying a Poisson family with a log-link.

The original version of SAINT used a parsimonious model for the spread describing just its level, slope and curvature. While this model performed respectably, it became clear from a practical point of view that the fit of target population mortality was simply unconvincing when plotted on log-scale against empirical data, a plot frequently reported to the FSA and the Board of Representatives. To improve the fit, the three regressor model was replaced with a five regressor model

$$r_{i,x} = \min(1, \max(0, x_i - x)/20), \quad i \in \{1,\dots,5\}, \tag{2.5.11}$$

where $(x_1,x_2,x_3,x_4,x_5) = (40,60,80,100,120)$, so that regressors are one until a given breakpoint after which they decrease linearly to zero over the course of 20 years. Even though a model with evenly spaced knots has obvious practical advantages,

there are now methods available to explore the optimal choice of the number and location of the knots in spline models, see Kaishev et al. (2016). The chief reason for choosing the regressors above is that r_2, r_3, and r_4 are equivalent to the three regressor model specified by the Danish FSA for their longevity benchmark, cf. Jarner and Møller (2015), giving credence to the credibility of (2.5.11) in the eyes of the regulator.

Spread Forecast

Since it is the trend that governs long-term mortality behaviour, the time dynamics used for forecasting should ensure that the spread remains bounded in probability. We use a (stable) vector autoregressive model,

$$y_t = Ay_{t-1} + \upsilon_t, \tag{2.5.12}$$

with A being a $p \times p$ matrix and υ_t mean-zero Gaussian errors. We note that even if data appears stationary, e.g. Figure 2.9, maximum likelihood estimation of the VAR-model (2.5.12) does not necessarily yield stationarity and more sophisticated models, e.g. including additional lags, may have to be used. Alternatively, stationarity can be imposed by putting a curb on the eigenvalues of A such that they lie within the unit circle.

In practice, restricting A to be a diagonal matrix with diagonal elements $0 \leq A_{ii} < 1$, works well in terms of stability and is easy to communicate to non-specialists as deviation half-life. For instance, the Danish mortality curve moves more or less in unison with the international trend and there are no signs of an impending catch-up. Simply imposing $A = 0.99I$, with I being the identity matrix of size p, corresponds to a deviation half-life of about 69 years and results in the forecast depicted in Figure 2.9.

The figure epitomizes the advantage of the SAINT projection methodology. In a typical country-specific projection, exemplified by the Lee-Carter forecast, the recent stagnation in Danish life expectancy makes the data window a critical component of the analysis. The Lee-Carter model extrapolates past trends and even if based on very recent data periods, it is likely to understate future improvements in especially old-age mortality. Using the SAINT methodology the long-term trend is determined by a large pooled set of countries which serves as a stable reference for small populations that exhibit substantial variability. Rather than basing long-term projections on irregular national rates of improvements, we can frame the development as short- to medium term deviations from the trend. Having a point of reference also makes it possible – even for non-specialists – to visually gauge if the projection is reasonable.

Although only the Danish data are modelled here, the SAINT framework is easily adapted to other populations of interest. All that is required is a separate VAR model for the spread between the (new) target population and the trend. The assumed

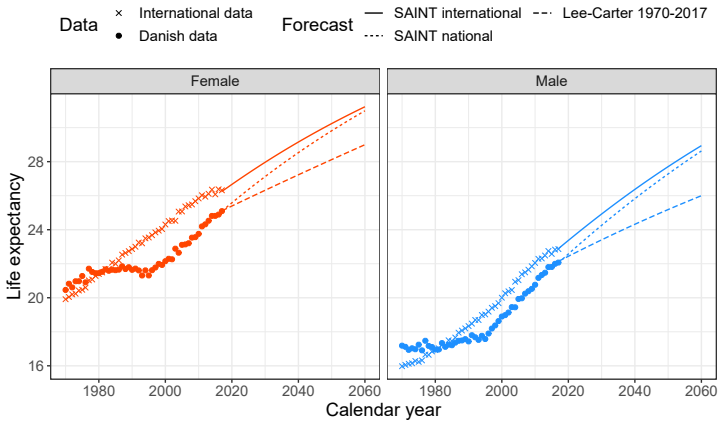

Figure 2.9: Actual and forecasted remaining period life expectancy for a 60-year-old.

stationarity of the spread will ensure that deviations from the trend are bounded. Multi-population analyses will therefore not exaggerate short-term differences or lead to diverging projections. All in all the SAINT methodology opens for a unified, flexible and robust forecasting approach, which is applicable to a wide range of populations despite their possibly unsteady development.

2.6 Concluding Remarks

A large and growing body of literature focuses on theoretical properties of mortality models and achieving accurate mortality forecasts. However, there are a number of desirable model properties besides theoretical ones that determine the success and applicability of a mortality model in practice. In this paper, we reviewed the lessons learned from more than a decade of experience with the SAINT model in use at ATP and the model modifications that have followed.

The main improvement over the original SAINT model was the generalization of frailty models from deterministic structures to a flexible class of stochastic mortality models, offering a general way of combining essentially any mortality model with frailty. This sets the work apart from the typical use of frailty which rely on matching parametric forms with closed-form expressions. We demonstrated how frailty extended mortality models dramatically improve the historical fit of even simple age-period structures and provide more realistic projections of improvement patterns on longer horizons. Obviously, frailty on its own is not enough to explain the complex dynamics of mortality, but helps capture essential features observed in the data otherwise addressed by ad hoc methods.

Although the original SAINT model was explicitly designed with stability in mind, we underestimated the effect that annual data updates could have on best

estimate projections. This lack of robustness was primarily rooted in the model's time dynamics whose parameters were estimated freely. In the trend, the long-run equilibrium relation between male and female mortality was overly sensitive to small changes in observed patterns, while the autoregressive model used to forecast the spread between reference and target population mortality did not guarantee stationarity. The issues were resolved by imposing sensible structural restrictions on the time dynamics of the model.

Even with well-behaved time dynamics, model stability is still challenged by outliers. In light of the COVID-19 crisis, it may not come as a surprise that a reference population – even when built from a large and geographically dispersed group of countries – will not guarantee stability under annual data updates. Nevertheless, we were taken by surprise by how deeply events like severe flu seasons impact reference mortality levels. Because of our globally connected world, infectious disease outbreaks cannot be diversified away by simply adding more countries to the data pool. We expect these types of fluctuations to be temporary rather than part of a lasting trend and measures to dampen their effect are indispensable in practical applications. This is an important issue for future research.

While accuracy, stability and flexibility requirements all pertain to model performance, there are certain properties in practical applications that do not. Governance rules entail an obligation on the part of the modeller to report, explain and justify outputs, some of which might not even have any direct practical implication, for example, in a pension fund context, how fitted death rates among the very young compare to actual rates. A poor fit in this age group will not affect the calculation of technical provisions in any way but may seriously detract from the model's credibility in the eyes of non-specialists. As modellers, we should be aware of this, partly external, desire for model explainability.

Acknowledgements

The authors are indebted to Esben Masotti Kryger for numerous stimulating discussions. The authors wish to thank two anonymous referees for their valuable input which helped improve the manuscript. The work was partly funded by Innovation Fund Denmark (IFD) under File No. 9065-00135B.

2.A Positive Stable Frailty

Hougaard (1986) introduced a family of generalized stable laws which includes the two most often used frailty distributions for mortality modelling, namely the Gamma and inverse Gaussian distributions, see e.g. Vaupel et al. (1979), Hougaard (1984), Butt and Haberman (2004), Jarner and Kryger (2011), and Spreeuw et al. (2013). The family is obtained by exponential tilting of stable densities with index $\alpha \in [0, 1)$.

The stable laws themselves only have moments of order strictly less than α, while moments of all orders exist for the exponentially tilted densities. From the original three-parameter family we obtain a two-parameter family by imposing the condition that mean frailty is one.

When Z follows a generalized stable law with index $\alpha \in [0,1)$, mean one and variance σ^2 the Laplace transform and mean frailty are given by

$$\mathcal{L}(s) = \exp\left(\frac{1-\alpha}{\alpha}\left[\frac{1-\left(1+\frac{\sigma^2 s}{1-\alpha}\right)^{\alpha}}{\sigma^2}\right]\right), \tag{2.A.1}$$

$$\mathbb{E}[Z|t,x] = \left(1+\frac{\sigma^2}{1-\alpha}\mathcal{M}_0(t,x)\right)^{\alpha-1} = \left(1+\frac{\alpha}{1-\alpha}\sigma^2\mathcal{M}(t,x)\right)^{\frac{\alpha-1}{\alpha}}. \tag{2.A.2}$$

The stated formulas are obtained from Hougaard (1986) using the parametrization $\theta = (1-\alpha)/\sigma^2$ and $\delta = [(1-\alpha)/\sigma^2]^{1-\alpha}$.

The generalized stable law specializes to the inverse Gaussian distribution for $\alpha = 1/2$ and to the Gamma distribution for the limiting case $\alpha = 0$ defined by continuity. While both Gamma and inverse Gaussian densities exist in closed form, generally (tilted) stable densities exist only as a series representation, cf. Hougaard (1984). Since the closed form expressions for the Laplace transform and mean frailty (2.A.1)-(2.A.2) are all we need for estimation and forecasting purposes, this is not problematic.

Arguably, the Gamma distribution is the most widely used frailty distribution. It is well-known that Gamma frailty in combination with Gompertz or Makeham baseline intensities lead to a cohort rate of the logistic type, and has been found to describe old-age mortality very well, see e.g. Thatcher (1999); Cairns et al. (2006). Gamma distributed frailty is also mathematically tractable and allows explicit calculations of many quantities of interest, e.g. frailty among survivors at a given age is Gamma distributed with known scale and shape parameters, cf. Vaupel et al. (1979).

It is generally found that Gamma frailty and the associated logistic form provides a better description of old-age mortality than inverse Gaussian frailty, see e.g. Butt and Haberman (2004) and Spreeuw et al. (2013). Furthermore Abbring and Berg (2007) show that for a large class of initial frailty distributions the frailty distribution among survivors converges to a Gamma distribution as the cumulated rate tends to infinity. Thus overall the Gamma distribution is a good default choice.

2.B Estimation of a Competing Risks Model

We consider the competing risk model of (2.4.9)–(2.4.10), i.e.

$$D(t,x) \sim \text{Poisson}\left(\left[\mu_s(\psi,\theta_t,\eta_x,\zeta,\xi) + \mu_b(\zeta_t,\xi_x)\right]E(t,x)\right), \tag{2.B.1}$$

where we have made μ_s's dependence on the vector of background parameters explicit. Selective and background mortality are given by

$$\mu_s(\psi, \theta_t, \eta_x, \zeta, \xi) = \nu'_\psi(\nu_\psi^{-1}\{\widetilde{\mathcal{M}}(t, x, \zeta, \xi\})F(\theta_t, \eta_x), \tag{2.B.2}$$

$$\mu_b(\zeta_t, \xi_x) = G(\zeta_t, \xi_x), \tag{2.B.3}$$

for a fixed value of ψ. Imagine that deaths were recorded according to the (hidden) sources

$$D_s(t, x) \sim \text{Poisson}\left(E(t, x)\mu_s(\psi, \theta_t, \eta_x, \zeta, \xi)\right), \tag{2.B.4}$$

$$D_b(t, x) \sim \text{Poisson}\left(E(t, x)\mu_b(\zeta_t, \xi_x)\right), \tag{2.B.5}$$

with D_s and D_b mutually exclusive so that $D(t, x) = D_s(t, x) + D_b(t, x)$. Even though D_s and D_b do not necessarily exist and hence are not "missing" in the usual sense of the word, we can still use the EM-algorithm based on the missing data interpretation of the model. Note that ζ and ξ are estimated from (2.B.5), while θ and η are estimated from (2.B.4) with ζ and ξ kept fixed at their current value.

Omitting time and age indices for ease of notation the expectation step is then to compute the complete data log-likelihood given death counts D and current parameter estimates from iteration $i - 1$,

$$Q(\theta, \eta, \zeta, \xi) = \mathbb{E}[\ell(\theta, \eta, \zeta, \xi) \mid D, \theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1}], \tag{2.B.6}$$

where a death is distributed according to a Bernoulli trial with a success parameter depending on the weight of the cause-specific death rate, i.e.

$$D_s|D, \theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1} \sim \text{Bin}\left(D, \frac{\mu_s(\theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1})}{\mu_s(\theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1}) + \mu_b(\zeta^{i-1}, \xi^{i-1})}\right), \tag{2.B.7}$$

$$D_b|D, \theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1} \sim \text{Bin}\left(D, \frac{\mu_b(\zeta^{i-1}, \xi^{i-1})}{\mu_s(\theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1}) + \mu_b(\zeta^{i-1}, \xi^{i-1})}\right), \tag{2.B.8}$$

under the assumption that the likelihood factorizes so that

$$\ell(\theta, \eta, \zeta, \xi) = \ell_s(\theta, \eta \mid D_s, \zeta^{i-1}, \xi^{i-1}) + \ell_b(\zeta, \xi \mid D_b). \tag{2.B.9}$$

The maximization step consists of maximizing (2.B.6) to obtain the i'th parameter estimate, i.e. estimating the two marginal mortality models with death counts

$$D_s = \mathbb{E}\left[D_s \mid D, \theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1}\right], \tag{2.B.10}$$

$$D_b = \mathbb{E}\left[D_b \mid D, \theta^{i-1}, \eta^{i-1}, \zeta^{i-1}, \xi^{i-1}\right]. \tag{2.B.11}$$

The E-step and M-step are iterated until converge.

In summary, the model (2.4.9)–(2.4.10) is estimated by optimizing the profile log-likelihood function (2.4.14), which for fixed frailty parameter ψ is computed by the following algorithm.

1. Initialize $\theta^0, \eta^0, \zeta^0, \xi^0$ and set $i = 1$.

2. For all t and x in the data window compute $\widetilde{\mathcal{M}}(t, x, \zeta^{i-1}, \xi^{i-1})$ and set

$$c(t,x) = \nu'_\psi(\nu_\psi^{-1}\{\widetilde{\mathcal{M}}(t, x, \zeta^{i-1}, \xi^{i-1})\}). \tag{2.B.12}$$

3. Compute θ^i and η^i as the maximum likelihood estimates of the model

$$D_s(t,x) \sim \text{Poisson}\left(c(t,x)E(t,x)F(\theta_t, \eta_x)\right) \tag{2.B.13}$$

with death counts

$$D_s(t,x) = D(t,x)\frac{c(t,x)F(\theta_t^{i-1}, \eta_x^{i-1})}{c(t,x)F(\theta_t^{i-1}, \eta_x^{i-1}) + G(\zeta_t^{i-1}, \xi_x^{i-1})}. \tag{2.B.14}$$

4. Compute ζ^i and ξ^i as the maximum likelihood estimates of the model

$$D_b(t,x) \sim \text{Poisson}\left(E(t,x)G(\zeta_t, \xi_x)\right) \tag{2.B.15}$$

with death counts

$$D_b(t,x) = D(t,x)\frac{G(\zeta_t^{i-1}, \xi_x^{i-1})}{c(t,x)F(\theta_t^{i-1}, \eta_x^{i-1}) + G(\zeta_t^{i-1}, \xi_x^{i-1})}. \tag{2.B.16}$$

5. Increase i by one.

6. Repeat steps 2–5 until convergence.

Chapter 3

Pitfalls and Merits of Cointegration-Based Mortality Models

This chapter contains the manuscript *Jarner and Jallbjørn (2020)*.

Abstract

In recent years, joint modelling of the mortality of related populations has received a surge of attention. Several of these models employ cointegration techniques to link underlying factors with the aim of producing coherent projections, i.e. projections with non-diverging mortality rates. Often, however, the factors being analysed are not fully identifiable and arbitrary identification constraints are (inadvertently) allowed to influence the analysis thereby compromising its validity. Taking the widely used Lee-Carter model as an example, we point out the limitations and pitfalls of cointegration analysis when applied to semi-identifiable factors. On the other hand, when properly applied cointegration theory offers a rigorous framework for identifying and testing long-run relations between populations. Although widely used as a model building block, cointegration as an inferential tool is often overlooked in mortality analysis. Our aim with this paper is to raise awareness of the inferential strength of cointegration and to identify the time series models and hypotheses most suitable for mortality analysis. The concluding application to UK mortality shows by example the insights that can be obtained from a full cointegration analysis.

JEL classification: C32, J11.
Keywords: *Mortality modelling, Lee-Carter model, CBD-model, cointegration, coherence, identification invariance.*

3.1 Introduction

Mortality models have many applications in areas such as demography, epidemiology, economics and actuarial sciences. In some applications we are interested in a single life expectancy projection, e.g. a unisex projection for a given country, but in many cases we are more interested in simultaneous projections for groups of related (sub)populations. Examples of the latter include joint modelling of males and females in a population, coherent forecasts for countries in a given region, projecting the life expectancy of smokers and non-smokers, modelling of insured lives relative to a national population, and assessing the effectiveness of a mortality hedge with the presence of basis risk; see e.g. Chen and Millossovich (2018), Kleinow (2015), Bergeron-Boucher et al. (2017), Janssen et al. (2013), Jarner and Kryger (2011), Cairns et al. (2011b), Dowd et al. (2011), and Cairns et al. (2011a).

Models applied independently to separate populations often lead to diverging forecasts. This is the case even for closely related populations. Tuljapurkar et al. (2000) found that applying the model of Lee and Carter (1992) separately to the G7 countries over a 50-year forecast horizon resulted in a life expectancy gap between the countries as large as eight years; despite the countries sharing long-term trends in mortality and convergence of social and economic factors. The projected divergence is due to small differences in the timing and magnitude of historical improvements being magnified by the separate analyses.

Joint mortality models are based on an assumption of non-divergence, or coherence, of mortality rates of the group of populations under consideration. Coherence is typically achieved by imposing a specific structure of the joint time series model used for forecasting factors driving mortality improvements in each population. Formally, the factors are assumed to cointegrate, i.e. to exhibit stationary relations preventing them from diverging. Cointegration theory offers a rigorous statistical framework for identifying and testing such stationary relations. However, this framework is rarely exploited in full since the structure is often imposed rather than tested, see e.g. Li and Lee (2005), Li and Hardy (2011), Dowd et al. (2011), and Cairns et al. (2011b).

The purpose of the present paper is twofold. First, we wish to advocate that cointegration analysis has more to offer than assuring coherence. Indeed, we will demonstrate the insights and "surprising" models that can arise from a full analysis. Second, we wish to highlight some of the pitfalls and limitations of cointegration analysis when applied to factors that are not fully identifiable, e.g. the mortality index of the popular Lee-Carter model. The overall message is that cointegration-based mortality models have much to offer as an inferential tool, but also that extreme care must be exercised when dealing with semi-identifiable factors often encountered.

3.1.1 Cointegration

Cointegration is rooted in econometrics. It was introduced by Engle and Granger (1987) as a methodology for testing for stationary relations between non-stationary time series. The basic idea is that if two variables share a (stochastic) trend, it might be possible to find a linear combination of the variables that cancels the trend resulting in a stationary process. The linear combination is referred to as a cointegrating relation. The Engle-Granger methodology is limited to only a single cointegrating relation, while the more general and comprehensive setup developed by Johansen (1995) allows for an arbitrary number of variables and cointegrating relations, at least in principle. In effect, each (linearly independent) cointegrating relation reduces the dimension of the "driving forces" of the system by one.

While the aim of an econometric analysis is to infer and interpret the (economic) system, the typical focus of a mortality analysis is to produce a plausible forecast with proper quantification of the uncertainty. With the aim of improving gender-specific mortality forecasts, Carter and Lee (1992) suggested cointegration as a possible tool. In recent years, there has been a proliferation of papers using cointegration techniques to obtain coherent forecasts, see e.g. Darkiewicz and Hoedemakers (2004), Lazar and Denuit (2009), Njenga and Sherris (2011), Gaille and Sherris (2011), Yang and Wang (2013), Hyndman et al. (2013), Zhou et al. (2014), Hunt and Blake (2015c), Salhi and Loisel (2017), and Li and Lu (2017). Many of the authors arrive at complex, high-dimensional models which are difficult to interpret, but potentially good at forecasting.

We find that, although cointegration can certainly be used as a tool to impose coherence, the real strength of cointegration lies in inference and interpretation. We believe that this aspect is largely absent in the actuarial literature and that important subject knowledge can be gained from a more statistical approach to mortality modelling, a point also made by Arnold and Sherris (2016). A number of the cited papers do in fact test for cointegration rank as part of their model selection, but formulating and testing hypotheses on parameters is not part of the analysis. The primary aim of this paper is to demonstrate the value of cointegration-based inference in a mortality context.

Cointegration theory is a technically sophisticated field and some preliminary work is needed to establish the type of cointegration models and hypotheses suitable for mortality modelling. Once established, we present a cointegration analysis of male and female UK mortality. We consider both a two-dimensional analysis based on the Lee-Carter model and a four-dimensional analysis based on the logistic two-factor model of Cairns et al. (2006). In principle, any number of factors can be analysed, but to aid interpretation and the formulation of hypotheses it is useful to consider only a moderate number of factors.

3.1.2 Identifiability

Many mortality models, including the Lee-Carter model and its many variants, are overparametrized and parameters are therefore only identifiable after adding one or more constraints. In the Lee-Carter model, for example, two constraints are needed to ensure identification of the time-varying index and the age-specific parameters. By definition, the fitted mortality rates are unaffected by the identification scheme, but the forecasted mortality rates are not necessarily unaffected. Forecasts are based on time series models for the time-varying parameters and these models might not be invariant to the identification scheme.

Several recent papers have addressed the issue of identifiability and forecasting in mortality models, see in particular Nielsen and Nielsen (2014), Kuang et al. (2008), Hunt and Blake (2015a,b), Hunt and Blake (2018), and Beutner et al. (2017). In summary, this body of work shows that for forecasts to be unaffected by the identification scheme the time series model should be flexible enough to "preserve" reparametrizations, i.e. forecasting and reparameterizing should be interchangeable operations. As Hunt and Blake (2018) point out, this seemingly innocent requirement can in fact be at odds with model structures imposed to achieve coherence. We will return to this point later in the paper.

Identifiability issues affect the interpretation of parameters and — unless properly addressed — might lead to conclusions resting entirely on arbitrary constraints. Non-trivial issues arise even in the standard setting of an age-period model of the Lee-Carter type, and the complexity of the issues increase rapidly with the number of time-varying indices, see Hunt and Blake (2015b). Further issues arise in joint models with cointegrating parameters, which is the focus of this paper. In this case, the semi-identifiability of the parameters severely limits the choice of meaningful time series models and hypotheses.

To guarantee identification invariant inference, Nielsen and Nielsen (2014) advocate an approach based on maximal invariants of reduced dimensionality in terms of which all estimation and forecasting must be formulated. Although theoretically elegant, researchers might be reluctant to adopt this idea, since the interpretation of the original parametrization is lost. Also, we fear that the approach adds to the impression held by many that identifiability concerns are an esoteric topic which overcomplicate simple problems. In contrast to Nielsen and Nielsen (2014), we do not develop any formal theory in this paper, but rather illustrate by examples some of the pitfalls and problems of semi-identifiability. Hopefully, our exposition will be both accessible and illuminating to a wide audience. We also prefer to retain the original parametrization of the models to make the examples as relevant and familiar as possible. In these respects, our work is similar in spirit to the analysis of the gravity model presented in Hunt and Blake (2018).

3.1.3 Outline

The rest of the paper is organized as follows. In Section 3.2 we introduce the mortality models we will use as examples throughout and we discuss identification issues in the familiar setup of the Lee-Carter model. Section 3.3 covers background information on cointegration theory, while Section 3.4 specializes the discussion of cointegration to mortality models and illustrates the problems with applying cointegration to semi-identifiable parameters. Section 3.5 contains a comprehensive analysis of UK mortality data applying cointegration techniques to both the semi-identifiable Lee-Carter model and the fully identified model of Cairns et al. (2006). Finally, Section 3.6 offers some concluding remarks.

3.2 Mortality Modelling

The object of study for mortality models is the age-specific death rates (ASDR's) in a given population. We assume that data consist of death counts, $D_{t,x}$, and corresponding exposures, $E_{t,x}$, over time-age cells of the form $[t, t+1) \times [x, x+1)$ for a range of calendar years t and (integer) ages x. We also assume that the underlying force of mortality, $\mu_{t,x}$, is constant over each of these cells. It then follows that $\mathbb{E}[D_{t,x}] = \mu_{t,x}\mathbb{E}[E_{t,x}]$, where $\mathbb{E}$ denotes the expectation operator.[1] The observed death rate is defined as the ratio $m_{t,x} = D_{t,x}/E_{t,x}$. This is commonly used as an (empirical) estimate of the underlying force of mortality. When considering more than one population we add an identifying superscript, e.g., $\mu^i_{t,x}$ denotes the force of mortality at time t and age x in population i.

Most mortality models capture the time evolution of mortality rates (the period effect) by one, or more, time-varying indices (factors), k_t. Below we introduce the one-factor model of Lee and Carter (1992) and the two-factor model of Cairns et al. (2006).

3.2.1 The Lee–Carter Model

The single population model proposed by Lee and Carter (1992) is used in a great number of mortality studies due to its simplicity and ease of interpretation; the ASDR's are modelled by a log-linear relation where an age-dependent a_x describes the general shape of the force of mortality and a single time-varying index, k_t, describes the speed of mortality improvements, governed by an age response b_x. The model is given by

$$\log m_{t,x} = \log \mu_{t,x} + \varepsilon_{t,x} = a_x + b_x k_t + \varepsilon_{t,x}, \tag{3.2.1}$$

[1] More precisely, $\mathbb{E}$ denotes the conditional expectation given $\mu_{t,x}$, since $\mu_{t,x}$ is itself a stochastic quantity. Thus, $\mathbb{E}$ averages over the random times of death given the force of mortality. Note that the exposure is also a stochastic quantity since it depends on the life spans. For modelling purposes, we (implicitly) condition on the exposures by treating them as fixed.

where $\varepsilon_{t,x}$ are homoscedastic error terms with mean 0 and variance σ_ε^2.

We note that the model is invariant under the parameter transformations

$$\{a_x\ ,\ b_x\ ,\ k_t\} \mapsto \{a_x - b_x c\ ,\ b_x/d\ ,\ d(k_t + c)\}\,, \tag{3.2.2}$$

where $c \in \mathbb{R}$ and $d \in \mathbb{R} \setminus \{0\}$, in the sense that these transformations all yield the same model for $\log m_{t,x}$. In other words, the parameters are not fully identifiable since k_t is only determined up to a linear transformation, b_x up to a multiplicative constant, and a_x up to a shift proportional to b_x. To ensure identification, the parameters are typically subject to the constraints

$$\sum_t k_t = 0, \quad \sum_x b_x = 1. \tag{3.2.3}$$

In Lee and Carter (1992) the parameters are estimated by ordinary least squares (OLS), i.e. by minimization of the quantity $\sum_{t,x}(\log m_{t,x} - a_x - b_x k_t)^2$. Under the constraints (3.2.3), $\hat{a}_x$ equals the time average of $\log m_{t,x}$, and $\hat{b}_x$ and $\hat{k}_t$ can be obtained from the first component of a singular value decomposition of the matrix $\{\log m_{t,x} - \hat{a}_x\}_{t,x}$. The parameter estimates thus obtained equal the maximum likelihood estimates under the additional assumption that the errors are normally distributed.[2]

The assumption of homoscedastic errors is questionable as we would expect observed death rates to fluctuate more when death counts are low. In addition, the use of OLS has the practical problem of how to handle cells with zero death counts frequent in small data sets. In the application to UK data we use instead the Poisson variant of the Lee-Carter model proposed by Brouhns et al. (2002). The two Lee-Carter variants have the same parametric structure and the points made later regarding identification issues in relation to forecasting and cointegration analysis therefore apply to both of them.

Forecasting

The time-varying index is typically modelled as a random walk with drift,

$$k_t = k_{t-1} + \theta + \varepsilon_t, \tag{3.2.4}$$

where θ is the drift and the ε_t's are i.i.d. $\mathcal{N}(0, \sigma^2)$. The drift and variance are estimated by the sample mean and sample variance, respectively, of the differences $\hat{k}_t - \hat{k}_{t-1}$.

Let T denote the last year of data. Forecasting is based on the conditional distribution of k_{T+h} given $k_T = \hat{k}_T$; we have for $h \geq 1$

$$k_{T+h} | k_T = \hat{k}_T \sim \mathcal{N}\left(\hat{k}_T + h\hat{\theta}\ ,\ h\hat{\sigma}^2\right), \tag{3.2.5}$$

[2]Lee and Carter (1992) contains a second stage adjustment of the parameters to reproduce the actual number of deaths each year. We do not consider this adjustment here.

from which forecasts and confidence intervals can be derived. In particular, a (median) forecast of future mortality rates is given by

$$\hat{\mu}_{T+h,x} = \exp\left(\hat{a}_x + \hat{b}_x\left[\hat{k}_T + h\hat{\theta}\right]\right) = \hat{\mu}_{T,x}\exp\left(h\hat{\theta}\hat{b}_x\right). \tag{3.2.6}$$

Coherence

The concept of coherent forecasts was introduced by Li and Lee (2005) and formalized by Hyndman et al. (2013). Mortality forecasts for two populations are said to be coherent if the relative mortality rates converge for each age x,

$$\frac{\hat{\mu}^1_{t,x}}{\hat{\mu}^2_{t,x}} \to R_x, \qquad t \to \infty, \tag{3.2.7}$$

for positive, age-specific constants R_x. When producing forecasts for populations with historically similar mortality evolutions, the concept of coherence formalizes the intuitively desirable property that the forecasts should reflect these similarities.

On the other hand, coherence is a rather strict requirement which will generally not be satisfied by forecasts obtained by applying e.g. the Lee-Carter model to separate populations. Indeed, it follows from (3.2.6) that in a Lee-Carter setting a necessary and sufficient condition for coherence is that $\hat{\theta}^1 = \hat{\theta}^2$ and $\hat{b}^1_x = \hat{b}^2_x$ for all x. In practice, of course, this will never happen. With the aim of obtaining coherent forecasts for a group of populations, Li and Lee (2005) proposed the *augmented common factor* model

$$\log m^i_{t,x} = \log \mu^i_{t,x} + \varepsilon^i_{t,x} = a^i_x + B_x K_t + b^i_x k^i_t + \varepsilon^i_{t,x}. \tag{3.2.8}$$

This model produces coherent forecasts when the population specific indices, k^i_t, are modelled as stationary processes, e.g. AR(1)-processes. The common factor, K_t, can be non-stationary, e.g. a random walk with drift as in the original Lee-Carter model.

The notion of coherence has undoubtedly been very influential in setting the standard for joint forecasts. Indeed, many joint mortality models have been devised with the specific aim of achieving coherence, as mentioned in the introduction. The model of Li and Lee (2005) can be seen as an early and very direct way to ensure coherence by equating the driving factors, while the more recent approaches typically combine a specific structure with cointegrating relations, e.g. Dowd et al. (2011). Generally, cointegrating relations do not guarantee coherence, although in the Lee-Carter setting the two concepts are closely linked. We will return to this point in Section 3.4.3.

It can be argued that coherence is too strict a requirement and that models enforcing coherence risk violating the historic pattern of covariation between populations, see Hunt and Blake (2018). Arguably, it is better to identify cointegrating relations which restrict the joint forecasts in plausible ways, than to insist on coherence. This point will be illustrated in the application section.

3.2.2 The Cairns-Blake-Dowd model

Originally made to accommodate the British pension market, the model of Cairns et al. (2006) focuses primarily on the post-age 60 mortality curve and the pricing of immediate life annuities. The model fits the mortality curve by a logistic curve[3]

$$\operatorname{logit}(q_{t,x}) \stackrel{\Delta}{=} \log\left(\frac{q_{t,x}}{1-q_{t,x}}\right) = k_{t,1} + k_{t,2}(x-\bar{x}), \tag{3.2.9}$$

where $\bar{x} = \frac{1}{N}\sum_i x_i$ is the arithmetic mean of the N ages considered and $q_{t,x} = 1 - \exp(-\mu_{t,x})$ is the probability for an individual aged x at time t to die before $t+1$. Note that the logit transform is well-defined since q lies between 0 and 1. We also note that $q_{t,x} \approx \mu_{t,x}$ for small $\mu_{t,x}$.

Under the CBD model the logit-transformed curve of death probabilities is linear in age with time-varying parameters. The first index is the *level* of the line, and a decreasing trend in $k_{t,1}$ thus represents an overall improvement in mortality over time. The second index is the *slope* of the line, and an increasing trend in $k_{t,2}$ thus implies a steepening of the mortality curve. The model is fully identified, since there are no invariant parameter transformations.

As is customary, we treat the model as a generalized linear model (GLM) within the binomial family with its canonical logit-link function. In principle, parameters can also be estimated by maximum likelihood assuming Poisson distributed death counts, see Currie (2016) for a comparison of the two approaches. The points made in this paper apply regardless of how parameters are estimated.

Forecasting

Forecasting is performed assuming a bivariate random walk with drift for the two time-varying indices

$$\begin{pmatrix} k_{t,1} \\ k_{t,2} \end{pmatrix} = \begin{pmatrix} k_{t-1,1} \\ k_{t-1,2} \end{pmatrix} + \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.2.10}$$

This projection method entails a dependency structure between the two time-varying indices by allowing for covariation, but omits the possibility of the indices being directly affected by the previous value of one another. As for the Lee-Carter model, the argument for the use of a random walk with drift is that it is often adequate to describe the data. Of course, if deemed necessary, more complicated ARIMA models can be used.

In the application section we shall model female and male mortality by two CBD models with cointegrating parameters. This will give rise to a sort of "logit coherence" rather than the usual (log) coherence.

[3]The use of a logistic model as a suitable choice for examining age patterns of human adult mortality is well established, see e.g. Thatcher (1999).

3.2.3 Identification Invariance

Identification issues arise in many fields of statistics, and mortality modelling is no exception. The problem arises from the fact that many (mortality) models are overparametrized such that there exist different sets of parameters yielding the same fit. From a statistical point of view all these sets are equally good, but to perform the analysis in practice the researcher has to choose one specific set. Therefore, constraints are imposed identifying one of the equivalent parameter sets over the others. The question now arises whether the (arbitrary) choice of constraints influence the forecast, and more generally the statistical inference.

To illustrate the problem, consider the familiar case of the Lee-Carter model as introduced above. Let $(\hat{a}_x, \hat{b}_x, \hat{k}_t)$ denote parameter estimates under the constraints (3.2.3), and consider the equivalent parameters $(\tilde{a}_x, \tilde{b}_x, \tilde{k}_t) = (\hat{a}_x - \hat{b}_x c, \hat{b}_x/d, d[\hat{k}_t + c])$ for given $c \in \mathbb{R}$ and $d \in \mathbb{R} \setminus \{0\}$. We forecast the time-varying index by the random walk with drift of (3.2.4). Using the mean difference as estimator for the drift we have $\tilde{\theta} = d\hat{\theta}$, and thereby $\tilde{\theta}\tilde{b}_x = \hat{\theta}\hat{b}_x$ for all x. Since also $\tilde{\mu}_{T,x} = \hat{\mu}_{T,x}$, it follows from (3.2.6) that for $h \geq 1$

$$\tilde{\mu}_{T+h,x} = \tilde{\mu}_{T,x} \exp\left(h\tilde{\theta}\tilde{b}_x\right) = \hat{\mu}_{T,x} \exp\left(h\hat{\theta}\hat{b}_x\right) = \hat{\mu}_{T+h,x}. \tag{3.2.11}$$

This shows that the forecast obtained by the standard Lee-Carter method is in fact invariant to the chosen identification scheme.

However, consider now the case where the time-varying index is modelled as the random walk with drift of (3.2.4), but with a fixed drift term $\theta = \theta_0$. We then have

$$\tilde{\mu}_{T+h,x} = \tilde{\mu}_{T,x} \exp\left(h\theta_0\tilde{b}_x\right) = \hat{\mu}_{T,x} \exp\left(h\theta_0\hat{b}_x/d\right), \tag{3.2.12}$$

which only equals $\hat{\mu}_{T+h,x}$ when $\theta_0\hat{b}_x = 0$. Hence, the forecasts will only be the same if $\theta_0 = 0$ (or $\hat{b}_x = 0$ for all x). Mathematically, the problem is that forecasting and reparametrization are no longer interchangeable operations or, in other words, the restricted forecasting model has different meaning for different parametrizations. One might argue that this is a contrived example as one would never consider this model, but very similar problems arise in cointegrated models where the identification of meaningful hypotheses is much less obvious.

Despite the ease with which specific problems related to lack of identification can be identified, it is surprisingly hard to formulate and justify a general principle that models and inferential procedures must adhere to.[4] Indeed, suggested principles often sound a bit vague: Hunt and Blake (2018) use the term "well-defined" for models

[4] Part of the problem seems to be that some researchers consider the constraints as an intrinsic part of the model, and not merely as (arbitrary) mathematical constraints needed for identification. From this perspective, different constraints imply different models and therefore "naturally" lead to different forecasts (even though the models are statistically identical in terms of describing the observed data).

giving the same projected mortality rates for any set of identifiability constraints, and Nielsen and Nielsen (2014) talk about "avoiding arbitrariness resulting from the identification process". We propose to use the term "identification invariance" when reparametrization and inference (including forecasting) are interchangeable, cf. Figure 3.1. We consider identification invariance a fundamental property of a sound statistical analysis.

Schematically, identification invariance is similar to the classical notion of "parametrization invariance", see e.g. Lindsey (1996). However, where parametrization invariance requires inferential invariance to all one-to-one reparametrizations, identification invariance requires only invariance to parameter transformations induced by different identification constraints. In practice, this amounts to inferential invariance to a specific set of linear parameter transformations. Note that, since time-varying parameters are typically modelled by linear time series models, the inference is generally not invariant to non-linear parameter transformations, i.e. the inference is generally not fully parametrization invariant.

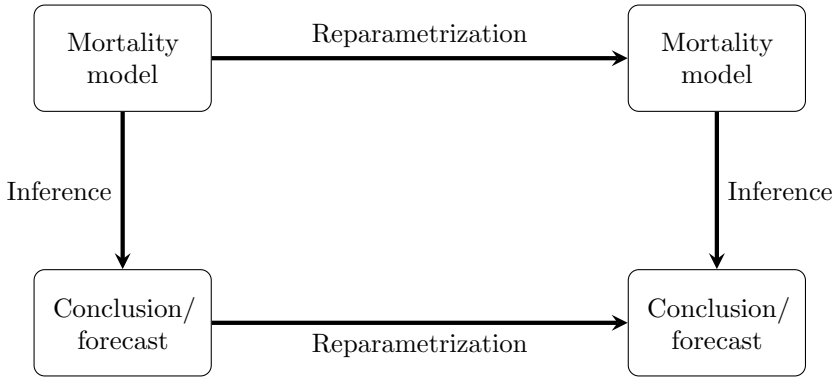

Figure 3.1: Illustration of identification invariance whereby the same inferential conclusion is reached for both the original and the reparameterized model. Here, *inference* refers to all aspects of the statistical analysis, e.g., parameter estimation, model selection, hypothesis testing, forecasting, prediction intervals etc.

3.3 Cointegration Theory

In the following we give a brief introduction to cointegrated vector autoregressive (VAR) models, including interpretation and testing of hypotheses; unless explicitly stated otherwise, the exposition relies on Johansen (1995). In the subsequent sections this framework will be applied to mortality modelling.

A p-dimensional VAR(k)-model is a model of the form

$$\mathbf{y}_t = \mathbf{\Pi}_1 \mathbf{y}_{t-1} + \ldots + \mathbf{\Pi}_k \mathbf{y}_{t-k} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{3.3.1}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ are i.i.d. errors and $\mathbf{D}_t$ contains all deterministic terms such as constant, trend and dummy variables. The evolution of each variable in the VAR-model is based on its own lagged values as well as the lagged values of the other variables in the system. This formulation highlights the short-term dynamics, while possible long-run relations between the variables are hard to discern. In order to study long-run relations we introduce the notion of cointegration. First a few preliminary definitions.

A *linear process* is defined by $\mathbf{y}_t = \sum_{i=0}^{\infty} \mathbf{C}_i \boldsymbol{\varepsilon}_{t-i}$, where $\boldsymbol{\varepsilon}_t$ are i.i.d. with mean zero and finite variance and $\mathbf{C}(z) = \sum_{i=0}^{\infty} \mathbf{C}_i z^i$ is convergent for $|z| < \delta$ for some $\delta > 1$. An $\mathsf{I}(0)$ process is a linear process with the additional requirement that $\sum_{i=0}^{\infty} \mathbf{C}_i \neq 0$, or such a process with a deterministic trend added.[5]

The difference operator, Δ, is defined by $\Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$. A stochastic process $\mathbf{y}_t$ is called integrated of order 1, or $\mathsf{I}(1)$, if $\Delta(\mathbf{y_t} - \mathbb{E}[\mathbf{y}_t])$ is an $\mathsf{I}(0)$ process. Loosely speaking, the stochastic component of an $\mathsf{I}(1)$ process behaves like a random walk.

Definition 3.3.1. *Let $\mathbf{y}_t$ be integrated of order 1. We say that $\mathbf{y}_t$ is cointegrated with cointegrating vector $\boldsymbol{\beta} \neq 0$ if $\boldsymbol{\beta}'\mathbf{y}_t - \mathbb{E}[\boldsymbol{\beta}'\mathbf{y}_t]$ admits a stationary distribution. The cointegrating rank is the number of linearly independent cointegrating vectors, and the cointegration space is the space spanned by the cointegrating vectors.*

3.3.1 The Vector Error Correction Model

We are interested in conditions for the VAR-model to be integrated of order 1. For this purpose we subtract $\mathbf{y}_{t-1}$ from both sides of (3.3.1) and rearrange terms to obtain the equivalent *vector error correction model* (VECM), where the increment is expressed in terms of differences, lagged differences and the level of the process itself

$$\Delta \mathbf{y}_t = \boldsymbol{\Pi} \mathbf{y}_{t-1} + \sum_{i=1}^{k-1} \boldsymbol{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{3.3.2a}$$

$$\boldsymbol{\Pi} = (\boldsymbol{\Pi}_1 + \ldots + \boldsymbol{\Pi}_k) - \mathbf{I}, \tag{3.3.2b}$$

$$\boldsymbol{\Gamma}_i = -(\boldsymbol{\Pi}_{i+1} + \ldots + \boldsymbol{\Pi}_k), \quad i = 1, \ldots, k-1. \tag{3.3.2c}$$

The behaviour of $\mathbf{y}_t$ is most easily studied in terms of its characteristic polynomial given by $\mathbf{A}(z) = (1-z)\mathbf{I} - \boldsymbol{\Pi} z - (1-z)\sum_{i=1}^{k-1} \boldsymbol{\Gamma}_i z^i$ with determinant $|\mathbf{A}(z)|$. If $\mathbf{A}$ has a unit root then $\boldsymbol{\Pi} = -\mathbf{A}(1)$ is singular and the process is non-stationary. In this case, $r = \operatorname{rank}(\boldsymbol{\Pi}) < p$ and there exist two $p \times r$ matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that

$$\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'. \tag{3.3.3}$$

This is essentially the requirement for $\mathsf{I}(1)$. However, to avoid explosive and seasonal roots and to ensure invertibility we also need the following technical condition.

[5] Sometimes a deterministic trend is allowed in the definition of a linear process. Here, however, we follow the terminology of Johansen (1995) whereby a linear process has mean zero.

Condition 3.3.2. *If* $|\mathbf{A}(z)| = 0$*, then either* $|z| > 1$ *or* $z = 1$*. Further, the matrix* $\boldsymbol{\alpha}_\perp' \boldsymbol{\Gamma} \boldsymbol{\beta}_\perp$ *has full rank, where* $\boldsymbol{\Gamma} = \mathbf{I} - \sum_{i=1}^{k-1} \boldsymbol{\Gamma}_i$ *and* $\boldsymbol{\alpha}_\perp$ *and* $\boldsymbol{\beta}_\perp$ *are* $p \times (p-r)$ *matrices spanning the orthogonal complement of span*$(\boldsymbol{\alpha})$ *and span*$(\boldsymbol{\beta})$*, respectively.*

None of the matrices $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_\perp$, $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_\perp$ are uniquely defined, but the conditions and conclusions do not depend on which versions we use. Let $\boldsymbol{\Gamma}(z) = \mathbf{I} - \sum_{i=1}^{k-1} \boldsymbol{\Gamma}_i z^i$ whereby $\boldsymbol{\Gamma}(1) = \boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}(\mathbf{L})\mathbf{y}_0 = \mathbf{y}_0 - \sum_{i=1}^{k-1} \boldsymbol{\Gamma}_i \mathbf{y}_{-i}$, where $\mathbf{L}$ is the lag operator. We can now formulate the celebrated Granger Representation Theorem in a version due to Hansen (2005).

Theorem 3.3.3. *If* $|\mathbf{A}(1)| = 0$ *and Condition 3.3.2 is satisfied, then* $\mathbf{y}_t$ *can be represented as the sum of a random walk and a stationary process*

$$\mathbf{y}_t = \mathbf{C} \sum_{i=1}^{t} (\boldsymbol{\varepsilon}_i + \boldsymbol{\Phi} \mathbf{D}_i) + \sum_{i=0}^{\infty} \mathbf{C}_i^* (\boldsymbol{\varepsilon}_{t-i} + \boldsymbol{\Phi} \mathbf{D}_{t-i}) + \mathbf{C}\boldsymbol{\Gamma}(\mathbf{L})\mathbf{y}_0, \tag{3.3.4}$$

where $\mathbf{C} = \boldsymbol{\beta}_\perp \left(\boldsymbol{\alpha}_\perp' \boldsymbol{\Gamma} \boldsymbol{\beta}_\perp\right)^{-1} \boldsymbol{\alpha}_\perp'$*, and* $\mathbf{C}_i^*$ *is defined recursively by*

$$\mathbf{C}_i^* = (\mathbf{I} + \boldsymbol{\Pi})\, \mathbf{C}_{i-1}^* + \sum_{j=1}^{k-1} \boldsymbol{\Gamma}_j \Delta \mathbf{C}_{i-j}^*, \quad i = 1, 2, \ldots,$$

with $\mathbf{C}_0^* = \mathbf{I} - \mathbf{C}$ *and* $\mathbf{C}_{-1}^* = \ldots = \mathbf{C}_{-k+1}^* = -\mathbf{C}$*. In particular, if* $r > 0$ *then* $\mathbf{y}_t$ *is a cointegrated* $\mathsf{I}(1)$ *process with cointegrating vectors* $\boldsymbol{\beta}$*.*

Intuitively, the process evolves as a random walk in $\operatorname{span}(\boldsymbol{\beta}_\perp)$ while at the same time it tries to establish the equilibrium relation for $\boldsymbol{\beta}'\mathbf{y}_t$ with a force that depends on the adjustment coefficients $\boldsymbol{\alpha}$ and the equilibrium error $\boldsymbol{\beta}'\mathbf{y}_t - \mathbb{E}\left[\boldsymbol{\beta}'\mathbf{y}_t\right]$.

The factorization (3.3.3) defines the cointegration space, but the individual cointegrating relations are not unique without further normalization. Johansen (1995) suggests letting the first part of $\boldsymbol{\beta}$ be an r-dimensional identity matrix making for a just-identified normalization and we adopt this approach throughout without further notification.

3.3.2 Deterministic Terms and Trends

The deterministic term is an important part of the specification of the model affecting both the trend of $\mathbf{y}_t$ and the test statistics for cointegration rank. Under the conditions of Theorem 3.3.3, the process $\mathbf{y}_t$ has, in general, a trend of the form $\mathbf{C}\boldsymbol{\Phi} \sum_{i=1}^{t} \mathbf{D}_i + \sum_{i=0}^{\infty} \mathbf{C}_i^* \boldsymbol{\Phi} \mathbf{D}_{t-i}$. Note that we refer to the deterministic part of $\mathbf{y}_t$ as a trend, regardless of its order.

In general, the deterministic terms accumulate to a trend one order higher. More precisely, however, it is only the terms $\mathbf{C}\boldsymbol{\Phi}\mathbf{D}_t$ that accumulate to a higher order. To

Table 3.1: Trending behaviour of the VECM for five nested models for the deterministic term of form (3.3.5). Starting with no restrictions, the models are defined by successively setting $\boldsymbol{\gamma}_1$, $\boldsymbol{\rho}_1$, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\rho}_0$ to zero.

Model	Deterministic term	Trend in $\mathbf{y}_t$	$\mathbb{E}[\Delta\mathbf{y}_t]$	$\mathbb{E}[\boldsymbol{\beta}'\mathbf{y}_t]$
$H_0(r)$	$\boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 t$	Quadratic	Linear	Linear
$H_0^*(r)$	$\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\boldsymbol{\rho}_1 t$	Linear	Constant	Linear
$H_1(r)$	$\boldsymbol{\theta}_0$	Linear	Constant	Constant
$H_1^*(r)$	$\boldsymbol{\alpha}\boldsymbol{\rho}_0$	Constant	Zero	Constant
$H_2(r)$	0	Zero	Zero	Zero

illustrate the implication of this in more detail, we consider deterministic terms of the form

$$\boldsymbol{\Phi}\mathbf{D}_t = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 t, \tag{3.3.5}$$

for p-dimensional vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Following Johansen (1995), we decompose each $\boldsymbol{\theta}_i$ as $\boldsymbol{\theta}_i = \boldsymbol{\alpha}_\perp\boldsymbol{\gamma}_i + \boldsymbol{\alpha}\boldsymbol{\rho}_i$. Since $\mathbf{C}\boldsymbol{\alpha} = 0$, it follows from the Granger representation (3.3.4) that only $\boldsymbol{\alpha}_\perp\boldsymbol{\gamma}_0 + \boldsymbol{\alpha}_\perp\boldsymbol{\gamma}_1 i$ enters into the i'th term of the random walk component. In particular, there is a quadratic trend in the level of process with coefficient $\frac{1}{2}\mathbf{C}\boldsymbol{\alpha}_\perp\boldsymbol{\gamma}_1$.

The decomposition of $\boldsymbol{\theta}_i$ gives rise to five nested models defined by restricting the number of non-zero terms. The models and the trending behaviour they entail are summarised in Table 3.1.

3.3.3 Cointegration Rank and Parameter Estimation

In some situations, the cointegration rank can be justified on the basis of prior knowledge. Often, however, the cointegration rank needs to be inferred from the data. Let $H(r)$ denote the hypothesis that $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ for two $p \times r$ matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Without further restrictions, this is equivalent to the hypothesis that $\text{rank}(\boldsymbol{\Pi}) \leq r$. This creates a set of nested hypotheses

$$H(0) \subset \cdots \subset H(r) \subset \cdots \subset H(p).$$

The cointegration rank can be determined by testing these hypotheses sequentially, starting from $H(0)$ and stopping when the first acceptance is encountered. The cointegration rank is r, say, if $H(0), \ldots, H(r-1)$ are rejected, while $H(r)$ is accepted. Johansen (1995) derives the likelihood-ratio test, known as the trace test, for performing these tests. The distribution of the test statistic is non-standard and it depends on the specification of the deterministic term. Critical values are tabulated in Section 15.3 of Johansen (1995) for the five models considered in Section 3.3.2. The testing procedure is complicated by the fact that we might need to infer the

model for the deterministic term and the cointegration rank simultaneously. We will return to this point in the application section.

Given the (maximal) cointegration rank and specification of the deterministic term, the maximum likelihood estimates of the parameters are obtained by reduced rank regression. For completeness, the estimates and the trace test statistic can be found in Appendix 3.A.

3.4 Cointegration in Mortality Models

In this section we first discuss the trend models most relevant in a mortality context. Next, we show how the identification issues of the Lee-Carter model severely limit the set of testable (cointegration) hypotheses. Finally, we comment on alternative approaches to cointegration within the Lee-Carter framework.

3.4.1 Linear Trend Models

We restrict our attention to the case of analysing the period effect within a given parametric mortality model. Assume that $\boldsymbol{k}_t$, consisting of the combined time-varying indices of (separate) age-period models, can be shown to form an I(1) process. This is a reasonable assumption, since we expect the period effect to cause at least one of the time-varying indices (for each population) to accumulate annual improvements over time and hence to behave like a random walk with drift. In general, the drift term itself could be time-varying giving rise to trends of all shapes and orders. Indeed, Arnold and Sherris (2016) find quadratic trends when analysing cause-specific mortality. However, for the purpose of this exposition we focus on linear trends only, which are, arguably, also the only type of trends suitable for robust forecasting.

To characterize the relevant models in more detail, let us consider deterministic terms of the form $\boldsymbol{\Phi}\mathbf{D}_t = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 t$, where $\boldsymbol{\theta}_i = \boldsymbol{\alpha}_\perp \boldsymbol{\gamma}_i + \boldsymbol{\alpha}\boldsymbol{\rho}_i$ for $i = 0, 1$, cf. Section 3.3.2. The absence of a quadratic trend implies that $\boldsymbol{\gamma}_1 = 0$. Hence, in the current context the largest model of interest is $H_0^*(r)$ of Table 3.1. Under this model, $\boldsymbol{k}_t$ has the representation

$$\boldsymbol{k}_t = \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 t + \mathbf{C}\sum_{i=1}^{t} \boldsymbol{\varepsilon}_i + \sum_{i=0}^{\infty} \mathbf{C}_i^* \boldsymbol{\varepsilon}_{t-i} + \mathbf{C}\boldsymbol{\Gamma}(\mathbf{L})\boldsymbol{k}_0; \tag{3.4.1}$$

expressions for the intercept, $\boldsymbol{\tau}_0$, and slope, $\boldsymbol{\tau}_1$, can be found in Hansen (2005). It can be shown that the cointegrating relations are *trend stationary*, i.e. they can be decomposed as a linear trend and a stationary process, $\boldsymbol{u}_t$, as $\boldsymbol{\beta}'\boldsymbol{k}_t = \boldsymbol{\beta}'\boldsymbol{\tau}_0 - \boldsymbol{\rho}_1 t + \boldsymbol{u}_t$. Thus parameters drift further and further apart over time even though they engage in an equilibrium correcting relationship.

The second model of interest for mortality modelling is $H_1(r)$ of Table 3.1. This model has $\boldsymbol{\rho}_1 = \boldsymbol{\gamma}_1 = 0$, i.e. the previous model with the further restriction that

the cointegrating relations do not trend. The level of the process still possesses a linear trend. Demographically, the lack of a trend in the cointegrating relations is appealing, but it cannot be assumed in advance. In the application section we will use the statistical setup of Johansen (1995) to test for $\boldsymbol{\rho}_1 = 0$.

Technically, it is also possible to test for further model restrictions, i.e. model $H_1^*(r)$ for absence of a linear trend altogether and model $H_2(r)$ for zero mean. However, neither of these latter models are relevant for modelling mortality data with period effects.

As described in Section 3.2, applications of the Lee-Carter model and the CBD-model often employ a simple random walk with drift to describe the time-varying indices. In the spirit of preserving as much of the marginal structure as possible, the natural candidate for joint modelling is therefore the VECM with zero lagged differences and no quadratic trend

$$\Delta \boldsymbol{k}_t = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\boldsymbol{k}_{t-1} + \boldsymbol{\rho}_1 t\right) + \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_t = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\boldsymbol{k}_{t-1} + \boldsymbol{\rho}_0 + \boldsymbol{\rho}_1 t\right) + \boldsymbol{\alpha}_\perp \boldsymbol{\gamma}_0 + \boldsymbol{\varepsilon}_t. \quad (3.4.2)$$

For this model, the linear trend of (3.4.1) is given by $\boldsymbol{\tau}_1 = \mathbf{C}\boldsymbol{\theta}_0 - \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\rho}_1$. Further, if $\boldsymbol{\rho}_1 = 0$ then $\boldsymbol{\tau}_1 = \mathbf{C}\boldsymbol{\theta}_0$ and $\mathbb{E}\left[\boldsymbol{\beta}'\boldsymbol{k}\right] = -\left(\boldsymbol{\beta}'\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\beta}'\boldsymbol{\theta}_0$, and we have

$$\Delta \boldsymbol{k}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{k}_{t-1} + \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_t = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\boldsymbol{k}_{t-1} - \mathbb{E}\left[\boldsymbol{\beta}'\boldsymbol{k}\right]\right) + \boldsymbol{\tau}_1 + \boldsymbol{\varepsilon}_t. \quad (3.4.3)$$

Models (3.4.2)–(3.4.3) will be our workhorse models in the application section.

3.4.2 Identification Invariance of Cointegrated Lee–Carter Models

The Lee-Carter model is the predominant single population mortality model. At first thought, it therefore seems natural to use the Lee-Carter model as the underlying model for a joint analysis of related populations. However, it turns out that the identifiability issues of the Lee-Carter model severely limit its usefulness for this purpose.

The lack of identifiability of the Lee-Carter model and the consequences for interpretation and forecasting have also been studied by other authors, see in particular Nielsen and Nielsen (2014) and Hunt and Blake (2018). In contrast to these contributions, we here focus on how the identifiability issues restrict the set of testable hypotheses, i.e. the hypotheses that can form part of a statistical analysis. More precisely, we are interested in characterizing the identification invariant cointegration models for the time-varying mortality index of two Lee-Carter models, i.e. the cointegration models for which forecasting and reparametrization of the underlying Lee-Carter models are interchangeable, cf. Section 3.2.3.

