Universidad Carlos III de Madrid
Doctorado en Ingeniería Matemática
Área de Ciencias y Técnicas Estadísticas

Ph.D. Thesis

Smoothing methods for the analysis of mortality development

Author: Carlo G. Camarda[†⋆]
Advisors: María L. Durbán Reguera[⋆]
Jutta Gampe[†]

[†]Max Planck Institute for Demographic Research, Rostock, Germany

[⋆]Departamento de Estadística, Universidad Carlos III de Madrid, Spain

Acknowledgments

I would like to acknowledge my debt to my colleagues and the staff of the Max Planck Institute for Demographic Research, for the friendly climate in which this dissertation was prepared. Among many excellent colleagues, I am immensely indebted to my supervisor Jutta Gampe. Without her fundamental assistance, her valuable suggestions, and her peculiar way of encouraging, this work could probably never have come into being.

I am particularly grateful to María Durbán from the Departamento de Estadística in the Universidad Carlos III de Madrid. This work benefited greatly from her useful suggestions and her positive and supportive attitude. Moreover, if this dissertation is part of the Doctorado en Ingeniería Matemática at the Carlos III it is primarily due to her help and aid.

Furthermore, I wish to express my gratitude to Paul Eilers: he helped me to develop new ideas for this dissertation and his suggestions improved enormously the computer programming of my work.

I would like to take this opportunity to extend my gratitude to James W. Vaupel for supporting my doctoral studies in the past years. Thanks to him, I could pursue my work in an excellent and pleasant environment such as the Max Planck Institute for Demographic Research. In this institute, I had the privilege to collaborate with numerous and helpful researchers. In particular, I would like to thank Jim Oeppen, Francesco Lagona, Elisabetta Barbi and Roland Rau for their interesting feedback and valuable suggestions.

I wish to extend my acknowledgement to all the members of the Laboratories of Statistical Demography and Survival and Longevity at the Max Planck Institute for Demographic Research, and to the researchers I met at the Departamento de Estadística in the Universidad Carlos III. Moreover, my dissertation benefited from the courses I attended at the International Max Planck Research School for Demography and at the Máster en Ingeniería Matemática in Madrid.

I would like to thank all the (former) Ph.D. fellows who shared this tough, but marvellous period of my life: I have been really fortunate to come across them. My thanks go especially to Dora Kostova, Dae-Jin Lee, Madgalena Muszynska, Anna Oksuzyan, Markéta Pechholdová, Sabine Schnabel, Kamil Sienkiewicz, Tiziana Torri, Harald Wilkoszewski and Sabine Zinn. I would like to share my pride with all of them.

Most important, I am extremely grateful to my mother and father. Their unreserved love and support for these years far from their home is what makes this dissertation valuable. I also wish to express my deepest appreciation to my brother Francesco, to my sisters, Lucia Chiara, Pamela, Luana and Giulia, and to *il piccolo* Antonio. Despite the actual distance, I have always felt them next to me. Moreover, I cannot forget to thank all my friends and relatives in Italy.

Finally, I would like to gratefully thank Tiziana for her forbearance and unselfish backup, and for being always close to me whilst I have spent the last years on this dissertation.

Resumen

La mortalidad, entendida como el riesgo de muerte, cambia con la edad, y además presenta cambios sistemticos con el tiempo, al menos durante los últimos 150 años. Comprender la dinámica de la mortalidad con respecto a la edad y al tiempo es un aspecto esencial de la demografía, ya que estos factores son las fuerzas que rigen los cambios en las problaciones. El continuo descenso de la mortalidad, y por lo tanto, el aumento de la longevidad, tiene importantes consecuencias, tanto para el individuo, como para la sociedad en su conjunto.

En el primer capítulo de esta tesis, se hace una revisión de los modelos clásicos que han venido siendo utilizados con el objetivo de capturar los cambios en mortalidad. Estos modelos abarcan desde las distribuciones paramétricas clásicas de Gomperz y Makeman, que sólo estudian los cambios en mortalidad de edades adultas, hasta los modelos de edad-periodo-cohorte, que sufren de problemas de identificabilidad. Como alternativa, el modelo bilineal introducido por Lee y Carter es considerado como el modelo estándar con el que nuevos modelos han de ser comparados.

El punto de partida de esta tesis son lo métodos de suavizado bidimensionales para datos de conteo que siguen una distrbución de Poisson, en concreto, los splines con penalizaciones o P-splines que se presentan en el segundo capítulo. En el caso unidimensional, este enfoque combina un número apropiado de bases de B-splines con una penalización sobre los coeficientes. Por un lado, los B-splines proporcionan la suficiente flexibilidad para capturar las tendencias presentes en los datos; y por otro, la penalización, aplicada sobre los coeficientes vecinos, aseguran la suavidad y reducen el número de parámetros, además de evitar los problemas de selección de número de nodos y el uso del método de *backfitting*. Los P-splines pueden entenderse como una genaralización de los modelos de regresión, en la que los B-splines actúan como regresores. El método de mínimos cuadrados (en el caso de datos normales), o el IRLS (*iteratively reweighted least-squares* para el caso generalizado) han sido modificados e incluyen la penalización controlada por un parámetro, el parámetro de suavizado. La penalización utilizada, está basada en una matriz de diferencias de order d (en general, $d = 2$), y fijado el parámetro de suavizado, los parámetros de regresión se estiman de modo sencillo, de forma simular al modelo clásico de regresión. En este mismo capítulo se muestra el cálculo de los errores estándar y los residuos asociados modelos de P-splines y se hace una revisión de los residuos más utilizados en el caso de datos de Poisson. Se propone el uso de los mapas de contorno de los residuos con respecto a la edad y año de muerte para localizar las zonas en las que los modelos utilizados no son capaces de capturar las tendencias, y poder así detectar aspectos demográficos interesantes. Mediante el uso de estas técnicas se ha demostrado que los P-splines capturan las tendencias de mortalidad de forma más propiada que los modelos de Lee-Carter, a pesar de que el número de parámetros utilizados en los modelos de P-splines es

muy inferior al utilizado por estos últimos.

El hecho de que el tamaño de las muestras con las que se trabaja sea grande, afecta de forma significativa a la inferencia, los intervalos de confianza son muy estrechos, y las medidas de bondad de ajuste usuales no aportan ninguna información, y por lo tanto, no son capaces de discriminar entre modelos de distinta complejidad. En el tercer capítulo de la tesis se proponen medias alternativas de bondad de ajuste. Primero se adaptan las medidas existentes, como el R^2 en el caso Normal, al caso de datos provenientes de familias exponenciales. La reducción proporcional de la incertidumbre debida a la inclusión de nuevos regresores en el modelo está basada en la divergencia de Kullback-Leibler. Además, se proponen medidas del tipo R^2 en el contexto de los P-splines, en concreto, se utiliza la relación entre el número de parámetros de un modelo y su dimensión efectiva para derivar una medida R^2 para modelos de suavizado. La idea básica ha sido considerar un modelo distinto bajo la hipótesis nula, que sea más apropiado para el caso de datos de mortalidad. Este modelo, es lineal o bilineal para el caso de datos unidimensionales y bidimensionales respectivamente. Se ha demostrado que el modelo bilineal está anidado en un modelo de P-splines, así como en un modelo de Lee-Carter, esta demostración está basada en la representación de los P-splines como modelos mixtos, y en el hecho de que la parte fija del modelo, corresponde con un modelos lineal o bilineal. Además se ha estudiado la relación entre esta nueva medida de bondad de ajuste y los métodos anteriormente mencionados (AIC, BIC), probándose que es muy similar al AIC. El comportamiento de esta medida ha sido evaluado mediante un ejercicio de simulación y con el anásis de datos procedentes del *Human Mortality Database* (HMD), en ambos casos, los modelos de P-splines dieron un mejor ajuste de los datos que los modelos de Lee-Carter.

En el cuarto capítulo se aborda un problema recurrente cuando se trabaja con datos históricos de mortalidad, o con países donde se recogen pocos datos, es la preferencia por dígitos, es decir, la tendencia a redondear números en torno a ciertos dígitos, en particular, en la distribución de muertes por edad aparecen picos en números que terminan en 0 (a veces en 5). Para solucionar este problema se ha propuesto un modelo que combina los conceptos de verosimilitud penalizada con el de modelos con función enlace compuesta: *composite link models*. Estos modelos permiten describir el modo en que la distribución latente de muertes por edad se mezcla con la preferencia de dígitos, mediante la redistribución de ciertos datos en torno a las edades preferidas, de modo que la distribución que se obtiene es preciamente la observada. La única restricción impuesta a la distribución latente es que sea suave, y se impone mediante una penalización similar a la utilizada en el caso de los P-splines. La estimación del modelo se ha llevade a cabo mediante una generalización del algoritmo IRLS, que incluye la matriz en la que se representan las probabilidades de redistribución. Estos modelos se han generalizado al caso en el que la preferencia puede aparecer entre dígitos que son vecinos, de modo que la tendencia a redondear no tiene por qué ser la misma para dígitos que terminan en un mismo número, sino que puede variar con la edad, como ocurre frecuentemente en datos demográficos. Las aplicaciones con datos simulados y datos reales han demostrado que este nuevo enfoque proporciona resultados excepcionales (Camarda, Eilers y Gampe (2008b)).

La reducción de la mortalidad a lo largo del tiempo puede considerarse como ganancia en esperanza de vida. Las muertes que ocurrían hace tiempo a edades tempranas, ocurren ahora

mucho más tarde. Esta manera de describir la mejora en mortalidad se ocupa de la distribución de la edad de muerte (la densidad) en vez del riesgo. El capítulo quinto de esta tesis está dedicado al desarrollo de métodos que permitar encontrar una transformación del eje de la edad para transformar una distribución de muerte por edad en otra. Nuevamente, estos métodos se han basado en la hipótesis de suavidad de esta transformación. Se ha considerado una transformación no-lineal mediante un modelo que utiliza la idea de suavizado y deformación del eje de la edad, esto modelos han sido llamados: *Warped Failure Time model* (WaFT). La metodología propuesta se basa en la elección de una distribución objetivo que se supone fija, y se busca una transformación tal que, una vez transformado el eje de la edad, la densidad de la distribuci ón observada se corresponde con la distribución objetivo. Se ha demostrado que el uso de los P-splines para representar la transformación permite controlar la suavidad de la misma de forma satisfactoria. Esta metodología ha sido extendida al caso en el que la distribución objetivo es desconocida, siendo estimada también mediante métodos de regresión no paramétrica. Los estudios de simulación han probado que los modelos WaFT pueden capturar transformaciones no-lineales, y el análisis de datos reales ha puesto de manifiesto que este tipo de modelos son necesarios, ya que una simple transformación lineal no es satisfactoria.

En resumen, esta tesis ha demostrado la utilidad de los métodos de suavizado, en particular de los P-splines, para el análisis de varios aspectos de relacionados con la mortalidad. Se ha propuesto una nueva medida de la variabilidad explicada para comparar distintos modelos en el caso de superficies de mortalidad, y se han desarrollado dos nuevos modelos: uno cuyo objetivo es salvar los problemas de preferencia de dígitos que pueden aparecer cuando se cuantifica el número de muertes a una cierta edad; y otro que ofrece un modo alternativo de explorar los cambios en la mortalidad centrándose en la ganancia (o pérdida) en esperanza de vida, como altrenativa al estudio del riesgo. Ambos modelos pueden ser utilizados de forma inmediata en otros contextos.

Preface

Populations change through three processes: mortality, fertility, and migration. Changes in mortality contribute considerably to population dynamics and variation in the levels of mortality lead to changes in the age distribution of the population. This has repercussions on almost all areas of a society, including its health-care system, health and life insurance, as well as pension schemes. The consequences of such transformations are also experienced on the more individual level such as changing kinship sizes, marriage squeezes, the value of children, genetic disease, family living arrangements and women's status.

Demographic research investigates levels and trends of mortality, fertility and migration processes and develops numerous techniques to measure and analyze them (Keyfitz and Caswell, 2005). While medical and epidemiological research usually deals with samples of moderate sizes, including quite detailed information on the individual level, demographic studies often use data on whole populations, or large subgroups within populations, with only a few, if any, additional covariates available. Hence demographic mortality studies are often performed on an aggregate level of analysis.

During the last decades, statistical perspective on demographic and mortality developments has received increased attention. This interest has lead to statistical techniques for modeling the data generation process that gave rise to demographic observations. Along this line of research, this dissertation attempts to further bridge the gap between demography and statistics, proposing novel statistical methods for investigating mortality processes on an aggregate level. The focus is on smoothing methods, in particular with regard to appropriate measures of fit for large samples, models based on transforming age-at-death distributions, and modeling digit preferences via smooth latent distributions.

The first chapter reviews traditional and well-established models in mortality analysis. First, source and structure of the mortality data used in this dissertation are introduced. The Lexis diagram is presented as a standard tool for summarizing demographic data. The fundamental Poisson assumption for the total number of deaths over a specified age- and year-interval will be introduced. Over the last two centuries, researchers aimed at reducing the dimensionality of the data to a smaller number of parameters by directly modeling some of the systematic patterns demographers have uncovered. Simple models for portraying mortality over age, and more sophisticated approaches for modeling mortality over both age and times will be reviewed in detail toward the end of the chapter.

Overparameterization is a typical feature in recent demographic models. The use of such an amount of parameters may often seem unnecessary. Therefore, smoothing approaches are a natural

alternative to analyzing mortality over age and time. Chapter 2 introduces smoothing methods in a demographic context. Among different methods, the so-called P-splines are particularly suitable for two-dimensional regression contexts. Introduced by Eilers and Marx (1996), this approach is well-established as a means of smoothing Poisson data such as death counts. The chapter gives a detailed introduction in both one- and two-dimensional settings. Particular emphasis is given to residual analysis and measurement of the variability for P-splines in a demographic context.

Mortality data on an aggregate level are characterized by (very) large sample sizes. For this reason, uninformative outcomes are evident in common goodness-of-fit measures. Following a review of the common measures of goodness-of-fit, Chapter 3 proposes a new measure that allows comparison of different mortality models even for large sample sizes. Particularly, we will propose a new measure which uses a null model specifically designed for mortality data. Several simulation studies and actual applications will demonstrate the performances of this new measure with special emphasis on previously introduced demographic models and P-spline approach.

The mentioned Poisson assumption can be relatively strong in demographic data and, in peculiar situations, the presence of overdispersion cannot be neglected. Digit preference, a tendency to round counts to pleasant digits, is a typical source of overdispersion for mortality data. Chapter 4 presents a new approach for dealing with this issue. In the last part of the chapter, we will propose a generalization of the original model which allows more general patterns of misreporting. Simulation studies and actual applications will be used to test both the original and the extended version of the model.

In Chapter 5, we consider a new approach to analyzing mortality data in a different way. This model operates directly on the probability density function of the life-times instead of the more common consideration of the hazard function. It can be considered an extension of the accelerated failure time model for comparison of density functions. With this model, one can study how the time-axis would have to be transformed so that one age-at death distribution conforms to another. Smoothing methodologies are employed for describing and estimating the transformation function. Simulated and actual examples illustrate the performances of this model, which allows alternative interpretations of observed mortality development over time.

A brief critical discussion of the various methods and models proposed in the dissertation is given in the final Chapter 6.

Contents

List of Figures

Chapter 1

Mortality data and models

Data collections for analyzing mortality have a longer history and are more developed than those for analyzing other demographic questions such as fertility and migration. On the other hand, relatively simple mathematical methods have traditionally been used to assess mortality trends. Classical demographic methods are inclined to stay very close to the data, which permit scholars to gain a detailed understanding of the data's strengths, weaknesses, and features. Moreover, demographers often treat data as fixed implicitly rather than as a realization of a stochastic process and methods have typically been based on the measurements of demographic rates by age and sex. Summary measures, such as life expectancy, can then be computed and evaluated. Classic methods from matrix algebra and differential and integral equations are also used to explain implications of the current mortality conditions into the future (Caswell, 2001; Keyfitz and Caswell, 2005). This chapter is primarily concerned with the basic data, measures and models used in demographic analysis of mortality development.

Specifically, Section 1.1 presents source and structure of the mortality data used in this study. Assumptions required for further modeling are in Section 1.2. Particular emphasis is given on the Poisson approximation of the death counts. Mortality data are largely informative when they are properly displayed and Section 1.3 presents the notion of mortality surface as a suitable tool for portraying mortality data. Demographic models for describing mortality over ages are illustrated in Section 1.4. Then Section 1.5 is devoted to approaches for modeling mortality over age and time with emphasis to the standard Lee-Carter model (Lee and Carter, 1992). Concerns about the over-parameterization of the common demographic models are considered in Section 1.6, leading to the non-parametric approach, which is more fully discussed in Chapter 2.

1.1 Data: sources and structure

To perform mortality research one needs access to accurate and reliable data that cover a long enough period so that trends in mortality can be identified and analyzed further. For comparative studies, this information has to be available for several countries or sub-groups. Human mortality can be defined also as risk of death and it changes with age. Moreover, mortality differs between males and females. The minimally required set of information to analyze mortality trends on a national level would be the number of individuals alive and deceased at all ages over a range of

years. Data separated for females and males would allow analysis by sex. Such data are nowadays routinely collected in most developed countries. Censuses, official vital statistics or population registers are the sources of such population data. However, continuous collection and updating of data from the official vital statistics for several countries may be very time-consuming and costly.

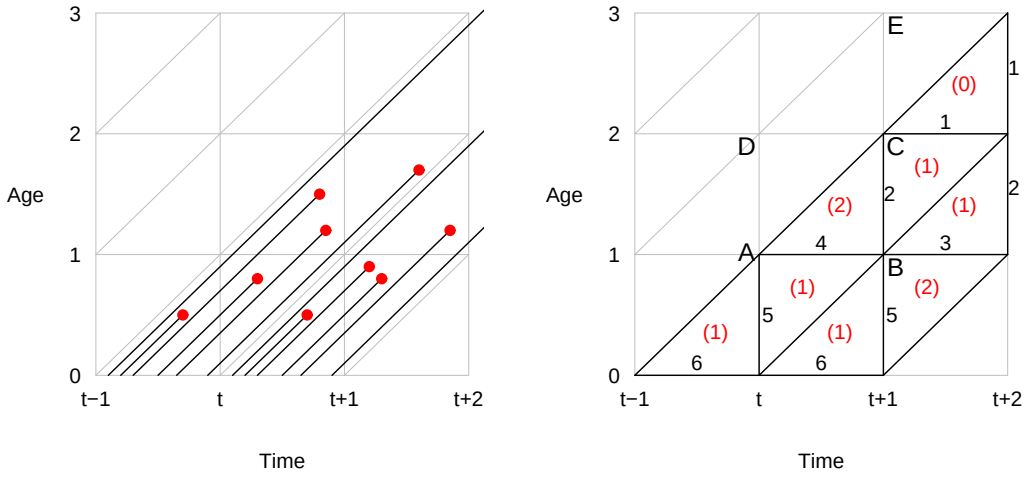

Figure 1.1: Schematic Lexis diagrams. Left panel: Lexis diagram containing life-times for birth cohorts of $t-1$ and t. Each individual is presented as a line in a time-age plane and red points depict the death for a given individual. Right panel: Lexis diagram containing counts of events pertaining to birth cohorts of $t-1$ and t.

To facilitate the investigation of human mortality, an international project, the Human Mortality Database (HMD), was recently initiated by the Department of Demography at the University of California Berkeley, USA, and the Max Planck Institute for Demographic Research, Rostock, Germany. This project provides detailed mortality and population data which can be accessed online and may be used freely for research purposes (Human Mortality Database, 2008).

Currently the HMD provides information on 34 countries[1]. For each country (or region), the HMD offers basic quantities in mortality studies, namely: the deceased and survivors by sex, age, year of death, and birth cohort. Though the age range covered is the same in all countries (from age 0 to 110+), the range of years covered differs from country to country. The longest series is provided for Sweden (1751–2006), whereas other countries have data from the nineteenth century (the other Scandinavian countries, Belgium, England, France, Italy, Netherlands, New Zealand and Switzerland). For some European countries, Japan, Australia, Canada, Taiwan and the USA, the series of data first start in the twentieth century (Human Mortality Database, 2008).

A standard tool for summarizing such data is the Lexis diagram (Lexis, 1875). In this diagram an individual life history is drawn as a line segment with slope 1. This line starts on the horizontal axis at the time of birth and ends at time of death. The value on the vertical axis is the individual's age. Hence a life-line starts at zero (birth) and ends at the age at death. In this way data are properly represented according to the three demographic coordinates: the time of death (period),

[1]For some countries, information is only available for some region, e.g. England and Wales.

the age at death, and the time of birth (cohort) of the deceased. Figure 1.1, left panel, shows a simplified example of the Lexis diagram. The individual life-lines can be grouped and hence the Lexis diagram also allows a systematic summary of aggregated death and population data by age, period and cohort. For instance, in Figure 1.1 (right panel), from the birth cohort of six births during period t: (1) death in t and five survivors to the beginning of the following period $t+1$; (2) deaths at age 0 in $t+1$ and three survivors to age 1; (1) death to the cohort at age 1 during $t+1$ and two survivors to the beginning of the period $t+2$.

1.2 Measures of mortality

Studying mortality data can be easily viewed as analysis of time to event data. The variable of interest is the life-span of an individual. We define X as the nonnegative and continuous random variable describing time from birth of an individual until death. Three functions characterize and describe the distribution of X: the probability density function, the survival function, and the hazard rate. If one of these functions is known, the other two can be uniquely determined.

The basic quantity to describe time-to-death distribution is the survival function, which is defined as the probability of an individual surviving beyond age x: $S(x) = Pr(X > x)$. The survival function is the complement of the cumulative distribution function, that is, $S(x) = 1 - F(x)$. Moreover, the survival function is the integral of the probability density function, $f(x)$, from x to infinity:

$$S(x) = Pr(X > x) = \int_x^\infty f(t)\, dt$$

Thus, $f(x) = -S'(x)$. Another fundamental quantity is the hazard function, also known as force of mortality in demography. It describes the instantaneous rate of death at age x, given survival until x. In formula:

$$h(x) = \lim_{\Delta x \to 0} \frac{Pr(x < X \leq x + \Delta x | X > x)}{\Delta x} = \frac{f(x)}{S(x)} = -\frac{d \ln S(x)}{d\, x}$$

An identity that relates survival function and hazard is given by

$$S(x) = \exp\left\{ -\int_0^x h(u)\, du \right\} = \exp\left\{ -H(x) \right\}.$$

The function $H(x) = \int_0^x h(u)\, du$ is called cumulative hazard function.

In practice, we may not observe the individual's full life-times. This can be the case if the survival time exceeds a certain value due to termination of the study, i.e. mortality data are available until a certain year. Such pattern of observations is called right-censoring. On the other hand, some individuals are actually not included in the study since they die before the beginning of the study, which is called left truncation. Many other and more complicated forms of incomplete observations are possible as well. For more details, see Klein and Moeschberger (2003, ch. 3).

When using aggregate data as provided by the HMD, we do not have precise individual information, but data are grouped into intervals, usually of length one year (cf. Figure 1.1, right panel), which could be viewed as a data set with censored individual information only.

1.2.1 Empirical death rates

To analyze mortality data, we have to make assumptions about the distribution of the random variable X and how it varies across individuals. The simplest assumption is that all individuals in one birth-cohort live their lives according to the same life-span distribution, hence, ignoring other sources of heterogeneity.

The choice of the distribution is commonly based on the hazard rate. Demography usually enjoys a wealth of data, hence, parsimonity is not a virtue in this field.

As human mortality, when considered over the full age-range, has a complicated pattern, the traditional assumption is that the hazard is constant over each one-year age-interval. It is, however, different between ages. This "piece-wise constant" assumption is the basis for the calculation of empirical death rates.

Constant hazards correspond to the assumption that X, conditionally on survival until the beginning of an age-interval, follows an exponential distribution with parameter h. The life-spans X_i of n individuals are independent and identically distributed (i.i.d.), i.e., $X_i \sim \text{Exp}(h)$, $i = 1, \ldots, n$. The maximum likelihood estimate (MLE) of the (constant) parameter h is given by

$$\hat{h} = \frac{Y}{L + C}\,, \tag{1.1}$$

where Y denotes the number of deaths that were observed during the interval, L is the total amount of time lived by those whose deaths were observed, and C is the total amount of time lived by those who were censored, i.e. survived beyond the end of the interval considered (Alho and Spencer, 2005, ch. 4).

The numerator of equation (1.1) corresponds, precisely, to the data collected from the squares of the Lexis diagram (e.g. $ABCD$ in Figure 1.1). Individuals spend varying times in any given square based on the time of the year they were born. This leads to fixed right and left censoring. Therefore, the exponential model provides a full estimation theory square by square, if the hazard is assumed to be constant in each square.

The denominator in (1.1) usually cannot be recovered exactly from aggregate data. In large populations, the person years lived during a year are typically approximated by the average of the population sizes in the beginning and at the end of the year. Of specific interest is the population in age x during t. In Figure 1.1, let P_{AD} and P_{BC} be the number of life-lines crossing segments AD and BC, respectively. Let Y_{ABCD} denote the number of deaths in the square $ABCD$. Then equation (1.1) is approximately $Y_{ABCD}/\{(P_{AD} + P_{BC})/2\}$. In general at age i, during year j we define as death rate the following ratio

$$m_{ij} = \frac{Y_{ij}}{E_{ij}} \tag{1.2}$$

where E_{ij} are the number of person-years aged i years during the year j and Y_{ij} are the number of deaths that occurred during the year j and attained age i. That is, we can define the death rate as the MLE of the hazard rate if the true hazard is constant within the time interval. Obviously, this assumption does not hold when intervals over age and/or time are too long.

The assumption of a constant hazard over short time intervals implies that the total number

of deaths over a specified age- and year-interval, Y_{ij}, is a Poisson distribution with mean $m_{ij} \cdot E_{ij}$

$$Y_{ij} \sim \text{Poisson}(E_{ij} \cdot m_{ij}) \tag{1.3}$$

and thus the model based on this assumption is, therefore, often called Poisson (regression) models (Keiding, 1990).

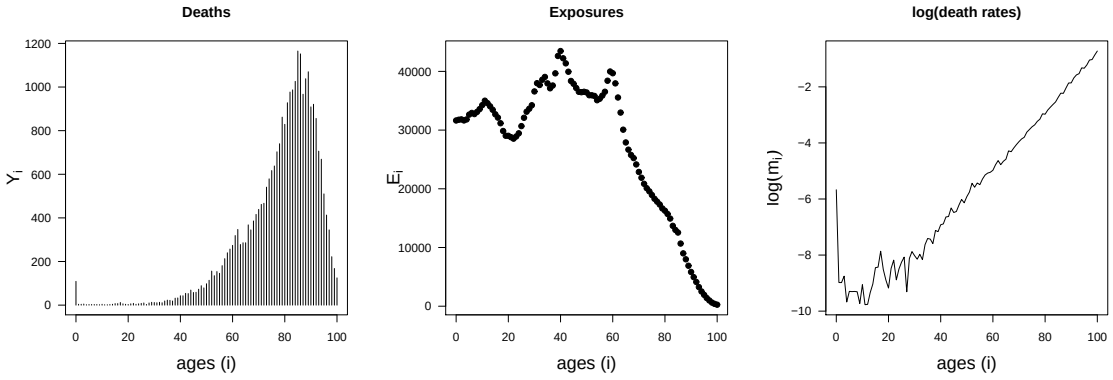

Figure 1.2: Deaths, exposures and death rates (logarithmic scale). Ages from 0 to 100. Denmark, females, 2006. Source: HMD.

1.3 Portraying mortality

Although density, survival function, and hazard all describe the same stochastic phenomenon, the force of mortality is more often used to portray mortality. The main reason is that the hazard more easily allows for capturing the change of the risk of death over age, due to its conditioning on the survivors to this particular age. It requires a lot of experience to read this information from the survival function. Therefore, empirical death rates over age and time are commonly used to describe mortality development over age and time.

Alternatively, pure death counts are also used, though in one calendar year they do not only reflect the effect of mortality, but also the size of the corresponding birth cohorts. Figure 1.2 presents death counts Y_{ij} and exposures E_{ij} , as well as the empirical death rates on logarithmic scale for the Danish females population for the year $j = 2006$ and for ages $i = 0, \ldots 100$. It is clear how both exposures and pure death counts are affected also by the previous cohort sizes.

When age-at-death distributions and estimates of the survival function in a calendar year are necessary for specific studies, the so-called period life-table approach is used for adjusting size of the birth cohorts (Keyfitz and Caswell, 2005). The hypothetical age-at-death distribution is calculated from the age-specific empirical death rates in one calendar year, which are derived from different birth cohorts. In this way, information on current mortality is summarized in a frequency distribution that would arise if a synthetic cohort was submitted to current death rates. A number of statistics can be derived, including the proportion of the synthetic cohort still alive and remaining life expectancy for people at different ages.

Furthermore, in what follows, empirical death rates over age and time can be used to portray

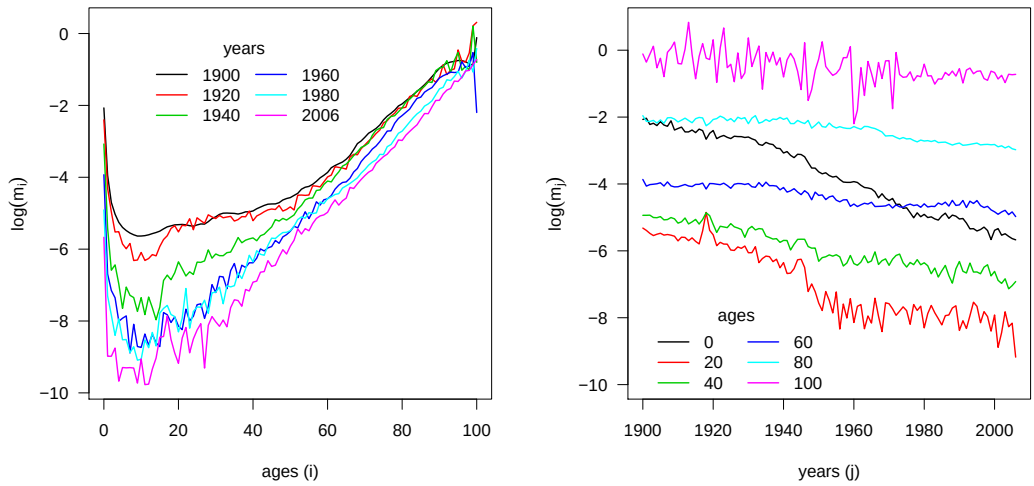

Figure 1.3: Death rates at selected years over ages (left panel) and selected ages over years (right panel), logarithmic scale. Denmark, females. Source: HMD.

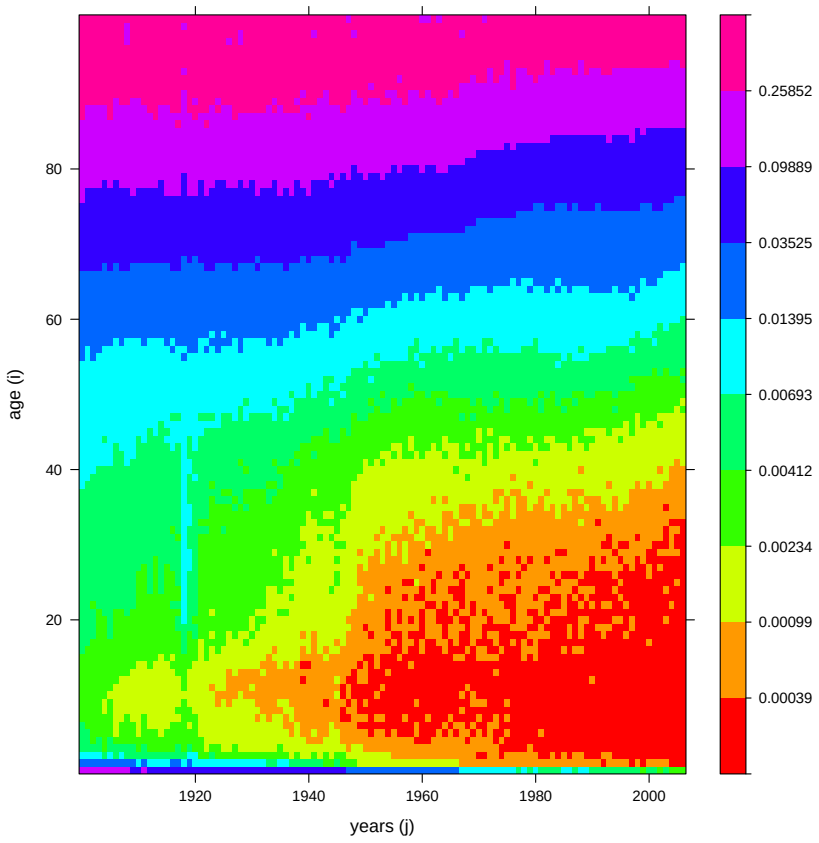

Figure 1.4: Death rates. Ages from 0 to 100. Denmark, females, 1900–2006. Source: HMD.

mortality change. In fixing a particular year, we can compute one death rate for each age. For a given age, there is one death rate for each year. For ease of manipulation and presentation, data are prepared as rectangular arrays. Given a population, for each calendar year and each age, we have the number of deaths and the number of exposures, as in equation (1.2). Deaths and exposures are arranged in $m \times n$ matrices $\boldsymbol{Y}$ and $\boldsymbol{E}$, respectively, whose rows are indexed by age and whose columns are indexed by year. Therefore, the matrix of empirical death rates is defined as $\boldsymbol{M} = \boldsymbol{Y}/\boldsymbol{E}$.

Population and mortality dynamics essentially develop over both period and ages. Arthur and Vaupel (1984) introduced the name "Lexis surface" to refer to the surface of demographic rates defined over age and time. In particular, in this thesis we work on death rates, hence we will refer to $\boldsymbol{M}$ as mortality surface (Vaupel et al., 1986).

Shaded contour maps permit visualization of mortality surfaces and offer a more comprehensive and more informative view than do graphs of death rates at selected ages over time or selected times over ages only. Figure 1.3 and 1.4 show the dynamics of the death rates, on logarithmic scale for the Danish female population from 1900 to 2006 and from age 0 to 100. The completeness of a shaded contour map for mortality surface in Figure 1.4 is evident in comparison to the information given by the unidimensional plots on Figure 1.3.

The basic information in the death rates and the related hazard function is, first of all, its qualitative behavior. As we see in Figure 1.3 (left panel) and 1.4, there is a general pattern in the human mortality over ages. During the infant ages, hazard functions are steeply decreased, dropping rapidly within the first years. A minimum is commonly reached at about ages 10–15. Afterward, especially for men, hazard rates show a hump at young-adult ages (usually called an accident-hump due to the main cause of death at those ages). Hazard rates then rise exponentially after approximately age 30 and a level off at ages above 80 (Preston, 1976; Thatcher et al., 1998; Vaupel, 1997). This type of hazard is similar to what in reliability engineering is often called "bath-tub shaped".

Qualitative considerations can be also given regarding the developments of mortality over time. For instance, in the last century, Figure 1.3 shows an overall decrease in mortality for the Danish female population, though of different pace for different ages.

Focusing on methodological aspects of the analysis of mortality developments, the choice of a certain population does not play a central role in this thesis. As mentioned, the Human Mortality Database (2008) offers numerous options and throughout this thesis we mainly deal with Danish females.

Data from the Human Mortality Database (2008) are sometimes estimates derived from aggregate data (e.g. five-year age groups, open age intervals such as 90+) and require various adjustments before being inserted into the database. Unlike other countries, Denmark presents outstanding data regarding quality: death counts and population have been accurately collected over single age and time intervals since 1916 and 1906, respectively. Prior corrections have thus not been carried out by the Human Mortality Database (2008) over these ages and years. Moreover, Denmark is a relatively small country[2] and therefore may show more variability in mortality

[2]The total population of Denmark in 2006 is 5,447,084 inhabitants, 2,750,422 females and 2,696,662 males.

due to smaller sample size.

Furthermore, in the following, we will largely focus on mortality changes over age 10 and after year 1930. Specifically, infant mortality presents features which would require particular methodology which exceed the scope of this thesis. In this direction, Krivobokova et al. (2006) suggested spatially-adaptive smoothing methodology and further investigation in this direction may shed some light in coping with such a steep decrease in infant mortality. The choice of 1930 aims at avoiding the presence of the abnormally high mortality during the Spanish Flu epidemic (Taubenberger and Morens, 2006). Moreover, the Danish population, especially women, was only partially affected by World War II. Again in a smoothing context, Kirkby and Currie (2007) have already proposed a model which successfully deals with period shocks such as wars, flu epidemics, hot summers or cold winters. These deviations from the smooth mortality surface can disproportionately affect the mortality of certain age groups in particular years. No attempt has been made here to further develop the analysis of such deviations.

Nevertheless, most of the presented methods which will be introduced in the following chapters can be extended to the Danish population as a whole, and for a longer series, as well as for other populations in the Human Mortality Database (2008). Interpretation of outcomes would require further care and considerations, which goes beyond our aims.

1.4 Mortality models over age

Although disaggregation over age and time of the mortality data provides a first impression of the phenomenon, a large set of numbers is cumbersome. Demographers have, thus, searched for more parsimonious representation of the variation of mortality over age and time.

A first systematic attempt in modeling hazard rates over age was done by Gompertz (1825). He observed that after a certain age, a "law of geometric progression pervades, in an approximate degree, large portions of different tables of mortality" (Gompertz, 1825, p. 514). Studying actuarial mortality tables, Gompertz discovered that in the age window of about 30–80 years, death rates increase exponentially with age (see Figure 1.3, left panel). Therefore, he suggested representing the hazard rates as

$$h(x) = a \cdot e^{b \cdot x}, \tag{1.4}$$

with parameters $a > 0$ and $b > 0$. The law has been applied in many countries during the last 180 years. It is a recurring pattern. Commonly, a represents the mortality at time zero (usually age 30) and b is the rate of increase of mortality and is frequently used as a measure of the rate of aging.

Using formulas in Section 1.2, we can derive the probability density function for the Gompertz distribution:

$$f(x) = ae^{bx} \exp\left[\frac{a}{b}(1 - e^{bx})\right]. \tag{1.5}$$

Makeham (1860) extended Gompertz' equation by adding a constant, an age-independent term, $c > 0$, to account for risks of death that do not depend on age:

$$h(x) = c + a \cdot e^{b \cdot x}. \tag{1.6}$$

Also, in this case the probability density function can be derived:

$$f(x) = ae^{bx} \exp\left[-c\,x + \frac{a}{b}(1 - e^{bx})\right].$$

Both Gompertz' and Makeham's model only intend to represent mortality at adult ages.

Successive attempts have tried to capture other three peculiarities of the human mortality over the age range, which can be seen in Figure 1.3 (left panel) and already mentioned in Section 1.3: a high value of the infant death rates (age 0), dropping rapidly within the first years, a hump at young-adult ages, and a leveling-off for ages above 80.

Logistic models have been proposed to portray this last feature in human mortality. Perks (1932) was the first to proposed a logistic modification of the Gompertz-Makeham models. A logistic function to model the late-life mortality deceleration can be given by

$$h(x) = c + \frac{a\,e^{bx}}{1 + \alpha\,e^{bx}}.$$

We can see that this includes Makeham's law as the special case when $\alpha = 0$. A similar logistic model has been proposed by Thatcher (1999).

Heligman and Pollard (1980) derived a descriptive model, covering the whole age range:

$$h(x) = A^{(x+B)^C} + De^{-E(\ln x - \ln F)^2} + \frac{GH^x}{1 + GH^x}$$

where $A, B, \ldots, H$ are the parameters in the model. It is easy to see that such parameterization can cause difficulties in the estimation procedure. Moreover, it would be hard to disentangle the physical meaning of each parameter.

A three-component, competing-risk mortality model, developed for animal survival data, has been proposed by Siler (1983). This model aims at portraying the whole of the age range with five parameters. On the other hand, Anson (1988) proposed a fifth degree polynomial to represent the hazard rate for humans. The Weibull (1951) model has been applied in a mortality context, though it was developed for the failure of technical systems due to wear and tear. A comprehensive review of the models for human population over ages has been provided by Gavrilov and Gavrilova (1991). For more information from an actuarial perspective see, Forfar et al. (1988).

1.5 Mortality models over age and over time

The models presented so far only capture the change of the hazard of death over age. To model how death rates change over time, possibly allowing different trends at different ages, several approaches have been used. We describe some of them in this section.

1.5.1 Relational models

Relational models use a tabulated "standard" mortality function and a mathematical rule for relating this standard to mortality in different populations, or within the same population at a different point in time. The standard mortality captures the complexity of age patterns of

mortality, while the model parameters describe deviations from the standard.