Let $\boldsymbol{k}_t = (k_t^1, k_t^2)'$ denote the vector of mortality indices of two Lee-Carter models with given identification schemes. Assume that we choose to model this as a VECM

with a linear trend of class $H_0^*(r)$,

$$\Delta \boldsymbol{k}_t = \sum_{i=1}^{l-1} \boldsymbol{\Gamma}_i \Delta \boldsymbol{k}_{t-i} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}' \boldsymbol{k}_{t-1} + \rho_1 t\right) + \boldsymbol{\theta_0} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.4.4}$$

Now, consider a reparametrization of the underlying Lee-Carter models or, equivalently, different identification schemes. According to (3.2.2) the vector of reparameterized mortality indices takes the form $\tilde{\boldsymbol{k}}_t = \mathbf{D}(\boldsymbol{k}_t + \boldsymbol{c})$, where $\mathbf{D} = \operatorname{diag}(d_1, d_2)$ with $d_1, d_2 \in \mathbb{R} \setminus \{0\}$, and $\boldsymbol{c} = (c_1, c_2)'$ with $c_1, c_2 \in \mathbb{R}$. Applying the same transformation to (3.4.4) yields

$$\Delta \tilde{\boldsymbol{k}}_t = \sum_{i=1}^{l-1} \mathbf{D}\boldsymbol{\Gamma}_i \mathbf{D}^{-1} \Delta \tilde{\boldsymbol{k}}_{t-i} + \mathbf{D}\boldsymbol{\alpha}\left(\boldsymbol{\beta}'\left(\mathbf{D}^{-1}\tilde{\boldsymbol{k}}_{t-1} - \boldsymbol{c}\right) + \rho_1 t\right) + \mathbf{D}\boldsymbol{\theta_0} + \mathbf{D}\boldsymbol{\varepsilon}_t \tag{3.4.5}$$

$$= \sum_{i=1}^{l-1} \tilde{\boldsymbol{\Gamma}}_i \Delta \tilde{\boldsymbol{k}}_{t-i} + \tilde{\boldsymbol{\alpha}}\left(\tilde{\boldsymbol{\beta}}' \tilde{\boldsymbol{k}}_{t-1} + \rho_1 t\right) + \tilde{\boldsymbol{\theta}}_0 + \tilde{\boldsymbol{\varepsilon}}_t, \tag{3.4.6}$$

where $\tilde{\boldsymbol{\Gamma}}_i = \mathbf{D}\boldsymbol{\Gamma}_i\mathbf{D}^{-1}$, $\tilde{\boldsymbol{\alpha}} = \mathbf{D}\boldsymbol{\alpha}$, $\tilde{\boldsymbol{\beta}} = \mathbf{D}^{-1}\boldsymbol{\beta}$, $\tilde{\boldsymbol{\theta}}_0 = \mathbf{D}\boldsymbol{\theta}_0 - \mathbf{D}\boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{c}$, and $\tilde{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D})$. Since $\mathbf{D}$ and $\mathbf{D}^{-1}$ have full rank, $\operatorname{rank}\left(\boldsymbol{\alpha}\boldsymbol{\beta}'\right) = \operatorname{rank}\left(\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}'\right)$, and it follows that (3.4.6) belongs to $H_0^*(r)$ for the same value of r as in (3.4.4). Thus, in the context of cointegrated Lee-Carter models, $H_0^*(r)$ is identification invariant, cf. Figure 3.1.

Similar calculations show that the first four model classes of Table 3.1 are all identification invariant, while the fifth, $H_2(r)$, is not (unless $r = 0$). In other words, constant and linear trends are preserved by linear transformations of the two indices being modelled—*as long as we allow all other parameters to vary freely.*

While it is generally valid to impose different models for the deterministic terms, e.g. $H_0^*(r)$ or $H_1(r)$, for freely varying parameters, it is generally *not* valid to test or impose further parameter constraints. Indeed, since $\mathbf{D}$ is a diagonal matrix with an arbitrary, non-zero diagonal and since $\tilde{\boldsymbol{\alpha}} = \mathbf{D}\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\beta}} = \mathbf{D}^{-1}\boldsymbol{\beta}$, constraints on either $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ will typically not be satisfied by the transformed model. In particular, the hypothesis of prime interest, $\boldsymbol{\beta} = (1, -1)'$, is not testable, i.e. the corresponding model is not identification invariant. Nor can we test hypotheses on the (relative) magnitude of the adjustment coefficients. In fact, apart from the cointegration rank, the only testable hypothesis of some demographic interest is $\rho_1 = \rho$ for given ρ. The ρ_1 parameter is identifiable and can be interpreted as a 'divergence' measure between (related) populations.

In summary, obeying identification invariance we can infer the cointegration rank and distinguish between the two types of linear trend models most relevant for mortality modelling (and also between other less relevant models). However, within a given model class we can in general not restrict parameters further without violating identification invariance, i.e. essentially all hypotheses of interest are non-testable.

Table 3.2: Overview of models for deterministic terms (drift) and hypotheses for parameter constraints. A check mark indicates that the model/hypothesis is identification invariant/testable in the context of a VECM for the mortality indices of two Lee-Carter models, and a minus sign that it is not.

Model	Drift	Invariant	Hypothesis	Testable
$H_0(r)$	$\boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 t$	✓	$\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\psi}$	÷
$H_0^*(r)$	$\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\rho_1 t$	✓	$\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\varphi}$	÷
$H_1(r)$	$\boldsymbol{\theta}_0$	✓	$\boldsymbol{\theta}_0 = \boldsymbol{\theta}$	÷
$H_1^*(r)$	$\boldsymbol{\alpha}\rho_0$	✓	$\rho_0 = \rho$	÷
$H_2(r)$	0	÷	$\rho_1 = \rho$	✓

The situation is summarized in Table 3.2. For someone interested purely in joint forecasting of two populations this might not pose a problem. However, if the aim is to analyse the nature of the joint behaviour in more detail, cointegrated Lee-Carter models are of limited use.

3.4.3 Alternative Approaches of the Lee–Carter Type

Arguably, the simplest way to avoid the problems due to lack of identifiability of the Lee-Carter model is to use another model. Indeed, if we base our (joint) analysis on a fully identifiable mortality model, e.g. the CBD model, all (joint) hypotheses are well-defined and testable. However, due to the familiarity and widespread use of the Lee-Carter model some researchers might be reluctant to follow this route. For that reason, we consider below two alternative approaches to obtain identification invariant inference within the Lee-Carter framework.

The first approach is to impose further restrictions on the parameters of the underlying Lee-Carter models, thereby implicitly restricting the set of identification invariant transformations of the mortality indices. As an example of this approach, Zhou et al. (2014) assume equality of the age response parameters of two Lee-Carter models, i.e. $b_x = b_x^1 = b_x^2$ for all x,

$$\log m_{t,x}^i = \log \mu_{t,x}^i + \varepsilon_{t,x}^i = a_x^i + b_x k_t^i + \varepsilon_{t,x}^i \quad \text{for } i = 1, 2. \tag{3.4.7}$$

This is similar in spirit to the augmented common factor model of Li and Lee (2005), cf. (3.2.8). Let $\boldsymbol{k}_t = (k_t^1, k_t^2)'$ denote the vector of mortality indices for given identification scheme. Due to the constraint on the b-parameters, different identification schemes lead to mortality indices of the form $\tilde{\boldsymbol{k}}_t = d(\boldsymbol{k}_t + \boldsymbol{c})$, with $d \in \mathbb{R} \setminus \{0\}$, and $\boldsymbol{c} = (c_1, c_2)'$ with c_1, $c_2 \in \mathbb{R}$. The point to note is that, in contrast to the situation in Section 3.4.2, the mortality indices are always scaled by the same constant.

If $\boldsymbol{k}_t$ is modelled by the VECM of (3.4.4) then $\tilde{\boldsymbol{k}}_t$ satisfies (3.4.6) with $\tilde{\boldsymbol{\alpha}} = d\boldsymbol{\alpha}$, $\tilde{\boldsymbol{\beta}} = d^{-1}\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\theta}}_0 = d\boldsymbol{\theta}_0 - d\boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{c}$. We notice that the transformed adjustment coefficients

and long-run relations are proportional to their previous values, hence hypotheses on the relative magnitude of these parameters are testable, while hypotheses on the "intercepts" $\boldsymbol{\theta}_0$ and ρ_0 are still not testable. In particular, $\boldsymbol{\beta} = (1, -1)'$, imposed—but not tested—by Zhou et al. (2014) in their analysis, is in fact a testable hypothesis.

In some sense the model of Zhou et al. (2014) "solves" the identification issue of separate Lee-Carter models by imposing just enough additional structure to allow the formulation of well-defined joint hypotheses of interest. However, it comes at the price of a very restrictive parameter structure. It is unlikely that identical age response parameters for two separate populations is a reasonable assumption, and in general the model must be expected to fit data rather poorly. Also, it is a somewhat indirect way to address the identification issue.[6]

In the generic mortality modelling setup considered so far one or more time-varying factors are first extracted from data and then forecasted by a time series model. It is tacitly assumed that the number of factors is low; the Lee-Carter model, for instance, uses a single factor to capture the mortality evolution over time.

An alternative approach is to consider the vector of log mortality rates as a multivariate time series (of high dimension) and model this series directly, i.e. to model directly the N-dimensional series $\mathbf{y}_t = (\log m_{t,x_1}, \ldots, \log m_{t,x_N})$. By construction, there are no factors and hence no identification issues related to factor identification, but due to the high dimension the time series models are more complex and harder to interpret.

Lazar and Denuit (2009) use the cointegration methodology of Section 3.2 to model $\mathbf{y}_t$. In this framework, the Lee-Carter model is a special case with cointegration rank $N-1$ corresponding to one common stochastic trend. Lazar and Denuit (2009) focus on single-population modelling, but the approach extends readily to multi-population modelling by stacking the $\mathbf{y}$-vectors. The VAR/VECM approach to modelling $\mathbf{y}_t$ is also explored in the recent papers by Salhi and Loisel (2017) and Li and Lu (2017).

As mentioned in Section 3.2.1, coherence has received much attention as a desirable property of mortality forecasts. In the Lee-Carter setting of Section 3.4.2 coherence corresponds to stationarity of $b_x^1 k_t^1 - b_x^2 k_t^2$ for all x. Note that, cointegrated mortality indices, i.e. stationarity of $\boldsymbol{\beta}' \boldsymbol{k}_t$ for *some* $\boldsymbol{\beta}$, does not in itself guarantee coherence. Indeed, for (strict) coherence we must have $\boldsymbol{b}_x \propto \boldsymbol{\beta}$ for all x, where $\boldsymbol{b}_x = (b_x^1, -b_x^2)'$. In practice, this will never be (strictly) satisfied, unless enforced by design as in Zhou et al. (2014). In contrast, when modelling $\mathbf{y}_t$ directly non-diverging rates for different populations (coherence) or for different ages within the same population

[6] In Zhou et al. (2014), the stated reason for assuming identical b-parameters is to obtain non-divergent mortality rates. It is unclear whether the authors realize that this assumption also ensures identification invariance. Indeed, Hunt and Blake (2018) in their otherwise careful paper seem to overlook this subtlety in their critique of the model by failing to acknowledge the restricted set of invariant transformations (the A-matrix of Equation (19) of Section 5 of Hunt and Blake (2018) ought to have identical, rather than freely varying, diagonal terms).

can more easily be obtained while preserving model flexibility, see e.g. the model of Li and Lu (2017).

The VAR/VECM approach to modelling $\mathbf{y}_t$ directly certainly has its merits as a flexible method for forecasting capable of capturing the dependency structure across ages. The approach also provides a useful framework for determining the number of common stochastic trends, i.e. the dimension of the driving dynamics. However, due to the high dimension the resulting models are often very complex and hard to interpret. For the purpose of gaining demographic insights by formulating and testing hypotheses we find that more parsimonious models with a limited set of interpretable factors are better suited.

3.5 Applications to UK Mortality Data

In this section we present two applications of cointegration-based mortality models. The applications focus on the inferential procedure, in particular hypothesis testing and interpretation of the models, rather than on the resulting forecasts. We use UK mortality data for males and females retrieved from the Human Mortality Database (2019).[7]

The first application is based on the Lee-Carter model and illustrates the care with which results must be interpreted due to semi-identifiability. The second application is based on the fully identified CBD-model for which a more detailed analysis is possible. In both cases, the analysis starts with a visual inspection of the mortality indices being modelled and tests for non-stationarity.

3.5.1 Lee–Carter Application

The period remaining life expectancy is a useful summary measure for the mortality conditions of a population at a given point in time. Figure 3.2 shows the period remaining life expectancy at birth and at age 60 for UK males and females, calculated using the observed death rates $m_{t,x} = D_{t,x}/E_{t,x}$. The most striking feature is the remarkable increase in life expectancy over the period; an increase also seen in most other developed countries. Another prominent feature is the degree of similarity between the life expectancy evolution of the two genders. This prompts the scientific question of whether the joint behaviour is the result of two random walk-type processes with similar drift and correlated innovations, or whether the two processes do in fact engage in a cointegrating relation? As a first attempt at answering this question, we consider a cointegration analysis of the mortality indices of two Lee-Carter models.

[7] Specifically, we use data for United Kingdom with HMD country code GBR_NP.

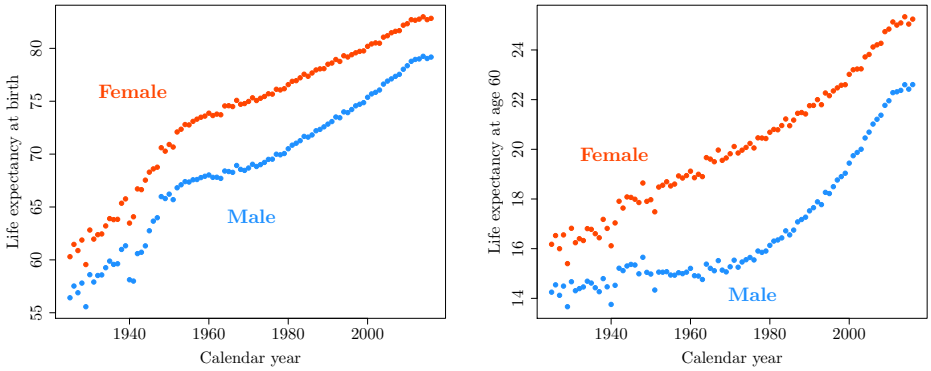

Figure 3.2: Empirical period remaining life expectancy for UK males and females at birth (left panel) and at age 60 (right panel) for the period 1930–2016.

From Figure 3.2 we can identify a number of different periods in the life expectancy evolution. After the steep and erratic initial part, the life expectancy evolution changes character during the early 1950s and improvements are hereafter smoother. Around 1970 improvements in male life expectancy pick up speed and the gender gap begins to narrow. Finally, in the last part of the series, around 2010, improvements slow down and life expectancy flattens for both sexes.

The choice of data period for the analysis is a compromise between including as much information as possible versus using only data adequately described by the models. Balancing these concerns we choose to use the period 1960–2016; a period so long that it enables us to capture potential equilibrium relations. As previously advertised we use the Poisson regression version of the Lee-Carter model with the identification constraints of (3.2.3). The Lee-Carter model is estimated separately for males and females over the period 1960–2016 and ages 0–100. Figure 3.3 shows the estimated mortality indices (k-index) and the age response parameters (b-parameters). The mortality indices evolve quite similarly over time, while the age response parameters are rather different between ages 20–80.

Cointegration Rank and Deterministic Structure

From unit root tests we conclude that $\hat{k}_t^{\male}$ and $\hat{k}_t^{\female}$ are indeed I(1)-processes (test results not shown).[8] The next objective of the analysis is to determine whether or not the two indices engage in an equilibrium correcting relationship, i.e. do $\hat{k}_t^{\male}$ and $\hat{k}_t^{\female}$ cointegrate?

[8]Specifically, we use the *Augmented Dickey-Fuller* test in conjunction with the *Phillips-Perron*, see e.g. Said and Dickey (1984) and Phillips and Perron (1988).

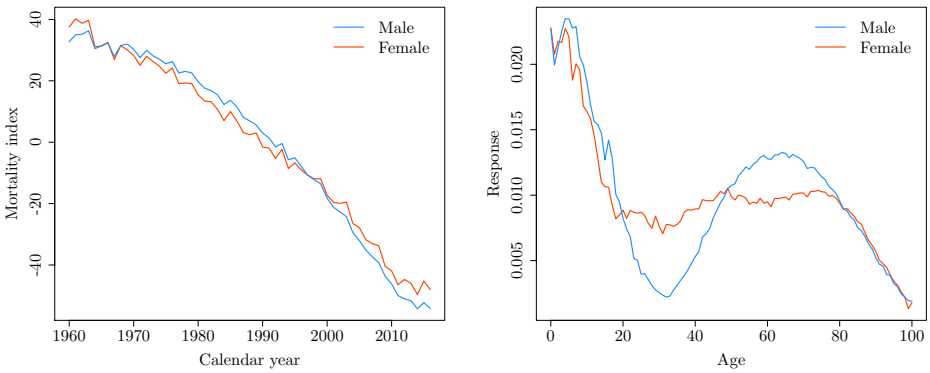

Figure 3.3: Parameters for Lee-Carter models fitted to UK male and female mortality over the period 1960–2016 and ages 0–100. Left panel shows the mortality index, $\hat{k}_t$, and right panel shows the age response parameters, $\hat{b}_x$.

To answer this question we employ the VECM of (3.4.2) and test for cointegration rank and deterministic structure in this model. For given deterministic structure, the test for cointegration rank is based on Johansen's trace statistic, cf. (3.A.15). The results are shown in Table 3.3. Recall that the null hypothesis of at most r cointegrated relations, $\text{rank}(\mathbf{\Pi}) \leq r$, is tested against an unrestricted $\mathbf{\Pi}$. Since the test for cointegration rank is very dependent on the assumed deterministic structure, the two must be determined jointly.

We first note that the two models $H_1(0)$ and $H_0^*(0)$ are identical, since $\text{rank}(\mathbf{\Pi}) = 0$ implies that $\boldsymbol{\alpha}$ is zero. Table 3.3 presents two tests for this model against, respectively, $H_1(2)$ and $H_0^*(2)$. Both of these are rejected at the 5%-significance level. Consequently, we conclude that the system cointegrates and we move on to test the two cointegration models $H_1(1)$ and $H_0^*(1)$. Both of these models are accepted and we therefore take the simpler of these, $H_1(1)$, as describing our data

$$\Delta \boldsymbol{k}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{k}_{t-1} + \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_t = \boldsymbol{\alpha}\left(\boldsymbol{\beta}' \boldsymbol{k}_{t-1} - \mathbb{E}\left[\boldsymbol{\beta}' \boldsymbol{k}\right]\right) + \boldsymbol{\tau}_1 + \boldsymbol{\varepsilon}_t, \tag{3.5.1}$$

where $\mathbb{E}\left[\boldsymbol{\beta}' \boldsymbol{k}\right] = -\left(\boldsymbol{\beta}' \boldsymbol{\alpha}\right)^{-1} \boldsymbol{\beta}' \boldsymbol{\theta}_0$ and $\boldsymbol{\tau}_1 = \boldsymbol{\beta}_\perp \left(\boldsymbol{\alpha}_\perp' \boldsymbol{\beta}_\perp\right)^{-1} \boldsymbol{\alpha}_\perp' \boldsymbol{\theta}_0$ are, respectively, the stationary mean of the cointegrating relation and the trend in the level of the process, cf. (3.4.3).

In an ordinary cointegration analysis model (3.5.1) would serve as a starting point for formulating and testing hypotheses of interest on the parameters. However, as described in Section 3.4.2, in the present context there are no further testable hypotheses and all there is left is to interpret the model.

Table 3.3: The trace test for cointegration rank for deterministic terms of class H_1 and H_0^*. The table shows the likelihood ratio test statistic for test of $H(r)$ in $H(2)$ for $r = 0, 1$. Critical values are the 95%-quantiles of the limiting distribution given in Section 15.3 of Johansen (1995).

Model	Cointegrating relations	Deterministic term	Trace statistic	Critical value
$H_1(0)$	0	$\boldsymbol{\theta}_0$	33.95	15.34
$H_1(1)$	1	$\boldsymbol{\theta}_0$	1.18	3.84
$H_0^*(0)$	0	$\boldsymbol{\theta}_0$	34.05	25.47
$H_0^*(1)$	1	$\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\rho_1 t$	1.28	12.39

Interpretation of the Cointegration Model

We obtain the following model, where we have used the just-identified normalization of Section 3.3.1 by which the first element of $\boldsymbol{\beta}$ is 1,

$$\begin{aligned}\begin{pmatrix}\Delta k_t^{\male}\\ \Delta k_t^{\female}\end{pmatrix} &= \begin{pmatrix}-0.084\\ 0.018\end{pmatrix}\begin{pmatrix}1 & -1.235\end{pmatrix}\begin{pmatrix}k_{t-1}^{\male}\\ k_{t-1}^{\female}\end{pmatrix} + \begin{pmatrix}-1.560\\ -1.526\end{pmatrix} + \boldsymbol{\varepsilon}_t\\ &= \begin{pmatrix}-0.084\\ 0.018\end{pmatrix}(s_{t-1} - 3.056) + \begin{pmatrix}-1.817\\ -1.471\end{pmatrix} + \boldsymbol{\varepsilon}_t,\end{aligned}$$

where $s_t = \hat{\boldsymbol{\beta}}'\boldsymbol{k}_t = k_t^{\male} - 1.235 k_t^{\female}$ and the drift term $\boldsymbol{\theta}_0$ is decomposed as in (3.5.1). Parameter estimates are obtained by reduced rank regression as described in Appendix 3.A.

Note that the adjustment coefficients have opposite signs for males and females such that the two indices are either pushed together or pushed apart in response to disequilibrium errors. However, since the female coefficient is much smaller than the male coefficient, the adjustments are primarily taken by the male index, while the female index evolves essentially like a random walk.

It is tempting to interpret the fact that the $\boldsymbol{\beta}$-coefficients are of the same magnitude but opposite signs as suggesting that male and female mortality "follow each other closely", i.e. as approximate coherence. However, this is not a valid conclusion since the (relative) magnitude of the $\boldsymbol{\beta}$-coefficient has no significance by itself, but is merely a result of the chosen identification schemes, cf. Section 3.4.2. The proper conclusion is that forecasted mortalities will be approximately coherent for ages with $\hat{b}_x^{\female} \approx 1.2\hat{b}_x^{\male}$, and for these ages only. From the right panel of Figure 3.3 we see that this relation is satisfied for only a small group of ages around age 25 and age 40.

Regarding the cointegrating relation, it follows from (3.5.1) that

$$s_t = (1 + \boldsymbol{\beta}'\boldsymbol{\alpha})\left(s_{t-1} - \mathbb{E}\left[\boldsymbol{\beta}'\boldsymbol{k}\right]\right) + \mathbb{E}\left[\boldsymbol{\beta}'\boldsymbol{k}\right] + u_t, \tag{3.5.2}$$

where $u_t = \boldsymbol{\beta}'\varepsilon_t$. Thus the cointegrating relation follows an autoregression with

AR-coefficient of $1+\boldsymbol{\beta}'\boldsymbol{\alpha}$.[9] We know from Section 3.4.2 that alternative identification schemes in the underlying Lee-Carter models lead to a new set of adjustment and long-run coefficients of the form $\tilde{\boldsymbol{\alpha}} = \mathbf{D}\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\beta}} = \mathbf{D}^{-1}\boldsymbol{\beta}$. Observing that $\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\alpha}} = \boldsymbol{\beta}'\boldsymbol{\alpha}$, it follows that the AR-coefficient is in fact invariant to the identification scheme(s). We can therefore conclude that the cointegrating relation is always mean-reverting with an estimated AR-coefficient of $1 + \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\alpha}} = 0.89$, implying a strong degree of mean reversion. Due to non-identifiability no further inference can be drawn about the nature of the joint behaviour under the cointegrated Lee-Carter model.

3.5.2 Cairns-Blake-Dowd Application

In contrast to the Lee-Carter model, the CBD model uses two, fully identified factors to describe the mortality evolution of the population under study. In the following we use the CBD model of Section 3.2.2 as the basis for a four-dimensional cointegration analyses of UK male and female mortality. The higher dimension opens for a richer, yet still interpretable, set of relations while factor identifiability enables the formulation of testable hypotheses.

The CBD model is intended for modelling of pensioners' mortality only and, consequently, we apply it to ages 60–100, rather than the full age span. Figure 3.4 shows the two mortality factors of the CBD model estimated separately for UK males and females over the period 1960–2016. The first factor represents the level of mortality and the second factor represents the slope of the mortality curve. Both factors show a clear trend over time. Not surprisingly, the level is generally declining reflecting an overall decrease in mortality over the past decades for both genders. We note that the profile of the level factor bears some resemblance to that of the mortality index of the Lee-Carter model displayed in Figure 3.3. However, in contrast to the mortality index of the Lee-Carter model the level factor of the CBD model represents the *absolute* level of mortality and it is therefore markedly higher for males than for females. The slope parameter, shown in the right panel of Figure 3.4, is generally upward trending indicating greater mortality improvements at younger ages than at older ages. While we observe a somewhat stable difference between male and female levels of mortality, there appears to be no obvious relation between the slope parameters for the two genders.

Cointegration Rank and Deterministic Structure

Proceeding as in Section 3.5.1, we first verify that the components of $\hat{\boldsymbol{k}}_t = (\hat{k}^{\male}_{t,1}, \hat{k}^{\female}_{t,1}, \hat{k}^{\male}_{t,2}, \hat{k}^{\female}_{t,2})'$ are I(1)-processes (test results not shown).

The next step of the analysis is to determine a suitable VECM for the composite index in which we can subsequently formulate and test specific hypotheses. We

[9] Note that, since $|\boldsymbol{A}(z)| = |1-z||1-z(1+\boldsymbol{\beta}'\boldsymbol{\alpha})|$ the first part of Condition 3.3.2 reduces to $|1+\boldsymbol{\beta}'\boldsymbol{\alpha}| < 1$, i.e. stationarity of (3.5.2).

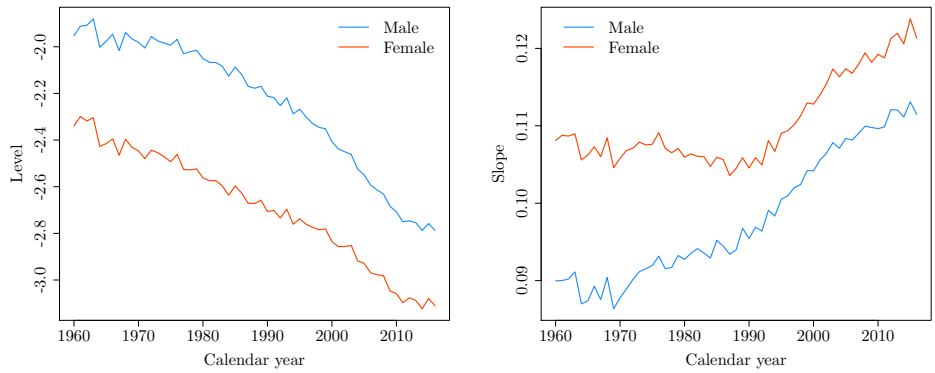

Figure 3.4: CBD model fitted to UK male and female mortality. Left panel shows the mortality level, $\hat{k}_{t,1}$, and right panel shows the mortality slope, $\hat{k}_{t,2}$.

employ again the VECM of (3.4.2) and test jointly for cointegration rank and deterministic structure in this model. Table 3.4 shows the results of Johansen's trace test for $\text{rank}(\mathbf{\Pi}) \leq r$ against an unrestricted $\mathbf{\Pi}$ for the two deterministic structures of relevance, H_1 and H_0^*. Figure 3.5 shows the relation between the various models; note that $H_1(0) = H_0^*(0)$ corresponds to a (multivariate) random walk with drift customarily used for forecasting in the CBD model, cf. Section 3.2.2.

$$\begin{array}{ccccccccc} H_1(0) & \subset & H_1(1) & \subset & H_1(2) & \subset & H_1(3) & \subset & H_1(4) \\ \| & & \cap & & \cap & & \cap & & \cap \\ H_0^*(0) & \subset & H_0^*(1) & \subset & H_0^*(2) & \subset & H_0^*(3) & \subset & H_0^*(4) \end{array}$$

Figure 3.5: Relation between the I(1)-models considered in Table 3.4.

It can be seen from Table 3.4 that the random walk hypothesis ($\text{rank}(\mathbf{\Pi}) = 0$) is rejected in both $H_1(4)$ and $H_0^*(4)$, and we conclude a cointegration rank of at least one. The smallest model of rank 1, $H_1(1)$, is also rejected, while $H_1(2)$ and $H_0^*(1)$ are both accepted. These two models represent the two smallest acceptable models, cf. Figure 3.5. Neither of the models are contained in the other, but $H_0^*(1)$ is arguably the "simpler" model introducing only a single additional trend term while $H_1(2)$ introduces an additional cointegrating relation. Consequently, we adopt $H_0^*(1)$ as our starting model, i.e. the VECM

$$\Delta \boldsymbol{k}_t = \boldsymbol{\alpha} \left(\boldsymbol{\beta}' \boldsymbol{k}_{t-1} + \rho_1 t \right) + \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_t, \tag{3.5.3}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are four-dimensional vectors.

Table 3.4: The trace test for cointegration rank for deterministic terms of class H_1 and H_0^*. The table shows the likelihood ratio test statistic for test of $H(r)$ in $H(4)$ for $r = 0, 1, 2, 3$. The critical value is the 95%-quantile of the limiting distribution given in Section 15.3 of Johansen (1995).

Model	Cointegrating relations	Deterministic term	Trace statistic	Critical value
$H_1(0)$	0	$\boldsymbol{\theta}_0$	80.29	47.21
$H_1(1)$	1	$\boldsymbol{\theta}_0$	38.87	29.38
$H_1(2)$	2	$\boldsymbol{\theta}_0$	12.95	15.34
$H_1(3)$	3	$\boldsymbol{\theta}_0$	0.51	3.84
$H_0^*(0)$	0	$\boldsymbol{\theta}_0$	80.31	62.61
$H_0^*(1)$	1	$\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\rho_1 t$	38.89	42.20
$H_0^*(2)$	2	$\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\boldsymbol{\rho}_1 t$	12.96	25.47
$H_0^*(3)$	3	$\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\boldsymbol{\rho}_1 t$	0.52	12.39

Hypothesis Testing and Interpretation

Having established that the system has a cointegration rank of one, we can now investigate hypotheses on the nature of the equilibrium correcting relation. We are primarily interested in investigating whether the mortality levels of the two genders enter the cointegrating relation with coefficients of the same magnitude and opposite signs, i.e. whether the distance between the level parameters is the quantity entering the stable relation. In addition, we are also interested in investigating the degree of dependence between the level and slope parameters which can be formulated as restrictions on the adjustment coefficients. Other hypotheses of interest could be formulated, but we restrict ourselves to these two.

We first formulate the hypothesis of main interest, namely that the first two components of $\boldsymbol{\beta}$ are of the same magnitude and opposite signs. The last two components of $\boldsymbol{\beta}$ are left unrestricted. At the same time we would like to test for the absence of a linear drift in the cointegrating relation ($\rho_1 = 0$), since this term muddles the interpretation of the stationary relation. Hence, we consider the composite hypothesis[10]

$$\mathscr{H}_0 \ : \ (\boldsymbol{\beta}', \rho_1) = (1, -1, \varphi_1, \varphi_2, 0), \tag{3.5.4}$$

where $\varphi_1, \varphi_2 \in \mathbb{R}$ are unrestricted parameters. The test statistic for this hypothesis is $\chi^2(2)$-distributed with a value of 0.005 and a critical value of 5.99 at a 5%-significance

[10]For an in-depth description of this type of test, we refer to Johansen and Juselius (1992).

level. Hence, we accept hypothesis (3.5.4) and obtain the model

$$\begin{pmatrix} \Delta k^{\male}_{t,1} \\ \Delta k^{\female}_{t,1} \\ \Delta k^{\male}_{t,2} \\ \Delta k^{\female}_{t,2} \end{pmatrix} = \begin{pmatrix} -0.105 \\ -0.042 \\ 0.002 \\ 0.002 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 21.66 \\ -12.44 \end{pmatrix}' \begin{pmatrix} k^{\male}_{t-1,1} \\ k^{\female}_{t-1,1} \\ k^{\male}_{t-1,2} \\ k^{\female}_{t-1,2} \end{pmatrix} + \begin{pmatrix} 0.1103 \\ 0.0368 \\ -0.0015 \\ -0.0024 \end{pmatrix} + \boldsymbol{\varepsilon}_t. \tag{3.5.5}$$

The second question of interest is whether the joint behaviour of the level and slope parameters can be simplified. In particular, whether the adjustment coefficients for the slope parameters, $k^{\male}_{t,2}$ and $k^{\female}_{t,2}$, as well as the female level parameter, $k^{\female}_{t,1}$, are in fact zero, i.e. whether the slope parameters and the female level parameter are weakly exogenous for the long-run coefficients $\boldsymbol{\beta}$. This would imply that while the slope parameters and the female level parameter influence the long-run relation, the long-run relation has no influence on them. To test the hypothesis of weak exogeneity while retaining the established model, we formulate a simultaneous linear restriction on both the adjustment coefficients $\boldsymbol{\alpha}$ and the long-run coefficients $\boldsymbol{\beta}$[11]

$$\mathscr{H}_0 \ : \ (\boldsymbol{\alpha}', \boldsymbol{\beta}', \rho_1) = (\psi_1, 0, 0, 0, 1, -1, \varphi_1, \varphi_2, 0), \tag{3.5.6}$$

where $\psi_1, \varphi_1, \varphi_2 \in \mathbb{R}$ are unrestricted parameters. The test statistic for this hypothesis is $\chi^2(5)$-distributed with a value of 9.54 and a critical value of 11.07 at a 5%-significance level. We conclude that both slope parameters are weakly exogenous and are therefore not impacted by the long-run relation. We obtain the following final model

$$\begin{aligned} \begin{pmatrix} \Delta k^{\male}_{t,1} \\ \Delta k^{\female}_{t,1} \\ \Delta k^{\male}_{t,2} \\ \Delta k^{\female}_{t,2} \end{pmatrix} &= \begin{pmatrix} -0.089 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 11.78 \\ -2.12 \end{pmatrix}' \begin{pmatrix} k^{\male}_{t-1,1} \\ k^{\female}_{t-1,1} \\ k^{\male}_{t-1,2} \\ k^{\female}_{t-1,2} \end{pmatrix} + \begin{pmatrix} 0.1061 \\ -0.0138 \\ 0.0004 \\ 0.0002 \end{pmatrix} + \boldsymbol{\varepsilon}_t, \\ &= \begin{pmatrix} -0.089 \\ 0 \\ 0 \\ 0 \end{pmatrix} (s_{t-1} - 1.395) + \begin{pmatrix} -0.0181 \\ -0.0138 \\ 0.0004 \\ 0.0002 \end{pmatrix} + \boldsymbol{\varepsilon}_t, \end{aligned} \tag{3.5.7}$$

where $s_t = \hat{\boldsymbol{\beta}}' \boldsymbol{k}_t = k^{\male}_{t,1} - k^{\female}_{t,1} + 11.78 k^{\male}_{t,2} - 2.12 k^{\female}_{t,2}$ and the drift term is decomposed in the stationary mean of the cointegrating relation and the trend in the level of the process, cf. (3.4.3). The cointegrating relation follows an AR-regression of form (3.5.2) with an AR-coefficient of $1 + \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\alpha}} = 0.91$.

The model can be interpreted as a "mixture" model where the original system is partitioned into a conditional and a marginal system. The marginal system,

[11] Likelihood ratio tests for composite hypotheses on cointegrating relations are covered in Johansen (1991).

consisting of the slope parameters and the female level parameter, evolves as a trivariate random walk with drift. Conditioned on these parameters, the male level parameter evolves as the sum of a random walk with drift (and innovations conditioned on the "marginal" innovations) and an error correction term. The error correction term seeks to maintain the long-run relation between the level and slope parameters, but it does so by affecting only the male level parameter.

Implications for Forecasting

Mortality modelling is often performed with the aim of forecasting and cointegration models are often enforced to ensure coherence. In this paper we wish to promote the broader use of cointegration as an inferential tool to obtain insights about mortality factor dynamics. Also, we wish to demonstrate that although cointegration does not necessarily imply coherence (in the strict sense of Section 3.2.1) it might still lead to strongly coupled, joint forecasts. Forecasts which might indeed be more plausible being inferred from data, rather than imposed. In this section we illustrate these points for the model (3.5.7) obtained above.

For forecasting purposes, it is useful to rewrite the model on VAR-form as

$$\boldsymbol{k}_t = \boldsymbol{\Pi}_1 \boldsymbol{k}_{t-1} + \boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_4 \left(\mathbf{0}, \boldsymbol{\Sigma} \right), \tag{3.5.8}$$

where $\boldsymbol{\Pi}_1 = \mathbf{I} + \boldsymbol{\alpha}\boldsymbol{\beta}'$, cf. (3.3.1). From this we can readily generate stochastic forecasts, and we also immediately have the forecasting distribution from which the mean forecast and confidence intervals can be derived,

$$\boldsymbol{k}_{T+h} | \boldsymbol{k}_T \sim \mathcal{N} \left(\boldsymbol{\Pi}_1^h \boldsymbol{k}_T + \sum_{i=0}^{h-1} \boldsymbol{\Pi}_1^i \boldsymbol{\theta}_0, \ \sum_{i=0}^{h-1} \boldsymbol{\Pi}_1^i \boldsymbol{\Sigma} (\boldsymbol{\Pi}_1^i)' \right), \tag{3.5.9}$$

where $h \geq 1$ is the horizon and $\boldsymbol{k}_T = \hat{\boldsymbol{k}}_T$ is the last value of the estimated indices.

Equation (3.5.9) is a general result valid for all VAR(1)-models with constant drift. This is useful for numerical computations, but the mean and variance structures are not easily discerned. Using the cointegrating relations it is possible to obtain more revealing expressions for the model at hand.

First, let $\mathbf{B} = \boldsymbol{\alpha} \left(\boldsymbol{\beta}' \boldsymbol{\alpha} \right)^{-1} \boldsymbol{\beta}'$ and $\mathbf{C} = \boldsymbol{\beta}_\perp \left(\boldsymbol{\alpha}_\perp' \boldsymbol{\beta}_\perp \right)^{-1} \boldsymbol{\alpha}_\perp'$ and notice that $\mathbf{I} = \mathbf{C} + \mathbf{B}$, cf. Chapter 3 of Johansen (1995). Next, observe that $\boldsymbol{\Pi}_1$ has eigenvalues 1 and $\lambda = 1 + \boldsymbol{\beta}' \boldsymbol{\alpha}$ with corresponding eigenspaces $\text{span}(\boldsymbol{\beta}_\perp)$ and $\text{span}(\boldsymbol{\alpha})$, respectively. Hence, for $\boldsymbol{v} \in \mathbb{R}^4$ and $i \geq 0$,

$$\boldsymbol{\Pi}_1^i \boldsymbol{v} = \boldsymbol{\Pi}_1^i \left(\mathbf{C} + \mathbf{B} \right) \boldsymbol{v} = \boldsymbol{\Pi}_1^i \left(\mathbf{C}\boldsymbol{v} + \mathbf{B}\boldsymbol{v} \right) = \mathbf{C}\boldsymbol{v} + \lambda^i \mathbf{B}\boldsymbol{v}, \tag{3.5.10}$$

since $\mathbf{C}\boldsymbol{v} \in \text{span}(\boldsymbol{\beta}_\perp)$ and $\mathbf{B}\boldsymbol{v} \in \text{span}(\boldsymbol{\alpha})$. In words, $\boldsymbol{\Pi}_1$ acts on $\boldsymbol{v}$ by leaving intact its $\boldsymbol{\beta}_\perp$-component, while shrinking its $\boldsymbol{\alpha}$-component by a factor of λ (< 1).

Using (3.5.10) and the formula for the h first terms of a geometric series $\sum_{i=0}^{h-1} \lambda^i = (1-\lambda^h)/(1-\lambda)$, we find the following expression for the mean of the forecasting distribution

$$\begin{aligned}\boldsymbol{\Pi}_1^h \boldsymbol{k}_T + \sum_{i=0}^{h-1} \boldsymbol{\Pi}_1^i \boldsymbol{\theta}_0 &= \mathbf{C}\boldsymbol{k}_T + \lambda^h \mathbf{B}\boldsymbol{k}_T + \mathbf{C}\boldsymbol{\theta}_0 h + (1-\lambda)^{-1}(1-\lambda^h)\mathbf{B}\boldsymbol{\theta}_0 \\ &= \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 h + \lambda^h \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\left[\boldsymbol{\beta}'\boldsymbol{k}_T - \mathbb{E}(\boldsymbol{\beta}'\boldsymbol{k})\right], \end{aligned} \tag{3.5.11}$$

where $\boldsymbol{\tau}_0 = \mathbf{C}\boldsymbol{k}_T + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\mathbb{E}(\boldsymbol{\beta}'\boldsymbol{k})$, $\boldsymbol{\tau}_1 = \mathbf{C}\boldsymbol{\theta}_0$ and $\mathbb{E}(\boldsymbol{\beta}'\boldsymbol{k}) = -(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\boldsymbol{\beta}'\boldsymbol{\theta}_0$. We see that asymptotically the mean of the process behaves like a random walk with a drift term that preserves the cointegrating relation. The last term of (3.5.11) shows how the initial disequilibrium error decays exponentially to zero. This term is not present in the Granger representation of Theorem 3.3.3, since there we assume that $\boldsymbol{\beta}'\boldsymbol{k}_T$ is distributed according to its stationary distribution, while in (3.5.11) we condition on the entire vector $\boldsymbol{k}_T$.[12]

For the variance of the forecasting distribution, it can be shown that

$$\lim_{h\to\infty} \frac{1}{h} \sum_{i=0}^{h-1} \boldsymbol{\Pi}_1^i \boldsymbol{\Sigma} (\boldsymbol{\Pi}_1^i)' = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}', \tag{3.5.12}$$

see pp. 68–69 of Johansen (1995). Thus, asymptotically the variance of the process accumulates linearly, with the variance from the random walk component on $\operatorname{span}(\boldsymbol{\beta}_\perp)$ dominating the total variability.

To give a more intuitive understanding of the role of cointegration, we conclude with a numerical example where we compare the (joint) forecast from the cointegration model (3.5.7) with separate, gender-specific forecasts based on the bivariate random walk model of Section 3.2.2. We compare the forecasted cohort remaining life expectancy as it depends on the projected mortality surface over a long horizon and it is therefore well suited to capture differences in dependency structures over time.

The two models describe and project the estimated CBD-parameters of Figure 3.4 differently. However, the models have similar deterministic structures and we therefore expect similar mean forecasts. This is confirmed by Figure 3.6 which shows the forecasting distributions of the cohort remaining life expectancy for a 60-year-old Briton based on 100,000 simulations. The female distributions align perfectly since the cointegration model (3.5.7) in fact results in a random walk forecast as well. The male distribution is shifted slightly towards its female counterpart in the case of the cointegrated model. This is the result of the male projection reacting to the perceived "disequilibrium".

[12] When $\boldsymbol{k}_t$ has the form (3.5.8) Theorem 3.3.3 reads (in terms of expected value) $\mathbb{E}\left[\boldsymbol{k}_{T+h}\right] = \mathbf{C}\boldsymbol{k}_T + \mathbf{C}\sum_{i=1}^{h}\boldsymbol{\theta}_0 + \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\boldsymbol{\alpha}\right)^{-1}\sum_{i=0}^{\infty}\left(\boldsymbol{I}+\boldsymbol{\alpha}\boldsymbol{\beta}'\right)^i\boldsymbol{\beta}'\boldsymbol{\theta}_0 = \mathbf{C}\boldsymbol{k}_T + \mathbf{C}\boldsymbol{\theta}_0 h - \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\boldsymbol{\alpha}\right)^{-1}\left(\boldsymbol{\beta}'\boldsymbol{\alpha}\right)^{-1}\boldsymbol{\beta}'\boldsymbol{\theta}_0 = \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 h$, i.e. (3.5.11) without the exponentially decaying error-correction term.

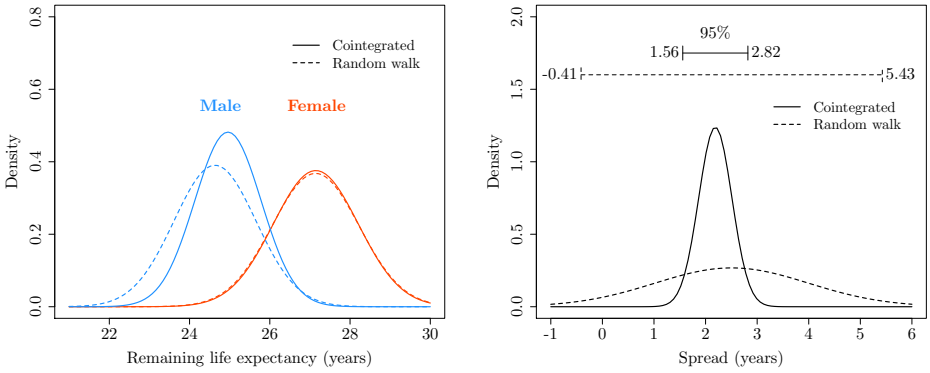

Figure 3.6: Forecasting distributions of the cohort remaining life expectancy for the random walk and the cointegration model based on 100,000 simulations. Left panel shows the gender-specific life expectancy, $e^c_{60}(2017)$, and the right panel shows the life expectancy difference (spread) between females and males.

As the theoretical analysis showed, the cointegration model generally behaves asymptotically like (coupled) random walks. Therefore, the marginal model for each sex is (at least asymptotically) close to the corresponding random walk model, both in terms of deterministic and stochastic behaviour. The dependency structure of the two models is, however, very different. By construction, the random walk model yields independent forecasts for men and women, while the cointegration model yields highly dependent forecasts. This implies a much more narrow distribution for the difference in life expectancy between men and women under the cointegration model than under the random walk model, as illustrated in the right panel of Figure 3.6.

In summary, the cointegration model yields forecasts that are similar to those obtained from the simpler random walk model for each sex, but with a more plausible dependency structure. The cross-gender dependency is achieved by a single cointegrating relation derived from data. The resulting forecasts are well-behaved and empirically justified, but they are not coherent (in the strict mathematical sense). This indicates that the current definition of coherence might be too strict and too specific a requirement, and that other types of "coherence" might be equally good—or even better—when judging the quality of joint forecasts.

3.6 Concluding Remarks

In this paper we have discussed the interlinked concepts of identifiability, coherence and cointegration in the context of multi-population mortality modelling. We have made the point that cointegration has an important role to play as an inferential tool to obtain insights into the joint dynamics of mortality factors. This role goes beyond

the typical usage of cointegration as merely a tool to obtain coherent forecasts – defined as forecasts for which the relative age-specific mortality rates converge over time.[13]

Since its introduction in Li and Lee (2005), the concept of coherence has served as the gold standard for joint forecasts of related populations, and many models have been designed with the explicit goal of achieving coherence. At first sight, coherence seems like a reasonable property, but on further inspection it appears somewhat arbitrary and specifically tailored to models of the Lee-Carter type with log-linear modelling of mortality rates. Joint forecasts based on other model types, e.g. logistic, can produce equally plausible dependency structures and thus be equally "coherent", even if they lack relative convergence. In our view, insisting on coherence is not a suitable starting point, nor a reasonable restriction, for a joint analysis and might in fact produce forecasts that are at odds with historic data.

In this paper we have focused on two types of linear trend models suitable for analysing the period effect of age-period mortality models, and we have shown how to interpret these models by use of the Granger decomposition. Cointegration is a technical field, and the analysis and interpretation of models and hypotheses are not straightforward. In many mortality models, e.g. Lee-Carter type models, the factor(s) are only semi-identifiable in which case additional difficulties (and pitfalls) arise. From a statistical point of view, it is in general meaningless to test, or impose restrictions, on the long-run coefficients in the context of semi-identifiable models. In contrast, fully identified models give access to the full inferential power of cointegration and in our opinion this is a strong argument in favour of these models if the aim is to gain subject matter insights.

Acknowledgements

The authors wish to thank two anonymous referees for valuable input for improving the presentation.

3.A Maximum Likelihood Estimation of the VECM

Consider the VAR(k) model in VECM form with $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ for $p \times r$ matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$,

$$\Delta \mathbf{y}_t = \sum_{i=1}^{k-1} \mathbf{\Gamma}_i \Delta \mathbf{y}_{t-i} + \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{y}_{t-1} + \mathbf{\Phi}\mathbf{D}_t + \boldsymbol{\varepsilon}_t, \tag{3.A.1}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}_p(0, \mathbf{\Sigma})$ and independent. Based on data for $t = 1, ..., T$, maximum likelihood estimates of the freely varying parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}_1, ...\mathbf{\Gamma}_{k-1}, \mathbf{\Phi}, \mathbf{\Sigma})$ can be

[13] Whether the goal of (statistical) models is to gain insights or to forecast is the object of a long-standing, and still highly relevant, debate, see e.g. Breiman (2001) and Shmueli (2010).

obtained by reduced rank regression. Following Juselius (2006), we introduce the shorthand notation

$$\mathbf{Z}_{0t} = \Delta \mathbf{y}_t, \tag{3.A.2}$$
$$\mathbf{Z}_{1t} = \mathbf{y}_{t-1}, \tag{3.A.3}$$
$$\mathbf{Z}_{2t} = \left[\Delta \mathbf{y}'_{t-1}, ..., \Delta \mathbf{y}'_{t-k+1}, \mathbf{D}'_t\right], \tag{3.A.4}$$

and write (3.A.1) on the form

$$\mathbf{Z}_{0t} = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{Z}_{1t} + \boldsymbol{\Psi}\mathbf{Z}_{2t} + \boldsymbol{\varepsilon}_t, \tag{3.A.5}$$

where $\boldsymbol{\Psi} = [\boldsymbol{\Gamma}_1, ..., \boldsymbol{\Gamma}_{k-1}, \boldsymbol{\Phi}]$. Define the product moment matrices

$$\mathbf{M}_{ij} = T^{-1} \sum_{t=1}^{T} \mathbf{Z}_{it}\mathbf{Z}'_{jt}, \quad i, j = 0, 1, 2, \tag{3.A.6}$$

and the sample-covariance matrices

$$\boldsymbol{S}_{ij} = \boldsymbol{M}_{ij} - \boldsymbol{M}_{i2}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{2j}, \quad i, j = 0, 1. \tag{3.A.7}$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is found by solving the eigenvalue problem

$$|\lambda \boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}| = 0, \tag{3.A.8}$$

for eigenvalues $1 > \hat{\lambda}_1 > \cdots > \hat{\lambda}_p > 0$ with associated eigenvectors $\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_p$ that satisfy

$$\hat{\lambda}_i \boldsymbol{S}_{11}\hat{\boldsymbol{v}}_i = \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}, \ i = 1, ..., p, \tag{3.A.9}$$
$$\hat{\boldsymbol{V}}'\boldsymbol{S}_{11}\hat{\boldsymbol{V}} = \boldsymbol{I}, \tag{3.A.10}$$

where $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_p)$. The cointegrating relations $\hat{\boldsymbol{\beta}}$ are given by the first r eigenvectors

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{v}}_1, ..., \hat{\boldsymbol{v}}_r). \tag{3.A.11}$$

The remaining parameters are estimated as

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}) = \boldsymbol{S}_{01}\boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta})^{-1}, \tag{3.A.12}$$
$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{S}_{00} - \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta})\boldsymbol{\alpha}', \tag{3.A.13}$$
$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{M}_{02}\boldsymbol{M}_{22}^{-1} - \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}, \tag{3.A.14}$$

with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}})$.

Trace Statistic for Test of Cointegration Rank

Recall that $H(r)$ denotes the hypothesis that $\text{rank}(\mathbf{\Pi}) \leq r$, or equivalently that $\mathbf{\Pi}$ has at most r non-zero eigenvalues. The likelihood ratio test statistic for $H(r)$ in $H(p)$, known as the *trace statistic*, is given by

$$LR_{trace}(r) = -T \sum_{i=r+1}^{p} \log(1 - \hat{\lambda}_i), \tag{3.A.15}$$

where $\hat{\lambda}_i$ are the eigenvalues found by solving (3.A.8). The asymptotic distribution of the trace statistic depends on the deterministic terms. Critical values are tabulated in Section 15.3 of Johansen (1995) for the linear models in Section 3.3.2. Intuitively, if $\text{rank}(\mathbf{\Pi}) = r$ then $LR_{trace}(r)$ will be small, since $\hat{\lambda}_{r+1}, \ldots, \hat{\lambda}_p$ will all be close to zero. Conversely, the test statistic will be large if $\text{rank}(\mathbf{\Pi}) > r$, since at least $\hat{\lambda}_{r+1}$ will deviate from zero.

Chapter 4

Sex Differential Dynamics in Coherent Mortality Models

This chapter contains the manuscript *Jallbjørn and Jarner (2022)*.

Abstract

The main purpose of coherent mortality models is to produce plausible, joint forecasts for related populations avoiding, e.g., crossing or diverging mortality trajectories; however, the coherence assumption is very restrictive and it enforces trends that may be at odds with data. In this paper we focus on coherent, two-sex mortality models and we prove, under suitable conditions, that the coherence assumption implies sex gap unimodality, i.e., we prove that the difference in life expectancy between women and men will first increase and then decrease. Moreover, we demonstrate that, in the model, the sex gap typically peaks when female life expectancy is between 30 to 50 years. This explains why coherent mortality models predict narrowing sex gaps for essentially all Western European countries and all jump-off years since the 1950s, despite the fact that the actual sex gap was widening until the 1980s. In light of these findings, we discuss the current role of coherence as the gold standard for multi-population mortality models.

JEL classification: C32, J11.
Keywords: *Life expectancy, Sex differential, Sex gap, Coherence, Mortality modelling, Projection, Forecasting.*

4.1 Introduction

The aim of coherent multi-population mortality models is to forecast the mortality of related populations preserving the structural differences observed in the past, e.g., preserving differences between countries, regional differences within a country, or higher mortality for men than for women. The idea is to prevent divergence or implausible crossings of mortality trajectories that can arise from forecasting each population individually. The concept of coherence was introduced by Li and Lee (2005), as an extension to the popular Lee-Carter model (Lee and Carter, 1992), and later formalized by Hyndman et al. (2013). Since then a number of coherent models have been proposed, both within the Lee-Carter framework (Li and Hardy, 2011; Li, 2013; Zhou et al., 2014; Kleinow, 2015), with added cohort effects (Jarner and Kryger, 2011; Cairns et al., 2011b; Börger and Aleksic, 2014), and based on the functional data approach (Hyndman et al., 2013; Shang and Hyndman, 2017). Today, coherence still serves as the gold standard for multi-population mortality models.

Technically, coherence means that the ratio of age-specific mortality rates of the populations being forecasted converges (to finite, age-specific constants). This requirement ties the forecasts together and it ensures that differences in aggregate characteristics, e.g., survival probabilities and life expectancies, remain bounded as desired. The flip-side, however, is that converging mortality sex ratios may not be supported by data and enforcing it can lead to unrealistic continuations of historic trends despite the intention, see Hunt and Blake (2018) and Jarner and Jallbjørn (2020). In this paper we investigate and exemplify the implications of coherence in two-sex mortality models, in particular, the implications for the dynamics of the life expectancy difference between the sexes.

The sex differential in life expectancy is a key statistic for summarizing and communicating discrepancies in sex-specific mortality curves. The differential has varied considerably over time, although its general shape has been fairly consistent from country to country. In the Western world, the gap has the shape of a (unimodal) hill; the differential widened substantially in favor of women throughout most of the 20th century, but the trend reversed around the 1980s and the gap has continued to narrow since then (Trovato and Lalu, 1996; Glei and Horiuchi, 2007).

Studies have sought to explain the observed trends in the sex gap through changes in behavioral, socioeconomic, and health factors, for example, smoking and drinking habits (Retherford, 1972; Mäkelä, 1998; Preston and Wang, 2006), labor market participation (Pampel and Zimmer, 1989; Trovato, 2005), risk behavior (Waldron, 1983), and cause-of-death contributions (Pampel, 2003; Trovato and Lalu, 2007; Booth, 2016). The rationale being that changes in external risk factors explain changes in mortality sex ratios, which in turn explain trends in the life expectancy

differential. This common-sense reasoning assumes that changing mortality sex ratios is the main driver of the sex gap, both when it increases and when it decreases. However, demographic analyses have challenged this assumption. In particular, Glei and Horiuchi (2007) and Cui et al. (2019) show that while the widening sex gap was indeed caused by changing mortality sex ratios, the narrowing of the sex gap was primarily caused by general mortality improvements for both sexes in combination with heterogeneity in the death distributions. In simpler terms, the expanding sex gap in the Western world was caused primarily by female mortality improving faster than male mortality (i.e., changing mortality sex ratios), while the subsequent narrowing of the gap was caused primarily by improvements in mortality for both sexes, with women retaining their relative advantage (i.e., improvements under stable mortality sex ratios).

In this paper, we demonstrate that coherent two-sex models generally imply unimodal sex gap dynamics. At first sight, this might seem as an attractive feature given that the sex gap in the Western world has also been unimodal, as described above. It turns out, however, that in practice the forecasts are almost always on the declining part of the sex gap trajectory. This in turn implies that coherent models are ill-suited to forecast the mortality in periods with increasing sex gaps. Coincidentally, the concept of coherence was introduced after a prolonged period of narrowing sex gaps and the continuation of this trend was seen as an argument in favor of coherent models, see Hyndman et al. (2013). That coherent models produce sensible forecasts only in periods with narrowing sex gaps does however question the status of coherence as a universally desirable feature in multi-population models. We will return to this point towards the end of the paper.

The rest of the paper is organized as follows. First, we illustrate the evolution of the sex gap in Western Europe since the 1950s and we survey the existing sex gap decomposition methods, formalizing the ratio and level effects responsible for the widening and the narrowing of the gap respectively. Second, we present our main mathematical result showing sex gap unimodality in (strongly) coherent models under certain conditions; we discuss the intuition behind the conditions and show by example when multi-modality can occur. Third, we analyze a dynamic Gompertz model as a simple example of a coherent model satisfying the conditions; we compute the typical sex gap trajectories that can arise under this model and we use this to explain why the forecasted sex gap will almost always be narrowing. Fourth, we apply the coherent models of Li and Lee (2005) and Hyndman et al. (2013) to Western European countries at selected jump-off years during the period with expanding sex gaps; with reference to the mathematical results, we discuss why the majority of these forecasts predict narrowing gaps. Finally, we end with some concluding remarks.

4.1.1 Data and Notation

Data is obtained from the Human Mortality Database (2022) and consists of death counts, $D(x,t)$, and central exposure-to-risk estimates, $E(x,t)$, on Lexis A-sets, that is, age-period cells of the form $[x,x+1)\times[t,t+1)$ for integer ages $x\in\{0,\dots,110\}$ and calendar years $t\in\{1950,\dots,2020\}$ for countries in Western Europe. The empirical death rate is estimated as

$$m(x,t)=D(x,t)/E(x,t). \tag{4.1.1}$$

Death rates for Western Europe are obtained by pooling death and exposure counts across individual countries. Throughout, (period) life expectancies are calculated by numerical integration whenever a continuous mortality curve is available, and under the assumption of piecewise constant mortality, whenever the mortality curve is only available at integer ages. That is,

$$e(x,t)=\frac{1}{S(x,t)}\int_x^\omega S(y,t)\,\mathrm{d}y=\int_x^\omega e^{-\int_x^y \mu(z,t)\,\mathrm{d}z}\,\mathrm{d}y=\sum_{i=x}^{\omega-1}\frac{1-e^{-\mu(i,t)}}{\mu(i,t)}e^{-\sum_{j=x}^{i-1}\mu(j,t)}, \tag{4.1.2}$$

where $S(x,t)=\exp(-\int_0^x \mu(y,t)\,\mathrm{d}y)$ is the (period) survival function, $\mu(x,t)$ the force of mortality at age x and time t, and $\omega\in(0,\infty)$ the age at truncation. Note that ω is not necessarily the maximum attainable age, and we do not assume that $\mu(x)=\infty$ for $x>\omega$. We return to the role of the truncation age later in the paper. The last equality in Equation (4.1.2) assumes piecewise constant mortality, a full derivation is given in Appendix 4.C.