Using once again a parametric approach, this class of models finds a way to explore mortality in both age and time directions. Brass (1971), who was among the first to suggest this approach, used a logit transformation of the probability of surviving. Transformed, these probabilities, from both standard and actual population, are then related via a simple regression function. Specifically, let $S_1(x)$ and $S_2(x)$ be the estimates of survival function of two different populations. Let $Y_1(x)$ and $Y_2(x)$ their logit transformations:

$$Y(x) = \ln\left[\frac{1 - S(x)}{S(x)}\right]$$

then it is possible to find constants α and β such that:

$$Y_1(x) \approx \alpha + \beta\, Y_2(x) \tag{1.7}$$

Keeping the relation (1.7), Zaba (1979) and Ewbank et al. (1983) extended Brass' approach, adding two parameters to properly represent the shape of mortality in childhood and adulthood:

$$Y(x;\kappa,\lambda) = \begin{cases} \frac{\left[\frac{S(x)}{1-S(x)}\right]^{\kappa} - 1}{2\kappa} & S(x) \geq 0.5 \\ \\ \frac{1 - \left[\frac{S(x)}{1-S(x)}\right]^{\lambda}}{2\lambda} & S(x) < 0.5 \end{cases}$$

An alternative perspective for relational models has been proposed by Himes et al. (1994). Let $Y_j(x)$ be the logit transformation of death rates at age x in population j, it is possible to find the solution to the equation:

$$Y_j(x) = \delta + \sum_x \beta_x I_x + \sum_j \gamma_j J_j$$

where I_x is a dummy variable for age x; J_j is dummy variable for population j. δ, β_x and γ_j are parameters to be estimated. Also in this case, the model is in a parametric setting. Moreover, note that for comparing two populations and for 50 ages, the model estimates parameters for 50 dummy variables.

Though simple in practice, relational models present several drawbacks. There is no systematic way of choosing the standard mortality pattern and they are needed only for comparison purposes. Besides, simple parametric approaches often do not capture features in mortality changes when these are not represented in the chosen standard distribution.

1.5.2 APC models

Age-Period-Cohort (APC) models have been developed in order to separate the changes of incidence data with the three demographic coordinates – age, period and cohort (see Figure 1.1). Mathematically, it can be written as a model for log-rates in which the effects of age, period, and

cohort are combined additively:

$$\begin{aligned} \ln(m_{ij}) = \alpha_i + \beta_j + \gamma_c + \epsilon_{ij} \qquad i &= 1, \ldots, m \\ j &= 1, \ldots, n \\ c &= 1, \ldots, m+n-1 \end{aligned} \tag{1.8}$$

where α_i, β_j and γ_c are the age (i), period (j) and cohort (c) effects, respectively. However, there is a difficulty with the interpretation of the fitted parameters since, of the $2m+2n-1$ parameters in (1.8), only $2m+2n-4$ are identifiable.

This model may be fitted either by weighted least squares or by Poisson maximum likelihood (Clayton and Schifflers, 1987). The log-likelihood contribution from observation of the quantity (Y_{ij}, E_{ij}) is given by

$$l(m_{ij}|Y_{ij}, E_{ij}) = Y_{ij} \ln(m_{ij}) - m_{ij} E_{ij}.$$

The log-likelihood for the entire mortality surface is the sum of such terms, because cells are assumed to be independent. Hence, model (1.8) can be fitted using software for Poisson regression for independent observations, allowing for an offset term. A common way of accommodating the non-linearity of age, period and cohort effects is to use one parameter per distinct value of i, j and c, by defining the variables as factors. The classical approach (largely employed in epidemiology) has been to define a tabulation sufficiently coarse to avoid an excess amount of parameters in the modeling. Since the three variables age, period and cohort are originally continuous variables, it seems natural to model their effects by parametric smooth functions. Several approaches have been proposed in this regard (Carstensen and Keiding, 2005; Currie et al., 2007; Heuer, 1997; Ogata et al., 2000).

The major problem of APC models is the identification problem introduced by the fact that cohort, age and period are linearly related: $c + i = j$. Hence, there is no unique solution to the parameter estimation, as the model is non-identifiable. Some constraints are usually used to identify a unique solution and the choice of the constraints remains always arbitrary. Clayton and Schifflers (1987) gave a careful exposition of the modeling problems, warning about the dangers of over-interpreting the fitted parameters and about the functions of the rates that could be (meaningfully) estimated. A more recent account can be found in Carstensen (2007), who used APC models for the Lexis diagram or Schmid and Held (2007), who used a Bayesian approach. A summary of the advances of APC models can be found in Smith (2008).

1.5.3 Lee-Carter model

Lee and Carter (1992) reduced the complexity of APC models by introducing the following bilinear model for the log-death-rates:

$$\begin{aligned} \ln(m_{ij}) = \alpha_i + \beta_i \cdot \gamma_j + \epsilon_{ij} \qquad i &= 1, \ldots, m \\ j &= 1, \ldots, n \end{aligned} \tag{1.9}$$

where α_i, β_i and γ_j are vectors of parameters to be estimated, and ϵ_{ij} represents the error term.

The variance ϵ_{ij} in Lee and Carter (1992) is assumed to be constant for all i and j. This assumption is relaxed in some of its variants which are presented later.

The Lee-Carter (LC for short) model is under-determined and requires additional constraints on the parameters to be successfully estimated. Usually, the model is centered by choosing the parameters α_i as the average death rates over time for each age group i: $\alpha_i = \frac{1}{n}\sum_{j=1}^{n} \ln(m_{ij})$. Consequently, we interpret β_i as fixed age effects or deviations from a standard pattern α_i for each age. γ_j is a time-varying mortality level index. The LC model can also be seen as a relational model, where the standard pattern of mortality is the α_i.

Since its introduction in 1992, the Lee-Carter model has been widely used in diverse demographic applications and it can be considered the standard in modeling and forecasting death rates. Although the LC model was first intended to forecast all-cause mortality in the United States, it is now widely used by researchers for modeling (and forecasting) all-cause and cause-specific mortality in diverse fields. In particular, the LC model is also used for modeling and describing changes in mortality dynamics. See, for instance, the study on the seven most developed countries (G7) by Tuljapurkar et al. (2000).

It is not our purpose to study mortality forecasting here, but an advantage of using the LC model lies in the fact that its time-index γ_j can be easily forecasted since it is able to condense the linear mortality decline in the past century (see next Figure 1.5). In particular, forecasting in the LC model is performed in two stages. In the first stage, α_i, β_i and γ_j are estimated using the actual mortality surface. In the second stage, fitted values of γ_j are modeled and extrapolated by an autoregressive integrated moving average (ARIMA) process, determined by the Box and Jenkins (1970) approach. Finally the extrapolated γ_j are combined with the previous estimation to forecast the future death rates.

It is worth pointing out that for an $m \times n$ death rates matrix, the LC model in equation (1.9) estimates $2m + n - 2$ parameters. To solve the problem, researchers have proposed general alternative approaches, which are briefly presented here.

Non-likelihood-based methods

In non-likelihood-based methods, we are not required to specify any probability distribution during model parameter estimation. Examples include Singular Value Decomposition (SVD, Good, 1969) methods proposed in Lee and Carter (1992) and the method of weighted least squares (WLS), suggested later in Wilmoth (1993).

Using SVD, we set $\alpha_i = \frac{1}{n}\sum_{j=1}^{n} \ln(m_{ij})$ and then we compute the SVD to the matrix of $[ln(m_{ij}) - \alpha_i]$. The first left and right singular vectors give initial estimates of β_i and κ_j, respectively. To satisfy the constraints for model identification, the estimates of β_i and κ_j are normalized. In the original paper, Lee and Carter (1992) adopted the constraints $\sum_i \beta_i = 1$ and $\sum_j \gamma_j = 0$. To further improve the fit, researchers later considered using a rank-p SVD approximation. For example, Renshaw and Haberman (2003b) considered $p = 2$ and Booth et al. (2002) considered $p = 5$. Under the rank-p SVD approximation, equation (1.9) is generalized to

$$\ln(m_{ij}) = \alpha_i + \sum_{k=1}^{p} \beta_i^k \cdot \gamma_j^k + \epsilon_{ij}$$

Higher p will fit the mortality surface better, but this procedure will enormously increase the amount of parameters to estimate and interpret (and forecast).

As the method of the SVD is purely a mathematical approximation applied to the log-death-rates, the fitted and actual number of deaths may not be the same. To reconcile the fitted and the observed number of deaths, we are required to make an adjustment to γ_j. Lee and Carter (1992) propose computing a new estimate of γ_j, for each year j, by searching for the value that makes the observed number of deaths equal to the predicted number of deaths. Other criteria have been proposed (see for example, Booth et al. (2002) and Lee and Miller (2001)).

In the weighted least squares (WLS) method the model parameters are derived by minimizing:

$$\sum_{i}^{m}\sum_{j}^{n} w_{ij}\left[\ln(m_{ij}) - \alpha_i - \beta_i\gamma_j\right]^2 ,$$

where w_{ij} can be taken as the reciprocal of the number of deaths at age i and in time period j.

Likelihood-based methods

To model death counts using likelihood approach, we need to specify the probability distribution for the death counts. As explained in Section 1.2.1, the total number of deaths over a specified age- and year-interval, Y_{ij}, is Poisson-distributed. Examples of likelihood-based approaches include the method of maximum likelihood estimation (MLE) considered by Wilmoth (1993) and implemented later by Brouhns et al. (2002), and the method of generalized linear models (GLMs) employed by Renshaw and Haberman (2006).

According to equation (1.3), the LC model in (1.9) can be written as follows:

$$Y_{ij} \sim \text{Poisson}\left(E_{ij} \cdot \exp(\alpha_i + \beta_i \cdot \gamma_j)\right)$$

Using MLE, we obtain estimates of the model parameters by maximizing the following log-likelihood

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{Y}, \boldsymbol{E}) = \sum_{i}\sum_{j}\left[Y_{ij} \cdot (\alpha_i + \beta_i\gamma_j) - E_{ij} \cdot (\exp(\alpha_i + \beta_i \cdot \gamma_j))\right] + c$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of the parameters α_i, β_i and γ_j. The c is a constant that is independent from the model parameters. The maximization can be accomplished via standard NewtonRaphson method. Brouhns et al. (2002) provide the associated derivations. The MLE does not require any re-estimation of γ_j.

In the method of GLMs, we use the log-link in modeling the "responses" Y_{ij}. The linear model can be written as

$$\ln(Y_{ij}) = \ln(E_{ij}) + \alpha_i + \beta_i \cdot \gamma_j$$

where $\ln(E_{ij})$ is the offset. An iterative algorithm is necessary in order to estimate the parameters: a first attempt was given by Renshaw and Haberman (2003a). GLM and MLE yield the same parameters estimates, if the same constraints for the parameters uniqueness are chosen.

The main advantage of using likelihood methods is that the errors are not assumed to be

homoscedastic. In contrast, the SVD assumed that the errors are normally distributed with constant variance, which is quite unrealistic: hazard rates have different variability over the whole of the age range. Furthermore, the Poisson approximation leads to a meaningful variance-covariance matrix, suitable diagnostic analysis and properties for forecasting death rates.

Alternative procedures have been proposed in order to improve and extend the LC model. Wang and Lu (2005) embedded the LC model in a binomial framework and computed interval estimates by a bootstrap approach. Czado et al. (2005) and Pedroza (2006) proposed an LC model in a Bayesian framework using Markov Chain Monte Carlo methods for parameters estimation. The extension proposed by Haberman and Renshaw (2008) deal with an age-period-cohort version of the LC model. Both de Jong and Tickle (2006), Delwarde et al. (2007) and Hyndman and Ullah (2007) have all used different smoothing approaches for overcoming LC model over-parameterization.

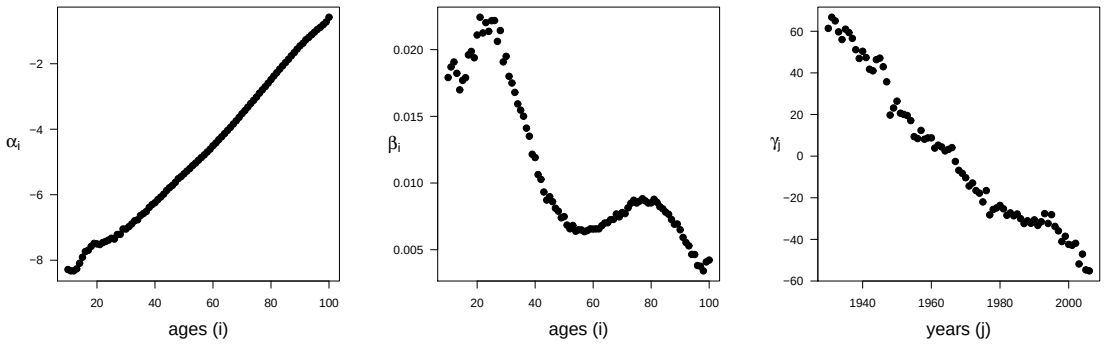

Figure 1.5: Lee-Carter estimates: α_i, β_i and γ_j. Ages from 10 to 100. Denmark, females, 1930–2006. Estimation procedure from Brouhns et al. (2002).

An application of the Lee-Carter model

Figure 1.5 shows the parameters α_i, β_i and γ_j of the LC model for the Danish mortality surface from ages 10 to 100 and years 1930-2006. In this example, we followed the methodology given by Brouhns et al. (2002). It is immediately apparent that parameters estimates follow specific regular trends. In particular, the left panel of Figure 1.5 presents the aforementioned linear trend which would be extrapolated in the forecasting applications.

Actual death rates along with fitted values are presented in Figure 1.6. Though the fitted values gives an overview of the mortality developments, the LC model still captures features in the trends which can be easily seen as random noise of the data. More accurate diagnostic analysis and goodness-of-fit measure are presented in Section 2.3.1 and in Chapter 3, respectively.

1.6 From over-parametric to smooth models

Both the LC and the APC model use individual parameters for each age and year (and possibly cohort). Overparameterization influences the estimation procedures and, from a more conceptual

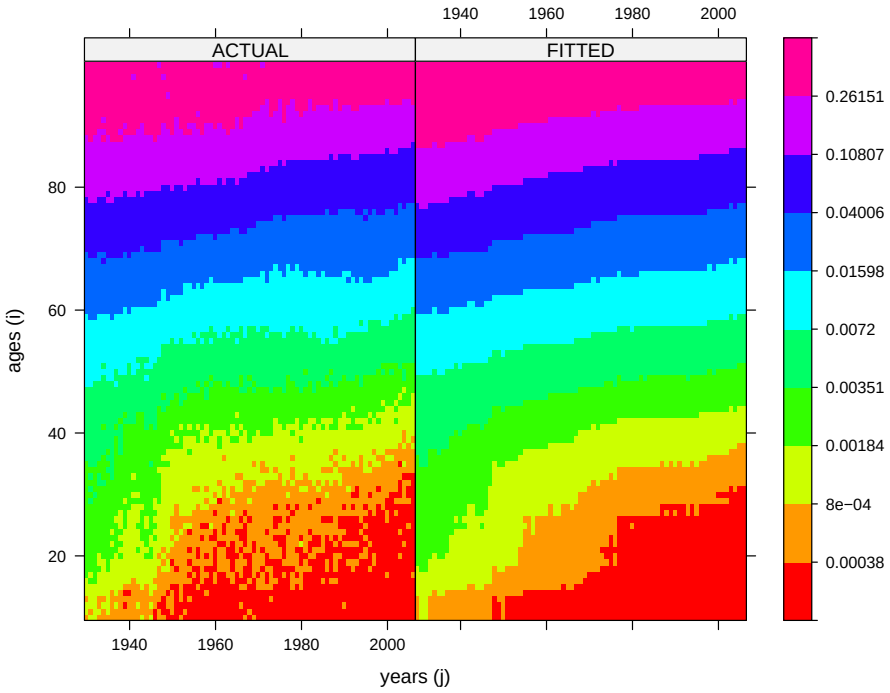

Figure 1.6: Actual and fitted death rates from Lee-Carter model. Ages from 30 to 100. Denmark, females, 1930–2006. Estimation procedure from Brouhns et al. (2002)

view, the use of such an amount of parameters may seem unnecessary, given the rather regular structure of human mortality development.

In developed countries with good data, large populations, and no extraordinary events (e.g. wars and epidemics), changes over time or over age normally show regular patterns, and erratic behavior is mainly caused solely by randomness of the rates (Figures 1.3 and 1.4). Therefore, more parsimonious models, albeit still flexible enough to pick up the age pattern and the time trend, should be able to capture the essence of the mortality surface. Smoothing approaches are a natural choice because mortality surfaces are themselves so informative that imposing a strong model structure seems unnecessary.

Smoothing methods for two-dimensional problems have been proposed by Cleveland and Devlin (1988), who use a generalization of the "Loess" methodology, and de Boor (2001) and Dierckx (1993), who employ a two-dimensional regression basis as the Kronecker product of B-splines. Gu and Wahba (1993) and Wood (2003) fit surfaces with thin plate splines. In addition, the mortality surface can be embedded in the framework of Generalized Additive Models (GAM) (Hastie and Tibshirani, 1990; Wood, 2006) if the more restrictive assumption of additive effects is justified.

An alternative appealing methodology was developed using two-dimensional regression splines, specifically B-splines with penalties, known as P-splines. Eilers and Marx (1996) deal with uni-dimensional regression and the extensions for bivariate regression have been presented in Eilers and Marx (2002b), Currie et al. (2004, 2006) and Eilers et al. (2006). A detailed description two-dimensional P-splines is presented in Section 2.2.

For demonstrative purposes, Figure 1.7 shows the Danish mortality surface along with fitted values by two-dimensional regression P-splines. At first glance, it is easy to check that outcomes in Figure 1.7 are more suitable than the LC model in describing mortality developments. Moreover, trends over ages and time look smoother and a P-spline approach employs approximately only 137 parameters, whereas the LC model estimates 257 for describing the same mortality surface[3].

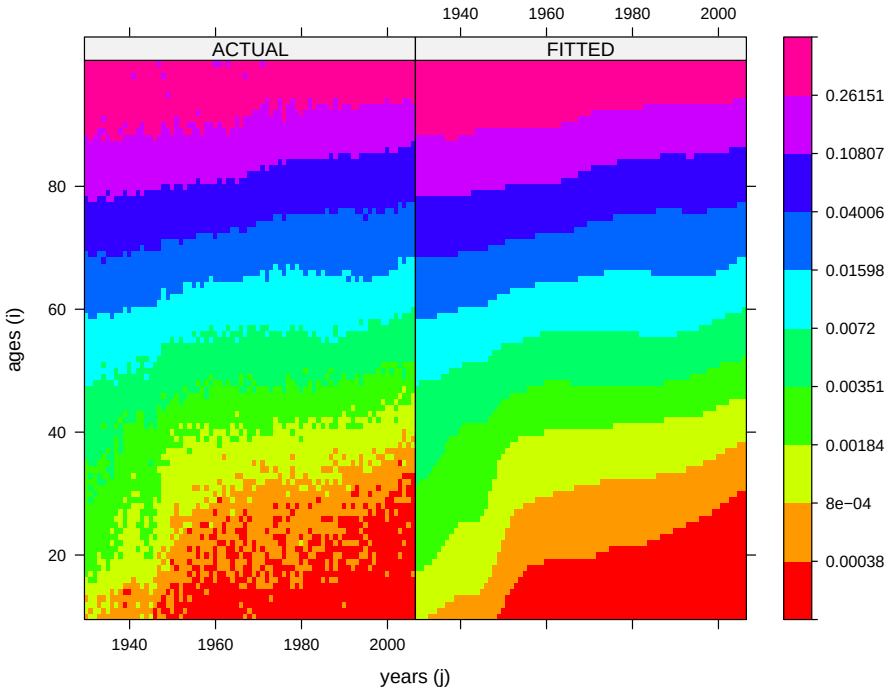

Figure 1.7: Actual and fitted death rates from 2D smoothing with P-splines. Ages from 0 to 100. Denmark, females, 1900–2006.

[3]For the concept of effective dimensions in a smoothing setting we refer to Section 2.1.3

Chapter 2

Smooth modeling of mortality

In the previous chapter, we presented several methods and models for analyzing mortality data. Many of the demographic approaches rely on parametric or over-parameterization assumptions, leading to rigid modeling structures and an unreasonable number of parameters. Hence, parsimonious approaches are needed in modeling mortality data.

Figure 1.7 in Section 1.6 already has shown outcomes of the smoothing methodology used in this thesis: two-dimensional P-splines regression. In this chapter we will present this approach in more detail. In the next section, we first introduce the P-spline methodology in the setting of unidimensional scatterplot smoothing. Thereafter, in Section 2.2, we will give more details on the generalization of the methodology in two-dimensional problems such as mortality surfaces. Finally Section 2.3 will be devoted to an analysis of the statistical tools and issues in measuring uncertainty in this context.

2.1 P-splines: an introduction

In the simplest case of a univariate response, which is normally distributed, a smooth relationship between the response $\boldsymbol{y}$ and a single predictor $\boldsymbol{x}$ is given as

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}\,, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}) \tag{2.1}$$

where the function f is assumed to be smooth. The aim is to estimate f given the observed pairs (x_i, y_i). Since death counts are Poisson-distributed, generalizations of the error distribution would be necessary for modeling mortality data.

2.1.1 Normal data

There are several methodologies for fitting f. Good summaries can be found in Hastie and Tibshirani (1990) and Simonoff (1996). Here we will focus on P-splines. Simplifying the scheme of O'Sullivan (1988), Eilers and Marx (1996) developed a method which combines (fixed-knot) B-splines with a roughness penalty. Generalized Linear Models (GLMs, McCullagh and Nelder, 1989) can nonparametrically be estimated with P-splines and we refer to Section 7 of Eilers and Marx (1996) for a fuller reference. Descriptions of the P-spline method can be found in the seminal

paper of Eilers and Marx (1996) as well as in Marx and Eilers (1998), Eilers and Marx (2002a) and in Currie and Durban (2002). A comprehensive study on the methodology is given in Currie et al. (2006). Wand (2003) reviews a mixed effect approach and provides a useful bibliography. Different applications can be found in Marx and Eilers (1999), Parise et al. (2001), Coull et al. (2001) and in Currie and Durban (2002).

Specifically, B-splines are bell-shaped curves composed of smoothly joint polynomial pieces. Polynomials of degree $q = 3$ will be used in the following [1]. The positions on the horizontal axis, where the pieces come together, are called "knots". We will use equally spaced knots of a distance h. For details on B-splines and related algorithms see de Boor (1978) and Dierckx (1993).

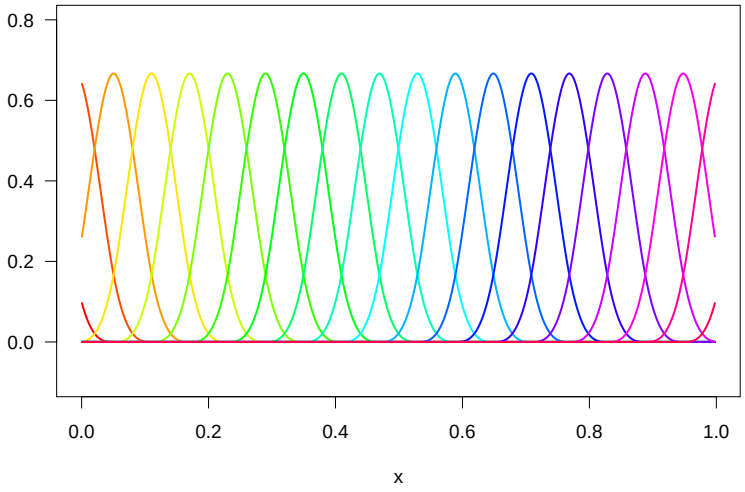

Figure 2.1: B-spline bases with equally-spaced knots, $k = 20$ and $q = 3$.

B-splines are a base of local functions that is well-suited for smoothing a scatterplot of pairs $(x_i, y_i), i = 1, \ldots, n$. Let $b_{ij}^q = B_j^q(x_i), j = 1, \ldots k(< n)$ be the value of the jth B-spline at x_i of degree q. $\boldsymbol{B} = [b_{ij}^q]$ denotes the matrix of covariates and $\boldsymbol{a}$ their respective regression coefficients. Figure 2.1 shows an example of $\boldsymbol{B}$, where the domain of x run from 0 to 1, $k = 20$ and $q = 3$. All bases have the same shape, but they are shifted horizontally by a multiple of the knot distance. Note that this is also true at the boundaries. This feature prevents boundary effects, as many types of kernel smoothers do not (Gasser and Müller, 1979; Marron and Ruppert, 1994).

If we rewrite equation (2.1) as

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \tag{2.2}$$

then, the smoothed function is found by minimizing

$$S = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|^2 \tag{2.3}$$

with the explicit solution

$$\hat{\boldsymbol{a}} = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{y} \tag{2.4}$$

[1] The choice of the degree of the polynomials is relatively irrelevant in case of P-spline models (Eilers and Marx, 2002a)

Given $\hat{\boldsymbol{a}}$, for any x, the fitted function will be $\hat{f}(x) = \sum_j \boldsymbol{B}_j(x)\hat{\boldsymbol{a}}_j = \hat{\boldsymbol{\mu}}$, and thus is a linear regression of $\boldsymbol{y}$ on $\boldsymbol{B}$. One can easily see that the higher the number of B-splines, the closer the smoothed curve is to the data. Conversely, a small number of B-splines leads to a smoother fitted curve.

The problem one faces now is to find an optimally smoothed curve. Following the approach outlined by Eilers and Marx (1996), we can choose a relatively large number of B-splines which would normally result in over-fitting. A penalty is put on the regression coefficients, in order to force the coefficients to vary more smoothly. We add to equation (2.3) a penalty weighted by a positive regularization parameter λ

$$S^* = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|^2 + \lambda\|\boldsymbol{D}_d\boldsymbol{a}\|. \tag{2.5}$$

The matrix $\boldsymbol{D}_d$ constructs dth order differences of $\boldsymbol{a}$

$$\boldsymbol{D}_d\boldsymbol{a} = \Delta^d\boldsymbol{a}\,.$$

As examples, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are as follows, when $k = 5$:

$$\boldsymbol{D}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}; \qquad \boldsymbol{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}. \tag{2.6}$$

High level programming languages have functions to apply the difference operator to a matrix; construction of $\boldsymbol{D}_d$ is then trivial, by (repeated) differencing of the identity matrix. Unless otherwise stated, we will use $d = 2$ in the following.

The solution of equation (2.5) is given by

$$\hat{\boldsymbol{a}} = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y} \tag{2.7}$$

where $\boldsymbol{P} = \lambda\boldsymbol{D}_d'\boldsymbol{D}_d$.

Figure 2.2 illustrates the capability of P-splines in smoothing scattered (x_i, y_i), $i = 1, \ldots, n$ simulated data[2]. The upper panel of Figure 2.2 is the solution of equation (2.7) with $k = 20$ B-splines of degree $q = 3$. The order of the penalty term is $d = 2$ and the smoothing parameter λ is equal to 10. These outcomes are compared with simple B-splines in which the coefficient vector $\boldsymbol{a}$ is unpenalized: $\lambda = 0$ (lower panel in Figure 2.2). The B-spline bases multiplied by the penalized and unpenalized coefficients $\boldsymbol{a}$ are also shown in the bottom part of both panels in Figure 2.2.

By changing λ the smoothness can be tuned (see Figure 2.3). Hence, the parameter λ controls the trade-off between smoothness and model fidelity. The number of equally spaced knots does not matter much, provided that enough of them are chosen to ensure greater flexibility than is needed (Eilers and Marx, 2002a).

[2]The data were simulated as follow: $y_i \sim \mathcal{N}(\mu_i, 0.16)$, $\mu_i = e^{x_i} + 0.4\,\sin(10\,x_i)$ and $x_i \sim \text{Unif}[0,1]\,, i = 1, \ldots, 100$.

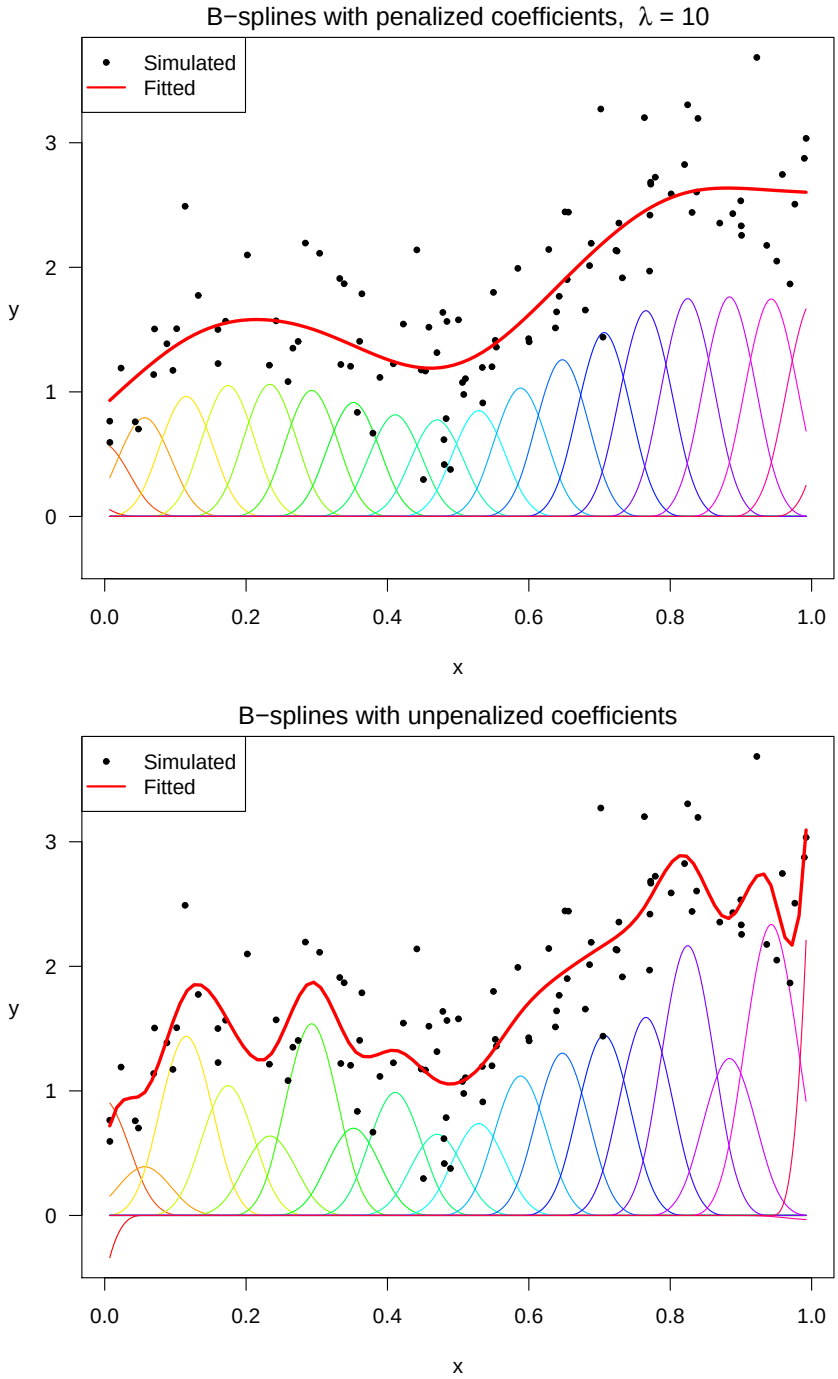

Figure 2.2: Penalized (upper panel) and unpenalized regression (lower panel). Simulated data. B-spline bases with equally-spaced knots, $k = 20$ and $q = 3$. $d = 2$ and $\lambda = 10$ for the penalized regression.

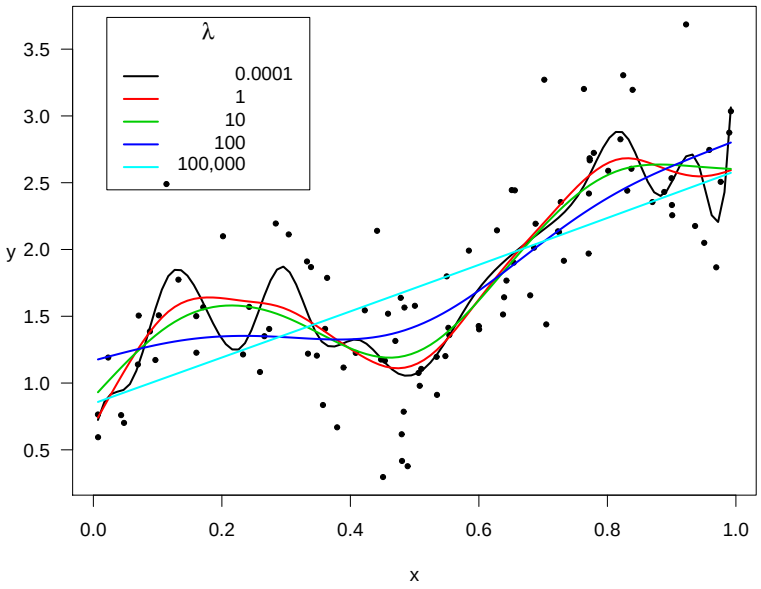

Figure 2.3: Smoothing of simulated data using P-splines with different parameters $\lambda = \{0.0001, 1, 10, 100, 100000\}$.

As in classic linear regression setting, from equation (2.7), we can specify the hat matrix, for a given value of λ:

$$\boldsymbol{H}_\lambda = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'. \tag{2.8}$$

Hence, if we focus on the fit at the observed points $x_1, \dots x_n$, a smoother such as P-splines can be also expressed as

$$\hat{\boldsymbol{\mu}} = \boldsymbol{H}_\lambda \boldsymbol{y}. \tag{2.9}$$

2.1.2 Count data

The P-spline methodology can be easily generalized to non-normally distributed data, such as Poisson counts. In GLMs, we introduce a linear predictor $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{a}$ and a (canonical) link function $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the expectation of $\boldsymbol{y}$, i.e. $E(\boldsymbol{y}) = \boldsymbol{\mu}$. Alternatively, we can write $\boldsymbol{\mu} = h(\boldsymbol{\eta})$ where $h(\cdot) = g^{-1}(\cdot)$, the inverse of the link function, sometimes called the response function.

With Normal data, minimizing least squares objective functions in equations (2.3) and (2.5) is equivalent to maximizing unpenalized and penalized log-likelihoods, respectively. In general, the penalized log-likelihood which will be maximized can be written as:

$$\begin{aligned} l^* &= l(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y}) - \frac{1}{2}\lambda \|\boldsymbol{D}_d \boldsymbol{a}\|^2 = \\ &= l(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y}) - \frac{1}{2}\boldsymbol{a}'\boldsymbol{P}\boldsymbol{a}. \end{aligned} \tag{2.10}$$

The factor $\frac{1}{2}$ is chosen for convenience only, such that it disappears after differentiation. $l(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y})$ is the usual log-likelihood for a GLM and $\boldsymbol{P} = \lambda \boldsymbol{D}_d' \boldsymbol{D}_d$.

Maximizing equation (2.10) gives the penalized likelihood equations

$$\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{P}\boldsymbol{a}$$

which can be solved with a penalized version of the iteratively reweighted least squares (IRWLS) algorithm[3] for the estimation of GLMs (Nelder and Wedderburn, 1972)

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}+\boldsymbol{P})\tilde{\boldsymbol{a}}=\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}\tilde{\boldsymbol{a}}+\boldsymbol{B}'(\boldsymbol{y}-\tilde{\boldsymbol{\mu}}) \tag{2.11}$$

where $\boldsymbol{B}$ is again the regression matrix, $\boldsymbol{P}$ is the penalty matrix, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{a}}$ denote current approximations to the solution and $\tilde{\boldsymbol{W}}$ is a diagonal matrix of weights

$$w_{ii}=\frac{1}{v_i}\left(\frac{\partial\,\mu_i}{\partial\,\eta_i}\right)^2,$$

where v_i is the variance of y_i, given μ_i. Note that the only difference with the standard procedure for fitting a GLM with B-splines as regressors is the modification of $\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}$ by $\boldsymbol{P}$ (which itself is constant for fixed λ) at each iteration.

In the case of Poisson errors, $\tilde{\boldsymbol{W}}=\mathrm{diag}(\tilde{\boldsymbol{\mu}})$ and the canonical link, $\ln(\cdot)$, will be used throughout this study. The algorithm (2.11) can also be written as

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}+\boldsymbol{P})\tilde{\boldsymbol{a}}=\boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \tag{2.12}$$

where $\tilde{\boldsymbol{z}}=(\boldsymbol{y}-\tilde{\boldsymbol{\mu}})/\tilde{\boldsymbol{\mu}}+\boldsymbol{B}\tilde{\boldsymbol{a}}$, which is defined as a working dependent variable. The formulation in (2.12) leads directly to the solution at the step $t+1$:

$$\hat{\boldsymbol{a}}_{t+1}=(\boldsymbol{B}'\hat{\boldsymbol{W}}_t\boldsymbol{B}+\boldsymbol{P})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}_t\hat{\boldsymbol{z}}_t. \tag{2.13}$$

Also, for non-Normal data, the hat matrix can be easily computed from the estimated linearized smoothing problem in (2.12):

$$\boldsymbol{H}_\lambda=\boldsymbol{B}(\boldsymbol{B}'\hat{\boldsymbol{W}}\boldsymbol{B}+\boldsymbol{P})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}\,, \tag{2.14}$$

where $\hat{\boldsymbol{W}}$ contains the weight of the last iterations after convergence.

When modeling mortality data, it is necessary to take into account the exposures in the presented regression setting. Specifically, we seek a smooth estimate of the actual death rates and from equation (1.3), the linear predictor $\boldsymbol{\eta}$ can be written as

$$\boldsymbol{\eta}=\ln(\boldsymbol{\mu})=\ln(E(\boldsymbol{y}))=\ln(\boldsymbol{e}\cdot\boldsymbol{m})=\ln(\boldsymbol{e})+\ln(\boldsymbol{m})=\ln(\boldsymbol{e})+\boldsymbol{B}\boldsymbol{a},$$

where $\boldsymbol{e}$, $\boldsymbol{y}$ and $\boldsymbol{m}$ are exposures, deaths and death rates, respectively, over a single dimension (age or time). The term $\boldsymbol{e}$ is called offset and can be easily incorporated in the regression system (2.12).

Figure 2.4 shows the estimated death rates of the Danish population at age 60, from 1930 to 2006., using a basis of 18 cubic B-splines and a smoothing parameter $\lambda=100$.

[3] See also Section 4.3.1 for a detailed description of the IRWLS.

Figure 2.4: Actual and fitted death rates from a P-spline approach, logarithmic scale. B-spline bases with equally-spaced knots, $k = 18$, $q = 3$, $d = 2$ and $\lambda = 100$. Denmark, females, age 60, years from 1930 to 2006.

2.1.3 Effective dimension of a smoother

An important concept in modeling is the effective dimension of the fitted model itself. This concept is clear and intuitive in the case of linear models: the number of parameters used in the model quantifies its dimension. However, in a non-parametric setting a different definition is needed.

A smoothing method, such as the P-splines, needs to balance a fundamental trade-off between the bias and the variance of the estimates. The smoothing parameter, λ, tunes this trade-off. In linear smoothers such as P-splines, it can be shown, as $n \to \infty$ and $\lambda \to 0$, under certain regularity conditions, $\hat{f}(x) \to f(x)$ (Hastie and Tibshirani, 1990, p. 40).

Let $\boldsymbol{b}_\lambda = \boldsymbol{f} - E(\boldsymbol{H}_\lambda \boldsymbol{y}) = \boldsymbol{f} - \boldsymbol{H}_\lambda \boldsymbol{f}$ denote the bias vector, where, for the P-splines case, $\boldsymbol{H}_\lambda$ is given in (2.8) and (2.14) and $f_i = f(x_i)$. Hastie and Tibshirani (1990, p. 46) show that we can measure the variance of a linear smoother in the following way

$$\operatorname{tr}(\boldsymbol{H}_\lambda \boldsymbol{H}'_\lambda) = \sigma^2 \sum_i^n \operatorname{var}(\hat{f}(x_i))$$

and a measure of squared bias is given by

$$\boldsymbol{b}'_\lambda \boldsymbol{b}_\lambda = \sum_i^n b^2_\lambda(x_i)\,.$$

Therefore, as the amount of smoothing increases, we expect the bias to increase while the variance decreases. Consequently, with an increasing amount of smoothing, $\operatorname{tr}(\boldsymbol{H}_\lambda \boldsymbol{H}'_\lambda)$ tends to decrease, while the elements of $\boldsymbol{b}_\lambda$ tend to increase. As a matter of fact, $\operatorname{tr}(\boldsymbol{H}_\lambda \boldsymbol{H}'_\lambda)$ can be considered as a quantity that we use to calibrate the amount of smoothing performed by P-splines.

This idea can be easily seen in a more simpler case, such as the classic linear regression

(Weisberg, 1985, pp. 110-111). In particular, the hat matrix for linear models is idempotent, i.e. $\text{tr}(\boldsymbol{H}_\lambda \boldsymbol{H}_\lambda) = \text{tr}(\boldsymbol{H}_\lambda) = \text{rank}(\boldsymbol{H}_\lambda)$, which is equal to the number of parameters in the fitted model.

Given this feature of the classic linear model, we can define the effective dimension, or degrees of freedom, to be

$$\text{ED}(\boldsymbol{a}, \lambda) = \text{tr}(\boldsymbol{H}_\lambda) \tag{2.15}$$

see Hastie and Tibshirani (1990, p. 52) and (Buja et al., 1989, p. 469). Equivalently, the ED is the sum of the eigenvalues of $\boldsymbol{H}_\lambda$. The relationship of this measure with the variance-bias trade-off in a smoother, will have a crucial importance in the smoothing parameter selection, as we will see in Section 2.1.4. Although the effective dimension is a function of both λ and the predictors in the model, ED is mainly determined by λ. It is noteworthy that ED is not a function of $\boldsymbol{Y}$, and this fact will ease the computation cost of selecting λ.

Eilers and Marx (1996, p. 94) also pointed out that in a P-spline setting the effective dimension will approach d, the order of difference. That is, d defines the degree of the "ultimately smooth" function for $\lambda \to +\infty$. In other words, in the limit of a very large λ, a linear ($d = 2$) or quadratic ($d = 3$) fit is obtained, as can be seen in Figure 2.3. Figure 2.5 shows the profile of the effective dimensions over a range of λ from the P-spline model fitted on a Danish female population at age 60 from 1930 to 2006 (Figure 2.4). Furthermore, it is worth pointing out that the ED profile shows an asymptotic behavior for both smaller and larger λ. In the former case, ED is practically equal to the number of B-splines ($k = 18$) for λ smaller than 0.01. We have effective dimension nearly equal to 2 for all the λ larger than 10,000,000.

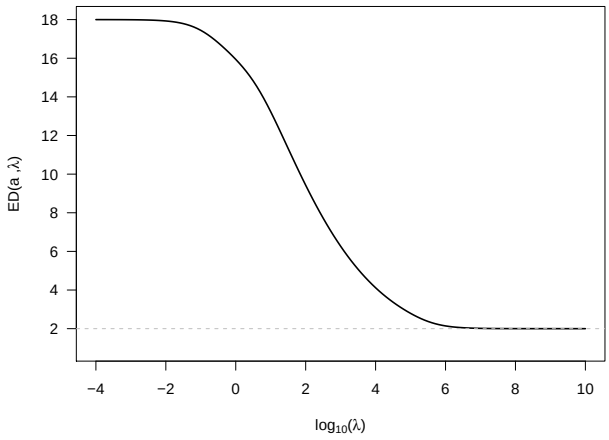

Figure 2.5: $\text{ED}(\boldsymbol{a}, \lambda)$ over increasing $\log_{10}(\lambda)$, cf. equation (2.15). Denmark, females, age 60, years from 1930 to 2006.

Alternative definitions of ED are also motivated by its analogy with classic linear normal models, namely $\text{tr}(2\boldsymbol{H}_\lambda - \boldsymbol{H}_\lambda \boldsymbol{H}_\lambda')$ and $\text{tr}(\boldsymbol{H}_\lambda \boldsymbol{H}_\lambda')$. Nevertheless, Hastie and Tibshirani (1990) suggest to approximate the overall effective dimension or degrees of freedom with the trace of the hat matrix of the smoother.

2.1.4 Smoothing parameter selection

Empirically, we are always faced with the trade-off between parsimony on the one hand and accuracy on the other. By parsimony, one means models with low effective dimensions or few parameters. By accuracy, one means the ability to reproduce observed data, commonly measured by suitable statistics.

In a P-spline approach, this trade-off is clearly driven by the choice of the smoothing parameter λ. Whereas the effective dimension of the model has been defined in Section 2.1.3, measures of discrepancy between actual and fitted values are borrowed by the GLMs framework.

A common measure of discrepancy in GLMs is the deviance. This measure is constructed from the logarithm of a ratio of likelihoods and it is proportional to twice the difference between the maximum log-likelihood achievable, commonly called "saturated model", and the log-likelihood achieved by the fitted model (McCullagh and Nelder, 1989, p. 33). For normally distributed data, the deviance is just the residual sum of squares, while for the Poisson, as it is in our cease, takes the following form

$$\text{Dev}(\boldsymbol{y};\boldsymbol{a},\lambda) = 2\sum_{i=1}^{n}\left\{y_i\, log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}. \tag{2.16}$$

The second term, a sum of differences between observed and fitted, is usually zero, because maximum likelihood estimators in Poisson models with log-link have the property of reproducing marginal totals, i.e. $\sum y_i = \sum \hat{\mu}_i$.

Alternatively and regardless of the complexity of the model, a measure of discrepancy between observed and fitted values in a Normal case is the mean squared error (MSE). This measure is given by (Hastie and Tibshirani, 1990, eq. 3.8)

$$MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}_\lambda(x_i)\right\}^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{\mu}_i\}^2 \,. \tag{2.17}$$

It can be shown that (2.17) is equivalent to the sum of variance and squared bias measures presented in Section 2.1.3 (see Hastie and Tibshirani, 1990, Section 3.4.2).

In a P-spline framework, we need some way to choose an "optimal" value for λ, which can balanced bias and variance in the model construction. Eilers and Marx (1996, 2002a) suggested using the Akaike's information criterion (AIC) (Akaike, 1973). The AIC is a common tool for model selection and it corrects the log-likelihood of a fitted model for the effective dimension. For a fuller treatment of the criterion we refer to Sakamoto et al. (1986). The expression for AIC is given by

$$\text{AIC}(\lambda) = \text{Dev}(\boldsymbol{y};\boldsymbol{a},\lambda) + 2\cdot\text{ED}(\boldsymbol{a},\lambda)\,, \tag{2.18}$$

where $\text{ED}(\boldsymbol{a},\lambda)$ and $\text{Dev}(\boldsymbol{y};\boldsymbol{a},\lambda)$ are given in equations (2.15) and (2.16), respectively and can be computed for a fixed λ.

Alternatively, one can use the Bayesian Information Criterion (BIC) (Schwarz, 1978), which penalizes model complexity more heavily than AIC, particularly when n is large (Chatfield, 2003). This penalization is practically done by increasing the factor multiplied by the model dimension

in the criterion and taking into account the value of n:

$$\mathrm{BIC}(\lambda) = \mathrm{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda) + \ln(n) \cdot \mathrm{ED}(\boldsymbol{a}, \lambda)\,. \tag{2.19}$$

In cases of normally distributed data, as alternative to AIC and BIC, Eilers and Marx (1996) suggested using cross-validation (CV) to find the optimal value of λ:

$$CV(\lambda) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(x_i)}{1 - H_{ii}(\lambda)}\right\}^2}\,.$$

Figure 2.6 presents the AIC and BIC profile over $\log_{10}(\lambda)$. Note how the λ picked by BIC is substantially higher than the λ selected by AIC.

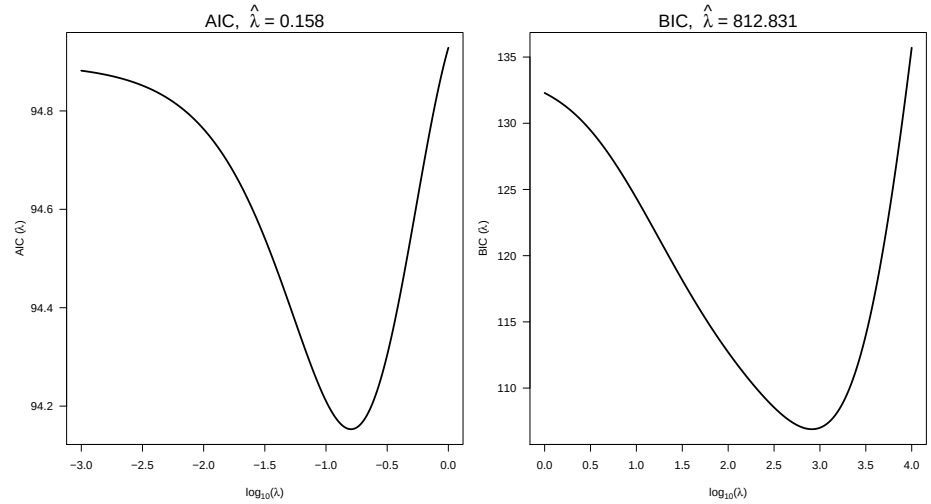

Figure 2.6: AIC and BIC over a range of $\log_{10}(\lambda)$, cf. equations (2.18) and (2.19). Denmark, females, age 60, years from 1930 to 2006.

Once λ is selected, the system of equations described in (2.12) has a unique solution. Figure 2.7 presents the estimated death rates of the Danish population at age 60, from 1930 to 2006, using both AIC and BIC to select the optimal smoothing parameter λ. As already pointed out by Currie et al. (2004, p. 285), stiffer fit, which is given by the BIC, is preferred when modeling mortality data with P-splines.

2.2 P-spline models for mortality surfaces

In order to model data in arrays, such as mortality data, we seek to construct a basis for two-dimensional regression with local support analogous to the one introduced in Section 2.1. A detailed description of this generalization can be found in Eilers and Marx (2002b), Currie et al. (2004), Currie et al. (2006) and Eilers et al. (2006).

Let $\boldsymbol{Y} = (y_{ij})$ be the $m \times n$ matrix of deaths at age i, $i = 1, \ldots m$, and year j, $j = 1, \ldots, n$. For the purpose of regression, we suppose that the data are arranged as a column vector, that is, $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$. Accordingly, we arrange the matrices of exposures $\boldsymbol{E} = (E_{ij})$ which are used as

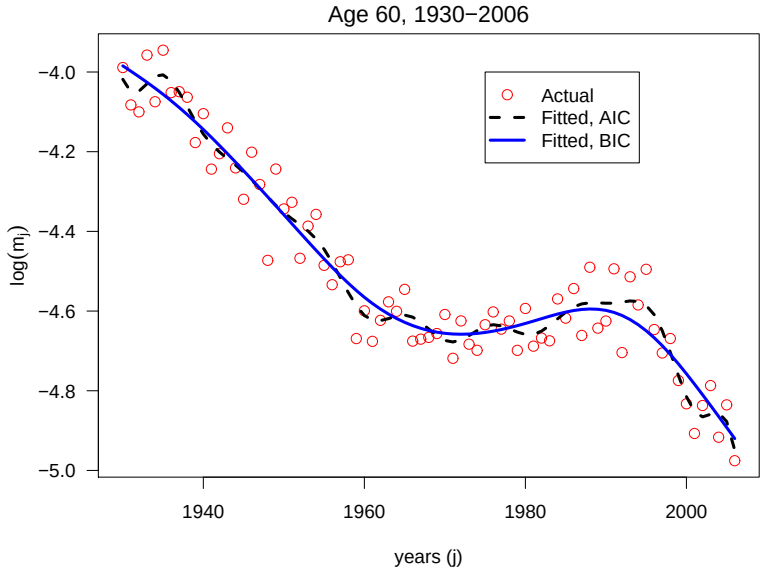

Figure 2.7: Actual and fitted death rates from a P-spline approach, logarithmic scale. B-spline bases with equally-spaced knots, $k = 18$, $q = 3$, $d = 2$ and λ selected by AIC and BIC. Denmark, females, age 60, years from 1930 to 2006.

offset in the Poisson setting, so that $\boldsymbol{e} = \text{vec}(\boldsymbol{E})$.

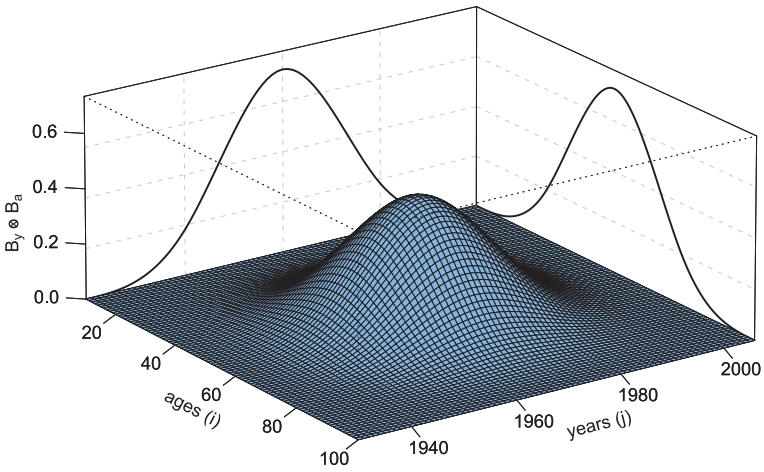

Figure 2.8: Two-dimensional Kronecker product of two cubic B-splines basis.

Let $\boldsymbol{B}_a = \boldsymbol{B}(\boldsymbol{x}_a)$, $n_a \times c_a$ be a regression matrix of B-splines based on age $\boldsymbol{x}_a$, and similarly, let $\boldsymbol{B}_y = \boldsymbol{B}(\boldsymbol{x}_y)$, $n_y \times c_y$, be a regression matrix of the explanatory variable for year $\boldsymbol{x}_y$. The regression matrix for our two-dimensional model is the Kronecker product

$$\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a \tag{2.20}$$

As in the unidimensional case, the number of columns in $\boldsymbol{B}_a$ and $\boldsymbol{B}_y$ is related to the number of knots chosen for the B-splines.

Figure 2.8 presents the Kronecker product of two cubic B-spline over the age-year grid. Following the same idea of the unidimensional case, we will use a relatively large number of

equally spaced B-splines over both domains. Figure 2.9 gives an impression of how a Kronecker product of several B-splines looks. The age-year grid is filled by a set of overlapping hills which are placed at regular intervals over the region. For clarity, only a subset of hills from a small basis is shown in Figure 2.9.

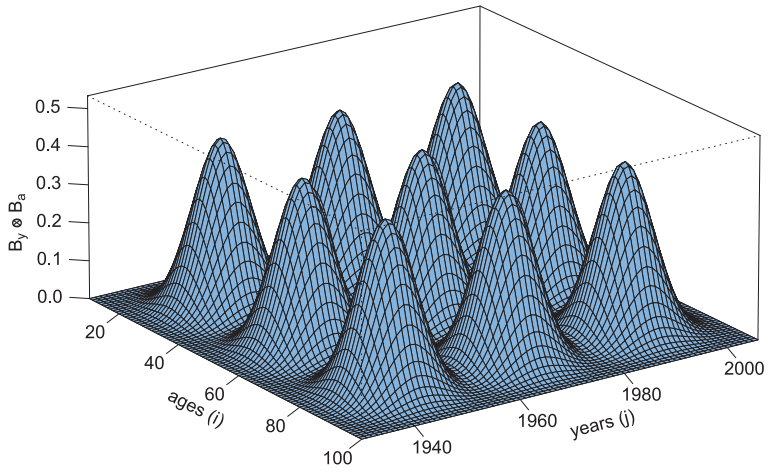

Figure 2.9: Two-dimensional Kronecker product of cubic B-splines basis.

In a regression setting, the matrix $\boldsymbol{B}$ has an associated vector of regression coefficients $\boldsymbol{a}$ of length $c_a c_y$. Therefore, the model can be written as

$$\boldsymbol{\mu} = E(\boldsymbol{y}) = (\boldsymbol{B}_y \otimes \boldsymbol{B}_a)\boldsymbol{a} = \boldsymbol{B}\boldsymbol{a}. \tag{2.21}$$

We arrange the elements of $\boldsymbol{a}$ in a $c_a \times c_y$ matrix $\boldsymbol{A}$, where $\boldsymbol{a} = \text{vec}(\boldsymbol{A})$ and the columns and rows of $\boldsymbol{A}$ are given by

$$\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{c_y}) \qquad \boldsymbol{A}' = (\boldsymbol{a}_1^r, \ldots, \boldsymbol{a}_{c_a}^r)\,,$$

then, instead of computing it as a vector, equation (2.21) can be written as

$$\begin{aligned} \boldsymbol{\mu} &= E(\boldsymbol{y}) = (\boldsymbol{B}_y \otimes \boldsymbol{B}_a)\boldsymbol{a} = \boldsymbol{B}\boldsymbol{a} \\ \boldsymbol{M} &= E(\boldsymbol{Y}) = \boldsymbol{B}_a \boldsymbol{A} \boldsymbol{B}_y \end{aligned} \tag{2.22}$$

It follows from the definition of the Kronecker product that the linear predictor of the columns of $\boldsymbol{Y}$ can be written as linear combinations of c_y smooths in age. The linear predictor corresponding to the jth column of $\boldsymbol{Y}$ can be expressed as

$$\sum_{k=1}^{c_y} b_{jk}^y \boldsymbol{B}_a \boldsymbol{a}_k$$

where $\boldsymbol{B}_y = (b_{ij}^y)$. Following the same idea of the unidimensional case, this result suggests that we should apply a roughness penalty to each of the columns of $\boldsymbol{A}$. An appropriate penalty will be given by

$$\sum_{j=1}^{c_y} \boldsymbol{a}_j' \boldsymbol{D}_a' \boldsymbol{D}_a \boldsymbol{a}_j = \boldsymbol{a}'(\boldsymbol{I}_{c_y} \otimes \boldsymbol{D}_a' \boldsymbol{D}_a)\boldsymbol{a}\,,$$

where $\boldsymbol{D}_a$ is the difference matrix acting on the columns of $\boldsymbol{A}$. In a similar fashion, by considering the linear predictor corresponding to the ith row of $\boldsymbol{Y}$, we can show that the corresponding penalty on the rows of $\boldsymbol{A}$ can be written as

$$\sum_{i=1}^{c_a} \boldsymbol{a}_i^{r'} \boldsymbol{D}_y' \boldsymbol{D}_y \boldsymbol{a}_i^r = \boldsymbol{a}'(\boldsymbol{D}_y' \boldsymbol{D}_y \otimes \boldsymbol{I}_{c_a})\boldsymbol{a}\,,$$

where $\boldsymbol{D}_y$ is the difference matrix acting on the rows of $\boldsymbol{A}$.

The regression coefficients $\boldsymbol{a}$ are estimated by maximizing the penalized log-likelihood (2.10) where $\boldsymbol{B}$ is given by equation (2.20) and the penalty term $\boldsymbol{P}$ by

$$\boldsymbol{P} = \lambda_a(\boldsymbol{I}_{c_y} \otimes \boldsymbol{D}_a' \boldsymbol{D}_a) + \lambda_y(\boldsymbol{D}_y' \boldsymbol{D}_y \otimes \boldsymbol{I}_{c_a})\,, \tag{2.23}$$

where λ_a and λ_y are the smoothing parameters used for age and year, respectively.

B-splines provide enough flexibility to capture surface trends. The additional penalty reduces the number of parameters leading to a rather parsimonious model with a smoothed fitted surface. The advantage of using two-dimensional P-splines lies also in the fact that different smoothing parameters can be chosen over ages and years, leading to great flexibility of the model.

In theory, we can use equation (2.13) to estimate $\boldsymbol{a}$. This will be possible in moderate-sized problems, but in the Danish example the parameter vector $\boldsymbol{a}$ has length $m \cdot n = 7007$ and this would require the usage of 7007×7007 matrices, and the penalized IRWLS algorithm quickly runs into storage and computational difficulties.

Eilers et al. (2006) and Currie et al. (2006) proposed an algorithm that takes advantage of the special structure of both the data as an rectangular array and the model matrix as a tensor product. The idea of this algorithm can be seen in the computation of the mean $\boldsymbol{\mu} = \text{vec}(\boldsymbol{M})$ in two dimensions, as in equation (2.22).

This avoids the construction of the large Kronecker product basis, saving space and time. In a similar fashion the two-dimensional "inner product" $\boldsymbol{B'WB}$ can be obtained efficiently. The key observation is that in a 4–dimensional representation, the elements of $\boldsymbol{B'WB}$ can be written

$$f_{jkj'k'} = \sum_h \sum_i w_{hi} b_{ij} b_{hk} b_{ij'} b_{hk'}$$

which can be rearranged as

$$f_{jkj'k'} = \sum_h b_{hk} b_{hk'} \sum_i w_{hi} b_{ij'} b_{ij}$$

Switching between four- and two-dimensional representations of matrices and arrays, one can completely avoid the construction of the large tensor product basis $\boldsymbol{B}$. These ideas, applied on a multidimensional grid algorithm, save both time and storage problems. More details can be found in the abovementioned papers, Eilers et al. (2006) and Currie et al. (2006).

2.2.1 A two-dimensional smoothing parameter selection

The penalty terms in (2.23) contains two different smoothing parameters, λ_a and λ_y. Since we still work in a regression setting, criteria such as AIC and BIC, can still be used for selecting

smoothing parameters in a two-dimensional approach, and the search mentioned earlier needs to be expanded into the two dimensions.

In a two dimensional setting, we also used the trace of the hat matrix as an effective dimension (eq. (2.15)) The hat matrix for Poisson data is defined in equation (2.14) in which the matrix $\boldsymbol{B}$ is the Kronecker product of $\boldsymbol{B}_y \otimes \boldsymbol{B}_a$ (cf. eq. (2.20))

The fitted mortality surface presented in Figure 1.7 (page 16), was estimated with a 26×22 grid of tensor product of cubic B-splines. The smoothing parameters λ_a and λ_y were selected by BIC and are equal to 955 and 10, respectively. Figure 2.10 displays the profile of both AIC and BIC.

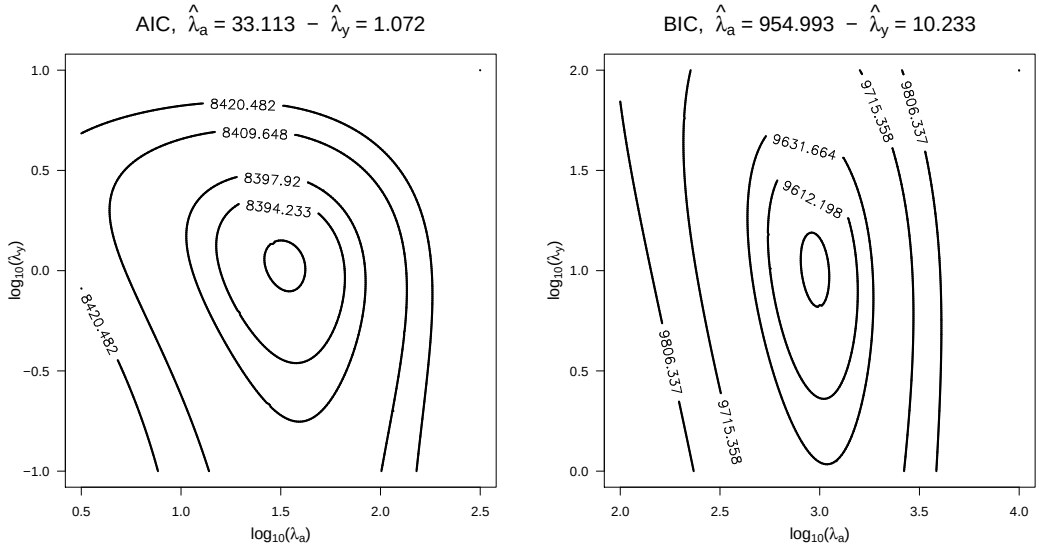

Figure 2.10: AIC (left panel) and BIC (right panel) over a two-dimensional grid of λ_a and λ_y.

Again, AIC selected smaller smoothing parameters than BIC, but the difference in fitted death rates are less evident when an entire mortality surface is estimated. Figure 2.11 shows death rates for the Danish females population at age 20 and 60. Fitted values are cross-section data from the estimated mortality surface, where, λ_a and λ_y are selected by both AIC and BIC.

In a two-dimensional setting, we can also compute the effective dimensions of the models from equation (2.15). Unlike in the fitted values, we note here the large difference in the effective dimensions of the fitted models selected by BIC and AIC: 137 with BIC, but 278 with AIC.

Furthermore, it is worth pointing out that we obtained quite different smoothing parameters over the two dimensions. As we could see in Figure 1.7 (page 16), trends over ages (10–100) are more regular than mortality developments from 1930 to 2006 for the Danish female population. Different smoothing over the two domains is commonly expected for mortality surfaces. Hence, though possible in a two-dimensional P-spline approach, isotropic smoothing ($\lambda_a = \lambda_y$) is not suitable when modeling mortality data.

Alternative two-dimensional smoothing methodologies, which rely on isotropic smoothing, are hence inadequate for fitting mortality data, too. Specifically, the general radial smoothers proposed by Ruppert et al. (2003, Section 13.4) and the Bayesian P-splines introduced by Lang and Brezger (2004) use only a single smoothing parameter over the two domains when fitting a

surface.

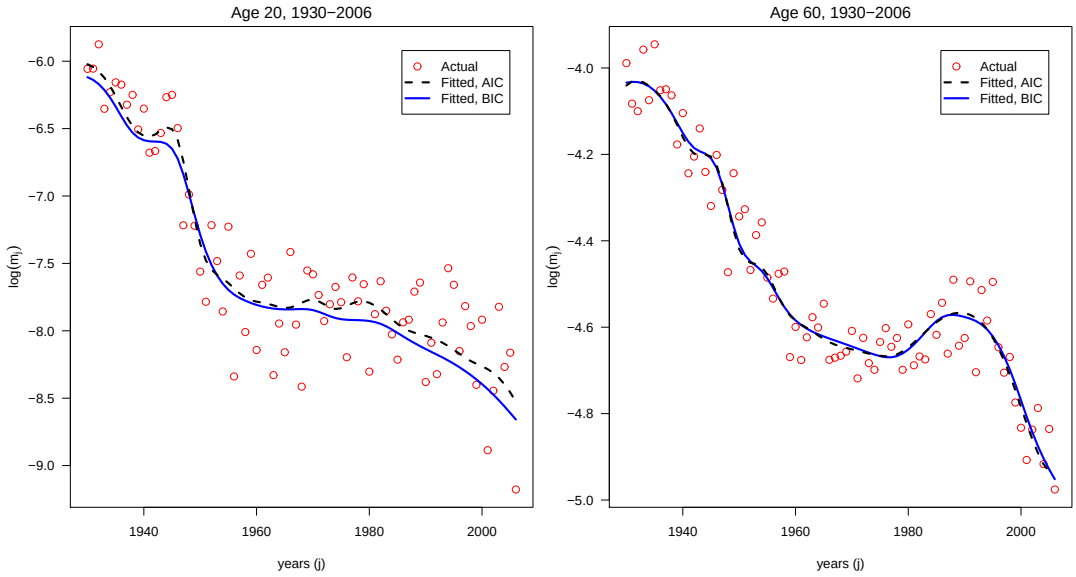

Figure 2.11: Actual and fitted death rates at age 20 (left panel) and age 60 (right panel), logarithmic scale. 2D smoothing with P-splines of the mortality surface. Ages from 10 to 100. Denmark, females, 1930–2006.

2.3 Measuring the uncertainty

One of the most appealing properties of the P-splines is their foundation on linear regression and GLMs. This fact avoids the need for backfitting and knot selection schemes as in Friedman and Silverman (1989) and Kooperberg and Stone (1991, 1992), and furthermore allows easy computation of diagnostics, compact results useful for prediction, computation of standard errors and fast cross-validation. Taking advantages of these properties, we will focus in the following sections on the diagnostics tools in the P-spline context with a particular emphasis on mortality surface modeling.

2.3.1 Residuals

In the process of statistical modeling, residuals provide information regarding assumptions about error terms and the appropriateness of the model. Any complete data analysis requires an examination of the residuals. Moreover, plots of residuals versus other quantities are used to find failures of assumptions. An early work on residuals is given by Anscombe (1961). Cox and Snell (1968, 1971) provide a systematic analysis of residuals in a linear model context. For standard normal models residuals are given by

$$\boldsymbol{r}_R = \boldsymbol{y} - \hat{\boldsymbol{\mu}}. \tag{2.24}$$

The subscription R denotes their usual name, *response residual*. In particular, errors in a linear model consist of unobservable random variables, assumed to have zero mean and uncorrelated elements, each with a common variance. We would like to assume residuals to behave as would

the unobservable errors, i.e. mean equal to zero and normally distributed.

In the GLMs framework, one of the first studies exclusively devoted to residuals was the one by Pierce and Schafer (1986). Widening the classic linear model for normally distributed data, Cox and Snell (1968, p. 258) pointed out a different kind of residuals in a more general framework, suggesting for Poisson data:

$$\boldsymbol{r} = \frac{\boldsymbol{y} - \hat{\boldsymbol{\mu}}}{\sqrt{\hat{\boldsymbol{\mu}}}}.$$

This type of residuals are commonly called *Pearson residuals*, and in general are defined by:

$$\boldsymbol{r}_P = \frac{\boldsymbol{y} - \hat{\boldsymbol{\mu}}}{\sqrt{V(\hat{\boldsymbol{\mu}})}} \tag{2.25}$$

Pearson residuals can be seen as response residuals scaled by the estimated standard deviation of $\boldsymbol{Y}$. In the Poisson case the distribution of the Pearson residuals is the signed square root of the component of the Pearson χ^2 goodness-of-fit statistic. In formula:

$$\sum r_P^2 = \chi^2$$

Pearson's statistic is normally used as a measure of residual variation.

Pearson residuals are often markedly skewed for non-Normal responses and may fail to have properties similar to those of Normal-theory residuals. In order to normalize the residuals, Anscombe (1953) defined a residual using a function $A(\boldsymbol{y})$ in place of $\boldsymbol{y}$. Given the likelihood in the GLMs, the function $A(\cdot)$ is given by

$$A(\cdot) = \int \frac{d\boldsymbol{\mu}}{V^{1/3}(\boldsymbol{\mu})}.$$

For details see Barndorff-Nielsen (1978).

For a response that follows a Poisson distribution we have

$$\int \frac{d\boldsymbol{\mu}}{\boldsymbol{\mu}^{1/3}} = \frac{3}{2}\boldsymbol{\mu}^{2/3}.$$

Thus, the residuals become $\boldsymbol{y}^{2/3} - \boldsymbol{\mu}^{2/3}$. Nevertheless, even if this transformation corrects for the skewness of the distribution, we still need to stabilize the variance of the residuals. This can be achieved by scaling the residuals, i.e. dividing by the square root of the variance of $A(\boldsymbol{y})$, which is, to the first order, $A'(\boldsymbol{\mu})\sqrt{V(\boldsymbol{\mu})}$. Employing also this second transformation, the so-called *Anscombe residuals* for Poisson distribution are given by

$$\boldsymbol{r}_A = \frac{\frac{3}{2}(\boldsymbol{y}^{2/3} - \hat{\boldsymbol{\mu}}^{2/3})}{\hat{\boldsymbol{\mu}}^{1/6}}. \tag{2.26}$$

In GLMs, the discrepancy between fitted and actual data can be measured by the deviance (cf. Section 2.1.4). Thus the individual unit of this quantity $d_i : \sum d_i = \text{Dev}$, can be used as

residuals. *Deviance residuals* are given by

$$\boldsymbol{r}_D = \text{sign}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})\sqrt{d_i} \tag{2.27}$$

This quantity increases with $y_i - \hat{\mu}_i$ and $\sum \boldsymbol{r}_D^2 = \text{Dev}$. In the Poisson GLM case the deviance residuals are:

$$\boldsymbol{r}_D = \text{sign}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})\{2(\boldsymbol{y}\log(\boldsymbol{y}/\hat{\boldsymbol{\mu}}) - \boldsymbol{y} + \hat{\boldsymbol{\mu}})\}^{1/2}.$$

Anscombe and deviance residuals are often similar although their equations look rather different. This similarity becomes evident when using a Taylor series expansion. The association can also be seen if we let $\boldsymbol{y} = c\,\boldsymbol{\mu}$. In the Poisson case, the mentioned residuals are

$$\begin{aligned} \boldsymbol{r}_A &= \frac{3}{2}\boldsymbol{\mu}^{1/2}(c^{2/3} - 1) \\ \boldsymbol{r}_D &= \text{sign}(c-1)\boldsymbol{\mu}^{1/2}[2(c\log c - c + 1)]^{1/2} \\ \boldsymbol{r}_P &= \boldsymbol{\mu}^{1/2}(c-1) \end{aligned} \tag{2.28}$$

Regardless of the value of $\boldsymbol{\mu}$, equations (2.28) are a functions of c. Therefore, the types of residuals can be compared by computing numerically those functions for a range of values of c (McCullagh and Nelder, 1989, p. 39).

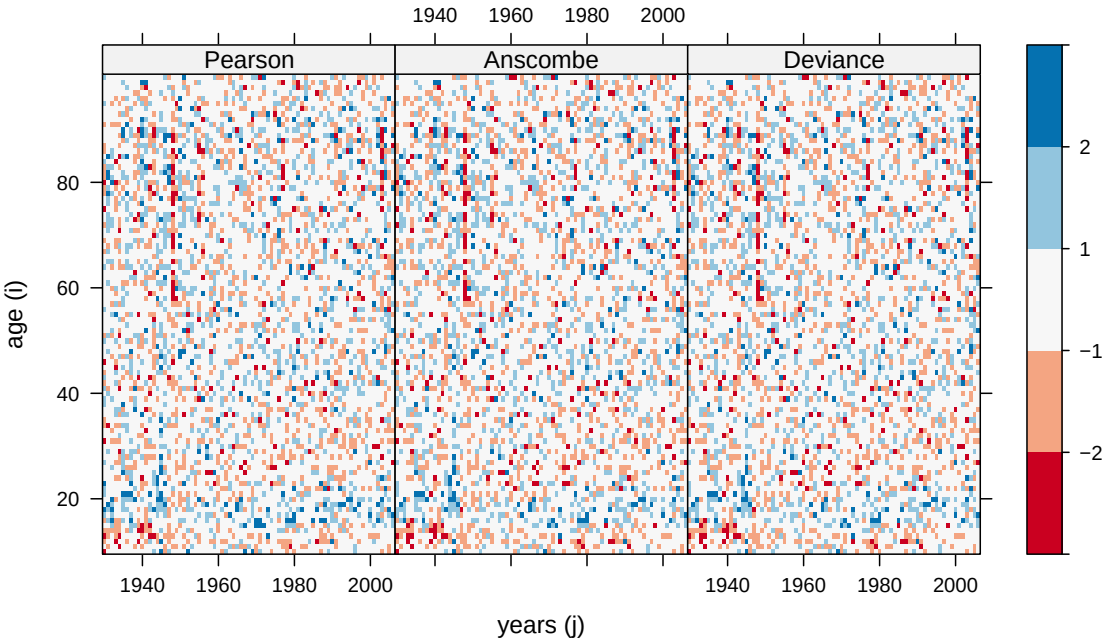

Figure 2.12: Pearson, Anscombe and deviance residuals over ages and years for death rates modeled with 2D smoothing with P-splines. Ages from 10 to 100. Denmark, females, 1930–2006.

The most common plot in the residual analysis is the plot of residuals versus fitted values.

Systematic features in this plot are of interest, e.g. residuals, that seem to increase or decrease in average magnitude with fitted values, might indicate nonconstant residual variance. In order to test the normality of the residuals, one can use a normal QQ plot. Alternatively, residuals can be plotted over the predictors of the model. This is done in Figure 2.12, in which residuals from a two-dimensional P-splines fit are plotted over age and time. This two-dimensional representation offers the opportunity to locate where the model cannot capture the actual data. Applied to mortality surfaces, this type of plot may reveal peculiar period shocks and cohort effects.

Figure 2.12 displays small discrepancies among the three residual functional forms. As expected, most residuals are around the value of zero without showing any particular systematic features. The only exceptions are the high values during World War II, during which the model overestimates the actual death rates (negative residuals depicted by red) and about age 20, where P-splines systematically underestimate death rates (positive residuals depicted by blue).

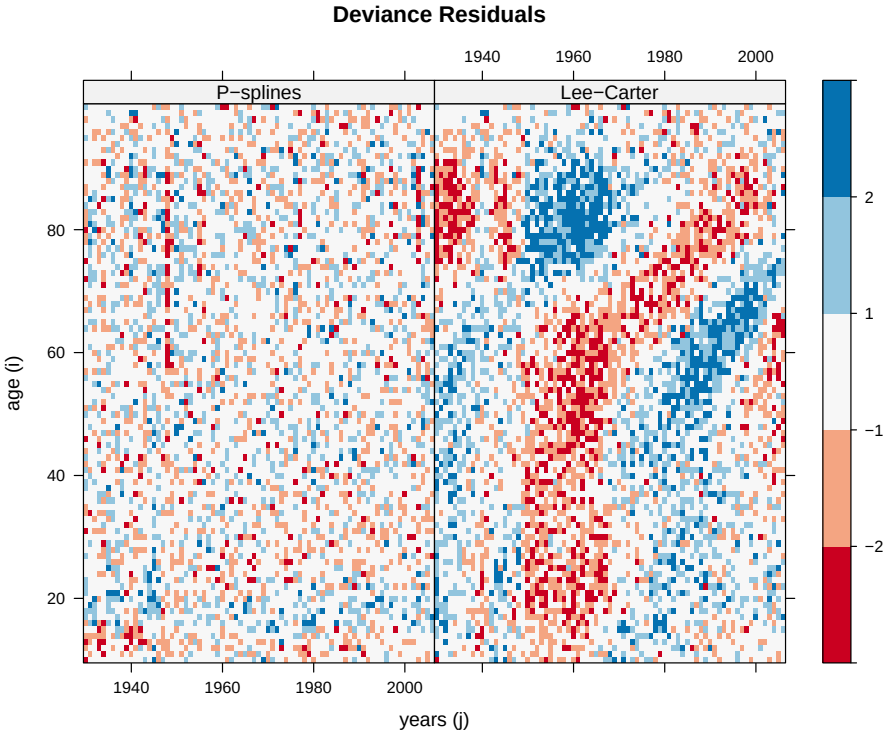

Figure 2.13: Deviance residuals over ages and years for death rates modeled with 2D smoothing with P-splines and Lee-Carter model. Ages from 10 to 100. Denmark, females, 1930–2006.

Residuals over age and time can also be used in comparing models over the same mortality surface. Figure 2.13 shows deviance residuals of both the Lee-Carter model and the two-dimensional P-splines smooth. It is easy to check that the P-spline approach can capture mortality developments more accurately than the LC model. Deviance residuals from the LC model clearly show systematic patterns. Moreover, we recall that the LC model, with its individual parameters for each age and year, estimates 257 parameters (see Section 1.5.3), whereas BIC selects a P-spline model with effective dimension equal to 137. Hence, a more parsimonious model such as the two-dimensional P-splines, can fit mortality surfaces better than the LC model.

2.3.2 Confidence Intervals

Two-dimensional smoothing with P-splines allows easy computation of the so-called hat matrix (see Section 2.1). Therefore, standard errors and confidence intervals can be easily obtained. In the particular case of Poisson-distributed data the approximating variance of $\boldsymbol{B}\hat{\boldsymbol{a}}$ is given by

$$\text{Var}(\boldsymbol{B}\hat{\boldsymbol{a}}) \approx \boldsymbol{B}(\boldsymbol{B}\hat{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}' \tag{2.29}$$

One justification of equation (2.29) is that the P-spline estimator of $\boldsymbol{Ba}$ can be derived from a Bayesian perspective (Lin and Zhang, 1999). Standard errors and confidence intervals for the fitted values are computed simultaneously from equation (2.29) and are included in Figure 2.14 for selected years and ages.

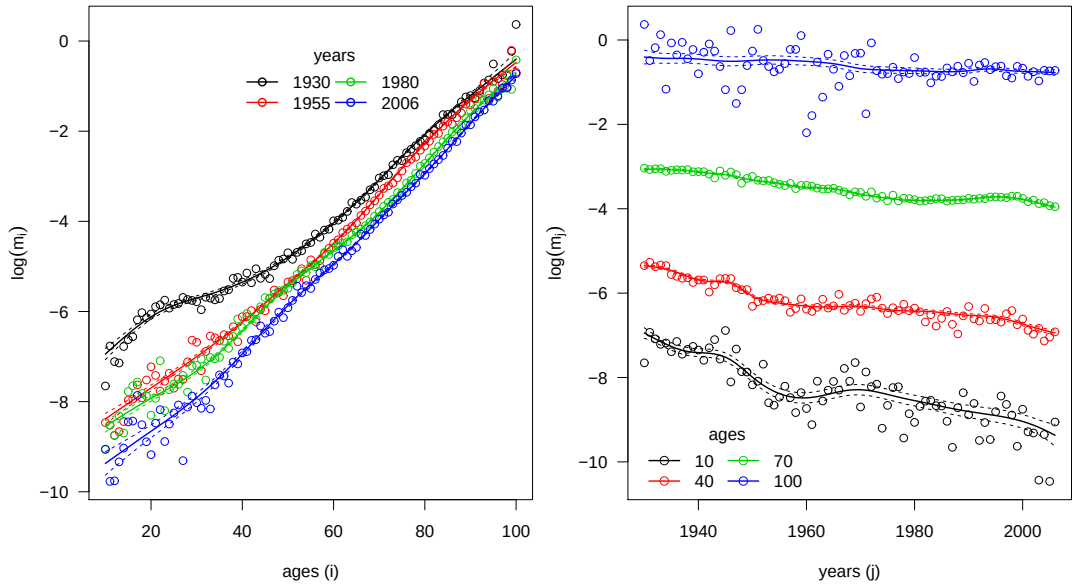

Figure 2.14: Actual and fitted death rates at selected years over ages (left panel) and selected ages over years (right panel), logarithmic scale. 2D smoothing with P-splines used for the estimation (solid lines) and the 99% confidence interval. Denmark, females.

The width of the confidence intervals indicates the level of uncertainty associated with the smooth surface (Figure 1.7). At first glance, the boundary of the confidence intervals in Figure 2.14 is extremely close to the fitted values. This is mainly due to the large number of death counts and to the large size of the population exposure. Nevertheless we see relatively larger confidence bends at old ages over the years. In this area, both exposures and deaths show moderate counts. Moreover a slightly wider confidence interval is observed at young ages in the last year (2006). This results from the small number of deaths along with a large exposure population.