4.2 Changes in Mortality, Life expectancy and Sex Differentials

In this section, we briefly survey the existing methods for decomposing changes in life expectancy and the sex differential. This provides the basis for the subsequent treatment of sex differentials in coherent mortality models.

4.2.1 Decomposing Changes in Life Expectancy

There are two main approaches for decomposing changes in life expectancy into constituent parts. The first approach, pioneered by Pollard (1982) and Arriaga (1984), focuses on the effect of changing mortality from one age-specific schedule to another, and is typically used to assess how different age groups contribute in driving life expectancy progress between two distinct points in time. The second approach, popularized by Keyfitz (1977a), examines the effect of a local change to the mortality curve by quantifying how various age-specific improvements in mortality,

$$\rho(x,t):=-\frac{\dot\mu(x,t)}{\mu(x,t)}=-\frac{\partial}{\partial t}\log\mu(x,t), \tag{4.2.1}$$

translate into changes in life expectancy. In (4.2.1) and throughout, a dot over a function is used to denote its derivative with respect to time as in Vaupel and Romo (2003).

Analyzing the effects of local change have been instrumental for understanding the linkage between the age pattern of mortality and the trends in $e(0,t)$ observed in data. Indeed, studies have shown that the dispersion of the life table death distribution is a main determinant of the pace at which life expectancy improves (Keyfitz, 1977a; Vaupel, 1986; Goldman and Lord, 1986).

The average number of life-years lost due to death (lifespan disparity) is given by

$$e^\dagger(t) = \int_0^\omega w(x,t)\,\mathrm{d}x, \tag{4.2.2}$$

where $w(x,t) = \mu(x,t)S(x,t)e(x,t)$ is the (life table) probability of dying at age x times the remaining life expectancy at that age. Lifespan disparity, $e^\dagger$, is a dispersion measure, i.e., it quantifies the effect of age of death being distributed across ages. At one extreme, if $e^\dagger$ is zero then everyone dies at the same age. Conversely, if $e^\dagger$ is large then the population experiences a high number of premature deaths and large gains in $e(0,t)$ can be made by reducing mortality. Improvements among the young are particularly important as more life years are lost upon death at these ages. This relation was formalized by Keyfitz (1977a) who showed that if the same rate of mortality improvement, $\rho(t)$, applies to all ages, then the change in life expectancy can be expressed as $\dot{e}(0,t) = \rho(t)e^\dagger(t)$. The absolute change in $e(0,t)$ can thus be interpreted as a product of the proportion of deaths averted (ρ) and the average number of life-years gained ($e^\dagger$) by those who now survive.

Vaupel and Romo (2003) generalized Keyfitz's formula to the case of age-dependent improvement rates, and suggested that a change in life expectancy at birth be decomposed into two components

$$\dot{e}(0,t) = \int_0^\omega \rho(x,t)w(x,t)\,\mathrm{d}x = \bar{\rho}(t)e^\dagger(t) + \mathrm{Cov}(\rho, e), \tag{4.2.3}$$

where the first term captures the main effect of improvement, while the second term arises due to heterogeneity in $\rho(x,t)$ at different ages. In (4.2.3), $\bar{\rho}$ denotes the average rate of improvement and $\mathrm{Cov}(\rho, e)$ is the covariance between improvement rates and life expectancy, see Vaupel and Romo (2003) for details. Equation (4.2.3) is often taken as the basis for deriving the dynamics of life expectancy sex differentials.

4.2.2 The Rise and Fall of Sex Differentials in Western Europe

The difference in life expectancy at birth between females and males in a given population (the sex gap) is defined as

$$\theta(t) = e_f(0,t) - e_m(0,t) = \int_0^\omega e^{-\int_0^x \mu_f(y,t)\,\mathrm{d}y}\,\mathrm{d}x - \int_0^\omega e^{-\int_0^x \mu_m(y,t)\,\mathrm{d}y}\,\mathrm{d}x, \tag{4.2.4}$$

where the subscripts f and m denote female and male quantities, respectively. In applications we might be interested also in (remaining) life expectancy and sex differential at other ages than 0; all formulas apply, mutatis mutandis, for a general age, but for ease of notation we develop the theory for age 0 only.

Figure 4.1 shows the evolution of the sex gap over time since 1950 in Western Europe. The overall pattern is the same across all countries with the gap being shaped as a countryside hill. That is, the gap initially widened, but has since fallen into a decline with differentials currently around 3–6 years. The timing of the turning point varies by country, occurring first in the United Kingdom circa 1970 and lastly in Spain towards the end of the 1990s.

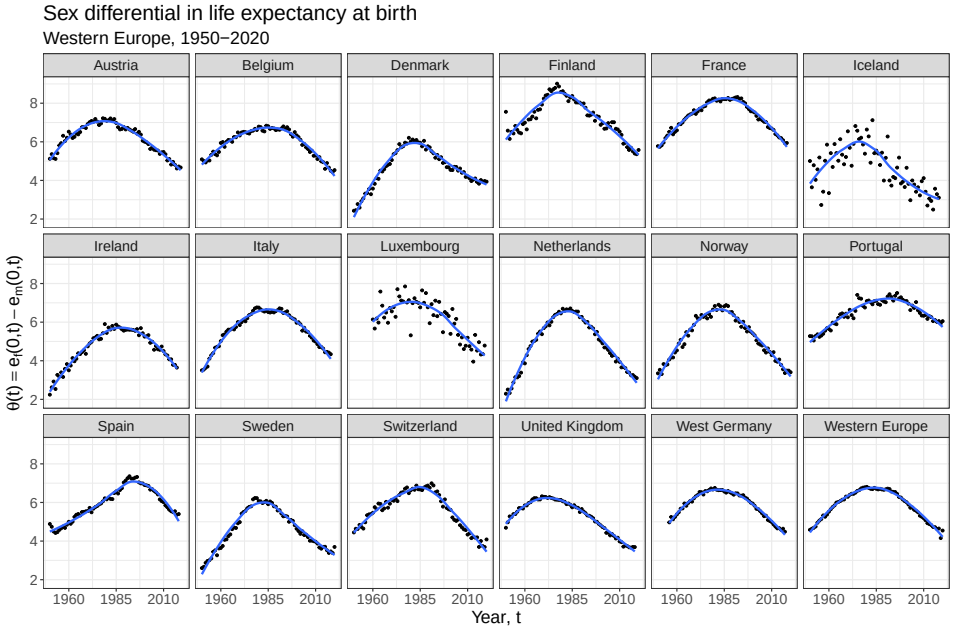

Figure 4.1: Life expectancy sex gap across Western Europe (black dots) with a fitted LOESS curve superimposed (blue line).

The rise and fall of sex differentials observed in the data has recently been studied using demographic decompositions. The principal method of Glei and Horiuchi (2007) separates change in θ into effects due to changing mortality ratios and effects due to changing mortality levels, namely

$$\dot{\theta}(t) = \int_0^\omega \frac{w_m(x,t) + w_f(x,t)}{2} \frac{\dot{c}(x,t)}{c(x,t)} \,\mathrm{d}x + \int_0^\omega [w_m(x,t) - w_f(x,t)] \frac{\dot{\zeta}(x,t)}{\zeta(x,t)} \,\mathrm{d}x, \tag{4.2.5}$$

where $c(x,t) = \mu_m(x,t)/\mu_f(x,t)$ is the mortality sex ratio, $\zeta(x,t) = \sqrt{\mu_m(x,t)\mu_f(x,t)}$ the geometric average of the two mortality rates, and, w_f and w_m are as in

Equation (4.2.2) for females and males, respectively. Note that $\dot{c}(x,t)/c(x,t) = \frac{\partial}{\partial t}\log c(x,t) = \rho_f(x,t) - \rho_m(x,t)$, and $\dot{\zeta}(x,t)/\zeta(x,t) = \frac{\partial}{\partial t}\log \zeta(x,t) = -[\rho_f(x,t) + \rho_m(x,t)]/2$. It follows that the ratio effect is positive when $\rho_f > \rho_m$, while the sign of the level effect depends on the relative size of mortality dispersion for the two sexes (as measured by w_f and w_m); thus, the two effects can work either in the same direction, or against each other.

Using (4.2.5) , Glei and Horiuchi (2007) show that the initial widening of θ seen in Figure 4.1 is caused primarily by women experiencing comparatively larger rates of mortality improvements than men ($\rho_f > \rho_m$), while the subsequent narrowing is largely attributable to differential dispersion between the sexes ($w_m > w_f$) in combination with general improvements. These results are echoed by Cui et al. (2019), who use (4.2.3) to separate $\dot{\theta}$ into three components of change through which they obtain conditions for the sex gap to be widening, narrowing, or at a turning point.

Pollard's Paradox

Glei and Horiuchi (2007) and Cui et al. (2019) both stress that differential dispersion plays a pivotal role in determining changes in θ. In fact, differential dispersion may give rise to somewhat counter-intuitive changes. Pollard (1982), for example, demonstrated that two populations may experience the same absolute change in mortality, but, at the same time, a widening life expectancy differential. The argument is as follows. For $i \in \{f, m\}$, consider an absolute decrease in mortality at all ages, that is, $\widetilde{\mu}_i(x) = \mu_i(x) - \varepsilon$ for some $\varepsilon > 0$. Let $\widetilde{S}_i(x) = S_i(x)e^{\varepsilon x}$ denote the new probability of surviving to age x. Assuming $\mu_f(x) < \mu_m(x)$ for all x, the new life expectancy differential is then larger than before,

$$\theta = \int_0^\omega [S_f(x) - S_m(x)]\,\mathrm{d}x < \int_0^\omega [S_f(x) - S_m(x)]\,e^{\varepsilon x}\,\mathrm{d}x = \theta(\varepsilon,\varepsilon), \tag{4.2.6}$$

where $\theta(\varepsilon_f, \varepsilon_m)$ denotes the life expectancy differential when female mortality is decreased by ε_f and male mortality by ε_m. If we subsequently increase female mortality slightly, we obtain a situation with $0 < \varepsilon_f < \varepsilon$ and $\theta < \theta(\varepsilon_f, \varepsilon)$, i.e., a narrowing mortality differential with a widening life expectancy differential. This phenomenon is sometimes referred to as Pollard's paradox.

In a similar fashion we can argue that a narrowing life expectancy differential does not guarantee narrowing mortality ratios. For instance, taking outset in Keyfitz's formula, when the rate of improvement is the same at all ages, we have

$$\dot{\theta}(t) = \rho_f(t)e^\dagger_f(t) - \rho_m(t)e^\dagger_m(t). \tag{4.2.7}$$

Suppose $e^\dagger_m(t) > e^\dagger_f(t)$, and let $k = e^\dagger_m(t)/e^\dagger_f(t) > 1$. If $\rho_m(t) < \rho_f(t) < k\rho_m(t)$, then $\dot{\theta}(t) < 0$, even though $\rho_m(t) < \rho_f(t)$. This situation could occur if female life

expectancy is close to ω in which case $e^{\dagger}_{f}(t)$ is small, but male life expectancy is not in which case $e^{\dagger}_{m}(t)$ is comparatively larger. In this scenario female mortality can improve at a faster pace than male mortality, but because males benefit from the improvements across a larger span of ages, they gain life expectancy faster than females. Generally, constant mortality ratios can occur together with both increasing and decreasing life expectancy differentials. Therefore, on its own, a narrowing life expectancy differential cannot be interpreted as male mortality rates "catching-up" to female mortality rates.

4.3 Sex Differentials in Coherent Mortality Models

In this section, we present our main mathematical result about sex gap unimodality under coherence. Since coherence implies that the modeled mortality schedules evolve in parallel, it is clear that the forecasted life expectancies approach the same maximal age, or age of truncation, when time approaches infinity. Likewise, it is clear that if we backcast, i.e., "run the model backwards", both mortality schedules will degenerate and the life expectancies converge to zero. Hence, in both limits the sex gap converges to zero. The question is, what happens in between these limits? It might appear obvious that the sex gap will first increase and then decrease, i.e., be unimodal. However, in full generality, this is not true; assuming only coherence, the sex gap can in general exhibit an arbitrary number of modalities. The condition we provide indicates that the mortality schedules have to have the same "shape", in a sense to be made precise later, to guarantee unimodality of the sex gap. Although the mathematical result may not cover most coherent mortality models used in practice, the implications of the result in terms of the location of the peak seems to be valid in much greater generality than proven. The result thereby provides an insight as to why coherent mortality models almost always forecast closing sex gaps.

4.3.1 Coherent Mortality Modeling

In demographic applications it is often required to make forecasts of related populations, e.g., Western, low-mortality countries, or males and females in a given population. Separate forecasting of even very similar populations runs the risk of exaggerating short-term differences leading to diverging projections, but such outcomes seem implausible if the populations have evolved in parallel in the past. For instance, in the case of females and males, we expect the mortality of both groups to keep improving, but we also expect shorter life spans of men relative to women to persist despite converging social and lifestyle factors (Kalben, 2000; Li and Lee, 2005; Zarulli et al., 2018; Jarner and Jallbjørn, 2020). The intuitively appealing property that forecasts of related populations should "stay together" is formalized by the concept of coherence and its use is often motivated by a desire for preserving historic relationships.

Box 4.1. On the definition of coherence

The literature is marked by some confusion regarding the precise, mathematical definition of coherence. Scholars seem to agree on the property as one ensuring non-diverging forecasts, but one can find contradictory definitions depending on the context in which the concept is used. In particular, it is often unclear whether coherence is a property concerning deterministic or stochastic forecasts, especially when authors define coherence as a property related to the mean forecast but apply the concept in a stochastic setting. In the original paper by Li and Lee (2005), coherence was directed at deterministic forecasts, namely "*to avoid long-run divergence in mean mortality forecasts*" (p. 577) by "*imposing shared rates of change by age*" (p. 575). The definition given by Hyndman et al. (2013) is often quoted as the one formalizing coherence and labels "*mortality forecasts as coherent when the forecast age-specific ratios of death rates for any two subpopulations converge to a set of appropriate constants*" (p. 262). This definition, however, seems inappropriate for stochastic models. The proper, mathematical definition of coherence in the spirit of Hyndman et al. (2013) would be to label forecasts as coherent if the age-specific mortality ratio converges to a stationary distribution π, that is,

$$\mu_1(x,t)/\mu_2(x,t) \xrightarrow{d} \pi_x, \text{ for each age } x.$$

In this paper, however, we use the usual deterministic definition given in (4.3.1).

The notion of coherence was introduced by Li and Lee (2005) and later formalized by Hyndman et al. (2013). Given a model for the mortality of two populations, $\mu_i(x,t)$, we let $\bar{\mu}_i(x,t)$ denote the forecast for population $i \in \{1,2\}$. The distinction is, that μ_i is typically a stochastic process, while $\bar{\mu}_i$ is a deterministic forecast, e.g., obtained as the median projection of μ_i. The mortality forecasts are said to be *coherent* if their ratio converges to positive, finite, age-specific constants $c(x)$, that is

$$\frac{\bar{\mu}_1(x,t)}{\bar{\mu}_2(x,t)} \xrightarrow[t\to\infty]{} c(x), \text{ for each age } x. \tag{4.3.1}$$

Forecasts for a group of populations are coherent, if the forecasts are pairwise coherent.

The model proposed by Li and Lee (2005), namely the augmented common-factor model or colloquially the Li-Lee model, models the observed death rate in population i as

$$\log m_i(x,t) = \log \mu_i(x,t) + \varepsilon_{x,t,i} = \alpha_{x,i} + B_x K_t + \beta_{x,i}\kappa_{t,i} + \varepsilon_{x,t,i}, \tag{4.3.2}$$

where K_t and $\kappa_{t,i}$ are stochastic processes modeling common and population-specific secular trends, respectively, and $\varepsilon_{x,t,i}$ is the observation error, i.e., the difference between the underlying mortality rates, $\mu_i(x,t)$, and the observed death rates,

$m_i(x,t)$. Median forecasts are obtained by inserting the estimates for the age-specific loadings, $\hat{\alpha}_{x,i}$, $\hat{B}_x$ and $\hat{\beta}_{x,i}$, and turning off the error terms, i.e., by

$$\log \bar{\mu}_i(x,t) = \hat{\alpha}_{x,i} + \hat{B}_x \bar{K}_t + \hat{\beta}_{x,i} \bar{\kappa}_{t,i}, \tag{4.3.3}$$

where $\bar{K}_t$ and $\bar{\kappa}_{t,i}$ denote median forecasts of the corresponding processes. The model is coherent (i.e., it produces coherent forecasts) when the $\kappa_{t,i}$'s are modeled as stationary, zero-mean processes, e.g., AR(1)-processes. This assumption ensures that each $\bar{\kappa}_{t,i}$ converges to zero, implying that asymptotically all population mortalities are subject to the same age-specific rates of improvements, which is the content of (4.3.1). The Li-Lee model is an archetypical coherent mortality model, and we recall it here to remind the reader of the type of models that we are considering. We return to this model in Section 4.5.

For the rest of the paper we focus on coherent, two-sex mortality models. Of course, mathematically, it makes no difference whether the mortalities are interpreted as sex-specific or not, but with the applications in mind and for ease of presentation, from now on we phrase everything in terms of female and male mortality.

4.3.2 Example: Sex Gap Unimodality for Truncated Exponential Distributions

To gain some intuition for the problem and method of proof, we start with a simple example in which the calculations can be made explicit. Assume female and male mortality are given by $\mu_f(x,t) = \mu/t$ and $\mu_m(x,t) = c\mu/t$, respectively, for $0 \leq x \leq \omega < \infty$ and $t \in (0,\infty)$, where $\mu > 0$ and $c > 1$ are given constants. Hence, the period life times are distributed as truncated exponential variates with life expectancy

$$e(\mu) = \int_0^\omega e^{-x\mu} \, \mathrm{d}x = \frac{1}{\mu}\left(1 - e^{-\omega\mu}\right), \tag{4.3.4}$$

expressed as a functional of the level of mortality. Defining $\theta(t)$ as in (4.2.4), we then have

$$\theta(t) = e(\mu/t) - e(c\mu/t) = \frac{t}{\mu}\left(1 - e^{-\omega\mu/t}\right) - \frac{t}{c\mu}\left(1 - e^{-\omega c\mu/t}\right). \tag{4.3.5}$$

For t tending to zero, both sexes die instantaneously after birth, while for t tending to infinity, both sexes become immortal on $[0,\omega]$. Thus, θ is zero in both limits, while strictly positive for $0 < t < \infty$. Consequently, since θ is smooth, it must have at least one stationary point, i.e., there must exist a t such that $\dot{\theta}(t) = 0$. If we can prove that this is the only stationary point, it follows that θ is unimodal.

Now, note that if θ were to have more than one stationary point they cannot all be local maxima, some of them have to be local minima or points of inflection. From this observation, it follows that if we can prove that all stationary points are

(local) maxima, there can be only one stationary point and we are done. This in turn follows, if we can prove that $\ddot{\theta}(t) < 0$ whenever $\dot{\theta}(t) = 0$. By direct calculations we find

$$\dot{\theta}(t) = \dot{e}(\mu/t) - \dot{e}(c\mu/t) = \frac{\mu}{t^2} \int_0^\omega x \left[e^{-x\mu/t} - ce^{-xc\mu/t} \right] \mathrm{d}x, \tag{4.3.6}$$

and

$$\ddot{\theta}(t) = \ddot{e}(\mu/t) - \ddot{e}(c\mu/t) = \frac{\omega^2 \mu}{t^3} \left(ce^{-\omega c\mu/t} - e^{-\omega\mu/t} \right). \tag{4.3.7}$$

Let t be a stationary point for θ, and define the functions $p(x) = e^{-x\mu/t}$ and $q(x) = ce^{-xc\mu/t}$ for $x \in [0, \omega]$. We want to prove $p(\omega) > q(\omega)$, since that implies $\ddot{\theta}(t) < 0$ by (4.3.7). By assumption, $\dot{\theta}(t) = 0$ and it therefore follows from (4.3.6) that $\int_0^\omega x[p(x) - q(x)]\,\mathrm{d}x = 0$. Since $p(0) = 1 < c = q(0)$ and since p and q can cross at most once (consider the straight lines $x \mapsto \log p(x)$ and $x \mapsto \log q(x)$) we must have that $p(\omega) > q(\omega)$. Because otherwise the integrand would be strictly negative Lebesgue almost surely on $[0, \omega]$, contradicting that the integral is zero. This concludes the proof.

In this specific example we could of course also have investigated the monotonicity properties of θ more directly to arrive at the same conclusion. However, the approach of examining stationary points better extends to the general situation in which explicit expressions for θ are not available.

Note that the existence of a maximum for θ hinges on θ being zero as time tends to infinity. In the example, both sexes experience rectangularization of the survival curve such that their life expectancies converge to the same, finite upper limit. However, if we remove the truncation the situation changes. Without truncation, the sex gap is $\theta(t) = t/\mu - t/(c\mu) = t(c-1)/(c\mu)$ with $\dot{\theta}(t) = (c-1)/(c\mu) > 0$. In this case the sex gap is monotonically increasing to infinity as t tends to infinity. Assuming that an upper limit on life expectancy exists is probably uncontroversial, but it is worth noting that it is this assumption that forces a diminishing sex gap in the limit.

4.3.3 Main Result

We are now ready to state our main mathematical result concerning sex gap unimodality in coherent models. Since our main purpose is to provide insights into the role of coherence; we will only prove the result under the assumption of uniform rates of improvement. More general versions of the result exist, but the conditions become less intuitive, harder to interpret, and tedious to verify.

A twice continuously differentiable function f is called (strongly) unimodal on $A \subseteq \mathbb{R}$, if there exists an $a \in A$ such that f is (strictly) increasing for $t < a$, and

(strictly) decreasing for $t > a$. Let $\mu : [0, \omega] \times \mathbb{R} \to (0, \infty)$ be a twice continuously differentiable function satisfying

$$\lim_{t \to -\infty} \mu(x,t) = \infty, \text{ and } \lim_{t \to \infty} \mu(x,t) = 0. \tag{4.3.8}$$

We think of μ as the backcasted/forecasted mortality surface in a given mortality model, but for ease of notation we leave out the bar over μ that we used in Section 4.3.1. We assume that the relevant time limits are plus/minus infinity, but other limits could also have been used, e.g., zero and infinity as in the example in Section 4.3.2.

Let $\mu_f(x,t) = \mu(x,t)$ and $\mu_m(x,t) = \mu(x,t)c(x)$ denote female and male mortality, respectively. Thus, $\mu_m(x,t)/\mu_f(x,t) = c(x)$ and the mortality forecasts are therefore strongly coherent, in the sense that the limit in (4.3.1) is replaced with an equality. In particular, females and males have the same rates of improvement at all ages and times. As previously, let $\theta(t) = e_f(0,t) - e_m(0,t)$ denote the sex gap. Finally, for $g \in \{f, m\}$, let $I_g(x,t) = \int_0^x \mu_g(u,t)\,\mathrm{d}u$, and let $I_g^{-1}(z,t)$ denote the age, x, for which $I_g(x,t) = z$.

Theorem 4.3.1. *Assume $-\frac{\partial}{\partial t} \log \mu(x,t) = \rho(t) > 0$ for all x, i.e., the rates of improvement is the same for all ages and strictly positive at all times. If $c(x) > 1$ for all x, and*

$$\frac{\partial}{\partial x} \log \left(\frac{\mu_m(I_m^{-1}(x,t),t)}{\mu_f(I_f^{-1}(x,t),t)} \right) \leq 0, \tag{4.3.9}$$

for all $t \in \mathbb{R}$, and all x where the argument is defined, then θ is strictly unimodal on $\mathbb{R}$.

The proof of Theorem 4.3.1 relies on a decision-theoretic argument and is given in Appendix 4.A. As in the example in the previous section, the proof consists in showing that stationary points are local maxima. In brief, the stationarity assumption, $\dot{e}_f(0,t) = \dot{e}_m(0,t)$, is used to construct two probability measures and the monotonicity assumption (4.3.9) is used to show that these two measures are stochastically ordered from which $\ddot{e}_f(0,t) < \ddot{e}_m(0,t)$, and thereby local maximality, can be deduced.

As demonstrated by the counterexample in Section 4.3.4, coherence on its own is not enough to ensure sex gap unimodality. Specifically, exceedingly large jumps in mortality levels can cause sex gap multimodality and, intuitively, the role of (4.3.9) is to prevent such jumps. Increases in mortality levels are revealed by the cumulative death rate $I(x)$, which can be interpreted as the expected number of deaths that would have occurred at age x had the event been repeatable. Thus $I^{-1}(x)$ is the age at which we would have experienced x deaths. Equation (4.3.9) looks at the rate of change in log-mortality differences, but with the age input transformed by $I^{-1}(\cdot)$. Because female mortality is lower than male mortality, $I_f^{-1}(x)$ will be higher than

$I_m^{-1}(x)$. Loosely speaking, condition (4.3.9) states that "high age" female mortality must increase faster than "low age" male mortality. This is typically, but not always, satisfied. As age and thereby mortality increases, $I^{-1}(x)$ flattens. If there is a sharp transition in the age-specific mortality curve, this flattening occurs at a comparatively much lower x-value for females than for males and means that (4.3.9) is comparing μ_f evaluated at a high age to μ_m evaluated at a much lower age. Figure 4.2 shows $I^{-1}(x)$ and $\mu(I^{-1}(x))$ in the two-level mortality example from Figure 4.3 and for two Gompertz mortality curves. Condition (4.3.9) is violated in the former case as $\mu(I^{-1}(x))$ does not "jump" at the same time for both sexes, turning the left-hand side of (4.3.9) positive once the male rate "jumps". Although counterexamples as these can be constructed, in practice, (modelled) female and male mortality schedules are sufficiently aligned that (4.3.9) is satisfied.

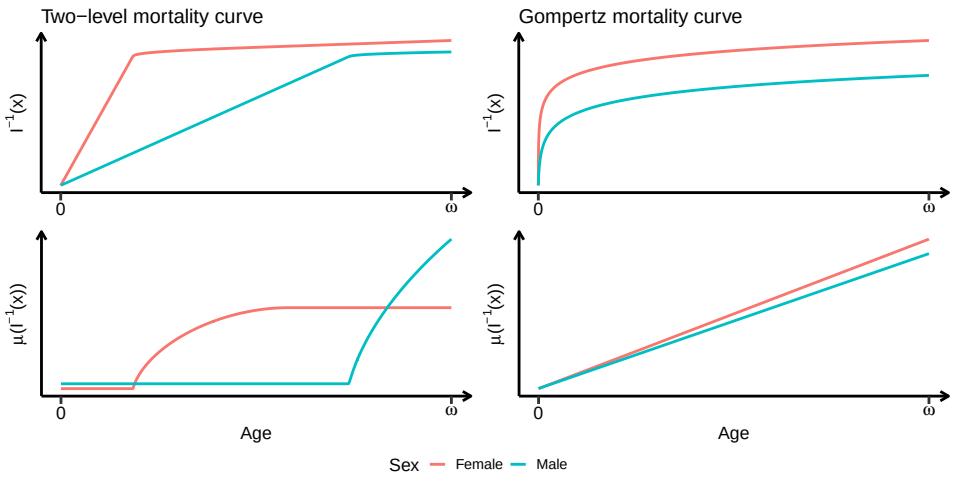

Figure 4.2: Illustration of $I^{-1}(x)$ and $\mu(I^{-1}(x))$ for the two-level mortality curve depicted in Figure 4.3 (left panels) and a Gompertz mortality curve (right panels).

Theorem 4.3.1 formalizes Pollard's paradox; under the stated conditions we have a situation with fixed mortality ratios and a sex gap that is both widening and narrowing, in two distinct epochs. In the terminology of Section 4.2, both the initial widening and the subsequent narrowing of the sex gap are caused by the level effect, since the ratio effect is absent by construction. Moreover, it is possible to characterize the point where the sex gap peaks in terms of the life expectancy of one of the sexes, e.g., females. This will be illustrated in Section 4.4.

4.3.4 Counterexample

The sex gap is not unimodal for arbitrary mortality profiles. As a counterexample, consider a two-level piecewise constant mortality curve made continuous through

quadratic interpolation

$$\mu(x) = \begin{cases} \mu & , \ x \in [0, x_0), \\ \mu\delta[3(x-x_0)^2/\varepsilon^2 - 2(x-x_0)^3/\varepsilon^3] & , \ x \in [x_0, x_0+\varepsilon], \\ \mu\delta & , \ x \in (x_0+\varepsilon, \omega], \end{cases} \tag{4.3.10}$$

for fixed constants $\mu, \delta, \varepsilon > 0$. That is, suppose the mortality schedule between ages 0 and x_0 is constant at level μ, but is "bumped" to a new level $\mu\delta$ over the age range x_0 to $x_0 + \varepsilon$ for some large δ and small ε. The left panel in Figure 4.3 shows an example mortality curve of this form, while the right panel shows the resulting life expectancy differential when mortality is subject to the same fractional improvement in the age-specific death rate at all ages; corresponding to $\mu(x,t) = \mu(x)\exp(-\rho t)$ for some $\rho > 0$. For reference, the unimodal curve shown as the dashed line is the life expectancy differential had $\omega = x_0$.

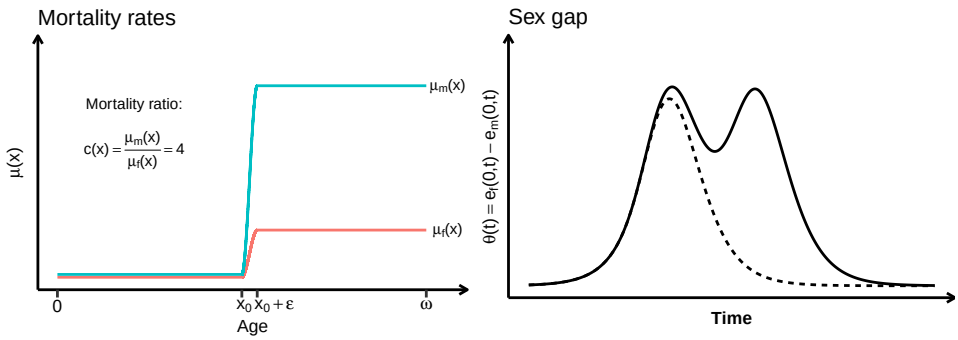

Figure 4.3: Mortality rates and the resulting life expectancy differential over time when the same, positive mortality improvement rate is applied to both sexes. In the right panel, the superimposed dashed line is the sex gap had the life expectancy calculation been truncated at x_0. The curves are calculated using $\mu(x) = \mu$ when $x \in [0, x_0)$ and $\mu(x) = \mu\delta$ when $x \in (x_0 + \varepsilon, \omega]$ with $\mu = e^{-1}$, $x_0 = 60$, $\delta = 50$, $\varepsilon = 5$, and $\omega = 120$. Continuity between x_0 and $x_0 + \varepsilon$ is achieved by quadratic interpolation as in (4.3.10). The means of interpolation is not important for the calculation; using any C^2-interpolating curve or applying a bump function to make $\mu(x)$ smooth and continuous yields virtually the same result.

The gap takes on a multimodal ("roller-coaster") shape. The additional humps are created by the jump in mortality after age x_0, which creates a false life expectancy barrier (if δ is sufficiently large). This barrier is broken by the females first, whereby the life expectancy differential starts to rise again, creating another modality. Thus, even with monotonically increasing hazards and $\mu_m(x,t)/\mu_f(x,t) > 1$, unimodality is not guaranteed. The example could be extended to create any number of modes by introducing more barriers. It is clear, however, that counterexamples are necessarily somewhat contrived and that they do not arise in "normal" situations.

4.4 The Dynamic Gompertz Model

Parametric models can be used for mortality projections by letting calendar year enter the model through its parameters. That is, for each calendar year the parameters of the model are estimated and then viewed as stochastic processes, forecasted by standard time series methods. By linking the stochastic processes driving the models of different populations, coherence and other dependency structures can be achieved, see e.g. Jarner and Kryger (2011), Cairns et al. (2011b), and Jarner and Jallbjørn (2020). In this section we will analyze a particularly simple, coherent mortality model of this type. The model is too simple to be of practical use, but it is useful as an illustration of the sex gap trajectories implied by coherence.

4.4.1 The Model

The Gompertz mortality law prescribes that mortality increases exponentially,

$$\mu(x) = e^{\alpha+\beta x}, \quad 0 \le x \le \omega. \tag{4.4.1}$$

This mortality profile is a remarkably good fit to adult mortality from age 20, say, onwards. For younger ages, and in fact also for very old-ages, the profile is not appropriate. A simple two-sex model for adult mortality can be constructed by fitting a Gompertz law to period, sex-specific mortality data resulting in estimated coefficients $(\alpha_{t,g}, \beta_{t,g})$ for $g \in \{f, m\}$ and t ranging over the years in the estimation window.

If the parameters are modeled as random walks with the same drift for both sexes, the median forecast for population g becomes

$$\bar{\mu}_g(x, T+h) = \exp\left\{\alpha_{T,g} + \beta_{T,g}x + h(\xi_\alpha + \xi_\beta x)\right\}, \tag{4.4.2}$$

where T denotes the projection jump-off year, $h \in \mathbb{N}_0$ is the forecasting horizon, and ξ_α and ξ_β are the (shared) drift terms of the α- and β-processes, respectively. By design, the forecasts in (4.4.2) are (strongly) coherent since the mortality sex ratios do not depend on h,

$$\frac{\bar{\mu}_m(x, T+h)}{\bar{\mu}_f(x, T+h)} = \exp\left\{(\alpha_{T,m} - \alpha_{T,f}) + (\beta_{T,m} - \beta_{T,f})x\right\} = c(x). \tag{4.4.3}$$

More complicated time-series models could also be used, but the current structure suffices for our purposes. Note that although h is assumed non-negative, we can evaluate (4.4.2) for all values of h. Specifically, we can think of a forecast as the right tail of an implied surface obtained by letting h range over all integers, negative and positive.

4.4.2 Sex Gap Unimodality Under Uniform Rates of Improvement

Assuming uniform rates of improvement, the continuous-time analog of the dynamic Gompertz model is given by

$$\mu(x,t) = e^{\alpha+\beta x-\rho t}, \quad 0 \le x \le \omega, \tag{4.4.4}$$

with level parameter α, slope parameter $\beta > 0$, and rate of improvement $\rho > 0$. As in Section 4.3.3, let $\mu_f(x,t) = \mu(x,t)$ and $\mu_m(x,t) = \mu(x,t)c(x)$, where $c(x) = \exp(\Delta_\alpha + \Delta_\beta x)$, with $\Delta_\beta > -\beta$, such that μ_m is also of form (4.4.4) with positive slope. Thus, mortality for both sexes follow a Gompertz mortality schedule with, in general, differing levels and slopes.

The cumulative death rate is

$$I(x,t) = \int_0^x \mu(u,t)\,\mathrm{d}u = e^{\alpha-\rho t}(e^{\beta x}-1)/\beta, \tag{4.4.5}$$

while its inverse for fixed time t is $I^{-1}(x,t) = \log(1 + x\beta/(e^{\alpha-\rho t}))/\beta$, whereby

$$\mu(I^{-1}(x,t),t) = e^{\alpha-\rho t} + \beta x. \tag{4.4.6}$$

It follows that

$$\frac{\partial}{\partial x}\log\left(\frac{\mu_m(I_m^{-1}(x,t),t)}{\mu_f(I_f^{-1}(x,t),t)}\right) = \frac{e^{\alpha-\rho t}\left[\Delta_\beta + \beta(1-e^{\Delta_\alpha})\right]}{[e^{\alpha-\rho t}+\beta x]\,[e^{\alpha+\Delta_\alpha-\rho t}+(\beta+\Delta_\beta)x]}. \tag{4.4.7}$$

In practice, we typically find excess male mortality relative to female mortality with $\Delta_\alpha > 0$ and $\Delta_\beta < 0$, that is, the level is higher for males than for females, but the curve is less steep. In this situation, the right-hand side of (4.4.7) is negative, since the numerator is negative while the denominator is positive. It then follows from Theorem 4.3.1 that the sex gap is strictly unimodal.

We note that Theorem 4.3.1 only provides sufficient conditions for unimodality. For instance, having $c(x) < 1$ at high ages does not imply that unimodality of the gap will not occur. In fact, using empiric estimates we typically have this situation, but still the sex gap is unimodal.

4.4.3 Sex Gap Trajectories

We define the sex gap trajectory as the pairs of female life expectancy and sex gap that can occur together on a given two-sex mortality surface,

$$\mathcal{T} = \{(e_f(0,t),\theta(t)) : t \in \mathbb{R}\}, \tag{4.4.8}$$

where as usual $\theta(t) = e_f(0,t) - e_m(0,t)$ denotes the life expectancy differential (sex gap). In the continuous-time version of the dynamic Gompertz model, we have

$$\mu_f(x,t) = e^{\alpha-\rho t}e^{\beta x}, \quad \mu_m(x,t) = e^{\alpha-\rho t}e^{\beta x}e^{\Delta_\alpha+\Delta_\beta x}. \tag{4.4.9}$$

Since $\rho > 0$, we see that as t goes from minus infinity to plus infinity the common factor, $\exp(\alpha - \rho t)$, takes all values in $(0, \infty)$, regardless of the value of α and ρ. It follows that for the dynamic Gompertz model the sex gap trajectory is a function of the female slope parameter and excess mortality only, i.e., $\mathcal{T} = \mathcal{T}(\beta, \Delta_\alpha, \Delta_\beta)$.

Based on mortality data from Western Europe 1950–2020, estimates of β are in a narrow range around 0.10, while excess mortality shows greater variability from country to country and over time. Figure 4.4 shows two sets of trajectories for typical values of β; the left plot corresponds to the average male excess mortality profile in the data, and the right plot corresponds to very high male excess mortality, as seen in, e.g., Finland.

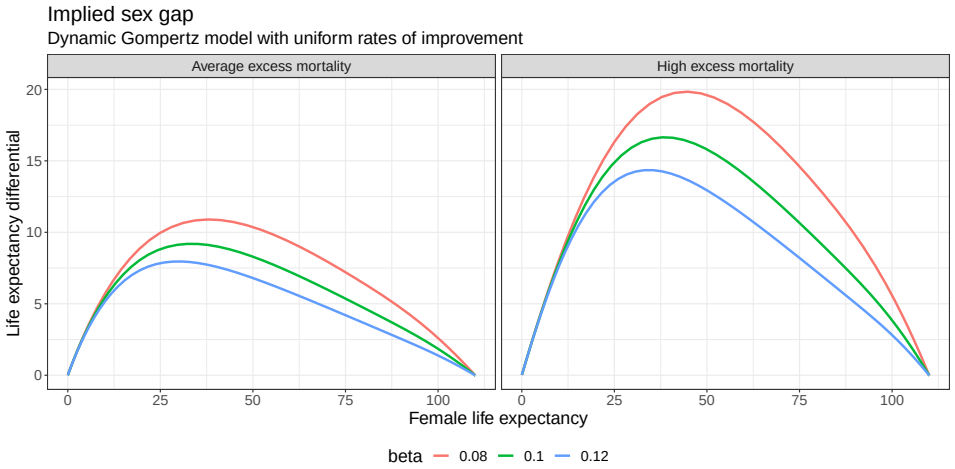

Figure 4.4: The sex gap trajectory implied by a dynamic Gompertz model for different values of the female slope parameter, β, for average and high excess male mortality. Left plot has $\Delta_\alpha = 1.25$ and $\Delta_\beta = -0.012$, right plot has $\Delta_\alpha = 2.2$ and $\Delta_\beta = -0.021$.

We notice that although the size of the sex gap is different in the two plots, the overall shape of the trajectory is almost the same in all cases. In particular, we notice that the sex gap peaks when female life expectancy is in the range 30 to 50 years. *This means that if a dynamic Gompertz model is fitted to data and forecasts are produced from a jump-off year where female life expectancy is larger than 50 years then the model will forecast a closing sex gap*; at least if the slope parameters are relatively constant (i.e., if $\xi_\beta \approx 0$).

Note that the speed with which the sex gap trajectory is traversed in the forecast depends on how fast female life expectancy evolves. We only know that as female life expectancy increases beyond 50 years, the sex gap closes. Since female life expectancy will not increase linearly, the shape of the sex gap in the forecast will typically not resemble the shape of the sex gap trajectories in Figure 4.4.

That the coherent, dynamic Gompertz model typically produces narrowing sex

gaps also for age-dependent rates of improvement, is illustrated in Figure 4.5. The forecasts are produced by (4.4.2) for dynamic Gompertz models fitted to different countries in different periods. The Gompertz models are calibrated using data in the age range $\{20, \ldots, 100\}$ to avoid the poor fit at younger ages to influence parameter estimates; for the same reason we look at the life expectancy differential at age 20, instead of at birth. Parameters of the Gompertz models are estimated by maximum likelihood assuming independent, Poisson-distributed death counts, $D(x,t)|E(x,t) \sim \text{Pois}(E(x,t)\exp(\alpha_t + \beta_t x))$. The model's parameters are calibrated to the latest 30 years of data available before projection jump-off. If data does not exist for all 30 years, only available data is used. We estimate and project the models for all countries in Western Europe and for periods with jump-off in 1960, 1980, 2000, and 2020, respectively.

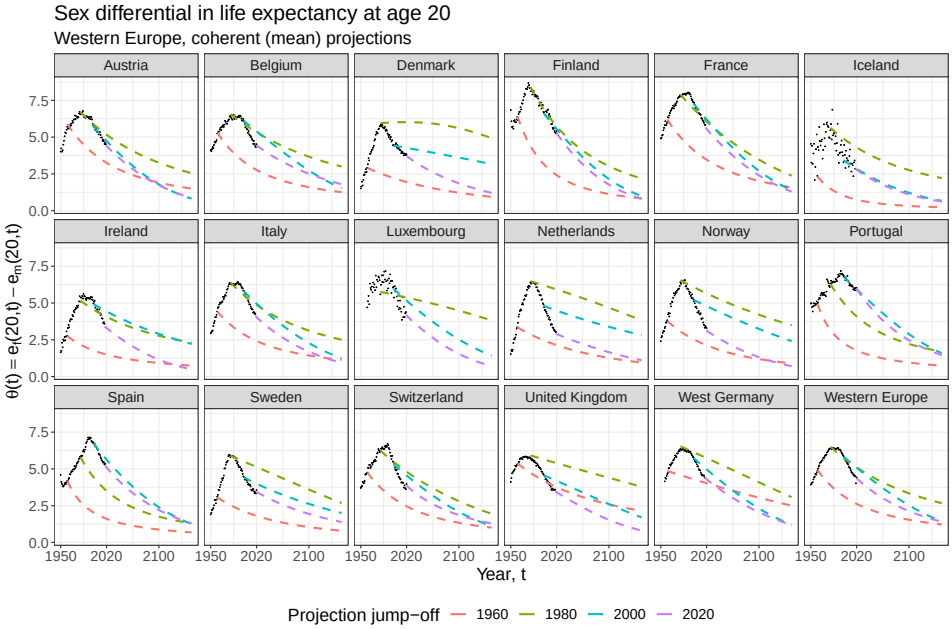

Figure 4.5: Forecast of the life expectancy sex gap in Western Europe using a dynamic Gompertz model with parameters forecasted as a random walk with drift, cf. Equation (4.4.2), and varying jump-off years.

As seen, almost all forecasts in Figure 4.5 produce narrowing gaps. In some periods the forecasts provide a good description of the future, e.g., the recent forecasts for Finland, France, Portugal and Spain, but in most cases the forecasted sex gap does not provide a sensible continuation of the historic trend. The problem is, that the behavior of data leading up to the jump-off year matters very little; it is the assumed coherence in the forecasts that produces the (implied) unimodal sex gap trajectory and it is the fact that female life expectancy is larger than 50 years at all jump-offs that places us on the declining part of that trajectory.

We note in closing, that the dynamic Gompertz model does not always forecast closing sex gaps. Periods with little or no mortality improvements for males (Denmark, 1980), or male mortality being very close to female mortality can lead to an increasing sex gap at jump-off. In the latter case, the peak of the implied sex gap trajectory can occur at a much higher age than indicated in Figure 4.4, cf. Box 4.2 "The outlier Ireland" for an example of this rarely happening situation.

4.5 The Forecast of Closing Sex Gaps by Coherent Mortality Models

In the previous section we demonstrated that the dynamic Gompertz model typically forecasts closing sex gaps, both under uniform rates of improvement, as predicted by Theorem 4.3.1, but also under age-dependent rates of improvement. In this section we show that more "realistic", coherent mortality models behave qualitatively similar to the dynamic Gompertz model. On this basis, we conclude that closing sex gap is indeed a general feature of coherent models. The analysis is in two parts. First, we relax the parametric structure of the Gompertz curve but keep the uniform rate of improvement. Second, we show closing sex gaps for coherent, semi-parametric models with age-dependent rates of improvement.

4.5.1 Location of the Sex Gap Zenith under Uniform Rates of Improvement

Based on the dynamic Gompertz model, we concluded in Section 4.4.3 that for Western European data the implied sex gap peaks at an earlier age than the observed female life expectancy, leading to closing sex gap forecasts. One might object to this analysis that the Gompertz model allows for only a very limited set of mortality profiles, and that the conclusion implicitly rests on this fact. Below we extend the analysis to general, smooth mortality profiles, retaining the assumption of uniform rates of improvement.

Consider a (strongly) coherent mortality surface with female and male mortality of the form $\mu_f(x,t) = \mu_f(x,T)\exp(-\rho(t-T))$ and $\mu_m(x,t) = \mu_f(x,t)c(x)$, respectively, where T is a fixed year, $\mu_f(x,T)$ is the female mortality at age x in year T, $c(x)$ is the male excess mortality at age x, and $\rho > 0$ is the (uniform) rate of improvement.

Recall from Equation (4.4.8) that the sex gap trajectory is defined as the set of female life expectancies and sex gaps that can occur together when t varies. By the same argument as in Section 4.4.3, this set does not depend on the rate of improvement. In other words, assuming a uniform rate of improvement, the sex gap trajectory depends only on the female mortality profile in any given year and the excess mortality profile. In particular, for any female mortality profile, $\mu_f(\cdot,T)$, and

any c-profile, we can define the (implied) sex gap zenith,

$$(a_{\max}, \theta_{\max}) = (e_f(0, t_{\max}), \theta(t_{\max})), \text{ where } t_{\max} = \arg\max \theta(t), \tag{4.5.1}$$

as the female life expectancy when the sex gap is at its maximum and the corresponding (maximal) value of the sex gap. The notation $(a_{\max}, \theta_{\max})$ is chosen to reflect the abscissa and ordinate at the maximum point, cf. Figure 4.4.

We compute and compare $(a_{\max}, \theta_{\max})$ for (i) a Gompertz model and (ii) a graduated mortality curve, where the mortality curves obtained by graduation closely represent the true underlying death rates. In short, the graduation procedure smooths the empirical death rates by a cubic smoothing spline, while the Kannisto model of old-age mortality is used at ages 80 and above. Further details are provided in Appendix 4.B.

The sex gap zenith is presented in Table 4.1 for the two models for the countries in Western Europe. For given year, mortality data for that year only is used to estimate female and male mortality profiles from which the c-profile is derived. Next, assuming a uniform rate of improvement, the implied sex gap trajectory (4.4.8) can be computed. Finally, the female life expectancy and the size of the sex gap at the zenith (4.5.1) of the trajectory are found. The computation is performed on data from three calendar years that captures the different epochs of the period. In the first year, 1950, all countries are on the widening part of the observed θ-curve. The second year, 1985, is the middle of the period where the observed θ-curve peaks, while the third year, 2020, is on the narrowing part of the curve, cf. Figure 4.1. If a country does not have data at one or both of the endpoints, the nearest data point is used instead.

Table 4.1: Female life expectancy ($a_{\max}$) when the implied sex gap trajectory is at its maximum ($\theta_{\max}$). For each country and model, the year refers to the data used to compute the implied sex gap trajectory, see main text for details. Life expectancies are truncated at the age of 110.

		Gompertz mortality curve						Graduated mortality curve					
		$a_{\max}$			$\theta_{\max}$			$a_{\max}$			$\theta_{\max}$		
Country	Data avail.	1950	1985	2020	1950	1985	2020	1950	1985	2020	1950	1985	2020
1. Austria	1947–2019	33.2	34.5	32.1	6.9	13.5	11.5	47.5	43.8	41.4	5.3	10.6	7.8
2. Belgium	1841–2020	32.8	35.0	31.8	7.6	11.5	9.8	43.0	45.0	52.0	6.2	8.3	6.2
3. Denmark	1835–2020	28.0	36.7	32.0	3.5	8.4	6.7	27.5	33.8	29.8	4.5	8.0	8.5
4. Finland	1878–2020	35.6	36.3	33.8	12.2	15.6	14.4	49.2	49.2	45.7	8.9	12.2	11.0
5. France	1816–2018	35.4	36.8	34.7	7.7	15.4	13.4	54.3	56.0	58.6	5.9	10.9	8.6
6. Iceland	1838–2018	34.1	33.9	29.9	8.9	13.0	8.8	37.6	44.5	43.1	9.2	9.4	6.2
7. Ireland	1950–2017	87.9	35.8	30.7	2.0	8.1	8.1	73.7	36.0	37.9	1.7	8.4	8.9
8. Italy	1872–2018	31.7	34.3	31.3	5.6	12.1	9.8	48.1	44.1	39.0	4.0	9.1	7.4
9. Luxembourg	1960–2019	32.1	35.8	31.8	9.8	10.0	9.5	32.4	38.3	31.3	9.1	10.3	10.1
10. Netherlands	1850–2019	28.0	36.2	30.9	3.7	9.4	5.2	30.9	66.2	37.7	4.6	7.0	4.7
11. Norway	1846–2020	29.7	34.8	30.1	5.4	12.2	7.4	34.3	39.3	40.3	6.2	10.0	7.6
12. Portugal	1940–2020	38.0	34.8	34.3	7.2	12.9	15.5	55.9	43.6	62.0	5.0	11.2	9.2
13. Spain	1908–2018	37.8	34.5	33.3	6.4	13.8	13.3	54.5	47.6	64.8	4.9	10.4	7.7
14. Sweden	1751–2020	28.8	34.3	30.3	4.0	10.1	7.1	29.6	51.5	37.8	5.5	7.3	8.4
15. Switzerland	1876–2020	32.1	34.8	30.3	6.7	11.7	9.2	41.2	42.9	38.3	5.7	9.6	7.9
16. United Kingdom	1922–2018	38.1	39.3	31.4	5.2	6.9	8.0	61.5	35.2	42.2	4.5	6.6	6.7
17. West Germany	1956–2017	31.5	34.7	32.0	8.7	11.4	10.4	39.2	54.4	55.1	7.2	8.1	6.0

Box 4.2. The outlier Ireland

The only outlier in Table 4.1 is Ireland, 1950. In this year, the observed life expectancy for females is 66.7 years, but according to Table 4.1, this is well below $\hat{e}_0^f$. Ireland is rather unique in a demographic context. Men outlived women at the beginning of the 20th century, and by the 1930s, Ireland was the only country in the West with higher survival rates for men than for women at any age (Coleman, 1992). This pattern has taken a while to reverse, and even though female mortality is lower than male mortality in 1950 (at most ages), the curves lie almost directly on top of each other as seen in the figure below. Interestingly, this fact makes the sex gap widen for a prolonged period, that is, the slope of θ is rather moderate before reaching the turning point. Once the turning is passed, however, the gap will close quite rapidly; the slope on the narrowing part is much steeper compared to the slope on the widening part of the curve as seen in the figure below. This feature needs to be understood in the context of life span disparity, $e^\dagger$, and the maximum attainable age, ω. Because the mortality profiles are similar not only in slope but also in level, $e_f^\dagger$ stays higher than $e_m^\dagger$ for longer. Female life expectancy at the turning point, $a_{\max}$, is, therefore, closer to ω than had the mortality levels been further apart, causing the subsequent "catch-up" to happen much faster.

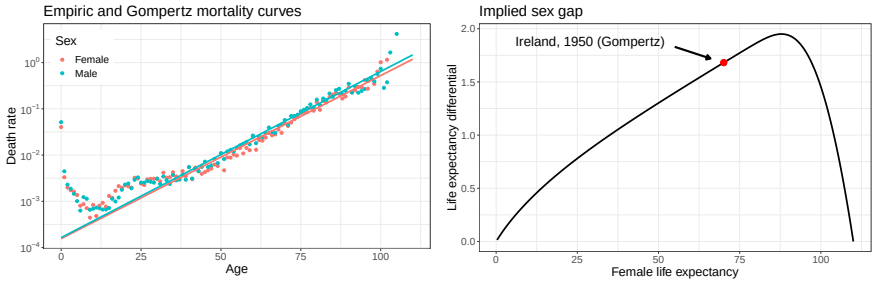

Figure: Mortality and implied sex gap for Ireland, 1950. In the left panel, the Gompertz fit is superimposed as solid lines.

The Gompertz model predicts a female life expectancy of around 30–40 years when the sex gap is at its zenith, varying slightly from country to country and period to period. For comparison, observed female life expectancy in 1950 is between 61.0 (Portugal) and 73.6 (Iceland) and increasing over the period. Thus, all countries are on the declining part of the implied θ-curve, and continued improvements in mortality will further narrow the gap. The only exception is Ireland, see Box 4.2.

Because the Gompertz curve constitutes a highly stylized mortality profile, the life expectancy predictions from this model are fairly consistent across the different countries and periods. The graduated mortality curve is more flexible, resulting

in more diverse predictions, in particular, regarding the female life expectancy at the sex gap zenith. The overall result is, however, qualitatively the same; with the exception of Ireland, the implied sex gap peaks when female life expectancy is below that observed at the estimation year (not shown). The takeaway message is that for essentially all countries and jump-off years, both models project narrowing sex gaps.

4.5.2 Coherence Implies Closing Sex Gaps

Until now we have demonstrated sex gap unimodality and closing sex gap forecasts under conditions imposed for mathematical tractability. In particular, strong coherence and uniform rates of improvement. Of course, coherent mortality models used in practice are neither strongly coherent, nor do they have uniform rates of improvement. Nevertheless, the conclusion extends to "realistic", coherent models also, and for the same reason. Calibrated to Western European mortality data, the sex gap implied by the models peaks when female life expectancy is (much) lower than levels observed after the Second World War.

Table 4.2 presents the slope of the sex gap following projection jump-off based on the dynamic Gompertz model (4.4.2), the Li-Lee model (4.3.2), and the product ratio model of Hyndman et al. (2013) for different jump-off years. For the Li-Lee model (4.3.2), we use AR(1)-processes to describe the sex-specific time-varying indices. The product ratio model is fitted and forecasted using the `Demography` package (available on CRAN) in R.

Table 4.2: Direction of the sex gap five years after projection jump-off for different coherent models and jump-off years. The table shows whether the gap is narrowing (↘) or widening (↗). The models are calibrated to data in the period 1950 (or the earliest year where data is available) to the listed jump-off year. No direction is shown if the model cannot be calibrated (i.e., lack of data).

	Gompertz model					Li-Lee model					The Product-Ratio model				
Country	1960	1965	1970	1975	1980	1960	1965	1970	1975	1980	1960	1965	1970	1975	1980
1. Austria	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	↗
2. Belgium	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
3. Denmark	↗	↗	↗	↗	↖	↗	↗	↗	↗	↗	↗	↗	↗	↖	↗
4. Finland	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
5. France	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
6. Iceland	↗	↗	↗	↗	↗	↖	↖	↗	↗	↗	↖	↗	↗	↗	↗
7. Ireland	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
8. Italy	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↖	↗
9. Luxembourg	–	↖	↖	↖	↗	–	↖	↗	↖	↖	–	↗	↖	↗	↖
10. Netherlands	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
11. Norway	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
12. Portugal	↗	↗	↗	↗	↗	↗	↗	↗	↖	↗	↗	↗	↗	↗	↗
13. Spain	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
14. Sweden	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
15. Switzerland	↗	↗	↗	↗	↗	↗	↗	↗	↗	↖	↗	↗	↗	↗	↗
16. United Kingdom	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
17. West Germany	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	–	↗	↖	↗	↗

Except for the United Kingdom, the actual sex gap was widening at all jump-off years, cf. Figure 4.1. Nevertheless, essentially all forecasts predict a narrowing sex gap. Further, for those (few) forecasts where the gap is projected to widen, the

historical trend is not reproduced. Rather, the apex of θ is predicted to be in the near future with a gap that remains approximately constant on a short horizon (not shown); the forecast trajectories resemble that of Denmark, 1980 seen in Figure 4.5 (green, dashed line).

In semi-parametric mortality models, e.g., the Lee-Carter model (Lee and Carter, 1992) or the Li-Lee model (Li and Lee, 2005), the fit to data is typically good due to a large number of parameters. Even parsimonious model structures such as the dynamic Gompertz model can capture many observed mortality patterns when parameters are allowed to vary freely. Consequently, in the estimation window, most mortality models allow flexible, if not freely varying, rates of improvement, $\rho(x,t)$. In the forecasting region, however, the improvement rates are often, directly or indirectly, constrained by (i) assuming temporal constancy, e.g., in models of the Lee-Carter type, and (ii) by imposing coherence:

$$\rho(x,t) \xrightarrow{\text{Time invariance}} \rho(x) \xrightarrow{\text{Coherence}} \rho_f(x) \simeq \rho_m(x).$$

Both assumptions might be at odds with (recent) trends in data, resulting in death rate (median) trajectories that may not conform with those observed in the past. Relaxing the assumption of time-invariant improvement rates can be achieved by various techniques, for example, by imposing convergence to a long-term target (Li, 2013), or applying frailty theory (Jarner and Jallbjørn, 2022). This, however, is not the focus of the present paper.

Although, formally only a condition in the limit, the coherence assumption built into many multi-population mortality models all but eliminates the ratio effect, identified by Glei and Horiuchi (2007) and Cui et al. (2019) as the main driver of the widening sex gap in Western Europe until the 1980's, cf. Section 4.2.2. In periods where the sex gap is narrowing, the ratio effect is less important and the coherence assumption agrees more with observed data. Narrowing sex gap projections are produced as intended but are not guaranteed to replicate the latest trends, cf. Figure 4.5. In summary, the implied sex gap trajectory of coherent models, being driven mainly by the level effect, is too inflexible to adequately describe and forecast the sex gap evolution since 1950.

4.5.3 Does Coherence Deserve Its Special Status?

The fact that the sex gap narrows in coherent projections at essentially all levels of mortality observed in the West over the last 70 years challenges the desirability of coherence as a modeling goal and adds to the recent criticism of coherence raised by Hunt and Blake (2018) and Jarner and Jallbjørn (2020). At the time of its development, coherence coincidentally appeared as a desirable property that (to some extent) continued the narrowing of the sex gap that had been observed for some time, unlike most independent methods that projected widening or diverging gaps, see

for example Figure 5 in Hyndman et al. (2013). One can understand that Hyndman et al. (2013) and others concluded that coherent forecasts were an improvement over independent forecasts, but this verdict seems somewhat anchored in the sex gap decline of the time. Imagine standing in a year between 1950–1980, observing sharp transitions from widening to narrowing gaps such as those in Figure 4.5. It then seems far less obvious that coherence is a property one should impose on a model and certainly questions whether coherence deserves the special status it has been given.