An overview of this issue is given in Figure 2.15, in which we display the shaded contour maps of both exposures and deaths from the Danish population as well as standard errors from the fitted model over ages and years. The image of the standard errors is essentially the negative image of the deaths, whereas the number of exposures does not affect the standard error patterns much.

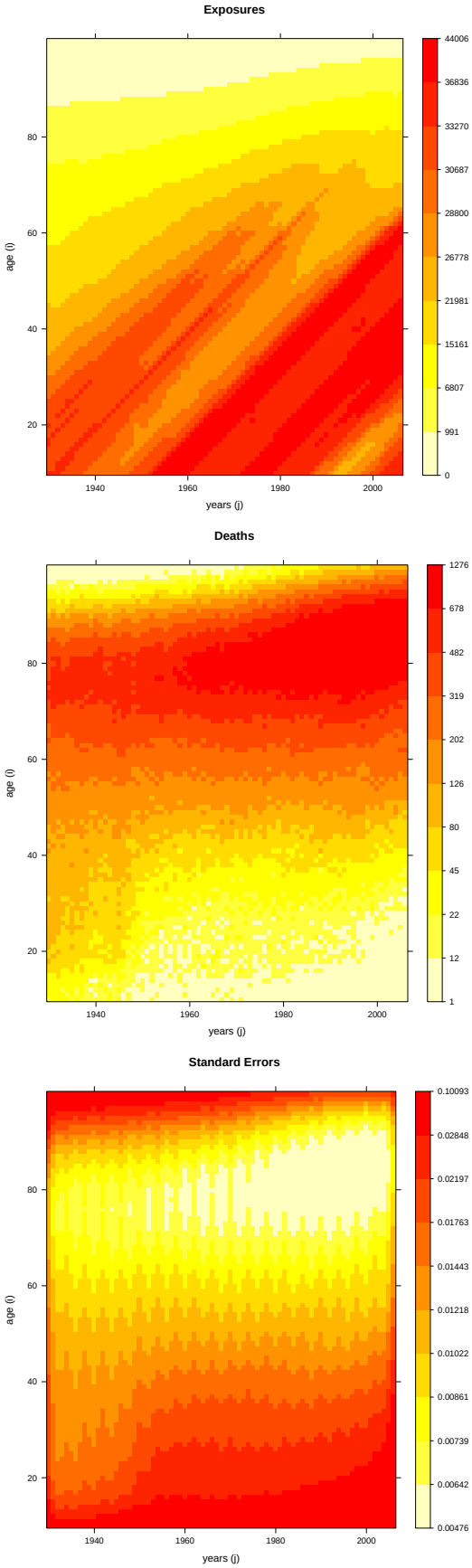

Figure 2.15: Actual exposures and deaths over ages and years as well as standard errors from two-dimensional smoothing with P-splines. Ages from 10 to 100. Denmark, females, 1930–2006.

2.4 P-splines in perspective

Smoothing methods are often overlooked in demographic analysis. We showed in Chapter 1 that commonly used demographic models either reduce the dimensionality of the data to quite few number of parameters or use overparameterized models such as the Lee-Carter model. Parsimonious and yet flexible models are the suitable tools for modeling the regular structure of human mortality development though.

In this chapter we presented a well-established smoothing methodology for fitting mortality over age and time: two-dimensional regression P-splines. This approach combines a relatively large number of B-splines with a roughness penalty. On one hand, B-splines provide enough flexibility to capture trends in the data. On the other hand, an additional penalty on neighbouring coefficients is used to ensure smoothness and reduces the number of parameters, leading to parsimonious models.

Furthermore, P-spline models can be easily embedded in a Poisson framework, therefore death counts can be straightforwardly fitted. An additional advantage of P-splines lies in the fact that different smoothing parameters are allowed over ages and time.

The foundation of P-splines on Generalized Linear Models allows easy computation of residuals and standard errors. We presented this facet of the model and plotted deviance residuals for mortality data in shaded contour maps over age and time. This procedure allows to locate where the model cannot capture the actual data and to understand additional demographic insights. However, both residuals and confidence intervals cannot properly capture the uncertainty in mortality data. The next chapter will present new advances to deal with this issue.

Given all the advantages, P-splines will be the standard smoothing methodology in the following chapters. In the next chapter two-dimensional P-spline will be compare with classic demographic models based on a suitable goodness-of-fit measures. Ideas and concepts such as the penalized likelihood will be employed in Chapter 4. Finally, in Chapter 5, the combination of B-splines and roughness penalty on the coefficient vector will be used to propose a new model for analyzing mortality data.

Chapter 3

Explained variation in models for mortality data

In the previous chapters we presented typical demographic approaches for modeling mortality data, and a smoothing method such as P-splines, which improves model fit, uses fewer degrees of freedom and is particularly flexible for modeling mortality surface. Nevertheless, demography can build on large samples, and this has implications for the statistical analysis of demographic data, including mortality studies. As Keyfitz (1966, p. 312) argues, in demographic studies, the "mere fact that over a period of a year the population is likely to be fifty or a hundredfold the deaths will result in a higher order of precision".

We have already noticed the consequences of this peculiarity in the residual analysis, and in construction of confidence intervals in a P-spline approach for mortality surface (see Section 2.3). Specific and uninformative outcomes are also evident in common goodness-of-fit (gof) measures. However, such measures are a necessary statistical tool for comparing mortality developments and, especially, to assess different models.

In this chapter, we first present the common measure of gof in the framework of GLMs. Extensions and adjustments of the classic R^2 are needed in models for non-Normal-distributed data. Section 3.2 introduces further extensions of gof measures in non-parametric settings such as the P-spline approach (see Chapter 2) and effective dimension of the smoother are considered when adjusting classic measures for GLMs. The presence of large counts in the mortality surface makes simple corrections essentially uninformative. In Section 3.3, we, thus, propose a new effective gof measure in models for mortality data: $R^2_{\text{(bi)lin}}$. The basic idea is to consider a null model which is specifically appropriate for mortality data. Particular emphasis will be given to the behavior of this measure in the Lee-Carter model and the P-spline approach. Simulation studies in one- and two-dimensional settings and applications for actual mortality surfaces are presented in Sections 3.4 and 3.5. A summary in Section 3.6 concludes the chapter.

3.1 Goodness-of-fit measure for Generalized Linear Models

The goodness-of-fit (gof) measures examine how well a statistical model fits a set of observations. Measures of gof typically summarize the discrepancy between observed values and the values

expected under the model in question. Such measures can be used in statistical hypothesis testing, or when investigating whether outcomes follow from a specified distribution.

In classic linear models, the most frequently used measure to express how well the model summaries features of the data is the well-known R^2. It is also called the "coefficient of determination" or the "percentage of variance explained". Its range is $0 \leq R^2 \leq 1$, with values closer to 1 indicating a better fit. It was developed to measure gof for linear regression models with homoscedastic errors. The concept of explained variation was generalized to heteroscedastic errors (Buse, 1973) and for logit, probit and tobit models (Veall and Zimmermann, 1996; Windmeijer, 1995).

One of the first studies on measuring explained variation in a GLM setting was undertaken by Cameron and Windmeijer (1997). They proposed an R^2 measure of goodness of fit for the exponential family. As a starting point, they defined a measure that took into account the proportional reduction in uncertainty due to the inclusion of regressors. Since in GLMs we have generalized the Normal distribution, the coefficient of determination should be interpreted as the fraction of uncertainty explained and no longer as a percentage of variance explained.

More specifically, Cameron and Windmeijer (1997) defined the R^2 for an exponential family regression model based on the Kullback-Leibler (KL) divergence (Kullback, 1959). A standard measure of the information from observations in a density $f(Y)$ is the expected information $E[\log(f(Y))]$ with the KL divergence measuring the discrepancy between two densities. Let $f_{\mu_1}(y)$ and $f_{\mu_2}(y)$ be two densities differing in mean μ only. The KL divergence is defined as

$$K(\mu_1, \mu_2) \equiv 2E_{\mu_1} \log \left[\frac{f_{\mu_1}(y)}{f_{\mu_2}(y)} \right] ,$$

where the factor 2 is multiplied for convenience and E_{μ_1} denotes that the expectation is taken with respect to the density $f_{\mu_1}(y)$. The KL measures how close μ_1 is to μ_2 and $K(\mu_1, \mu_2) \geq 0$ with equality iff $f_{\mu_1}(y) \equiv f_{\mu_2}(y)$.

If we define $f_y(y)$ as the density for which the mean is set to the realized y, the deviation of y from the mean μ is given by

$$K(y, \mu) \equiv 2E_y \log \left[\frac{f_y(y)}{f_\mu(y)} \right] = 2 \int f_y(y) \log \left[\frac{f_y(y)}{f_\mu(y)} \right] dy \, . \tag{3.1}$$

Hastie (1987) and Vos (1991) proved that if $f_y(y)$ is within the exponential family equation (3.1) is reduced to

$$K(y, \mu) = 2 \log \left[\frac{f_y(y)}{f_\mu(y)} \right] .$$

In an estimated regression model, with n individual estimated means $\hat{\mu}_i = \mu(x_i' \hat{\beta})$, the KL divergence between vectors $\boldsymbol{y}$ and $\boldsymbol{\mu}$ is equal to twice the difference between the maximum log-likelihood achievable, $l(\boldsymbol{y}, \boldsymbol{y})$, and the log-likelihood achieved by the fitted model $l(\hat{\boldsymbol{\mu}}, \boldsymbol{y})$

$$K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left[\log f_{y_i}(y_i) - \log f_{\hat{\mu}_i}(y_i) \right] = 2 \left[l(\boldsymbol{y}, \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) \right] \tag{3.2}$$

A particular case would be the constant only model where the fitted mean would be an n–vector $\hat{\boldsymbol{\mu}}_0$, and the KL divergence, $K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)$, can be interpreted as the information in the sample data on $\boldsymbol{y}$ potentially recoverable by inclusion of expectation with respect to the observed values $\boldsymbol{y}$.

Using the mentioned attributes of the KL divergence, Cameron and Windmeijer (1997) proposed an R^2 for the class of exponential family regression models

$$R^2_{KL} = 1 - \frac{K(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)} \tag{3.3}$$

given that $K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)$ is minimized when $\hat{\boldsymbol{\mu}}_0$ is the maximum likelihood estimate.

Since the expression for $K(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ in (3.2) is equivalent to the definition of the deviance (McCullagh and Nelder, 1989, p. 33), R^2_{KL} can be interpreted as being based on deviance residuals (cf. equation (2.27)). Therefore, R^2_{KL} is related to the analysis of deviance in the same way as the classic R^2 is related to the analysis of variance. We can re-write the R^2_{KL} as

$$R^2_{KL} = \frac{K(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_0)}{K(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0)}\,, \tag{3.4}$$

which reveals another interesting aspect of this measure: using the canonical link in exponential family models, R^2_{KL} measures the fraction of uncertainty explained by the fitted model, if uncertainty is quantified by the deviance.

As we have seen in Section 2.1, standard tools to quantify the discrepancy between observed and fitted values for Poisson models are deviance and the Pearson statistics. These concepts can be used to define different R^2 measures. One idea is to compare the sum of squared Pearson residuals for two different models: the fitted model and the most restricted model in which only an intercept is included, which is estimated by $\bar{y}$. Cameron and Windmeijer (1996) proposed

$$R^2_{PEA} = 1 - \frac{\sum_i^n (y_i - \hat{\mu}_i)^2/\hat{\mu}_i}{\sum_i^n (y_i - \bar{y})^2/\bar{y}} \tag{3.5}$$

The choice of $\bar{y}$ as weight in the denominator is a generalization for the Poisson case of the weighted R^2 proposed by Buse (1973).

Instead of using Pearson residuals, we can alternatively construct an R^2 measure based on deviance residuals (see Section 2.3.1). Let $\bar{y}$ be the predicted mean for a Poisson model with just an intercept: then, the deviance is $\mathrm{Dev}(\boldsymbol{y}, \bar{\boldsymbol{y}}) = 2\sum_i^n y_i \log(y_i/\bar{y})$. From this formulation, the deviance-R^2 for the Poisson model is

$$R^2_{DEV} = 1 - \frac{\sum_i^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}{\sum_i^n y_i \log(y_i/\bar{y})}. \tag{3.6}$$

For the canonical link, the term $\sum_i (y_i - \hat{\mu}_i)$ reduces to 0. Though equation (3.6) is equivalent to the R^2 based on the KL divergence in (3.3), we opt for a different subscript to emphasize that deviance residuals are the basic quantities in (3.6).

3.1.1 Adjustments according to the number of parameters

The R^2 measures we presented so far do not consider the number of covariates used in the regression models. Some studies underline this aspect in situations with small sample sizes relative to the number of covariates in the model (Mittlböck and Waldhör, 2000; Waldhör et al., 1998). In these cases, R^2 measures may be seriously inflated and may need to be adjusted. As we have seen in Section 2.1.3, in a non-parametric setting, effective dimensions are an equivalent concept to the number of covariates. Hence gof measures for non-parametric models should take into account the effective dimensions of the fitted model.

In a GLM context, Waldhör et al. (1998) proposed to correct both the deviance and Pearson R^2 in the following way:

$$\begin{aligned} R^2_{PEA,adj} &= 1 - \frac{(n-p-1)^{-1}\sum_i^n (y_i - \hat{\mu}_i)^2/\hat{\mu}_i}{(n-1)^{-1}\sum_i^n (y_i - \bar{y})^2/\bar{y}} \\ R^2_{DEV,adj} &= 1 - \frac{(n-p-1)^{-1}\sum_i^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}{(n-1)^{-1}\sum_i^n y_i \log(y_i/\bar{y})} \end{aligned}$$

where p is the number of estimated covariates additional to the intercept. This type of adjustment stems from the normal linear model and is appropriate when using the sum-of-squares approach to quantifying deviation. In Poisson regression models, this adjustment would be just an approximation.

Two adjusted R^2 measures for Poisson regression models based on deviance residuals are presented in Mittlböck and Waldhör (2000):

$$\begin{aligned} R^2_{DEV,adj1} &= 1 - \frac{l(\boldsymbol{y}) - [l(\hat{\boldsymbol{\mu}}) - p/2]}{l(\boldsymbol{y}) - l(\bar{\boldsymbol{y}})} \\ &= 1 - \frac{l(\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}) + p/2}{l(\boldsymbol{y}) - l(\bar{\boldsymbol{y}})} \\ R^2_{DEV,adj2} &= 1 - \frac{l(\boldsymbol{y}) - [l(\hat{\boldsymbol{\mu}}) - (p+1)/2]}{l(\boldsymbol{y}) - [l(\bar{\boldsymbol{y}}) - 1/2]} \\ &= 1 - \frac{l(\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}) + (p+1)/2}{l(\boldsymbol{y}) - l(\bar{\boldsymbol{y}}) + 1/2} \end{aligned} \tag{3.7}$$

It is easy to see how $R^2_{DEV,adj2}$ is always closer to zero than is $R^2_{DEV,adj1}$.

Mittlböck and Waldhör (2000) compared these two measures by simulation with different population values. They showed that $R^2_{DEV,adj1}$ performs remarkably well where the Poisson regression is based on a small sample and/or many covariates. Moreover, while the equations in (3.7) work well in a GLM setting, further extensions are needed in the case of smoothing models. These are presented in the next section.

3.2 Extending R^2 measures for smoothers

In the previous section, we presented R^2 measures for GLMs. Although the usage of likelihood ratio statistics in a smoothing context needs particular care (Ruppert et al., 2003), we follow the

same arguments and construct R^2 measures for Poisson-distributed mortality data fitted with P-splines.

Deaths counts are Poisson-distributed data and, therefore, we use a measure based on deviance residuals since Pierce and Schafer (1986) illustrated that deviance residuals are more suitable for this type of data. We replace in equations (3.7) the number of covariates p by the effective dimension ED:

$$R^2_{DEV,SMO,1} = 1 - \frac{\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\} + \frac{\text{ED}-1}{2}}{\sum_{i=1}^{n}\{y_i \log(y_i/\bar{y})\}} \tag{3.8}$$

and

$$R^2_{DEV,SMO,2} = 1 - \frac{\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\} + \frac{\text{ED}}{2}}{\sum_{i=1}^{n}\{y_i \log(y_i/\bar{y})\} + \frac{1}{2}} \tag{3.9}$$

where ED is estimated as in Section 2.1.3. Equations (3.8) and (3.9) can be naturally computed in both unidimensional and two-dimensional contexts. Distinct smoothing techniques can be compared using these measures, as long as they allow for easy computation of the used effective dimension of the model.

As a first example, the two R^2 measures are computed for Danish mortality data to which a surface is fitted with two-dimensional P-splines (see Figure 1.7 and 2.2). $R^2_{DEV,SMO,1}$ and $R^2_{DEV,SMO,2}$ are equal to 0.995722 and 0.995721, respectively. The difference between these values is minimal and it seems that the fraction of uncertainty explained by the model is close to 100%. Table 3.1 presents values of the R^2 measure given in (3.9) for different period and age ranges. For comparison, results of the R^2 measure from the Lee-Carter (LC) model are also given. All the values in Table 3.1 are in the range $[0.989038, 0.995721]$ with marginally smaller values for the LC model. Of course, the similarity between these outcomes does not reveal the important differences in explaining variation between the P-spline and the LC model. For a specific Danish mortality surface, such differences between the two approaches are evident from the fitted values in Figure 1.6 and 1.7 and from the residual pattern in Figure 2.12.

Danish Data			P-splines	Lee-Carter
females	1930–2006	10–100	0.995721	0.992671
males	1930–2006	10–100	0.995166	0.992583
females	1930–2006	50–100	0.993654	0.989038
males	1930–2006	50–100	0.993951	0.990885
females	1950–2006	50–100	0.994261	0.991518
males	1950–2006	50–100	0.994089	0.991426

Table 3.1: $R^2_{DEV,SMO,2}$ values for the Danish population by different period and age ranges, as well as models.

The presence of a large number of death counts in the mortality surface leads to rather small deviance residuals, which are the basic elements of these R^2 measures. Consequently, equations (3.8) and (3.9) will always generate figures significantly close to 1, which are essentially uninformative. An explanation for this drawback of equations (3.8) and (3.9) refers directly to the null model in the denominators of these measures. The model in the denominators incor-

porates only the intercept, and is a reasonable null model in a GLM framework. A different and peculiar null model is needed in models for mortality data. Specifically, it is appropriate to compare different mortality models to a "limit" model, which is nested in all the selected models, and which is more complex than a simple constant plane. For instance, the mortality surface for the Danish population is a 91×77 matrix. It is pleonastic and uninformative to check whether a P-spline model explains the variation in the data more than the overall mean of the matrix.

3.3 Alternative R^2 measures for mortality models

An alternative strategy for constructing gof measure for mortality data is to choose, as null model, the linear and bilinear models for unidimensional and two-dimensional models, respectively. That is, we consider a model in which age and time and possibly their interaction are the only regressors employed. This approach is appealing since both P-splines and the Lee-Carter model (see Section 1.5.3) can be seen as extensions of this proposed null model. This will allow comparison of different models relative to the linear or bilinear null model.

In the next sections, we show that a linear model is nested within a P-spline model or a Lee-Carter model, and therefore it is natural to use it as a null model. In Section 3.3.3 we will demonstrate that this decomposition can be used for the alternative R^2 measure.

3.3.1 P-splines with a transformed basis

In a P-spline setting, one can extract a linear component from the fitted trend, and fit the remaining variation by a smooth curve with a penalty that penalizes departures from zero. An early reference about this topic can be found in Green (1985). Verbyla et al. (1999) and Currie et al. (2006) have discussed this idea, too.

As demonstrated in Section 2.1.1, we can write a P-splines model in the following way:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \epsilon, \qquad \epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{3.10}$$

where $\boldsymbol{B} = \boldsymbol{B}(x)$, $n \times k$ is the regression matrix of the B-splines and an additional difference penalty $\boldsymbol{P}$ on the coefficients $\boldsymbol{a}$ is used to enforce smoothness. Given these components, the smoothed function is found by minimizing

$$S^* = \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}\|^2 + \boldsymbol{P}.$$

where $\boldsymbol{P} = \lambda \boldsymbol{D}_d' \boldsymbol{D}_d$ and $\boldsymbol{D}_d$ is the difference matrix acting on the coefficients $\boldsymbol{a}$ (see Section 2.1).

A linear or bilinear model can be seen as a nested model in the more general P-spline framework, and in the following, we demonstrate explicitly this association. Specifically, we will present how a P-spline model can be decomposed in two unique and distinct components, one of which is the linear model.

In particular, we show how to represent equation (3.10) in the following alternative way:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \epsilon, \qquad \boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}),\ \epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{3.11}$$

where $\boldsymbol{G}$ is a covariance matrix which depends on λ and in the following, we assume the simple structure $\boldsymbol{G} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}$, with unknown variance $\sigma^2_{\boldsymbol{\alpha}}$. In this way, we separate the fixed part, which does not depend on the smoothing parameter, and the remaining variation which will be smoothed via P-splines. We show how the fixed part can be a simple linear (bilinear) model in a unidimensional (two-dimensional) setting.

It is worth pointing out that this is the common representation of P-splines as mixed models (Currie and Durban, 2002; Currie et al., 2006; Durban et al., 2006) and that the fixed-effects model-matrix $\boldsymbol{X}$ is nested in the full model-matrix $\boldsymbol{B}$.

We assume that a second order penalty is used, i.e. $d = 2$. The aim is to find the unique matrix $\boldsymbol{T}$ such that

$$\boldsymbol{BT} \equiv [\boldsymbol{X} : \boldsymbol{Z}] \Rightarrow \boldsymbol{Ba} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} \tag{3.12}$$

The idea is to use the Singular Value Decomposition (SVD, Good, 1969) of the penalty $\boldsymbol{P}$ to partition the difference penalty into a null penalty (for the fixed part) and a diagonal penalty (for the random part).

The SVD of the square matrix $\boldsymbol{D'D}$ can be written as

$$\boldsymbol{D'D} = \boldsymbol{V\Lambda V'},$$

where $\boldsymbol{V}$ can be decomposed into two matrices: $\boldsymbol{V}_n$ and $\boldsymbol{V}_s$. The former is the part of the matrix $\boldsymbol{V}$ corresponding to the zero eigenvalues of $\boldsymbol{D'D}$ (fixed part). Since we are using a second order penalty we will have only two zero eigenvalues. The matrix $\boldsymbol{V}_s$ corresponds to the nonzero eigenvalues (random part). Therefore the fixed part would be:

$$\boldsymbol{X} = \boldsymbol{B}\,\boldsymbol{V}_n \tag{3.13}$$

For the random part, we consider the diagonal matrix $\boldsymbol{\Lambda}$ where we remove the elements corresponding to the fixed part $\boldsymbol{X}$:

$$\boldsymbol{\Lambda} = \left[\begin{array}{c|c} \tilde{\boldsymbol{\Sigma}} & \\ \hline & \mathbf{0}_{2\times 2} \end{array}\right]$$

The new diagonal matrix $\tilde{\boldsymbol{\Sigma}}$ contains the non-zero eigenvalue: therefore, the random-effects part can be written as:

$$\boldsymbol{Z} = \boldsymbol{B}\,\boldsymbol{V}_s\,\tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \tag{3.14}$$

The mentioned matrix $\boldsymbol{T}$ will then be:

$$\boldsymbol{T} = \left[\boldsymbol{V}_n \,:\, \boldsymbol{V}_s\,\tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\right]$$

and consequently $\boldsymbol{TB} = [\boldsymbol{X} : \boldsymbol{Z}]$ giving the parameterization in equation (3.12) where

$$\boldsymbol{\beta} = \boldsymbol{V}_n'\boldsymbol{a} \qquad \text{and} \qquad \boldsymbol{\alpha} = [\boldsymbol{V}_s\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}]\boldsymbol{a}.$$

and the penalty term is given by

$$\boldsymbol{a'D'Da} \rightarrow \boldsymbol{\alpha'\alpha}$$

In this way quadratic and cubic fixed-effects can be chosen with $d = 3$ and $d = 4$, respectively (Verbyla et al., 1999, p. 308). This representation can be generalized in a Poisson case in a straightforward manner with the additional weight matrix and link function as in Section 2.1. In a unidimensional setting, the fixed part for mortality data which can be used as null model for constructing gof measure will be a simple linear term such that

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \tag{3.15}$$

where the second column of $\boldsymbol{X}$ will be either the age or year values.

In a two-dimensional setting, the previous considerations can be easily generalized as in the P-spline approach (see Section 2.2). We now have $\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a$ with penalty $\boldsymbol{P}$ given by

$$\boldsymbol{P} = \lambda_a \boldsymbol{I}_{c_y} \otimes \boldsymbol{D}_a' \boldsymbol{D}_a + \lambda_y \boldsymbol{D}_y' \boldsymbol{D}_y \otimes \boldsymbol{I}_{c_a}. \tag{3.16}$$

Taking the SVD of $\boldsymbol{D}_a' \boldsymbol{D}_a$ we obtain $\boldsymbol{V}_a \boldsymbol{\Sigma}_a \boldsymbol{V}_a'$ and partitioning the matrix

$$\boldsymbol{V}_a = [\boldsymbol{V}_{as} : \boldsymbol{V}_{an}]$$

where $\boldsymbol{V}_{as}$ corresponds to the non-zero eigenvalues and $\boldsymbol{V}_{an}$ to the zero eigenvalues.

Assuming a second order penalty in both dimensions, $\boldsymbol{\Sigma}_a$ has two zero eigenvalues and $\boldsymbol{V}_{an}$ has two columns. Let $\boldsymbol{\Sigma}_{as}$ contain the positive eigenvalues of $\boldsymbol{\Sigma}_a$. In the same way we decompose $\boldsymbol{D}_y' \boldsymbol{D}_y$ obtaining $\boldsymbol{V}_y = [\boldsymbol{V}_{ys} : \boldsymbol{V}_{yn}]$ and $\boldsymbol{\Sigma}_{ys}$.

Then we have the fixed part:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{B}(\boldsymbol{V}_{yn} \otimes \boldsymbol{V}_{an}) \\ &= \boldsymbol{B}_y \boldsymbol{V}_{yn} \otimes \boldsymbol{B}_a \boldsymbol{V}_{an} \\ &= \boldsymbol{X}_y \otimes \boldsymbol{X}_a. \end{aligned} \tag{3.17}$$

And the random part is given by

$$\begin{aligned} \boldsymbol{Z} &= \boldsymbol{B}(\boldsymbol{V}_{ys} \otimes \boldsymbol{V}_{as}) \tilde{\boldsymbol{\Sigma}}^{-1/2} \\ &= (\boldsymbol{B}_y \boldsymbol{V}_{ys} \otimes \boldsymbol{B}_a \boldsymbol{V}_{as}) \tilde{\boldsymbol{\Sigma}}^{-1/2} \end{aligned} \tag{3.18}$$

where the diagonal matrix $\tilde{\boldsymbol{\Sigma}}$ is a block-diagonal containing the non-zero eigenvalues in both dimensions.

The new basis $\boldsymbol{T}$ is given by the combination of equations (3.17) and (3.18):

$$\boldsymbol{T} = [\boldsymbol{V}_{yn} \otimes \boldsymbol{V}_{an} : \boldsymbol{V}_{ys} \otimes \boldsymbol{V}_{an} : \boldsymbol{V}_{yn} \otimes \boldsymbol{V}_{as} : \boldsymbol{V}_{ys} \otimes \boldsymbol{V}_{as}]$$

We can prove that $\boldsymbol{T}$ is orthogonal, so for the two-dimensional case in (3.12) we have

$$\begin{aligned}\boldsymbol{\beta} &= (\boldsymbol{V}_{yn} \otimes \boldsymbol{V}_{an})'\boldsymbol{a} \\ \boldsymbol{\alpha} &= (\boldsymbol{V}_{ys} \otimes \boldsymbol{V}_{an} : \boldsymbol{V}_{yn} \otimes \boldsymbol{V}_{as} : \boldsymbol{V}_{ys} \otimes \boldsymbol{V}_{as})'\boldsymbol{a}.\end{aligned}$$

The penalty is then given by

$$\boldsymbol{a}'\boldsymbol{P}\boldsymbol{a} = \boldsymbol{\omega}'\boldsymbol{T}'\boldsymbol{P}\boldsymbol{T}\boldsymbol{\omega}$$

where $\boldsymbol{\omega}' = [\boldsymbol{\beta}' : \boldsymbol{\alpha}']$.

The fixed part for our null model in a two-dimensional case for mortality surfaces is then given by

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & a_1 & y_1 & a_1 \cdot y_1 \\ 1 & a_2 & y_1 & a_2 \cdot y_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_1 & a_m \cdot y_1 \\ 1 & a_1 & y_2 & a_1 \cdot y_2 \\ 1 & a_2 & y_2 & a_2 \cdot y_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_2 & a_m \cdot y_2 \\ 1 & a_1 & y_3 & a_1 \cdot y_3 \\ 1 & a_2 & y_3 & a_2 \cdot y_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_3 & a_m \cdot y_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_1 & y_n & a_1 \cdot y_n \\ 1 & a_2 & y_n & a_2 \cdot y_n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_m & y_n & a_m \cdot y_n \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \tag{3.19}$$

where $a_i, i = 1, \ldots, m$ and $y_j, j = 1, \ldots, n$ are age and year values, respectively. In this way we have as fixed part, and, consequently as null model, a bilinear surface in which age and time interact. The dimension of the model is equal to four, i.e. the number of columns of $\boldsymbol{X}$. The linear model can be easily fitted using a two-dimensional P-spline framework by choosing large smoothing parameters for both age and year. In the example we considered $\lambda_a = \lambda_y = 10^8$ worked well and lead to an effective dimension of about four.

3.3.2 The Lee-Carter as a simple bilinear model

In a two-dimensional setting, the Lee-Carter (LC) model is widely used in modeling mortality surface and it is a commonly used model for mortality development (see Section 1.5.3, page 11). Therefore, it is useful to also apply alternative gof measures for mortality data to this model. In this section, we show how the basic bilinear structure is nested in the LC model, too.

The LC model (page 13) is given by:

$$Y_{ij} \sim \text{Poisson}\left(E_{ij} \cdot \exp(\alpha_i + \beta_i \cdot \gamma_j)\right)$$

where α_i, β_i and γ_j are vectors of parameters that have to be estimated. Using the canonical link function for Poisson-distributed data, the linear predictor of the LC model is given by

$$\eta_{ij} = \alpha_i + \beta_i \cdot \gamma_j. \tag{3.20}$$

It can be proved that equation (3.20) is a general case of the fixed part of the model in (3.12) in the two-dimensional case, where the linear predictor is given in equation (3.19). We let the Lee-Carter vectors of parameters, α_i, β_i and γ_j, vary linearly over ages and years, that is

$$\begin{aligned} \alpha_i &= \theta_1 + \theta_2 \cdot a_i \\ \beta_i &= \theta_3 + \theta_4 \cdot a_i \\ \gamma_j &= \theta_5 + \theta_6 \cdot y_j. \end{aligned} \tag{3.21}$$

The linear predictor in (3.20) then becomes:

$$\begin{aligned} \eta_{ij} &= \theta_1 + \theta_2 a_i + (\theta_3 + \theta_4 a_i)(\theta_5 + \theta_6 y_j) \\ &= (\theta_1 + \theta_3\theta_5) + (\theta_2 + \theta_4\theta_5)a_i + \theta_3\theta_6 y_j + \theta_4\theta_6 a_i y_j\,, \end{aligned}$$

which is equivalent to linear part of the mixed model representation of P-splines models in equation (3.19) if

$$\begin{aligned} \beta_1 &= \theta_1 + \theta_3\theta_5 \\ \beta_2 &= \theta_2 + \theta_4\theta_5 \\ \beta_3 &= \theta_3\theta_6 \\ \beta_3 &= \theta_4\theta_6. \end{aligned}$$

3.3.3 $R^2_{(\text{bi})\text{lin}}$: a goodness-of-fit measure for mortality data

In Sections 3.3.1 and 3.3.2, we showed that both the two-dimensional P-spline model and the Lee-Carter model can be seen as extensions of a bilinear surface, where ages and years interact. Such a parsimonious surface can be used as a null model to compare the explained variability from more sophisticated models in an appropriate way. That is, we replace the constant surface as a null model in the denominator in (3.9), by either an estimated linear or bilinear model as given in (3.15) and (3.19).

We define by $\hat{\mu}_i^1$ and $\hat{\mu}_i^0$ the estimated values by the fitted model and the null model, respectively. In a similar fashion, ED^1 and ED^0 denote the effective dimension of the two models. Recalling explicitly equation (3.9), we propose as appropriated gof measure for mortality models

the following equation:

$$R^2_{\text{(bi)lin}} = 1 - \frac{\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i^1) - (y_i - \hat{\mu}_i^1)\} + \frac{\text{ED}^1}{2}}{\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i^0) - (y_i - \hat{\mu}_i^0)\} + \frac{\text{ED}^0}{2}} \tag{3.22}$$

where n denotes the total number of measurement values in the data. As mentioned, the null model is defined by the linear predictor in (3.19) or (3.15). The variation explained by the fitted model is now compared to the (bi)linear model and therefore we use (bi)lin as subscription in the measure.

Equation (3.22) can be alternatively written as

$$R^2_{\text{(bi)lin}} = 1 - \frac{\text{Dev}^1(\boldsymbol{y}; \boldsymbol{a}^1, \lambda) + \frac{\text{ED}^1(\boldsymbol{a}^1, \lambda)}{2}}{\text{Dev}^0(\boldsymbol{y}; \boldsymbol{a}^0) + \frac{\text{ED}^0(\boldsymbol{a}^0)}{2}} \tag{3.23}$$

where $\boldsymbol{a}$ are the coefficients of the model and λ is the smoothing parameter, for smoothing models. Again, the superscripts 0 and 1 denote quantities computed from the null and fitted model, respectively.

As in the other R^2 measures (see Section 3.1), values of (3.23) closer to 1 indicate a better fit compared to the (bi)linear null model. Moreover, $\frac{\text{ED}^0(\boldsymbol{a}^0)}{2}$ is equal to 1 in a unidimensional setting and to 2 in a two-dimensional setting, and it does not depend on the smoothing parameter λ in a smoothing setting.

Relations between $R^2_{\text{(bi)lin}}$ and information criteria

Actual associations between our $R^2_{\text{(bi)lin}}$ and information criteria for a smoother can shed additional light on the meaning and implications of the proposed R^2 measure. We have already presented several information criteria for selection of smoothing parameters for different data (in Section 2.1). Here we focus on the commonly used criteria in the case of Poisson data: Akaike's Information Criterion and Bayesian Information Criterion. Let's recall their formulas:

$$\begin{aligned} \text{AIC}(\lambda) &= \text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda) + 2 \cdot \text{ED}(\boldsymbol{a}, \lambda) \\ \text{BIC}(\lambda) &= \text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda) + \ln(n) \cdot \text{ED}(\boldsymbol{a}, \lambda)\,, \end{aligned}$$

and, consequently, the deviance can be written equivalently as:

$$\begin{aligned} \text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda) &= \text{AIC}(\lambda) - 2 \cdot \text{ED}(\boldsymbol{a}, \lambda) \\ \text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda) &= \text{BIC}(\lambda) - \ln(n) \cdot \text{ED}(\boldsymbol{a}, \lambda)\,. \end{aligned} \tag{3.24}$$

One can show that AIC and BIC are linked with $R^2_{\text{(bi)lin}}$. Substituting the first equation of (3.24) in the measure presented in (3.23), we obtain:

$$
\begin{aligned}
R^2_{\text{(bi)lin}} &= 1 - \frac{\text{AIC}^1 - 2 \cdot \text{ED}^1 + \frac{\text{ED}^1}{2}}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\
&= 1 + \frac{\frac{3}{2} \cdot \text{ED}^1 + \text{AIC}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\
&= \left[1 + \frac{\frac{3}{2} \cdot \text{ED}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}}\right] - \left[\frac{1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}}\right] \cdot \text{AIC}^1
\end{aligned} \tag{3.25}
$$

which shows the relation between $R^2_{\text{(bi)lin}}$ and AIC. In a similar fashion, substituting the second equation of (3.24) in (3.23), we obtain the relationship between $R^2_{\text{(bi)lin}}$ and BIC:

$$
\begin{aligned}
R^2_{\text{(bi)lin}} &= 1 - \frac{\text{BIC}^1 - \ln(n) \cdot \text{ED}^1 + \frac{\text{ED}^1}{2}}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\
&= 1 + \frac{\left[\ln(n) - \frac{1}{2}\right] \cdot \text{ED}^1 + \text{BIC}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\
&= 1 + \frac{\left[\ln(n) - \ln\left(e^{\frac{1}{2}}\right)\right] \cdot \text{ED}^1 + \text{BIC}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}} \\
&= \left[1 + \frac{\ln\left(\frac{n}{\sqrt{e}}\right) \cdot \text{ED}^1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}}\right] - \left[\frac{1}{\text{Dev}^0 + \frac{\text{ED}^0}{2}}\right] \cdot \text{BIC}^1
\end{aligned} \tag{3.26}
$$

For simplicity of notation, on the right side of (3.25) and (3.26) we have dropped the arguments in the brackets.

From equations (3.25) and (3.26), the $R^2_{\text{(bi)lin}}$ is a linear transformation of both AIC and BIC. It is noteworthy that the slope of this transformation depends merely on the null model. As a result, if we would use $R^2_{\text{(bi)lin}}$ as a criterion for smoothing parameter selection, i.e. maximizing $R^2_{\text{(bi)lin}}$, we would obtain an optimal value that is different from the one obtained by minimizing the AIC and BIC.

The important point to note here is the presence of the deviance of the null model, Dev^0, in the second terms of the intercepts in equations (3.25) and (3.26). Especially in a two-dimensional setting and with mortality data, Dev^0 can be substantially higher than any quantity in the numerator, leading to an intercept close to 1. Therefore, in the presence of larger Dev^0, the profiles $R^2_{\text{(bi)lin}}$ and AIC (or BIC) will be more and more similar over a grid of smoothing parameters that differ only in sign.

Furthermore, fitted values picked by minimizing AIC will always result in a larger $R^2_{\text{(bi)lin}}$ with respect to those picked by minimizing BIC, especially for a large mortality surface (see the different penalization produced by AIC and BIC in Section 2.1.4).

3.4 Simulation studies

In this section, we present different simulation studies which demonstrate the performance of the proposed $R^2_{(\text{bi})\text{lin}}$. Its features will be considered in both unidimensional and two-dimensional contexts. Simulation settings are chosen such that they resemble mortality data, based on Poisson data and different sample sizes.

3.4.1 The unidimensional case

Though the proposed measure $R^2_{(\text{bi})\text{lin}}$ reveals its capability to identify how well a mortality model fits in a two-dimensional setting, in this section we will illustrate the performances of $R^2_{(\text{bi})\text{lin}}$ over a single variable only. A univariate P-spline model is fit based on smoothing parameters selected by BIC (see Section 2.1, page 17).

The $R^2_{(\text{bi})\text{lin}}$ measure is constructed from a given fitted model and a null model. In a unidimensional setting, the latter is represented in equation (3.15). Equation (3.15) is a simple linear predictor where the time-axis is the only covariate. As mentioned above, we will fit this null model by simply applying a P-spline approach with a sufficiently large smoothing parameter. Consequently ED^0 will always be equal to 2.

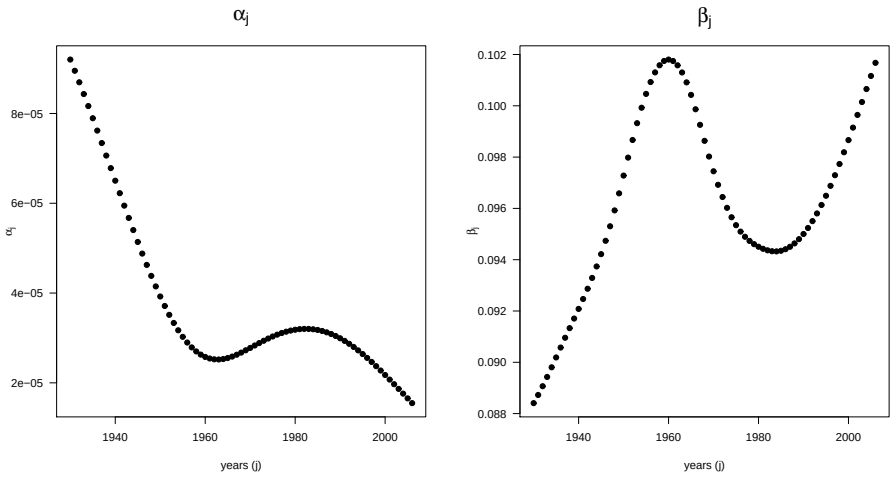

Figure 3.1: Gompertz parameters, α_j and β_j, over time j used in the simulation setting, cf. equations (3.27) and (3.28)

Death counts were simulated from a Poisson distribution with rates following a Gompertz distribution (Section 1.4, page 8). In particular, we simulated mortality surfaces from the following setting

$$\begin{aligned} Y_{ij} \sim \text{Poisson}\left(E_{ij} \cdot \exp(\eta_{ij})\right) \qquad i &= 30, \ldots, 100 \\ j &= 1930, \ldots, 2006\,, \end{aligned} \tag{3.27}$$

where the linear predictor varies over time j:

$$\eta_{ij} = \ln(\alpha_j) + \beta_j \cdot i \tag{3.28}$$

and i are the ages, $30, \dots 100$. The values of the parameters α_j and β_j over year j are shown in Figure 3.1 and they are chosen to mimic a realistic scenario.