From the adversarial perspective, it is—at least in theory—possible for coherent models to accommodate temporary ratio effects of varying (but bounded) size and thereby produce trends in the sex gap that are in keeping with those recently observed. In most applications, however, the ratio effect diminishes quickly, leading to a misalignment between historic and projected trends as demonstrated in Table 4.2. Arguably, a diminishing ratio effect is intentional, because otherwise, the rationale for imposing coherence would be undermined. Even though coherence may be a sensible restriction to impose when extrapolating the present mortality regime (for some populations), it does not seem to be so historically and may not be so in the future either. In our view, this suggests a revision of coherence as the guiding principle for multi-population mortality modelling.

Moreover, as is also pointed out by Jarner and Jallbjørn (2020), coherence seems too tailored to the log-linear common factor models for which the property was introduced. Even though single population models can be coupled via cointegration techniques to achieve coherence, the choice of scale, i.e., ratios of mortality rates, remains somewhat arbitrary and effectively restricts the coherence label to models of the Lee-Carter type. Therefore, we see a need for a broader definition that covers a larger class of models and permits less restrictive dependency structures.

As a minimum, the coherence requirement should be relaxed to allow for observed patterns of covariation to continue in forecasts. A possible approach to tackle this issue could be to redefine coherence into a property that concerns modeling on the parameter scale rather than modelling on the (log) data scale. In particular, it seems more natural to identify cointegrating relations between (time-varying) parameters, i.e., identifying linear combinations of the parameters that are stationary, rather than requiring mean-reversion of mortality log differences. For further discussion on this point, see Jarner and Jallbjørn (2020).

4.6 Conclusion

The notion of coherence has been one of the most influential ideas in multi-population mortality modeling. When projecting groups of related populations, e.g., males and females, or countries of similar affluence, which have evolved in parallel in

the past, it is natural to expect these populations to "stay together" in the future also. Coherence solves the problem of diverging, or crossing, forecasts that can arise when projecting, even very similar, populations separately. It does so by requiring converging mortality ratios, or, equivalently, asymptotically equal rates of improvement at each age. At first sight, this seems an innocent and reasonable requirement, but in practice coherent models enforce a rigid structure on the forecasts that can be at odds with trends in data.

In this paper we discussed the implications of coherence in two-sex mortality models with focus on the dynamics of the sex differential in life expectancy (sex gap). We provided both theoretical and empirical evidence to support the conclusion that coherent models forecast closing sex gaps for Western European countries for almost all jump-off years since 1950. Despite the fact that the actual sex gap was widening until the 1980s. Coincidentally, coherence was introduced after almost 20 years of narrowing sex gaps, and a continued sex gap closing was seen as a desirable feature of coherent models. However, the inadequacy of coherent models in the first half of the period from 1950 till today, lead us to question coherence as a general modelling principle.

Technically, we prove in the paper that strongly coherent models with a uniform rate of improvement produce a unimodal sex gap trajectory. Further, we demonstrate that for Western European levels of male excess mortality (relative to female mortality) the implied sex gap trajectory typically peaks when female life expectancy is in the range 30 to 50 years. Although formally proven only for a specific subclass of coherent models, this insight applies in much greater generality and it explains why forecasts from coherent two-sex models are almost always on the declining part of the trajectory.

We also discussed the effects responsible for the observed widening and narrowing sex gap in Western Europe from 1950 till today. In particular, with appeal to Glei and Horiuchi (2007) we identified changing sex ratios (the ratio effect) as the main driver of the widening sex gap, and differential disperson combined with general improvements (the level effect) as the main driver of the subsequent narrowing of the sex gap. In light of the similarity between the observed and implied sex gap trajectories in Figure 4.1 and Figure 4.4, respectively, it is a priori surprising that coherent models cannot be used to describe the observed sex gap dynamics over the entire period. However, the implied sex gap trajectory peaks too early, in terms of female life expectancy, and therefore cannot be aligned with the widening part of the observed sex gap trajectory. Moreover, even on the narrowing part of the observed sex gap trajectory, coherent forecasts often have a kink at jump-off, see Figure 4.5. In conclusion, the implied sex gap trajectory of coherent models being driven purely, or mainly, by the level effect is not sufficiently flexible to describe the observed sex gap dynamics.

To prove unimodality of the sex gap, we relied on the assumption that life expectancies are bounded from above, implying that we eventually approach a life expectancy plateau and that the sex gap therefore vanishes in the limit. Even though life expectancy does not appear to be approaching a maximum at present (Oeppen and Vaupel, 2002) and even though women have been shown to "always" live longer than men (Kalben, 2000; Zarulli et al., 2018), it does seem reasonable that a biological or genetic barrier to an infinite lifespan exists; however, if an upper limit does not exist a vanishing sex gap is not guaranteed. It is also conceivable that the present level of old-age mortality acts like a de facto barrier, but that continued improvements can "break through" the barrier and move it to even higher ages.

Although we have presented the results and the analysis as pertaining to coherent models and their properties, the unimodality result can also be given a real-world interpretation. In Western Europe, males have historically had lower rates of improvement than females, but currently the two sexes enjoy similar rates of improvement. Combined with the unimodality result, this indicates that the observed sex gap will continue to close in the future but it also points to factors that need to change for the gap to start widening again. That is, (i) if mortality undergoes a (perhaps temporary) regime change in which female rates of improvement substantially outweigh male improvements, for instance, if certain diseases that primarily target females are reduced or eradicated, or if sex-specific risk behaviour changes in favour of women; (ii) if the current lifetime barrier is broken through, the sex gap dynamics could "start over" and result in a multimodal curve similar to that exemplified in Figure 4.3. It is also conceivable that a widening gap in favor of men could be brought about. This can happen, e.g., if the sex-specific mortality curves continue to steepen in such a way that high-age male mortality falls below high-age female mortality.

The paper focused on coherence in two-sex models. However, our main result applies also to other coherent, multi-population models whenever an ordering of the population-specific mortality rates exists. A noteworthy instance is mortality rates between different socio-economic groups, a case that has attracted much recent attention, see e.g. Bennett et al. (2018) and Cairns et al. (2019). Historically, the life expectancy gap between the most and least affluent has widened for some time, and the trend is expected to continue over the coming years by domain experts. However, under the assumption of coherence, the gap may inadvertently be projected to close right after projection jump-off. Other applications with mortality orderings expected to persist over time include modelling of rich countries relative to poor countries, insured lives relative to non-insured lives, and smokers relative to non-smokers. In all these cases, coherent models are likely to produce immediately narrowing gaps irrespective of the historic development leading up to jump-off.

Finally, since the main result hinges on excess mortality of one population relative to another, it cannot be applied to situations where such an ordering does not exist.

For example, when modeling a group of neighboring countries of similar affluence, we cannot conclude that this will generally lead to closing gaps for all pairs of countries. Arguably, even if we could, narrowing gaps would perhaps be less of a worry in this context, at least in the long run.

Acknowledgements

The authors wish to thank three anonymous referees for helpful comments that improved the manuscript. This research was partly funded by Innovation Fund Denmark (IFD) grant number 9065-00135B.

4.A Proofs and Lemmas

4.A.1 Stochastic Dominance

The following lemma gives a general result that is used to assess the sign of the second derivative of the sex gap at a stationary point. The lemma has ties to the concept of stochastic dominance, see e.g. Lindvall (2002). We say that a real-valued random variable Y stochastically dominates another real-valued random variable X if $P(Y \leq z) \leq P(X \leq z)$ for all z. Informally, this means that the distribution of X is to the left of the distribution of Y. Various inequalities based on this partial ordering can be derived. The result we need is stated below.

Lemma 4.A.1. *Let f_1 and f_2 be strictly positive functions with compact support $[0, K_1]$ and $[0, K_2]$, respectively, with $0 < K_1 < K_2 < \infty$, satisfying*

$$\int_0^{K_1} f_1(x)\,\mathrm{d}x = \int_0^{K_2} f_2(x)\,\mathrm{d}x. \tag{4.A.1}$$

Assuming that f_1 and f_2 have continuous first derivatives such that

$$\frac{\partial}{\partial x}\log f_1(x) \leq \frac{\partial}{\partial x}\log f_2(x) \text{ for } 0 < x < K_1, \tag{4.A.2}$$

then for any strictly increasing, differentiable function $h : [0, K_2] \to \mathbb{R}$,

$$\int_0^{K_1} h(x) f_1(x)\,\mathrm{d}x < \int_0^{K_2} h(x) f_2(x)\,\mathrm{d}x. \tag{4.A.3}$$

Proof. Let m denote the common value in (4.A.1), and let ν_1 and ν_2 denote measures concentrated on $[0, K_1]$ and $[0, K_2]$, respectively, so that $g_1(x) = f_1(x)/m$ and $g_2(x) = f_2(x)/m$ are the Radon-Nikodym derivatives of ν_1 and ν_2 with respect to Lebesgue measure. Further, let $G_1(x) = \int_0^x g_1(u)\,\mathrm{d}u$ and $G_2(x) = \int_0^x g_2(u)\,\mathrm{d}u$ be the corresponding cdf's.

It suffices to show that $G_2(x) \leq G_1(x),\ \forall x \in [0, K_2]$, with strict inequality at some x, because then

$$\begin{aligned}\int_0^{K_1} h \,\mathrm{d}\nu_1 - \int_0^{K_2} h \,\mathrm{d}\nu_2 &= \int_0^{K_1} \int_0^x h'(u) \,\mathrm{d}u \,\mathrm{d}G_1(x) - \int_0^{K_2} \int_0^x h'(u) \,\mathrm{d}u \,\mathrm{d}G_2(x) \\ &= \int_0^{K_1} h'(x) \left(1 - G_1(x)\right) \mathrm{d}x - \int_0^{K_2} h'(x) \left(1 - G_2(x)\right) \mathrm{d}x \\ &= \int_0^{K_2} h'(x) \left(G_2(x) - G_1(x)\right) \mathrm{d}x < 0, \end{aligned} \tag{4.A.4}$$

and (4.A.3) follows immediately. In the above, the first equality follows by writing

$$h(x) = h(0) + \int_0^x h'(u) \,\mathrm{d}u, \tag{4.A.5}$$

and the second by reversing the order of integration. The latter computation is valid by Tonelli's theorem since h' is non-negative.

We know that $G_1(0) = G_2(0) = 0$ and $G_2(x) \leq G_1(x) = 1$ for $x \in [K_1, K_2]$. For $0 < z < w < K_1$ we have from (4.A.2) that

$$\log \frac{g_1(w)}{g_1(z)} = \int_z^w \frac{\partial}{\partial x} \log f_1(x) \,\mathrm{d}x \leq \int_z^w \frac{\partial}{\partial x} \log f_2(x) \,\mathrm{d}x = \log \frac{g_2(w)}{g_2(z)}, \tag{4.A.6}$$

and thereby $g_1(w)/g_1(z) \leq g_2(w)/g_2(z)$, or, equivalently, $g_2(z)/g_1(z) \leq g_2(w)/g_1(w)$. Hence, $r(x) = g_2(x)/g_1(x)$ is monotonically increasing on $(0, K_1)$. Define the sets

$$A = \{x \mid x \in (0, K_1), r(x) < 1\}, \tag{4.A.7}$$

$$B = \{x \mid x \in (0, K_1), r(x) = 1\}, \tag{4.A.8}$$

$$C = \{x \mid x \in (0, K_1), r(x) > 1\}, \tag{4.A.9}$$

that collectively cover $(0, K_1)$. Since

$$\int_0^{K_1} g_1(x) \,\mathrm{d}x = \int_0^{K_2} g_2(x) \,\mathrm{d}x > \int_0^{K_1} g_2(x) \,\mathrm{d}x = \int_0^{K_1} r(x) g_1(x) \,\mathrm{d}x, \tag{4.A.10}$$

A cannot be empty. Due to the monotonicity of r, C cannot be empty without B being empty. If both B and C are empty or if only C is empty, we have that

$$G_2(x) = \int_0^x r(u) g_1(u) \,\mathrm{d}u < \int_0^x g_1(u) \,\mathrm{d}u = G_1(x), \tag{4.A.11}$$

for all $x \in (0, K_1)$. If C is non-empty, suppose for contradiction that there exists an $x \in C$ for which $G_2(x) \geq G_1(x)$. Then, due to the monotonicity of r, we would have $G_2(K_1) \geq G_1(K_1)$, but this contradicts (4.A.10). Thus, we conclude that $G_2(x) < G_1(x)$ for all $x \in (0, K_1)$. □

4.A.2 Proof of Theorem 4.3.1

For ease of reference, we recall the central definitions and assumptions used in Theorem 4.3.1. Let $\mu : [0, \omega] \times \mathbb{R} \to (0, \infty)$ be a twice continuously differentiable function satisfying

$$\lim_{t \to -\infty} \mu(x,t) = \infty, \text{ and } \lim_{t \to \infty} \mu(x,t) = 0. \tag{4.A.12}$$

Let $\mu_f(x,t) = \mu(x,t)$ and $\mu_m(x,t) = \mu(x,t)c(x)$ denote female and male mortality, respectively. Let $\theta(t) = e_f(0,t) - e_m(0,t)$ denote the sex gap. Note that θ is twice continuously differentiable. The proof of Theorem 4.3.1 relies on Lemma 4.A.1 and the following two lemmas.

Lemma 4.A.2. *Assume $c(x) > 1$ for all x. The sex gap, θ, attains its global maximum on $\mathbb{R}$. In particular, θ has at least one stationary point.*

Proof. We first note that $\theta(t) > 0$ for all $t \in \mathbb{R}$ with limits

$$\lim_{t \to -\infty} \theta(t) = \lim_{t \to \infty} \theta(t) = 0 \tag{4.A.13}$$

This follows from (4.A.12) and the assumption $\mu_f(x,t) < \mu_m(x,t)$ for all x and t. Choose an arbitrary real number $t_0 \in \mathbb{R}$. Because the limit of θ in either end of the real line is zero, there exist numbers $-\infty < t_l < t_0 < t_u < \infty$ such that $\theta(t) \leq \theta(t_0)$ for $t \leq t_l$, and $\theta(t) \leq \theta(t_0)$ for $t \geq t_u$. Since θ is continuous it attains a maximal value at, say, t_m on the compact interval $[t_l, t_u]$, cf. Rudin (1976, Theorem 4.16). This is also the global maximum by construction since $t_0 \in [t_l, t_u]$. Hence, since θ is continuously differentiable, it follows that $\dot{\theta}(t_m) = 0$. □

Lemma 4.A.3. *Assume $c(x) > 1$ for all x. If stationary points of θ are local maxima, that is, if*

$$\forall t \in \mathbb{R} : \dot{\theta}(t) = 0 \Rightarrow \ddot{\theta}(t) < 0, \tag{4.A.14}$$

then θ is strongly unimodal on $\mathbb{R}$.

Proof. Assume (4.A.14) holds. We start by showing that θ can have at most one stationary point, or equivalently, that $\dot{\theta}$ can have at most one root. Assume for contradiction that $\dot{\theta}$ has at least two roots $a, b \in \mathbb{R}$ with $a < b$. By (4.A.14), $\ddot{\theta}(a) < 0$ and therefore there exists $\varepsilon > 0$ such that $\dot{\theta}$ is strictly negative on $(a, a + \varepsilon]$. Let $c = \inf\{t > a : \dot{\theta}(t) \geq 0\}$ be the first time $\dot{\theta}$ passes zero after a. The set defining c is disjoint with $(a, a + \varepsilon]$ and it is non-empty because it contains b, by assumption. In particular, $c \neq a$. By continuity, $\dot{\theta}(c) = 0$, because if $\dot{\theta}(c) > 0$ there would exist $a + \varepsilon < t < c$ with $\dot{\theta}(t) = 0$ contradicting the definition of c, see e.g. Rudin (1976, Theorem 5.12). For the same reason, $\dot{\theta}(t) < 0$ for all $t \in (a, c)$. Now, since θ is twice differentiable at c, the limit

$$\lim_{t \to c^-} \frac{\dot{\theta}(t) - \dot{\theta}(c)}{t - c} = \lim_{t \to c^-} \frac{\dot{\theta}(t)}{t - c} \tag{4.A.15}$$

exists and equals $\ddot{\theta}(c)$. But since $\dot{\theta}(t) < 0$ in an interval to the left of c, we find $\ddot{\theta}(c) \geq 0$ contradicting (4.A.14). We conclude that θ can have at most one stationary point.

By Lemma 4.A.2 we know that θ attains its unique maximum at some $t_m \in \mathbb{R}$, and as just shown this is the only stationary point. Consequently, $\dot{\theta}(t) > 0$ for all $t < t_m$, and $\dot{\theta}(t) < 0$ for all $t > t_m$, and we conclude the θ is strongly unimodal. □

Theorem 4.3.1. *Assume $-\frac{\partial}{\partial t} \log \mu(x,t) = \rho(t) > 0$ for all x, i.e., the rates of improvement is the same for all ages and strictly positive at all times. If $c(x) > 1$ for all x, and*

$$\frac{\partial}{\partial x} \log \left(\frac{\mu_m(I_m^{-1}(x,t),t)}{\mu_f(I_f^{-1}(x,t),t)} \right) \leq 0, \tag{4.3.9}$$

for all $t \in \mathbb{R}$, and all x where the argument is defined, then θ is strictly unimodal on $\mathbb{R}$.

Proof. By Lemma 4.A.3 it suffices to prove condition (4.A.14). The first and second derivatives of θ are given by

$$\dot{\theta}(t) = \rho(t) \int_0^\omega \left[I_f(x,t) e^{-I_f(x,t)} - I_m(x,t) e^{-I_m(x,t)} \right] \mathrm{d}x, \tag{4.A.16}$$

$$\begin{aligned} \ddot{\theta}(t) = \dot{\rho}(t) &\int_0^\omega \left[I_f(x,t) e^{-I_f(x,t)} - I_m(x,t) e^{-I_m(x,t)} \right] \mathrm{d}x \\ &+ \rho^2(t) \int_0^\omega \left[(I_f^2(x,t) - I_f(x,t)) e^{-I_f(x,t)} - (I_m^2(x,t) - I_m(x,t)) e^{-I_m(x,t)} \right] \mathrm{d}x. \end{aligned} \tag{4.A.17}$$

Fix a $t \in \mathbb{R}$ for which $\dot{\theta}(t) = 0$. We then get from (4.A.16) that $\ddot{\theta}(t)$ reduces to

$$\ddot{\theta}(t) = \rho^2(t) \int_0^\omega \left[I_f^2(x,t) e^{-I_f(x,t)} - I_m^2(x,t) e^{-I_m(x,t)} \right] \mathrm{d}x. \tag{4.A.18}$$

Thus, to establish (4.A.14) it suffices to establish that the integral in (4.A.18) is negative.

The proof relies on an invocation of Lemma 4.A.1. To bring $\dot{\theta}(t)$ and $\ddot{\theta}(t)$ on a form more suitable for this purpose, we introduce, with slight abuse of notation, the variable transform $z = I(x,t)$ and write

$$\dot{\theta}(t) = \rho(t) \left[\int_0^{I_f(\omega,t)} f_f(z,t) \, \mathrm{d}z - \int_0^{I_m(\omega,t)} f_m(z,t) \, \mathrm{d}z \right], \tag{4.A.19}$$

$$\ddot{\theta}(t) = \rho^2(t) \left[\int_0^{I_f(\omega,t)} z f_f(z,t) \, \mathrm{d}z - \int_0^{I_m(\omega,t)} z f_m(z,t) \, \mathrm{d}z \right], \tag{4.A.20}$$

where $f_g(z,t) = ze^{-z}/\mu_g(I_g^{-1}(z,t),t)$, for $g \in \{f,m\}$, is the (transformed) integrand of the integrals in (4.A.16). We note that I is invertible by the inverse function theorem, see e.g. Rudin (1976, Theorem 9.24).

Now, since assumption (4.A.2) of Lemma 4.A.1 is equivalent to assumption (4.3.9) of Theorem 4.3.1, that is,

$$\frac{\partial}{\partial z}\log\left(\frac{f_f(z,t)}{f_m(z,t)}\right) = \frac{\partial}{\partial z}\log\left(\frac{\mu_m(I_m^{-1}(z,t),t)}{\mu_f(I_f^{-1}(z,t),t)}\right) \leq 0, \tag{4.A.21}$$

and since $\mu_m(x,t) > \mu_f(x,t)$ for all x implies that $I_m(\omega,t) > I_f(\omega,t)$, we can invoke Lemma 4.A.1 to get that

$$\int_0^{I_f(\omega,t)} z f_f(z,t)\,\mathrm{d}z < \int_0^{I_m(\omega,t)} z f_m(z,t)\,\mathrm{d}z. \tag{4.A.22}$$

It immediately follows that the integral in (4.A.20), and thereby the integral in (4.A.18), is negative, and we are done. □

4.B Graduation of $\mu(x)$

Suppose we have data on deaths, $D(x)$, and exposures, $E(x)$, for integer ages $x \in \{0,\ldots,110\}$. By smoothing the empirical death rates, $m(x) = D(x)/E(x)$, we wish to obtain an improved representation of the true underlying mortality curve, $\mu(x)$. An example of the graduation procedure described below applied to data is shown in Appendix 4.6.

4.B.1 General Smoothing

Within the observed age range (0–110) we fit a cubic smoothing spline to the observed death rates, i.e., we model m by an additive noise model $m(x) = f(x) + \varepsilon(x)$ where $\varepsilon(x)$ is assumed iid with mean zero and the cubic smoothing spline estimate $\hat{f}$ of f minimizes

$$\sum_{x\in\mathcal{X}}\left[m(x) - \hat{f}(x)\right]^2 + \lambda\int \hat{f}''(u)^2\,\mathrm{d}u \tag{4.B.1}$$

where the smoothing parameter λ is fitted through cross-validation and

$$\mathcal{X} = \{0,1,10,20,30,40,50,60,70,80,110\}$$

is the set of knots used. From $\hat{f}$, we can obtain smoothed estimates $\hat{\mu}_{\text{spline}}(x) = \hat{f}(x)$ for general (non-integer) x.

4.B.2 Old-Age Smoothing

For the highest ages, we smooth the mortality curve by fitting a logistic function to the observed death rates for ages 80 and above. That is, we fit the Kannisto model

$$\operatorname{logit}\mu(x;a,b) = a + b(x-80), \tag{4.B.2}$$

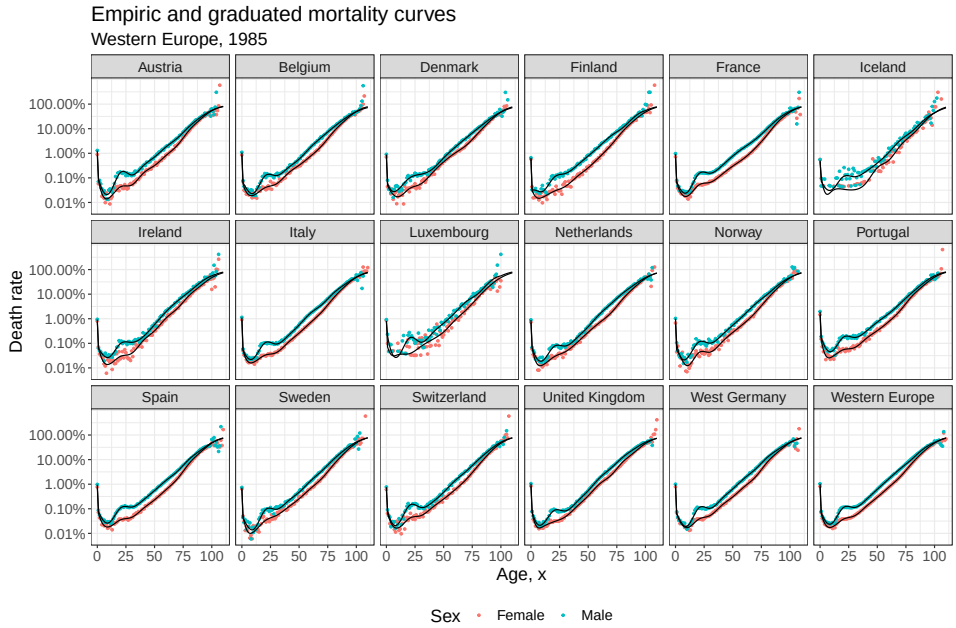

Figure 4.6: Empiric (dots) and graduated (line) death rates. The graduated mortality curve is computed through (4.B.4).

with parameters $a, b \in \mathbb{R}$, cf. Thatcher et al. (1998). The parameters are found by maximizing a Poisson log-likelihood

$$\log L(a,b) = \sum_{x=80}^{99} \left[D(x) \log \mu(x;a,b) - E(x)\mu(x;a,b) \right], \tag{4.B.3}$$

for data in the age-range 80–99. Substituting the maximum likelihood estimates $\hat{a}$ and $\hat{b}$ into (4.B.2) yields smoothed death rates $\hat{\mu}_{\text{logit}}(x) = \mu(x;\hat{a},\hat{b})$.

4.B.3 Mortality for the Combined Curve

We calculate smooth death rates for the entire age range by the weighted average

$$\hat{\mu}(x) = \hat{\mu}_{\text{spline}}(x)(1 - w(x)) + \hat{\mu}_{\text{logit}}(x)w(x), \tag{4.B.4}$$

with weights $w(x) = \min(1, \max((x-80)/20, 0))$ for $x \in [0, 110]$.

4.C Life Expectancy under Piecewise Constancy

Suppose that mortality is piecewise constant over squares of the form $[x, x+1)\times[t, t+1)$ for integer ages $x \in \{0, \ldots, 110\} = \mathcal{X}$ and calendar years $t \in \mathcal{T}$, that is, suppose

$$\mu(x + \Delta x, t + \Delta t) = \mu(x,t), \quad \Delta x, \Delta t \in [0,1),$$

for all $x \in \mathcal{X}, t \in \mathcal{T}$. The assumption of piecewise constancy is convenient for many calculations of life table related quantities. In particular, for any integer age of truncation, $\omega \in \{0, \dots, 111\}$, the (truncated) life expectancy can be evaluated analytically as:

$$\begin{aligned}
e(x,t) &= \int_x^{\omega} e^{-\int_x^y \mu(z,t)\,\mathrm{d}z}\,\mathrm{d}y \\
&= \sum_{i=x}^{\omega-1} \int_i^{i+1} e^{-\int_x^y \mu(z,t)\,\mathrm{d}z}\,\mathrm{d}y \\
&= \sum_{i=x}^{\omega-1} \left(\int_i^{i+1} e^{-\int_i^y \mu(z,t)\,\mathrm{d}z}\,\mathrm{d}y \right) e^{-\int_x^i \mu(z,t)\,\mathrm{d}z} \\
&= \sum_{i=x}^{\omega-1} \left(\int_i^{i+1} e^{-\mu(i,t)(y-i)}\,\mathrm{d}y \right) e^{-\sum_{j=x}^{i-1} \int_j^{j+1} \mu(z,t)\,\mathrm{d}z} \\
&= \sum_{i=x}^{\omega-1} \left[-\frac{e^{-\mu(i,t)(y-i)}}{\mu(i,t)} \right]_i^{i+1} e^{-\sum_{j=x}^{i-1} \mu(j,t)} \\
&= \sum_{i=x}^{\omega-1} \frac{1 - e^{-\mu(i,t)}}{\mu(i,t)} e^{-\sum_{j=x}^{i-1} \mu(j,t)}.
\end{aligned}$$

We use $\omega = 110$ throughout the paper.

Chapter 5

Forecasting, Interventions and Selection: The Benefits of a Causal Mortality Model

This chapter contains the manuscript *Jallbjørn et al. (2022)*.

Abstract

Integrating epidemiological information into mortality models offers the promise of improved forecasting performance and opens the possibility for assessing the gain of preventive measures that reduce disease risk. While probabilistic models can be used to forecast mortality, predicting how a system behaves under external manipulation is a causal query that requires a causal model. Using the framework of potential outcomes, we discuss how mortality forecasts are affected by interventions and we address the assumptions and data needed to operationalize such an analysis. We bring attention to a challenge unique to population-level mortality models. Common forecasting methods treat risk prevalence as an exogenous process, determined outside the mortality model. While ignoring (part of) the inter-dependency between risk and death makes the joint system easier to forecast, it comes at the cost of the model's ability to relay selection-induced effects. Using techniques from causal mediation theory, we pinpoint the selection effect usually missing in studies on cause-of-death elimination and when analyzing actions that modify risk prevalence. In particular, we decompose the total effect of an intervention into a part directly attributable to the action and a part due to selection effects. We illustrate the effects using U.S. data.

Keywords: *Mortality modelling, risk factors, cause elimination, interventions, causality.*

5.1 Introduction

Soaring life expectancies throughout the industrialized world have prompted a rapid increase in research on mortality modelling and forecasting. Predominantly, models focus on achieving accurate out-of-sample forecasts of future all-cause mortality relying on age, calendar time, and birth-cohort as sole predictors, e.g. Lee and Carter (1992), Cairns et al. (2006), and Renshaw and Haberman (2006). These models serve their purpose well and have proven extremely difficult to beat when it comes to predicting future death rates. None of them, however, include information about the causal mechanisms responsible for past trends, but instead assume secular linear trends at an aggregate level.

To foster a deeper understanding of mortality and its drivers, recent studies have focused on enriching traditional models by taking into account risk behaviour that exerts a strong influence on health. The aim is to disentangle the effects of general health improvements from risk behaviour and its changes over successive generations. Booth and Tickle (2008) characterize models exploiting relations between risk prevalence and death as explanatory models but warn that such relationships are still imperfectly understood. Nonetheless, various advances to make more precise and better substantiated forecasts by integrating health and lifestyle related trends have been made in recent years, see in particular Wang and Preston (2009), King and Soneji (2011), Janssen et al. (2013), Preston et al. (2014), and Foreman et al. (2018).

Thus far, however, little work exists on how explanatory models open the possibility for assessing the impact of health interventions in demographic forecasts. Studying how human longevity can be improved by curing or reducing the prevalence of existing diseases or by modifying risk behaviour is of great interest with many potential applications across a range of disciplines including demography, actuarial risk management, and health economics. In practice, such analyses are difficult to carry out because of their inherent causal nature, necessitating explicit assumptions about the data generating process and parameters that can be adjusted to represent interventions.

Moreover, current frameworks are generally incapable of analyzing the impact of interventions in a realistic and consistent manner. Explanatory models has hitherto regarded risk prevalence as an exogenous process, creating a one-sided dependence structure where risk prevalence affects mortality but not vice versa. This approach is incongruent with reality where the presence of individual risk factor differences play a key role in determining who die at a given point in time, which then in turn affects how much risk is left in the population at later times. Zeger and Liang (1991) use the term *feedback* to refer to such inter-dependencies, where the response at time t influences the risk factor distribution at future times $s > t$. Because the

model's ability to relay selection-induced effects rely on the feedback mechanism through which the (risk) composition of the surviving population is changed in accordance with the (risk) composition among those who die, appropriate joint forecasting methods are required if interventions are to conform with real-world implementations.

The purpose of this paper is to discuss how mortality forecasts are affected by interventions and to show by example the assumptions and data needed for such an analysis to be operationalized. To facilitate this discussion we use the framework of potential outcomes, see e.g. Imbens and Rubin (2015) and Hernán and Robins (2020). In particular, we seek to pinpoint the selection-induced feedback effect usually missing in studies on alternative mortality scenarios. Based on techniques from causal mediation theory, we propose a method that decomposes the total effect of cause-of-death elimination into a part directly attributable to death rate deletion and a part due to selection effects. The overall message is that if we want to make accurate statements about the effect of an intervention, that is if we are to quantify both the direct *and* the indirect effects of an intervention, risk prevalence must be endogenous to the mortality model. We consider this work a first step towards demographic forecasting of mortality under interventions.

5.2 When do we need a Causal Mortality Model?

The most pertinent use of a model describing the link between risk behaviour and cause-specific mortality is to obtain more precise and better substantiated mortality projections. In many countries, life expectancy has historically evolved in a complex fashion with periods of near stagnation followed by rapid increases. This history is hard to reconcile with current mortality models in which the assumption of secular linear trends is often at odds with reality (Janssen and Kunst, 2007; Jarner and Jallbjørn, 2022). In contrast, separating lifestyle-related risks and diseases from mortality allows for a detailed understanding of the historic evolution which in turn could lead to more solidly founded forward-looking projections.

In general modelling and forecasting (cause-specific) mortality is a problem of prediction, for which a probabilistic model suffices. That is, if we wish to predict mortality given the prevalence of concurrent health risks, it suffices to consider a probabilistic model of the conditional hazard. In theory, it is also possible to use such a model to study the impact of interventions by perturbing the distribution of variables that have been conditioned on as is done in conventional stress-tests. But this analysis will only give a realistic picture of the consequences of an intervention if the conditional distribution of mortality does not change when the predictors change.

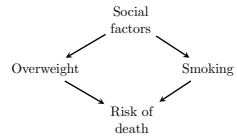

Figure 5.1: An example of confounding.

For example, suppose that we are interested in studying the effect of the number of cigarettes smoked daily, S, on the risk of death, D. There is an abundance of evidence in the literature to support the conclusion that the more one smokes, the higher one's risk of death becomes. We can represent this cause-effect relationship concisely as 'Smoking $\rightarrow$ Risk of death'. Suppose we focus on the regression task of learning $s \mapsto \mathbb{E}[D|S = s]$. It is then tempting to interpret $\mathbb{E}[D|S = s] - \mathbb{E}[D|S = 0]$ as the expected increase in survival for a smoker, who quits smoking. But such an interpretation is invalid in the presence of confounding factors. A confounding factor is any variable that influences both risk exposure and response. Here we could imagine the relationship depicted in Figure 5.1. A lack of commitment to not smoke tends to co-occur with a higher susceptibility to being obese through various underlying social factors. Since associations between the risk factor 'Smoking' and the outcome may arise not only through the number of cigarettes smoked daily but also through the underlying factors that determine general risk behaviour, the parameters associated with $\mathbb{E}[D|S = s]$ may have no causal meaning.

In comparison, the distinctive feature of a causal model is its ability to represent and update its prediction endogenously in response to changing conditions. This ability stems from the property that the conditional distribution of a variable given its causes is stable under interventions that only affect other variables. That is, the generative mechanism for a variable not targeted by an intervention is left intact. Several comprehensive books have been devoted to the topic of causal inference and discovery, see e.g. Spirtes et al. (2000), Pearl (2009), Imbens and Rubin (2015), Peters et al. (2017), and Hernán and Robins (2020), and we will thus not give an in-depth account of the concepts and methods here. In the present paper we focus on defining and decomposing the effects of interventions in demographic forecasts of mortality with a particular emphasis on the role of selection.

5.2.1 When is a mortality model causal?

To give precise answers to causal questions, we need to invoke restrictive assumptions about the data generating process. A formal account of the ones needed can be found in, e.g., Robins (1998), Robins et al. (1999), and Hernán and Robins (2020). These may hold by virtue of study design, for instance in a randomized controlled trial, but generally we can only identify causal effects from observational data when there is no unmeasured confounding. Because no regard is paid to confounding factors in a standard regression analysis, these estimates cannot be endowed with a causal interpretation. This issue is also well recognized in the mortality forecasting literature, for instance by King and Soneji (2011) who carefully remark: "*Indeed, none of our results should be seen as claims about the causal effects of obesity, smoking, or any*

other factor.".

Estimating causal effects is an ambitious task requiring specialized methods, subject matter expertise, and detailed individual-level health data. To operationalize a causal mortality model, dose-response relationships must be based on epidemiological evidence from the literature, supported by trial and cohort data. Recent advances in the field of epidemiology, spearheaded by the Global Burden of Disease (GBD) initiative, may assist in bridging this gap between mortality and its determinants. In particular, Murray et al. (2020) gives a standardized and comprehensive account of how 87 risk factors interact and affect different causes of death, covering in total 560 risk-outcome pairs based on a systematic review of partial studies. Using data from the GBD, it is possible to construct a causally interpretable mortality model with the aim of forecasting country specific mortality under varying scenarios. We take up this task in the second part of the paper.

The modelling choices we make are close in spirit to the seminal work of Foreman et al. (2018), who also build an explanatory model based on the GBD estimates with the aim of substantiating mortality forecasts and exploring alternative health scenarios. While Foreman et al. (2018) does evaluate better/worse scenarios, these are made using a more conventional stress testing procedure, where the improvement rates of the risk factors are varied. They stress that such scenarios are to be understood as "[...] *a signal on the scope for policy change*", rather than actual alternative scenarios. We wish to expand on their method of analysis and explain which additional model components are needed to evaluate actual interventions.

5.2.2 Feedback between mortality and risk prevalence

Building and calibrating a causal mortality model comes with the general challenges intrinsic to causal modelling and discovery, some of which are described in the previous section. In this paper we bring attention to an additional challenge which is unique to predicting mortality under interventions in population-level models, namely the second-order effects that arise due to selection over time and how these can be quantified.

In general, perturbing any one part of a system has a ripple effect as the initial disturbance propagates to the remaining system over time. This is also true when concerned with the dynamics of mortality which is subject to a selection-induced feedback mechanism: if the risk composition among the survivors change, the composition among those who die also change and vice versa. We aim to explicate the effect relayed through the feedback mechanism under two types of interventions, namely i) on actions that target the death rates directly (i.e., cause-of-death elimination), and ii) on actions that modify behavioural risk factor prevalence thus targeting the death rates indirectly.

Cause-of-death elimination

The impact of eradicating certain causes of death is a topic widely debated in the literature, dating back to Bernoulli's discussion of a hypothetical world without small-pox, presented before the French Academy of Sciences in 1760 (Karn, 1931). The pivotal assumption made by Bernoulli, namely that individuals "saved" from the eliminated cause are as susceptible to dying from the non-eliminated causes as the general population, still permeates most cause-deleted lifetable calculations today. Indeed, the prevailing methodology is to directly manipulate the cause-specific death rates of interest, while leaving remaining rates unaffected. This is commonly referred to as cause elimination under an assumption of independent competing risks, an approach that typically overstates the actual effect because it fails to account for subsequent selection, cf. Keyfitz (1977b).

Some papers recognize the issue of dependence among competing causes in their estimates of cause-deleted life tables, e.g. Manton and Poss (1979), Mackenbach et al. (1999), Kaishev et al. (2007), Dimitrova et al. (2013), Alai et al. (2015), and Li and Lu (2019), but they do not explain the pathways through which dependence originates. We give one way of explaining the dependence by linking individual risk behaviour to cause-specific mortality. It is this link that allows us to explicate the consequences of selection following cause elimination.

As a motivating example, suppose that we are able to prevent all deaths due to lung cancer by some unusually successful targeted laser therapy[1]. In this hypothetical world there will, at least initially, be fewer deaths compared to the world where the cause still operates. But since everyone eventually dies, deaths are ultimately redistributed among remaining causes. What is left is then to quantify how soon those "saved" die from something else, and what they die from instead. Because individuals who die from lung cancer are predominantly smokers, improvements in the treatment of lung cancer will indirectly affect (and most likely increase) the mortality rates for other tobacco-attributable causes such as heart diseases, following the progressive build-up of smokers in the population. The initial decrease in the aggregate death rate due to lung cancer being eradicated is thus partly offset by a subsequent "harvesting" of the "saved" smokers. We will return to this point later in the paper.

[1]The intervention 'prevent all deaths due to cause k' is ambiguous and it is crucial to define the precise manner in which such elimination is achieved. Here, we focus on interventions that affect the *treatment* of lung cancer, rather than interventions that affect the underlying risk factors. If, in this example, the elimination of lung cancer was achieved by convincing the entire population to never have smoked, the impact of the intervention would be far greater, because not only the lung cancer rate but all tobacco-attributable death rates would be lowered.

5.3 The Feedback Mechanism

To define interventions and their consequences on mortality, we will conceptualize how risk mechanisms at the level of individuals transfer to the level of populations in the framework of potential outcomes. We establish the basic relations in this section and use them to obtain a visual representation of the feedback mechanism through which mortality and risk prevalence influence each other. We then give an instructive example that demonstrates the role of the mechanism in a scenario of cause-of-death elimination.

For ease of exposition, we consider the dynamics of a single ageing (birth) cohort followed until some maximum attainable age $\omega \in (0, \infty)$. Since age and calendar time advance synchronously in this case, we omit dependence on time in the following. Consider $i = 1, \ldots, n$ independent lives endowed with an age-varying vector of categorical[2] covariates $Z^{(i)}(x)$ and denote by an overbar the history of the covariate process, i.e. $\bar{Z}(x) = (Z(u) : 0 \leq u \leq x)$. Let $\bar{z}(x) = (z(x) : 0 \leq u \leq x)$ be a possible (fixed) covariate trajectory and define $X^{\bar{z}}$ as the individual's *potential* life time had covariate exposure been $\bar{z}$ with $\bar{z}$'s dependence on x suppressed. For each possible trajectory of $\bar{z}$ the distribution of $X^{\bar{z}}$ is completely characterized by the hazard rate

$$\mu^{\bar{z}}(x) := \lim_{dx \to 0^+} \mathsf{P}\left(x \leq X^{\bar{z}} < x + dx \mid X^{\bar{z}} \geq x\right)/dx. \tag{5.3.1}$$

The superscript identifies that $\mu^{\bar{z}}$ is the hazard function for $X^{\bar{z}}$. We can relate the potential outcome to the observed outcome by making the assumption of *consistency*, namely that the two coincide for the observed covariate trajectory, i.e. $X^{(i)} = X^{\bar{z}}$ when $\bar{Z}^{(i)}(x) = \bar{z}(x)$ for all $0 \leq x \leq X^{(i)}$.

All-cause mortality is decomposed by considering $k \in \{1, \ldots, K\}$ mutually exclusive and exhaustive causes of death. This is a situation of competing risks, where different causes compete to end the life of an individual and occurrence of one event precludes occurrence of the remaining. The cause-specific hazard $\mu_k^{\bar{z}}$ characterizes the instantaneous rate of death from cause k in the presence of competing causes and is constructed such that $\sum_{k=1}^{K} \mu_k^{\bar{z}}(x) = \mu^{\bar{z}}(x)$ for all x.

5.3.1 The relative risk model

To facilitate estimation and inference, some structure must be imposed on the hazard. In epidemiological and biostatistical applications where the inferential goal is to establish causal explanations for the etiology of disease and death, mortality is often studied in the relative risk framework. Motivated by the framework used in the GBD, we adopt a simplified paradigm[3] in which the death rate is related to $\bar{z}$ in a

[2] Covariates are, especially for large cohort studies, often reported as categorical variables even when the underlying exposure is continuous.

[3] We note that (5.3.2) is a partly conditional model in the sense that it does not condition on the entire covariate history of the individual but only on the concurrent exposure. Nonetheless,

multiplicative fashion adhering to the form

$$\mu_k^{\bar{z}}(x) = \mu_{0k}(x) \exp\left(\beta_k^\top z(x)\right), \tag{5.3.2}$$

where μ_{0k} is the baseline hazard describing mortality for an individual with no excess risk, and β_k are coefficients governing the effect of risk exposure. Equation (5.3.2) can be classified as a marginal structural model in the sense that it provides a structural (and thereby *causal*) description of the marginal distribution of $X^{\bar{z}}$. This interpretation will be important when arguing about the effects of interventions.

5.3.2 Mortality from the individual's point of view

To visualize the inter-dependence between an individual's exposure to risk and probability of death, notice that the life time $X^{(i)}$ gives rise to the multivariate counting processes $N^{(i)}(x) = (N_1^{(i)}(x), \ldots, N_K^{(i)}(x))$ where $N_k^{(i)}(x) = \mathbb{I}(X^{(i)} \leq x, \delta = k)$ is an indicator function registering whether or not an individual has died at age x from cause k, with δ designating cause. If we let $D_k^{(i)}(x)$ denote the increment of $N_k^{(i)}(x)$ over an infinitesimally small interval $[x, x + dx)$, that is the number of deaths observed in the interval, then, informally,

$$\mathsf{P}(D_k^{(i)}(x) = 1 \mid \text{individual } i \text{ alive just before age } x) = \mu_k^{(i)}(x) Y^{(i)}(x) dx \tag{5.3.3}$$

where $Y^{(i)}(x) = \mathbb{I}(X^{(i)} \geq x)$ denotes the at-risk indicator. Thus, the expected (local) change in the death process is a function of the individual's covariates and their survivorship status. We note that $Y^{(i)}$ depends on the set of causes operating, but we suppress this in the notation for now.

Let $0 = x_0 < x_1 < \cdots < x_{J-1} < x_J = \omega$ be a partition of $[0, \omega]$. We can then depict the functional relationships of the joint model from the individual's point of view graphically as in the left panel of Figure 5.2. The graph is to be read as follows. At a given age x_j, we first observe if the individual is alive. In the affirmative, $Z^{(i)}(x_j)$ is defined with its current value depending on its previous state. The number of events between x_j and x_{j+1} is

$$D_k^{(i)}(x_j) = \int_0^\omega \mathbb{I}(x_j \leq x < x_{j+1}) \, \mathrm{d}N_k^{(i)}(x) = \mathbb{I}(X^{(i)} \in [x_j, x_{j+1}), \delta = k), \tag{5.3.4}$$

for $j \in \{0, \ldots, J-1\}$, and is observed immediately after $Y^{(i)}(x_j)$ and $Z^{(i)}(x_j)$. Once an individual experiences an event at some age x_j, all future variables $Y^{(i)}(x_{j'})$ and $D^{(i)}(x_{j'})$ are deterministically zero while covariates $Z^{(i)}(x_{j'})$ are undefined for $j' > j$. From a modelling perspective, the graph shows that the forecasting procedure can be modularized by determining first the covariate dynamics (without regard to the specific time of death) and subsequently the death rate given the covariates. Thus, forecasting risk exposure prior to mortality is an admissible strategy at the level of individuals.

(5.3.2) can also accommodate prior behaviours by making them explicit levels of the categorical covariates.

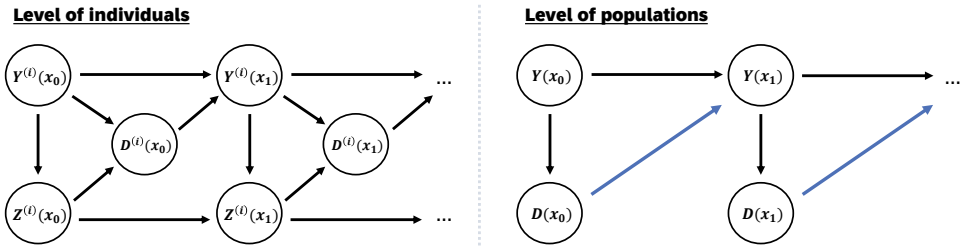

Figure 5.2: Directed acyclic graphs describing the functional relationship between death and risk prevalence: (left panel) individual-level model, (right panel) population-level model obtained by marginalizing over surviving individuals. The arrow $D(x_j) \to Y(x_{j+1})$ represents the feedback mechanism.

5.3.3 Moving to the population level

In demographic and actuarial studies of mortality the focus is on the aggregate age-specific death rate. To relate the individual level model to the population level we marginalize (5.3.2) over surviving individuals. The corresponding counting process is $N(x) = (N_1(x), \dots, N_K(x))$ with increment $D(x) = (D_1(x), \dots, D_K(x))$ where $N_k(x) = \sum_{i=1}^n N_k^{(i)}(x)$ and $D_k(x) = \sum_{i=1}^n D_k^{(i)}(x)$. Importantly, the intensity process for $N_k(x)$ is

$$\sum_{i=1}^n \mu_k^{z^{(i)}}(x) Y^{(i)}(x) = \mu_{0k}(x) \sum_{i=1}^n \exp\left(\beta_k^\top z^{(i)}(x)\right) Y^{(i)}(x). \tag{5.3.5}$$

Since covariates are categorical, we can assume a grouping of the individuals based on their covariate configuration. We denote the $G \in \mathbb{N}_+$ different subgroups of individuals by $g \in \{1, \dots, G\} = \mathcal{G}$ and their covariate configuration by z_g. This grouping depends on age, because it is performed on the basis of age-varying covariates. The proportion of surviving individuals in group g at age x is $\pi_g(x) = Y_g(x) / \sum_{g \in \mathcal{G}} Y_g(x)$ where $Y_g(x) = \sum_{i=1}^n Y^{(i)}(x) \mathbb{I}(z^{(i)}(x) = z_g)$ is the number of individuals alive in group g. Now, with $R_{kg} = \exp(\beta_k^\top z_g)$ being the combined relative risk of group g for cause k we can write (5.3.5) as a weighted average

$$\mu_{0k}(x) \sum_{i=1}^n \exp\left(\beta_k^\top z^{(i)}(x)\right) Y^{(i)}(x) = \mu_{0k}(x) \sum_{g \in \mathcal{G}} \pi_g(x) R_{kg} =: \mu_k^\pi(x), \tag{5.3.6}$$

where $\pi(x) \in \{p \in [0,1]^G \mid \sum_{g=1}^G p_g = 1\}$ is the *potential* risk factor composition identified by the superscript π.

The right panel of Figure 5.2 illustrates the aggregate equivalent model obtained from collapsing the individual level variables. Importantly, $Y^{(i)}$ and $Z^{(i)}$ are collapsed into a single node $Y(x) = (Y_1(x), \dots, Y_G(x))$ representing the number of individuals alive in the G groups at age x. The arrow $D(x_j) \to Y(x_{j+1})$ encodes feedback; the risk composition among those who die at age x_j affects the risk composition among

those left alive at age x_{j+1}. As a consequence, the forecast of risk prevalence Y cannot be modularized without cost in the same way that the forecast of individual level risk factors $Z^{(i)}$ can. If risk prevalence is exogenous in the sense that $Y(x_{j+1}) \perp\!\!\!\perp D(x_j) \mid Y(x_j)$ then there is no edge $D(x_j) \to Y(x_{j+1})$ whereby the model explicitly ignores the feedback effect. Consequently, the risk composition becomes invariant to any intervention that would cause a change in death rates. Thus, feedback must be recognized if perturbations are to reflect real-world implementations of interventions.

An example: Changes in population composition due to cause elimination

It is instructive to see how eradication of a cause-of-death impacts the remaining causes in the simplest possible setting. Consider therefore a closed population consisting of two homogeneous subgroups, differing only by their exposure to a binary risk factor $Z \in \{0, 1\}$. Suppose that there are two causes operating in this world, governed by the individual-level model

$$\mu_k^z(x) = \begin{cases} \mu_{0k}(x), & \text{if } z = 0, \\ \mu_{0k}(x) R_k, & \text{if } z = 1, \end{cases} \tag{5.3.7}$$

with $R_k > 1, \mu_{0k}(x) > 0$ for all $x \geq 0$ and both $k \in \{1, 2\}$. At the population level, the cause-specific hazard is a weighted average of the healthy ($z = 0$) and the unhealthy ($z = 1$) subpopulations

$$\mu_k^\pi(x) = \pi(x)\mu_k^{z=1}(x) + (1 - \pi(x))\mu_k^{z=0}(x), \tag{5.3.8}$$

where $\pi(x)$ is the proportion of unhealthy individuals at age x, namely

$$\pi(x) = \frac{\pi(0)S^{z=1}(x)}{(1-\pi(0))S^{z=0}(x) + \pi(0)S^{z=1}(x)} = \left[\frac{1-\pi(0)}{\pi(0)}\frac{S^{z=0}(x)}{S^{z=1}(x)} + 1\right]^{-1}, \tag{5.3.9}$$

with survival function $S^z(x) = \exp\{-\int_0^x (\mu_1^z(u) + \mu_2^z(u))\,\mathrm{d}u\}$ and $\pi(0) \in [0, 1]$ being some initial state.

Now, consider the "reference" world with both causes operating, $\mathcal{K} = \{1, 2\}$, and a hypothetical world in which cause 1 has been eradicated, $\mathcal{K}^* = \{2\}$. Because of competing risks, we need to be mindful that the cause k hazard is evaluated in the presence of other causes. We make this explicit in the notation now with π^{K} identifying the risk proportion and $S^{\mathsf{K},z}$ the survival function in a world where a specific (sub)set of causes $\mathsf{K} \subseteq \{1, 2\}$ are operating. By the assumptions above, we have that

$$\frac{S^{\mathcal{K},z=0}(x)}{S^{\mathcal{K},z=1}(x)} = \exp\left\{\int_0^x \left[\mu_{01}(u)(R_1 - 1) + \mu_{02}(u)(R_2 - 1)\right]\mathrm{d}u\right\} \geq \frac{S^{\mathcal{K}^*,z=0}(x)}{S^{\mathcal{K}^*,z=1}(x)}, \tag{5.3.10}$$

which, combined with (5.3.9), implies that $\pi^{\mathcal{K}}(x) < \pi^{\mathcal{K}^*}(x)$ for all $x > 0$. Thus, eradicating cause 1 weakens the selection mechanism resulting in a progressive

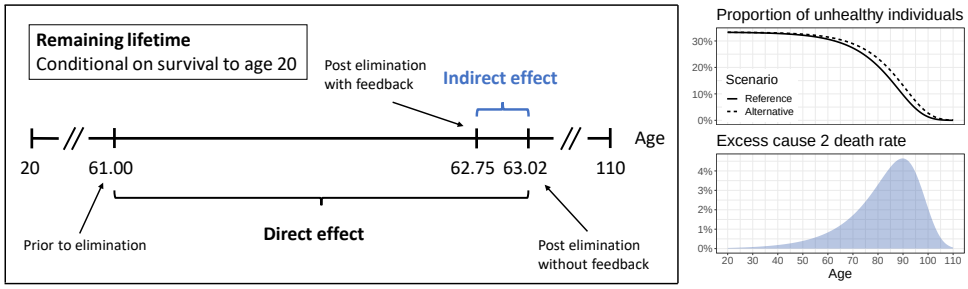

Figure 5.3: Example effect of cause elimination for a cohort aged 20. The initial proportion of unhealthy individuals is $\pi(0) = 1/3$. Baseline mortality curves are given by $\mu_{01}(x) = \exp(-11.69 + 0.074x)$ and $\mu_{02}(x) = \exp(-10.58 + 0.088x)$ and relative risks by $R_1 = 5$ and $R_2 = 2.5$. The parameters are calibrated to reflect current death rates due to cancer and residual causes. The upper right panel visualizes the build-up of unhealthy individuals following cause-1 elimination, while the lower right panel pictures the subsequent harvesting. The left panel shows the remaining life expectancies prior and post elimination. In this example, the feedback effect reduces the life expectancy gained by about a quarter of a year.

build-up of unhealthy individuals, which makes the cause 2 death rate rise at the population level

$$\mu_2^{\pi^{\mathcal{K}^*}}(x) - \mu_2^{\pi^{\mathcal{K}}}(x) = \mu_{02}(x)(R_2 - 1)\left(\pi^{\mathcal{K}^*}(x) - \pi^{\mathcal{K}}(x)\right) > 0, \quad x > 0. \tag{5.3.11}$$

Equation (5.3.11) describes an indirect effect of cause removal brought about by a change to the risk composition through the feedback mechanism. We will formalize the distinction between direct and indirect effects of cause removal later. For now, notice that if the model ignored the feedback effect in the sense that π^{K} was given in advance, and not on the basis of (5.3.9), then the indirect effect would be zero since $\pi^{\mathcal{K}}(x) = \pi^{\mathcal{K}^*}(x)$ for all x. The effects of cause removal are exemplified in Figure 5.3.

5.4 A Causal Mortality Model

For the remainder of the paper, we switch focus to a model spanning multiple birth cohorts and therefore consider a population defined in the rectangular age-period region

$$\mathcal{R}_{\text{data}} = \{(x,t) \mid x_{\min} \le x \le x_{\max}; t_{\min} \le t \le t_{\max}\}. \tag{5.4.1}$$

To give a proper justification for the declining mortality rates observed over the past centuries, one needs to model the influence of both individual and contextual factors on the risk of death. Contextual factors are the general living conditions to which all individuals are exposed, while individual factors may be divided into two types – observable and unobservable. We do not consider unobserved heterogeneity in the following, although this could be modelled using standard frailty theory as

in Vaupel et al. (1979). Our focus is instead on individual differences relating to (observable) health and lifestyle related behaviour. Following the notation outlined in the previous section, we assume that the cause-specific death rate under a potential covariate trajectory $\bar{z}(x)$ follows the relative risk regression model

$$\mu_k^{\bar{z}}(x,t;C(t)) = \mu_{0k}(x;C(t)) \exp\left(\beta_k^\top(x) z(x)\right). \tag{5.4.2}$$

Here, the pair $(x,t) \in \mathcal{R}_{\text{data}}$ identifies the cohort in question. Individual risk exposure is captured as a multiplicative effect on the baseline rate. The relative risk coefficients β_k vary with age but not over time. Age-related changes are consistent with current epidemiological research which indicates that the relative effect of (most) risk exposures dissipate over the course of a life span. Time invariance is, however, only justifiable over short- to medium horizons as it renders the model unable to capture temporal changes in the effect of exposure.

The process $C(t)$ describes the evolution of contextual variables such as improvements in food security, water supply, and sanitation, innovations in public health and medicine, the development of a health care system, GDP, and so on. It plays a dual role as a time-varying confounder process that must be controlled for to ensure that the β_k's have a causal interpretation, and acts as an effect modifier by stratifying the baseline death rate. Because the evolution of these environmental factors exhibits high co-linearity with the calendar year in which they are measured, we typically equate $C(t) = t$ whereby calendar time acts as a surrogate confounder. This means that death rates may change over time due to various unobserved factors, which is then captured by distributional changes as a function of time.

The equivalent of (5.4.2) at the population level obtained through aggregation under the assumption of categorical covariates reads

$$\mu_k^\pi(x,t;C(t)) = \mu_{0k}(x;C(t)) \sum_{g \in \mathcal{G}} \pi_g(x,t) R_{kg}(x), \tag{5.4.3}$$

where $R_{kg}(x) = \exp(\beta_k^\top(x) z_g)$ collects the risks of individuals in group g associated with each covariate and thus describes the combined relative risk of a given group. Both the individual and population level models fit within the relative risk framework. Whereas (5.4.2) relies on a specific covariate configuration, the aggregate model collapses the entire heterogeneous population into a single risk weighted individual with relative risk

$$R_k^\pi(x,t) = \sum_{g \in \mathcal{G}} \pi_g(x,t) R_{kg}(x). \tag{5.4.4}$$

5.5 Forecasting, Interventions and Selection

We turn to the impact of interventions on demographic mortality forecasts. Demographic forecasting is centered around the projection of population level quantities

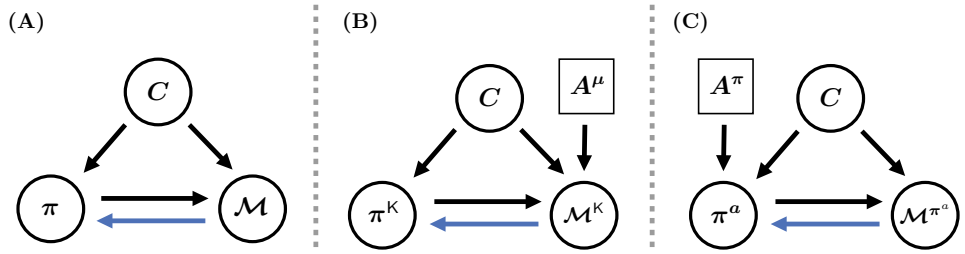

Figure 5.4: Rolled graphs of models with and without interventions. The blue arrow encodes feedback. Panel (A) shows the unintervened setting describing the relationship between π, $\mathcal{M}$, and C. Panels (B) and (C) are graphs of models where the death process and the risk process are indexed by possible actions. In panel (B) the action is on the set of causes operating and in panel (C) the action is an intervention on the risk distribution.

such as the aggregate death rate. Predominantly, the stochastic processes considered are indexed by discrete time and standard time series methods are used for prediction. Even though we also view our data as time series in the following, all the points made can be extended to the continuous time case.

Our focus will be on time points in the forecast region $\tau \subseteq \mathbb{Z}$. We have three processes we need to consider jointly. The multivariate process containing the cause-specific death rates $\mathcal{M} = \{\mathcal{M}(t)\}_{t\in\tau}$ where $\mathcal{M}(t) = (\mu_1(t), \ldots, \mu_K(t))^\top$ with components $\mu_k(t) = (\mu_k(x_{\min}, t), \ldots, \mu_k(x_{\max}, t))^\top$, the multivariate risk prevalence process $\pi = \{\pi(t)\}_{t\in\tau}$ where $\pi(t) = (\pi(x_{\min}, t), \ldots, \pi(x_{\max}, t))^\top$, and the confounder process $C = \{C(t)\}_{t\in\tau}$. To understand how interventions affect this system, we must describe how the processes influence each other, and, in particular, whether or not one process has predictive power over another. The concept of Granger causality (Granger, 1969) known from econometrics formalizes the notion of influence between processes, and is particularly useful for studying dynamic relationships in multivariate time series. We give a precise definition and explain how the concept is used to obtain graphical representations in Appendix 5.A.

We can represent the system by the graph shown in Figure 5.4 Panel (A). In the graph each process is represented as a single node with time being implicit. Two nodes are joined by a directed edge whenever a process at time t is predictive for another process at a future time $s > t$. For instance, in a model with feedback the cycle $\pi \to \mathcal{M} \to \pi$ represents a mutual dependence between $\mathcal{M}$ and π. The level of risk faced by the population at time t affects the death rate experienced between t and $t+1$, which in turn affects risk prevalence at time $t+1$. Conversely, the absence of an edge implies that a process is not predictive for another. Thus, in a model without feedback there is no arrow pointing from $\mathcal{M}$ to π because $\pi(t+h) \perp\!\!\!\perp \overline{\mathcal{M}}(t) \mid (\overline{\pi}(t), \overline{C}(t))$ for any $h \in \mathbb{N}_+$. This relation is asymmetric in the sense that the risk composition always predicts the death rate $\mathcal{M}(t+1) \not\perp\!\!\!\perp \overline{\pi}(t) \mid (\overline{\mathcal{M}}(t), \overline{C}(t))$.

5.5.1 Cause-of-death elimination

Inspired by Eichler and Didelez (2007, 2010), we consider a set of actions $A^\mu = (A_1, \ldots, A_K)$ that act on components of $\mathcal{M}(t)$ through all points in time $t \in \tau$. For our purposes, each A_k takes values in $\{0, 1\}$ describing two different regimes. Having $A_k = 0$ corresponds to no action on the k'th component, while $A_k = 1$ is an atomic intervention that forces $\mu_k(t)$ to be zero for all $t \in \tau$. More general interventions could also be considered but will not be pursued in the present paper.

We assume that intervening on the k'th component of $\mathcal{M}$ does not affect the remaining components or the remaining processes in the system other than through past variables that may develop differently depending on the intervention.[4] We can then represent an intervention on $\mathcal{M}$ by augmenting the graph in Figure 5.4 Panel (A) with an additional source node A^μ pointing into $\mathcal{M}$ as shown in Panel (B). Because A^μ is a decision variable it is represented graphically by a box and indicates possible eradication of certain causes of death. The variable $\mathcal{M}^{\mathsf{K}}$ is a potential outcome indexed by the set of causes operating following the action $A^\mu = a$. Risk prevalence π^{K} is likewise indexed by this set as it may develop differently depending on which components in $\mathcal{M}$ that are affected.

It follows that without feedback there is no causal effect of intervening in $\mathcal{M}$ on $\pi(t)$ for any $t \in \tau$. Thus the figure with the blue feedback edge removed represents a model where the action of cause removal only has a direct effect on the risk of death, because there is no indirect effect through variables between time points t and $t + h$, $h \in \mathbb{N}_+$. In other words, while the action does prevent specific types of death, thereby increasing the *absolute* number of deaths at later points in time because of competing risks, it changes neither the *relative* risk composition nor the death rates of non-eliminated causes. When the model includes feedback the risk process acts as an intermediate variable that mediates an additional effect through the loop $\mathcal{M} \to \pi \to \mathcal{M}$. In this case, cause-elimination weakens the selection mechanism and leads to larger (relative) concentration of high-risk individuals at future time points.

A decomposition of the death rate

The causal contrast of interest is the difference between the death rate in the reference world where all causes are operating compared to the rate in a world where only a subset of causes are operating. The all-cause death rate is

$$\mu^{\mathsf{K}}(x, t; C(t)) = \sum_{k \in \mathsf{K}} \mu_{0k}(x; C(t)) \sum_{g \in \mathcal{G}} \pi_g^{\mathsf{K}}(x, t) R_{kg}(x). \tag{5.5.1}$$

[4] Formally, we can write this assumption as $(C(t), \pi(t), \mathcal{M}_{-k}(t)) \perp\!\!\!\perp A_k | (\bar{C}(t-1), \bar{\pi}(t-1), \bar{\mathcal{M}}(t-1))$ for all $t \in \tau$ where $\mathcal{M}_{-k}(t)$ denotes $\mathcal{M}(t)$ without the k'th component. It is important to note that A_k is not a stochastic variable thus altering slightly the meaning of the $\perp\!\!\!\perp$–symbol. Here, $\perp\!\!\!\perp$ expresses that the distribution of $(C(t), \pi(t), \mathcal{M}_{-k}(t))$ is the same regardless of the value of A_k, cf. Dawid (2002).

Here, μ^{K} is a single-world quantity where the K-index refers to both the set of causes entering the sum and the world in which π is evaluated. Examining the impact of an intervention $A^{\mu} = a^*$ that leaves only a subset of causes $\mathcal{K}^* \subsetneq \mathcal{K} = \{1, \ldots, K\}$ operating comes down to evaluating the difference

$$\mathrm{TE}(x,t) = \mu^{\mathcal{K}}(x,t) - \mu^{\mathcal{K}^*}(x,t), \tag{5.5.2}$$

which constitutes the total causal effect. We have left the conditioning on C implicit for readability. Comparing the total effect in the model without feedback to the total effect in the model with feedback does tell us something about how much of the effect is mediated via the risk process, but it does not give us a clean decomposition. Instead, we consider the standard definitions of natural direct and indirect effects from the mediation literature adapted to the present setup, cf. Robins and Greenland (1992) and Pearl (2001).

We seek to measure the direct effect of the action $A^{\mu} = a^*$ associated with the arrow $A^{\mu} \to \mathcal{M}^{\mathcal{K}^*}$ separately from the indirect effect associated with the loop $\pi^{\mathcal{K}^*} \to \mathcal{M}^{\mathcal{K}^*} \to \pi^{\mathcal{K}^*}$. To this end, we introduce a cross-world model. Cross-world models specify a joint distribution of processes corresponding to different values of the action $A^{\mu} = a^*$. We introduce the cross-world quantity $\mu^{\mathcal{K}^*,\pi^{\mathcal{K}}}$ indexed by both $\mathcal{K}^*$ and $\mathcal{K}$ to denote the death rate in a world where causes $\mathcal{K} \setminus \mathcal{K}^*$ are eliminated, but where the risk process develops as if all causes were still operating. To achieve this, we need "to run" $\mathcal{M}^{\mathcal{K}}$ simultaneously to drive the risk process $\pi^{\mathcal{K}}$. Note that $\mu^{\mathcal{K},\pi^{\mathcal{K}}} = \mu^{\mathcal{K}}$ and $\mu^{\mathcal{K}^*,\pi^{\mathcal{K}^*}} = \mu^{\mathcal{K}^*}$.

The total effect (5.5.2) may now be decomposed into a part directly attributable to cause removal, i.e. the expected change in μ induced by replacing the set of causes $\mathcal{K}$ with $\mathcal{K}^*$ while keeping the "mediator" fixed at its reference value $\pi^{\mathcal{K}}$, and an indirect effect relayed through the mediating variable. We write

$$\mathrm{TE}(x,t) = \underbrace{\mu^{\mathcal{K},\pi^{\mathcal{K}}}(x,t) - \mu^{\mathcal{K}^*,\pi^{\mathcal{K}}}(x,t)}_{\stackrel{\text{def}}{=} \mathrm{DE(x,t)}} + \underbrace{\mu^{\mathcal{K}^*,\pi^{\mathcal{K}}}(x,t) - \mu^{\mathcal{K}^*,\pi^{\mathcal{K}^*}}(x,t)}_{\stackrel{\text{def}}{=} \mathrm{IE(x,t)}}, \tag{5.5.3}$$

where the natural direct (DE) and indirect (IE) effects are given by

$$\mathrm{DE}(x,t) = \sum_{k \in \mathcal{K} \setminus \mathcal{K}^*} \mu_{0k}(x; C(t)) \sum_{g \in \mathcal{G}} \pi_g^{\mathcal{K}}(x,t) R_{kg}(x), \tag{5.5.4}$$

$$\mathrm{IE}(x,t) = \sum_{k \in \mathcal{K}^*} \mu_{0k}(x; C(t)) \sum_{g \in \mathcal{G}} \left[\pi_g^{\mathcal{K}}(x,t) - \pi_g^{\mathcal{K}^*}(x,t) \right] R_{kg}(x). \tag{5.5.5}$$

We note that (5.5.4) marks the change in μ caused by simply subtracting the death rates of causes $\mathcal{K} \setminus \mathcal{K}^*$ from the all-cause rate without adjusting risk prevalence. This action coincides with the notion of cause removal in the setting of independent competing risks in which elimination does not alter the composition of the surviving population.

5.5.2 Alternative risk prevalence distributions

Another type of intervention deals with the effect on mortality brought about by changing risk prevalence from the reference distribution π to some alternative distribution π^a. We consider a set of interventions $A^\pi = \{A(t)\}_{t\in T}$ for a subset of time points $T \subseteq \tau$. Each $A(t)$ can be represented as a point in the $G-1$ dimensional probability simplex $\{p \in [0,1]^G \mid \sum_{g=1}^G p_g = 1\}$, augmented by an additional state $\emptyset$ that represents no action. We assume that an intervention on $\pi(t)$ is i) not predictive for earlier or remaining contemporaneous variables; and that ii) future variables are unaffected by the intervention other than through past variables.[5] An intervention on π is then represented graphically as in Figure 5.4 Panel (C) with π and $\mathcal{M}$ indexed by the action $A^\pi = a$.

Measuring again the total causal effect on the risk difference scale, we have

$$\mu_k^{\pi^\emptyset}(x,t;C(t)) - \mu_k^{\pi^a}(x,t;C(t)) = \mu_{0k}(x;C(t)) \sum_{g\in\mathcal{G}} \left[\pi_g^\emptyset(x,t) - \pi_g^a(x,t)\right] R_{kg}(x) \tag{5.5.6}$$

for cause k. The total effect can be decomposed in a similar manner to what we did when the intervention was on $\mathcal{M}$. We introduce the cross-world quantity π^{a,a^*} where the action on π is a, but the death rates behave as if it were a^*. The cross-world process $\pi^{a,\emptyset}$ is thus the risk process when the action is a but with the death process developing as if no intervention has been made. Omitting the dependency on C for readability, we can then write the total effect of the action $A^\pi = a$ on cause k as

$$\mu_k^{\pi^\emptyset}(x,t) - \mu_k^{\pi^a}(x,t) = \underbrace{\mu_k^{\pi^{\emptyset,\emptyset}}(x,t) - \mu_k^{\pi^{a,\emptyset}}(x,t)}_{\text{natural direct effect}} + \underbrace{\mu_k^{\pi^{a,\emptyset}}(x,t) - \mu_k^{\pi^{a,a}}(x,t)}_{\text{natural indirect effect}}. \tag{5.5.7}$$

Because all effects are effectively mediated via π itself, the decomposition is complicated to interpret. The direct effect is the effect as if there were no feedback. It describes the change in the death rate following one or more perturbations of the risk prevalence distribution in a world where mortality does not influence risk prevalence. In other words, the risk prevalence prediction is completely unaffected by the fact that the risk composition among those dying in the $\pi^\emptyset$-regime is different from that in the π^a-regime.

The indirect effect describes a self-exciting change to the death process due to it developing differently within the π^a-regime. It is helpful to have a concrete example in mind. Suppose that we intervene on a "marginal" risk factor distribution. For example, consider a situation where the proportion of obese has been substantially increased. Because of competing risks, influencing the risk of any one event will also influence the risk of the remaining (on a population level). Therefore, we expect

[5] Formally, these conditions read $(\overline{\mathcal{M}}(t), \overline{\pi}(t-1), \overline{C}(t)) \perp\!\!\!\perp A(t)$ and $\{\overline{Q}(t+h)\}_{h\in\mathbb{N}} \perp\!\!\!\perp A(t) \mid \overline{Q}(t)$, where $Q(t) = (\mathcal{M}(t), \pi(t), C(t))^\top$.

smokers to die out faster on average compared to the reference scenario, because their risk of death from obesity related causes has increased. Conversely, we expect the death rates of tobacco-attributable causes to drop – even the death rates of causes that are solely attributable to smoking behaviour such as chronic obstructive pulmonary diseases. This is a selection-induced false protectivity phenomena; a seemingly reverse association where an increase in obesity appears protective for tobacco-attributable causes. The feedback mechanism is responsible for capturing such change.

The strength of a model that preserves the feedback mechanism lies in its internal consistency. The system is always able to determine risk prevalence endogenously within the model based on a potential risk composition at projection jump-off, for example given a single initial shock, but is also capable of describing a gradual shift in prevalence towards some target distribution. Using a model for which risk prevalence is exogenous, complex and detailed scenarios are difficult to produce in a consistent manner, as we are only able to quantify mortality *given* all underlying risk factors and their mutual temporal development throughout the entire forecasting region. The problem of obtaining consistent scenarios of competing death rates is then simply recast into a problem of producing a consistent and realistic development of the risk factors.

5.6 An Application to US Data: Illustrating the Direct and Indirect Effects of Cause-of-Death Elimination

We consider an application of the methodology outlined in the previous sections to U.S. risk and mortality data. To keep the exposition concise we restrict the analysis to the SNAP risk factors: smoking, poor nutrition, excess alcohol consumption, and insufficient physical activity. These four modifiable lifestyle related risks are associated with most causes of death.

5.6.1 Data sources

We use the relative risk estimates of Murray et al. (2020), part of the Global Burden of Disease initiative, to describe the link between risk exposure and mortality. The estimates are reported as time homogeneous quantities by sex and 5-year age groups. To get single age estimates we perform linear interpolation with the age bucket centroids as fixed points, see Appendix 5.C.

For smoking the risk-outcome relationship is listed by either current number of cigarettes smoked daily or by pack-years. Pack-years collapses smoking intensity and duration into a single variable, so that we do not have to condition on the entire smoking history of an individual. One pack-year is the equivalent of having smoked one pack of cigarettes (20) a day for a year. For a given sex, age, and risk-outcome

pair the exposure category is listed in jumps of 10. We use natural cubic spline interpolation between categories to obtain a continuous dose-response curve, see Appendix 5.C.

As an indicator for nutritional status we use the Body Mass Index[6] (BMI), which is the dominant metric for categorizing individuals in terms of weight excess or deficiency. The relative risk is reported per five-unit change in BMI with 20 to 25 kg/m^2 being the baseline category. Risks for alcohol consumption are reported directly in terms of grams consumed per day while the relative risk for physical activity is measured in metabolic equivalents (METs) with one MET being the rate of energy expenditure at rest.

Cause of death data is extracted from CDC WONDER (2020) and contains U.S. specific mortality and population data through the years 1999–2018. The data is based on death certificates on which a single underlying cause of death is registered. Matching the data with risks from the GBD study of Murray et al. (2020), we consider in total 35 causes of death known to be influenced by the risk factors. These causes make up about two-thirds of the total age-specific deaths in the population above the age of 35. To obtain an exhaustive list of causes such that the sum of the cause-specific rates equals the all-cause rate, remaining causes are collected and aggregated into a 'residual' category and assigned a relative risk of one for all risk factors.

Risk prevalence data is collected from the IPUMS National Health Interview Survey (NHIS) database (IPUMS, 2019). The NHIS is a large cross-sectional survey conducted annually by the U.S. government and contains comprehensive health and behaviour data at the level of individuals. The IPUMS NHIS data relies on sampling weights to produce representative estimates. Each unit of study can thus be inflated such that the sum of the weighted units constitutes the entire U.S. population. The present analysis is based on adult individuals covering ages 20–84 and years 1999–2018. Observations with missing data are placed into the baseline category. Pack-years of smoking exposure is constructed by assuming that the amount someone currently smokes has not changed since they began smoking. Exposure among former smokers is estimated using years since cessation and average cigarette consumption of the respective cohort. Figure 5.5 shows the evolution of risk prevalence over time.

5.6.2 Baseline model

For modelling purposes we assume that the (true) hazard rate μ_k is constant over the squares $[x, x+1) \times [t, t+1)$ for integer ages x and calendar years t. We consider

[6]Body Mass Index $:= \frac{\text{weight in kilograms}}{(\text{height in meters})^2}$. A BMI below 18.5 is considered underweight and a BMI in the range 25–29.99 is considered overweight. A BMI of 30 or above is classified as obese, subdivided into three categories: 30–34.99 is Class I, 35–39.99 is Class II, and 40 or greater is Class III.

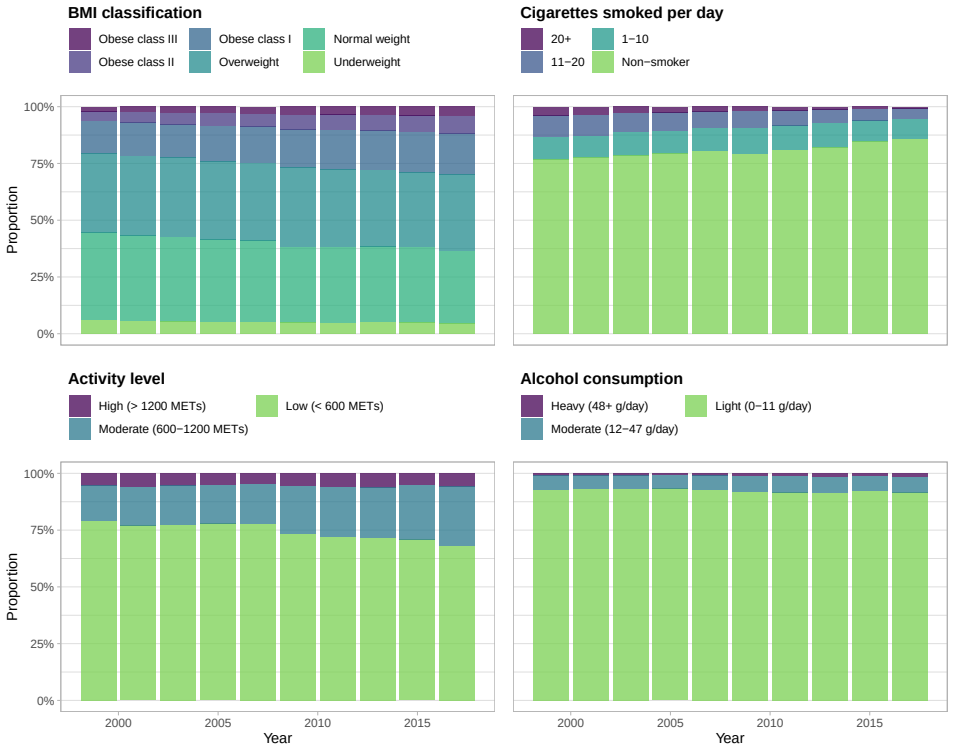

Figure 5.5: U.S. risk proportions of BMI, smoking, alcohol consumption, and physical activity based on IPUMS data for both sexes and ages 20–84. The data shown in the figure is aggregated for the purpose of visual presentation. More granular data is used in the application.

data on the form of cause-specific death counts, $D_k(x,t)$, with corresponding central exposure to risk estimates, $E(x,t)$, and group-wise risk factor prevalence proportions, $\pi_g(x,t)$, over age-time cells in the age-period grid, $\mathcal{R}_{\text{data}}$. From these quantities we can define the empirical cause-specific death rate

$$m_k(x,t) = \frac{D_k(x,t)}{E(x,t)}, \tag{5.6.1}$$

an estimate of the underlying hazard μ_k. Furthermore, we assume to have collected data on $\beta_k(x)$, making $R_k(x,t)$ a known quantity. Consequently, only the baseline in (5.4.3), parametrized in terms of some vector θ, needs to be estimated. This is typically done via maximum likelihood, and it is customary to assume that

$$D_k(x,t) \mid E(x,t), R_k(x,t), C(t) \overset{\text{indep.}}{\sim} \text{Pois}\left(E(x,t)R_k(x,t)\mu_{0k}(x;C(t),\theta)\right). \tag{5.6.2}$$

Contrary to all-cause mortality that is generally well-behaved as a function of age, cause-specific mortality may exhibit several structural changes over the age span.

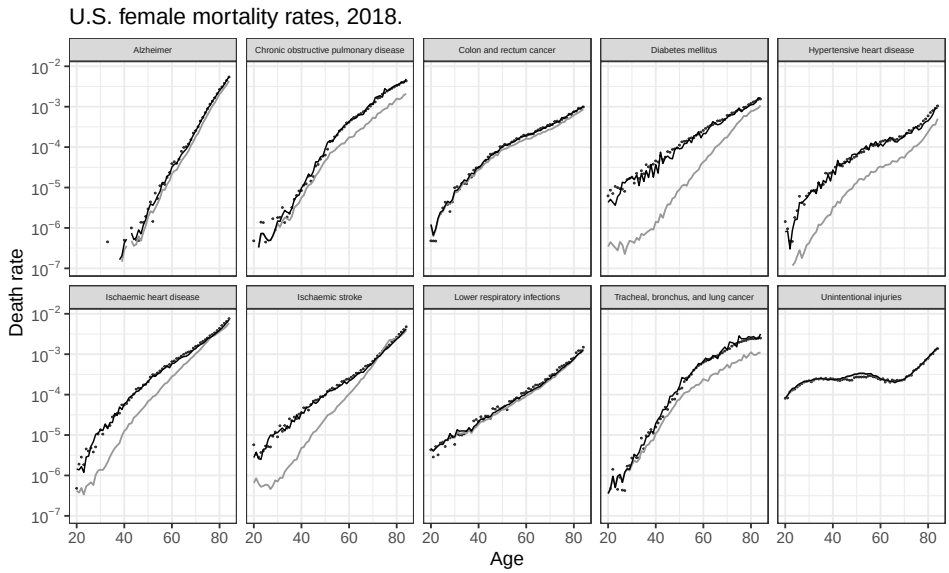

Figure 5.6: Empirical (dotted), fitted (solid black), and baseline (solid grey) death rates for the top 10 leading causes of death in the dataset. The gap between the black and grey lines expresses the excess risk faced by the population due to deviations from baseline levels of exposure in the risk factors considered.

Sudden rapid increases, periods of constancy, and even declines are not unusual. We could in principle use different functional forms to model μ_{0k} depending on cause, however stating just a single parametric form that generalizes well to most settings might be preferable in terms of interpretability. A simple yet widely used parametric form is the log-linear model

$$\mu_{0k}(x, t; C(t), \theta) = \exp\left(\theta_{0kx} + \theta_{1kx} t\right) = \mu_{0k}(x, t; \theta), \tag{5.6.3}$$

which has been applied in settings similar to ours, for instance by King and Soneji (2011) and Foreman et al. (2018). The model is easy to estimate (see Appendix 5.B), flexible enough to capture the different shapes associated with cause-specific mortality, and reflects that age is generally the most important driver of mortality regardless of risk exposure. We use (5.6.3) as the baseline model in what follows. Figure 5.6 shows the empirical female cause-specific death rates for the last year in the estimation period with fits of the aggregate and baseline rates superimposed.

5.6.3 Joint forecasting

Generally speaking, straightforward extrapolative approaches for forecasting risk prevalence are not recommended as they lead to an unabated continuation of historical trends and likely poor out-of-sample performance. Many researches resort to models that are specifically tailored to project the risk prevalence distributions in question,

but existing methods are confined to working on the marginals and do not capture selection-induced feedback effects either. Developing a scalable joint forecasting procedure is an important topic of research, but it is beyond the scope of this paper. For the demonstration we have in mind we make do with a somewhat elementary state-transition model.

We aim at extrapolating the cohorts available in our sample until they reach age $x_{\max}$. We assume that there is no migration in or out of the composite population. Define the (one-step) survival probabilities

$$p_g(x,t) = \exp\left(-\sum_{k=1}^{K} \mu_{0k}(x,t) R_{kg}(x)\right), \tag{5.6.4}$$

for group g and denote by $m_{i,j}(x,t)$ the probability of the cohort aged x at time t changing its "risk-group membership" from i to j. We employ a cohort state-transition model

$$Y(x+1,t+1) = M(x,t)Y(x,t), \tag{5.6.5}$$

with $Y(x,t) \in \mathbb{N}^G$ being the number of individuals in the G groups and $M(x,t) \in [0,1]^{G\times G}$ a matrix of transition probabilities with elements $M_{i,j}(x,t) = p_i(x,t)m_{i,j}(x,t)$. Note that migration rates are only applied to the surviving population. The estimated transition matrices are stated in Appendix 5.D.

Example death rate forecast

Figure 5.7 shows the empirical and forecasted female rates for ischaemic heart disease, the leading cause of death in the dataset, and diabetes for select ages. To gauge the effect of including additional covariates in the forecast, the superimposed dashed lines are reference projections using the baseline model (5.6.3) fitted without additional covariates.

Mortality projections are usually based on empirical regularities such as smooth age profiles and small incremental mortality improvements, but the present projection depends heavily on the cohort specific exposures causing it to exhibit a rather erratic behaviour.[7] Figures 5.6 and 5.7 suggest that this is particularly so for diabetes as a large proportion of mortality is attributable to obesity and cohorts evidently differ substantially in their exposure.

On the other hand, the inclusion of covariates caters for the fact that (baseline) mortality levels ought to be consistently declining over time. While all-cause mortality

[7]For practical applications some smoothing is warranted. The type of smoothness violations seen here prompted the Bayesian modelling approach developed by Girosi and King (2008), applied by e.g. King and Soneji (2011) and Foreman et al. (2018), that down-weighs risk factor information if contradicted by observed empirical patterns. These papers also use smoothed prevalence estimates, whereas we simply apply the raw data.

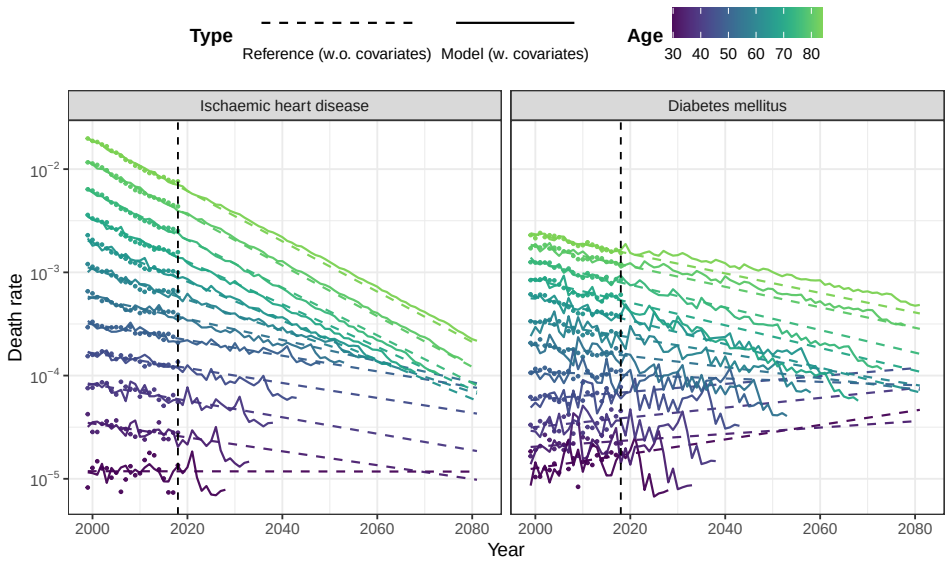

Figure 5.7: Historical (dotted) and estimated and projected (lines) female rates for diabetes mellitus and ischaemic heart disease. The dashed lines are reference projections using the baseline model fitted without additional covariates.

adheres to this pattern, historical cause-specific rates may have actually increased over time – even recently as Figure 5.7 shows – for some causes and ages. This is an ever-present issue widely acknowledged in cause-specific forecasting. The problem is that increasing rates generally do not express that health care and treatment options have worsened, but that mortality improvements have been substantially offset by changes to the risk prevalence distribution. A purely extrapolative model is not able to explain this development and will simply continue the observed trend unabated as seen in the reference forecast for the youngest age groups in Figure 5.7. In the long run this may result in the aggregate all-cause forecast being dominated by the causes that have increased historically.

The causal model with covariate information is, in contrast, equipped to analyze the historic evolution of death rates at a granular level and may disentangle the effects of lifestyle related habits changing from generation to generation from general health care improvements. Indeed, the model successfully separates risk prevalence and mortality in this case by yielding negative slopes for the baseline for all age groups considered in Figure 5.7. This shows that baseline mortality has improved, despite the immediate trend in the raw rates suggesting otherwise, and hints at why cause-specific modelling without additional information or assumptions ought to be avoided.

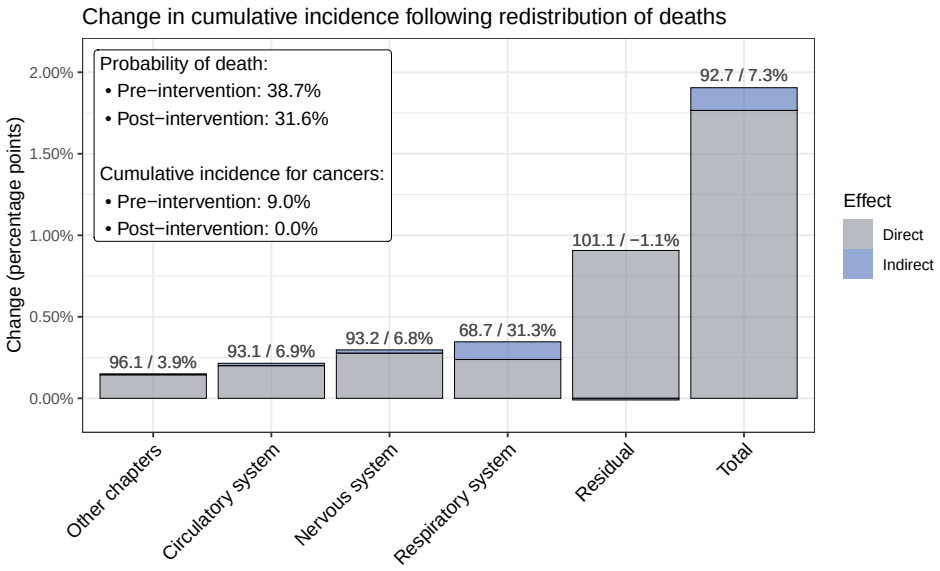

Figure 5.8: Change to the cumulative incidence of select disease chapters following the elimination of deaths due to neoplasms affected by smoking and/or obesity for the female cohort aged 60 in 2018. The bars explain how the 9 percent of the cohort who previously died from cancer are redistributed into other categories. The percentages listed on top of the bars differentiate the part of the change due to direct and indirect effects respectively.

5.6.4 Cause-of-death elimination: What happens if cancer were eradicated?

We now seek to answer the central question posed in the beginning of the paper. If certain causes of death are eradicated, how soon will the individuals "saved" die from something else and what will they die from instead? We illustrate this query using the U.S. dataset by considering an elimination of deaths due to cancers. To answer the questions precisely, we look at the cumulative incidence

$$F_k(u+x \mid x) = \frac{1}{S(x)} \int_0^u S(v+x)\mu_k(v+x)\,\mathrm{d}v, \tag{5.6.6}$$

i.e., the probability of dying from cause k before or at age $u+x$ conditionally on being alive at age x.

Figure 5.8 shows as an example how cumulative incidence for the cohort aged 60 in 2018 ($t_{\max}$) is affected by the intervention. The figure explains the probability of dying before or at age 84 ($x_{\max}$). The cumulative incidences add up to the total probability of death which is 38.7 percent prior to elimination. Cancers make up roughly a quarter of all deaths with the cumulative incidence being 9 percent. An elimination therefore initially causes the total death count to decrease to 29.7 percent of the original cohort, while adjusting for the subsequent redistribution of deaths

brings it up to 31.6 percent. The figure decomposes the redistribution into a part attributable to competing risks and a part due to feedback, using the framework developed and calibrated over the previous subsections.

The decomposition allows us to compare the model with feedback to its non-feedback alternative, namely the same model but with the feedback mechanism disengaged. Without feedback, individuals saved from cancer die according to the rates observed in the population prior to the intervention. Because of competing risks this leads to a rise in the cumulative incidence for every non-eliminated cause (grey bars). This change occurs despite the fact that the corresponding death rates are unaltered. Because the risk prevalence distribution among individuals who previously died from cancer is not the same as that of the general population, there is an additional effect (blue bars). In the model with feedback, higher-than-average-risk individuals are carried forward in the system, causing an increase in the death rates among remaining causes attributed to the SNAP risks. The number of deaths due to diseases of the respiratory system is particularly amplified.

Overall, a simple deletion of cancer death rates that disregards feedback will understate the total probability of death by 7.3 percent among those that are saved and by more than 30 percent at the cause-specific level. These percentages are naturally bounded by the number of risk factors included in the model. As additional risk factors are introduced, and as the departure from population homogeneity becomes more pronounced, selection-induced effects will carry even more weight.

Comparison to other non-feedback alternatives?

One might also take interest in comparing the method we have applied here to other non-feedback alternatives. Such a comparison is, however, beside the point that we are trying to make. We do not claim that our model is superior in predicting the reference scenario compared to other models. In fact, other models with carefully "sculptured" risk prevalence projections likely have better out-of-sample performance compared to the method we have used.

Our focus has been on finding and discussing the magnitude of second-order effects. We have shown that non-feedback models are unable to quantify these, making a comparison between the effect of an intervention based on our proposed method and an alternative somewhat fruitless. Often – especially when it comes to policy making – it is tacitly assumed that second-order effects are small and can be more or less deliberately ignored. However, to reveal the extent of such an assumption, we need a method that allows us to realistically and consistently analyze the impact of interventions. We have detailed how to do so in this paper.

5.7 Concluding Remarks

In this paper we discussed how mortality forecasts were affected by interventions in structural models that link individual risk behaviour to cause-specific mortality. We saw that when these risk mechanisms were specified at the level of populations, the model's ability to relay selection effects hinged on a feedback mechanism controlling how risk prevalence changed in response to differential mortality. We made the point that perturbations of the system only conformed with real-world implementations of interventions when risk prevalence was endogenous to the model.

We considered how death rates changed following the eradication of certain causes of death. The prevalent approach directly manipulates the death rates of interest, with little or no regard for subsequent effects on non-eliminated rates. However, since individuals "saved" cannot be expected to follow the same pattern of mortality as that observed in the population prior to the intervention, these methods are too generous in their estimate of mortality reduction – but by how much? To disentangle and quantify the magnitude of indirect effects we applied techniques from causal mediation theory. This method gave us a straightforwardly interpretable decomposition of the total effect of cause-elimination with a part directly attributable to death rate deletion and a part due to disrupting the selection mechanism. The latter effect is, however, only quantifiable when risk prevalence is endogenous to the mortality model.

From a methodological perspective, our analysis of indirect effects is limited to those induced by changes in behavioural risks. Other health indicators, such as existing or developing medical conditions that have an impact on the length of life, may also be important contributing factors. To give one example, consider the COVID-19 vaccine which is highly effective at preventing serious disease, hospitalization, and death. Those who would have died from COVID without the vaccine may instead have a milder disease course although they could potentially still suffer from 'long COVID'. Such a delayed effect on their risk of death could be modelled using Barker frailty (Palloni and Beltrán-Sánchez, 2017). Those who would have survived even without the vaccine potentially never even contract the disease, thus producing a feedback that raises the vitality of the population. Further, by avoiding overcrowded hospitals due to COVID-related admissions, there could also be an effect on the general access to health care. This example shows that assessing all higher-order effects of an intervention can be extremely challenging and requires a comprehensive modelling framework.

From a practical perspective, mortality models with integrated epidemiological information are still in their infancy. A major challenge when building mortality models that include covariates is the substantial data demand. Data is typically not available at a sufficient granular level to warrant a model at the level of individuals,

and it is in fact rarely the case that a single authoritative source contains a complete set of the covariate distributions of interest, not even at an aggregate level. Instead, researchers often have to collect (aggregate) prevalence data of marginal distributions from multiple sources. In time, however, as the availability and quality of detailed risk data continues to improve, causal models will inevitably gain a footing and contribute to more precise and better substantiated long-term projections of mortality. Moreover, the ability to formulate scenarios of interest in a straightforward and verbal manner is key to engaging non-specialist and making results accessible to a wider audience.

Acknowledgements

The work was partly funded by Innovation Fund Denmark (IFD) under File No. 9065-00135B.

5.A Granger Causality

In the following we give a brief overview of Granger causality and its use for describing conditional (in)dependence relations. For an in-depth account of causal reasoning in (graphical) time series models, we refer the interested reader to Eichler and Didelez (2007, 2010) for the discrete time case and Didelez (2000) for the continuous time analogue.

Granger causality was introduced by Granger (1969) and is a popular tool not only within its origin of econometrics, but also for causal time series analysis. Consider a multivariate time series $Q = \{Q(t)\}_{t\in\mathbb{Z}}$ with $Q(t) = (Q_1(t), \dots, Q_d(t))^\top$. Let $V = \{1, \dots, d\}$ be the index set and define for any $U \subseteq V$ the subprocess $Q_U(t) = (Q_u(t) : u \in U)$. Further, denote by an overbar $\overline{Q}_U(t) = \{Q_U(s)\}_{s\le t}$ the history of the series. Let A and B be two disjoint subsets of V. We say that Q_A is *Granger non-causal* for Q_B up to horizon $h \in \mathbb{N}$ (w.r.t. Q) if

$$Q_B(t+l) \perp\!\!\!\perp \overline{Q}_A(t) \mid \overline{Q}_{V\setminus A}(t), \quad \forall\, l \in \{1, \dots, h\}, t \in \mathbb{Z}. \tag{5.A.1}$$

Here the $\perp\!\!\!\perp$-symbol denotes independence. The formulation (5.A.1) of Granger causality tacitly assumes that all relevant variables for predicting Q are available in Q. This differs from the original formulation in which the information available is that of the "entire universe".

If (5.A.1) holds for $h = 1$, we say that Q_A is Granger non-causal for Q_B and we write $Q_A \nrightarrow Q_B$. Thus, a process Q_A is Granger non-causal for another process Q_B if the past of Q_A up to time t does not give a better prediction of Q_B at time $t+1$ given all information available up to time t but without that of Q_A. If Q_A is Granger non-causal for Q_B at all horizons we write $Q_A \overset{(\infty)}{\nrightarrow} Q_B$.

5.A.1 Graphical representation

We can use the concept of Granger (non-)causality to obtain a graphical representation of the conditional independence relations of the time series. A graph $\mathcal{G} = (V, E)$ consists of a finite set of *nodes* V and a finite set of *edges* E. We only consider graphs containing *directed* edges, that is $E \subseteq V \times V$ is a subset of ordered pairs of nodes. We allow for multiple edges between two nodes if they are of different orientation in which case there is a *loop*.

Instead of a full time graph in which time is made explicit, we are primarily interested in a "rolled" version, also sometimes called a summary graph. To construct such a graph based on the time series Q, we partition the index set V into mutually disjoint subsets $A_1, \ldots, A_q$, $q \leq d$, and associate with the corresponding sub-processes the nodes $V = \{1, \ldots, q\}$. We join two nodes $a, b \in V$ by a directed edge if Q_{A_a} is Granger causal for Q_{A_b} at *some* horizon. Conversely, the absence of an edge implies that $Q_{A_a} \overset{(\infty)}{\nrightarrow} Q_{A_b}$. Although not explicitly shown in the graphs, we assume that all nodes have self-loops. Figure 5.9 gives an example of a rolled graph and a corresponding unrolled version.

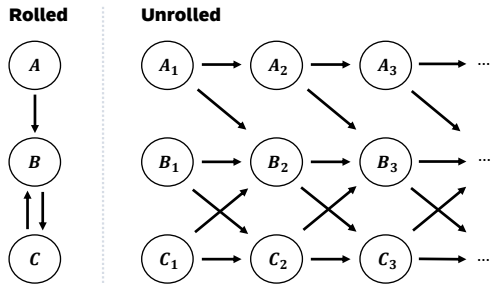

Figure 5.9: An example of a rolling graph with three nodes A, B, and C and a corresponding unrolled version. Since, e.g., $A \to B$ in the rolled version, the unrolled version could contain edges from A_t to B_s for any $s > t$.

5.B Estimation of Baseline Parameters

Suppose we have data on cause-specific death counts, D_k, exposure-to-risk estimates, E, and relative risk coefficients, R_k, each of dimension $d_x \times d_t$ with $d_x = x_{\max} - x_{\min} + 1$ being the length of the age span and $d_t = t_{\max} - t_{\min} + 1$ being the length of the time span. The model (5.6.2) with baseline hazard function (5.6.3) is

$$D_k(x,t) \mid E(x,t), R_k(x,t) \overset{\text{indep.}}{\sim} \text{Pois}\left(E(x,t)R_k(x,t)\exp(\theta_{0kx} + \theta_{1kx}t)\right). \quad \text{(5.B.1)}$$

Since the predictor is linear we have the entire machinery of generalized linear models at our disposal. Using a Poisson error structure, canonical logarithmic link function, and stacking data into column vectors, that is, $d_k = \text{vec}(D_k)$, $e = \text{vec}(E)$ and

$r_k = \text{vec}(R_k)$, we have that

$$\log \mathbb{E}[d_k \mid e, r_k] = \eta + \log(e \circ r_k) \tag{5.B.2}$$

where the latter term on the right-hand side is treated as an offset while $\eta = M\theta$ is the linear predictor with θ being the vector containing the parameters and

$$M = \left[1_{d_t} : (t_{\min}, \ldots, t_{\max})^\top\right] \otimes \mathsf{I}_{d_a} \tag{5.B.3}$$

being the model matrix. In the above, 1_d a d-dimensional vector of ones, I_d a d-dimensional identity matrix, $\circ$ the Hadamard product, and $\otimes$ the Kronecker product.

5.C Interpolation of Relative Risks: Examples

The relative risk estimates of Murray et al. (2020) are reported as risk-outcome pairs by sex, age category, and exposure category. All quantities are time homogeneous.

For every risk-outcome pair the age category is listed in the groups 20–24, 25–29, ..., 90–94, 95–120. To obtain relative risk estimates for every (integer) age, we perform linear interpolation with the age bucket centroids $\mathcal{X} = \{20, 22, 27, \ldots, 92, 107.5, 120\}$ as fixed points. Thus, for fixed risk-outcome pair and sex and an age $x \in [x_0, x_1)$ where x_0 and x_1 are two consecutive numbers in $\mathcal{X}$, the relative risk at age x given by

$$\text{RR}(x_0) + (x - x_0)\frac{\text{RR}(x_1) - \text{RR}(x_0)}{x_1 - x_0},$$

where $\text{RR}(\cdot)$ supplies the relative risk estimate available in the data. An example is given in the left panel of Figure 5.10.

For the risks "Number of cigarettes smoked daily" and "Pack years" the exposure categories are listed in jumps of 10, specifically by 0, 10, 20, 30, 40, 50, or 60 cigarettes per day and 0, 10, ..., 90, or 100 pack years. To obtain a dose-response curve for any (integer) number of cigarettes smoked per day or number of pack years, we extend the exposure categories using natural cubic spline interpolation for fixed sex, age, and cause-of-death. A similar approach is used in the appendix of Murray et al. (2020). An example is given in the right panel of Figure 5.10.

5.D Transition Matrices

Figure 5.5 shows a drift in the prevalence distributions for smoking, obesity and physical activity, whereas alcohol consumption remains roughly constant over the period. We opt for a migration model that captures the main effect, namely the

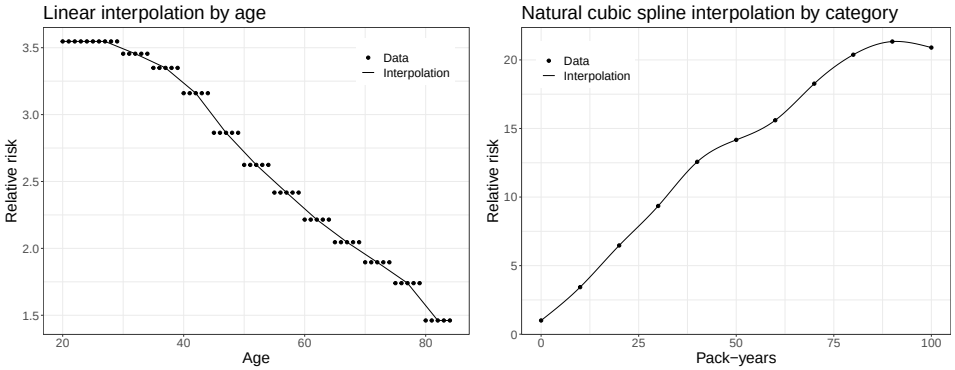

Figure 5.10: Left panel: Female relative risk for diabetes mellitus by age (dots) using linear interpolation with age-bucket centroids as fix points (solid line). Right panel: Age 60 female relative risk for tracheal, bronchus, and lung cancer by pack-years (dots) using natural cubic spline interpolation (solid line).

net migration flow, using just the data at hand. We construct the number of net migration events (NM) by balancing the equation of population change

$$Y(x+1, t+1) = Y(x,t) - D(x,t) + NM(x,t),$$

and by imposing that transitions occur only between neighbouring categories of BMI and physical activity (in one year), while transitions for smoking are described in terms of the probability of cessation. In the above, $Y(x,t)$ is the number of individuals alive at age x and time t while $D(x,t)$ is the number of deaths. We estimate transitions for each risk factor independently of one another and for both sexes and all ages and calendar years combined. The resulting transition matrices are shown below.

Table 5.1: One-year smoking summary transition matrix.

(pct.)	Smoking transition matrix			
	Non-smoker	1–10	11–20	20+
Non-smoker	100	0	0	0
1–10	2.35	97.65	0	0
11–20	4.36	0	95.64	0
20+	9.21	0	0	90.79

Table 5.2: One-year BMI summary transition matrix.

(pct.)	BMI transition matrix				
	Underweight	Normal weight	Overweight	Obese class I	Obese class II-III
Underweight	94.55	5.55	0	0	0
Normal weight	0	97.60	2.40	0	0
Overweight	0	0	98.25	1.75	0
Obese class I	0	0	0	98.23	1.77
Obese class II-III	0	0	0	0	100

Table 5.3: One-year physical activity summary transition matrix in terms of MET-minute-categories given by the cut points $0, 600, 1200, 1800, 2400, 3000, 3600$, and 4200.

(pct.)	Physical activity transition matrix							
	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
Cat. 1	95.37	4.63	0	0	0	0	0	0
Cat. 2	0	100	0	0	0	0	0	0
Cat. 3	0	21.83	78.17	0	0	0	0	0
Cat. 4	0	0	43.43	56.57	0	0	0	0
Cat. 5	0	0	0	19.73	80.27	0	0	0
Cat. 6	0	0	0	0	14.52	85.48	0	0
Cat. 7	0	0	0	0	0	3.69	95.73	0.58
Cat. 8	0	0	0	0	0	0	0	100

Chapter 6

Aggregated Structural Causal Models

This chapter contains the manuscript *Jallbjørn and Hansen (2022)*.

Abstract

Most approaches to causal inference assume a single dataset of i.i.d. observations covering all variables. However, in practice, samples from the joint distribution of two or more variables may not be available. In this paper, we consider the situation where data consists of density estimates on marginal distributions of variables observed across multiple populations, and discuss how heterogeneity amongst these can be leveraged in concert with causal knowledge to construct a joint causal model. For this purpose, we introduce the basic ideas of how structural causal models at the level of individuals can be transferred into structural causal models at the level of populations, specifically in terms of the marginal distributions of the variables. The resulting population-level models will be called aggregated structural causal models, which we argue are causally consistent with their individual level analogues. We present an algorithm for determining a directed acyclic graph that represents the distributions necessary for formulating the aggregated model as a structural one.

Keywords: *Causality, Aggregation, Causal Consistency, Policy Interventions, Ecological Inference.*

6.1 Introduction

Causal modelling is used within a wide range of disciplines to predict the behaviour of a system when subject to external manipulation (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017). An increasingly important application concerns the quantification of population-based interventional strategies, for example predicting the efficacy of behaviour change in context of public health decision-making (WHO, 2009). However, obtaining representative data samples to enable identification of causal relationships is seldom feasible when dealing with populations on, say, a national scale. In these cases, interventions are studied under a presumed causal structure, justified by systematic reviews and meta-analyses, and the effects are calibrated to publicly available registry data. Since these data sources generally do not carry granular information on the joint distribution of all the required variables, we are tasked with inferring interventional effects from aggregate, population-level data.

As a motivating example, suppose that we want to assess the burden of mortality from heart diseases attributable to the combined effects of various risk factors. One can find a wealth of evidence in the literature on both direct and distal causes as well as potential confounders, see Figure 6.1. The figure illustrates the relationship between the distributions of the background variables that act as confounders, the modifiable lifestyle risks that comprise our interventional target and the physiological response variables through which effects of the behavioural risks are mediated. If we want to quantify the effect of lowering the number of smokers, say, on the prevalence of ischaemic heart disease, we need estimates of the conditional distributions corresponding to the edges in Figure 6.1. It is, however, often difficult or impractical to

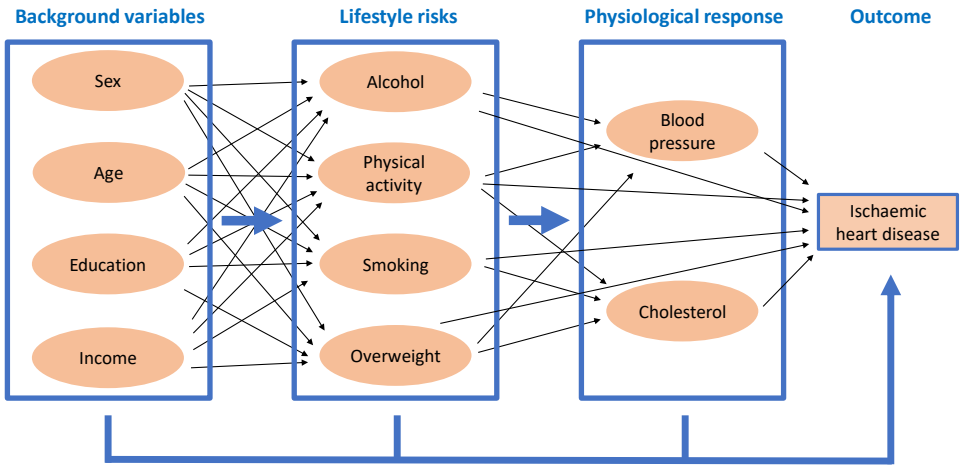

Figure 6.1: The causal links between some but not all variables associated with ischaemic heart diseases, adapted from WHO (2009).

obtain a single dataset covering the full joint distribution, but multiple datasets for the marginal distributions may be available from registries and surveys – and such data are often available for several spatially or temporally separated populations. Moreover, data are often aggregated and thus only available in the form of tables of relative frequencies or in the form of histograms – the latter being an empirical estimate of a density. Models of relative frequencies, also known as compositional data, or densities arise in many contexts, see for example Petersen et al. (2022) and the references therein. The goal of this paper is to leverage such aggregated data for causal inference – and specifically clarify under which assumptions such causal inference is valid. We make two contributions.

Our first contribution is a simple, but valuable and practical, regression procedure for estimating causal effects from data on the marginal distributions alone, provided that we are willing to assume the causal structure and have access to datasets containing measurements from different populations. Considering the problem of estimating the effect of intervening on smoking prevalence in Figure 6.1, then since we "know" the causal structure, we show how our procedure is able to come up with a more qualified coupling of these distributions than the independence coupling typically used in epidemiological applications (Ezzati et al., 2003).

Our second contribution is a novel aggregated causal model. Based on the ideas exploited by the regression procedure, we ask if it is possible to cluster the variables in a causal model in such a way that aggregate data on each cluster, but across populations, is sufficient for estimating the entire causal model. To this end we relate the individual level causal model to the population level aggregated causal model. This way of looking at model abstraction is fundamentally different compared to what is typically done. We propose that the aggregation should happen at the abstract level of probability distributions and not at the level of variables. An immediate advantage is that we achieve causal consistency between the micro- and macro level in the sense that both models agree on the interventional distributions they entail. In contrast, this is hardly ever possible when aggregation happens at the level of variables (Rubenstein et al., 2017; Beckers and Halpern, 2019).

Considering Figure 6.1 again, the nodes represent at the individual level variables. If we have access to the joint distribution for certain pairs of variables, for example, the ones contained in the blue boxes, we can collapse these into single cluster nodes. From aggregated data separately on background variables, lifestyle risks, physiological responses and prevalence of ischaemic heart diseases we are able to obtain a causal model at the population level that can predict how interventions of these cluster nodes (changing the distribution of risk factors in the population, say) affect the prevalence of ischaemic heart diseases.

6.1.1 Related work

Ecological Inference The problem of reconstructing a joint probability distribution given several marginal distributions observed across different populations is a version of the ecological inference problem (Greenland and Robins, 1994; Morgenstern, 1995; Wakefield, 2008). The solution we suggest in the discrete case in Section 6.3 is strongly linked to the seminal method of Goodman (1953, 1959). Modern solutions to the ecological inference problem are mostly rooted in Bayesian methods (King, 1997; Wakefield, 2004; Flaxman et al., 2015). Recently, the problem has also been phrased as an optimal transport problem in which the transportation plan is the unknown one seeks to recover (Muzellec et al., 2017; Frogner and Poggio, 2019).

Merging Data using Causal Information Causal models have previously been suggested as a way to merge information from different datasets (Schölkopf et al., 2012; Peters et al., 2016; Mooij et al., 2020). Tsamardinos et al. (2012) coined the term *integrative causal analysis* for when a joint causal model is constructed on the basis of multiple heterogeneous datasets covering only aspects of the full distribution. Related work considers learning equivalence classes of graphs compatible with data (Triantafillou et al., 2010; Tillman and Spirtes, 2011; Hyttinen et al., 2013), merging marginal SCM's (Janzing, 2018; Gresele et al., 2022), and obtaining a joint model using the principle of maximum entropy (Mejia et al., 2022). In contrast to this body of work, we do not require observations on overlapping subsets of variables. Also, we are not trying to learn the causal structure in this work, but instead assume that it is known through, e.g., various partial studies.

Causal Abstraction and Consistency A question that often arises when causal consequences have to be understood in terms of aggregated features of micro-level data, is whether or not two causal models describing the same system at different levels of granularity yield conclusions that are consistent with each other (Chalupka et al., 2015, 2016). Recent approaches have developed a formalism in terms of variable transformations to answer whether such causal coarsenings or abstractions are sensible (Rubenstein et al., 2017; Beckers and Halpern, 2019). However, in practice, only few transformations yield causally consistent representations. This fact has led to the development of the related notion of approximate abstraction, where the abstracted model only approximates the underlying system up to some error (Beckers et al., 2019). The concept of causal consistency, and the resulting abstraction error, has also been formulated and studied using a category-theoretic approach (Rischel and Weichwald, 2021; Otsuka and Saigo, 2022). We return to the topic later in the paper.

6.1.2 Outline

After presenting the relevant notions of causal modelling (§ 6.2), we consider how causal knowledge may be leveraged to estimate a joint distribution from marginal density data (§ 6.3). We then discuss how causal models at the level of individuals can be transferred into causal models at the level of populations in terms of marginal distributions of the variables, and explicate the causal consequences (§ 6.4). Finally, we end on some concluding remarks (§ 6.5).

6.2 Causal Graphical Models

Throughout, let $(\Omega, \mathcal{F}, P)$ be a common background probability space. We consider a collection of random variables $X = (X_v : v \in V)$ indexed by a finite set V and a probability measure $\mathbb{P}$ over X. Each X_v takes values in a measurable space $(\mathcal{X}_v, \mathcal{A}_v)$ and $\mathbb{P}$ is defined on the Cartesian product of these spaces $(\mathcal{X}, \mathcal{A}) := (\times_{v\in V}\mathcal{X}_v, \otimes_{v\in V}\mathcal{A}_v)$. Even though the points made in this paper can be generalized to any marginally continuous[1] $\mathbb{P}$, we assume for the sake of exposition that the state spaces $\mathcal{X}_v$ are all discrete with the counting measure on $\mathcal{X}$ serving as the dominating measure. In particular, each X_v then has distribution $\mathbb{P}_v$ with probability mass function $p(x_v) = P(X_v = x_v)$. We will use the shorthand notation $(\mathcal{X}_I, \mathcal{A}_I) = (\times_{v\in I}\mathcal{X}_v, \otimes_{v\in I}\mathcal{A}_v)$ for any non-empty subset $I \subseteq V$. Similarly, we will write $X_I = (X_v : v \in I)$ with possible values $x_I \in \mathcal{X}_I$.

Causal Graphical Model (CGM) We assume that $\mathbb{P}$ is Markovian and faithful with respect to a directed acyclic graph (DAG) $\mathcal{D} = (V, E)$ with vertex set V equal to the indices of X and edge set E. That is, given disjoint sets $A, B, C \subseteq V$ then $A \perp_{\mathcal{D}} B \mid C \iff X_A \perp\!\!\!\perp X_B \mid X_C$, where $\perp_{\mathcal{D}}$ denotes d-separation in $\mathcal{D}$, see for example Pearl (2009). Consequently, $\mathbb{P}$ factorizes so that its probability mass function p has the form

$$p(x) = P(X = x) = \prod_{v\in V} p\left(x_v | x_{\mathrm{pa}(v)}\right), \tag{6.2.1}$$

where $\mathrm{pa}(v) := \{j \in V : j \to v\}$ are the graphical parents of v and $x_v \mapsto p(x_v|x_{\mathrm{pa}(v)})$ is the probability mass function of the conditional distribution $X_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}$.