In order to understand the role of the sample size in the outcomes of $R^2_{\text{(bi)lin}}$, we simply modify the values into the matrix of exposures E_{ij}. Specifically, we designed two different matrices in which $E_{ij} = 5,000$ and $E_{ij} = 25,000$ for all $i = 30, \dots, 100$ and $j = 1930, \dots, 2006$. In this way we can generate two different mortality surfaces in which the true model is the same, whereas the variability is different.

We will pick two particular ages only ($i = 40$ and $i = 80$) and analyze the performance of $R^2_{\text{(bi)lin}}$ over time, $j = 1930, \dots, 2006$. We have four series of death rates (2 ages and 2 exposures) which will be smoothed using P-spline methodology with 18 cubic B-spline bases. The proposed $R^2_{\text{(bi)lin}}$ is then computed for each fitted model.

We repeated the procedure 1,000 times. Figure 3.2 shows the outcomes of a single simulation at the given ages $i = 40$ and $i = 80$ for the different exposure matrices. True, fitted and null models are plotted. Values for $R^2_{\text{(bi)lin}}$ and $R^2_{DEV,SMO,2}$ are also presented.

Table 3.2 presents the median values for $R^2_{\text{(bi)lin}}$ and $R^2_{DEV,SMO,2}$ as well as median values for deviance, effective dimensions and selected smoothing parameter from the 1,000 simulations by different exposures and ages. An overview of the distributions of these parameters is given in Figure 3.3.

	Median value of:				
Simulated data	$R^2_{\text{(bi)lin}}$	$R^2_{DEV,SMO,2}$	$\text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda)$	$\text{ED}(\boldsymbol{a}, \lambda)$	λ
$i = 40$, $E_{ij} = 25,000$	0.325	0.822	74.338	4.441	1000.000
$i = 40$, $E_{ij} = 5,000$	0.061	0.477	77.424	2.778	3162.278
$i = 80$, $E_{ij} = 25,000$	0.860	0.991	68.655	9.633	630.957
$i = 80$, $E_{ij} = 5,000$	0.545	0.957	72.280	7.280	630.957

Table 3.2: Median values of $R^2_{\text{(bi)lin}}$, $R^2_{DEV,SMO,2}$, $\text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda)$, $\text{ED}(\boldsymbol{a}, \lambda)$ and λ from the 1,000 simulations at age 40 and 80 over years $j = 1930, \dots, 2006$. Different exposure matrices are used, cf. equations (3.27) and (3.28).

As expected, $R^2_{\text{(bi)lin}}$ always is smaller than $R^2_{DEV,SMO,2}$. $R^2_{\text{(bi)lin}}$ measures how much more variation is captured by the model relative to the null linear model. The outcomes of $R^2_{DEV,SMO,2}$ are close to 1 (0.991 and 0.956), especially at age 80, due to the large number of deaths at this age.

Moreover, Figure 3.3 shows that $R^2_{\text{(bi)lin}}$ differs more strongly between the setting than $R^2_{DEV,SMO,2}$. In particular, results from $R^2_{DEV,SMO,2}$ at age $i = 80$ are all very close to 1, i.e. all models capture almost 100% of the variation.

The smoothing parameter at age 40 with $E_{ij} = 5,000$ is on average considerably larger, leading to fitted values often similar to the linear null model. This might also be due to the fact that the variability in the data is larger for smaller exposure. In these cases, the $R^2_{\text{(bi)lin}}$ will generally be close to 0, i.e. the fitted model does not capture more variability than the linear null model.

It is a easy to see that the values of $R^2_{\text{(bi)lin}}$ are mainly influenced by the difference in effective dimensions in the fitted models. On the other hand, the deviance does not show substantial

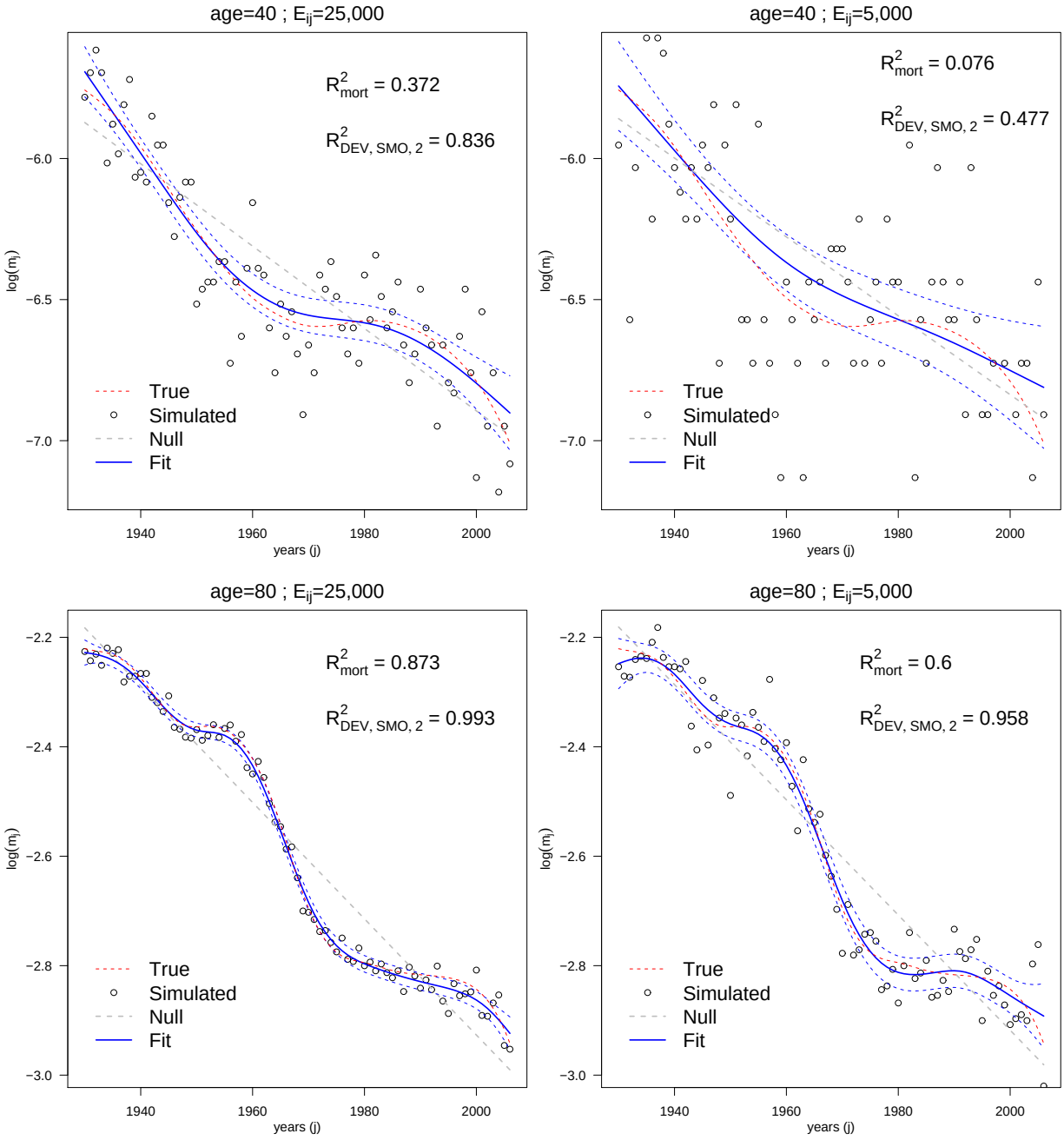

Figure 3.2: True, simulated and fitted deaths rates (with 95% confidence interval) along with the null model at age 40 and 80 over years $j = 1930, \ldots, 2006$, logarithmic scale. P-spline approach is used to fit the data, and BIC for selecting the smoothing parameters.

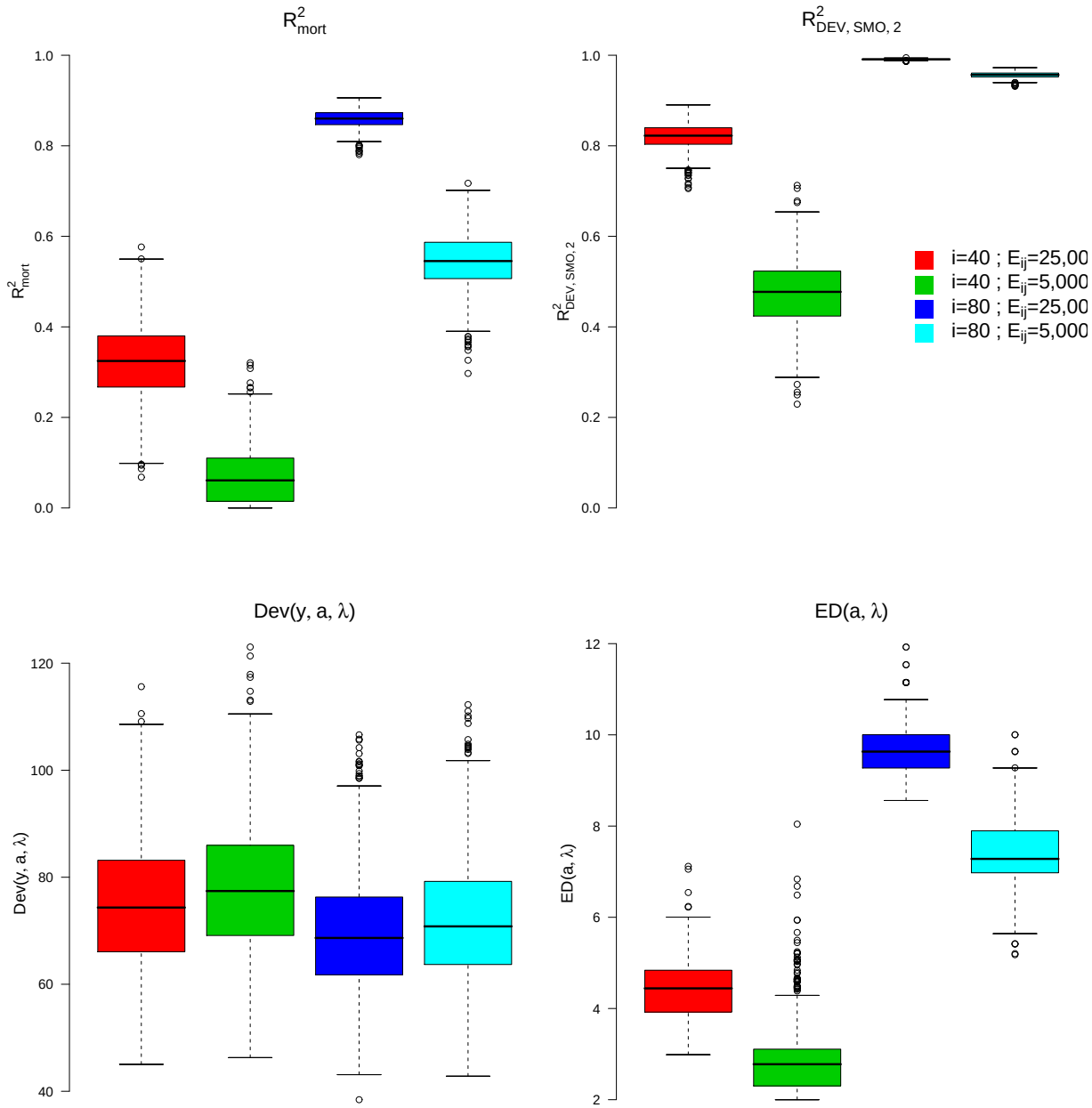

Figure 3.3: Summary of 1,000 simulations. Box-plots of $R^2_{(\text{bi})\text{lin}}$, $R^2_{DEV,SMO,2}$, $\text{Dev}(\boldsymbol{y};\boldsymbol{a},\lambda)$ and $\text{ED}(\boldsymbol{a},\lambda)$ for ages $i = 40$ and $i = 80$ and different exposure matrices, cf. equations (3.27) and (3.28).

differences between the four settings (see Figure 3.3, bottom-left panel).

In conclusion, we consider it is often more meaningful and appealing to perceive how much our fitted model improves with respect to a known model, instead of to the overall mean.

3.4.2 The two-dimensional case

In this section, we will present results from both $R^2_{\text{(bi)lin}}$ and $R^2_{DEV,SMO,2}$ in a simulated two-dimensional setting. As explained in Sections 3.3.1 and 3.3.2, both the Lee-Carter model and the two-dimensional regression P-splines can be considered as extensions of the simple bilinear model over age and time, as specified in equation (3.19). Such a bilinear model will be used in this simulation setting as null model for the $R^2_{\text{(bi)lin}}$ measure.

Our study in a two-dimensional case employs equations (3.27) and (3.28) for simulating 1,000 mortality surfaces which follow Gompertz distribution with parameters α_j and β_j varying over time as display in Figure 3.1. Figure 3.4 presents an example of such simulation in which the true mortality surface is accompanied by possible simulated surfaces with different exposure matrices.

These mortality surfaces are then fitted by two-dimensional regression P-splines and the LC model (see Sections 2.2 and 1.5.3). In particular, we selected the smoothing parameters by BIC in the P-spline approach and we followed the Poisson likelihood approach given by Brouhns et al. (2002) for fitting the LC model. Finally, both $R^2_{\text{(bi)lin}}$ and $R^2_{DEV,SMO,2}$ are computed for each of the 1,000 simulations.

Table 3.3 shows the median values from the 1,000 simulations of both $R^2_{\text{(bi)lin}}$ and $R^2_{DEV,SMO,2}$, as well as the median deviances from both the P-spline and LC approach. Two-dimensional regression P-splines allow distinct smoothing parameters for each surface and consequently different effective dimensions (Table 3.3 presents also the median values of the effective dimensions). Note that the LC model always employs $2 \cdot m + n - 2 = 215$ parameters, which will have an important impact on $R^2_{\text{(bi)lin}}$ and $R^2_{DEV,SMO,2}$.

$R^2_{\text{(bi)lin}}$ is substantially higher for the two-dimensional P-spline approach than for the LC model (0.931 vs. 0.750 and 0.734 vs. 0.591 for the two simulation settings). We might conclude that the LC model performs much worse than two-dimensional P-spline regression on the given simulation setting. This difference is more evident than it appears looking directly at $R^2_{DEV,SMO,2}$, in which all the values are close to 1.

Regarding model comparison, one could check the differences in the deviances between the two approaches (5407 vs. 19656 and 5444 vs. 8134). Again in this case the discrepancy between simulated and fitted values is much larger in the LC model. This is due to the comparatively rigid structure of the LC model. Though the effective dimensions in the LC model is substantially larger, as comparison to the P-spline model, the LC model is incapable of capturing variability in the data better.

It is worth pointing out that $R^2_{\text{(bi)lin}}$ combines both deviance and effective dimensions of a fitted model in a single value, which reveals straightforwardly the gof of the model. Also in this measure, the sample size plays an important role in the results.

The differences between the LC model and the P-spline approach are even clearer in Figure 3.5, where two particular ages from the two different simulated mortality surfaces are por-

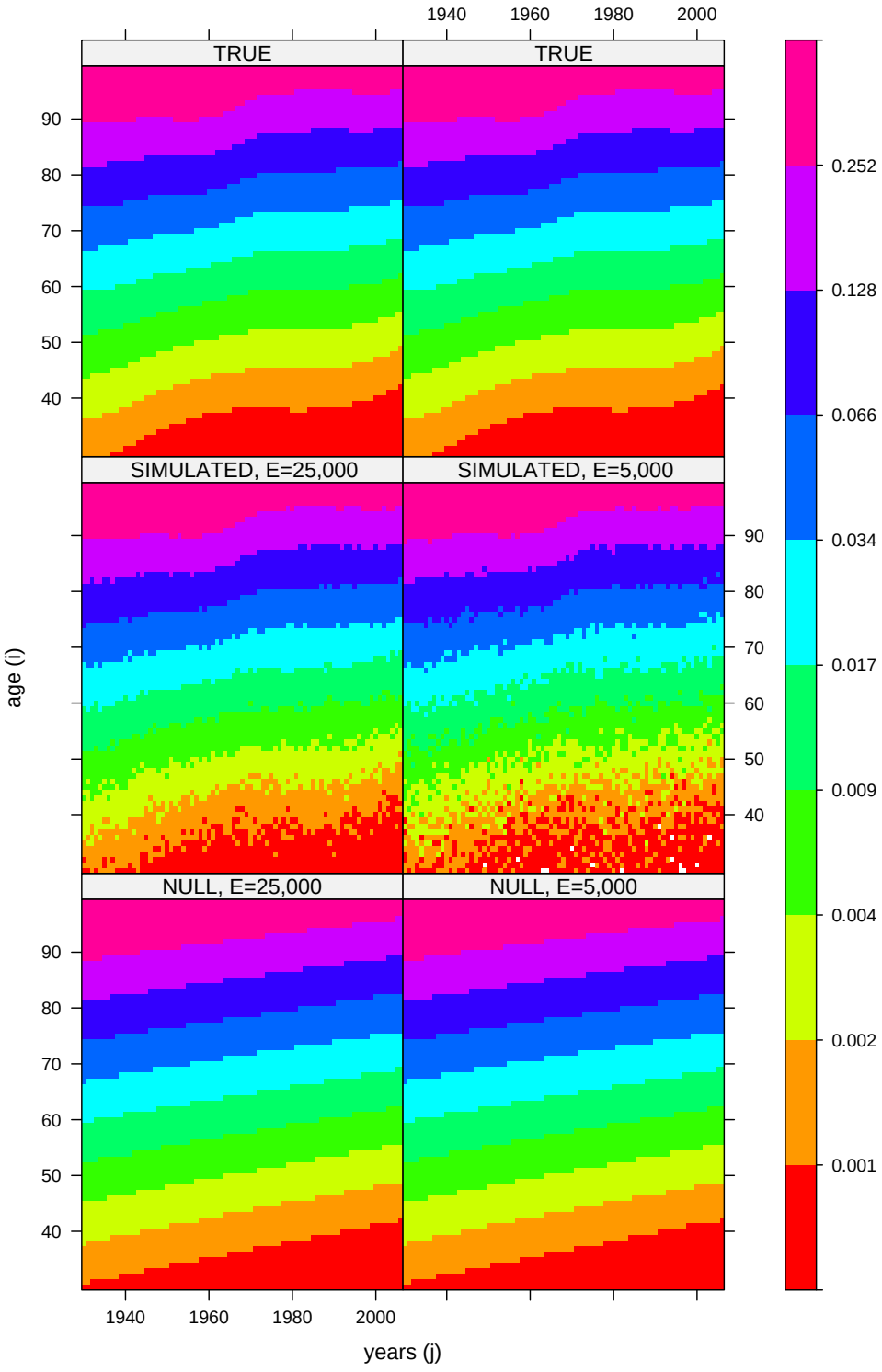

Figure 3.4: True and simulated death rates over age and years with different exposure matrices. Bilinear model from the simulation setting is also plotted.

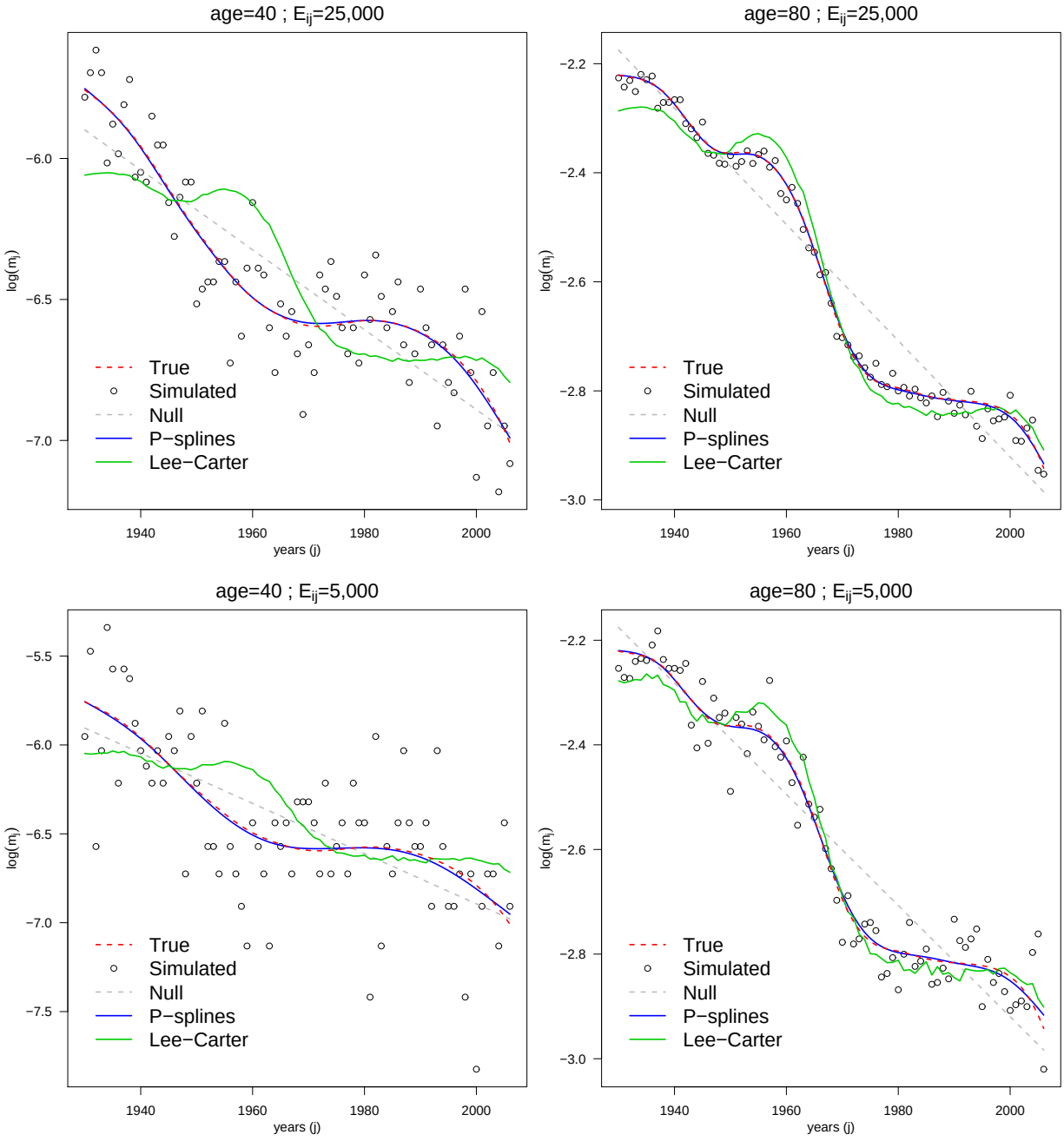

Figure 3.5: True, simulated and fitted deaths rates with P-splines and LC model along with the null model, logarithmic scale. Age 40 and 80 over years $j = 1930, \ldots, 2006$.

median values of		Simulation setting $E_{ij} = 25,000$	$E_{ij} = 5,000$
P-splines	$R^2_{(\text{bi})\text{lin}}$	0.931	0.734
	$R^2_{DEV,SMO,2}$	0.999	0.999
	$\text{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda)$	5406.800	5443.667
	$\text{ED}(\boldsymbol{a}, \lambda)$	23.224	21.366
Lee-Carter	$R^2_{(\text{bi})\text{lin}}$	0.750	0.591
	$R^2_{DEV,SMO,2}$	0.999	0.998
	$\text{Dev}(\boldsymbol{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$	19656.059	8133.536
	$\text{ED}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$	215.000	215.000

Table 3.3: Median values of $R^2_{(\text{bi})\text{lin}}$, $R^2_{DEV,SMO,2}$, deviance and effective dimensions/parameters from the 1,000 simulations fitted with P-spline approach and LC. Different exposure matrices are used, cf. equations (3.27) and (3.28).

trayed. The LC model clearly misfits the data and, additionally, it produces under-smoothed fitted values.

3.5 Applications to the Danish data

In this section, we study the performance of $R^2_{(\text{bi})\text{lin}}$ for the Danish female population introduced in Chapter 1.

It was already clear from the residual analysis, that the P-spline approach outperforms the LC model, even though the former employs fewer effective dimension (cf. Section 2.3.1). Nevertheless in Table 3.1, modified R^2 measures for smoothers did not reveal remarkable differences in gof between these two approaches.

The proposed $R^2_{(\text{bi})\text{lin}}$ aims to overcome this issue. Figure 3.6 shows shaded contour maps of the actual Danish female death rates, along with the fitted values and the null model. Figure 3.7 illustrates, for selected ages, the actual death rates and fitted values from the LC model and P-spline approach, together with the fitted values from the null model. P-splines follow the mortality development over years more closely than the LC model and the LC model clearly under-smoothed the actual death rates.

$R^2_{(\text{bi})\text{lin}}$ values are 0.8279138 and 0.7052449 for the P-spline approach and LC model, respectively. The difference is here more perceptible and informative than with $R^2_{DEV,SMO,2}$.

Finally, Table 3.4 presents different outcomes from (3.23) for the Danish population, taking into consideration different period and age ranges. The different values between P-splines and the LC model are clear in all the fitted mortality surfaces from which we conclude that P-splines give a better fit to these data in all scenarios. In contrast to Table 3.1, the range of the outcomes in Table 3.4 is [0.454684, 0.827914].

3.5.1 $R^2_{(\text{bi})\text{lin}}$ and information criteria

In Section 3.3.3, we presented the relations between $R^2_{(\text{bi})\text{lin}}$ and the information criteria for selecting smoothing parameters in a smoothing context: that is, AIC and BIC. For the Danish female

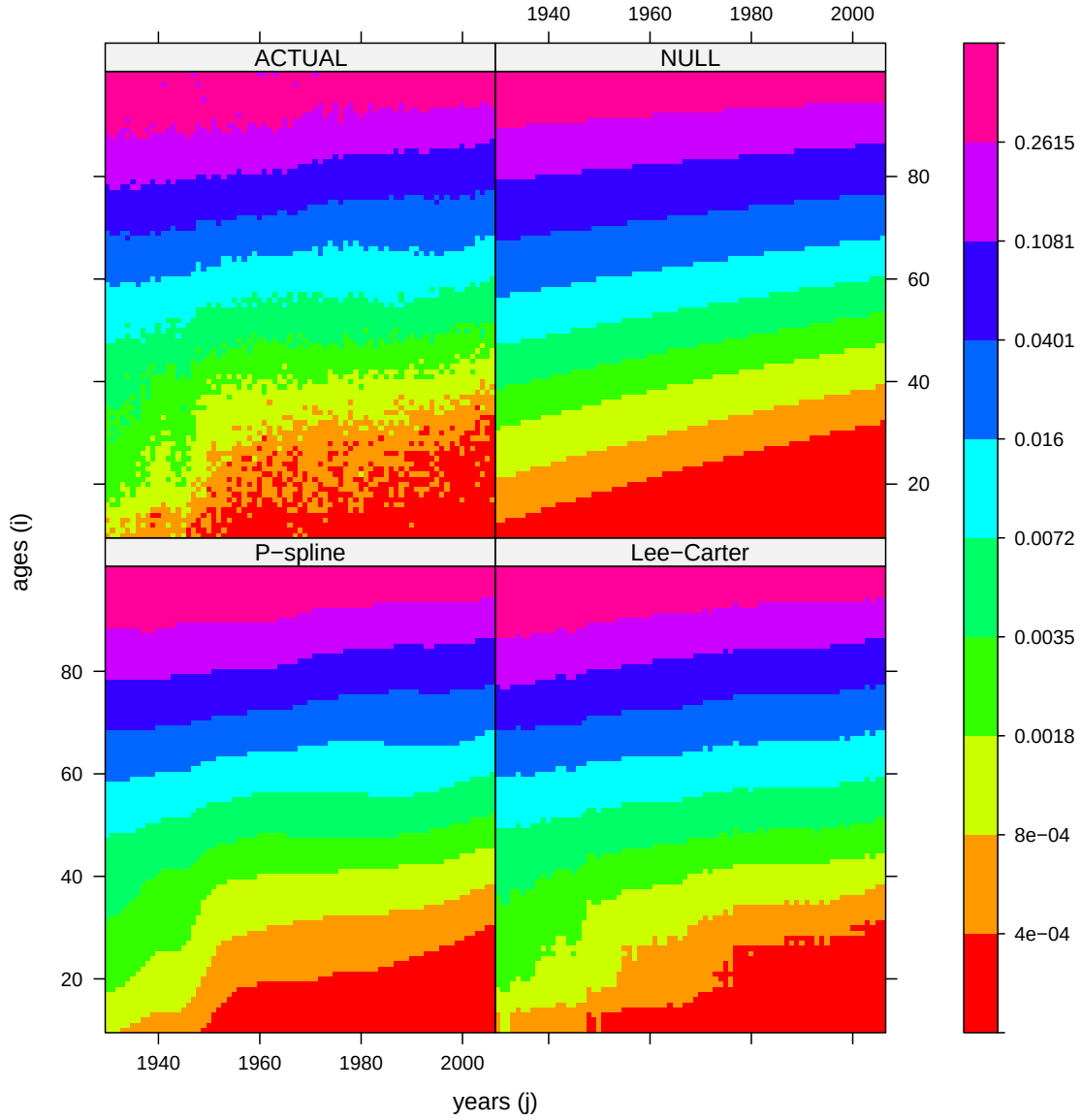

Figure 3.6: Actual and fitted deaths rates with P-splines and LC model along with the null model given in equation (3.19). Denmark, females.

Danish Data			P-splines	Lee-Carter
females	1930–2006	10–100	0.827914	0.705245
males	1930–2006	10–100	0.822210	0.727210
females	1930–2006	50–100	0.702440	0.486016
males	1930–2006	50–100	0.638110	0.454684
females	1950–2006	50–100	0.720131	0.586349
males	1950–2006	50–100	0.684898	0.542944

Table 3.4: $R^2_{\text{(bi)lin}}$ values for the Danish population by different period and age ranges.

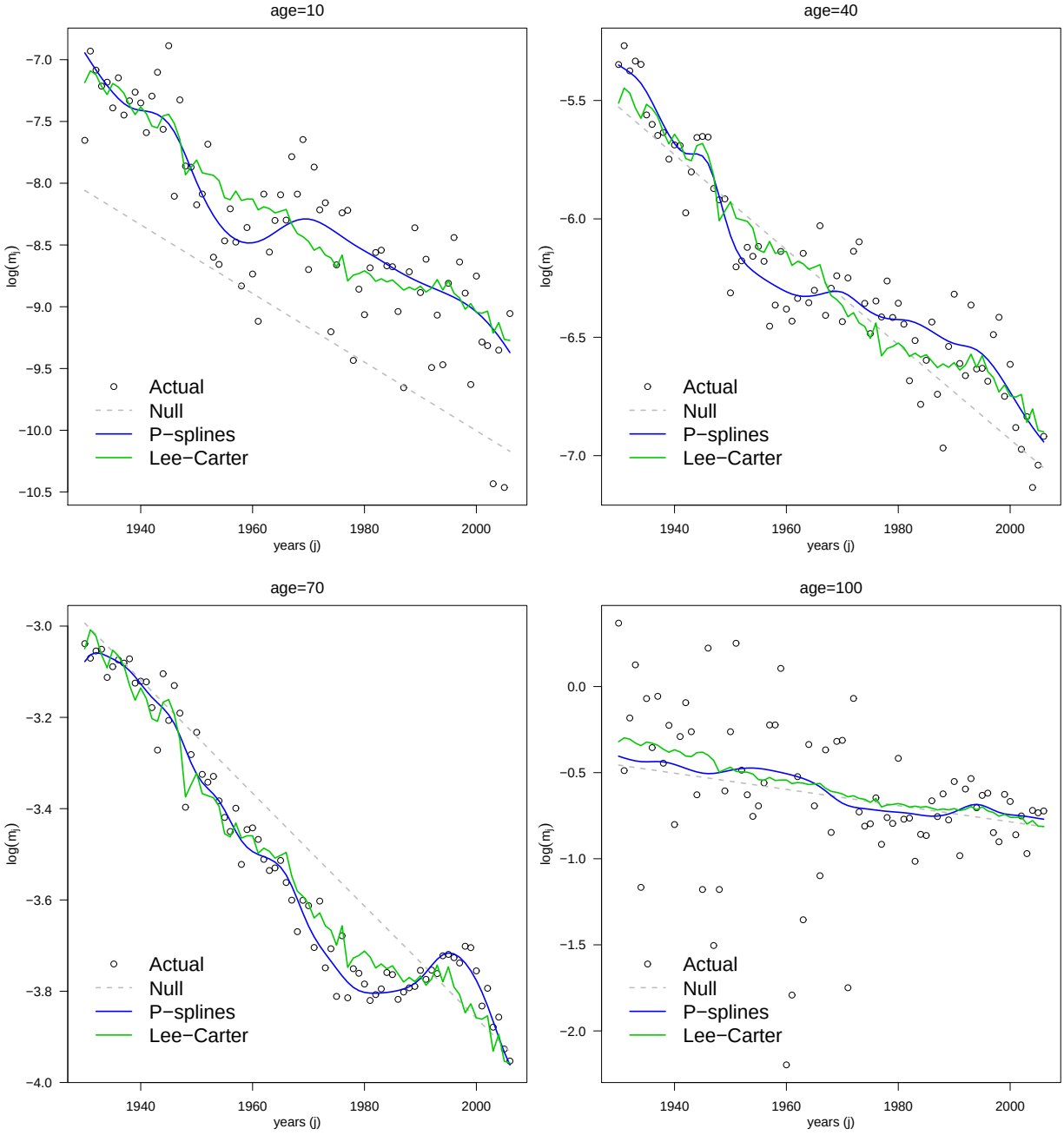

Figure 3.7: Actual and fitted death rates at selected ages over years, logarithmic scale. 2D smoothing with P-splines and LC model used for the estimation. Null model given in equation (3.19). Denmark, females.

population, Figure 3.8 shows the contour plots of AIC, BIC and $R^2_{(\text{bi})\text{lin}}$ over the same grid of λ_a and λ_y for the Danish female population from 1930 to 2006 and from age 30 to 100.

As mentioned in Section 2.2.1, BIC is more convenient to use in selecting smoothing parameters in a mortality setting, i.e. death counts are so large that effective dimension of the model needs to be penalized more. Smoothing parameters selected by AIC and $R^2_{(\text{bi})\text{lin}}$ (Fig. 3.8) are smaller than the smoothing parameters picked by BIC.

As explained in equations (3.25) and (3.26), outcomes from $R^2_{(\text{bi})\text{lin}}$ are closer to the AIC than the BIC. Consequently, $R^2_{(\text{bi})\text{lin}}$ will always be slightly higher for a fitted mortality surface in which λ_a and λ_y are selected by AIC. In this case $R^2_{(\text{bi})\text{lin}}$ is equal to 0.835897 when AIC is used, in contrast with 0.827914 of the fitted mortality surface selected by BIC.

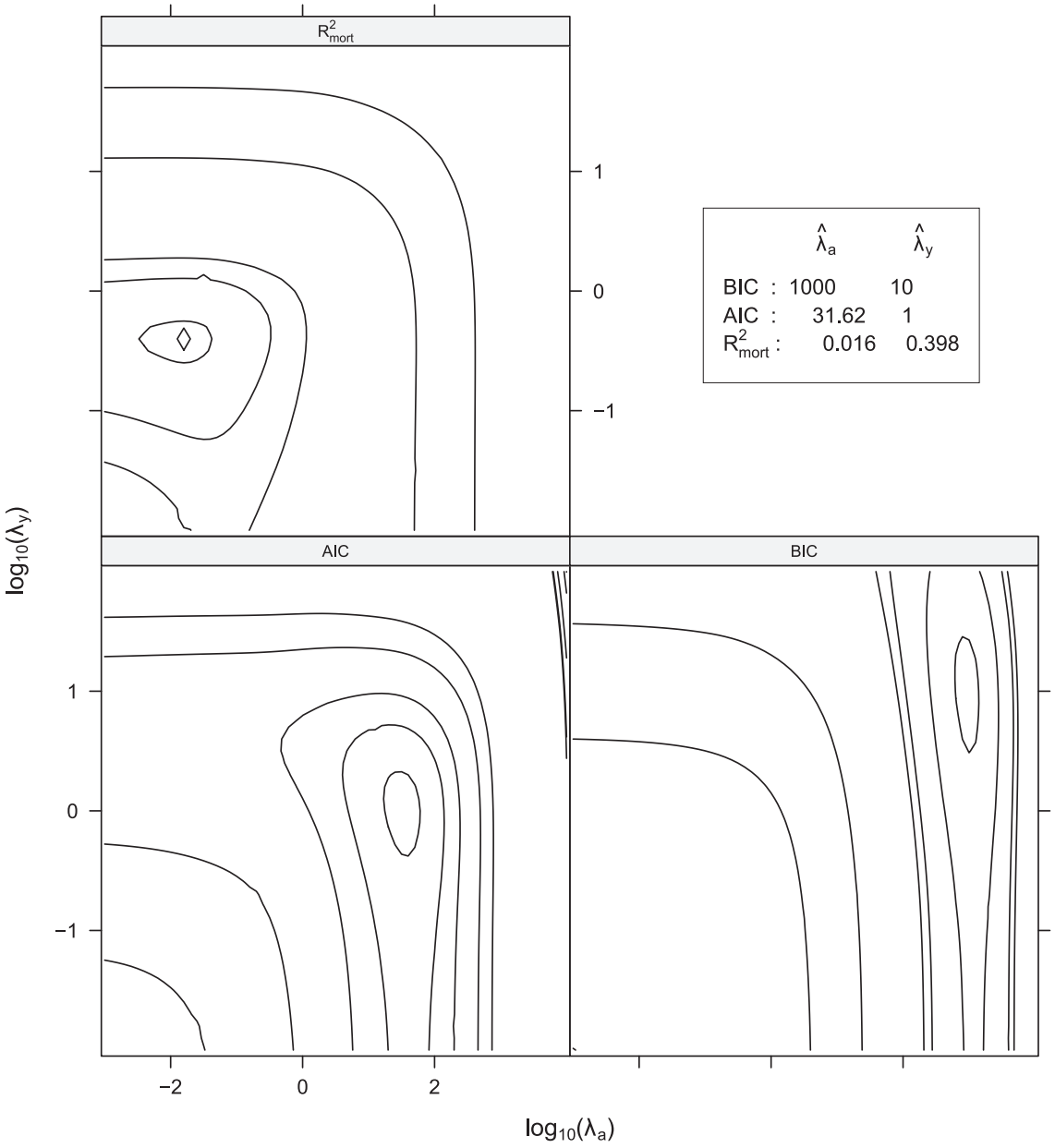

Figure 3.8: BIC, AIC and $R^2_{(\text{bi})\text{lin}}$ over a two-dimensional grid of λ_a and λ_y. Ages from 10 to 100. Denmark, females, 1930–2006

3.6 Summary

In this chapter, we proposed a new gof measure for mortality data, $R^2_{\text{(bi)lin}}$, an important tool for comparing models and data in this specific setting. First, we noticed that classic measures are essentially uninformative in the mortality context for two specific reasons. On one hand, mortality data present an often large number of death counts. On the other hand, the classic gof measure aims to compare fitted models with a null model, which is a simple overall mean of the data. Consequently, commonly used R^2 measures always give outcomes that are close to one, regardless of the applied model and of the actual data.

We also presented various generalizations of the R^2 for normal data which are suitable for non-normal data. Further corrections are also needed to account for the number of parameters in the fitted model. Moreover, working with smoothing techniques, effective dimension of the model have to be included in gof measures. Nevertheless, none of these adjustments is enough to allow informative comparison of explained variability of different models in mortality data.

The proposed measure is based on an alternative null model, specifically designed for mortality data, that is a linear or bilinear model for unidimensional or two-dimensional models, respectively. The equation (3.23) is a particular variant of commonly used gof measures in which we incorporate the effective dimension (number of parameters) used by the fitted model and an original denominator, which includes deviance and an effective dimension of a linear model.

The attractive feature of this new measure lies in the fact that the selected null model is nested in widely used demographic and smoothing models for analysis of mortality development. Specifically, both the Lee-Carter model and the two-dimensional regression P-splines can be viewed as extensions of the bilinear model over age and time.

Whereas differences in the classic gof measures, even after several adjustments, are hardly perceptible, $R^2_{\text{(bi)lin}}$ can be easily used to select and assess models and mortality data. In particular, we showed that, though the Lee-Carter model employs a considerably large number of parameters, P-spline methodology can better capture changes in mortality. For instance, for Danish females for the years 1930–2006 and for ages 10–100, the LC model explains 70% more variability present in the actual data than does the bilinear model, while two-dimensional P-splines improves the bilinear null model by 83%. This difference in the outcomes summarizes remarkably well what can be seen in the residual analysis, something that was not evident in common gof measures.

Chapter 4

Smooth estimates of age misreporting

In Chapter 1, we assumed that the total number of deaths over a specified age- and year-interval is a Poisson distribution. Demographic models can be thus viewed as Poisson regression models. Specifically, P-splines presented in Chapter 2 employed the Poisson assumption for fitting mortality surfaces, i.e. the mean and the variance are characterized by the same parameter. Therefore, we considered the variability of data already established by the Poisson distribution.

Poisson models are widely used in the regression analysis of count data (e.g. Frome, 1983; Frome et al., 1973; Haberman, 1974; Holford, 1983). Successive events, such as death counts in an age- and time-interval, "occur independently and at the same rate, the Poisson model is appropriate for the number of events observed" (McCullagh and Nelder, 1989, p. 193). At the same time, it is recognized that counts often display extra-Poisson variation, or overdispersion, relative to a Poisson model (among others, see Breslow, 1984; Cameron and Trivedi, 1986).

A peculiar source of overdispersion in mortality data is the so-called digit preference (DP) or age heaping. DP is defined as the tendency to round counts or measurements to pleasant digits. This usually leads to errors and bias in reported ages and spiky age-at-death distribution. This is particularly seen in the analysis of historical data or of countries with relatively poor data. Because of the large samples these digit preferences will always be picked up as a "signal" by statistical methods, including smoothing techniques, and therefore will distort analyses.

Different techniques have been developed to deal with this problem. Digit preference is most likely to be seen whenever laymen are involved. Hence, age misreporting has long been an issue in demography (Coale and Li, 1991; Das Gupta, 1975; Myers, 1940; Siegel and Swanson, 2004). Suggested solutions to compensate for age heaping are the application of summary indices to quantify the extent of misreporting and ad hoc procedures to reduce digit preferences and adjust age distributions. Mari Bhat (1990) proposed a model to estimate transition probabilities of age misstatement based on iterative adjustments and generalized stable population relationships.

Besides the quantification of digit preferences and the assessment of their consequences, only a few studies aim to model, estimate and correct the process of misreporting (Crawford et al., 2002; Edouard and Senthilselvan, 1997; Heitjan and Rubin, 1990; Pickering, 1992; Ridout and Morgan, 1991). Lately a novel and general methodology for dealing with this issue has been proposed by Camarda et al. (2008b).

They have proposed a general model for estimating the actual underlying age distribution as

well as the preference pattern which, by transferring observations from adjacent digits, leads to the finally observed pattern. This is achieved by combining the concept of penalized likelihood with the composite link model (CLM, Eilers, 2007; Thompson and Baker, 1981). Additionally, an L_1-ridge regression (Tibshirani, 1996) allows extraction of both the latent distribution and the pattern of misreporting probabilities.

In this chapter, the model proposed by Camarda et al. (2008b) will be presented in detail. Specifically, we will first introduce a typical example of age heaping, after which the essence of the CLM is introduced in Section 4.2, including the specific form of the composition matrix. Estimation of the model is covered in Section 4.3, including difference penalties for assuring smoothness, the estimation of the preference pattern, and the choice of optimal smoothing parameters. Computational details are then given in Section 4.4. In Section 4.5, we illustrate the approach via simulated data and present some demographic applications.

Although the model proposed by Camarda et al. (2008b) is in many respects very flexible, the fact that observations can only shift to their nearest neighbors is rather limiting. In Section 4.6, we generalize their approach to more general patterns of misreporting, i.e. allowing for exchanges between digits that are more than one category apart. Further extensions are presented in Section 4.7.

4.1 An example of digit preference

As an example for manifest age misreporting, Figure 4.1 shows the age distribution of adult Portuguese females (ages 30 to 89) who died during 1940 (Human Mortality Database, 2008). Systematic peaks at ages ending in 0 and, less prominently, 5 are typical features for countries with less accurate vital registration, which certainly was the case in Portugal almost seven decades ago.

Flanking the peaks, troughs are found at ages ending in 9, 1, 4, and 6. Moreover this phenomenon seems particularly severe at older ages. Also, even numbers in general seem to be preferred over odd digits.

Current age-at-death distributions are the result of the number of births and deaths, and migration flows in the past. Individual years may show particular outcomes, like epidemics, when birth cohorts are considerably smaller than the years before and after the crisis, or years of armed conflicts, when deaths are higher, especially among men. Thus, there is the possibility of irregularities in an age distribution, however, the specific reasons for such irregularities are usually well-understood from the historic records. In the absence of such specific past events, the assumption of a smooth age distribution is reasonable, implying that the peaks and gaps are the result of certain preferences in reported ages. If spikes or troughs in the distribution are due to events in the past, rather than digit preference, these digits will be excluded from the smoothing procedure.

The observed frequencies, therefore, can be viewed as the outcome of a misreporting process that transforms a smooth but latent age distribution into observed data. The counts at the preferred digits are composed of the actual values at these ages, plus the misclassified cases from the neighboring categories due to the prevalent preference pattern.

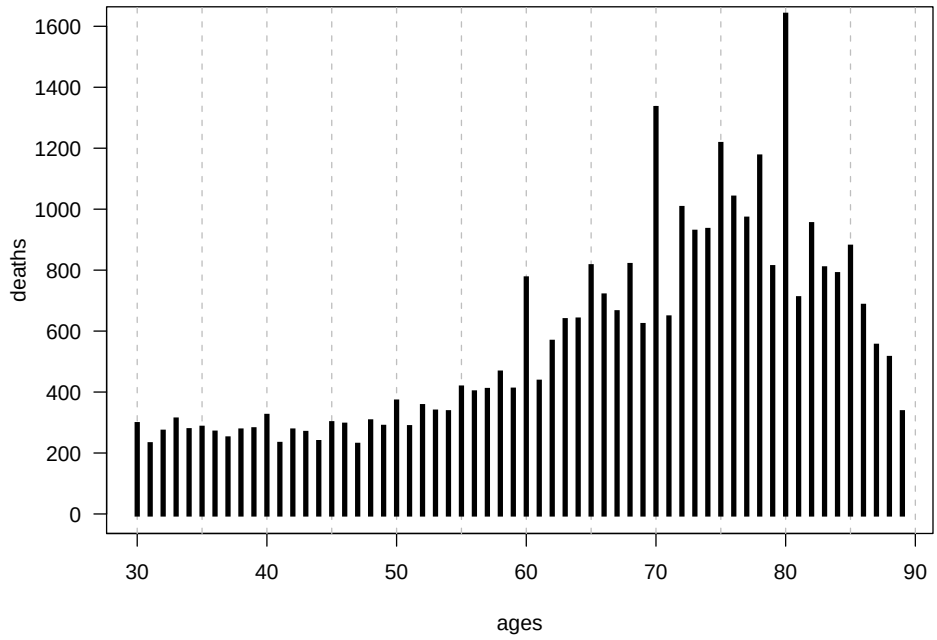

Figure 4.1: Age-at-Death distribution for Portugal, females, 1940.

4.2 The Composite Link Model

Camarda et al. (2008b) assumed a smooth discrete sequence $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$, which is the unknown latent distribution. In this specific demographic setting, $j = 1, \ldots, J$ are the ages. To ensure non-negative elements of $\boldsymbol{\gamma}$, this sequence is denoted as $\boldsymbol{\gamma} = \exp(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ smooth, that is neighboring elements of $\boldsymbol{\beta}$ being of similar size. The elements γ_j, $j = 1, \ldots, J$ are the counts that would be expected, if there were no age heaping. However, the mechanism, which actually generates observations, operates by linearly composing the values in $\boldsymbol{\gamma}$ to a vector $\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$. The observed counts $\boldsymbol{y}$ are realizations from Poisson variables with $E(\boldsymbol{y}) = \boldsymbol{\mu}$, i.e.

$$P(y_j) = \frac{\mu_j^{y_j}\, e^{-\mu_j}}{y_j!}\,. \tag{4.1}$$

The composition matrix $\boldsymbol{C}$ embodies the digit preference mechanism by partly redistributing certain elements of $\boldsymbol{\gamma}$ to neighboring, preferred values in $\boldsymbol{\mu}$. In a general CLM the composition matrix $\boldsymbol{C}$ need not be a square matrix, as several categories could be lumped together. Modeling age heaping, because expected counts are redistributed only partly, the matrix $\boldsymbol{C}$ is of dimension $J \times J$.

4.2.1 The composition matrix C

The composition matrix $\boldsymbol{C}$ describes how the latent distribution $\boldsymbol{\gamma}$ was mixed before generating the data and it is characteristic for the predominant preference pattern. Consequently, for mod-

eling digit preferences, the matrix $\boldsymbol{C}$ needs to be defined according to the assumptions of the misreporting process. Eilers and Borgdorff (2004) allowed misreporting only for a few selected digits, with probabilities that did not change with the size of the underlying number, e.g. the probability for a transfer from $10x + 7$ to $10x + 8$ was assumed to be the same for all $x \epsilon \mathbb{N}$.

The mentioned limitation of the model proposed by Camarda et al. (2008b) lies in the assumption that misreporting will only move observations to the immediate neighboring digits, both to the left and the right. For example, observations are allowed to move from 9 to 10, but also from 9 to 8, but not from 12 to 10. In Section 4.6, we generalize this approach and consider preferences that move observations by two or more steps, e.g. observations shifted from 8 to 10, or from 12 to 15, if more than 2 steps are allowed.

Let denote by p_{jk} the proportion of γ_k that is moved from category k to category j. Allowing only one-step transitions implies that $p_{jk} = 0$ for $|j - k| > 1$. These proportions are summarized in the $J \times J$ composition matrix $\boldsymbol{C}$:

$$\boldsymbol{C} = \begin{pmatrix} 1-p_{21} & p_{12} & 0 & 0 & \cdots & 0 \\ p_{21} & 1-p_{12}-p_{32} & p_{23} & \cdots & & \vdots \\ 0 & p_{32} & 1-p_{23}-p_{43} & p_{34} & \cdots & \vdots \\ 0 & 0 & p_{43} & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1-p_{J-2,J-1}-p_{J,J-1} & p_{J-1,J} \\ 0 & \cdots & \cdots & 0 & p_{J,J-1} & 1-p_{J-1,J} \end{pmatrix} \tag{4.2}$$

The diagonal elements $c_{jj} = 1 - p_{j-1,j} - p_{j+1,j}$ of $\boldsymbol{C}$ specify the proportions of the γ_j that do not get redistributed. Note that all columns in $\boldsymbol{C}$ add up to 1. The matrix $\boldsymbol{C}$ can be adapted also when a particular digit needs to be excluded from the redistribution process.

It is obvious that the $2 \cdot (J-1)$ unknown elements p_{jk} cannot be estimated without imposing additional restrictions. A penalized weighted least-squares approach is suggested by Camarda et al. (2008b) and is discussed in detail in Section 4.3.3.

4.3 Estimating the CLM and the preference pattern

4.3.1 The CLM algorithm

Thompson and Baker (1981) present the CLM and the estimation algorithm very succinctly, and Eilers (2007) extended the approach to smooth latent distributions estimated by penalized likelihood. Camarda et al. (2008b) briefly described the CLM for modeling digit preference. For easier reference, in this section we provide details on the CLM and then describe the estimation procedure. Working with death counts, the focus is on the Poisson distribution.

In case of no digit preference one would be able to directly observe counts z_j, $j = 1, \ldots, J$,

following a Poisson distribution such that

$$P(z_j) = \frac{\gamma_j^{z_j} e^{-\gamma_j}}{z_j!}$$

Assuming smoothness of the elements of $\boldsymbol{\beta}$ immediately implies smoothness of $\gamma_j = \exp\{\beta_j\}$. In case one aims to model a flexible functional dependence of the latent means $\boldsymbol{\gamma}$ on some covariate, a regression on B-splines can be included into this framework, as well (see Section 2.1). This leads to the more general formulation $\boldsymbol{\gamma} = \exp\{\boldsymbol{X}\boldsymbol{\beta}\}$, where the design matrix $\boldsymbol{X}$ contains the basis elements covering the range of $\boldsymbol{z}$, and the vector $\boldsymbol{\beta}$ gives the weights by which the individual B-splines in the basis get multiplied. Again, smoothness of the vector $\boldsymbol{\beta}$ implies smoothness of $\boldsymbol{\gamma}$. In what follows, the model matrix is defined as $\boldsymbol{X} = \boldsymbol{I}$, the identity matrix.

Without any digit preference, this model can be seen as a generalized linear model (GLM) (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972). As in Section 2.1.2 a link function $g(\gamma)$ and a linear predictor η_j are introduced:

$$g(\gamma_j) = \eta_j = \sum_{k=1}^{K} x_{jk}\beta_k \,.$$

The method of maximum likelihood is used to estimate $\boldsymbol{\beta}$. McCullagh and Nelder (1989, equations 2.12 and 2.13) showed that the ML equations are:

$$\sum_{j=1}^{J} \frac{(z_j - \gamma_j)}{v(\gamma_j)} \frac{\partial \gamma_j}{\partial \beta_k} = \sum_{j=1}^{J} \frac{(z_j - \gamma_j)}{v(\gamma_j)} \frac{\partial \gamma_j}{\partial \eta_k} x_{jk} = 0 \tag{4.3}$$

where $v(\gamma_j)$ is the variance when $E(z_j) = \gamma_j$. With the Poisson distribution and the log link, $\eta_j = \ln(\gamma_j)$, it follows that $\partial\gamma_j/\partial\eta_j = \gamma_j$ and $v(\gamma_j) = \gamma_j$. Thus, equation (4.3) simplifies to:

$$\sum_{j=1}^{J} (z_j - \gamma_j) x_{jk} = 0$$

These equations are nonlinear in $\boldsymbol{\beta}$ and an iterative procedure is needed to solve them.

Assume that approximate values $\tilde{\beta}_k$, and corresponding $\tilde{\gamma}_j$ are known. For small changes in $\boldsymbol{\beta}$ one has

$$\Delta\gamma_j = \gamma_j - \tilde{\gamma}_j \approx \sum_{k=1}^{K} \frac{\partial \gamma_j}{\partial \beta_k} \Delta\beta_k = \gamma_j \sum_{k=1}^{K} x_{jk}(\beta_k - \tilde{\beta}_k)$$

and

$$\sum_{j=1}^{J} \sum_{l=1}^{K} \tilde{\gamma}_j x_{jk} x_{jl} \Delta\beta_l = \sum_{j=1}^{J} x_{jk}(z_j - \tilde{\gamma}_j)$$

Add $\sum_j \sum_l \tilde{\gamma} x_{jk} x_{jl} \beta_l = \sum_j \tilde{\gamma} x_{jk} \tilde{\eta}_j$ to both side, to get

$$\sum_{l=1}^{K} \sum_{j=1}^{J} \tilde{\gamma}_j x_{jk} x_{jl} \beta_l = \sum_{j=1}^{J} \tilde{\gamma}_j x_{jk} \left(\frac{z_j - \tilde{\gamma}_j}{\tilde{\gamma}_j} + \tilde{\eta}_j \right)$$

with $\tilde{\eta}_j = \sum_{k=1} x_{jk}\tilde{\beta}_k$.

It is easy to recognize the last system of equations as weighted linear regression of a "working variable" $(z_j - \tilde{\gamma}_j)/\tilde{\gamma}_j + \tilde{\eta}_j$ on X, with weights $\tilde{\gamma}_j$. This is a special case of the iteratively reweighted least squares (IRWLS) algorithm for the estimation of GLMs (Nelder and Wedderburn, 1972) also used in this thesis in Section 2.1.2. In matrix notation it is

$$\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{z} - \tilde{\boldsymbol{\gamma}}) + \boldsymbol{X}\tilde{\boldsymbol{\beta}}\} \tag{4.4}$$

with $\tilde{\boldsymbol{W}} = \mathrm{diag}(\tilde{\gamma})$.

However, one does not observe the vector $\boldsymbol{z}$, but another variable $\boldsymbol{y}$, with $\boldsymbol{\mu} = E(\boldsymbol{y}) = \boldsymbol{C}\boldsymbol{\gamma}$, or $\mu_i = \sum_j c_{ij}\gamma_j$. The maximum likelihood equations (4.3) can be adapted in the following way:

$$\sum_{i=1}^{I} \frac{(y_i - \mu_i)}{v(\mu_i)} \frac{\partial \mu_i}{\partial \beta_k} = 0$$

As stated in equation (4.1) the observed elements of vector $\boldsymbol{y}$ are realizations from Poisson variables.

The derivatives of the expected values μ_i can be written as

$$\frac{\partial \mu_i}{\partial \beta_k} = \sum_{j=1}^{J} c_{ij} \frac{\partial \gamma_j}{\partial \beta_k} = \sum_{j=1}^{J} c_{ij} x_{jk} \gamma_j$$

and the likelihood equations will thus be

$$\sum_{i=1}^{I} (y_i - \mu_i)\breve{x}_{ik} = 0\,,$$

where $\breve{x}_{ik} = \sum_j c_{ij}x_{jk}\gamma_j/\mu_i$. The matrix with elements $\breve{x}_{ik}$ can be interpreted as a "working X". The IRWLS equations become:

$$\sum_{k=1}^{K}\sum_{i=1}^{I} \tilde{\mu}_i \breve{x}_{ik}\breve{x}_{il}\tilde{\beta}_l = \sum_{i=1}^{I} \tilde{\mu}_i \breve{x}_{ik} \left(\frac{y_i - \tilde{\mu}_i}{\tilde{\mu}_i} + \sum_{k=1}^{K} \breve{x}_{ik}\tilde{\beta}_k \right)$$

or, in matrix notation:

$$\breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} = \breve{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \breve{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\} \tag{4.5}$$

where $\tilde{\boldsymbol{W}} = \mathtt{diag}(\tilde{\mu})$. A detailed derivation of (4.5) can be found in Eilers (2007).

4.3.2 Smooth latent distribution in a CLM

When $\boldsymbol{X} = \boldsymbol{I}$, then $\ln(\boldsymbol{\gamma}) = \boldsymbol{\beta}$ and smoothness of $\boldsymbol{\beta}$ implies smoothness of $\boldsymbol{\gamma}$. As we have seen in Section 2.1, we can smooth the solution vector of the coefficients $\boldsymbol{\beta}$ by subtracting a roughness penalty from the log-likelihood (Eilers and Marx, 1996). Specifically the roughness of vector $\boldsymbol{\beta}$ can be measured with differences. The simplest form of differences which is used to capture the

smoothness into a vector is:

$$S_1 = \sum_{k=2}^{K} (\Delta\beta_k)^2 = \sum_{k=2}^{K} (\beta_k - \beta_{k-1})^2$$

A rough vector $\boldsymbol{\beta}$ will show large differences between neighboring elements and hence give a high value of S_1, while a smooth vector will make S_1 low. Ultimate smoothness is obtained when all elements of $\boldsymbol{\beta}$ are equal and $S_1 = 0$.

Higher order differences can be used as well and introduce stronger smoothness:

$$S_2 = \sum_{k=3}^{K} (\Delta^2\beta_k)^2 = \sum_{k=3}^{K} (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2$$

or

$$S_3 = \sum_{k=4}^{K} (\Delta^3\beta_k)^2 = \sum_{k=4}^{K} (\beta_k - 3\beta_{k-1} + 3\beta_{k-2} - \beta_{k-3})^2$$

$S_2 = 0$ is obtained when $\beta_k = c_1 k + c_0$, for arbitrary c_0 and c_1, while S_3 will be zero for any $\boldsymbol{\beta}$ that is a quadratic in k.

In matrix notation, let $\boldsymbol{D}_d$ be the matrix that computes dth differences: $\Delta^d\boldsymbol{\beta}$. For some examples of difference matrices, we refer to equation (2.6) on page 19.

The roughness measure with differences of order d can be thus written as

$$\boldsymbol{S}_d = \boldsymbol{\beta}'\boldsymbol{D}_d'\boldsymbol{D}_d\boldsymbol{\beta} = \|\boldsymbol{D}_d\boldsymbol{\beta}\|^2 \tag{4.6}$$

where $\boldsymbol{D}_d \in \mathbb{R}^{(K-d)\times K}$ is the matrix that computes dth order differences. Like in the P-spline approach we use $d = 2$ in the following.

The partial derivatives of this penalty w.r.t. the unknown parameters are given by

$$\frac{\partial \boldsymbol{S}_d}{\partial \boldsymbol{\beta}} = 2\boldsymbol{D}_d'\boldsymbol{D}_d\boldsymbol{\beta}\,.$$

Both in GLMs and CLMs, the solution vector $\boldsymbol{\beta}$ can be forced to be smooth by subtracting a roughness penalty from the log-likelihood $l(\boldsymbol{\beta};\boldsymbol{y})$ (see section 2.1 and Eilers and Marx (1996)). This penalty is the roughness measure (4.6), weighted by the smoothing parameter λ:

$$l^* = l(\boldsymbol{\beta};\boldsymbol{y}) - \frac{\lambda}{2}\|\boldsymbol{D}_d\boldsymbol{\beta}\|^2\,.$$

This penalty can be easily introduced into the likelihood for the CLM, leading to the following system of equations

$$(\check{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\check{\boldsymbol{X}} + \lambda\,\boldsymbol{D}^T\boldsymbol{D})\tilde{\boldsymbol{\beta}} = \check{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\{\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \check{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\}. \tag{4.7}$$

As in the P-spline approach the smoothing parameter λ balances model fidelity, as expressed by the log-likelihood $l(\boldsymbol{\beta};\boldsymbol{y})$, and smoothness of the parameter estimates, as expressed by the penalty term. For a given value of λ, equations (4.7) can be solved iteratively. Methods for optimal choice

of λ are discussed in Section 2.1.4 for the general P-spline approach and Section 4.3.4 introduces the specific criterion that was proposed in Camarda et al. (2008b).

4.3.3 Finding the misreporting proportions

In order to estimate the proportions p_{jk} of misreported counts in the matrix $\boldsymbol{C}$ (cf. equation (4.2)), Camarda et al. (2008b) suggested solving a constrained weighted least squares regression within the IRWLS procedure. From the structure of the composition matrix $\boldsymbol{C}$ in equation (4.2), the vector of expected values $\boldsymbol{\mu}$ can be written as

$$\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma} = \boldsymbol{\gamma} + \boldsymbol{\Gamma}\boldsymbol{q}, \tag{4.8}$$

where $\boldsymbol{p} = (p_{12}, p_{23}, \ldots, p_{J-1,J}; p_{21}, \ldots, p_{J,J-1})^T$, the left-to-right and the right-to-left transfer probabilities concatenated into a vector of length $2 \cdot (J-1)$. Correspondingly, the $J \times 2 \cdot (J-1)$-matrix $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_2 & 0 & \cdots & 0 & -\gamma_1 & 0 & \cdots & 0 \\ -\gamma_2 & \gamma_3 & & \vdots & \gamma_1 & -\gamma_2 & & \vdots \\ 0 & -\gamma_3 & \ddots & 0 & 0 & \gamma_2 & \ddots & 0 \\ \vdots & & \ddots & \gamma_J & \vdots & & \ddots & -\gamma_{J-1} \\ 0 & \cdots & 0 & -\gamma_J & 0 & \cdots & \cdots & \gamma_{J-1} \end{pmatrix}, \tag{4.9}$$

where the last column containing γ_J is not included as p_J is obviously set to 0.

Since $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\mu})$, the distribution of $(\boldsymbol{y} - \boldsymbol{\gamma})$ can be approximated as

$$(\boldsymbol{y} - \boldsymbol{\gamma}) \approx N(\boldsymbol{\Gamma}\boldsymbol{p}, \texttt{diag}(\boldsymbol{\mu})). \tag{4.10}$$

In a linear model framework equation (4.10) would be estimated with a simple weighted least-squares approach. However, as the number of unknowns in $\boldsymbol{p}$, namely $2 \cdot (J-1)$, is considerably larger than the number J of available data points, additional restrictions have to be imposed on $\boldsymbol{p}$.

In order to capture the most significant probabilities, the size of the misreporting probabilities $\boldsymbol{p}$ has to be constrained. Simple continuous shrinkage methods, such as ridge regression (Hoerl and Kennard, 1988), are, in this case, a possible approach:

$$(\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma} + \kappa\boldsymbol{I})\boldsymbol{p} = \boldsymbol{\Gamma}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\gamma}) \tag{4.11}$$

The ridge-term $\kappa\boldsymbol{I}$ penalizes the squared norm $\boldsymbol{p}'\boldsymbol{p}$ of the coefficient vector $\boldsymbol{p}$, the resulting estimates tending to have elements of similar sizes, which is unlike what we would expect for digit preference patterns. An increasing parameter κ brings together p_j to smaller values. Furthermore, ridge regression is not able to produce a more parsimonious model, since it shrinks all the p_j at the same pace, and hence either all or none of the estimated p_j will be approximately zero.

For a typical DP pattern, we would expect that particular digits attract observations from their neighbors, while the rest of the p_{jk} will be close to zero because, for them, no preference is evident. Therefore, an automatic selection of the misreporting probabilities should be combined with a continuous shrinkage of the model.

Equation (4.11) can be seen as a specific case of the power ridge regression introduced by Frank and Friedman (1993). They introduced a general ridge penalty L_q-norm of the parameters and equation (4.11) corresponds to $q = 2$. Within this framework, instead of the L_2 norm $\boldsymbol{p}'\boldsymbol{p}$, Camarda et al. (2008b) introduced an L_1 penalty into the weighted least-squares problem (4.10). As pointed out by Tibshirani (1996), this penalty tends to select a small number of elements, p_{jk}, that exhibit the strongest effects, while possibly shrinking some others to zero. Commonly called "lasso", this penalized regression model leads to a penalty term equal to $\kappa \sum |p_j|$.

Numerical optimization in this framework is not immediately straightforward since the objective function contains a quadratic term and a sum of absolute value. Following the proposal of Schlossmacher (1973), Camarda et al. (2008b) avoided quadratic programming or other methods that move away from the (iterative) least squares algorithm, because a well-defined effective dimension will be needed to compute information criteria such as AIC (see Section 4.3.4). Specifically, they set $\sum_j |p_j| = \sum p_j^2/|p_j|$, turning a sum of absolute values into a weighted sum of squares. Of course, to compute the weights, ones needs to know $\boldsymbol{p}$. Camarda et al. (2008b) solved this problem by iteration, using weights $1/|\tilde{p}_j|$, where $\tilde{\boldsymbol{p}}$ is an approximation of the solution.

In formulas the iterative solution is given by the following system of equations:

$$(\boldsymbol{\Gamma}'\boldsymbol{V}\boldsymbol{\Gamma} + \kappa\tilde{\boldsymbol{Q}})\boldsymbol{p} = \boldsymbol{\Gamma}'\boldsymbol{V}(\boldsymbol{y} - \boldsymbol{\gamma}), \tag{4.12}$$

where $\boldsymbol{V} = \text{diag}(1/\boldsymbol{\mu})$ and the matrix $\boldsymbol{Q}$ is

$$\boldsymbol{Q} = \begin{pmatrix} \frac{1}{|p_{12}|+\epsilon} & 0 & . & . & . & . & . & . \\ 0 & \frac{1}{|p_{23}|+\epsilon} & 0 & . & . & . & . & . \\ . & 0 & \ddots & 0 & . & . & . & . \\ . & . & 0 & \frac{1}{|p_{j-1,j}|+\epsilon} & 0 & . & . & . \\ . & . & . & 0 & \frac{1}{|p_{21}|+\epsilon} & 0 & . & . \\ . & . & . & . & 0 & \frac{1}{|p_{32}|+\epsilon} & 0 & . \\ . & . & . & . & . & 0 & \ddots & 0 \\ . & . & . & . & . & . & 0 & \frac{1}{|p_{J,J-1}|+\epsilon} \end{pmatrix}. \tag{4.13}$$

A small number ϵ is introduced to prevent numerical instabilities when elements of $\boldsymbol{p}$ become very small. In our experience, $\epsilon = 10^{-6}$ worked well.

The additional parameter κ in (4.12) constrains the size of misreporting proportions p_{jk} and has to be estimated similarly to the smoothing parameter λ (see Section 4.3.4). In practice, Camarda et al. (2008b) suggested alternately estimating $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ for a few iterations and afterward, they start updating $\boldsymbol{p}$ from (4.12).

4.3.4 Optimal smoothing

The estimating equations for the penalized CLM in (4.7) and for the preference pattern in (4.12) depend on the combination of the two smoothing parameters λ and κ. Once λ and κ are fixed, the estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{p}}$ are determined. To choose the optimal (λ, κ)-combination, Camarda et al.

(2008b) minimize Akaike's Information Criterion (AIC):

$$\mathrm{AIC}(\lambda, \kappa) = \mathrm{Dev}(\boldsymbol{y}; \boldsymbol{\beta}, \lambda, \kappa) + 2\,\mathrm{ED}(\boldsymbol{\beta}, \lambda, \kappa)\,. \tag{4.14}$$

$\mathrm{Dev}(\boldsymbol{y}; \boldsymbol{\beta}, \lambda, \kappa)$ is the deviance of the Poisson model (cf. equation (2.16) on page 25), and $\mathrm{ED}(\boldsymbol{\beta}, \lambda, \kappa)$ is the effective dimension of the model for given (λ, κ). In a similar fashion, as described in Section 2.1.3, Camarda et al. (2008b) denoted the effective dimension as the sum of the two model components, i.e. $\mathrm{ED}(\boldsymbol{\beta}, \lambda, \kappa) = \mathrm{ED}_1(\boldsymbol{\beta}, \lambda) + \mathrm{ED}_2(\boldsymbol{\beta}, \kappa)$, where $\mathrm{ED}_1(\boldsymbol{\beta}, \lambda)$ denotes the effective dimension of the penalized CLM, and $\mathrm{ED}_2(\boldsymbol{\beta}, \kappa)$ refers to the penalized WLS-regression. Specifically, we have

$$\mathrm{ED}_1(\boldsymbol{\beta}, \lambda) = \mathrm{trace}\{\check{\boldsymbol{X}}(\check{\boldsymbol{X}}'\hat{\boldsymbol{W}}\check{\boldsymbol{X}} + \lambda \boldsymbol{P})^{-1}(\check{\boldsymbol{X}}'\hat{\boldsymbol{W}})\} \tag{4.15}$$

and

$$\mathrm{ED}_2(\boldsymbol{\beta}, \kappa) = \mathrm{trace}\{\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{V}\boldsymbol{\Gamma} + \kappa \boldsymbol{Q})^{-1}\,\boldsymbol{\Gamma}'\boldsymbol{V}\}\,. \tag{4.16}$$

An efficient 2D grid-search for λ and κ is adequate for finding the minimum of the AIC (see Figure 4.3 in Section 4.5). Both IRWLS iterations and penalized WLS were implemented in `R` (R Development Core Team, 2008) and the next section will present the computational details.

4.4 Software considerations

As explained in Section 4.3, the proposed model is built up from two components: the penalized composite link model and the constrained weighted least squares regression. We show how to implement these two components within an IWRLS algorithm framework. We work in `R` (R Development Core Team, 2008) because of its widespread use, but the following code should be easily understood by someone who is unfamiliar with this language, if the following notation is known.

The symbol `<-` is an assignment statement, an asterisk, `*`, means element-by-element multiplication, `%*%` means matrix multiplication, `t()` means transpose and `solve()` estimates a generic least square model, i.e. one would write `x <- solve(A, b)` to solve $A \cdot x = b$ for x, where b can be either a vector or a matrix.

4.4.1 The Penalized CLM component

The IRWLS related to the system of equations in (4.5) can be solved given a computed $\boldsymbol{C}$ matrix (cf. (4.2)). Given a vector of starting values for $\boldsymbol{p}$ (`p`):

```
# Compositional matrix C
build.C <- function(p){
  ld  <- length(p)/2
  p.u <- p[1:ld]
  p.l <- p[1:ld+ld]
  C <- diag(c(1-p.l, 1))
  diag(C)[2:(ld+1)] <- diag(C)[2:(ld+1)] - p.u
```

```
  diag(C[-1,]) <- p.l
  diag(C[,-1]) <- p.u
  return(C)
}
```

The following routine is used for each iteration for fitting the penalized CLM in equation (4.7). As a first step, one needs to create the following objects: the penalty matrix $\boldsymbol{P}$, the composition matrix $\boldsymbol{C}$ and initial values for the linear predictor $\boldsymbol{\eta}$ (in the code `eta.old`).

Sequentially, the routine updates the latent distribution ($\boldsymbol{\gamma}$, `gamma`) the expected values ($\boldsymbol{\mu}$, `mu`), the "working model matrix" ($\breve{X}$, `X`), the weight matrix ($\tilde{W}$, `W`), the RHS (`tXr`) and the LHS of equations (4.7), with and without the additional penalty matrix (`G` and `GpP`). Finally with `eta.new`, this routine solves the system of equations in (4.7), updating the linear predictors $\boldsymbol{\eta}$.

```
# Updating the Poisson P-GLM part
UpdatePOI <- function(C, P, eta.old){
  gamma   <- exp(eta.old)
  mu      <- C %*% gamma
  X       <- C * ((1 / mu) %*% t(gamma))
  W       <- diag(as.vector(mu))
  tXr     <- t(X) %*% (y - mu + C %*% (gamma * eta.old))
  G       <- t(X) %*% W %*% X
  GpP     <- G + P
  eta.new <- solve(GpP, tXr)
  return(list(eta=eta.new, gamma=gamma, mu=mu, p=p))
}
```

This `R`-function will produce new estimates of the linear predictor ($\boldsymbol{\eta}$, `eta`), the updated vectors of the latent distribution ($\boldsymbol{\gamma}$, `gamma`), expected values ($\boldsymbol{\mu}$, `mu`) and misreporting probabilities ($\boldsymbol{p}$, `p`).

4.4.2 The constrained WLS component

The system of equations (4.12), which solves the constrained WLS regression, needs both the model matrix $\boldsymbol{\Gamma}$ (eq. (4.9)) and the penalty matrix $\boldsymbol{Q}$ for the L_1 "lasso" regression problem (cf. equation (4.13)).

Given a vector of updated expected values for the latent distribution $\boldsymbol{\gamma}$, the following routine builds up the $J \times 2 \cdot (J-1)$-matrix $\boldsymbol{\Gamma}$ (cf. (4.9)):

```
# model matrix GAMMA for P-WLS
build.G <- function(gamma){
  m <- length(gamma)
  GAMMA1  <- rbind(diag(gamma[2:m]), rep(0, m-1))
  diag(GAMMA1[2:m, ]) <- -gamma[2:m]
  GAMMA2  <- rbind(diag(-gamma[1:(m-1)]), rep(0, m-1))
```

```
  diag(GAMMA2[2:m, ]) <- gamma[1:(m-1)]
  GAMMA <- cbind(GAMMA1, GAMMA2)
  return(GAMMA)
}
```

A vector of misreporting probabilities $\boldsymbol{p}$ is needed to build up the matrix $\boldsymbol{Q}$ in equation (4.13), which will be used in the constrained WLS problem. In R:

```
# Penalty matrix for the P-WLS (named Q)
build.Q <- function(p){
  diagonal <- 1 / (1e-06 + abs(p))
  Q <- diag(as.vector(diagonal))
  return(Q)
}
```

The small number `1e-06` is externally supplied, as mentioned in Section 4.3.3.

Given a parameter κ (`kappa`), the system of equations in (4.12) can be solved. From the previous routine, expected values for the observation ($\boldsymbol{\mu}$, `mu`) and for the latent distribution ($\boldsymbol{\gamma}$, `gamma`) and misreporting probabilities ($\boldsymbol{p}$, `p`) are already computed.

The following routine updates the response of the WLS regression, $\boldsymbol{y} - \boldsymbol{\gamma}$ (here `r.wls`), the weight matrix ($\boldsymbol{W}$, `W.wls`), the RHS (`rhs.wls`) and LHS (`lhs.wls`) of equation (4.12), without and with the shrinkage matrix. The L_1 model for the misreporting probabilities $\boldsymbol{p}$ is finally solved.

```
# Updating the constrained WLS
UpdateWLS <- function(y, gamma, mu, kappa, p){
  r.wls   <- y - gamma
  GAMMA   <- build.G(gamma)
  W.wls   <- diag(as.vector(1 / mu))
  rhs.wls <- t(GAMMA) %*% W.wls %*% r.wls
  lhs.wls <- t(GAMMA) %*% W.wls %*% GAMMA
  Q       <- build.Q(p=p)
  lhspQ   <- lhs.wls + kappa*Q
  p       <- solve(lhspQ, rhs.wls)
  return(list(p=p))
}
```

The $2 \cdot (J-1)$ vector of misreporting probabilities $\boldsymbol{p}$ is computed and can be used as an argument for computing the matrix $\boldsymbol{C}$ as explained above. Given a (λ, κ)-combination, the complete procedure described in this section is reiterated until convergence.

In order to speed up the grid-search over the smoothing parameters, an elegant and efficient grid-search is employed: starting for the largest values of λ and κ, every new iterative procedure above described will use a previously estimated linear predictor $\boldsymbol{\eta}$ as starting values for the new (λ, κ)-combination. In this way, at the beginning of the process, a strong penalty is given for the smooth latent distribution and the misreporting probabilities. This will produce rough estimates

fairly quickly and then estimates with smaller λ and κ will only refine already fitted values, saving a lot of time.

4.5 Simulation and applications

4.5.1 Simulation study

To demonstrate the performance of the model, Camarda et al. (2008b) applied it to simulated scenarios with a bimodal true distribution. In this section, we present the outcomes of the model for two different scenarios. The first is a distribution in which the observations are redistributed according to a simple preference pattern. We refer to it as *simple* simulation setting. The second scenario mimics a typical demographic age-at-death distribution from a Gompertz model. For abbreviation, we refer to it as *demographic* simulation setting.

Simple simulation setting

Figure 4.2 shows one possible true distribution, i.e. the vector $\boldsymbol{\gamma}$ together with the simulated $\boldsymbol{y}$ such that $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$ and the estimated values $\hat{\boldsymbol{\gamma}}$. The assumed digit preference in this example attracted additional observations to 5 and 10 from both neighboring categories (see Table 4.1).

Transfer pattern	$4 \rightarrow 5$	$6 \rightarrow 5$	$9 \rightarrow 10$	$11 \rightarrow 10$
Probabilities	0.35	0.25	0.30	0.20

Table 4.1: Choice of transfer patterns for the simulation setting in Section 4.5.1.

The estimated values $\hat{\boldsymbol{\gamma}}$ were obtained from the optimal combination of (λ, κ), as picked from the AIC profile shown in Figure 4.3, left image. The image on the right hand-side demonstrates the effect of the L_1 penalty. On the horizontal axis, the value of $\log \kappa$, i.e. the weight of the L_1 penalty, is given. For big values of κ all proportions p_{jk} are shrunk to zero. For small values of κ most proportions are far too large, but for increasing values of κ many of them are quickly damped down to zero, leaving the important ones in the model. The optimally chosen $\hat{\kappa}$ practically selects the true proportions, which are depicted by the horizontally dashed lines. Note that the only assumption that is made about the underlying true distribution is smoothness.

As pointed out in Section 4.3.3, the model actually estimates $2 \cdot (J-1)$ misreporting proportions, which have not been restricted to be positive. A negative value of p_{jk} implies that category j *receives* a negative proportion of γ_k, that is, digit preference actually moves observations *away* from category j to k, but the amount is expressed as a proportion of the receiving category k. This seemingly paradoxical behavior is a consequence of the L_1 penalty: depending on whether $\gamma_j < \gamma_{j+1}$ or $\gamma_j > \gamma_{j+1}$, that is, whether the true distribution is increasing or decreasing at γ_j, the same preference leads to a smaller L_1 penalty when expressed via $p_{j,j+1}$ or $p_{j+1,j}$, one of them being negative.

However, a final result, in which the net proportions are stated as positive numbers, is desirable. Therefore, the following transformation is used to convert $2 \cdot (J-1)$ parameters to $J-1$

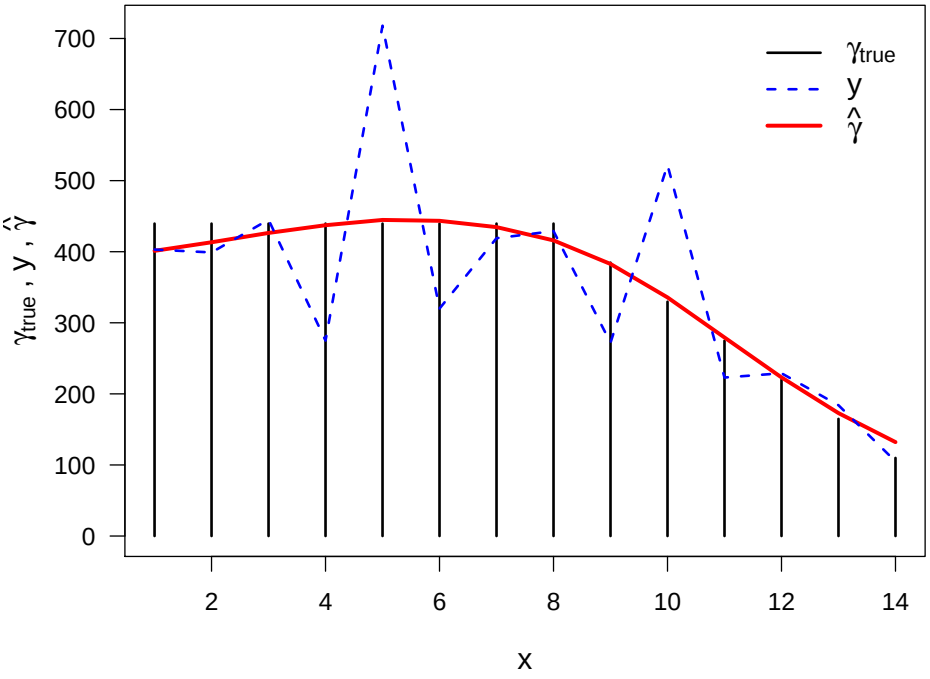

Figure 4.2: Raw data, true values and estimates for simulated data (*simple* scenario).