The triple $(X, \mathcal{D}, \mathcal{P})$ defines a causal graphical model, where $\mathcal{P} = (\mathcal{P}^v)_{v\in V}$ is the collection of conditional distributions $\mathcal{P}^v = (\mathbb{P}_v(\cdot|X_{\mathrm{pa}(v)} = x))_{x\in\mathcal{X}_{\mathrm{pa}(v)}} = \mathbb{P}_{v|\,\mathrm{pa}(v)}$. Importantly, the graphical relations are equipped with a causal interpretation in the sense that each $\mathcal{P}^v$ represents a stochastic assignment of X_v according to the values of its parents. This assignment process remains invariant under perturbations that do not affect X_v.

[1] Marginal continuity of $\mathbb{P}$ means that $\mathbb{P}$ is absolutely continuous with respect to a product measure $\nu = \otimes_{v\in V}\nu_v$ for σ-finite measures ν_v on $(\mathcal{X}_v, \mathcal{A}_v), v \in V$. When a dominating measure for $\mathbb{P}$ has been determined, the density of $\mathbb{P}$ is defined by the Radon-Nikodym derivative $\mathrm{d}\mathbb{P}/\,\mathrm{d}\nu$.

Interventions The causal interpretation induces interventional distributions which we, for the applications we have in mind, define in general form, see, for example, Peters et al. (2017, Definition 6.32). At the level of individuals, using Pearl's do-operator, e.g. Pearl (2009), an intervention is an action $\text{do}(X_k = x_k)$ that fixes X_k to a target value x_k. It is unreasonable to only consider interventions at the level of populations where all individuals get assigned the same fixed value of the variable(s) we intervene upon. More realistically, we get to assign values to individuals from a distribution with positive variance. A general intervention on X_k leads to a product decomposition similar to (6.2.1) except that the term $p(x_k|x_{\text{pa}(k)})$ is replaced by $q(x_k|x_{\widetilde{\text{pa}}(k)})$, that is,

$$p^{\text{do}\left(X_k:=q(\cdot|x_{\widetilde{\text{pa}}(k)})\right)}(x) = q(x_k|x_{\widetilde{\text{pa}}(k)}) \prod_{v\in V\setminus\{k\}} p(x_v|x_{\text{pa}(v)}), \tag{6.2.2}$$

with $\sum_{x_k\in\mathcal{X}_k} q(x_k|x_{\widetilde{\text{pa}}(k)}) = 1$ and the (possibly) modified parents not introducing any cycles in the graph. Further, we denote by $\mathbb{I}^{\text{Ind}}$ the set of all possible interventions in the CGM, and we let $\leq_X$ be the partial ordering in which $i \leq_X j$ for $i, j \in \mathbb{I}^{\text{Ind}}$ if and only if i intervenes on a subset of the nodes that j intervenes on and assigns them the same distributions as j.

Graph terminology We briefly introduce some additional graphical terminology needed later. We write $\text{ch}(v) := \{j \in V : v \to j\}$ to denote the children of $v \in V$. For $I \subseteq V$, we define the expressions $\text{Pa}(I) = \cup_{i\in I} \text{pa}(i) \setminus I$ and $\text{Ch}(I) = \cup_{i\in I} \text{ch}(I) \setminus I$. We use the notation $\text{an}(I)$ to denote all the ancestors of I *not* containing I itself, and $\text{An}(I) = \text{an}(I) \cup I$ to denote the ancestors containing I. When not clear from context, we will add a subscript to identify the graph to which the expressions refer, for example, $\text{pa}_{\mathcal{D}}(v)$ refers to the parents of v in $\mathcal{D}$. A bijective mapping $\pi : V \to \{1, \ldots, |V|\}$ is said to be a topological ordering if $\pi(\alpha) < \pi(\beta)$ whenever β is a descendant of α, for nodes $\alpha, \beta \in V$. Due to acyclicity, every DAG has at least one topological ordering, and we tacitly assume throughout that V (and any subset $I \subseteq V$) is ordered in this way.

6.3 Aggregated Regression

In order to compute the effect of an intervention we need the causal graph as well as the conditional distributions in (6.2.1). If we know the causal graph we can estimate the conditional distribution of X_v given its parents from a dataset containing i.i.d. observations from the joint distribution $\mathbb{P}_{\{v\}\cup\text{pa}(v)}$. In practice, we may not have such data. In the example illustrated by Figure 6.1 we have data on lifestyle risks from one dataset, while data on physiological responses are from another dataset. In this section, we consider the situation where we have access to multiple datasets at an aggregated level but separately for a node and its parents, e.g., annual data on the prevalence of high cholesterol and separately annual data on the distribution of

lifestyle risks. In general, we suppose that we have access to empirical estimates from multiple populations of the two marginal[2] distributions $\mathbb{P}_v$ and $\mathbb{P}_{\mathrm{pa}(v)}$. Since there exist many joint distributions on $\mathcal{X}_v \times \mathcal{X}_{\mathrm{pa}(v)}$ that match such marginals, further assumptions must be imposed to obtain a unique solution. The solution we present is based on assuming that the variations between populations can be described by independent noise variables.

6.3.1 A Coupling via Independent Noise

We suppose that we observe the marginals across multiple (heterogeneous) populations, and we explicate how population-level noise enters the system by a functional characterization of the data generating process in terms of each parent-child relation in $\mathcal{D}$. We do so with a structural causal model $\mathcal{M} = (\nu_{\varepsilon,\eta}, \mathcal{S})$, consisting of a distribution $\nu_{\varepsilon,\eta}$ over $\varepsilon = (\varepsilon_v : v \in V)$ and $\eta = (\eta_v : v \in V)$ and $|V|$ structural assignments $\mathcal{S}$:

$$X_v := F_v\left(X_{\mathrm{pa}(v)}, \varepsilon_v, \eta_v\right), \quad v \in V. \tag{6.3.1}$$

To emphasize the hierarchical interpretation, we should think of the ε's as variations at the individual level and the η's as variations across populations. Each F_v represents a stable mechanism that does not change across populations (unless specifically altered by an intervention). We will assume independence of all ε's and η's, that is, $\nu_{\varepsilon,\eta}$ is a product measure.

Fixing η, the structural assignments define a distribution $\mathbb{P}^\eta$ of X indexed by η. Here and throughout, the η superscript is fixed and specific in each population but varies between populations. The main idea is to exploit the variation in the observed marginal distributions $\mathbb{P}^\eta_v$ and $\mathbb{P}^\eta_{\mathrm{pa}(v)}$ across populations to estimate (sufficient aspects of) the F_v–map to make it possible to reconstruct $\mathbb{E}(\mathbb{P}^\eta_{\{v\}\cup\mathrm{pa}(v)})$.

Inferring the complete F_v–map from population-level data in a discrete setting would require restrictive assumptions on the data generating process. However, fixing η and $X_{\mathrm{pa}(v)}$ in (6.3.1), observe that F_v determines the conditional distribution in (6.2.1) by

$$\begin{aligned} F_v(x_{\mathrm{pa}(v)}, \nu_{\varepsilon_v}, \eta_v)(A) &= \nu_{\varepsilon_v}\left(F_v(x_{\mathrm{pa}(v)}, \cdot, \eta_v)^{-1}(A)\right) \\ &= P^{\eta_v}\left(X_v \in A \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}\right), \end{aligned} \tag{6.3.2}$$

for $A \in \mathcal{A}_v$. Thus, in the discrete setting, the problem of inferring a joint distribution amounts to estimating a number of conditional probabilities.

An application of the law of total probability in combination with (6.3.1) implies

$$p^\eta(x_v) = \sum_{x_{\mathrm{pa}(v)} \in \mathcal{X}_{\mathrm{pa}(v)}} p^{\eta_v}\left(x_v | x_{\mathrm{pa}(v)}\right) p^\eta\left(x_{\mathrm{pa}(v)}\right), \tag{6.3.3}$$

[2]We refer to any distribution that can be obtained by marginalizing the full joint distribution of X as a marginal distribution.

where p^η is used as short-hand for the probability mass function for fixed η. Since we have access to empirical versions of $\mathbb{P}^\eta_v$ and $\mathbb{P}^\eta_{\mathrm{pa}(v)}$, the linear structure of (6.3.3) invites for a regression-based approach to determine the conditional $\mathbb{P}^{\eta_v}_{v|\,\mathrm{pa}(v)}$. Since $p^\eta\left(x_{\mathrm{pa}(v)}\right)$ only depends on $(\eta_j)_{j\in\mathrm{an}(v)}$, $p^{\eta_v}\left(x_v|x_{\mathrm{pa}(v)}\right)$ and $p^\eta\left(x_{\mathrm{pa}(v)}\right)$ are independent by independence of the η's, whence a regression of observations of $\mathbb{P}^\eta_v$ on observations of $\mathbb{P}^\eta_{\mathrm{pa}(v)}$ yields an estimate of the expected conditional distribution

$$\int p^{\eta_v}(x_v|x_{\mathrm{pa}(v)})\nu_{\eta_v}(\mathrm{d}\eta_v). \tag{6.3.4}$$

In the special case where η_v is degenerate, and where the conditional distribution thus does not depend on η, the regression will identify the conditional distribution common across populations. In such a case the conditional distribution is said to be invariant;

$$\left(X^\zeta_v|X^\zeta_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}\right) \overset{d}{=} \left(X^\psi_v|X^\psi_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}\right), \tag{6.3.5}$$

for any two populations ζ and ψ and all $x_{\mathrm{pa}(v)} \in \mathcal{X}_{\mathrm{pa}(v)}$. This might be a reasonable assumption in some scenarios. Indeed, (6.3.5) is the assumption that underlies the principle of invariant causal prediction formulated by Peters et al. (2016). In such a setting, one considers samples (X^η, Y^η) where X^η is a vector of predictor variables for some target Y^η. Equation (6.3.5) is then only assumed to hold for the target variable given its causal predictors.

In an observational setting, we would not generally expect (6.3.5) to hold for all parent-child relations. Allowing for some variation means that we only identify the expectation of the conditional distributions across populations, but this expectation might still be a decent quantity to use with (6.3.3); in certain applications we may be convinced that the conditional for a given population deviates only little from the average. For example, in the context of Figure 6.1, it might be reasonable to assume that most of the variation in the lifestyle risks observed across populations is due to variation in demographic and socio-economic structures. That is, variation in $\mathbb{P}^\eta_{\mathrm{Lifestyle}}$ is due mainly to variation in $\mathbb{P}^\eta_{\mathrm{Background}}$ and only to a small extent to variation in $\eta_{\mathrm{Lifestyle}}$. In any case, the dispersion around the mean is picked up the regression residuals, which could be used to provide some variability bands.

As for the regression, the coefficients should represent probabilities and thus fall in the unit interval to enhance model interpretability. This can be achieved by solving the constrained non-negative least squares problem[3]

$$\begin{aligned}\hat{\beta} := \quad & \arg\min_{\beta\succeq 0} \tfrac{1}{2}||Z\beta - y||_2^2 \\ & \text{Subject to } C\beta = \mathbf{1}_{|\mathcal{X}_{\mathrm{pa}(v)}|},\end{aligned} \tag{6.3.6}$$

[3]The least squares problem (6.3.6) can be solved by recasting it as a quadratic program under the same positivity and equality constraints but with objective $\frac{1}{2}\beta^\top Q\beta - d^\top\beta$ where $Q = Z^\top Z$ and $d = Z^\top y$.

where $y = \text{vec}([\hat{p}^\eta(x)]_{\eta \in H, x \in \mathcal{X}_v})$ contains the empirical estimates of $p^\eta(x_v)$ for each population and value in the domain, while $Z = \boldsymbol{I}_{|\mathcal{X}_v|} \otimes [\hat{p}^\eta(x)]_{\eta \in H, x \in \mathcal{X}_{\text{pa}(v)}}$ is the design matrix. Here, the vec-operator stacks the columns of a matrix into a vector, $\otimes$ denotes the Kroenecker product, $\boldsymbol{I}_d$ is an identity matrix of size d, $\mathbf{1}_d$ a d-dimensional vector of ones, while $C = \mathbf{1}^\top_{|\mathcal{X}_v|} \otimes \boldsymbol{I}_{|\mathcal{X}_{\text{pa}(v)}|}$ specifies the sum constraint.

Alternatively, the regression coefficients can be parametrized in such a way that they automatically satisfy the desired constraints. The multinomial-logit parametrization achieved through the softmax-function

$$\beta_i = \frac{\exp(\tilde{\beta}_i)}{\sum_{j \in \mathcal{J}_i} \exp(\tilde{\beta}_j)}, \tag{6.3.7}$$

produces non-negative regression coefficients summing to unity where appropriate. Under (6.3.7), the least-squares objective $\frac{1}{2}||Z\beta(\tilde{\beta}) - y||_2^2$ is non-linear as the regression coefficients become a function of the free parameters $\tilde{\beta}$.

Ties to Ecological Regressions

The ecologic regression model proposed by Goodman (1953, 1959) is equivalent to the contents of (6.3.6) without constraints on the coefficients. Goodman, however, does not justify why the regression estimates ought to represent a conditional probability distribution apart from asserting that the model should only be used "*in very special circumstances*". Framing the problem in terms of knowledge about the causal graph as done here gives one way of rationalizing whether or not the regression produces a valid approximation. We emphasize that the assumption of independent η's is key to justify the regression procedure.

An alternate solution to the ecological inference problem is the method of bounds proposed by Duncan and Davis (1953). Although not a unique solution, observing a set of marginals deterministically restricts the set of joint distribution that are consistent with them, implying that the conditional probabilities we seek to estimate may be bounded more narrowly than the unit interval. This suggests that the box constraint of (6.3.6) could be modified to convey the deterministic bounds for the conditional probabilities. Determining cell bounds for k-way contingency tables is described in the statistical data privacy literature, see Dobra and Fienberg (2001) and Dobra and Fienberg (2009).

6.3.2 Numerical Case-Study: Risk Factor Interventions

In this section, we present some numerical results for the example from the introduction, that is, we consider the system described by the DAG in Figure 6.1. For the risk factors, we use U.S. data from IPUMS (2022), NCD-RisC (2017), and CDC (2022), which, after preprocessing, consists of dichotomous marginal distributions, see Appendix 6.E. The data covers the period 1999–2018, that is, 20 different populations.

We use (6.3.6) to determine the missing conditional distributions. Using the logistic parametrization (6.3.7) yields virtually the same results for this data and is therefore not shown.

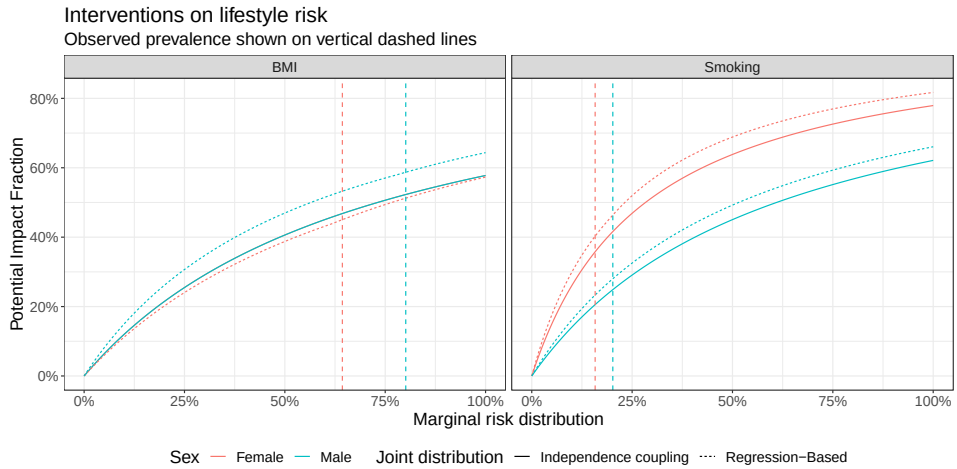

Figure 6.2: PIFs for U.S. females and males aged 50–54 in 2018. Using the estimates of Murray et al. (2020), the relative risks for BMI are the same for both sexes, causing the two PIF curves to coincide when (6.3.8) is evaluated under an assumption of independent marginals.

The burden of disease from certain lifestyle related risks is defined as an interventional query through the potential impact fraction (PIF) (Ezzati et al., 2003):

$$p \mapsto \frac{\mathbb{E}_p[\mathrm{RR}] - \mathbb{E}_{\tilde{p}}[\mathrm{RR}]}{\mathbb{E}_p[\mathrm{RR}]}, \tag{6.3.8}$$

where RR is the relative risk at a given level of exposure, while p and $\tilde{p}$ describe the distributions under which the expectation is evaluated. Using the relative risk estimates of Murray et al. (2020), we compute (6.3.8) in Figure 6.2 for interventions on the smoking and BMI marginals. This is slightly ambiguous as an intervention replaces the entire conditional distribution $\mathbb{P}_{\mathrm{Lifestyle}|\mathrm{Background}}$. We discuss how to intervene on multiple nodes simultaneously in Section 6.4.5. For now, the interventions are implemented as a proportional change to the original values. The reference distribution $\tilde{p}$ is defined as the theoretical minimum exposure distribution as is customary, that is, $\mathbb{P}_{\mathrm{Smoke}} = \delta_{\mathrm{Non\text{-}smoker}}$ in the smoking scenario and $\mathbb{P}_{\mathrm{BMI}} = \delta_{\mathrm{Normal}}$ in the BMI scenario, with δ denoting the Dirac measure.

We generally expect the lifestyle risks to exhibit positive correlation as they are all affected by the same background variables. Moreover, part of their effect on the outcome is mediated through 'blood pressure' and 'cholesterol'. Thus, we expect the independence coupling to understate the risk burden, which is also what Figure 6.2 generally shows; the estimated PIF on the dotted line is able to take the joint

effect of the risks into account, although, at least in this case where the number of variables and categories are low, the difference between evaluating (6.3.8) under the independence assumption and the proposed method is not drastic.

6.3.3 A Mixture-Model Example

Even though we generally focus on discrete distributions, it is illuminating to think about how the idea from the discrete setup translates to a setting with parametric assumptions, in which case the objective is to identify the causal parameters of the model.

As an instructive example, consider the structural equations

$$\begin{aligned}
X_1 &\sim \text{Bern}(p(\eta_1)) \text{ with } p(\eta_1) = P(X_1 = 1 \mid \eta_1),\\
X_2 &:= \beta_{21} X_1 + \sigma_2 \varepsilon_2 + \tau_2 \eta_2,\\
X_3 &:= \beta_{31} X_1 + \beta_{32} X_2 + \sigma_3 \varepsilon_3 + \tau_3 \eta_3,
\end{aligned}$$

where $\varepsilon_2, \varepsilon_3, \eta_2, \eta_3$ are i.i.d. $\mathcal{N}(0,1)$ variables (independent of X_1). Suppose that samples from the joint distribution $\mathbb{P}_{\{1,2,3\}}$ are not available, but that we instead have density data on the marginals $(\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3)$ across multiple populations.

Written out, we have samples from

$$\begin{aligned}
\mathbb{P}_1^\eta &= p(\eta_1)\delta_1 + (1 - p(\eta_1))\delta_0,\\
\mathbb{P}_2^\eta &= p(\eta_1)\mathcal{N}(\beta_{21} + \tau_2\eta_2, \sigma_2^2) + (1 - p(\eta_1))\mathcal{N}(\tau_2\eta_2, \sigma_2^2),\\
\mathbb{P}_3^\eta &= p(\eta_1)\mathcal{N}(\beta_{31} + \beta_{32}(\beta_{21} + \tau_2\eta_2) + \tau_3\eta_3, \beta_{32}^2\sigma_2^2 + \sigma_3^2)\\
&\quad + (1 - p(\eta_1))\mathcal{N}(\beta_{32}\tau_2\eta_2 + \tau_3\eta_3, \beta_{32}^2\sigma_2^2 + \sigma_3^2),
\end{aligned}$$

where δ denotes the Dirac measure. In particular, we imagine data of the form X_{1ih} for individuals $i = 1, \ldots, n_{1h}$ and populations $h = 1, \ldots, H$. To simplify, let us assume that the populations are from the same location but at different time points $t_1, \ldots, t_H$. We note that the data will generally *not* be on the same individuals, so we do not have time series data at the level of individuals. The empirical distribution of X_1 can be represented as

$$\hat{p}(t_h) = \frac{1}{n_{1j}} \sum_{i=1}^{n_{1h}} X_{1ih},$$

which form a time series of probability estimates. For X_{2ih} we have a similar data structure, and for each time point, t_h, we have samples from the mixture

$$p(t_h)\mathcal{N}\left(\beta_{21} + \tau_2\eta_2(t_h), \sigma_2^2\right) + (1 - p(t_h))\mathcal{N}\left(\tau_2\eta_2(t_h), \sigma_2^2\right).$$

From the empirical distribution of $X_{21h}, \ldots, X_{2n_{2h}h}$ we can fit the parameters β_{21}, $\tau_2\eta_2(t_h)$ and σ_2^2 that enter into the mixture distribution. We can use the empirical

distribution of $X_{11h}, \ldots, X_{1n_{1h}h}$ to determine the mixing coefficients $p(t_h)$, so these do not have to be estimated. The time series of $\widehat{\tau_2\eta_2}(t_h)$ can be used to estimate τ_2, but just from the parameters estimated at time point t_h we have enough information to deduce the F_2-map and thus the coupling of $\mathbb{P}_1$ and $\mathbb{P}_2$. This is, of course, due to the strong assumptions of Gaussian distributions, homogeneous variance, and additive effects. For $\mathbb{P}_3$, however, there is not enough information in a single time point to estimate all parameters. There are three unknown parameters in the mean values, and, fitting a marginal mixture distribution, only two equations to determine them from. Here, we really need the time series structure to identify all parameters.

6.4 Aggregating the Causal Model

We will now discuss how to transfer a causal model at the level of individuals into a causal model at the level of populations in terms of the marginal distributions of X. The resulting aggregate model may be used to formulate interventional queries concerning the distribution of individual-level behaviour. In particular, we will argue that conclusions drawn from the aggregate model will be consistent with the ones drawn from the individual level model.

The techniques outlined in the previous section describe how and when conditional distributions may be estimated from marginal density data. We do not necessarily need to find these conditionals for every $v \in V$. It may suffice to determine conditionals of certain bundles of variables, namely clusters of parents. We can then specify the causal relation among the clusters, but leave their internal structure unspecified.

Let a bar over a family of sets $\mathcal{I}$ denote its union, that is, $\overline{\mathcal{I}} := \cup_{I\in\mathcal{I}} I$. To specify an "aggregated" graph $\mathcal{D}^*$ that still represents $\mathbb{P}$, we introduce the following properties:

P1 (Coverage) $\mathcal{I} = (I_k)_{k\in\mathcal{K}}$ is a family of non-empty subsets $I_k \subseteq V$ s.t. $\overline{\mathcal{I}} = V$.

P2 (Parent preservation) $A \cup \mathrm{Pa}_{\mathcal{D}}(B) \subseteq \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}$ where $A \subseteq I$ and $B = I \setminus A$ for every $I \in \mathcal{I}$.

P3 (Running intersection) For any node J on a path between I and I' in $\mathcal{D}^*$ then $I \cap I' \subseteq J$, where $I, I', J \in \mathcal{I}$.

P4 (Reduction) $\mathcal{I}$ is reduced, that is, $\nexists I, I' \in \mathcal{I}$ such that $I \subseteq I'$.

Definition 6.4.1 (Parental Decomposition)**.** *A parental decomposition (PD) of a DAG $\mathcal{D} = (V, E)$ is a DAG $\mathcal{D}^* = (\mathcal{I}, \mathcal{E})$ satisfying* P1–P4.

The motivation behind this construction is to specify how we can meaningfully collapse nodes into "cluster-nodes". In particular, observe that:

Lemma 6.4.2. *If $\mathcal{D}$ is a DAG, then $\mathcal{D}$ is a PD of itself.*

By meaningful collapse, we mean that if a given DAG $\mathcal{D}^*$ satisfy the above properties, then the joint distribution factorizes according to $\mathcal{D}^*$ into conditional distributions of clusters of parents given other clusters, like in the individual-level model (6.2.1). However, because we do not require $\mathcal{I}$ to be composed of *disjoint* subsets of V, see for example the middle panel of Figure 6.4, the computations are a bit more cumbersome. The factorization reads:

Proposition 6.4.3. *If $\mathbb{P}$ factorizes according to $\mathcal{D} = (V, E)$, then for a PD $\mathcal{D}^* = (\mathcal{I}, \mathcal{E})$ of $\mathcal{D}$ it holds that*

$$p(x) = \frac{\prod_{I\in\mathcal{I}} p(x_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}|x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}})}{\prod_{v\in V} p(x_v|x_{\mathrm{pa}_{\mathcal{D}}(v)})^{|C(v)|-1}} = \frac{\prod_{I\in\mathcal{I}} p(x_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}|x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}})}{\prod_{v\in V} p(x_v|x_{C_v^{\mathrm{pa}}})^{|C(v)|-1}}, \tag{6.4.1}$$

for all $x \in \mathcal{X}$, where $C_v = \{I \in \mathcal{I} : v \in I, v \notin \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}\}$ and $C_v^{\mathrm{pa}} = \{\cap_{I\in C_v} \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}\} \cup \{\cap_{I\in C_v}\{\alpha \in I : \pi(\alpha) < \pi(v)\}\}$.

In Definition 6.4.1, we require $\mathcal{I}$ to be reduced to avoid superfluous nodes; otherwise some distributions could be specified as marginalizations of others. For example, it would in a sense be redundant to include $\mathbb{P}_1$ if $\mathbb{P}_{\{1,2\}}$ is already included.

6.4.1 The Aggregated Structural Causal Model

The PD of an arbitrary DAG is obviously not unique. For example, there is always a trivial decomposition consisting of a single node containing the entire vertex set $\mathcal{I} = \{V\}$ and no edges. More meaningful representations of the conditional independences encoded in $\mathcal{D}$ depend on the type of analysis we have in mind.

If we are only interested in interventions on specific variables, the PD can be chosen to reflect this. For example, suppose that we want to quantify the causal influence of a treatment X on a target Y in the presence of a confounder Z that influences both X and Y, and another confounder C that influences Z and Y. In this case, the interventional distribution can be calculated using backdoor adjustment,

$$p^{\mathrm{do}\,(X=q(\cdot|z))}(y) = \sum_{x,z} p(y|x,z)q(x|z)p(z), \tag{6.4.2}$$

without reference to C. That is, if we want to intervene on X it suffices to condition on Z.

Typically the focus is on the more general – and more ambitious – goal of obtaining a model that allows us to quantify all interventions at once via a structural representation. For this purpose, let us introduce the additional property

P5 (Connectivity) For every $I \in \mathcal{I}$ then $J \perp_{\mathcal{D}} J'$ for all $J, J' \in \mathrm{pa}_{\mathcal{D}^*}(I)$ with $J \neq J'$.

Definition 6.4.4. *A PD$^+$ of a DAG $\mathcal{D} = (V, E)$ is a DAG $\mathcal{D}^* = (\mathcal{I}, \mathcal{E})$ satisfying* P1–P5.

We can define a structural causal model for the probability measures $(\mathbb{P}_I^\eta : I \in \mathcal{I})$ on such a graph, making $\mathbb{P}_I^\eta$ a function of its graphical parents and exogenous (noise) variables η_I.

Definition 6.4.5 (Aggregated Structural Causal Model)**.** *Let $\mathcal{M}^{\mathrm{Ind}} = (\nu_{\varepsilon,\eta}, \mathcal{S}^{\mathrm{Ind}})$ be an SCM for X with associated DAG $\mathcal{D}$ and let $\mathcal{D}^*$ be a PD$^+$ of $\mathcal{D}$. An aggregated structural causal model (ASCM) $\mathcal{M}^{\mathrm{Agg}} = (\nu_\eta, \mathcal{S}^{\mathrm{Agg}})$ consists of the product distribution ν_η and $|\mathcal{I}|$ structural assignments $\mathcal{S}^{\mathrm{Agg}}$:*

$$I \in \mathcal{I}: \quad \mathbb{P}_I^\eta(x_I) := G_I\left(\mathbb{P}^\eta_{\mathrm{pa}_{\mathcal{D}^*}(I)}, \eta_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}\right)(x_I), \quad x_I \in \mathcal{X}_I, \tag{6.4.3}$$

where

$$G_I\left(\mathbb{P}^\eta_{\mathrm{pa}_{\mathcal{D}^*}(I)}, \eta_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}\right)(x_I) = \sum_{x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}\setminus I}} p^{\eta_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}}\left(x_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}} | x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}\right) \cdot \prod_{J\in\mathrm{pa}_{\mathcal{D}^*}(I)} p^\eta(x_J). \tag{6.4.4}$$

Intuitively, G_I arises by fixing η in (6.3.1), the structural equations for X, and computing the distribution of X_I as transformations of ε. A general formulation of the ASCM is possible using compositions of Markov kernels, cf. Appendix 6.A. We return to how the PD$^+$ can be chosen in Section 6.4.4.

By construction, the ASCM replicates the observational distribution over X implied by the SCM. The joint distribution can be computed by noticing that

$$G_I\left(\otimes_{J\in\mathrm{pa}_{\mathcal{D}^*}(I)}\delta_{x_J}, \eta_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}\right)(x_I) = p^{\eta_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}}\left(x_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}} | x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}\right), \tag{6.4.5}$$

and using (6.4.1). Since $\mathcal{D}^*$ is a PD$^+$, the denominator in (6.4.1) is only relevant when $\mathcal{D}^*$ consists of multiple components that contain overlapping subsets of variables.

6.4.2 Interventions

An intervention $\mathrm{do}(i)$ in the ASCM maps $\mathcal{M}^{\mathrm{Agg}}$ to $\mathcal{M}^{\mathrm{Agg};\mathrm{do}(i)}$ by replacing the structural assignments for $\mathbb{P}_I^\eta$. We denote the entailed interventional distribution $\mathbb{P}^{\eta;\mathrm{do}(i)}$.

To illustrate the basic principles for interventions in the ASCM, consider as an example the SCM

$$X_1^\eta := F_1(\varepsilon_1, \eta_1), \quad X_2^\eta := F_2(X_1^\eta, \varepsilon_2, \eta_2), \quad X_3^\eta := F_3(X_1^\eta, X_2^\eta, \varepsilon_3, \eta_3),$$

for which a corresponding ASCM reads

$$\mathbb{P}^{\eta}_{\{1,2\}} := G_{\{1,2\}}\left(\eta_{\{1,2\}}\right), \quad \mathbb{P}^{\eta}_{3} := G_3\left(\mathbb{P}^{\eta}_{\{1,2\}}, \eta_3\right).$$

At the level of individuals, interventions are defined by fixing, for example, $X_1 = x_1$, which results in the modified SCM

$$X_1^{\eta} := x_1, \quad X_2^{\eta} := F_2(x_1, \varepsilon_2, \eta_2), \quad X_3^{\eta} := F_3(x_1, X_2^{\eta}, \varepsilon_3, \eta_3),$$

while an intervention on X_2 results in

$$X_1^{\eta} := F_1(\varepsilon_1, \eta_1), \quad X_2^{\eta} := x_2, \quad X_3^{\eta} := F_3(X_1^{\eta}, x_2, \varepsilon_3, \eta_3).$$

In the ASCM, intervening on either X_1 or X_2 is done using a distribution $\mathbb{P}^{\mathrm{do}}_{\{1,2\}}$, which corresponds to modifying (a number of) conditional distributions:

$$\begin{aligned}
&\text{Intervention on } X_1 : \mathbb{P}^{\eta}_{\{1,2\}} := \mathbb{P}^{\mathrm{do}}_{1} \otimes G^{*}_{\{1,2\}}\left(\mathbb{P}^{\mathrm{do}}_{1}, \eta_2\right),\ \mathbb{P}^{\eta}_{3} := G_3\left(\mathbb{P}^{\eta}_{\{1,2\}}, \eta_3\right),\\
&\text{Intervention on } X_2 : \mathbb{P}^{\eta}_{\{1,2\}} := G^{*}_{\{1,2\}}\left(\eta_1\right) \otimes \mathbb{P}^{\mathrm{do}}_{2} \qquad ,\ \mathbb{P}^{\eta}_{3} := G_3\left(\mathbb{P}^{\eta}_{\{1,2\}}, \eta_3\right).
\end{aligned}$$

Not all interventions in the ASCM are meaningful. There may exist – arguably hypothetical – interventions in the aggregated model for which there is no individual-level analogue. For example, in the bottom DAG in the middle panel of Figure 6.4, we could mathematically impose $\mathbb{P}_{\{1,2\}} := \mathbb{P}^{\mathrm{do}}_{\{1,2\}}$ and $\mathbb{P}_{\{1,3\}} := \mathbb{P}^{\mathrm{do}}_{\{1,3\}}$ that do not agree on their respective $\mathbb{P}_1$–marginals. We want to avoid this situation and restrict attention to the set of interventions induced by the individual-level model, namely those that fulfil the deterministic constraints of the system.

Assumption 6.4.6. *The distributions $(P^{\eta}_I : I \in \mathcal{I})$ implied by an ASCM $\mathcal{M}^{\mathrm{Agg};\mathrm{do}(i)}$, $i \in \mathbb{I}^{\mathrm{Agg}}$, through (6.4.3) satisfy for all $v \in V$ that for any pair $I, I' \in \mathcal{I}$ both containing $v \in V$, that is $\{v\} \subseteq I$ and $\{v\} \subseteq I'$, then*

$$\mathbb{P}^{\eta}_I(A_v \times \mathcal{X}_{I\setminus\{v\}}) = \mathbb{P}^{\eta}_{I'}(A_v \times \mathcal{X}_{I'\setminus\{v\}}), \quad A_v \in \mathcal{A}_v,$$

with $\mathbb{P}^{\eta}_I$ and $\mathbb{P}^{\eta}_{I'}$ defined on $\mathcal{X}_v \times \mathcal{X}_{I\setminus\{v\}}$ and $\mathcal{X}_v \times \mathcal{X}_{I'\setminus\{v\}}$, respectively.

Thus, the set of admissible interventions in the aggregate model $\mathbb{I}^{\mathrm{Agg}}$ consists of interventions for which nodes with overlapping subsets of variables agree on their implied marginals.

6.4.3 Causal Consistency

Rubenstein et al. (2017) and Beckers and Halpern (2019) developed the notion of causal consistency, which describes whether two models at different levels of abstraction can sensibly be thought of as models of the same system. If the models

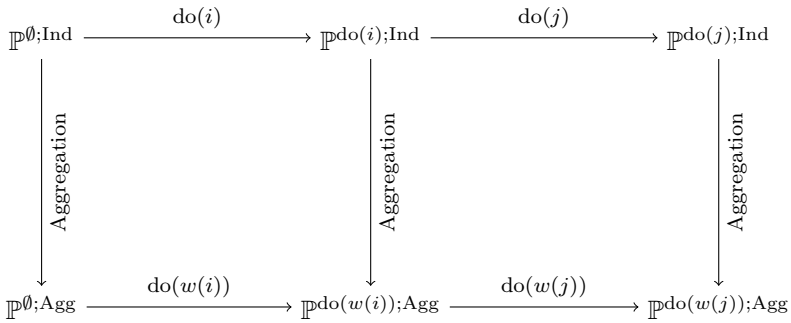

Figure 6.3: Schematic of causal consistency whereby the same conclusions are reached for both the original and the aggregated model. In Rubenstein et al. (2017) the 'aggregation' arrows correspond to the function τ, whereas they represent the transformation from SCM to ASCM in this paper. The distribution $\mathbb{P}^{\emptyset}$ refers to the observational distribution, while $i, j \in \mathbb{I}^{\text{Ind}}$ are interventions such that $i \leq_X j$.

are consistent, we expect them to produce the same interventional distributions regardless of the order in which we do the aggregation and the interventions.

Formally, Rubenstein et al. (2017) defines causal consistency between two structural causal models with variables X and X' as commutativity of the implied probability distributions on a (specific) set of partially ordered perfect, deterministic interventions[4] $\mathbb{I}^{\text{Ind}}$:

$$\tau\left(\mathbb{P}^{\text{do}(i);\text{Ind}}\right) = \mathbb{P}^{\text{do}(w(i));\text{Agg}}, \quad \forall i \in \mathbb{I}^{\text{Ind}}, \tag{6.4.6}$$

where $\tau : X \to X'$ is a variable mapping and $w : \mathbb{I}^{\text{Ind}} \to \mathbb{I}^{\text{Agg}}$ a surjective, order-preserving map between interventions. When (6.4.6) holds then the diagram in Figure 6.3 commutes (Rubenstein et al., 2017, Thm. 6), and the operations of intervening and aggregating are interchangeable. That is, performing an intervention $j \in \mathbb{I}^{\text{Agg}}$ in the aggregate model is equivalent to computing the push-forward measure of the distribution resulting from any intervention $i \in w^{-1}(\{j\})$ in the individual-level model using τ. Order-preservingness of w ensures that consistency also applies to compositions of interventions, that is, the right square in Figure 6.3 commutes.

The above notion of consistency is, however, quite restrictive and does not extend well to realistic settings. As τ maps variables into aggregate summaries, most models, especially non-linear ones, incur an irreversible loss of information when transferred to the aggregate level so that actions cannot be back-transformed to the individual level without substantial ambiguity. This has led to the development of the related notion of approximate abstraction or consistency (Beckers et al., 2019; Rischel and Weichwald, 2021). As the name suggests, the requirement on the transformations is

[4]With perfect, deterministic interventions the target distribution of X_I is a one-point measure δ_{x_I}.

relaxed so that the aggregate model captures the underlying system only up to some error.

With the ASCM, we look at model abstraction in a fundamentally different way in the sense that the aggregation happens at the level of distributions rather than at the level of variables. This means that the only detail that is "abstracted away" is the specific labelling of individuals. So long as we are not concerned with which particular unit is assigned what particular value $x_I \in \mathcal{X}_I$, but only that individuals are assigned according to the correct target distribution, this loss of information is irrelevant.

While the transformation from SCM to ASCM cannot be defined in terms of a variable mapping τ, and so does not fit into Equation (6.4.6), consistency between the two models can still be defined as the diagram in Figure 6.3 commuting. Because $\mathbb{I}^{\text{Agg}}$ is restricted to interventions induced by the SCM, we have the following lemma.

Lemma 6.4.7. *Under Assumption 6.4.6, there exists a surjective, order-preserving map* $w : \mathbb{I}^{\text{Ind}} \to \mathbb{I}^{\text{Agg}}$ *such that* $\mathbb{P}^{\text{do}(i);\text{Ind}} = \mathbb{P}^{\text{do}(w(i));\text{Agg}}, \ \forall i \in \mathbb{I}^{\text{Ind}}$.

Thus, all interventions in the ASCM have at least one corresponding intervention in the SCM, and consistency follows as a corollary.

Corollary 6.4.8. *Under Assumption 6.4.6, the ASCM is causally consistent with the SCM in the sense that the diagram in Figure 6.3 commutes.*

6.4.4 Constructing a Generic PD$^+$

In the following, we describe the most important steps to Algorithm 1 which details one way of constructing a PD$^+$ of an arbitrary DAG $\mathcal{D}$. We label the resulting graph the aggregated DAG, or ADAG for short, and denote it by $\mathcal{D}^+$. The ADAG is constructed in two steps. In the first step we find PD$^+$'s of each sink node $s \in \{v \in V : \text{ch}(v) = \emptyset\}$, namely directed rooted trees, where s is the root and all edges point towards it (i.e., an anti-arborescence). Next, $\mathcal{D}^+$ is formed by the join of the resulting trees, including a pruning step to remove superfluous nodes. Additional details are given in Appendix 6.B.

Algorithm 1 can be viewed as a greedy algorithm. For some graphs, there exists decompositions that offer an improvement over the ADAG in terms width, $\max_{I \in \mathcal{I}} |I|$. Relaxing the requirement on how to choose the parents of a given $I \in \mathcal{I}$ does, however, introduce a tricky path dependence whereby the optimal choice at a given step depends on subsequent options. DAGs of similar structure will then have different optimal aggregated versions depending on the number of nodes at each level in the graph, see Appendix 6.D for examples.

Algorithm 1: DAG to ADAG

Input : A DAG $\mathcal{D} = (V, E)$.
Output : The ADAG $\mathcal{D}^+ = (\mathcal{I}, \mathcal{E})$ of $\mathcal{D}$ w.r.t. π.

$\mathcal{S} \leftarrow$ sink nodes of V;
for $s \in \mathcal{S}$ **do**
  $\mathcal{T}_s \leftarrow$ TIBAL of s;
  $\mathcal{T}_s \leftarrow$ `separateLineage`$(\mathcal{T}_s)$;
end
return `JoinAndReduce`$_\pi\left((\mathcal{T}_s)_{s\in\mathcal{S}}\right)$

For the first step, we use a recursive pairing strategy to build PD$^+$'s, revolving around the following sets.

Definition 6.4.9 (Ancestral Lineage)**.** *Let $s \in \{v \in V : \mathrm{ch}(v) = \emptyset\}$ be a terminal vertex in $\mathcal{D}$. For each such s, the ancestral lineage is defined recursively by the sets*

$$I_{k,s} = \{\mathrm{Pa}(I_{k-1,s} \setminus A) \cup A : A = \mathrm{An}(\mathrm{Pa}(I_{k-1,s})) \cap I_{k-1,s}\}, \tag{6.4.7}$$

for $k = 1, \ldots, L_s$, with $I_{0,s} = \{s\}$, and L_s being the depth of the lineage so that $I_{L_s+n,s} = \emptyset$ for any $n \in \mathbb{N}_+$.

Starting from the bottom of the causal hierarchy, (6.4.7) traverses through the graph generating for each I-node another node consisting of the union of parents. Importantly, variables that reappear in a parent set at a higher level are always retained, see for example the left panel of Figure 6.4.

Definition 6.4.10 (TIBAL)**.** *Let $\mathcal{I}_s = (I_{k,s})_{k\in\{0,\ldots,L_s\}}$ be the ancestral linage defined by a sink node s of $\mathcal{D}$. The tree implied by the ancestral lineage (TIBAL) is $\mathcal{T}_s = (\mathcal{I}_s, \mathcal{E}_s)$ with vertices $\mathcal{I}_s = (I_{k,s})_{k\in\{0,\ldots,L_s\}}$ and edges $\mathcal{E}_s = ((I_{k,s}, I_{k-1,s}))_{k\in\{1,\ldots,L_s\}}$.*

The ancestral lineage covers the sets we are after in the sense that a TIBAL of a sink s is a PD$^+$ of $\mathcal{D}_{\mathrm{An}(s)}$.

Proposition 6.4.11. *If $\mathcal{D} = (V, E)$ has exactly one terminal vertex and $\mathcal{T} = (\mathcal{I}, \mathcal{E})$ is the tree implied by the ancestral lineage, then $\mathcal{T}$ is a PD$^+$ of $\mathcal{D}$.*

Because the ancestral lineage does not take unconditional separation statements in the original graph into consideration we need an additional step to ensure that these are properly displayed in $\mathcal{T}$. For instance, if the input graph describes a three variable V-structure $1 \to 3 \leftarrow 2$, we want $\mathcal{T}$ to show that $\mathbb{P}_3$ is a function of the marginals $(\mathbb{P}_1, \mathbb{P}_2)$ only since $\mathbb{P}_{\{1,2\}} = \mathbb{P}_1 \otimes \mathbb{P}_2$. This leads to the function `separateLineage` described in Appendix 6.B that improves the PD$^+$ in terms of width, $\max_{I\in\mathcal{I}} |I|$.

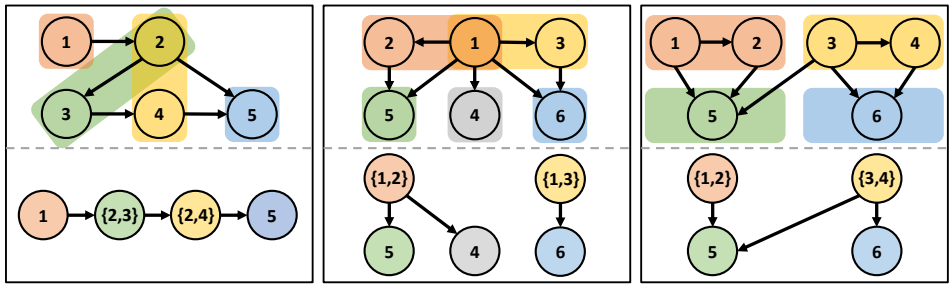

Figure 6.4: Examples of how DAGs (top graphs) transfer to ADAGs (bottom graphs). *Left panel:* The ancestral lineage retains variables appearing in multiple clusters. *Middle panel:* A DAG for which the corresponding ADAG is determined up to a topological ordering; the 4-node can be a "child" of either $\{1, 2\}$ or $\{1, 3\}$. If there was an additional arrow $2 \to 3$ in the DAG, we prefer the arguably simpler graph structure $\{1, 2\} \to 4$ over the construction $\{1, 2\} \to \{1, 3\} \to 4$. *Right panel:* A DAG in which the unconditional separation statements require special treatment.

Lemma 6.4.12. *If $\mathcal{D}$ has exactly one terminal vertex and $\mathcal{T}$ is the tree implied by the ancestral lineage, then* `separateLineage`$(\mathcal{T}) =: \tilde{\mathcal{T}} = (\mathcal{I}, \mathcal{E})$ *is a PD$^+$ of $\mathcal{D}$ satisfying $\forall I \in \mathcal{I} : \nexists I' \subsetneq I$ such that $I' \perp_{\mathcal{D}} (I \setminus I')$, making $\tilde{\mathcal{T}}$ the PD$^+$ of smallest possible width under the restriction that the sink of $\mathcal{D}$ is also a sink in $\tilde{\mathcal{T}}$ and that P2 reads: $A \cup \mathrm{Pa}_{\mathcal{D}}(B) = \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}$ where $A \subseteq \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I)) \cap I$ and $B = I \setminus A$ for every $I \in \mathcal{I}$.*

Next, the ADAG is formed by joining each such tree of sinks in $\mathcal{D}$. For this purpose we define the operation JoinAndReduce$_\pi$ that merges together any number of input trees while removing all vertices that are contained in other vertices. The subscript π refers to the order in which reduction is performed as for given $I \in \mathcal{I}$ there may be multiple sets $I' \in \mathcal{I}$ that satisfy $I \subseteq I'$. However, the pruning operation is not inherently "safe" and there are a few non-trivial pitfalls to avoid in order to not create cycles or produce a graph that violates P3 or P5, see Appendix 6.B.

Proposition 6.4.13. *Algorithm 1 produces a PD$^+$ of an input DAG $\mathcal{D}$.*

If only the marginal distribution of each variable is needed to define the aggregated model, this is what the ADAG shows.

Proposition 6.4.14. *If $\mathcal{D} = (V, E)$ is a DAG whose underlying undirected graph is a forest, then the ADAG $\mathcal{D}^+$ is equal to $\mathcal{D}$ in the sense that $\mathcal{D}^+ = (V, E)$.*

This property suggest a natural use case for the aggregated model; when the conditional independence structure is a tree, we do not require measurements on the simultaneous distribution for *any* pair of variables to fit the ASCM. Obtaining the data needed to fit the population-level model may therefore be considerably easier than collecting the data required to fit the individual-level SCM.

Separation statements in the ADAG are tricky since two I-nodes can have shared variables among their ancestors and descendants even if the nodes appear in different connected components, see for example the middle panel in Figure 6.4.

Proposition 6.4.15. *If $\mathcal{D}^+ = (\mathcal{I}, \mathcal{E})$ is the ADAG determined from $\mathcal{D}$ and $A, B, C \subseteq \mathcal{I}$ are three mutually reduced families of sets, then*

$$A \perp_{\mathcal{D}^+} B \mid C \Longrightarrow (\overline{A} \setminus (\overline{C} \cup D)) \perp_{\mathcal{D}} (\overline{B} \setminus (\overline{C} \cup D)) \mid (\overline{C} \cup D),$$

where $D = \mathrm{IS}\left(\overline{\mathrm{An}_{\mathcal{D}^+}(A)}, \overline{\mathrm{An}_{\mathcal{D}^+}(B)}, \overline{\mathrm{An}_{\mathcal{D}^+}(C)}\right) \setminus \overline{\mathrm{an}_{\mathcal{D}^+}(\{J \in C : J \in \mathrm{An}_{\mathcal{D}^+}(\{A, B\})\})}$ *with* $\mathrm{IS}(A, B, C) = (A \cap B) \cup (A \cap C) \cup (B \cap C)$.

6.4.5 Case-Study: Risk Factor Interventions (Continued)

We continue with the example from Section 6.3.2, and follow up on the idea that relations among nodes in a cluster do not necessarily have to be explicitly modelled. In particular, we want to demonstrate how one can make sensible interventions on the lifestyle cluster-node of Figure 6.1.

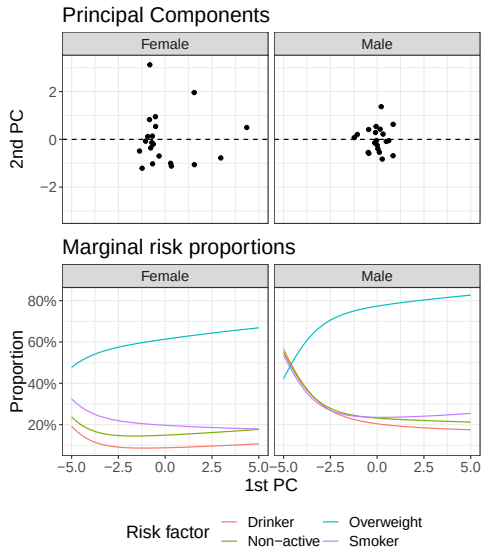

Figure 6.5: PCA for the ilr-transformed risk factor data (top panel). The bottom panel shows the induced prevalence distributions when intervening in the direction of the first principal component (dashed line, top panel).

The idea we pursue here is to apply a principal components transformation to the data, and intervene in the direction of the first component that describes the direction of maximum variance (dashed line, top panel, Figure 6.5). Because the data is compositional, we need to apply an appropriate transformation to it before applying PCA. That is, we have to transform the data from the standard simplex to the real space. The usual transformations for this purpose are either the centred-log-ratio transformation (Aitchison, 1986) or the isometric-log-ratio transformation (Egozcue, 2003). We use the latter, see Appendix 6.F. For dimensionality reduction, only the first two principal components are kept. Once the intervention has been applied in the principal component space, the results are transformed back to the original space.

Figure 6.5 shows the results for the risk factor data. Based on the induced marginal distributions shown in the bottom panel, it is clear that the first principal component

picks up the secular trend in the data. People are becoming increasingly overweight over time, while other harmful lifestyle behaviours are reduced. Thus, an intervention in the direction of the first principal component can be phrased as the question: "what if the current development in lifestyle behaviour continues?".

6.5 Conclusion

We have shown how a structural causal model may be specified at the level of probability distributions rather than at the level of variables, and that conclusions drawn from the aggregated model are consistent with the ones from its individual level analogue. We have shown that the aggregated model can be estimated from marginal density data observed across multiple populations given knowledge about the causal graph. The proposed regression-based approach was formulated in a discrete variable setting, motivated by applications to data from statistics bureaus that provide discretized marginal distributions, that is, contingency tables of low dimension.

Even with low dimensional data, the number of explanatory variables used to estimate a given conditional probability may be substantial compared to the number of available data points (i.e., populations) and may lead to inefficient prediction. One might therefore want to study a regularized version of the problem, while remaining aware of whether the regression coefficients have an appropriate interpretation (as conditionals). For example, for dimension reduction, we could apply PCA to the explanatory variables and use only a subset of the principal components as regressors.

A direction for future work would be to explicate solutions for when the variables studied are continuous. With parametric assumptions this amounts to identifying individual-level parameters from aggregated data by leveraging the heterogeneity induced through the population-level noise variables. We gave an example of how to approach the problem in a linear structural causal model, namely by writing out the marginal distributions for which we have data. Still, it is not obvious how one can best exploit all of the available information. Ideally, we want a top-down strategy for estimating the parameters, taking as a starting point the variables at the highest level of the causal hierarchy, and iteratively work our way down through the system. The problem is then how we can estimate the parameters entering the structural equation for $\mathbb{P}_v$, while also taking into account data on (or estimated parameters for) its causal parents.

On the practical side, it remains an interpretational challenge how one would translate an actual policy change into a $\mathbb{P}^{\text{do}}$–intervention. Indeed, in the case study of risk factor interventions, we did not specify how the (joint) interventions arise, but simply examined the consequences of bringing $\mathbb{P}_{\text{Lifestyle risks}}$ to $\mathbb{P}^{\text{do}}_{\text{Lifestyle risks}}$. For the purposes of policy-planning, associating a monetary cost with the possible

interventions is another interesting direction for further research. This would open for cost-effectiveness analyses in which one could compare the cost of a gain in a pre-specified health outcome for different interventions. For example, this would enable us to answer a questions like: "given that our interventional target is the set of modifiable lifestyle risks, what is then the intervention that produces the most life years gained at the cheapest cost?".

Acknowledgements

The work was partly funded by Innovation Fund Denmark (IFD) under File No. 9065-00135B.

6.A The ASCM in the General Case

Suppose that X is generated according to an SCM with structural equations

$$X_v = F_v(X_{\mathrm{pa}(v)}, \varepsilon_v, \eta_v), \quad v \in V, \tag{6.A.1}$$

where the joint distribution $\nu_{\varepsilon,\eta} = \bigotimes_{v\in V}(\nu_{\varepsilon_v} \otimes \nu_{\eta_v})$ over the noise variables $\varepsilon = (\varepsilon_v)_{v\in V}$ and $\eta = (\eta_v)_{v\in V}$ is a product distribution. We assume that ν_{ε_v} and ν_{η_v} are dominated by either the counting measure if the respective state space is discrete or the Lebesgue measure otherwise for every $v \in V$. Thus, $\nu_{\varepsilon,\eta}$ is marginally continuous and all distributions have densities by the Radon-Nikodym theorem.

6.A.1 Markov Kernels

To formulate the SCM at the level of probability distributions, the notion of a Markov kernel proves useful. The following exposition is based on Rønn-Nielsen and Hansen (2014). Let $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}, \mathcal{B})$, and $(\mathcal{Z}, \mathcal{C})$ be measurable spaces. A Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$ is a family of probability measures $(P^x)_{x\in\mathcal{X}}$ on $(\mathcal{Y}, \mathcal{B})$ such that for every fixed $B \in \mathcal{B}$ the map $x \mapsto P^x(B)$ is $\mathcal{A}$-measurable. If $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and P is a Markov kernel from X_2 to $\mathcal{Y}$, we can extend P to be a Markov kernel $\tilde{P}$ from $\mathcal{X}$ to $\mathcal{Y}$ as $P^{x_2} = \tilde{P}^{(x_1,x_2)}$. We do not distinguish between P and its extension. Further, we let $P \circledast Q$ denote the composition of Markov kernels P from $\mathcal{X}$ to $\mathcal{Y}$ and Q from $\mathcal{Y}$ to $\mathcal{Z}$, which is again a Markov kernel from $\mathcal{X}$ to $\mathcal{Y} \times \mathcal{Z}$, determined as

$$(P \circledast Q)^x(B \times C) = \int_B Q^y(C)\,\mathrm{d}P^x(y), \quad B \times C \in \mathcal{B} \otimes \mathcal{C}. \tag{6.A.2}$$

The composition is associative by Tonelli's theorem. A probability measure ν on $\mathcal{X}$ can be viewed as a Markov kernel from a singleton set to $\mathcal{X}$, and we can therefore write the integration of $(P^x)_{x\in\mathcal{X}}$ w.r.t. ν as

$$(\nu \circledast P)(A \times B) = \int_A P^x(B)\,\mathrm{d}\nu(x), \quad A \times B \in \mathcal{A} \otimes \mathcal{B} \tag{6.A.3}$$

yielding a unique joint probability measure on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$.

6.A.2 The Aggregated Structural Causal Model

Conditionally on η, denote the distribution of X_I for any ordered subset $I \subseteq V$ as

$$\mu_I(A_I) = P(X_I \in A_I \mid \eta), \quad A_I \in \mathcal{A}_I. \tag{6.A.4}$$

By fixing η in (6.A.1), the joint distribution μ on $\mathcal{X}$ can be written as a recursive kernel factorization $\mu = \circledast_{v\in V} F_v(\cdot, \nu_{\varepsilon_v}, \eta_v)$ where each F_v plays the role of a Markov kernel from $\mathcal{X}_{\mathrm{pa}(v)}$ to $\mathcal{X}_v$, representing the conditional distribution $X_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}$. Here, $F_v(x_{\mathrm{pa}(v)}, \nu_{\varepsilon_v}, \eta_v)$ is used as shorthand for the push-forward measure of ν_{ε_v} obtained using the map $\varepsilon \mapsto F_v(x_{\mathrm{pa}(v)}, \varepsilon, \eta_v)$. Further, since there exists a topological ordering of V such that $\alpha < \beta$ for all $\alpha \in \mathrm{An}(I)$ and $\beta \in V \setminus \mathrm{An}(I)$, we can factorize μ as

$$\mu = \left(\circledast_{v\in \mathrm{An}(I)} F_v(\cdot, \nu_{\varepsilon_v}, \eta_v)\right) \circledast \left(\circledast_{v\in V\setminus \mathrm{An}(I)} F_v(\cdot, \nu_{\varepsilon_v}, \eta_v)\right) = \mu_{\mathrm{An}(I)} \circledast Q. \tag{6.A.5}$$

Thus, any μ_I-marginal can be obtained by computing it as transformations of $\nu_{\varepsilon_{\mathrm{An}(I)}}$ by marginalizing $\mu_{\mathrm{An}(I)}$ in (6.A.5) over $\mathrm{An}(I) \setminus I$.