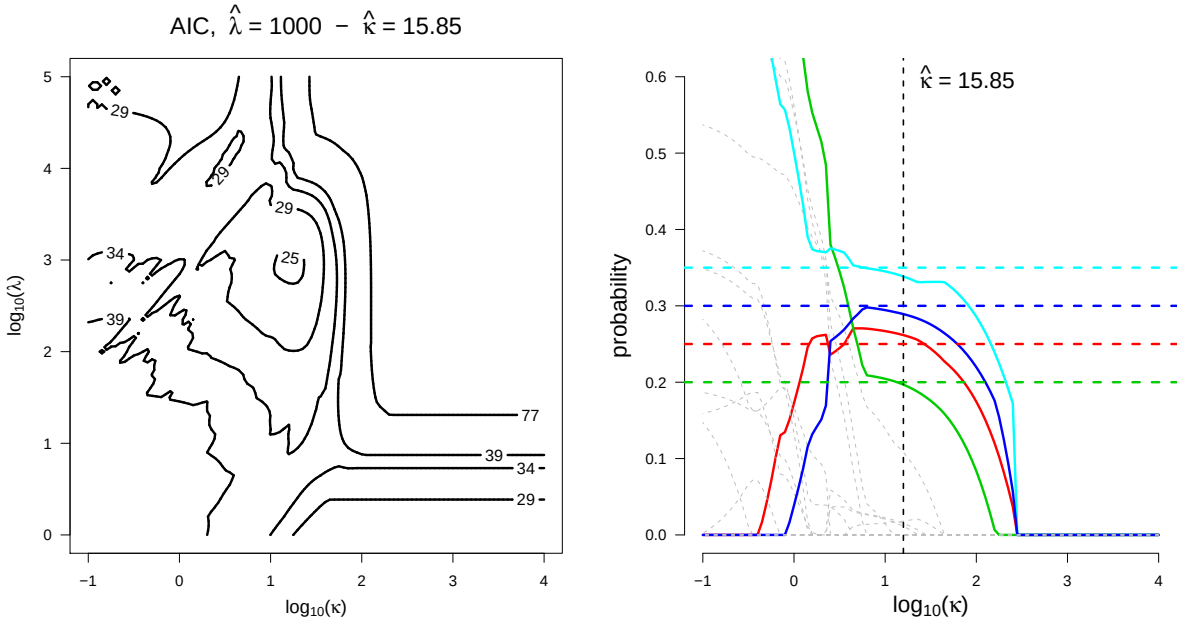

Figure 4.3: Left panel: AIC contour plot for the simulated data in Figure 4.2. Right panel: change of estimated misreporting probabilities with κ. The probabilities that are non-zero in the simulation are represented by thick and colored lines, the zero probabilities by thin gray lines.

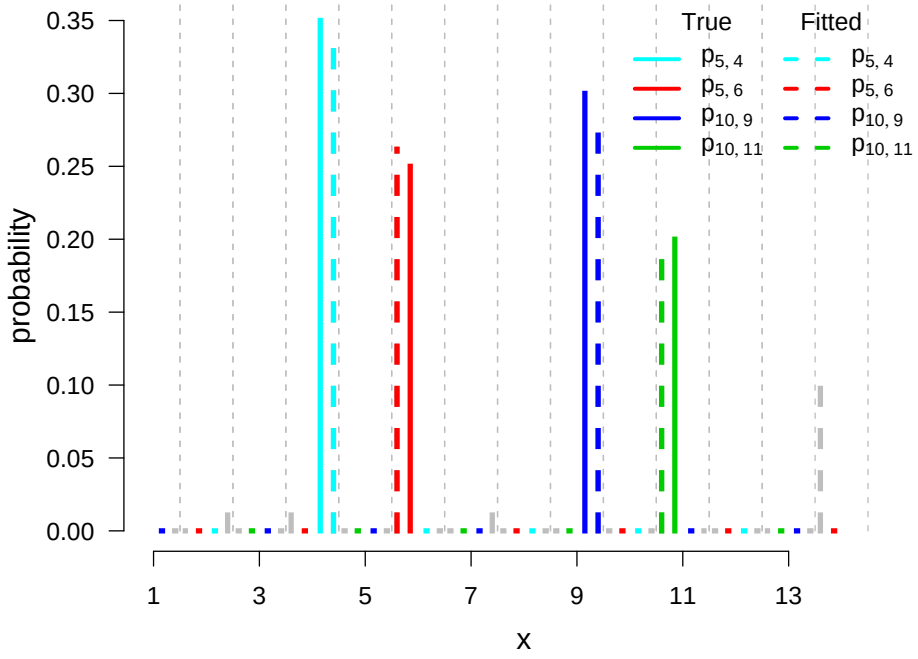

Figure 4.4: True misreporting probabilities and estimates for simulated data in Figure 4.2.

positive proportions:

$$\text{if} \quad p_{j,j-1} < 0 \quad \Rightarrow \quad p_{j-1,j} = -\frac{\delta_j}{\gamma_j} \quad \text{and} \quad p_{j,j-1} = 0$$
$$\text{if} \quad p_{j-1,j} < 0 \quad \Rightarrow \quad p_{j,j-1} = \frac{\delta_j}{\gamma_{j+1}} \quad \text{and} \quad p_{j-1,j} = 0$$

for $j = 2, \ldots, J-2$ and where $\delta_j = \mu_j - \gamma_j + \gamma_j \cdot c_{j-1,j} - \gamma_{j-1} \cdot c_{j,j-1}$. This procedure is simplified for the first step $j = 1$ and the last $j = J-1$.

The right image in Figure 4.3 shows these transformed and hence positive estimates. Additionally, Figure 4.4 summarizes true and estimated misreporting probabilities for the simulation example.

Demographic simulation setting

The distribution used to simulate counts which mimic actual death counts is the Gompertz distribution (see Section 1.4). Specifically, the true latent distribution is given by the probability density function of the Gompertz in equation (1.5) (p. 8) with parameters $a = 0.00095$ and $b = 0.08$.

We applied to the true latent distribution, $\boldsymbol{\gamma}$, a preference pattern which attracted additional observations to ages ending in 5 and 0, with probabilities always equal to 0.2 and 0.3, respectively.

Figure 4.5 presents the true latent distribution $\boldsymbol{\gamma}$ from this Gompertz distribution along with the simulated data and the fitted values. In this particular case, the smoothing parameters λ and

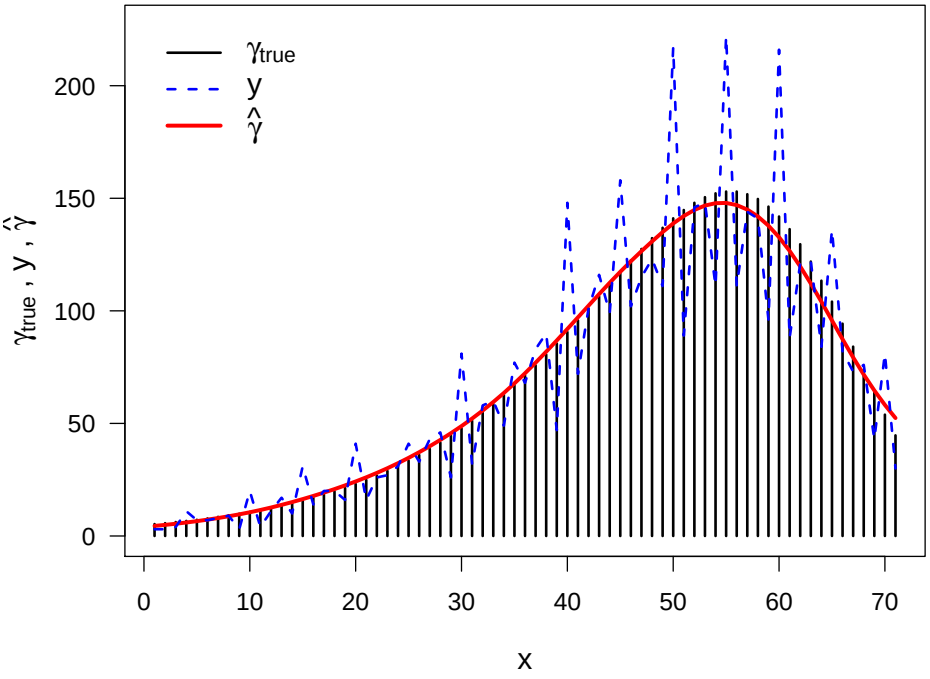

Figure 4.5: Raw data, true values and estimates for simulated data (*demographic* scenario).

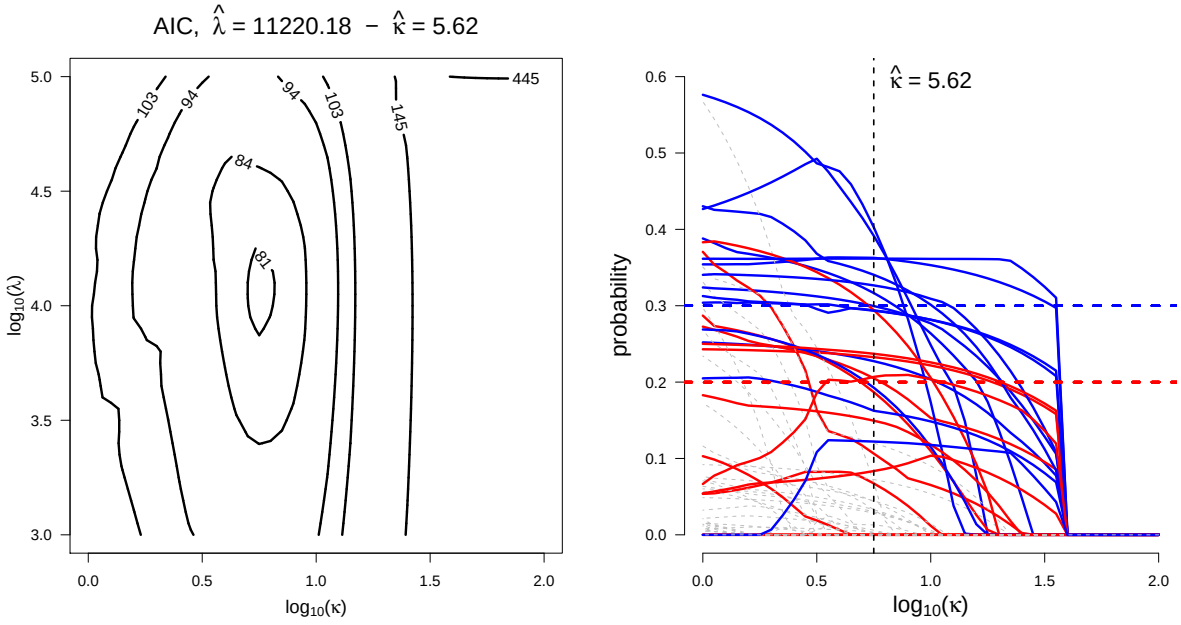

Figure 4.6: Left panel: AIC contour plot for the simulated data in Figure 4.5. Right panel: change of estimated misreporting probabilities with κ. The probabilities that are non-zero in the simulation are represented by thick and colored lines, the zero probabilities by thin gray lines. The values picked by $\hat{\kappa}$ are plotted with the same colors in Figure 4.7

κ are equal to 11,220 and 5.62, respectively. The left panel of Figure 4.6 presents the AIC profile for this simulation setting. The right panel shows the behavior of the misreporting probabilities over the values of $\log \kappa$. With ticker and colored lines we depicted the misreporting probabilities which are expected to be different from zero in this simulation study.

Figure 4.7 shows the estimated misreporting pattern: the probabilities detected by the model are practically only the ones at ages ending with 0 and 5 (depicted with thicker and colored lines) and the estimated values are close to 0.2 and 0.3. Exceptions can be found for younger ages in which the observations that have been redistributed are too few for being estimated properly.

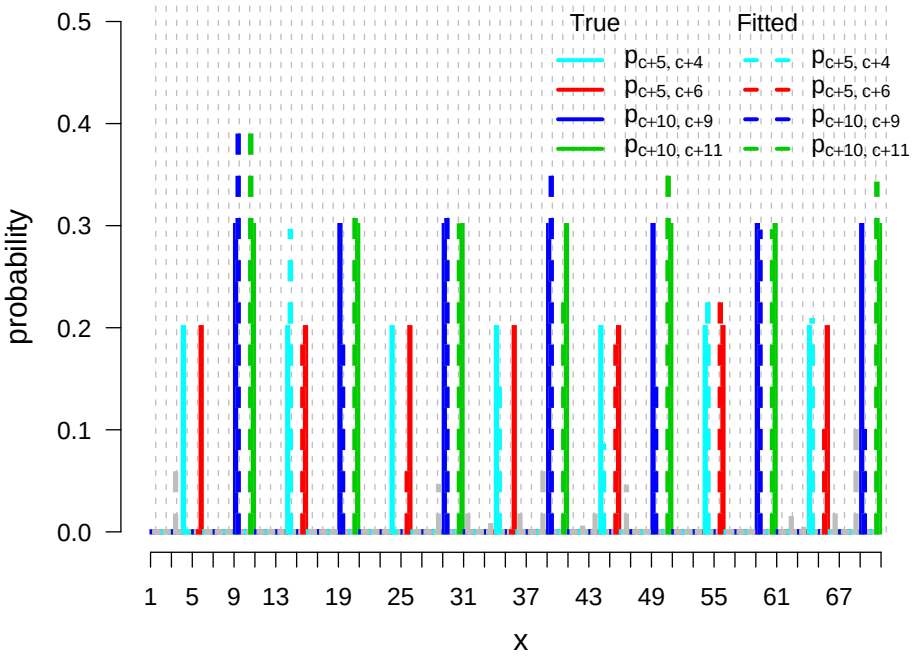

Figure 4.7: True misreporting probabilities and estimates for simulated data in Figure 4.2. The values c in the legend stands for the sequence $c = \{0, 10, 20, 30, 40, 50, 60\}$.

4.5.2 Portuguese ages at death

For the Portuguese age-at-death distribution introduced in Section 4.1, the model produces the results shown in Figure 4.8. The smooth fitted curve shows a smooth density without any age heapings. The AIC is minimized for λ and κ equal to 10^4 and 15.85, respectively.

The misreporting probabilities are portrayed in Figure 4.9. As expected, digit preference mainly attracts observations to ages that end in 5 or 10, the latter ones showing the strongest effects. The amount of misreporting increases with age, and this fits well with the demographic experience that accurate age reporting is more problematic at the high ages. Also, for ages that are multiples of 10, there is a slightly higher tendency to receive counts from their respective right neighbors.

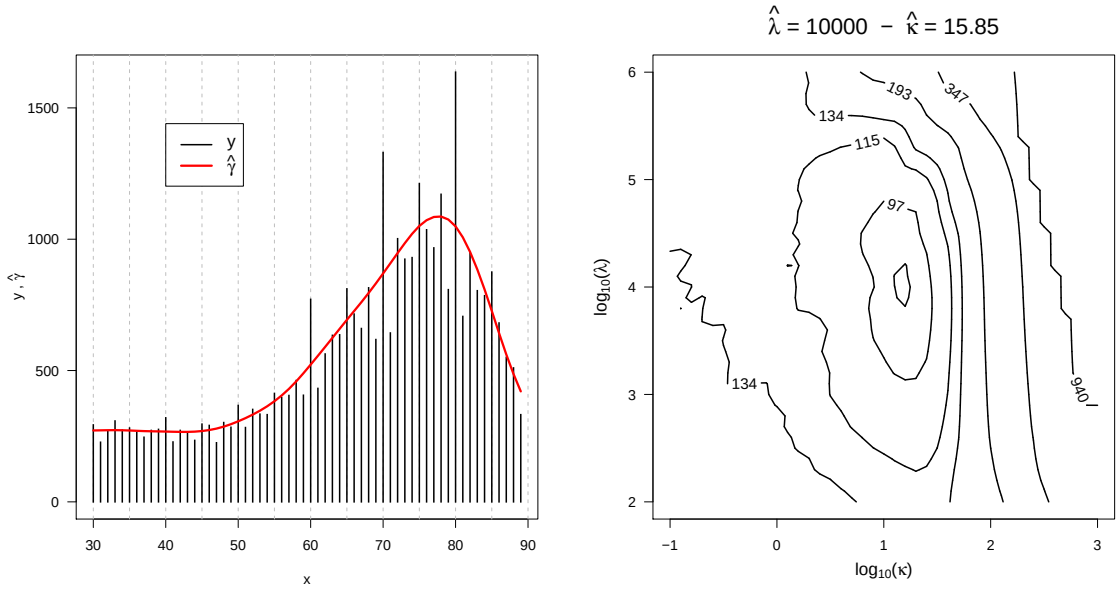

Figure 4.8: Results for the Portuguese data, cf. Figure 4.1. Observed and estimated distribution of age at death (left panel). AIC contour plot (right panel).

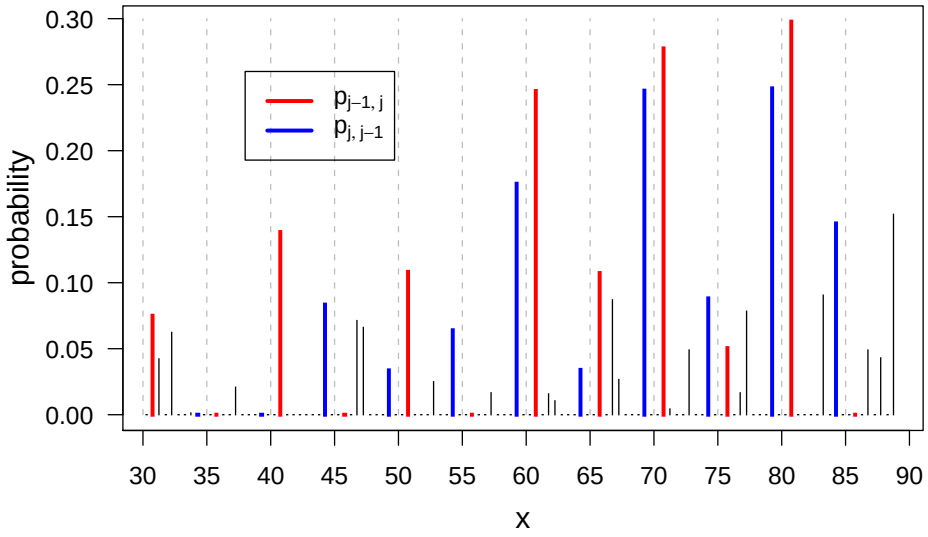

Figure 4.9: Misreporting probabilities for the Portuguese data. Probabilities to digits multiples of 5 and 10 are depicted in thicker and colored lines.

As pointed out by Camarda et al. (2008b), the model is computationally quite intensive when J is considerably large. In their paper, an example had $J = 204$. With mortality data J is commonly smaller and, for instance, for the Portuguese age-at-death distribution, J is only equal to 60 with $J+2(J-1) = 178$ parameters. In any case, the penalties for the latent distribution and the misreporting probabilities worked properly in reducing the effective dimensions, as well as in capturing the actual weight distribution with an impressive precision. Based on equations (4.15) and (4.16), the effective dimension of the fitted model for the Portuguese data is equal to 74.

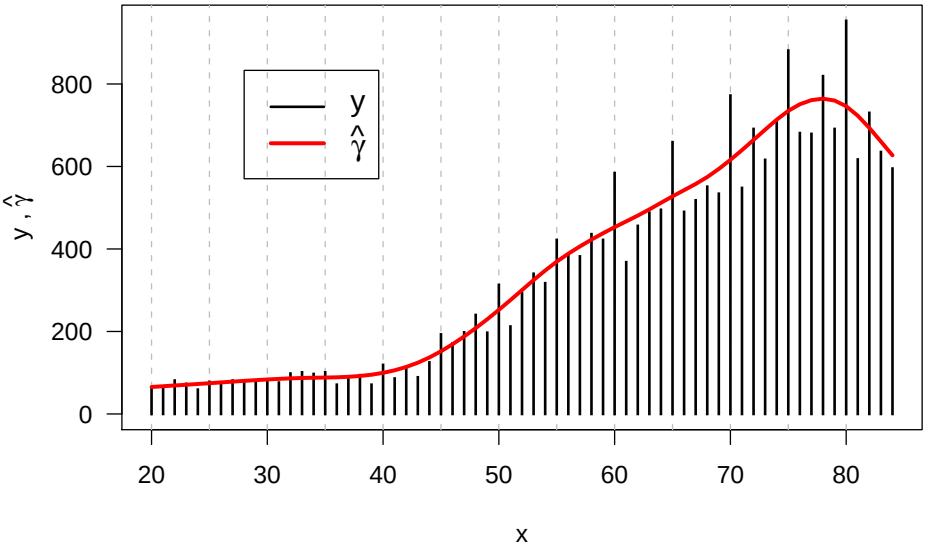

Figure 4.10: Observed and estimated distribution of age at death for the Greek data.

4.5.3 Greek ages at death

For a different purpose than modeling the misreporting pattern, Kostaki and Panousis (2001) presented demographic data with evident age heaping. Specifically the age-at-death distribution for the Greek female population in 1960, from age 20 to 84 shows systematic peaks at ages ending in 0 and, less prominently, 5, as shown in Figure 4.10.

Also in this case, the model proposed by Camarda et al. (2008b) is a suitable statistical tool for extracting the latent distribution (see Figure 4.10), as well as the pattern of misclassification as demonstrated in Figure 4.11.

4.6 More general misreporting patterns

Although this model is in many respects very flexible, the fact that observations can only shift to their nearest neighbors is rather limiting. In the original paper, Camarda et al. (2008b) overcame this seemingly over-simplistic assumption showing a modified simulation example.

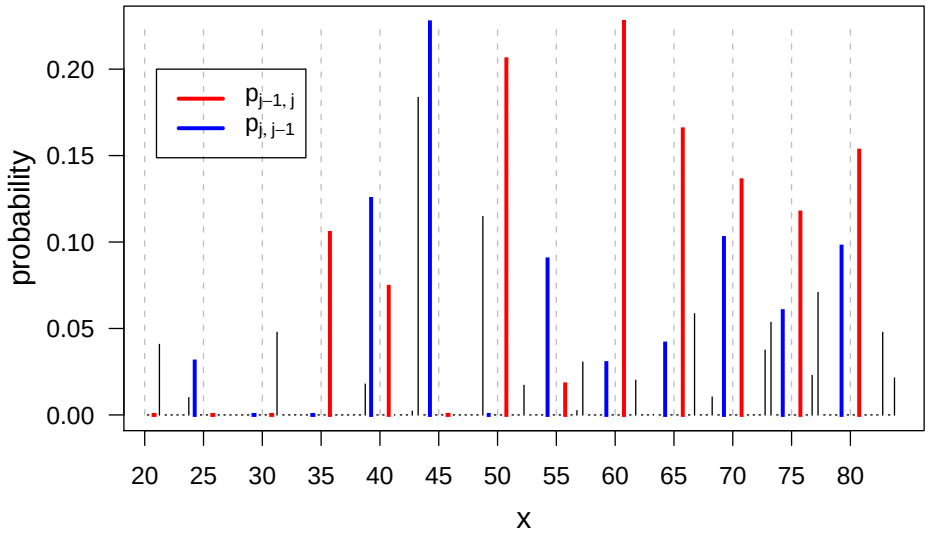

Figure 4.11: Misreporting probabilities for the Greek data. Probabilities to digits multiples of 5 and 10 are depicted in thicker and colored lines.

In this simulation setting, specific digits could attract observations both from their next neighbors and also from digits that are two steps away. The model has thus reduced the two-step misreporting probabilities to two one-step probabilities, i.e. instead of shifting the corresponding proportions in one sweep by two steps, which the original model does not provide for, they get assigned to their next neighbors first. However, these proportions then get stacked on top of the one-step estimates to the preferred target digits. Hence the original model can "decompose" more complex preference patterns into subsequent simpler steps.

Nevertheless the interpretation of the estimated misreporting probabilities is not as straightforward as in the one-step probabilities examples. Here we directly generalize the model starting from the compositional matrix $\boldsymbol{C}$ (cf. (4.2)).

We include a more general pattern of misreporting, i.e. allow for exchanges between digits that are more than one category apart. For a simple case, in which we include two categories and $J = 8$, the matrix $\boldsymbol{C}$ needs to be modified as follows

$$\boldsymbol{C} = \begin{pmatrix} 1-p_{21}-p_{31} & p_{12} & p_{13} & 0 \\ p_{21} & 1-p_{12}-p_{32}-p_{42} & p_{23} & p_{24} \\ p_{31} & p_{32} & 1-p_{13}-p_{23}-p_{43}-p_{53} & p_{34} \\ 0 & p_{42} & p_{43} & 1-p_{24}-p_{34}-p_{54}-p_{64} \\ 0 & 0 & p_{53} & p_{54} \\ 0 & 0 & 0 & p_{64} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
p_{35} & 0 & 0 & 0 \\
p_{45} & p_{46} & 0 & 0 \\
1-p_{35}-p_{45}-p_{65}-p_{75} & p_{56} & p_{57} & 0 \\
p_{65} & 1-p_{46}-p_{56}-p_{76}-p_{86} & p_{67} & p_{68} \\
p_{75} & p_{76} & 1-p_{57}-p_{67}-p_{87} & p_{78} \\
0 & p_{86} & p_{87} & 1-p_{68}-p_{78}
\end{pmatrix}.$$

Also in this example, the diagonal elements $c_{jj} = 1 - p_{j-2,j} - p_{j-1,j} - p_{j+1,j} - p_{j+2,j}$ specify the proportions of the γ_j that did not get redistributed. Moreover, as in the original model, p_{jk} denotes the proportion of γ_k that is moved from age k to age j. The additional feature that we propose is that two-step transitions are allowed, implying that $p_{jk} = 0$ only for $|j-k| > 2$.

For instances, p_{42} indicates the probability that death counts from age 2 will move to age 4. In this way, it is possible to estimate a more general pattern of misreporting probabilities. On the other hand, the number of parameters increases enormously, i.e. we need to estimate $2(J-1)+2(J-2)$ probabilities as well as the true latent distribution. Furthermore, $\mathbf{\Gamma}$ (cf. (4.9) in the previous model) becomes a $J \times [2(J-1)+2(J-2)]$ matrix with the following complex structure:

$$\mathbf{\Gamma} = \begin{pmatrix}
\gamma_2 & 0 & 0 & 0 & 0 & -\gamma_1 & 0 & 0 & 0 & 0 \\
-\gamma_2 & \gamma_3 & 0 & 0 & \vdots & \gamma_1 & -\gamma_2 & 0 & 0 & \vdots \\
0 & -\gamma_3 & \gamma_4 & 0 & \vdots & 0 & \gamma_2 & -\gamma_3 & 0 & \vdots \\
0 & 0 & -\gamma_4 & \ddots & \vdots & 0 & 0 & \gamma_3 & \ddots & \vdots \\
0 & 0 & 0 & \ddots & \vdots & 0 & 0 & 0 & \ddots & \vdots \\
\vdots & \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \vdots & \ddots & 0 \\
\vdots & \vdots & \vdots & \ddots & \gamma_J & \vdots & \vdots & \vdots & \ddots & -\gamma_{J-1} \\
0 & 0 & 0 & 0 & -\gamma_J & 0 & 0 & 0 & 0 & \gamma_{J-1}
\end{pmatrix}$$

$$\begin{pmatrix}
\gamma_3 & 0 & 0 & 0 & 0 & -\gamma_1 & 0 & 0 & 0 & 0 \\
0 & \gamma_4 & 0 & 0 & \vdots & 0 & -\gamma_2 & 0 & 0 & \vdots \\
-\gamma_3 & 0 & \gamma_5 & 0 & \vdots & \gamma_1 & 0 & -\gamma_3 & 0 & \vdots \\
0 & -\gamma_4 & 0 & \gamma_6 & \vdots & 0 & \gamma_2 & 0 & \ddots & \vdots \\
0 & 0 & -\gamma_5 & \ddots & \vdots & 0 & 0 & \gamma_3 & \ddots & \vdots \\
\vdots & \vdots & 0 & \ddots & \gamma_J & \vdots & \vdots & \vdots & \ddots & -\gamma_{J-2} \\
\vdots & \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & 0 & 0 & -\gamma_J & 0 & 0 & 0 & 0 & \gamma_{J-2}
\end{pmatrix}.$$

Consequently the vector of misreporting probabilities used in the constrained WLS component

of the model (cf. Section 4.3.3) can be written in the following way:

$$\boldsymbol{p} = (p_{12}, p_{23}, \ldots, p_{J-1,J}; p_{21}, p_{31}, \ldots, p_{J,J-1}; p_{13}, p_{24}, \ldots, p_{J-2,J}; p_{31}, p_{42}, \ldots, p_{J,J-2};)^T .$$

An essential aspect of this extension lies in the fact that, despite the matrices $\boldsymbol{C}$ and $\boldsymbol{\Gamma}$, we can still use formulas and algorithms of the original model. Specifically, we will use the system of equations given in (4.7) and (4.11) for the latent distribution and the misreporting probabilities, respectively. Furthermore the AIC is still able to select a suitable (λ, κ)-combination (cf. Fig. 4.12 in Section 4.6.1).

No changes are needed in the constrained WLS component. The "lasso" model can still guarantee a selection of the misreporting parameters. The L_1 penalty can cope with commonly called "large p, small n" problems such as our extension, which contains a huge amount of misreporting probabilities.

Also in the extension of the model, we have not restricted misreporting probabilities to be positive. Hence we need to generalize the transformation presented in Section 4.5.1 in order to obtain as final results only the net proportions as positive numbers. We separated the effects of the misreporting probabilities from the next neighbors and the probabilities from digits that are two steps away. Hence, we first construct a matrix $\boldsymbol{C}^1$, which includes only p_{jk} such as $|j-k| = 1$ (cf. (4.2)). In this way, we can compute the expected values as if there was only one-step misreporting, $\boldsymbol{\mu}^1$. The transformation in Section 4.5.1 is then used to obtain net proportions as positive numbers only for the next neighbors. We define the quantities δ_j as

$$\delta_j = (\mu_j - \mu_j^1) + \gamma_j \cdot c_{i-2,i} - \gamma_{j-2} \cdot c_{i,i-2} .$$

For $j = 3, \ldots, J-3$, the following transformation is applied in order to convert the $2(J-2)$ misreporting probabilities from digits that are two steps away in $J-2$ positive proportions:

$$\begin{aligned} \text{if} \quad \delta_j > 0 \quad &\Rightarrow \quad p_{j,j+2} = \frac{\delta_j}{\gamma_{j+2}} \quad \text{and} \quad p_{j+2,j} = 0 \\ \text{if} \quad \delta_j < 0 \quad &\Rightarrow \quad p_{j+2,j} = -\frac{\delta_j}{\gamma_j} \quad \text{and} \quad p_{j,j+2} = 0 \end{aligned}$$

This procedure is simplified for the first two steps $j = 1, 2$ and the last two $j = J-2, J-1$.

4.6.1 Simulation study

In this section, we demonstrate the performance of the new approach on simulated data. We use similar setting in Camarda et al. (2008b, Section 5.1).

A normal distribution is used as true latent distribution $\boldsymbol{\gamma}$. Digits 10 and 20 attract observation from categories both one and two categories away. The structure of the transfer pattern is given in Table 4.2.

Figure 4.12 (left panel) shows one possible true distribution along with the simulated $\boldsymbol{y}$ such that $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$, where $\boldsymbol{C}$ is given in equation (4.6). Estimated values $\hat{\boldsymbol{\gamma}}$ are plotted, too, and they reproduce the true smooth latent distribution extremely well. The selected smoothing

Transfer pattern	8→10	9→10	11→10	12→10	18→20	19→20	21→20	22→20
Probabilities	0.30	0.40	0.35	0.25	0.30	0.40	0.35	0.25

Table 4.2: Choice of transfer patterns for the extended simulation setting.

parameters, $\hat{\lambda}$ and $\hat{\kappa}$ are equal to 3981.07 and 19.95, respectively. The right panel of Figure 4.12 presents the AIC profile obtained for this simulation setting.

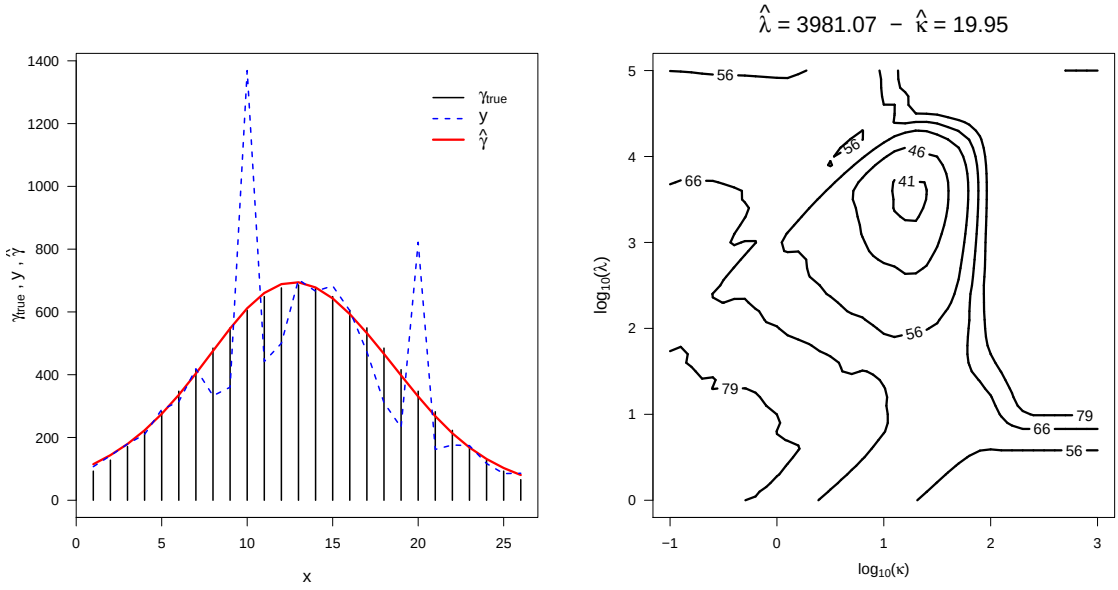

Figure 4.12: Results for simulated data in Section 4.6.1 (cf. Table 4.2). Raw data, true values and estimates (left panel). AIC contour plot (right panel).

The advantage of the generalization is particularly evident for the estimated misreporting patterns. Now we can easily disentangle the contribution of both next neighbors and the misreported counts from two categories apart. Figure 4.13 presents both true misreporting probabilities, as in Table 4.2 and the fitted ones. The similarity between true and fitted values is remarkable, despite the complexity of the misreporting pattern.

Figure 4.14 shows the effect of the L_1 penalty on the generalization of the model. The total number of fitted misreporting probabilities is equal to $2(J-1)+2(J-2)$ and, after the mentioned transformation, we have only $2 \cdot J - 3$ positive proportions. The constrained WLS can cope exceptionally well with this amount of parameters: the optimally chosen $\hat{\kappa}$ practically selects only the true proportions (cf. Table 4.2).

4.6.2 Application of actual data

In Figure 4.1, we presented age-at-death distribution for Portuguese females in 1940 from age 30 to 89. In Section 4.5.2, we applied the model proposed by Camarda et al. (2008b) to these data (see Fig. 4.8 and 4.9). Nevertheless, it is extremely likely that ages such as 30, 45, 65, 75 and 85 also attract death counts from two categories away. Therefore generalization presented in Section 4.6 was applied to these data.

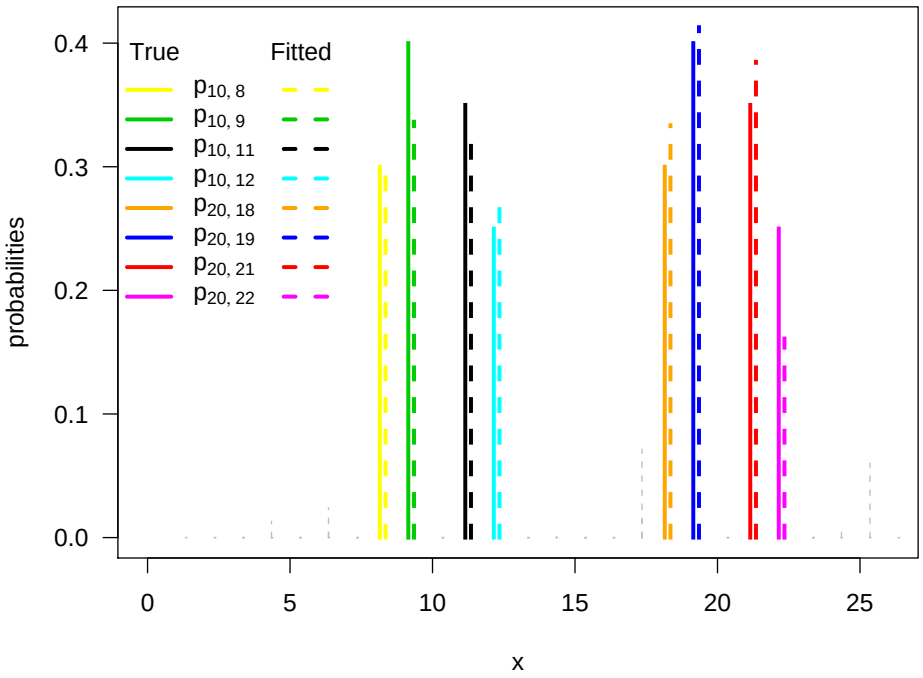

Figure 4.13: True and fitted misreporting probabilities for simulated data in Section 4.6.1 (cf. Table 4.2).

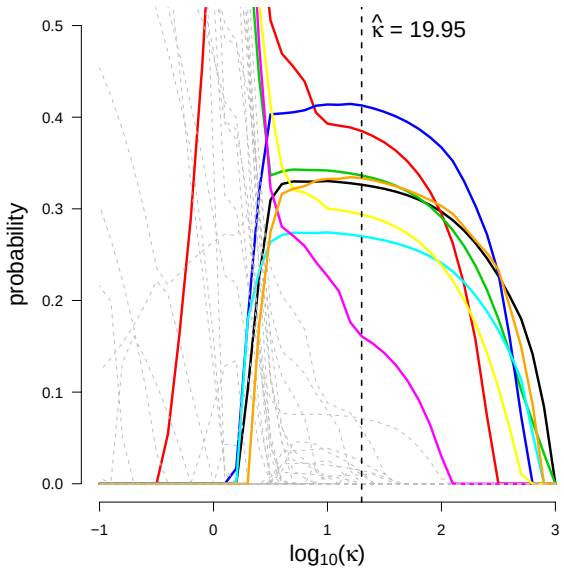

Figure 4.14: Change of estimated misreporting probabilities with κ. The probabilities that are non-zero in the simulation are represented by thick and colored lines, the zero probabilities by thin gray lines. Simulated data in Section 4.6.1 (cf. Table 4.2).

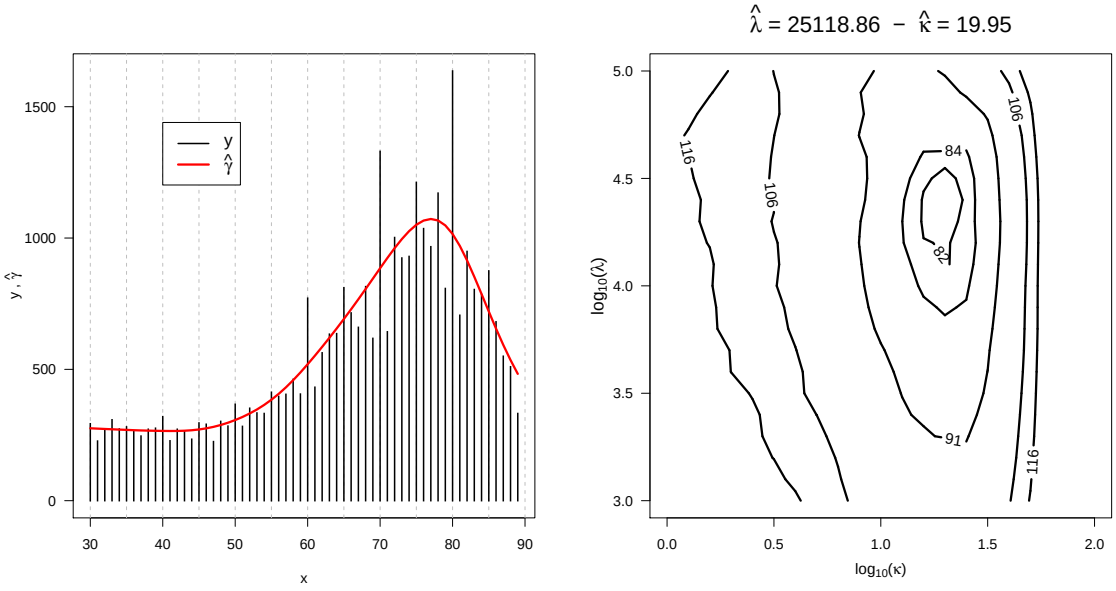

Figure 4.15: True and fitted misreporting probabilities for simulated data in Section 4.6.1 (cf. Table 4.2).