The ASCM defines μ_I as a function of its graphical parents in a PD^+ $\mathcal{D}^+$ and exogenous noise variables η_I through assignments of the form

$$\mu_I(A_I) = G_I(\mu_{\mathrm{pa}_{\mathcal{D}^+}(I)}, \eta_I)(A_I), \quad A_I \in \mathcal{A}_I. \tag{6.A.6}$$

In the above, G_I specifies a composition of Markov kernels

$$G_I(\mu_{\mathrm{pa}_{\mathcal{D}^+}(I)}, \eta_I)(A_I) = \left(\left(\otimes_{J\in \mathrm{pa}_{\mathcal{D}^+}(I)} \mu_J\right) \circledast \left(\circledast_{v\in I\setminus \overline{\mathrm{pa}_{\mathcal{D}^+}(I)}} F_v(\cdot, \nu_{\varepsilon_v}, \eta_v)\right)\right)(C), \tag{6.A.7}$$

where $C = \bigtimes_{v\in I\cup \overline{\mathrm{pa}_{\mathcal{D}^+}(I)}} C_v$ is a product set with C_v being $A_v \in \mathcal{A}_v$ if $v \in I$ and equal to $\mathcal{X}_v$ otherwise. In this way, distributions appearing among the parents are generally marginalized out unless they also appear in I. It is possible for I to have more than one parent, but this happens if and only if $\mu_{\overline{\mathrm{pa}_{\mathcal{D}^+}(I)}} = \otimes_{J\in \mathrm{pa}_{\mathcal{D}^+}(I)} \mu_J$, cf. P5. If the first argument in (6.A.6) is empty, we can write $\mu_I = G_I(\eta_I) = \circledast_{v\in I} F_v(\cdot, \nu_{\varepsilon_v}, \eta_v)$ for short. Note that when $I \in \mathcal{I}$ is a singleton set $\{v\} \subseteq V$, we have

$$\mu_v(A_v) = (\mu_{\mathrm{pa}(v)} \circledast F_v(\cdot, \nu_{\varepsilon_v}, \eta_v))(\mathcal{X}_{\mathrm{pa}(v)} \times A_v) = \int F_v(x, \nu_{\varepsilon_v}, \eta_v)(A_v)\, \mathrm{d}\mu_{\mathrm{pa}(v)}(x), \tag{6.A.8}$$

for $A_v \in \mathcal{A}_v$.

6.B Establishing the ADAG

The procedure for defining the aggregated DAG (i.e., the ADAG) is given in Algorithm 1. Below, we expand on the steps in the algorithm.

The primary step of Algorithm 1 involves a construction of PD$^+$'s of each sink node of the input DAG $\mathcal{D}$. The tree implied by an ancestral lineage of a sink node $s \in \mathcal{S} := \{v \in V : \mathrm{ch}_{\mathcal{D}}(v) = \emptyset\}$ is $\mathcal{T}_s = (\mathcal{I}_s, \mathcal{E}_s)$ with vertices $\mathcal{I}_s = (I_{k,s})_{k\in\{0,\dots,L_s\}}$ and edges $\mathcal{E}_s = ((I_{k,s}, I_{k-1,s}))_{k\in\{1,\dots,L_s\}}$ where the ancestral lineages $(I_{k,s})_{k,s}$ are given by Definition 6.4.9. The resulting trees form PD$^+$'s of sinks in $\mathcal{D}$, cf. Proposition 6.4.11, and can be improved in terms of tree width by taking unconditional separation statements in $\mathcal{D}$ into account – without violating P1–P5. The following result is useful for finding the relevant separation statements.

Lemma 6.B.1. *Let $\mathcal{D} = (V, E)$ be a DAG. If $A, B \subseteq V$ then $\mathrm{An}_{\mathcal{D}}(A) \cap \mathrm{An}_{\mathcal{D}}(B) = \emptyset \iff A \perp_{\mathcal{D}} B$.*

The lemma suggests that a bottom-up search for separation statements is advantageous. Given an ancestral lineage $\mathcal{I} = (I_k)_{k\in\{0,\dots,L\}}$, we need to check whether each $I \in \mathcal{I}$ should be split into multiple nodes.

Algorithm 2: `separateSet`

Input : DAG $\mathcal{D} = (V, \mathcal{E})$, Set $I \subseteq V$.
Output : A family of sets $\mathcal{A}$ satisfying the properties in Lemma 6.B.2.

1 Construct the undirected graph $\mathcal{U} = (I, \{\{\alpha, \beta\} : \mathrm{An}_{\mathcal{D}}(\alpha) \cap \mathrm{An}_{\mathcal{D}}(\beta) \neq \emptyset, \alpha, \beta \in I\})$;
2 $\mathcal{A} \leftarrow$ connected components of $\mathcal{U}$;
3 **return** $\mathcal{A}$;

Lemma 6.B.2. *Consider a DAG $\mathcal{D} = (V, E)$ and a set of nodes $I \subseteq V$. The* `separateSet`*-algorithm constructs a family of sets $(A_j : j \in \mathcal{J})$ such that $\cup_{j\in\mathcal{J}} A_j = I$, and $A_j \perp_{\mathcal{D}} A_{j'}$ for all $j, j' \in \mathcal{J}$ with $j \neq j'$ and such that $\forall j \in \mathcal{J}$ there does not exist an $A \subsetneq A_j : A \perp_{\mathcal{D}} (A_j \setminus A)$.*

If an I-set is to be divided, new separate branches of the lineage can be formed without ambiguities since the variables would have no common ancestors in the original graph.

Corollary 6.B.3. *Let $\mathcal{T} = (\mathcal{I}, \mathcal{E})$ be a PD$^+$ of a DAG $\mathcal{D}$ with a single terminal vertex under the restriction that $A = \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I)) \cap I$ for every $I \in \mathcal{I}$ in P2. Suppose there is an $I_0 \in \mathcal{I}$ such that $\mathrm{pa}_{\mathcal{T}}(I_k) = I_{k+1}$ for all $k \in \{0, \dots, L\}$. If there is an $A \subsetneq I_0$ such that $B = I_0 \setminus A$ satisfies $\mathrm{An}_{\mathcal{D}}(A) \cap \mathrm{An}_{\mathcal{D}}(B) = \emptyset$, it holds that $\mathcal{T}$ is still a PD$^+$ in the modified graph where the nodes $(I_k)_{k\in\{0,\dots,L\}}$ are replaced*

by nodes $((I_{k,1})_{k\in\{0,\dots,L_1\}}, (I_{k,2})_{k\in\{0,\dots,L_2\}})$*, defined by* $I_{k+n,1} := I_k \cap \mathrm{An}_{\mathcal{D}}(A)$ *and* $I_{k+n,2} := I_k \cap \mathrm{An}_{\mathcal{D}}(B)$ *for* $k \in \{0,\dots,L\}$ *and the edges* (I_{k+1}, I_k) *are replaced by edges* $(I_{k+1,1}, I_{k,1})$ *and* $(I_{k+1,2}, I_{k,2})$ *for* $k \in \{0,\dots,L-1\}$*, while the edge* (I_0, I_{-1}) *is replaced by edges* $(I_{0,1}, I_{-1})$ *and* $(I_{0,2}, I_{-1})$*, whenever the sets involved are non-empty.*

This leads to the function `separateLineage` described in Algorithm 3 that produces an improved PD^+ of a sink in $\mathcal{D}$ in terms of width, cf. Lemma 6.4.12.

Algorithm 3: `separateLineage`

Input : TIBAL $\mathcal{T} = (\mathcal{I}, \mathcal{E})$ of a sink s and the DAG $\mathcal{D}$ from which it is constructed.

Output : A directed tree $(\mathcal{I}, \mathcal{E})$ being a PD^+ of $\mathcal{D}_{\mathrm{An}(s)}$ satisfying $\forall I \in \mathcal{I} : \nexists I' \subsetneq I$ such that $I' \perp_{\mathcal{D}} (I \setminus I')$.

```
L ← depth of T;
if L ≥ 1 then
    for k ∈ {1,...,L} do
        for I ∈ {H ∈ I : dist_T(H, {s}) = k} do
            (A_j)_{j∈{1,...,J}} ← separateSet(D, I);
            if J ≥ 2 then
                for j ∈ {2,...,J} do
                    T ← modify T as in Corollary 6.B.3 with I playing the
                     role of I_0 and A_j the role of A;
                end
            end
        end
    end
end
return T;
```

6.B.1 Joining the Trees

Once PD^+'s of sinks $\mathcal{S}$ in $\mathcal{D}$ have been established, the first step of the algorithm is complete. The ADAG is then defined as the join of the resulting trees $\mathcal{D}^+ := \mathrm{JoinAndReduce}_\pi((\mathcal{T}_s)_{s\in\mathcal{S}})$ The full proposed algorithm is described in Algorithm 4 for a given topological ordering π of $\mathcal{D}$.

As a first step, we set $\mathcal{D}^* = \sqcup_{s\in\mathcal{S}}\mathcal{T}_s$ denoting the disjoint union of trees that treats nodes and edges as distinct regardless of their labels. Next, we modify the vertex set to make it reduced. Consider two sets $I, I' \in \mathcal{I}$ for which $I \subseteq I'$. Because $I \subseteq I'$ implies that $\mathrm{An}_{\mathcal{D}}(I) \subseteq \mathrm{An}_{\mathcal{D}}(I')$ we should generally (i) add edges from I' to nodes among $\mathrm{ch}_{\mathcal{D}^*}(I)$, and (ii) remove nodes and edges among $\mathrm{An}_{\mathcal{D}^*}(I)$. There are, however, a few pitfalls to avoid.

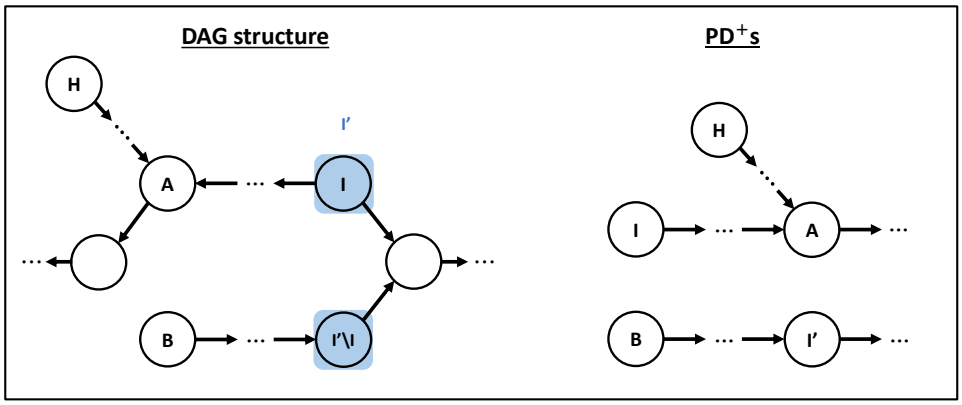

Figure 6.6: DAG structure leading to issues when pruning superfluous nodes. The nodes in the graphs represent clusters of nodes. If there is an α such that (i) $\alpha \in A$ and $\alpha \in B$ but $\alpha \notin \mathrm{ch}(I)$ in the PD^+ and/or $\alpha \notin I'$ then P3 is violated, while (ii) $\alpha \in B$ and $\alpha \in H$ would cause a violation of P5, upon replacing the arrow $I \to \mathrm{ch}(I)$ with $I' \to \mathrm{ch}(I)$.

Firstly, for given $I \in \mathcal{I}$ there may be multiple sets $I' \in \mathcal{I}$ that satisfy $I \subseteq I'$. To decide which I' should become the new parent of $\mathrm{ch}_{\mathcal{D}^*}(I)$, we rely on the following adaptation of the topological ordering.

Definition 6.B.4 (Lineage ordering)**.** *Given a DAG $\mathcal{D} = (V, E)$ and a reduced collection of sets $\mathcal{I}$ satisfying $\cup_{I \in \mathcal{I}} I = V$, we say that $I \in \mathcal{I}$ is greater than $\mathcal{I}' \in \mathcal{I}$ in the lineage ordering ρ_π with respect to a topological ordering π of $\mathcal{D}$ if (i) $I \subsetneq I'$, or if (ii) for every $\alpha \in I \setminus I'$ and some $\beta \in I' \setminus I$ then $\pi(\beta) < \pi(\alpha)$.*

The reduction operation follows the lineage ordering in the sense that if multiple I' satisfy $I \subseteq I'$, we always attach the children of I to the I' lowest in the ordering. In the case that I and I' are equal and when there does not exist an $I'' \supsetneq I$, then the above steps work as the graph union that simply merges vertices and edges with shared labels.

Secondly, we might encounter nodes $I \subseteq I'$ for which pruning according to the above rules eventually causes a cycle in $\mathcal{D}^*$ or results in a violation of P3 or P5. See Figure 6.6 for the generic DAG structure leading to the above or Figures 6.7–6.9 for concrete examples. There are more than one solution to this problem, and how such conflicts should be resolved is not obvious. We give one way of merging the graphs in Algorithm 4. See also the proof of Proposition 6.4.13 for additional details.

Algorithm 4: JoinAndReduce$_\pi$

Input : PD$^+$'s $(\mathcal{T}_s)_{s\in\mathcal{S}}$ of sinks $\mathcal{S}$ of $\mathcal{D}$ satisfying the postcondition of Algorithm 3 and a topological ordering π of the DAG $\mathcal{D}$ from which they are constructed.
Output : A PD$^+$ of $\mathcal{D}$.

$\mathcal{D}^* \leftarrow \sqcup_{s\in\mathcal{S}}\mathcal{T}_s = (\mathcal{I},\mathcal{E})$;
$\rho_\pi \leftarrow$ lineage ordering of $\mathcal{I}$ w.r.t. π;
Arrange $\mathcal{I}$ in descending order according to ρ_π;
$i \leftarrow 1$;
while $i < |\mathcal{I}|$ **do**
  $I \leftarrow i$'th element of $\mathcal{I}$;
  $I' \leftarrow$ element in $\mathcal{I}$ of lowest order in ρ_π such that $I \subseteq I'$ if it exists; $I' \leftarrow \emptyset$ otherwise;
  $i \leftarrow i+1$;
  if $I' = \emptyset$ **then** $i \leftarrow i+1$ **end**;
  else
    if $\overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)} \subseteq I' \cap \overline{\mathrm{ch}_{\mathcal{D}^*}(I)}$ *and* $\left(\overline{\mathrm{An}_{\mathcal{D}^*}(\mathrm{de}_{\mathcal{D}^*}(I))} \setminus \overline{\mathrm{An}_{\mathcal{D}^*}(I)}\right) \cap \overline{\mathrm{An}_{\mathcal{D}^*}(I')} = \emptyset$ **then**
      for $J \in \mathrm{ch}_{\mathcal{D}^*}(I)$ **do** add edge (I', J) to $\mathcal{E}$ **end**;
      Remove I from $\mathcal{I}$ and all edges involving I from $\mathcal{E}$;
    end
    else
      if $\overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)} \nsubseteq I' \cap \overline{\mathrm{ch}_{\mathcal{D}^*}(I)}$ **then**
        $I^* \leftarrow I \cup (\overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)})$;
        $\tilde{I} \leftarrow$ element in $\mathcal{I}$ of lowest order in ρ_π such that $I^* \subseteq \tilde{I}$ if it exists; $\tilde{I} \leftarrow \emptyset$ otherwise;
        if $\tilde{I} = \emptyset$ **then** $\mathcal{D}^* \leftarrow$ `addNodes`$(\mathcal{D}^*, I^*, \mathcal{D})$ **else** $I^* \leftarrow \tilde{I}$ **end**;
        for $J \in \mathrm{ch}_{\mathcal{D}^*}(I)$ **do** add edge (I^*, J) to $\mathcal{E}$ **end**;
        for $J \in \mathrm{de}_{\mathcal{D}^*}(I)$ **do** $J \leftarrow J \cup (\cap_{\{A: A \text{ on path from } I^* \text{ to } J \text{ in } \mathcal{D}^*\}} A)$ in $\mathcal{I}$ **end**;
        Remove I from $\mathcal{I}$ and all edges involving I from $\mathcal{E}$;
      end
      else
        $I^* \leftarrow I$;
      end
      if $\left(\overline{\mathrm{An}_{\mathcal{D}^*}(\mathrm{de}_{\mathcal{D}^*}(I))} \setminus \overline{\mathrm{An}_{\mathcal{D}^*}(I)}\right) \cap \overline{\mathrm{An}_{\mathcal{D}^*}(I')} \neq \emptyset$ **then**
        $I^{**} \leftarrow I^*$;
        for $J \in \mathrm{de}_{\mathcal{D}^*}(I^*)$ **do**
          $A \leftarrow$ parent of J in $\mathcal{D}^*$ for which there exists a path from I^* to J;
          for $B \in \{H \in \mathrm{pa}_{\mathcal{D}^*}(J) : \nexists$ *path from* I^* *to* J *through* $H, \overline{\mathrm{An}_{\mathcal{D}^*}(H)} \cap \overline{\mathrm{An}_{\mathcal{D}^*}(I^*)} \neq \emptyset\}$ **do**
            $C \leftarrow \mathrm{pa}_{\mathcal{D}}(J \setminus (A \cup B)) \setminus (\overline{\mathrm{pa}_{\mathcal{D}^*}(J)} \setminus B)$;
            if $C \neq \emptyset$ **then**
              Set $A \leftarrow A \cup C$ in $\mathcal{I}$;
              $I^{**} \leftarrow I^{**} \cup C$;
            end
            Delete the edge (B, J) in $\mathcal{E}$;
          end
        end
        $\tilde{I} \leftarrow$ element in $\mathcal{I}$ of lowest order in ρ_π such that $I^{**} \subseteq \tilde{I}$ if it exists; $\tilde{I} \leftarrow \emptyset$ otherwise;
        if $\tilde{I} = \emptyset$ **then** $\mathcal{D}^* \leftarrow$ `addNodes`$(\mathcal{D}^*, I^{**}, \mathcal{D})$ **else** $I^{**} \leftarrow \tilde{I}$ **end**;
        for $J \in \mathrm{ch}_{\mathcal{D}^*}(I^*)$ **do** add edge (I^{**}, J) to $\mathcal{E}$ **end**;
        for $J \in \mathrm{de}_{\mathcal{D}^*}(I^*)$ **do** set $J := J \cup (\cap_{\{A: A \text{ on path from } I^{**} \text{ to } J \text{ in } \mathcal{D}^*\}} A)$ in $\mathcal{I}$ **end**;
        Remove I^* from $\mathcal{I}$ and all edges involving I^* from $\mathcal{E}$;
      end
      $\rho_\pi \leftarrow$ lineage ordering of $\mathcal{I}$ w.r.t. π;
      Arrange $\mathcal{I}$ in descending order according to ρ_π;
      $i \leftarrow 1$;
    end
  end
end

Function `addNodes`$(\mathcal{D}^*, I^*, \mathcal{D})$**:**
  $\mathcal{T} \leftarrow$ TIBAL of I^* (treating I^* as a sink node in $\mathcal{D}$);
  $\mathcal{T} \leftarrow$ `separateLineage`$(\mathcal{T})$;
  $\mathcal{D}^* \leftarrow \mathcal{D}^* \sqcup \mathcal{T}$;
return $\mathcal{D}^*$

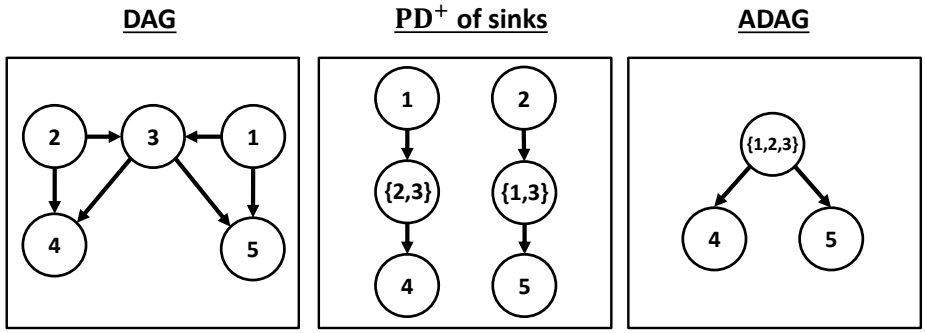

Figure 6.7: In the middle graph there are redundancies since $\{1\} \subseteq \{1,3\}$ and $\{2\} \subseteq \{2,3\}$. Removing the nodes $\{1\}$ and $\{2\}$, and replacing the edge $(\{1\},\{2,3\})$ by $(\{1,3\},\{2,3\})$ and the edge $(\{2\},\{1,3\})$ by $(\{2,3\},\{1,3\})$ introduces a cycle (and either replacement causes a violation of P3).

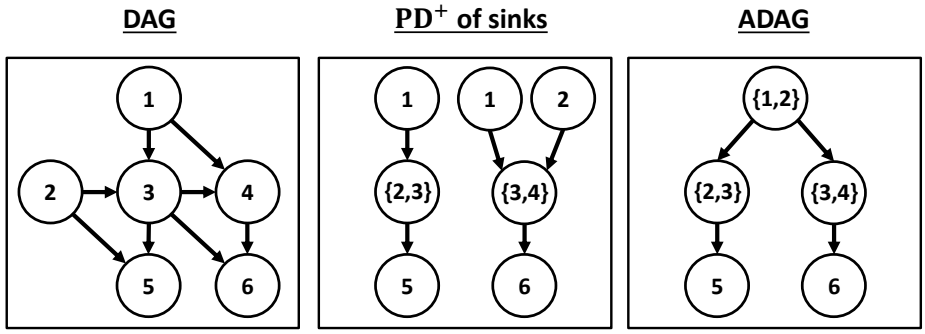

Figure 6.8: In the middle graph there is a redundancy since $\{2\} \subseteq \{2,3\}$. Joining the two $\{1\}$ nodes, removing the $\{2\}$ node and replacing the edge $(\{2\},\{3,4\})$ by $(\{2,3\},\{3,4\})$ causes a violation of P5.

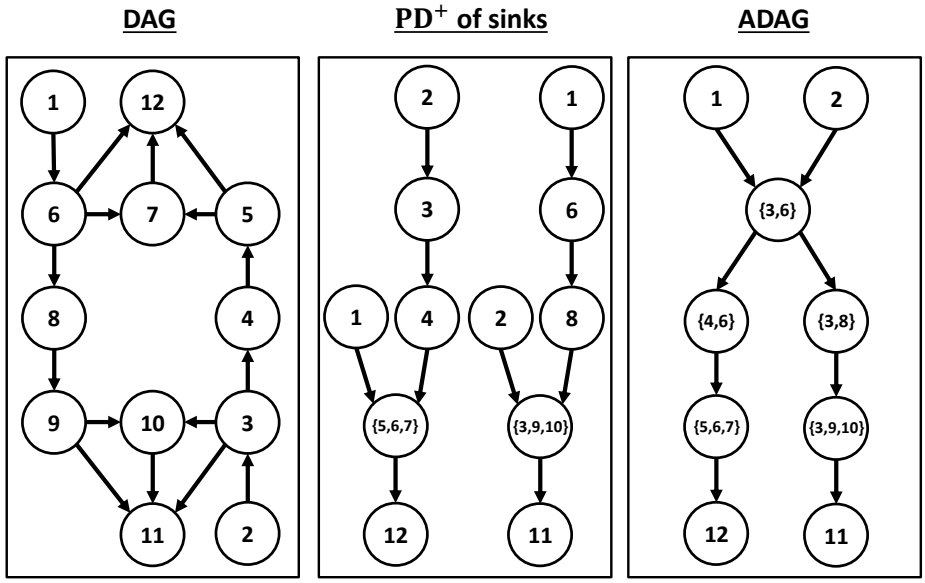

Figure 6.9: DAG for which there are multiple "problem sets" requiring special treatment in Algorithm 4.

6.C Proofs

Lemma 6.4.2. *If $\mathcal{D}$ is a DAG, then $\mathcal{D}$ is a PD of itself.*

Proof. Immediate from the definition. □

Proposition 6.4.3. *If $\mathbb{P}$ factorizes according to $\mathcal{D} = (V, E)$, then for a PD $\mathcal{D}^* = (\mathcal{I}, \mathcal{E})$ of $\mathcal{D}$ it holds that*

$$p(x) = \frac{\prod_{I\in\mathcal{I}} p(x_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}|x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}})}{\prod_{v\in V} p(x_v|x_{\mathrm{pa}_{\mathcal{D}}(v)})^{|C(v)|-1}} = \frac{\prod_{I\in\mathcal{I}} p(x_{I\setminus\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}|x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}})}{\prod_{v\in V} p(x_v|x_{C_v^{\mathrm{pa}}})^{|C(v)|-1}}, \tag{6.4.1}$$

for all $x \in \mathcal{X}$, where $C_v = \{I \in \mathcal{I} : v \in I, v \notin \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}\}$ and $C_v^{\mathrm{pa}} = \{\cap_{I\in C_v} \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}\} \cup \{\cap_{I\in C_v}\{\alpha \in I : \pi(\alpha) < \pi(v)\}\}$.

Proof. For $I \in \mathcal{I}$ let $B_I := \{v \in I : v \notin \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}\}$ and $A_I = I \setminus B_I$. For all $v \in \overline{\mathrm{pa}_{\mathcal{D}^*}(I)} \setminus (A_I \cup \mathrm{Pa}_{\mathcal{D}}(B_I))$ it holds that $v \notin \mathrm{de}_{\mathcal{D}}(B_I)$ any $I \in \mathcal{I}$. Otherwise, there exists a $\beta \in B_I$ such that $\beta \in \mathrm{an}_{\mathcal{D}}(v)$, which, due to P1–P2, implies the existence of a node among the ancestors of I in $\mathcal{D}^*$ also containing β, but then $\beta \in \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}$ due to P3, contradicting that $\beta \in B_I$.

The above implies that for each $I \in \mathcal{I}$ we can write $p(x_I|x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}) = p(x_{B_I}|x_{\mathrm{pa}_{\mathcal{D}}(B_I)})$ whenever B_I is non-empty. Due to P1 and P2, $(B_I : I \in \mathcal{I})$ covers only and all nodes of V, although some nodes $v \in V$ may appear in multiple B_I-sets. We therefore have that

$$\begin{aligned}
\prod_{I\in\mathcal{I}} p\left(x_I|x_{\overline{\mathrm{pa}_{\mathcal{D}^*}(I)}}\right) &= \prod_{B_I:I\in\mathcal{I}} p\left(x_{B_I}|x_{\mathrm{pa}_{\mathcal{D}}(B_I)}\right) \\
&= \prod_{B_I:I\in\mathcal{I}} \prod_{v\in B_I} p\left(x_v|x_{\mathrm{pa}_{\mathcal{D}}(B_I)\cup\{\beta\in B_I:\pi(\beta)<\pi(v)\}}\right) \\
&= \prod_{v\in V} p\left(x_v|x_{\mathrm{pa}_{\mathcal{D}}(v)}\right)^{|\{B_I:I\in\mathcal{I},v\in B_I\}|} \\
&= \prod_{v\in V} p\left(x_v|x_{\mathrm{pa}_{\mathcal{D}}(v)}\right)^{|C_v|} \\
&= p(x) \prod_{v\in V} p\left(x_v|x_{\mathrm{pa}_{\mathcal{D}}(v)}\right)^{|C_v|-1} \\
&= p(x) \prod_{v\in V} p\left(x_v|x_{C_v^{\mathrm{pa}}}\right)^{|C_v|-1},
\end{aligned}$$

where the final equality follows from the fact that C_v^{pa} always contains all of v's parents but none of its descendants. □

Lemma 6.4.7. *Under Assumption 6.4.6, there exists a surjective, order-preserving map* $w : \mathbb{I}^{\text{Ind}} \to \mathbb{I}^{\text{Agg}}$ *such that* $\mathbb{P}^{\text{do}(i);\text{Ind}} = \mathbb{P}^{\text{do}(w(i));\text{Agg}}$, $\forall i \in \mathbb{I}^{\text{Ind}}$.

Proof. Choose an intervention $j \in \mathbb{I}^{\text{Agg}}$. This intervention affects the distribution of a subset of variables $I \subseteq \overline{\mathcal{I}} = V$. Due to Assumption 6.4.6 all overlapping marginal distributions implied by the ASCM $\mathcal{M}^{\text{Agg};\text{do}(j)}$ are equal. Thus, the intervention $\text{do}(j)$ induces the interventional distribution

$$\mathbb{P}^{\text{Agg},\text{do}(j)}(x) = \widetilde{p}\left(x_I | x_{\widetilde{\text{pa}}_{\mathcal{D}}(I)}\right) \prod_{v \in V \setminus I} p\left(x_v | x_{\text{pa}_{\mathcal{D}}(v)}\right), \quad x \in \mathcal{X},$$

through (6.4.1) using the conditionals defined by (6.4.5). The same interventional distribution can be produced by the stochastic intervention $\text{do}(X_I := \widetilde{p}(\cdot | x_{\widetilde{\text{pa}}(I)}))$ in the SCM. Let this intervention have index $i \in \mathbb{I}^{\text{Ind}}$. The mapping w can then be defined as an index-mapping such that $i \in w^{-1}(\{j\})$ and such that the partial ordering $\leq_X$ on $\mathbb{I}^{\text{Ind}}$ is preserved. □

Corollary 6.4.8. *Under Assumption 6.4.6, the ASCM is causally consistent with the SCM in the sense that the diagram in Figure 6.3 commutes.*

Proof. The argument is analogous to the proof of Theorem 6 in Rubenstein et al. (2017). Let $i, j \in \mathbb{I}^{\text{Ind}}$ be interventions such that $i \leq_X j$. Commutativity of the left square of the diagram follows from Lemma 6.4.7. Moreover, we have that $\mathbb{P}^{\text{Ind},\text{do}(i)} = \mathbb{P}^{\text{Agg},\text{do}(w(i))}$ and $\mathbb{P}^{\text{Ind},\text{do}(j)} = \mathbb{P}^{\text{Agg},\text{do}(w(j))}$. Existence of the arrow $\mathbb{P}^{\text{Agg},\text{do}(w(i))} \xrightarrow{\text{do}(w(j))} \mathbb{P}^{\text{Agg},\text{do}(w(j))}$ follows from w being order-preserving. □

Proposition 6.4.11. *If* $\mathcal{D} = (V, E)$ *has exactly one terminal vertex and* $\mathcal{T} = (\mathcal{I}, \mathcal{E})$ *is the tree implied by the ancestral lineage, then* $\mathcal{T}$ *is a PD*$^+$ *of* $\mathcal{D}$.

Proof. Since $\mathcal{D}$ only has one terminal vertex, the ancestral lineage is given by the sets

$$I_k = \left\{ \text{Pa}_{\mathcal{D}}(I_{k-1} \setminus A) \cup A : A = \text{An}_{\mathcal{D}}(\text{Pa}_{\mathcal{D}}(I_{k-1})) \cap I_{k-1} \right\}, \tag{6.C.1}$$

for $k \in \{1, \ldots, L\}$ with $I_0 = \{s\}$ being the sink node. Thus, $\mathcal{T}$ is the path graph $I_0 \leftarrow \cdots \leftarrow I_L$.

P1, P2 and P5 follow directly from the construction of $\mathcal{T}$ from (6.C.1).

For P3, let J be a node on the unique path from I_{k+n} to I_k, $n \geq 1$, in $\mathcal{T}$. It is enough to establish that if $\alpha \in I_k \cap I_{k+n}$ then $\alpha \in J$. Let $A_k := \text{An}_{\mathcal{D}}(\text{Pa}_{\mathcal{D}}(I_k)) \cap I_k$ for $k \geq 1$. If $n = 1$ there is nothing to show. If $n = 2$, suppose for contradiction that $\alpha \notin A_{k+1} \subseteq I_{k+1}$. Then, since $\alpha \in I_{k+2}$, there exists $\beta \in I_{k+1}$ such that $\alpha \in \text{pa}_{\mathcal{D}}(\beta)$. Thus, either $\beta \in A_k$ or there exists a $\gamma \in I_k \setminus A_k$ such that $\beta \in \text{pa}_{\mathcal{D}}(\gamma)$. Since $\alpha \in I_k$

by assumption, both cases yield a contradiction. For $n > 2$, successive application of this rule shows that $\alpha \in J$ for any J on the path from I_{k+n} to I_k.

If $|V| \leq 2$ then P4 is immediate, so suppose $|V| > 2$. We will start by arguing that the set A in (6.C.1) never makes up an entire set in the lineage, that is, we cannot have $I_k = \mathrm{An}(\mathrm{Pa}(I_k)) \cap I_k$ for any $k \in \{1, \dots, L\}$. Let k be given. If $I_k = \mathrm{An}(\mathrm{Pa}(I_k)) \cap I_k$, or equivalently $I_k \subseteq \mathrm{An}(\mathrm{Pa}(I_k))$, then for every node $\alpha \in I_k$ there is a directed path $\alpha \to \cdots \to \beta$ in $\mathcal{D}$ where β is another node in I_k. Suppose $|I_k| = n \in \mathbb{N}_+$ and let $\{1, \dots, n\}$ be the labels of I_k. Note that n cannot be one, otherwise $\mathrm{An}(\mathrm{Pa}(I_k))$ would be empty, so suppose $n \geq 2$. The first node, 1, cannot have a path leading back to itself, otherwise there would be a cycle. Without loss of generality, let there be a directed path $1 \to \cdots \to 2$. Likewise, 2 cannot have a path leading back to 1 or to itself. If $n = 2$ this yields a contradiction, if not let there be a directed path $2 \to \cdots \to 3$. Continuing this process up until the n'th variable, we see that there must be a path $n \to \cdots \to i$ for some $i \in \{1, \dots, n\}$, introducing a cycle. Thus, we cannot have $I_k = \mathrm{An}(\mathrm{Pa}(I_k)) \cap I_k$ for any k meaning that $A_k := \mathrm{An}(\mathrm{Pa}(I_k)) \cap I_k$ is always a proper (possibly empty) subset of I_k.

Consequently, at least one element from I_{k-1} is always removed by the parent condition when constructing I_k (and at least one element, namely a parent, is added). Further, due to P3, once a variable is removed, it will not reappear. Therefore, we have for any $k \in \{1, \dots, L\}$ that

(i) at least one element in I_k is not in I_{k+n} for any $n \in \{1, \dots, L-k\}$, namely a child among the nodes in I_{k+1};

(ii) at least one element in I_k is not in I_{k-n} for any $n \in \{1, \dots, k\}$, namely a parent among the nodes in I_{k-1}.

It follows from (i) that we cannot have $I_{k-n} \subseteq I_k$ for any $n \in \{1, \dots, k\}$ and from (ii) that we cannot have $I_{k+n} \subseteq I_k$ for any $n \in \{1, \dots, L-k\}$. P4 therefore holds, and we conclude that the tree implied by the ancestral lineage is a PD^+ of $\mathcal{D}$. □

Lemma 6.4.12. *If $\mathcal{D}$ has exactly one terminal vertex and $\mathcal{T}$ is the tree implied by the ancestral lineage, then* `separateLineage`$(\mathcal{T}) =: \tilde{\mathcal{T}} = (\mathcal{I}, \mathcal{E})$ *is a PD^+ of $\mathcal{D}$ satisfying $\forall I \in \mathcal{I} : \nexists I' \subsetneq I$ such that $I' \perp_{\mathcal{D}} (I \setminus I')$, making $\tilde{\mathcal{T}}$ the PD^+ of smallest possible width under the restriction that the sink of $\mathcal{D}$ is also a sink in $\tilde{\mathcal{T}}$ and that P2 reads: $A \cup \mathrm{Pa}_{\mathcal{D}}(B) = \overline{\mathrm{pa}_{\mathcal{D}^*}(I)}$ where $A \subseteq \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I)) \cap I$ and $B = I \setminus A$ for every $I \in \mathcal{I}$.*

Proof. We start by showing that $\tilde{\mathcal{T}} :=$ `separateLineage`$(\mathcal{T}) = (\mathcal{I}, \mathcal{E})$ is a PD^+ of $\mathcal{D}$ with the additional property that there does not exists an $I' \subsetneq I$ such that $I' \perp_{\mathcal{D}} (I \setminus I')$. We refer to the latter condition as $(\star)$. The proof is by induction on

the depth of the ancestral lineage. Note that since $\mathcal{D}$ only has one terminal vertex, the ancestral lineage is given by the sets

$$I_k = \{\mathrm{Pa}_{\mathcal{D}}(I_{k-1} \setminus A) \cup A : A = \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I_{k-1})) \cap I_{k-1}\}, \tag{6.C.2}$$

for $k \in \{1, \dots, L\}$ with $I_0 = \{s\}$ being the sink node and L being the depth.

For the base case ($L = 0$), the algorithm returns $\mathcal{D}$ itself; a graph containing only a single node and no edges. In this case, the algorithm's postcondition is trivially satisfied.

For the induction case ($L \geq 1$), suppose that after n iterations of the outer for-loop, then $\mathcal{T}$ is a PD^+ of $\mathcal{D}_{\mathrm{An}(s)}$ satisfying $(\star)$ for all $I \in \{H \in \mathcal{I} : \mathrm{dist}_{\mathcal{T}}(H, I_0) \leq n\}$.

In the second loop a set $I \in \{H \in \mathcal{I} : \mathrm{dist}_{\mathcal{T}}(H, I_0) = n + 1\}$ is used as the basis for all computations. Due to the loop-invariant, it is not affected by any of the previous $w - 1$ iterations, and will not affect subsequent iterations of the second loop. That is, all computations in the second loop could, in principle, be performed in parallel. In the w'th iteration, I is split into disjoint sets $(A_j)_{j \in \mathcal{J}}$ through the function `separateSet`. These sets have the properties described by Lemma 6.B.2. For each $j \in \mathcal{J}$ the set splitting procedure opens for a branching of the lineage as described by Corollary 6.B.3. If $|\mathcal{J}| = 1$, no branching is performed; I is left unaltered for every $k \in \{n+1, \dots, L\}$. If $|\mathcal{J}| \geq 2$, new separate branches are created in the sense of Corollary 6.B.3. By the induction hypothesis, Lemma 6.B.2 and successive application of Corollary 6.B.3, the loop-invariant holds at the end of the $n + 1$'th iteration of the outer loop. Upon termination at loop-exit ($n + 1 = L$), the invariant implies the postcondition.

Since $\tilde{\mathcal{T}}$ has the property $(\star)$, the decomposition following the sets specified by (6.C.2) cannot be improved further in terms of width under P5. What is left is then to argue that these sets cannot be chosen differently under the restricted version of P2 in the lemma, P2^r say, which states that for every I-node the parent set in the tree should be a clustering of the nodes $A \cup \mathrm{Pa}_{\mathcal{D}}(B)$ where $A \subseteq \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I)) \cap I$ and $B = I \setminus A$. We note that the ancestral lineage selects $A = \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I)) \cap I$, that is, it always retains every node that reappears upstream.

The statement is trivial if $|V| \leq 2$ or if $\mathrm{pa}_{\mathcal{D}}(s) = V \setminus \{s\}$, because to satisfy P2^r we always have $I_0 = \{s\}$ and $I_1 = \mathrm{pa}_{\mathcal{D}}(s)$, so suppose that $|V| > 2$ and $\mathrm{pa}_{\mathcal{D}}(s) \subsetneq V \setminus \{s\}$. Let $A_k := \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I_{k-1})) \cap I_{k-1}$ where I_k is defined as in (6.C.2). If $A_k = \emptyset$ for all levels $k \geq 1$, then (6.C.2) reduces to $I_k = \mathrm{Pa}_{\mathcal{D}}(I_{k-1})$ and we are done. If A_k is not empty at every k, consider the first time, $k \geq 2$, where A_k is non-empty and where we wish to retain only some or no variables $A_k^* \subsetneq A_k$, redefining I_k to be

$$I_k^* = A_k^* \cup \mathrm{Pa}_{\mathcal{D}}(B_k \cup (A_k \setminus A_k^*)),$$

where $B_k := I_{k-1} \setminus A_k$. For any excluded $\alpha \in A_k \setminus A_k^*$ there exists a $\beta \in I_k^*$ such that $\alpha \in \mathrm{an}_{\mathcal{D}}(\beta)$, because otherwise α would not be in the set defining A_k. Due to

P2, this implies the existence of an I^*_{k+n}, $n \geq 1$, containing α. Since $\alpha \in A_k$ implies that $\alpha \in I_{k-1}$, P3 states that α has to be included in I^*_k. Thus, no elements among A_k may be excluded from I_k, concluding the proof. □

Proposition 6.4.13. *Algorithm 1 produces a PD$^+$ of an input DAG $\mathcal{D}$.*

Proof. By Lemma 6.4.12, the graphs plugged into JoinAndReduce$_\pi$ satisfy the precondition. To satisfy the postcondition, we will show that when the while loop in JoinAndReduce$_\pi$ finishes, D^* will be a PD$^+$ of $\mathcal{D}$. If $|\mathcal{I}| = 1$ this is trivial, so suppose $|\mathcal{I}| \geq 2$. We propose the following loop invariant; at the end of any iteration of the while loop then:

(a) $\mathcal{D}^*$ is a DAG;

(b) $\mathcal{D}^*$ satisfies P1, P2, P3 and P5;

(c) The i first elements in $\mathcal{I}$ are mutually reduced.

Before the loop iterates the first time $i = 1$ and the invariant trivially holds. Suppose the invariant holds at the beginning of some iteration $i > 1$. If $I' = \emptyset$, nothing happens save an increase of i by one. Since $\mathcal{I}$ was sorted in descending order according to ρ_π, there does not exist an $I_l \in \mathcal{I}$ such that $I_l \subseteq I$. Since $I' = \emptyset$ there does not exist an $I_u \in \mathcal{I}$ such that $I_u \supseteq I$. Therefore, the invariant holds. Suppose $I' \neq \emptyset$. We have 2 cases.

Case I: $\overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)} \subseteq I' \cap \overline{\mathrm{ch}_{\mathcal{D}^*}(I)}$ and $\left(\overline{\mathrm{An}_{\mathcal{D}^*}(\mathrm{de}_{\mathcal{D}^*}(I))} \setminus \overline{\mathrm{An}_{\mathcal{D}^*}(I)}\right) \cap \overline{\mathrm{An}_{\mathcal{D}^*}(I')} = \emptyset$. Since $I \subseteq I'$, we get from the invariant that the ancestors of I' contain all the ancestors of I. Removing I and replacing the edge (I, J) with (I', J) for every $J \in \mathrm{ch}_{\mathcal{D}^*}(I)$ thus always preserves P1, P2 and P4. Since P3 holds at the beginning of the iteration, P3 is intact following the replacement of the edge(s) if and only if for every $A \in \mathrm{de}_{\mathcal{D}^*}(I)$ and $B \in \mathrm{An}_{\mathcal{D}^*}(I')$ then $A \cap B \subseteq J$ for every node J on the path from B to A in the modified graph; or equivalently, if and only if $\overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)} \subseteq I' \cap \overline{\mathrm{ch}_{\mathcal{D}^*}(I)}$. Further, since P5 holds at the beginning of the iteration, P5 is intact following the replacement of the edge(s) if and only if for every $A \in \mathrm{An}_{\mathcal{D}^*}(\mathrm{de}_{\mathcal{D}^*}(I)) \setminus \mathrm{An}_{\mathcal{D}^*}(I)$ and $B \in \mathrm{An}_{\mathcal{D}^*}(I')$ then $A \perp_{\mathcal{D}} B$; or equivalently, if and only if $\overline{\mathrm{An}_{\mathcal{D}^*}(\mathrm{de}_{\mathcal{D}^*}(I))} \setminus \overline{\mathrm{An}_{\mathcal{D}^*}(I)} \cap \overline{\mathrm{An}_{\mathcal{D}^*}(I')} = \emptyset$, cf. Lemma 6.B.1. Thus, in case I, we have ruled out the possibility that the edge replacement causes a violation of P3 and P5. We conclude that (a) and (b) hold. Since only I is deleted from $\mathcal{I}$, i is left unchanged and takes the same value as at the beginning of the iteration, and we get from the invariant that (c) also holds.

Case II: $\overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)} \not\subseteq I' \cap \overline{\mathrm{ch}_{\mathcal{D}^*}(I)}$ or $\left(\overline{\mathrm{An}_{\mathcal{D}^*}(\mathrm{de}_{\mathcal{D}^*}(I))} \setminus \overline{\mathrm{An}_{\mathcal{D}^*}(I)}\right) \cap \overline{\mathrm{An}_{\mathcal{D}^*}(I')} \neq \emptyset$. In the first case, pruning as above would yield a violation of P3. We

introduce the node (if it does not already exist, lines 17–19) $I^* \supseteq \overline{\mathrm{An}_{\mathcal{D}^*}(I')} \cap \overline{\mathrm{de}_{\mathcal{D}^*}(I)}$ that becomes the new parent of the children of I (instead of I', line 20), and update all descendants of I so that P3 holds (line 21). The second case deals with sets that would cause a violation of P5. Note that such a violation may occur whether or not pruning would cause a violation of P3. To rectify the situation, we look for nodes J among the descendants of I (or I^* in the modified graph if P3 would have been violated) for which there exists distinct paths into J with a non-empty intersection (lines 29–39). If we come across such a set B, say, we remove the edge (B, J) to maintain P5 (line 37). If B contains parents of J that are not among J's remaining parents, this would cause a violation of P2. We therefore add missing parents to the parent of J for which there is a path from I (or I^*) to J and also to I (or I^*, lines 30-36). This potentially causes a violation of P3, so all descendants of I (or I^*) have to be updated so that P3 holds (line 44). After these operations (a) and (b) hold. Because we have modified sets of $\mathcal{I}$, we have to update the lineage ordering, rearrange $\mathcal{I}$, and start over with $i = 1$ (lines 46–48). Thus, the invariant also holds at the end of the iteration in case II.

Observe that if there are no "problem sets", that is, if we are always in case I when $I \subseteq I'$, then, after each iteration of the while loop, $|\mathcal{I}| - i$ decreases by one. Thus, if the loop continues to iterate, eventually $i = |\mathcal{I}|$, whereupon the loop terminates. If there is an I and an I' such that we enter case II, i resets and $|\mathcal{I}|$ may increase. However, because all sets modified by operations in case II always grow in size, then eventually, if the loop continues to iterate, we will not find ourselves in case II again. Upon termination of the while loop, we have from the invariant that the final contents of D^* is a DAG satisfying P1–P5, and we are done. □

Proposition 6.4.14. *If $\mathcal{D} = (V, E)$ is a DAG whose underlying undirected graph is a forest, then the ADAG $\mathcal{D}^+$ is equal to $\mathcal{D}$ in the sense that $\mathcal{D}^+ = (V, E)$.*

Proof. Let the skeleton $\mathrm{ske}(\mathcal{G})$ of a graph $\mathcal{G}$ be the undirected graph where two nodes are connected in $\mathrm{ske}(\mathcal{G})$ if and only if they are adjacent in $\mathcal{G}$, that is, $\mathrm{ske}(\mathcal{G})$ is the underlying undirected graph of $\mathcal{G}$. If $\mathrm{ske}(\mathcal{D})$ is a forest, the ADAG is a disjoint union of the ADAG's of the connected components of $\mathcal{D}$. It suffices to consider the case where $\mathrm{ske}(\mathcal{D})$ is a tree

Let $\mathcal{S} = \{v \in V : \mathrm{ch}_{\mathcal{D}}(v) = \emptyset\}$ be the set of sink nodes and fix an $s \in \mathcal{S}$. We start by showing that the ADAG of $\mathcal{D}_{\mathrm{An}(s)}$, namely $\mathcal{D}^{+_s} = \texttt{separateLineage}(\mathcal{T}_s)$ where $\mathcal{T}_s$ is the TIBAL of s, is equal to $\mathcal{D}_{\mathrm{An}(s)}$ itself. Notice that since $\mathrm{ske}(\mathcal{D})$ is a tree then $\mathrm{ske}(\mathcal{D}_{\mathrm{An}(s)})$ is also a tree. It follows that the ancestral lineage simplifies to the union of parents, that is, $I_0 = \{s\}$ and

$$I_k = \mathrm{Pa}_{\mathcal{D}}(I_{k-1}), \quad k \in \{1, \ldots, L_s\}.$$

Otherwise there is an I_k, $k \geq 2$, with $|I_k| \geq 2$ and $A_k := \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I_k)) \cap I_k \neq \emptyset$ (since $A_k \subsetneq I_k$ cf. the proof of Proposition 6.4.11), such that an $\alpha \in A_k$ and a $\beta \in I_k \setminus A_k$ share at least one common descendant, contradicting that $\mathrm{ske}(\mathcal{D}_{\mathrm{An}(s)})$ is a tree. Moreover, for every I_k with $|I_k| > 1$ then $\alpha \perp_{\mathcal{D}} \beta$ for every pair $\alpha, \beta \in I_k, \alpha \neq \beta$, because otherwise, if $\alpha \sim \beta$, there would be a cycle in $\mathrm{ske}(\mathcal{D}_{\mathrm{An}(s)})$ as α and β always share at least one common descendant. It follows from Lemma 6.4.12 that $\mathcal{D}^{+_s}$ is equal to $\mathcal{D}_{\mathrm{An}(s)}$.

If $|\mathcal{S}| = 1$ we are done since Algorithm 4 would return its input. Suppose $|\mathcal{S}| > 1$. Since the width of $\mathcal{D}^{+_s}$ for any $s \in \mathcal{S}$ is one, there are no "cluster nodes", and Algorithm 4 works as the graph union, merging vertices and edges with shared labels. Thus,

$$\mathcal{D} = \cup_{s \in \mathcal{S}} \mathcal{D}_{\mathrm{An}(s)} = \cup_{s \in \mathcal{S}} \mathcal{D}^{+_s} = \mathcal{D}^{+},$$

concluding the proof. □

Proposition 6.4.15. *If $\mathcal{D}^{+} = (\mathcal{I}, \mathcal{E})$ is the ADAG determined from $\mathcal{D}$ and $A, B, C \subseteq \mathcal{I}$ are three mutually reduced families of sets, then*

$$A \perp_{\mathcal{D}^{+}} B \mid C \Longrightarrow (\overline{A} \setminus (\overline{C} \cup D)) \perp_{\mathcal{D}} (\overline{B} \setminus (\overline{C} \cup D)) \mid (\overline{C} \cup D),$$

where $D = \mathrm{IS}\left(\overline{\mathrm{An}_{\mathcal{D}^{+}}(A)}, \overline{\mathrm{An}_{\mathcal{D}^{+}}(B)}, \overline{\mathrm{An}_{\mathcal{D}^{+}}(C)}\right) \setminus \overline{\mathrm{an}_{\mathcal{D}^{+}}(\{J \in C : J \in \mathrm{An}_{\mathcal{D}^{+}}(\{A, B\})\})}$ *with* $\mathrm{IS}(A, B, C) = (A \cap B) \cup (A \cap C) \cup (B \cap C)$.

Proof. Let $\mathcal{D}^{m}$ denote the moral graph of a DAG $\mathcal{D}$, that is, the undirected graph with the same vertex set as $\mathcal{D}$ and nodes α and β that are adjacent in $\mathcal{D}^{m}$ if and only if $\alpha \to \beta$ or $\beta \to \alpha$ or α and β have a common child (Lauritzen, 1996). Similarly for an ADAG $\mathcal{D}^{+} = (\mathcal{I}, \mathcal{E})$, or any subgraph of it, we define the ADAG-moral graph $(\mathcal{D}^{+})^{m^{+}}$ as the undirected graph with vertex set $\cup_{I \in \mathcal{I}} I$ and nodes α and β that are adjacent in $(\mathcal{D}^{+})^{m^{+}}$ if and only if either

1. $\alpha, \beta \in I$ for any $I \in \mathcal{I}$ or,
2. $\alpha \in I, \beta \in I'$ s.t. $\alpha \neq \beta$ and there is an edge $I \to I'$, or an edge $I' \to I$, or I and I' have a common child I''.

We start by showing that separation in $\mathcal{D}^{+}$ implies separation in the ADAG-moral graph. Notice that the ADAG-moral graph can be constructed in two steps. The first step is the usual moralization operation; all parents are married and directions deleted. Next, the graph is "unclustered" so that two nodes are adjacent $\alpha \sim \beta$ if they appear in the same cluster node, that is $\alpha, \beta \in I$, or if they appear in different but adjacent cluster nodes, that is, $\alpha \in I, \beta \in I'$ with $I \sim I'$. Thus, if $\mathcal{I}$ is only composed of disjoint subsets of V, we clearly have

$$A \perp_{\mathcal{D}^{+}} B \mid C \Longrightarrow \overline{A} \perp_{\left(\mathcal{D}^{+}_{\mathrm{An}(\{A,B,C\})}\right)^{m^{+}}} \overline{B} \mid \overline{C},$$

and there is no need to condition on additional variables. However, when some members of $\mathcal{I}$ have a non-empty intersection, the "unclustering" step in the construction above may result in a walk w from $\alpha \in \overline{A} \setminus (\overline{B} \cup \overline{C})$ to $\beta \in \overline{B} \setminus (\overline{A} \cup \overline{C})$ that circumvents $\overline{C}$ in $(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})})^{m^+}$. We need to rule out the possibility of these walks, so generally we want to show that

$$A \perp_{\mathcal{D}^+} B \mid C \Longrightarrow (\overline{A} \setminus (\overline{C} \cup D)) \perp_{\left(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})}\right)^{m^+}} (\overline{B} \setminus (\overline{C} \cup D)) \mid (\overline{C} \cup D),$$

where

$$D = \left(\left(\overline{\mathrm{An}_{\mathcal{D}^+}(A)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(B)}\right) \cup \left(\overline{\mathrm{An}_{\mathcal{D}^+}(C)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(A)}\right) \cup \left(\overline{\mathrm{An}_{\mathcal{D}^+}(C)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(B)}\right)\right) \setminus \overline{\mathrm{an}_{\mathcal{D}^+}(\{J \in C : J \in \mathrm{An}_{\mathcal{D}^+}(\{A,B\})\})}.$$

Starting with the case $C = \emptyset$, consider a pair $\alpha \in \overline{A} \setminus \overline{\mathrm{An}_{\mathcal{D}^+}(B)}$ and $\beta \in \overline{B} \setminus \overline{\mathrm{An}_{\mathcal{D}^+}(A)}$ such that there exists a walk $w = \{\alpha := w_1, \ldots, w_n =: \beta\}$ in $(\mathcal{D}^+_{\mathrm{An}(\{A,B\})})^{m^+}$. Then there must exist an adjacent pair w_j, w_{j+1} such that $w_j \in \overline{\mathrm{An}_{\mathcal{D}^+}(A)} \setminus \overline{\mathrm{An}_{\mathcal{D}^+}(B)}$ and $w_{j+1} \in \overline{\mathrm{An}_{\mathcal{D}^+}(B)}$. Since A is separated from B by the empty set in $\mathcal{D}^+$, we have by definition of the ADAG-moral graph that adjacency between w_j and w_{j+1} occurs if and only if $w_j \in J, w_{j+1} \in J'$ for $J \in \mathrm{An}_{\mathcal{D}^+}(A), J' \in \mathrm{An}_{\mathcal{D}^+}(B)$ and either $w_j \in J'$ or $w_{j+1} \in J$. Hence, any walk between α and β must pass through $\overline{\mathrm{An}_{\mathcal{D}^+}(A)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(B)}$, and we conclude that $\overline{\mathrm{An}_{\mathcal{D}^+}(A)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(B)}$ separates $\overline{A} \setminus \overline{\mathrm{An}_{\mathcal{D}^+}(B)}$ from $\overline{B} \setminus \overline{\mathrm{An}_{\mathcal{D}^+}(A)}$ in $(\mathcal{D}^+_{\mathrm{An}(\{A,B\})})^{m^+}$.

When C is non-empty, we can use an analogous argument to show that conditioning further on $(\overline{\mathrm{An}_{\mathcal{D}^+}(C)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(A)}) \cup (\overline{\mathrm{An}_{\mathcal{D}^+}(C)} \cap \overline{\mathrm{An}_{\mathcal{D}^+}(B)})$ would block off any walk from from $\alpha \in \overline{A} \setminus (\overline{B} \cup \overline{C})$ to $\beta \in \overline{B} \setminus (\overline{A} \cup \overline{C})$ that circumvents $\overline{C}$ in $(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})})^{m^+}$. However, conditioning on every common variable among the ancestors is unnecessary; we can exclude some variables from this set. Suppose $J \in C$ is an ancestor of A (or B) and let J' be an ancestor of J in $\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})}$. Then there does not exist a path from J' to A (or B) that circumvents J, because otherwise there would be a node on the path from J to A (or B) with incoming arrows from two nodes, both sharing J' as an ancestor, contradicting P5. Thus, we do not need to include $\overline{J'}$ in the conditioning set D.

Next, we wish to show that separation in the ADAG-moral graph implies separation in the moral graph of $\mathcal{D}$ in the sense that $(\mathcal{D}_{\mathrm{An}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D)})^m = (V, E)$ is a sub-graph of $(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})})^{m^+} = (V^+, E^+)$ so that $V \subseteq V^+$ and $E \subseteq E^+$. Notice first that

$$(\overline{A} \setminus (\overline{C} \cup D)) \cup (\overline{B} \setminus (\overline{C} \cup D) \cup (\overline{C} \cup D) = \overline{A} \cup \overline{B} \cup \overline{C} \cup D.$$

Consider the moral graph $(\mathcal{D}_{\mathrm{An}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D)})^m$ with vertex set $V = \mathrm{An}_{\mathcal{D}}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D)$. By definition, adjacency between two nodes $\alpha, \beta \in V$ in this graph occurs if and

only if either (i) $\alpha \to \beta$, or (ii) $\alpha \leftarrow \beta$, or (iii) $\alpha \to \gamma$ and $\beta \to \gamma$ for some $\gamma \in V$ in $\mathcal{D}_{\mathrm{An}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D)}$.

It follows from the definition of the ADAG that

$$\mathrm{An}_{\mathcal{D}}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D) \subseteq \overline{\mathrm{An}_{\mathcal{D}^+}(\{A, B, C\})},$$

and thus $V \subseteq V^+$. We want to argue that if either (i)–(iii) is true then α and β are also adjacent in $(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})})^{m^+}$. If $\alpha \to \beta$ in $\mathcal{D}_{\mathrm{An}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D)}$ we must have that either $\alpha, \beta \in I$ for some $I \in \mathcal{I}_{\mathrm{An}(\{A,B,C\})}$ or $\alpha \in I$ and $\beta \in I'$ where (I, I') is an edge in $\mathcal{E}_{\mathrm{An}(\{A,B,C\})}$. The same is true with the roles are reversed, i.e. when $\alpha \leftarrow \beta$. If α and β have a common child γ then either α and β appear in the same node I or $\alpha \in I$ and $\beta \in I'$ appear in two different nodes while γ may appear in either I, I' or in a third node I''. If $\gamma \in I$ then there must be an arrow $I' \to I$, if $\gamma \in I'$ then there must be an arrow $I' \to I$, and if $\gamma \in I''$ then there must be arrows $I \to I''$ and $I' \to I''$. By definition, α is adjacent to β in $(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})})^{m^+}$ in all of the above cases and thus $E \subseteq E^+$.

The statement in the proposition now follows as a corollary since

$$\begin{aligned} A \perp_{\mathcal{D}^*} B \mid C &\implies (\overline{A} \setminus (\overline{C} \cup D)) \perp_{\left(\mathcal{D}^+_{\mathrm{An}(\{A,B,C\})}\right)^{m^+}} (\overline{B} \setminus (\overline{C} \cup D)) \mid (\overline{C} \cup D) \\ &\implies (\overline{A} \setminus (\overline{C} \cup D)) \perp_{\left(\mathcal{D}_{\mathrm{An}(\overline{A} \cup \overline{B} \cup \overline{C} \cup D))}\right)^{m}} (\overline{B} \setminus (\overline{C} \cup D)) \mid (\overline{C} \cup D) \\ &\implies (\overline{A} \setminus (\overline{C} \cup D)) \perp_{\mathcal{D}} (\overline{B} \setminus (\overline{C} \cup D)) \mid (\overline{C} \cup D), \end{aligned}$$

with the ultimate implication due to separation in the moral graph of the smallest ancestral set containing all the variables involved being equivalent to d-separation.

□

Lemma 6.B.1. *Let $\mathcal{D} = (V, E)$ be a DAG. If $A, B \subseteq V$ then $\mathrm{An}_{\mathcal{D}}(A) \cap \mathrm{An}_{\mathcal{D}}(B) = \emptyset \iff A \perp_{\mathcal{D}} B$.*

Proof. It suffices to show that $\alpha \perp_{\mathcal{D}} \beta$ for singletons α and β, because $A \perp_{\mathcal{D}} B$ if and only if $\alpha \perp_{\mathcal{D}} \beta$ for every pair of $\alpha \in A$ and $\beta \in B$. This follows since $\perp_{\mathcal{D}}$ is a compositional graphoid, satisfying decomposition and composition in particular, that is for disjoint subsets A, B, C of V then

$$A \perp_{\mathcal{D}} (B \cup C) \iff A \perp_{\mathcal{D}} B \text{ and } A \perp_{\mathcal{D}} C.$$

The forward direction is immediate from the definition of d-separation. For the backward direction we proceed by contraposition. Suppose that $\mathrm{An}(\alpha) \cap \mathrm{An}(\beta) \neq \emptyset$. Then either (i) $\{\alpha\} \subseteq \mathrm{An}(\alpha) \cap \mathrm{An}(\beta)$ in which case there is a directed path from α to β, (ii) $\{\beta\} \subseteq \mathrm{An}\,(\alpha) \cap \mathrm{An}(\beta)$ in which case there is a directed path from β to α,

or (iii) α and β share a common ancestor $\gamma \in V$ from which there is a directed path to both α and β. Thus, α is not d-separated from β by the empty set and the proof is complete. □

Lemma 6.B.2. *Consider a DAG $\mathcal{D} = (V, E)$ and a set of nodes $I \subseteq V$. The* `separateSet`*-algorithm constructs a family of sets $(A_j : j \in \mathcal{J})$ such that $\cup_{j\in\mathcal{J}} A_j = I$, and $A_j \perp_{\mathcal{D}} A_{j'}$ for all $j, j' \in \mathcal{J}$ with $j \neq j'$ and such that $\forall j \in \mathcal{J}$ there does not exist an $A \subsetneq A_j : A \perp_{\mathcal{D}} (A_j \setminus A)$.*

Proof. It follows from Lemma 6.B.1 that the undirected graph

$$\mathcal{U} = (I, \{\{\alpha, \beta\} : \mathrm{An}_{\mathcal{D}}(\alpha) \cap \mathrm{An}_{\mathcal{D}}(\beta) \neq \emptyset, \alpha, \beta \in I\}),$$

is equivalent to the undirected graph

$$(I, \{\{\alpha, \beta\} : \alpha \not\perp_{\mathcal{D}} \beta, \alpha, \beta \in I\}).$$

A connected component of $\mathcal{U}$ is a subgraph in which any vertex is reachable from any other vertex by traversing edges. Denote by $\mathcal{A} = (A_j, j \in \mathcal{J})$ the connected components of $\mathcal{U}$. The component sets are non-empty, pairwise disjoint and collectively cover I. The first two properties of the statement follow directly from these facts. For the third property, suppose that there is an A_j for which there is proper subset $A \subsetneq A_j$ such that $A \perp_{\mathcal{D}} (A_j \setminus A)$. But then $\beta \in (A_j \setminus A)$ is not reachable from $\alpha \in A$ in $\mathcal{U}$ contradicting that $\alpha \in A_j$. □

Corollary 6.B.3. *Let $\mathcal{T} = (\mathcal{I}, \mathcal{E})$ be a PD$^+$ of a DAG $\mathcal{D}$ with a single terminal vertex under the restriction that $A = \mathrm{An}_{\mathcal{D}}(\mathrm{Pa}_{\mathcal{D}}(I)) \cap I$ for every $I \in \mathcal{I}$ in P2. Suppose there is an $I_0 \in \mathcal{I}$ such that $\mathrm{pa}_{\mathcal{T}}(I_k) = I_{k+1}$ for all $k \in \{0, \ldots, L\}$. If there is an $A \subsetneq I_0$ such that $B = I_0 \setminus A$ satisfies $\mathrm{An}_{\mathcal{D}}(A) \cap \mathrm{An}_{\mathcal{D}}(B) = \emptyset$, it holds that $\mathcal{T}$ is still a PD$^+$ in the modified graph where the nodes $(I_k)_{k\in\{0,\ldots,L\}}$ are replaced by nodes $((I_{k,1})_{k\in\{0,\ldots,L_1\}}, (I_{k,2})_{k\in\{0,\ldots,L_2\}})$, defined by $I_{k+n,1} := I_k \cap \mathrm{An}_{\mathcal{D}}(A)$ and $I_{k+n,2} := I_k \cap \mathrm{An}_{\mathcal{D}}(B)$ for $k \in \{0, \ldots, L\}$ and the edges (I_{k+1}, I_k) are replaced by edges $(I_{k+1,1}, I_{k,1})$ and $(I_{k+1,2}, I_{k,2})$ for $k \in \{0, \ldots, L-1\}$, while the edge (I_0, I_{-1}) is replaced by edges $(I_{0,1}, I_{-1})$ and $(I_{0,2}, I_{-1})$, whenever the sets involved are non-empty.*

Proof. Since any node in $\mathcal{T}$ has at most one child, the statement concerning the edge $I_0 \to I_{-1}$ is well-defined. To see this, suppose $I \in \mathcal{I}$ has at least two children $I', I'' \in \mathcal{I}$. If I' and I'' share a common descendant then P5 is violated. If I' and I'' do not have any descendants in common then $\mathcal{D}$ must have more than one sink node.

We want to argue that the modified graph is a PD$^+$. Clearly, P1–P3 and P5 are satisfied by the construction of the two paths, but P4 requires an argument. Let $p = (I_L, \ldots, I_0, I_{-1}, \ldots, I_{-n})$ be the unique path in $\mathcal{T}$ from I_L to the sink node $I_{-n} := \{s\}$. Since $\mathcal{T}$ is a PD$^+$, we have that $I_k \perp_{\mathcal{D}} I$ for all $k \in \{0, \ldots, L\}$

and $I \in \mathcal{I}$ that are not in p. Thus, it suffices to consider whether the family of sets $((I_{k,1})_{k\in\{0,\dots,L_1\}}, (I_{k,2})_{k\in\{0,\dots,L_2\}}, (I_{-k})_{k\in\{1,\dots,n\}})$ is reduced. Since $\mathcal{T}$ is a PD^+, $(I_{-k})_{k\in\{1,\dots,n\}}$, $(I_{k,1})_{k\in\{0,\dots,L_1\}}$ and $(I_{k,2})_{k\in\{0,\dots,L_2\}}$ are reduced. Further, since A and B have no common ancestors, the intersection of any two sets $I_{k,1}$ and $I_{k,2}$, $k, k' \in \{0,\dots,L\}$ is always empty and we cannot have $I_{k,1} \subseteq I_{k',2}$ or $I_{k,1} \supseteq I_{k',2}$ (unless both sets are empty). For given $w \in \{1,2\}$, if $I_{k,w} \subseteq I_{-l}$ for any $k \in \{1,\dots,L_w\}$ and $l \in \{1,\dots,n\}$, then, due to P3, $I_{k,w} \subseteq I_{k-1,w}$ contradicting that the two are mutually reduced. If $I_{0,w} \subseteq I_{-l}$, then $I_{0,w} \subseteq I_{-1}$ by P3, but, due to the restricted version of P2 in the corollary, there will always be a $\beta \in I_{0,w}$ such that $\beta \in \mathrm{Pa}_{\mathcal{D}}(I_{-1})$, so this cannot be. Finally $I_{-l} \subseteq I_{k,w}$ any $\{0,\dots,L_w\}$ is not possible due to the construction of the I-sets following the restricted version of P2 in the corollary. We conclude that P5 holds for the modified vertex set and we are done.

□

6.D Alternative Decompositions

Figures 6.10–6.11 give examples different decompositions satisfying P2. In the graphs, each node represents a cluster of nodes with the number detailing how many connected components each node contains. The colored boxes represent the parents appearing in the corresponding parental decomposition. Figure 6.10 gives an example of how two DAGs of similar structure lead to different optimal aggregated structures (in terms of width) depending on the number of nodes within the clusters. Figure 6.11 gives an example of how the choice of a parents at a given level can depend on future options.

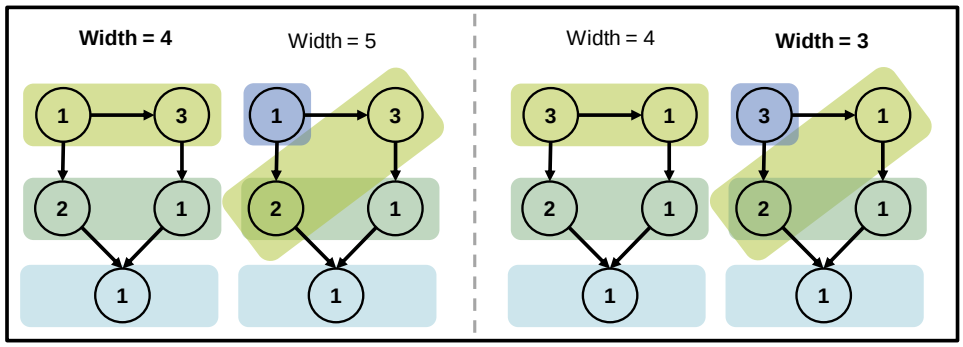

Figure 6.10: The graphs have the same overall structure. However, different sets of parents are preferred in terms of width depending on the number of nodes in each cluster.

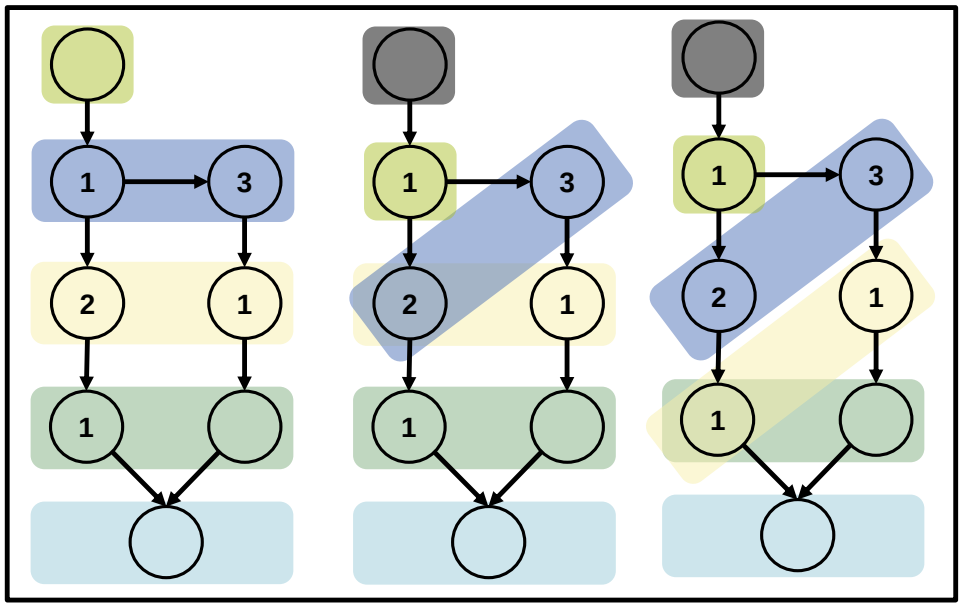

Figure 6.11: The three viable parent sets in this graph structure. Choosing whether to place the yellow box horizontally or diagonally depends on how one can subsequently place the dark blue box.

6.E Description of Data

In this section, we describe the data and preprocessing steps used in the applications in Section 6.3.2 and Section 6.4.5.

Data on the distributions of education level, income, alcohol consumption, activity level, smoking behaviour and overweight is from IPUMS (2022). The Integrated Public Use Microdata Series (IPUMS) database provides survey data at the individual level based on the annual National Health Interview Survey. Specifically we have used the following variables, mapped into dichotomous versions:

- Income: The variable 'POORYN' which indicates whether family income was above or below poverty level is used as is.
- Education: The variable 'EDUCREC2' reports the respondent's highest attained level of education. We follow the standardized education recoding and classify a respondent's level of education as low if the highest attained level is 8th Grade or lower.
- Alcohol: The variable 'ALCDAYSYR', describing the frequency with which respondents consumed alcohol, reported as number of days in the past year, is combined with 'ALCAMT', describing average number of drinks on days drank,

to determine daily average alcohol consumption. We classify respondents as light drinkers if average consumption is less than 12 grams of alcohol per day.

- Physical activity: We use a combination of the variables 'MOD10DMIN' and 'VIG10DMIN', that describe the duration of moderate and vigorous activity of 10+ minutes in minutes. We convert the time spent exercising into MET-minutes by multiplying moderate activity minutes by 5 and vigorous activity minutes by 9. We categorize respondents as being in the low physical activity category if the number of daily MET-minutes falls below 600.
- Smoking: We use variable 'SMOKESTATUS2' to determine the current smoking status of respondents.
- Overweight: We use the variable 'BMI' reporting the Body Mass Index, a measure of body fat based on height and weight, to determine whether or not an individual is overweight. Following the standard WHO classification, an individual is categorized as being overweight if the BMI exceeds 25.

Missing observations are distributed proportionally among categories.

Data on the prevalence of adults who have high blood cholesterol is from CDC (2022). The data is based on the Behavioral Risk Factor Surveillance System which is an annual health-related telephone survey among U.S. residents. The respondents are categorized as having high blood cholesterol if, after having their cholesterol checked, they have ever been told by a health professional that it was high. After 1999, the dataset covers unisex values for the years 1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017 and 2019. To obtain prevalence estimates for all calendar years in the period 1999–2018, a quadratic regression was fitted and used as a smoothed estimate.

Data on the prevalence of adults who have raised blood pressure is from NCD-RisC (2017). Raised blood pressure is defined as having a systolic blood pressure at or above 140 mm Hg, or a diastolic blood pressure at or above 90 mm Hg. The dataset covers all years in the period 1999–2015. To obtain data for 2016–2018, a quadratic regression was fitted and used to predict missing values.

Data on relative-risks are obtained from Murray et al. (2020). The relative-risks are reported in five-year age groups at various levels of exposure. We match the risk distributions with the relative-risks estimates as follows. Light consumers of alcohol are assigned the baseline relative risk of unity, while non-light consumers are assigned the relative risk corresponding to a consumption of 36 g/day. Individuals at low levels of activity are assigned a relative risk corresponding to 0 MET-minutes of activity, while individuals at non-low levels are assigned a relative risk corresponding to 1200 MET-minutes of activity. Non-smokers are assigned the baseline relative risk of unity, while smokers are assigned the relative risk corresponding to smoking

20 cigarettes a day. Non-overweight individuals are assigned the baseline relative risk of unity, while overweight individuals are assigned a relative risk corresponding to a Class I Obese (i.e., a BMI of 30–35). Individuals with non-high cholesterol are assigned the baseline relative risk of unity, while individuals with high cholesterol are assigned the relative risk associated with a heightened level of LDL cholesterol of 1 mmol/L. Individuals with non-high blood pressure are assigned the baseline relative risk of unity, while individuals with high blood pressure are assigned the relative risk associated with a level heightened by 10 mm Hg.

6.F PCA for ilr-transformed Data

Consider a matrix X containing n observations of data from the d-dimensional simplex, that is, each row of X corresponds to an observation from

$$\mathbb{S}^d = \left\{ x = (x_1, \dots, x_d)^\top : x \succ 0, \sum_{i=1}^{d} x_i = 1 \right\}.$$

We want to transform each observation from $\mathbb{S}^d$ onto $\mathbb{R}^d$ (or $\mathbb{R}^{d-1}$) so that standard multivariate tools are at our disposal. In particular, we want to apply principal component analysis. The usual transformations for this purpose are the centered-log-ratio (clr) transformation (Aitchison, 1986) and the isometric-log-ratio (ilr) transformation (Egozcue, 2003). The clr-transform centers the data around the geometric mean. It is defined as

$$y = \operatorname{clr}(x) := \left(\log \frac{x_1}{\sqrt[d]{\prod_{i=1}^d x_j}}, \dots, \log \frac{x_d}{\sqrt[d]{\prod_{i=1}^d x_j}} \right)^\top,$$

sending each composition $x \in \mathbb{S}^d$ to a vector $y \in \mathbb{R}^d$ satisfying $\sum_{i=1}^d y_i = 0$. Thus, the data are collinear; such a vector constitutes a $d-1$ dimensional subspace of $\mathbb{R}^d$. The ilr-transformation is based on a choice of orthonormal basis $V = (v_1, \dots, v_{d-1})$ of this subspace so that the transformed data $\operatorname{ilr}(x) := V^\top \operatorname{clr}(x)$ are non-collinear. Egozcue (2003) suggests using

$$\mathbb{R}^d \ni v_i = \sqrt{\frac{i}{i+1}} \left(\underbrace{\frac{1}{i}, \dots, \frac{1}{i}}_{i \text{ elements}}, -1, 0, \dots, 0 \right)^\top, \quad i = 1, \dots, d-1,$$

which we also adopt here.

Let 1_d denote a d-dimensional vector of ones and I_d a d-dimensional identity

matrix. In matrix notation, we then have

$$\begin{aligned} \text{Original data} : X &\in \mathbb{R}^{n\times d}, \\ \text{clr-transformed data} : Y = \log(X)C^\top &\in \mathbb{R}^{n\times d}, \\ \text{ilr-transformed data} : Z = YV &\in \mathbb{R}^{n\times(d-1)}, \end{aligned}$$

where $C = I_d - \frac{1}{d}1_d 1_d^\top$. The (full) principal components decomposition of Z is then

$$Z^* = (Z - 1_n L(Z)^\top)W,$$

where $L(Z)$ contains the empirical means of the columns of Z, and W is the weights-matrix obtained from an eigendecomposition $W\Lambda W^\top$ of $(Z^\top - L(Z)1_n^\top)(Z - 1_n L(Z)^\top)$, where Λ is a diagonal matrix of eigenvalues and the columns of W are the corresponding eigenvectors (i.e., the PC-loadings). Robust (covariance) estimation is also an option when working in ilr-space, see Filzmoser et al. (2009).

For dimensionality reduction, only the $p < d$ first principal components are used. We can then back-transform the (truncated) transformed data from ilr-space to clr-space, and finally back onto the simplex:

$$\hat{X} = \mathfrak{C}\left[\exp\left\{\left(Z_p^* W_p^\top + 1_n L(Z)\right) V^\top\right\}\right],$$

where the p-subscript is used to denote extraction of the p first columns of a matrix, exp is applied element-wise and $\mathfrak{C}[M] = M \oslash ((M1_d) \otimes 1_d^\top)$ for a $n \times d$ matrix M with $\oslash$ denoting Hadamard-division and $\otimes$ the Kroenecker product.

Bibliography

Abbring, J. and Berg, G. van den (2007). The Unobserved Heterogeneity Distribution in Duration Analysis. *Biometrika*, **94**(1), pp. 87–99. DOI: 10.1093/biomet/asm013.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK: Chapman and Hall. DOI: 10.1007/978-94-009-4109-0.

Alai, D. H., Arnold, S., and Sherris, M. (2015). Modelling cause-of-death mortality and the impact of cause-elimination. *Annals of Actuarial Science*, **9**(1), pp. 167–186. DOI: 10.1017/S174849951400027X.

Antonio, K. et al. (2017). Producing the Dutch and Belgian mortality projections: a stochastic multi-population standard. *European Actuarial Journal*, **7**(2), pp. 297–336. DOI: 10.1007/s13385-017-0159-x.

Arnold, S., Jijiie, A., Jondeau, E., and Rockinger, M. (2019). Periodic or generational actuarial tables: Which one to choose? *European Actuarial Journal*, **9**(2), pp. 519–554. DOI: 10.2139/ssrn.3099103.

Arnold, S. and Sherris, M. (2016). International Cause-Specific Mortality Rates: New Insights from a Cointegration Analysis. *ASTIN Bulletin: The Journal of the IAA*, **46**(1), pp. 9–38. DOI: 10.1017/asb.2015.24.

Arriaga, E. E. (1984). Measuring and explaining the change in life expectancies. *Demography*, **21**(1), pp. 83–96. DOI: 10.2307/2061029.

Beard, R. E. (1959). Appendix: Note on Some Mathematical Mortality Models. In: *Ciba Foundation Symposium - The Lifespan of Animals (Colloquia on Ageing, Vol. 5)*. Boston: Little, Brown and Company, pp. 302–311. DOI: 10.1002/9780470715253.app1.

Beckers, S., Eberhardt, F., and Halpern, J. Y. (2019). "Approximate Causal Abstractions". In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Beckers, S. and Halpern, J. Y. (2019). "Abstracting Causal Models". In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI Press. DOI: 10.1609/aaai.v33i01.33012678.

Bennett, J. E., Pearson-Stuttard, J., Kontis, V., Capewell, S., Wolfe, I., and Ezzati, M. (2018). Contributions of diseases and injuries to widening life expectancy inequalities in England from 2001 to 2016: a population-based analysis of vital registration data. *The Lancet Public Health*, **3**(12), e586–e597. DOI: 10.1016/s2468-2667(18)30214-7.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J., and Vaupel, J. W. (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, **37**(1), pp. 527–566. DOI: 10.4054/DemRes.2017.37.17.

Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **24**(3), pp. 179–195. DOI: 10.2307/2987782.

Beutner, E., Reese, S., and Urbain, J.-P. (2017). Identifiability issues of age–period and age–period–cohort models of the Lee–Carter type. *Insurance: Mathematics and Economics*, **75**, pp. 117–125. DOI: 10.1016/j.insmatheco.2017.04.006.

Bongaarts, J. (2005). Long-range trends in adult mortality: Models and projection methods. *Demography*, **42**(1), pp. 23–49. DOI: 10.1353/dem.2005.0003.

Booth, H. (2016). Epidemiologic Transition in Australia–the last hundred years. *Canadian Studies in Population*, **43**(1-2), pp. 23–47. DOI: 10.25336/P6VP5J.

Booth, H., Maindonald, J., and Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, **56**(3), pp. 325–336. DOI: 10.1080/00324720215935.

Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, **3**(1-2), pp. 3–43. DOI: 10.1017/S1748499500000440.

Börger, M. (2010). Deterministic shock vs. stochastic value-at-risk—an analysis of the Solvency II standard model approach to longevity risk. *Blätter der DGVFM*, **31**(2), pp. 225–259. DOI: 10.1007/s11857-010-0125-z.

Börger, M. and Aleksic, M.-C. (2014). Coherent projections of age, period, and cohort dependent mortality improvements. *Washington DC: Paper presented at the 30th International Congress of Actuaries*.

Börger, M., Fleischer, D., and Kuksin, N. (2014). Modeling the mortality trend under modern solvency regimes. *ASTIN Bulletin: The Journal of the IAA*, **44**(1), pp. 1–38. DOI: 10.1017/asb.2013.24.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3), pp. 199–215. DOI: 10.1214/ss/1009213726.

Brouhns, N., Denuit, M., and Keilegom, I. V. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, **2005**(3), pp. 212–224. DOI: 10.1080/03461230510009754.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A Poisson Log-Bilinear Regression Approach to the Construction of Projected Lifetables. *Insurance: Mathematics and Economics*, **31**(3), pp. 373–393. DOI: 10.1016/S0167-6687(02)00185-3.

Butt, Z. and Haberman, S. (2004). Application of Frailty-Based Mortality Models Using Generalized Linear Models. *ASTIN Bulletin: The Journal of the IAA*, **34**(1), pp. 175–197. DOI: 10.2143/AST.34.1.504961.

Cairns, A., Blake, D., and Dowd, K. (2006). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, **73**(4), pp. 687–718. DOI: 10.1111/j.1539-6975.2006.00195.x.

Cairns, A., Blake, D., and Dowd, K. (2008). Modelling and management of mortality risk: A review. *Scandinavian Actuarial Journal*, **2008**(2-3), pp. 79–113. DOI: 10.1080/03461230802173608.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., Epstein, D., Ong, A., and Balevich, I. (2009). A Quantitative Comparison of Stochastic Mortality Models Using Data From England & Wales and the United States. *North American Actuarial Journal*, **13**, pp. 1–35. DOI: 10.1080/10920277.2009.10597538.

Cairns, A., Dowd, K., Blake, D., and Coughlan, G. D. (2011a). Longevity hedge effectiveness: a decomposition. *Quantitative Finance*, **14**(2), pp. 217–235. DOI: 10.1080/14697688.2012.748986.

Cairns, A. J. (2000). A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, **27**(3), pp. 313–330. DOI: 10.1016/s0167-6687(00)00055-x.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011b). Bayesian Stochastic Mortality Modelling for Two Populations. *ASTIN Bulletin: The Journal of the IAA*, **41**(1), pp. 29–59. DOI: 10.2143/AST.41.1.2084385.

Cairns, A. J., Kallestrup-Lamb, M., Rosenskjold, C., Blake, D., and Dowd, K. (2019). Modelling socio-economic differences in mortality using a new affluence index. *ASTIN Bulletin: The Journal of the IAA*, **49**(3), pp. 555–590. DOI: 10.1017/asb.2019.14.

Carter, L. R. and Lee, R. D. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting*, **8**(3), pp. 393–411. DOI: 10.1016/0169-2070(92)90055-E.

CDC (2022). *Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. BRFSS Prevalence & Trends Data [online]*. https://www.cdc.gov/brfss/brfssprevalence/.

CDC WONDER (2020). *Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2018 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2018, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.* `http://wonder.cdc.gov/ucd-icd10.html`.

Chalupka, K., Perona, P., and Eberhardt, F. (2015). "Visual Causal Feature Learning". In: *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Chalupka, K., Perona, P., and Eberhardt, F. (2016). "Multi-Level Cause-Effect Systems". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. PMLR.

Chen, R. Y. and Millossovich, P. (2018). Sex-specific mortality forecasting for UK countries: a coherent approach. *European Actuarial Journal*, **8**(1), pp. 69–95. DOI: `10.1007/s13385-017-0164-0`.

Coleman, D. A. (1992). The Demographic Transition in Ireland in International Context. *Proceedings of the British Academy*, **79**, pp. 53–57.

Continuous Mortality Investigation (2016). *CMI Mortality Projections Model-Working Paper 90, Institute and Faculty of Actuaries*. `https://www.actuaries.org.uk/system/files/field/document/CMI%20WP090%20v03%202016-08-31%20-%20CMI%20Model%20consultation.pdf`.

Cui, Q., Canudas-Romo, V., and Booth, H. (2019). The Mechanism Underlying Change in the Sex Gap in Life Expectancy at Birth: An Extended Decomposition. *Demography*, **56**(6), pp. 2307–2321. DOI: `10.1007/s13524-019-00832-z`.

Currie, I. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, **2016**(4), pp. 356–383. DOI: `10.1080/03461238.2014.928230`.

Darkiewicz, G. and Hoedemakers, T. (2004). *How the Co-integration Analysis Can Help in Mortality Forecasting*. Katholieke Universiteit Leuven, Department of Applied Economics.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**(2), pp. 161–189. DOI: `10.1111/j.1751-5823.2002.tb00354.x`.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Didelez, V. (2000). *Graphical models for event history analysis based on local independence*. Berlin, Germany: Logos.

Dimitrova, D., Haberman, S., and Kaishev, V. (2013). Dependent competing risks: Cause elimination and its impact on survival. *Insurance: Mathematics and Economics*, **53**(2). DOI: 10.1016/j.insmatheco.2013.07.008.

Dobra, A. and Fienberg, S. E. (2009). The generalised shuttle algorithm. In: *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, pp. 135–156. DOI: 10.1017/cbo9780511642401.009.

Dobra, A. and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statistical Journal of the United Nations Economic Commission for Europe*, **18**(4), pp. 363–371. DOI: 10.3233/sju-2001-18411.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., and Khalaf-Allah, M. (2011). A Gravity Model of Mortality Rates for Two Related Populations. *North American Actuarial Journal*, **15**(2), pp. 334–356. DOI: 10.1080/10920277.2011.10597624.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. 1st ed. 2008. Statistics for Biology and Health. New York, NY: Springer New York. ISBN: 1-281-10794-8.

Duncan, O. D. and Davis, B. (1953). An Alternative to Ecological Correlation. *American Sociological Review*, **18**(6), p. 665. DOI: 10.2307/2088122.

Egozcue, J. J. (2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, **35**(3), pp. 279–300. DOI: 10.1023/a:1023818214614.

Eichler, M. and Didelez, V. (2007). "Causal Reasoning in Graphical Time Series Models". In: *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, pp. 109–116.

Eichler, M. and Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, **16**(1), pp. 3–32. DOI: 10.1007/s10985-009-9143-3.

Engle, R. and Granger, C. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, **55**(2), pp. 251–276.

Ezzati, M., Vander Hoorn, S., Rodgers, A., Lopez, A. D., Mathers, C. D., Murray, C. J., Group, C. R. A. C., et al. (2003). Estimates of global and regional potential

health gains from reducing multiple major risk factors. *The Lancet*, **362**(9380), pp. 271–280.

Filzmoser, P., Hron, K., and Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, **20**(6), pp. 621–632. DOI: 10.1002/env.966.

Finanstilsynet (2010). *Brev om tilsyn med livsforsikringsselskabers og tværgående pensionskassers levetidsforudsætninger*. URL: https://www.finanstilsynet.dk/-/media/Tal-og-fakta/2010/Brev_LP.pdf.

Flaxman, S. R., Wang, Y.-X., and Smola, A. J. (2015). "Who Supported Obama in 2012?" In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. DOI: 10.1145/2783258.2783300.

Foreman, K. et al. (2018). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. *The Lancet*, **392**(10159), pp. 2052–2090. DOI: 10.1016/S0140-6736(18)31694-5.

Frogner, C. and Poggio, T. (2019). "Fast and Flexible Inference of Joint Distributions from their Marginals". In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR. URL: https://proceedings.mlr.press/v97/frogner19a.html.

Gaille, S. and Sherris, M. (2011). Modelling Mortality with Common Stochastic Long-Run Trends. *The Geneva Papers on Risk and Insurance - Issues and Practice*, **36**(4), pp. 595–621. DOI: 10.1057/gpp.2011.19.

Gao, G. and Shi, Y. (2021). Age-coherent extensions of the Lee-Carter model. *Scandinavian Actuarial Journal*, **2021**(10), pp. 1–19. DOI: 10.1080/03461238.2021.1918578.

Girosi, F. and King, G. (2008). *Demographic forecasting*. Princeton University Press.

Glei, D. A. and Horiuchi, S. (2007). The narrowing sex differential in life expectancy in high-income populations: Effects of differences in the age pattern of mortality. *Population studies*, **61**(2), pp. 141–159. DOI: 10.1080/00324720701331433.

Goldman, N. and Lord, G. (1986). A new look at entropy and the life table. *Demography*, **23**(2), pp. 275–282. DOI: 10.2307/2061621.

Gompertz, B. (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, **115**, pp. 513–583. DOI: 10.1098/rstl.1825.0026.

Goodman, L. A. (1953). Ecological Regressions and Behavior of Individuals. *American Sociological Review*, **18**(6), p. 663. DOI: 10.2307/2088121.

Goodman, L. A. (1959). Some Alternatives to Ecological Correlation. *American Journal of Sociology*, **64**(6), pp. 610–625. DOI: 10.1086/222597.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pp. 424–438. DOI: 10.2307/1912791.

Greenland, S. and Robins, J. (1994). Invited Commentary: Ecologic Studies–Biases, Misconceptions, and Counterexamples. *American Journal of Epidemiology*, **139**(8), pp. 747–760. DOI: 10.1093/oxfordjournals.aje.a117069.

Gresele, L., Kügelgen, J. V., Kübler, J., Kirschbaum, E., Schölkopf, B., and Janzing, D. (2022). "Causal Inference Through the Structural Causal Marginal Problem". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR. URL: https://proceedings.mlr.press/v162/gresele22a.html.

Haberman, S. and Renshaw, A. (2012). Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and Economics*, **50**(3), pp. 309–333. DOI: 10.1016/j.insmatheco.2011.11.005.

Hansen, P. R. (2005). Granger's Representation Theorem: A Closed-Form Expression for I(1) Processes. *Econometrics Jorunal*, **8**(1), pp. 23–38. DOI: 10.1111/j.1368-423X.2005.00149.x.

Hári, N., Waegenaere, A. D., Melenberg, B., and Nijman, T. E. (2008). Longevity risk in portfolios of pension annuities. *Insurance: Mathematics and Economics*, **42**(2), pp. 505–519. DOI: 10.1016/j.insmatheco.2007.01.012.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What if*. Boca Raton: Chapman & Hall/CRC.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, **71**(1), pp. 75–83. DOI: 10.1093/biomet/71.1.75.

Hougaard, P. (1986). Survival Models for Heterogeneous Populations Derived from Stable Distributions. *Biometrika*, **73**(2), pp. 387–396.

Hougaard, P. (2012). *Analysis of Multivariate Survival Data*. Berlin: Springer Science & Business Media.

Human Mortality Database (2019). http://www.mortality.org. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded April 2019.

Human Mortality Database (2021). http://www.mortality.org. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded May 2021.

Human Mortality Database (2022). http://www.mortality.org. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded January 2022.

Hunt, A. and Blake, D. (2015a). Identifiability in age/period mortality models (WP). *The Pensions Institute, Cass Business School, City University London*, **PI-1508**.

Hunt, A. and Blake, D. (2015b). Identifiability in age/period/cohort mortality models (WP). *The Pensions Institute, Cass Business School, City University London*, **PI-1509**.

Hunt, A. and Blake, D. (2015c). Modelling longevity bonds: Analysing the Swiss Re Kortis bond. *Insurance: Mathematics and Economics*, **63**, pp. 12–29. DOI: 10.1016/j.insmatheco.2015.03.017.

Hunt, A. and Blake, D. (2017). Modelling mortality for pension schemes. *ASTIN Bulletin: The Journal of the IAA*, **47**(2), pp. 601–629. DOI: 10.1017/asb.2016.40.

Hunt, A. and Blake, D. (2018). Identifiability, cointegration and the gravity model. *Insurance: Mathematics and Economics*, **78**, pp. 360–368. DOI: 10.1016/j.insmatheco.2017.09.014.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, **50**(1), pp. 261–283. DOI: 10.1007/s13524-012-0145-5.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., and Järvisalo, M. (2013). "Discovering Cyclic Causal Models with Latent Variables: A General SAT-Based Procedure". In: *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge, UK: Cambridge University Press. ISBN: 9781139025751.

IMF (2012). *Chapter 4: The Financial Impact of Longevity Risk*. USA: International Monetary Fund, pp. 1–32. ISBN: 9781616352479. DOI: 10.5089/9781616352479.082.

IPUMS (2019). *Blewett, Lynn A. and Drew, Julia A. R. and King, Miriam L. and Williams, Kari C.W. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]. Minneapolis, MN: IPUMS, 2019.* https://www.nhis.ipums.org. DOI: 10.18128/D070.V6.4.

IPUMS (2022). *Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Natalie Del Ponte and Pat Convey. IPUMS Health Surveys: National Health Interview Survey, Version 7.1 [dataset]. Minneapolis, MN: IPUMS, 2021.* https://www.nhis.ipums.org. DOI: 10.18128/D070.V7.1. URL: http://www.nhis.ipums.org.

Jallbjørn, S. and Hansen, N. R. (2022). Aggregated Structural Causal Models. *Working paper.*

Jallbjørn, S., Jarner, S. F., and Hansen, N. R. (2022). Forecasting, Interventions and Selection: The Benefits of a Causal Mortality Model. *Submitted for publication.*

Jallbjørn, S. and Jarner, S. F. (2022). Sex Differential Dynamics in Coherent Mortality Models. *Forecasting*, **4**(4), pp. 819–844. DOI: 10.3390/forecast4040045.

Janssen, F. (2018). Advances in mortality forecasting: Introduction. *Genus*, **74**(1), pp. 1–12. DOI: 10.1186/s41118-018-0045-7.

Janssen, F. and Kunst, A. (2007). The choice among past trends as a basis for the prediction of future trends in old-age mortality. *Population Studies*, **61**(3), pp. 315–326. DOI: 10.1080/00324720701571632.

Janssen, F., Wissen, L., and Kunst, A. (2013). Including the Smoking Epidemic in Internationally Coherent Mortality Projections. *Demography*, **50**(4), pp. 1341–1362. DOI: 10.1007/s13524-012-0185-x.

Janzing, D. (2018). Merging joint distributions via causal model classes with low VC dimension. *preprint arXiv:1804.03206*.

Jarner, S. F. and Jallbjørn, S. (2020). Pitfalls and merits of cointegration-based mortality models. *Insurance: Mathematics and Economics*, **90**, pp. 80–93. DOI: 10.1016/j.insmatheco.2019.10.005.

Jarner, S. F. and Jallbjørn, S. (2022). The SAINT Model: A Decade Later. *ASTIN Bulletin: The Journal of the IAA*, **52**(2), pp. 483–517. DOI: 10.1017/asb.2021.37.

Jarner, S. F. and Kryger, E. M. (2011). Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin: The Journal of the IAA*, **41**(2), pp. 377–418. DOI: 10.2143/AST.41.2.2136982.

Jarner, S. F., Kryger, E. M., and Dengsøe, C. (2008). The evolution of death rates and life expectancy in Denmark. *Scandinavian Actuarial Journal*, **2008**(2-3), pp. 147–173. DOI: 10.1080/03461230802079193.

Jarner, S. F. (2014). Stochastic frailty models for modeling and forecasting mortality. *arXiv preprint arXiv:2109.02584*, pp. 1–37.

Jarner, S. F. and Møller, T. (2015). A partial internal model for longevity risk. *Scandinavian Actuarial Journal*, **2015**(4), pp. 352–382. DOI: 10.1080/03461238.2013.836561.

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, **59**(6), pp. 1551–1580. DOI: 10.2307/2938278.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press. ISBN: 978-0198774501.

Johansen, S. and Juselius, K. (1992). Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of Econometrics*, **53**(1), pp. 211–244. DOI: 10.1016/0304-4076(92)90086-7.

Juel, K. (2008). Life expectancy and mortality in Denmark compared to Sweden. What is the effect of smoking and alcohol? [Middellevetid og dødelighed i Danmark sammenlignet med i Sverige: Hvad betyder rygning og alkohol?] *Ugeskrift for Laeger*, **170**(33), pp. 2423–2427.

Juselius, K. (2006). *The Cointegrated VAR Model: Methodology and Applications*. Oxford University Press.

Kaishev, V., Dimitrova, D., and Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions. *Insurance: Mathematics and Economics*, **41**, pp. 339–361. DOI: 10.1016/j.insmatheco.2006.11.006.

Kaishev, V., Dimitrova, D., Haberman, S., and Verrall, R. (2016). Geometrically designed, variable knot regression splines. *Computational Statistics*, **31**(3), pp. 1079–1105. DOI: 10.1007/s00180-015-0621-7.

Kalben, B. B. (2000). Why Men Die Younger: Causes of Mortality Differences by Sex. *North American Actuarial Journal*, **4**(4), pp. 83–111. DOI: 10.1080/10920277.2000.10595939.

Kallestrup-Lamb, M., Kjærgaard, S., and Rosenskjold, C. P. T. (2020). Insight into stagnating adult life expectancy: Analyzing cause of death patterns across socioeconomic groups. *Health Economics*, **29**(12), pp. 1728–1743. DOI: 10.1002/hec.4166.

Kannisto, V., Lauritsen, J., Thatcher, A. R., and Vaupel, J. W. (1994). Reductions in Mortality at Advanced Ages: Several Decades of Evidence from 27 Countries. *Population and development review*, **20**(4), pp. 793–810. DOI: 10.2307/2137662.

Karn, M. N. (1931). An inquiry into various death-rates and the comparative influence of certain diseases on the duration of life. *Annals of Eugenics*, **4**(3-4), pp. 279–302. DOI: 10.1111/j.1469-1809.1931.tb02080.x.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, **332**(1627), pp. 487–509. DOI: 10.1098/rsta.1990.0128.

Keyfitz, N. (1977a). *Applied Mathematical Demography*. New York: John Wiley and Sons.

Keyfitz, N. (1977b). What difference would it make if cancer were eradicated? An examination of the Taeuber paradox. *Demography*, **14**(4), pp. 411–418. DOI: 10.2307/2060587.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press. ISBN: 0-691-01240-7.

King, G. and Soneji, S. (2011). The future of death in America. *Demographic research*, **25**, pp. 1–38. DOI: 10.4054/DemRes.2011.25.1.

Kjærgaard, S., Canudas-Romo, V., and Vaupel, J. (2016). "The importance of the reference populations for coherent mortality forecasting models". In: European Population Conference; Conference date: 31-08-2016 Through 03-09-2016.

Kleinow, T. (2015). A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, **63**, pp. 147–152. DOI: 10.1016/j.insmatheco.2015.03.023.

Koissi, M.-C., Shapiro, A. F., and Högnäs, G. (2006). Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. *Insurance: Mathematics and Economics*, **38**(1), pp. 1–20. DOI: 10.1016/j.insmatheco.2005.06.008.

Kuang, D., Nielsen, B., and Nielsen, J. (2008). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika*, **95**(4), pp. 987–991. DOI: 10.1093/biomet/asn038.

Lauritzen, S. L. (1996). *Graphical Models*. Vol. 17. Oxford: Clarendon Press. ISBN: 0-19-852219-3.

Lazar, D. and Denuit, M. (2009). A multivariate time series approach to projected life tables. *Applied Stochastic Models in Business and Industry*, **25**(6), pp. 806–823. ISSN: 1526-4025. DOI: 10.1002/asmb.781.

Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, **38**(4), pp. 537–549. DOI: 10.1353/dem.2001.0036.

Lee, R. D. and Carter, L. R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, **87**(419), pp. 659–675. DOI: 10.2307/2290201.

Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Strassburg: Trübner.

Li, H. and Li, J. (2017). Optimizing the Lee-Carter Approach in the Presence of Structural Changes in Time and Age Patterns of Mortality Improvements. *Demography*, **54**(3), pp. 1073–1095. DOI: 10.1007/s13524-017-0579-x.

Li, H. and Lu, Y. (2017). Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin: The Journal of the IAA*, **47**(2), pp. 563–600. DOI: 10.1017/asb.2016.37.

Li, H. and Lu, Y. (2019). Modeling cause-of-death mortality using hierarchical Archimedean copula. *Scandinavian Actuarial Journal*, **2019**(3), pp. 247–272. DOI: 10.1080/03461238.2018.1546224.

Li, J. (2013). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population Studies*, **67**(1), pp. 111–126. DOI: 10.1080/00324728.2012.689316.

Li, J., Li, J. S.-H., Tan, C. I., and Tickle, L. (2019). Assessing basis risk in index-based longevity swap transactions. *Annals of Actuarial Science*, **13**(1), pp. 166–197. DOI: 10.1017/S1748499518000179.

Li, J. and Liu, J. (2019). A logistic two-population mortality projection model for modelling mortality at advanced ages for both sexes. *Scandinavian Actuarial Journal*, **2019**(2), pp. 97–112. DOI: 10.1080/03461238.2018.1511464.

Li, J. and Hardy, M. (2011). Measuring Basis Risk in Longevity Hedges. *North American Actuarial Journal*, **15**(2), pp. 177–200. DOI: 10.1080/10920277.2011.10597616.

Li, J. S.-H., Hardy, M. R., and Tan, K. S. (2009). Uncertainty in Mortality Forecasting: An Extension to the Classical Lee-Carter Approach. *ASTIN Bulletin: The Journal of the IAA*, **39**(1), pp. 137–164.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, **42**(3), pp. 575–594. DOI: 10.1353/dem.2005.0021.

Li, N., Lee, R., and Gerland, P. (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, **50**(6), pp. 2037–2051. DOI: 10.1007/s13524-013-0232-2.

Lindahl-Jacobsen, R., Rau, R., Jeune, B., Canudas-Romo, V., Lenart, A., Christensen, K., and Vaupel, J. W. (2016). Rise, stagnation, and rise of Danish women's life expectancy. *Proceedings of the National Academy of Sciences*, **113**(15), pp. 4015–4020. DOI: 10.1073/pnas.1602783113.

Lindsey, J. K. (1996). Parametric Statistical Inference. In: Oxford University Press. ISBN: 978-0198523598.

Lindvall, T. (2002). *Lectures on the coupling method*. New York: Dover Publications.

Mackenbach, J., Kunst, A., Lautenbach, H., Oei, Y., and Bijlsma, F. (1999). Gains in life expectancy after elimination of major causes of death: revised estimates taking into account the effect of competing causes. *Journal of Epidemiology and Community Health*, **53**(1), pp. 32–37. DOI: 10.1136/jech.53.1.32.

Makeham, W. M. (1867). On the Law of Mortality. *Journal of the Institute of Actuaries*, **13**(6), pp. 325–358. DOI: 10.1017/S2046166600003238.

Mäkelä, P. (1998). Alcohol-related mortality by age and sex and its impact on life expectancy: Estimates based on the Finnish death register. *The European Journal of Public Health*, **8**(1), pp. 43–51. DOI: 10.1093/eurpub/8.1.43.

Manton, K. G. and Poss, S. S. (1979). Effects of dependency among causes of death for cause elimination life table strategies. *Demography*, **16**(2), pp. 313–327. DOI: 10.2307/2061145.

Mejia, S. H. G., Kirschbaum, E., and Janzing, D. (2022). "Obtaining Causal Information by Merging Datasets with MAXENT". In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*. PMLR.

Menzietti, M., Morabito, M. F., and Stranges, M. (2019). Mortality Projections for Small Populations: An Application to the Maltese Elderly. *Risks*, **7**(2), pp. 1–35. DOI: 10.3390/risks7020035.

Moivre, A. de (1725). *Annuities upon lives or, the valuation of annuities upon any number of lives; as also, of reversions. To which is added, an appendix concerning the expectations of life, and probabilities of survivorship.*

Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, **21**(99), pp. 1–108. URL: http://jmlr.org/papers/v21/17-123.html.

Morgenstern, H. (1995). Ecologic Studies in Epidemiology: Concepts, Principles, and Methods. *Annu. Rev. Public Health*, **16**(1), pp. 61–81. DOI: 10.1146/annurev.pu.16.050195.000425.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I.,

et al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, **396**(10258), pp. 1223–1249. DOI: 10.1016/S0140-6736(20)30752-2.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. (2017). "Tsallis Regularized Optimal Transport and Ecological Inference". In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. AAAI Press. DOI: 10.1609/aaai.v31i1.10854.

NCD-RisC (2017). Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. *The Lancet*, **389**(10064), pp. 37–55. DOI: 10.1016/s0140-6736(16)31919-5.

Nielsen, B. and Nielsen, J. (2014). Identification and Forecasting in Mortality Models. *The Scientific World Journal*, **2014**. DOI: 10.1155/2014/347043.

Njenga, C. N. and Sherris, M. (2011). Longevity Risk and the Econometric Analysis of Mortality Trends and Volatility. *Asia-Pacific Journal of Risk and Insurance*, **5**(2), pp. 1–39. DOI: 10.2202/2153-3792.1115.

OECD (2021). *Pensions at a Glance 2021: OECD and G20 Indicators*. Paris: OECD Publishing, pp. 1–224. DOI: 10.1787/a957e891-en.

Oeppen, J. and Vaupel, J. W. (2002). Broken limits to life expectancy. *Science*, **296**(5570), pp. 1029–1031. DOI: 10.1126/science.1069675.

Olivieri, A. (2006). Heterogeneity in Survival Models - Applications to Pensions and Life Annuities. *Belgian Actuarial Bulletin*, **6**(1), pp. 23–39. DOI: 10.2139/ssrn.913770.

Olshansky, S. J., Carnes, B. A., and Mandell, M. S. (2009). Future trends in human longevity: Implications for investments, pensions and the global economy. *Pensions: An International Journal*, **14**(3), pp. 149–163. DOI: 10.1057/pm.2009.12.

Otsuka, J. and Saigo, H. (2022). "On the Equivalence of Causal Models: A Category-Theoretic Approach". In: *Proceedings of the 1st Conference on Causal Learning and Reasoning*. PMLR. URL: https://proceedings.mlr.press/v177/otsuka22a.html.

Palloni, A. and Beltrán-Sánchez, H. (2017). Discrete Barker Frailty and warped mortality dynamics at older ages. *Demography*, **54**(2), pp. 655–671. DOI: 10.1007/s13524-017-0548-4.

Pampel, F. C. (2003). Declining sex differences in mortality from lung cancer in high-income nations. *Demography*, **40**(1), pp. 45–65. DOI: 10.1353/dem.2003.0007.

Pampel, F. C. and Zimmer, C. (1989). Female labour force activity and the sex differential in mortality: Comparisons across developed nations, 1950–1980. *European Journal of Population/Revue européenne de Démographie*, **5**(3), pp. 281–304. DOI: 10.1007/BF01796820.

Pearl, J. (2001). "Direct and Indirect Effects". In: *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 411–420. ISBN: 1558608001. DOI: 10.5555/2074022.2074073.

Pearl, J. (2009). *Causality*. Cambridge, UK: Cambridge University Press. ISBN: 978-0-521-89560-6. DOI: 10.1017/CBO9780511803161.

Perks, W. (1932). On Some Experiments in the Graduation of Mortality Statistics. *Journal of the Institute of Actuaries*, **63**(1), pp. 12–57. DOI: 10.1017/S0020268100046680.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **78**(5), pp. 947–1012.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: The MIT Press. ISBN: 978-0-262-03731-0.

Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling Probability Density Functions as Data Objects. *Econometrics and Statistics*, **21**, pp. 159–178. DOI: 10.1016/j.ecosta.2021.04.004.

Phillips, P. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, **75**(2), pp. 335–346. DOI: 10.1093/biomet/75.2.335.

Pitacco, E., Denuit, M., Haberman, S., and Olivieri, A. (2009). *Modelling longevity dynamics for pensions and annuity business*. London: Oxford University Press. ISBN: 978-0-19-954727-2.

Plat, R. (2009). Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, **45**(1), pp. 123–132. DOI: 10.1016/j.insmatheco.2009.05.002.

Pollard, J. H. (1982). The expectation of life and its relationship to mortality. *Journal of the Institute of Actuaries*, **109**(2), pp. 225–240. DOI: 10.1017/S0020268100036258.

Preston, S., Stokes, A., Mehta, N., and Cao, B. (2014). Projecting the Effect of Changes in Smoking and Obesity on Future Life Expectancy in the United States. *Demography*, **51**(1), pp. 27–49. DOI: 10.1007/s13524-013-0246-9.

Preston, S. H. and Wang, H. (2006). Sex mortality differences in the United States: The role of cohort smoking patterns. *Demography*, **43**(4), pp. 631–646. DOI: 10.1353/dem.2006.0037.

Renshaw, A. E. and Haberman, S. (2006). A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance: Mathematics and Economics*, **38**(3), pp. 556–570. DOI: 10.1016/j.insmatheco.2005.12.001.

Retherford, R. D. (1972). Tobacco smoking and the sex mortality differential. *Demography*, **9**(2), pp. 203–215. DOI: 10.2307/2060633.

Riley, J. C. (2001). *Rising Life Expectancy: A Global History*. Cambridge: Cambridge University Press. ISBN: 0521802458.

Rischel, E. F. and Weichwald, S. (2021). "Compositional Abstraction Error and a Category of Causal Models". In: *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Robins, J. M. (1998). Structural Nested Failure Time Models. *Encyclopedia of Biostatistics*, **6**, pp. 4372–4389. DOI: 10.1002/9781118445112.stat06059.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pp. 143–155. DOI: 10.1097/00001648-199203000-00013.

Robins, J. M., Greenland, S., and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, **94**(447), pp. 687–700. DOI: 10.1080/01621459.1999.10474168.

Rønn-Nielsen, A. and Hansen, E. (2014). *Conditioning and Markov properties*. Department of Mathematical Sciences.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). "Causal Consistency of Structural Equation Models". In: *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. UAI Press. URL: http://auai.org/uai2017/proceedings/papers/11.pdf.

Rudin, W. (1976). *Principles of mathematical analysis*. Vol. 3. New York: McGraw-Hill.

Said, S. and Dickey, D. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, **71**(3), pp. 599–607. DOI: 10.2139/ssrn.2882101.

Salhi, Y. and Loisel, S. (2017). Basis risk modelling: A cointegration-based approach. *Statistics*, **51**(1), pp. 205–221. DOI: 10.1080/02331888.2016.1259806.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). "On causal and anticausal learning". In: *Proceedings of the 29th Conference on Machine Learning*. Omnipress.

Shang, H. L. and Hyndman, R. J. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, **26**(2), pp. 330–343. DOI: 10.1080/10618600.2016.1237877.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, **25**(3), pp. 289–310. DOI: 10.1214/10-STS330.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. Cambridge, MA: The MIT Press. ISBN: 0-262-19440-6.

Spreeuw, J., Nielsen, J., and Jarner, S. F. (2013). A nonparametric visual test of mixed hazard models. *Statistics and Operations Research Transactions*, **37**(2), pp. 153–174.

Thatcher, A. R. (1999). The long-term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society: Series A, (Statistics in Society)*, **162**(1), pp. 5–43. DOI: 10.1111/1467-985x.00119.

Thatcher, A. R., Kannisto, V., and Vaupel, J. W. (1998). *The force of mortality at ages 80 to 120*. Odense, Denmark: Odense University Press.

Tillman, R. and Spirtes, P. (2011). "Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables". In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. PMLR. URL: https://proceedings.mlr.press/v15/tillman11a.html.

Triantafillou, S., Tsamardinos, I., and Tollis, I. (2010). "Learning Causal Structure from Overlapping Variable Sets". In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. PMLR. URL: https://proceedings.mlr.press/v9/triantafillou10a.html.

Trovato, F. (2005). Narrowing sex differential in life expectancy in Canada and Austria: Comparative analysis. *Vienna Yearbook of Population Research*, **1**, pp. 17–52. DOI: 10.1553/populationyearbook2005s17.

Trovato, F. and Lalu, N. (2007). From divergence to convergence: The sex differential in life expectancy in Canada, 1971–2000. *Canadian Review of Sociology/Revue canadienne de sociologie*, **44**(1), pp. 101–122. DOI: 10.1111/j.1755-618X.2007.tb01149.x.

Trovato, F. and Lalu, N. (1996). Narrowing sex differentials in life expectancy in the industrialized world: Early 1970's to early 1990's. *Social biology*, **43**(1-2), pp. 20–37. DOI: 10.1080/19485565.1996.9988911.

Tsamardinos, I., Triantafillou, S., and Lagani, V. (2012). Towards Integrative Causal Analysis of Heterogeneous Data Sets and Studies. *The Journal of Machine Learning Research*, **13**(1), pp. 1097–1157.

Tuljapurkar, S., Li, N., and Boe, C. (2000). A universal pattern of mortality decline in the G7 countries. *Nature*, **405**(6788), pp. 789–792. DOI: 10.1038/35015561.

Vaupel, J. W. (1986). How change in age-specific mortality affects life expectancy. *Population studies*, **40**(1), pp. 147–157. DOI: 10.1080/0032472031000141896.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**(3), pp. 439–454. DOI: 10.2307/2061224.

Vaupel, J. W. and Romo, V. C. (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography*, **40**(2), pp. 201–216. DOI: 10.1353/dem.2003.0018.

Vaupel, J. W., Villavicencio, F., and Bergeron-Boucher, M.-P. (2021). Demographic perspectives on the rise of longevity. *Proceedings of the National Academy of Sciences*, **118**(9), e2019536118. DOI: 10.1073/pnas.2019536118.

Villegas, A. M. and Haberman, S. (2014). On the Modeling and Forecasting of Socioeconomic Mortality Differentials: An Application to Deprivation and Mortality in England. *North American Actuarial Journal*, **18**(1), pp. 168–193. DOI: 10.1080/10920277.2013.866034.

Villegas, A. M., Haberman, S., Kaishev, V. K., and Millossovich, P. (2017). A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin: The Journal of the IAA*, **47**(3), pp. 631–679. DOI: 10.1017/asb.2017.18.

Vékás, P. (2019). Rotation of the age pattern of mortality improvements in the European Union. *Central European Journal of Operations Research*, **28**(1), pp. 1–18. DOI: 10.1007/s10100-019-00617-0.

Wakefield, J. (2004). Ecological inference for 2 x 2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **167**(3), pp. 385–425. DOI: 10.1111/j.1467-985x.2004.02046_1.x.

Wakefield, J. (2008). Ecologic studies revisited. *Annu. Rev. Public Health*, **29**(1), pp. 75–90.

Waldron, I. (1983). Sex differences in human mortality: The role of genetic factors. *Social Science & Medicine*, **17**(6), pp. 321–333. DOI: 10.1016/0277-9536(83)90234-4.

Wan, C. and Bertschi, L. (2015). Swiss coherent mortality model as a basis for developing longevity de-risking solutions for Swiss pension funds: A practical approach. *Insurance: Mathematics and Economics*, **63**, pp. 66–75. DOI: 10.1016/J.INSMATHECO.2015.03.025.

Wang, H. and Preston, S. (2009). Forecasting United States mortality using cohort smoking histories. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, pp. 393–398. DOI: 10.1073/pnas.0811809106.

Wang, S. and Brown, R. (1998). A Frailty Model for Projection of Human Mortality Improvements. *Journal of Actuarial Practice*, **6**, pp. 221–241.

WHO (2009). Global Health Risks: Mortality and burden of disease attributable to selected major risks. *World Health Organization*, pp. 1–70. URL: https://apps.who.int/iris/handle/10665/44203.

Wienke, A. (2010). *Frailty models in Survival Analysis*. Chapman & Hall/CRC biostatistics series. Boca Raton: Taylor & Francis. ISBN: 0-429-13960-8.

Yang, S. S. and Wang, C.-W. (2013). Pricing and securitization of multi-country longevity risk with mortality dependence. *Insurance: Mathematics and Economics*, **52**(2), pp. 157–169. DOI: 10.1016/j.insmatheco.2012.10.004.

Zarulli, V., Jones, J. A. B., Oksuzyan, A., Lindahl-Jacobsen, R., Christensen, K., and Vaupel, J. W. (2018). Women live longer than men even during severe famines and epidemics. *Proceedings of the National Academy of Sciences*, **115**(4), E832–E840. DOI: 10.1073/pnas.1701535115.

Zeger, S. L. and Liang, K.-Y. (1991). Feedback models for discrete and continuous time series. *Statistica Sinica*, **1**, pp. 51–64.

Zhou, R., Wang, Y., Kaufhold, K., Li, J. S.-H., and Tan, K. S. (2014). Modeling period effects in multi-population mortality models: Applications to Solvency II. *North American Actuarial Journal*, **18**(1), pp. 150–167. DOI: 10.1080/10920277.2013.872553.