Figure 4.15 (left panel) shows the actual age-at-death distribution along with the fitted latent distribution. The difference between these outcomes and the fitted values from the simpler model (cf. Fig. 4.8) is negligible. On the other hand, with the proposed extension, we are able to separate the misreporting probabilities from the next neighbors and the misreporting probabilities from digits that are two steps away. The right panel of Figure 4.15 shows the AIC profile which presents a clear minimum and selects smoothing parameters λ and κ equal to 25,119 and 19.95, respectively.

Since we are dealing with $2 \cdot 60 - 3 = 117$ net and positive misreporting proportions, we used an alternative graphical device for portraying such a large amount of parameters. First, it would be more convenient to display the probabilities in four different plots: p_{jk} such as $|j-k| = 1$ (one-step misreporting from right and left digits) and p_{jk} such as $|j-k| = 2$ (two-step misreporting from right and left digits). Moreover, instead of plotting histograms, as in the previous sections, we use shaded contour maps similar to the ones used for mortality surfaces. Here, we employ different colors for different levels of misreporting probabilities.

Figure 4.16 shows the fitted misreporting probabilities for the Portuguese data over the units and the age decades. Higher probabilities are depicted with darker colors, whereas light grey indicates misreporting probabilities equal to zero. It is easily seen that digits ending with 10 and 5 shows the cells with the darkest colors and counts are misreported from both one and two digits away. Death counts are particularly misreported downward. Therefore, we have darker colored cells in the left panels. The introduction in the model of misreporting probabilities from more than one digit away helps to capture misstatements in ages ending with 5 and especially at older ages. A tendency of misreporting age-at-death toward ages ending with 2 and 8 is also evident.

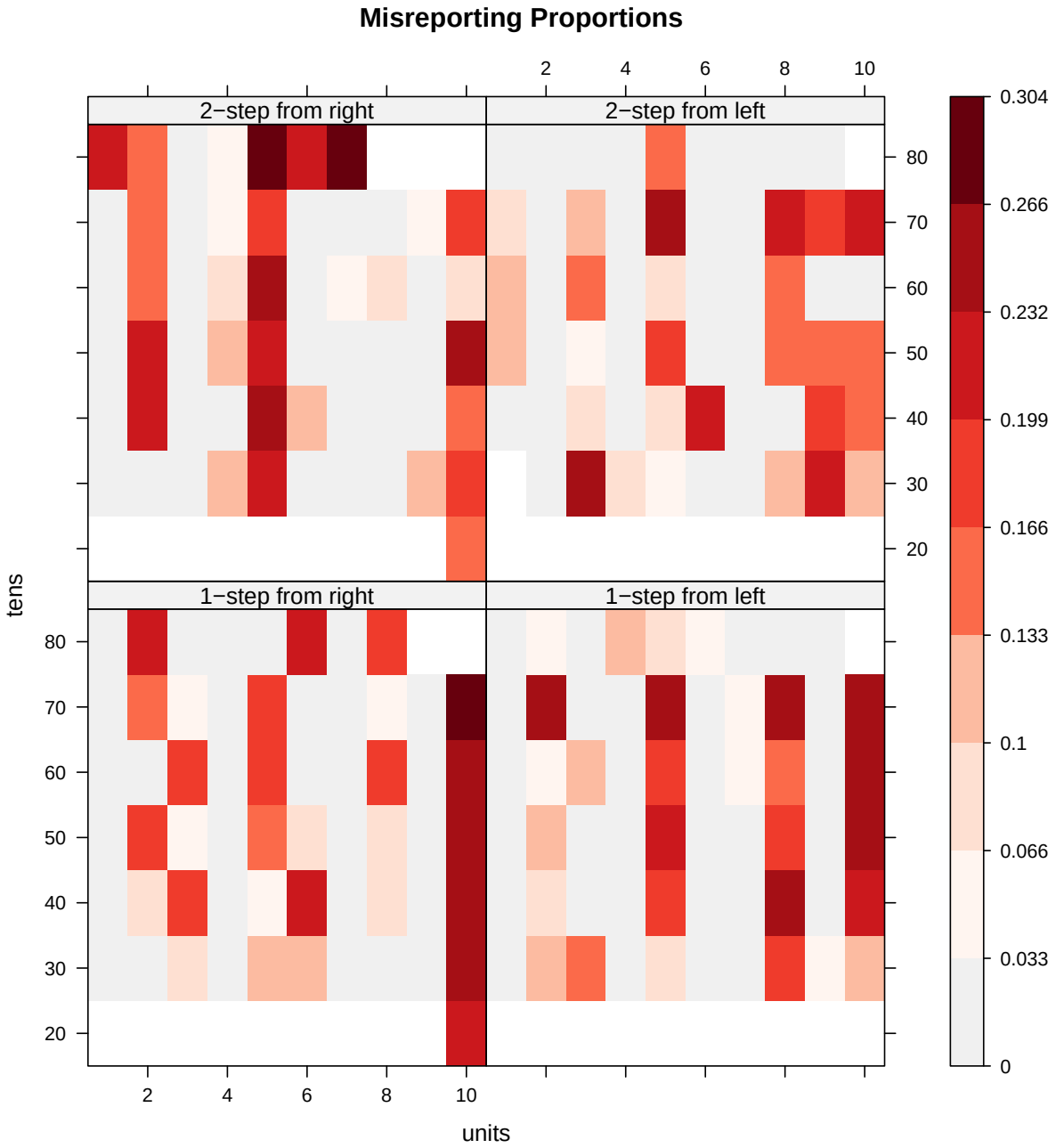

Figure 4.16: Fitted misreporting probabilities over units and tens of ages for Portuguese data (cf. Fig. 4.1). Generalization of the model presented in Section 4.6. Higher probabilities are depicted with darker colors. Light grey indicates misreporting probabilities equal to zero.

4.7 Further extensions

The method proposed by Camarda et al. (2008b) demonstrates how digit preferences can be modeled by combining the composite link model with the concept of penalized likelihood. The only assumption that is made about the underlying true distribution is smoothness. The approach directly addresses the process that leads to heaping of certain values. Extracting the latent distribution will be most important in many applications, however, the pattern of misclassification may also be of interest in itself. The model presented by Camarda et al. (2008b) goes beyond the mere quantification of digit preference provided by many indices and allows the analysis of both aspects.

The misreporting pattern was allowed to partly redistribute observations from any digit to its adjacent neighbors. Again a penalty, in this case an L_1 penalty, restrains the problem and makes estimation feasible. Allowing this rather flexible preference pattern the tendency to misreport need not be the same for identical end-digits, but may vary over the measurement range, which is often seen in real data.

Though the model proposed by Camarda et al. (2008b) allows estimation of the latent distribution with more complicated transfer patterns, an evident limitation lies in the assumption that misreporting will only move observations to the next neighboring digits. We propose an extension which can include more complex preference patterns, i.e. allow for exchanges between digits that are more than one category apart.

The simulation study and actual application presented in Sections 4.6.1 and 4.6.2 deal only with exchanges from one and two digits away. Further generalizations are also possible by plugging additional misreporting probabilities into the $\boldsymbol{C}$ matrix. Naturally, this kind of extension increases enormously the number of parameters and their interpretability.

We envision a different generalization of the model. In case of mortality data with age heaping, digit preferences may improve over time. In this case, the transfer probabilities for different years are expected to change smoothly. This trend can be handled by an additional penalty that controls the temporal pattern in the misreporting pattern. Two-dimensional smoothing with penalized likelihood, such as presented in Section 2.2, is a natural approach in this context.

Chapter 5

A Warped Failure Time Model for Human Mortality

Life-span distributions can be described by three alternative, albeit interchangeable measures (see Section 1.2): the probability density function, the survival function, and the hazard function. Similarly, mortality developments may not only be described by studying death rates, but also by investigating the corresponding frequency distribution of the ages-at-death. Even though we analyze the same phenomenon, different views on the same problem may shed additional light on our understanding of changes in human mortality.

Traditionally, mortality dynamics are studied by direct investigation of the hazard over time (see models in Chapter 1). A different approach would be to ask how the age-axis would need to be transformed so that one age-at-death distribution would conform to another (which could be a different country at the same point in time, or the same country at different points in time).

This approach is similar to the accelerated failure time (AFT) model which is routinely used in failure time analysis of technical devices (among others, see Kalbfleisch and Prentice, 2002, ch. 7). In AFT models the time-axis is linearly transformed, implying uniformly slower or faster aging, usually introduced by some experimental conditions.

However, the assumption that all cohorts postpone death or speed up their lives at a constant rate across the age range is too simplistic for human mortality studies. Mortality changes are not driven by simple gains in longevity that shift and uniformly re-scale the age-at-death distributions from one year to another. Therefore, more general transformations will have to be considered.

The idea of transforming the axis of the independent variable nonlinearly to achieve good alignment of functions has been proposed in diverse fields and in regression setting is commonly called "warping". Two procedures have a long tradition. Marker registration or procrustes analysis involve identifying the timing of specified features in the function, and then transforming time so that these marker events occur at the same time. For a fuller treatment, we refer the reader to Bookstein (1991). Dynamic time warping is a registration technique frequently used in engineering literature. This methodology aims to estimate the shift or warping function from one curve to another to align two functions. An early reference on the methodology is given by Sakoe and Chiba (1978). Kneip and Gasser (1988, 1992) and Wang and Gasser (1997) give statistical details and different perspectives on this approach.

Silverman (1995) developed a technique that does not require markers in a "Functional Data Analysis" (FDA) framework (Ramsay and Silverman, 2005), and Ramsay and Li (1998) made use of the smooth monotone function introduced by Ramsay (1998b) for a more general approach. On the other hand, Dijksterhuis and Eilers (1997), and also Ledauphin et al. (2006) proposed procedures to project time intensity curves on a nonparametric prototype curve with a linear registration of time and intensity domains. A parametric model for aligning chromatograms has been proposed by Eilers (2004a).

In classic regression, analysis transformation of the independent axis to align actual data to the Normal distribution has been studied since Tukey (1957) and Box and Cox (1964). For a review of the proposed parametric families in the literature, we refer to Sachia (1992). Alternatively, nonparametric estimation of the transformation for regression analysis has also been proposed (among others see Breiman and Friedman, 1985; Tibshirani, 1988). Transformation of distributions has been also used in density estimation methodology. Early parametric approaches are presented in Wand et al. (1991) and Ruppert and Wand (1992). Ruppert and Cline (1994) generalized the previous approaches using a smooth, monotonic transformation function.

Warping ideas have also been introduced into the analysis of mortality data by Eilers (2004b). Focusing on adult mortality Eilers (2004b) uses Normal distribution and seeks for a non-parametric transformation of the age-axis, such that the actual age-at-death distribution becomes essentially Normal. He called the model "Shifted Warped Normal" (SWaN).

Camarda et al. (2008a) recently proposed a new model for analyzing mortality development using warping ideas. They presented an extension of the accelerated failure time model for comparison of density functions and called the model Warped Failure Time (WaFT) model. The aim of the WaFT model is to estimate a warping function such that the age-axis is transformed, and one age-at-death distribution conforms to another. They assumed only smoothness for the warping function. Moreover, a penalized Poisson likelihood approach was proposed to estimate the model.

In this chapter, we present the model proposed by Camarda et al. (2008a) in detail. Specifically, we first introduce the general idea of transformation of random variables that will give the basic structure for the model. Section 5.2 illustrates the WaFT model for describing the mean of Poisson-distributed death counts. Particular emphasis is given to the representation of the warping function. Estimation of the model is covered in Section 5.3, including difference penalties for assuring smoothness of the warping function and the choice of optimal smoothing parameters. In Section 5.5, we illustrate the approach via simulated data and present some demographic applications. Conclusions and an outlook for future work are given in Section 5.6.

5.1 Comparing age-at-death distributions

To investigate the changes in mortality that lead to the different pattern, we want to transform the age-axis, such that, the two densities coincide. More specifically, we define one distribution as the target, with density $f(y)$, and want to obtain the transformation function $w(x)$ so that the

density of the other distribution, $g(x)$, conforms to the target density on the warped axis, i.e.,

$$g(x) = f(w(x)) \cdot |w'(x)|\,, \tag{5.1}$$

with $y = w(x)$.

This general rule for transforming the x-axis in order to match two densities is used in the next section to build up the model. Note that one needs to account for the derivative of the transformation function for correcting the transformed target distribution, $f(w(x))$. This correction will play an important role in choosing the representation of $w(x)$ in Section 5.2.1.

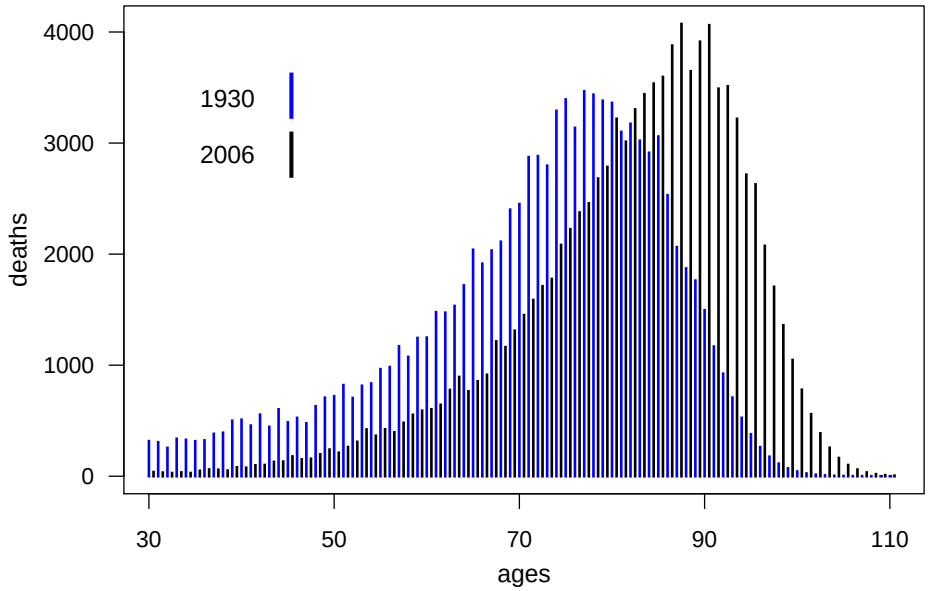

Figure 5.1: Life-table age-at-death distribution of Danish females for the years 1930 and 2006.

Figure 5.1 shows the age-at-death distributions, as derived from period life-tables (Keyfitz and Caswell, 2005), for Danish women older than 30 in 1930 and in 2006. Instead of pure death counts, Camarda et al. (2008a) used a period life-table approach for adjusting the effect of birth cohort sizes in the age-at-death distributions (see Section 1.3, page 5). In this way, they exclude exposure population in the approach.

As mentioned, a linear $w(x)$ which shifts and/or stretches the age-axis to conform one distribution to another would be too simplistic for human mortality. Figure 5.1 clearly shows this issue: for instance, the age-at-death distribution in 1930 is neither a shifted nor a stretched version of the age-at-death distribution in 2006. On the other hand, parametric assumption of the transformation function will impose too much structure on the model and may lead to misleading interpretation.

Following these considerations, Camarda et al. (2008a) suggested freeing the transformation

function $w(x)$ from any rigid shape. Therefore, the assumption will be the smoothness of $w(x)$. Moreover, they assume that the transformation function $w(x)$ is increasing. This assumption has the effect that different ages follow the same ordering after the transformation, i.e., if $x < x'$, then $w(x) < w(x')$. This assumption also guarantees that $w(x)$ can be inverted. Therefore the absolute values in (5.1) can be dropped. This approach leads to a nonlinear transformation of the independent axis which is commonly called "warping". On the other hand, we deal with a transformation of time as in the AFT models. Hence Camarda et al. (2008a) called their model Warped Failure Time (WaFT) model.

5.2 The Warped Failure Time Model

The WaFT model is not restricted to any particular target distribution, $f(x;\boldsymbol{\theta})$. The parameter(s) $\boldsymbol{\theta}$ can be fixed, but are mostly be estimated from data. In the following, we nevertheless consider the parameters as $\boldsymbol{\theta}$ fixed. Section 5.5 will present different approaches for estimating $f(\cdot)$ both parametrically and non-parametrically.

The observed death counts at age x_i, $i = 1, \ldots, n$ are denoted by y_i. Ages x_i can be seen as midpoints of particular bins which collapse all the death counts in the bin to a certain age. If p_i is the probability of finding a raw observation in cell i, i.e. in the bin around age x_i, the likelihood of the given histogram is proportional to the multinomial likelihood $\prod_{i=1}^{n} p_i^{y_i}$. Bishop et al. (1975) demonstrated that one can equivalently work with the likelihood of n Poisson variables with expectations $\mu_i = p_i y_+$, where $y_+ = \sum_i^n y_i$.

In other words, we can write the likelihood of the observed death count y_i as

$$P(y_i) = \frac{\mu_i^{y_i}\, e^{-\mu_i}}{y_i!}\,. \tag{5.2}$$

Using the relation (5.1), the values μ_i derive from the density $g(x)$ that generated the data. Therefore

$$\begin{aligned} \mu_i = E(y_i) &= \gamma \cdot f(w(x_i;\boldsymbol{\alpha}),\boldsymbol{\theta}) \cdot \frac{\partial}{\partial x} w(x_i;\boldsymbol{\alpha}) \\ &= \gamma \cdot f(w(x_i;\boldsymbol{\alpha}),\boldsymbol{\theta}) \cdot v(x_i;\boldsymbol{\alpha})\,, \end{aligned} \tag{5.3}$$

where γ is a normalizing constant such that $\sum_i y_i = \sum_i \mu_i$ and $f(\cdot)$ is the target density. The warping function $w(x_i;\boldsymbol{\alpha})$ is to be determined such that, after transforming the age-axis, the density matches the specified target.

Different from equation (5.1), the model in (5.3) does not compute the absolute values for the partial derivative of the warping function, $v(x_i;\boldsymbol{\alpha})$. This is because monotonicity of $w(x_i;\boldsymbol{\alpha})$ is assumed. We will see that, in actual demographic applications, this assumption holds and we do not have to impose any additional constraints.

5.2.1 Warping function representation

The representation of the warping function is a crucial key in building up the WaFT model (5.3). To allow for arbitrary shape of $w(\cdot)$, Camarda et al. (2008a) suggested representing the warping

function by a linear combination of K B-splines of degree q. Their knots are equally spaced by a distance h:

$$w(x;\boldsymbol{\alpha}) = \sum_{k=1}^{K} B_k^q(x)\,\alpha_k\,. \tag{5.4}$$

As showed in Section 2.1, B-splines are a base of local functions that is well-suited for a non-parametric estimation of a function. A relatively large number of knots will ensure the flexibility we need to nonlinearly transform the age-axis. Smoothness of the warping function will be enforced by a difference penalty on neighbouring coefficients α_k, as Section 5.3.2 show.

This representation naturally allows for incorporating the derivative of the warping function $w(x;\boldsymbol{\alpha})$, which is needed in the transformation approach in (5.3). Moreover, when considering a warping function for transforming the age-axis, one can be interested in $w(\cdot)$, as well as in the relative change of $w(\cdot)$. In particular, in our case x represents age, in which the first derivative of $w(\cdot)$ can be interpreted as speed.

A linear combination of B-splines can be easily derived with respect to x and the formula is given by

$$\frac{\partial}{\partial x} w(x;\boldsymbol{\alpha}) = v(x;\boldsymbol{\alpha}) = \frac{1}{h}\sum_{k=1}^{K} B_k^{q-1}(x)\,[\alpha_k - \alpha_{k+1}]\,, \tag{5.5}$$

c.f. Eilers and Marx (1996, p. 91). Note that the basis of $v(x;\boldsymbol{\alpha})$ is again a B-spline basis of a degree less than for $w(x;\boldsymbol{\alpha})$ in (5.4).

Equivalently, operating directly on the coefficients α_k, we can write (5.5) in the following way:

$$\frac{\partial}{\partial x} w(x;\boldsymbol{\alpha}) = v(x;\boldsymbol{\alpha}) = \sum_{k=1}^{K} C_k^q(x)\,\alpha_k, \tag{5.6}$$

where $C_k^q(x) = \frac{1}{h}[B_k^{q-1}(x_i) - B_{k-1}^{q-1}(x_i)]$.

As an example of a derivative from functions represented as a linear combination of B-spline, we use the simulated data presented in Section 2.1.1 (page 17 of this thesis). Figure 5.2 shows both the function itself and its derivative along with the true functions. We employ a P-spline approach to smooth the fitted coefficients, and cross-validation is used to select the smoothing parameter since Normal-distributed data are used for this simulation example (see Sections 2.1.1 and 2.1.4).

Figure 5.2 presents also the B-splines basis of degree $q = 3$ and $q = 2$ for the function and its derivative, respectively. In the latter case, the B-splines are multiplied by $\frac{1}{h}$. It is worth pointing out that the smoothing parameter selection is carried out directly on the function and information about the derivative is not taken into account in this approach.

5.3 A Penalized Poisson Likelihood Approach

In order to estimate the coefficients $\boldsymbol{\alpha}$ in (5.3), Camarda et al. (2008a) proposed a penalized Poisson likelihood approach. For the sake of clarity, we first introduce the estimation procedure of the WaFT model without any penalization, and then, in Section 5.3.2, we will present the approach used to ensure smoothness of the warping function.

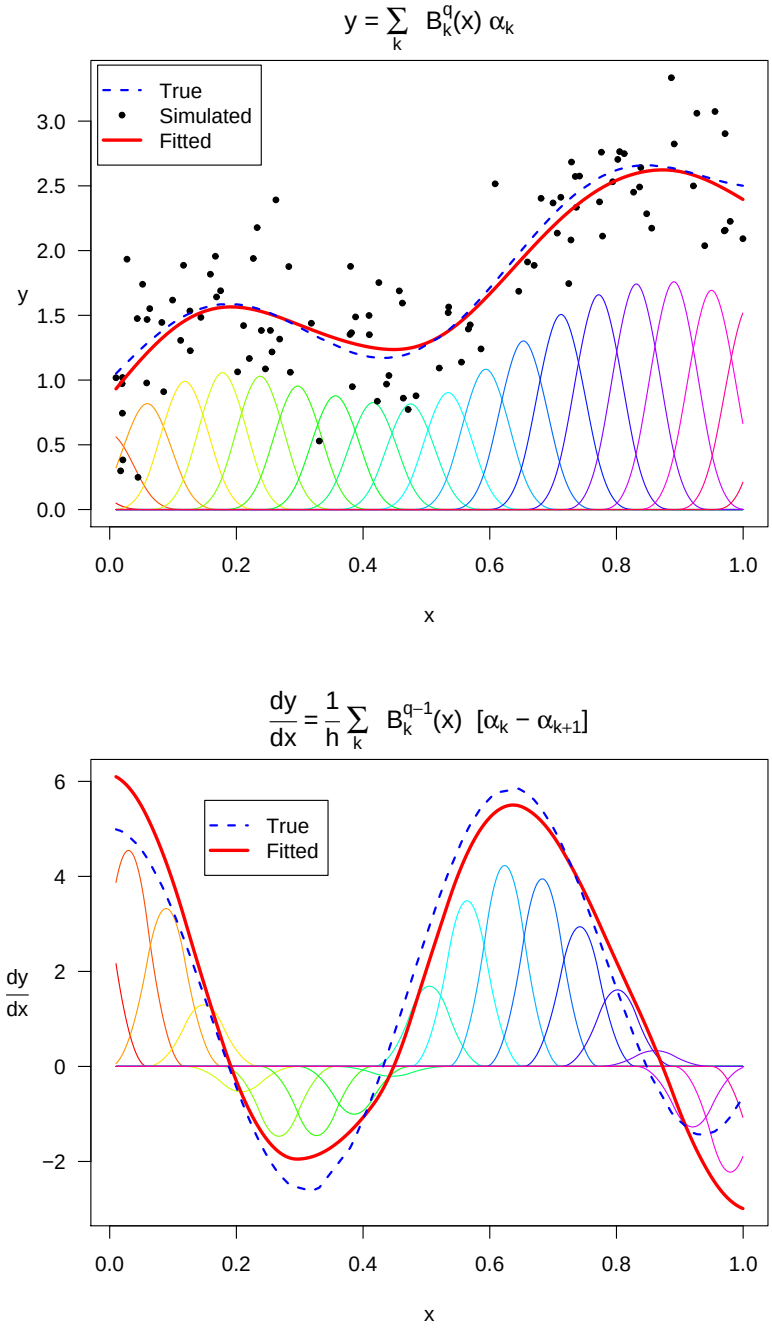

Figure 5.2: P-spline regression for simulated data. Fitted and true values for the function (upper panel) and its derivative (lower panel). B-spline bases with equally-spaced knots, $k = 20$, $q = 3$, $d = 2$ and λ selected by CV.

5.3.1 An algorithm for the WaFT model

The Poisson log-likelihood function, with a log-link, for the model in (5.3) is given by

$$l(\boldsymbol{\alpha}|\boldsymbol{y};\boldsymbol{\theta}) \propto \sum_{i=1}^{n} [y_i \, \ln(\mu_i) - \mu_i] = \sum_{i=1}^{n} [y_i \, \eta_i - \exp(\eta_i)] \, . \tag{5.7}$$

The linear predictor, η_i can be easily derived from (5.3):

$$\eta_i = \ln(\mu_i) = \ln(\gamma) + \ln\left[f(w(x_i;\boldsymbol{\alpha}),\boldsymbol{\theta})\right] + \ln\left[v(x_i;\boldsymbol{\alpha})\right] \tag{5.8}$$

The estimates of the solution to the first-order conditions are obtained by differentiating the log-likelihood (5.7) with respect to the elements of $\boldsymbol{\alpha}$:

$$\frac{\partial}{\partial \boldsymbol{\alpha}} l(\boldsymbol{\alpha}|\boldsymbol{y};\boldsymbol{\theta}) = \sum_{i}^{n} (y_i - \mu_i) \frac{\partial \eta_i}{\partial \alpha_k} = 0 \, , \tag{5.9}$$

where the partial derivative of the linear predictor (5.8) is given by

$$\frac{\partial \eta_i}{\partial \alpha_k} = \frac{\partial \ln(\mu_i)}{\partial \alpha_k} = \frac{\frac{\partial f(w(x_i;\boldsymbol{\alpha}),\boldsymbol{\theta})}{\partial w(x_i;\boldsymbol{\alpha})}}{f(w(x_i;\boldsymbol{\alpha}),\boldsymbol{\theta})} \cdot \frac{\partial w(x_i;\boldsymbol{\alpha})}{\partial \alpha_k} + \frac{\frac{\partial v(x_i;\boldsymbol{\alpha})}{\partial \alpha_k}}{v(x_i;\boldsymbol{\alpha})} \, . \tag{5.10}$$

In matrix notation, we can more succinctly write (5.10) in the following way:

$$\boldsymbol{Q} = \mathtt{diag}\left(\frac{\boldsymbol{f'}}{\boldsymbol{f}}\right) \cdot \boldsymbol{B}(\boldsymbol{x}) + \mathtt{diag}\left(\frac{1}{\boldsymbol{v}}\right) \cdot \boldsymbol{C}(\boldsymbol{x}) \, . \tag{5.11}$$

Here $\boldsymbol{B}(\boldsymbol{x}) = [B_k^q(x_i)]_{ik}$ and $\boldsymbol{C}(\boldsymbol{x}) = [C_k^q(x_i)]_{ik}$ as specified in the previous Section 5.2.1. Namely, $\boldsymbol{B}(\boldsymbol{x})$ is the B-splines basis of the warping function and $\boldsymbol{C}(\boldsymbol{x})$ is given by the differences between B-splines in the basis of the function $v(\cdot)$, as in equation (5.6). The vectors $\boldsymbol{f}$ and $\boldsymbol{f'}$ are, respectively, target function and its derivative with respect to $w(\cdot)$, evaluated on the transformed axis.

Given (5.9) and (5.11), the iteratively reweighted least squares (IRWLS) algorithm can be adapted to solve these equations[1]. Specifically, equation (5.11) is the basic structure for the working model matrix, which clearly depends on the coefficients $\boldsymbol{\alpha}$ and needs to be updated with each iteration. Moreover, the normalizing constant γ needs to be included in the algorithm. In matrix notation, we have the following modified IRWLS algorithm:

$$(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{X}})\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{W}}(\tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}) \, , \tag{5.12}$$

Here the final model matrix is $\tilde{\boldsymbol{X}} = [\boldsymbol{1} : \tilde{\boldsymbol{Q}}]$. The first column is the regressor for the constant γ, which does not depend on $\boldsymbol{\alpha}$. The coefficient vector includes both B-spline coefficients and normalizing vector, i.e. $\boldsymbol{\beta}' = [\ln(\gamma), \boldsymbol{\alpha}']$. The diagonal matrix $\tilde{\boldsymbol{W}} = \mathtt{diag}(\tilde{\boldsymbol{\mu}})$ is the common weight matrix in a Poisson GLM setting.

[1] See Section 4.3.1 for a detailed description of the IRWLS algorithm.

5.3.2 Smoothing the warping function

In Section 2.1, we introduced the P-spline approach for smoothing univariate functions (Eilers and Marx, 1996). This methodology combines (fixed-knot) B-splines with a roughness penalty and can be incorporated into the WaFT model.

Following the P-splines scheme, the number of B-spline basis K for the warping function is chosen purposely high. Specifically, in the following we use $K = 15$ B-splines of degree $q = 3$. One can easily see that this number of B-splines will lead to undersmoothed outcomes. To ensure smoothness, a roughness penalty on the coefficient vector $\boldsymbol{\alpha}$ is used as shown in Section 2.1.1.

In particular, the roughness of vector $\boldsymbol{\alpha}$ can be measured with differences of order d and the following penalty can be computed:

$$\boldsymbol{S}_d = \boldsymbol{\alpha}' \boldsymbol{D}_d' \boldsymbol{D}_d \boldsymbol{\alpha} = \|\boldsymbol{D}_d \boldsymbol{\alpha}\|^2 ,$$

cf. equation (4.6), p. 69.

This penalty, weighted by a positive regularization parameter λ, can be introduced into the likelihood for the WaFT model. The system of equations in 5.12 is then modified as follows:

$$(\tilde{\boldsymbol{X}}' \tilde{\boldsymbol{W}} \tilde{\boldsymbol{X}} + \lambda \boldsymbol{P}) \tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}' \tilde{\boldsymbol{W}} (\tilde{\boldsymbol{W}}^{-1} (\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}}) , \tag{5.13}$$

where

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 \\ 0 & \breve{\boldsymbol{P}} \end{pmatrix}$$

and $\breve{\boldsymbol{P}} = \boldsymbol{D}_d' \boldsymbol{D}_d$. The specific structure of the matrix $\boldsymbol{P}$ is due to the normalizing constant γ, which is not penalized in the WaFT model. As in the standard P-spline approach, the smoothness of the warping function can be controlled via the value of the parameter λ.

5.3.3 Optimal smoothing

The estimating equations for the penalized likelihood in (5.13) depend on the smoothing parameter λ. Once λ is fixed, the estimates $\hat{\boldsymbol{\alpha}}$ and, consequently, $w(x_i; \hat{\boldsymbol{\alpha}})$ are determined. In Section 2.1.4, we already presented different criteria for selecting λ in a typical P-spline setting. For the WaFT model, Camarda et al. (2008a) minimize the Bayesian Information Criterion (AIC) to choose the optimal amount of smoothing for the warping function:

$$\mathrm{BIC}(\lambda) = \mathrm{Dev}(\boldsymbol{y}; \boldsymbol{\alpha}, \lambda) + \ln(n) \cdot \mathrm{ED}(\boldsymbol{\alpha}, \lambda) . \tag{5.14}$$

$\mathrm{Dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda)$ is the deviance of the Poisson model (see equation (2.16)) and ED is the effective dimension of the model for a given smoothing parameter. As explained in Section 2.1.3, for the effective dimension Camarda et al. (2008a) follow the suggestion of Hastie and Tibshirani (1990) to take the trace of the hat-matrix from the estimated linearized smoothing problem in (5.13):

$$\mathrm{ED}(\boldsymbol{\alpha}, \lambda) = \mathrm{tr}(\boldsymbol{H}_\lambda) = \mathrm{tr}\left\{ \breve{\boldsymbol{X}} (\breve{\boldsymbol{X}}' \hat{\boldsymbol{W}} \breve{\boldsymbol{X}} + \lambda \boldsymbol{P})^{-1} (\breve{\boldsymbol{X}}' \hat{\boldsymbol{W}}) \right\} , \tag{5.15}$$

where $\hat{\boldsymbol{W}}$ is the matrix of weights at the last iteration of the IRWLS.

An efficient grid-search is sufficient: $\ln(\lambda)$ varies on a grid, and the value that minimizes BIC is picked (see Figure in Section 5.5). The presented penalized Poisson likelihood approach was implemented in `R` (R Development Core Team, 2008), and some computational details will be presented in the next section.

5.4 Software considerations

The estimation procedure for the WaFT model is a modified version of the IRWLS algorithm. We show how to implement the procedure in `R` (R Development Core Team, 2008) since its use is widespread, the following code should be easily understood by someone who is unfamiliar with this language, so long as the notation presented in Section 4.4 is known.

Furthermore, we denote with `B`, the B-spline basis $\boldsymbol{B}(\boldsymbol{x})$, `C` represents the matrix $\boldsymbol{C}(\boldsymbol{x})$ and `P` stands for the penalty $\boldsymbol{P}$, already multiplied by a certain value λ, which is externally supplied. Finally, we assume that our target distribution is known and can be evaluated for arbitrary points on the age-axis. Specifically, `dens(x,par)` is the target distribution function, where `x` and `par` are the age-axis on which $f(\cdot)$ is evaluated, and the estimated parameter(s) $\hat{\boldsymbol{\theta}}$, respectively. In the same way, the function `dens1(x,par)` represents the derivative of the target density with respect to the (transformed) age-axis.

5.4.1 Starting specifications

The WaFT model itself is rather computationally intensive, therefore proper starting values for the B-spline coefficients $\boldsymbol{\alpha}$ and for the normalizing constant γ help speed up the iterative procedure, and to ensure convergence. The idea is to first estimate a warping function that only shifts the distribution, so that the modes of the two densities coincide. The starting B-spline coefficients, $\boldsymbol{\alpha}^{(1)}$, of such warping function are given by the linear system of equations:

$$(\boldsymbol{B}(\boldsymbol{x})'\,\boldsymbol{B}(\boldsymbol{x}) + \boldsymbol{P})\,\boldsymbol{\alpha}^{(1)} = \boldsymbol{B}(\boldsymbol{x})\,\boldsymbol{w}^{(1)}\,,$$

where $\boldsymbol{w}^{(1)}$ is the simple shift warping function. Given that `w1` represents $\boldsymbol{w}^{(1)}$, in `R`, we can compute the following objects:

```
# Starting B-spline coefficients
alpha.st <- solve(t(B) %*% B + P, t(B) %*% w1)
# Starting warping function
w.st <- B %*% alpha.st
# Starting derivative of the warping function
v.st <- C %*% alpha.st
```

Since the starting warping function is a straight line with slope 1, i.e. $w^{(1)}(\boldsymbol{x}) = \boldsymbol{x} + c$, we expect that `v.st` (starting $\boldsymbol{v}$) is simply a column vector of ones of dimension $n \times 1$. Summing

over all observations on both side of model (5.3), we have:

$$\sum_i^n y_i = \gamma \cdot \sum_i^n \left[f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta}) \cdot v(x_i; \boldsymbol{\alpha}) \right] ,$$

and consequently, applying the log-link function:

$$\ln(\gamma) = \ln \left[\sum_i^n y_i \right] - \ln \left[\sum_i^n \left[f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta}) \cdot v(x_i; \boldsymbol{\alpha}) \right] \right] ,$$

which simplifies to

$$\ln(\gamma) = \ln \left[\sum_i^n y_i \right] - \ln \left[\sum_i^n f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta}) \right] .$$

for $v(x_i; \boldsymbol{\alpha}^{(1)})$, i.e. derivative of the warping function at the starting coefficients $\boldsymbol{\alpha}^{(1)}$ practically all equal to 1.

This result can then be used to find proper starting values for $\ln(\gamma)$ (`ln.gamma`), given $\hat{\boldsymbol{\theta}}$ (`theta.hat`):

```
# Target density over starting warping function
f.st <- dens(w.st, theta.hat)
# Starting normalizing constant
ln.gamma <- log(sum(y)) - log(sum(f.st))
# Starting coefficient vector
beta <- c(ln.gamma, alpha.st)
```

The last line combines both $\ln(\gamma)$ and $\boldsymbol{\alpha}$ in the coefficient vector $\boldsymbol{\beta}$ (`beta`).

5.4.2 Fitting the WaFT model

In this section, we present an R-function for updating the penalized system of equations in (5.13). In particular, it updates both coefficients $\boldsymbol{\beta}$ and expected values $\boldsymbol{\mu}$. The routine requires starting values for $\boldsymbol{\beta}$, the B-spline bases for the warping function and its derivative and the penalty matrix $\boldsymbol{P}$.

The function first selects the log of the normalizing constant $\ln(\gamma)$ (`ln.gamma`) and the B-spline coefficients $\boldsymbol{\alpha}$ (`alpha`) from the coefficient vector $\boldsymbol{\beta}$. The warping function $\boldsymbol{w}$ (`w`), and its derivative $\boldsymbol{v}$ (`v`) are then computed following equations (5.4) and (5.6).

The known functions `dens()` and `dens1()` are employed to evaluate target density (`f`) and its derivative (`f1`) over the updated transformed age-axis. Equation (5.8) is used to compute the linear predictor $\boldsymbol{\eta}$ (`eta`) and via the response function $\exp(\cdot)$, we obtain an estimate of the expected value $\boldsymbol{\mu}$ (`mu`). The diagonal weight matrix ($\boldsymbol{W}$, `W`) is then set up.

Equation (5.11) is used to compute the matrix $\boldsymbol{Q}$ (`Q`) and subsequently, for the IRWLS in (5.13), we updated the RHS (`RHS`) and the LHS, without and with additional penalty matrix (`LHS` and `LHSpP`). Finally, the R-function `solve()` is employed to update the coefficient vector $\boldsymbol{\beta}$ (`beta.new`).

```
# Updating the WaFT system of equations
UpdateWaFT <- function(y, beta, B, C, P){
    nb       <- ncol(B)
    ln.gamma <- log(beta[1])
    alpha    <- beta[1:nb+1]
    w        <- B %*% alpha
    v        <- C %*% alpha
    f        <- dens(w, theta.hat)
    f1       <- dens1(w, theta.hat)
    eta      <- ln.gamma + log(f) + log(v)
    mu       <- exp(eta)
    W        <- diag(mu)
    Q        <- diag(f1/f) %*% B  + diag(1/v) %*% C
    X        <- cbind(1, Q)
    LHS      <- t(X) %*% W %*% X
    LHSpP    <- G + P
    RHS      <- t(X)%*%(y - mu) + LHS%*%beta
    beta.new <- solve(LHSpP, RHS)
    return(list(beta=beta.new, mu=mu))
}
```

Given a certain λ, the previous function is re-iterated until convergence.

In order to speed up the grid-search over the smoothing parameter, we employed an efficient grid-search, already described in Section 4.4 for a different purpose. Here we start from the estimated starting value as described in Section 5.4.1, with a large value of λ, and then the actual penalized IRWLS algorithm will use the previously estimated coefficients $\boldsymbol{\alpha}$ as starting values for the new and now smaller λ.

When comparing age-at-death distribution for human mortality, an additional practice can be employed in order to speed up the process and guarantee the convergence. Specifically, it is applied when comparing two distributions from the same population, but distant in time and consequently significantly different in their shape. In such a case, we propose fitting the WaFT model for an age-at-death distribution closer in time and then using estimated coefficients $\boldsymbol{\alpha}$ as starting values for the more distant distribution, using a sort of *chain* principle in the time-dimension.

5.5 Simulation and applications

5.5.1 Simulation study

To demonstrate the performance of the model, we applied it to simulated data. In this section we will present the results of the model for two different simulated scenarios. Both of them mimic demographic data, but they differ from each other with respect to the estimating procedure for the target distribution.

The first simulation scenario uses a target distribution with known parameters that will be

then directly employed in fitting of the WaFT model. We will refer to it as *parametric* simulation setting. The second scenario will assume a known target distribution represented in a non-parametric setting. For abbreviation, we refer to it as *non-parametric* simulation setting.

Parametric simulation setting

In order to resemble demographic data, we assume a Gompertz target distribution (see Section 1.4). The probability density distribution is given by:

$$f(y_i) = ae^{bx_i} \exp\left[\frac{a}{b}(1 - e^{bx_i})\right]. \tag{5.16}$$

As explained in Section 1.2, we can derive the cumulative distribution function from (5.16). For a Gompertz distribution, this allows us to simulate life-times analytically. In our example, we assume the parameters $a = 0.005$ and $b = 0.12$ and we simulate $n = 10,000$ life-times. Figure 5.3 (left panel) presents an example from this simulated setting with the true target Gompertz distribution (in black), along with the histogram of the simulated life-times (in grey).

A known nonlinear warping function operates to transform the Gompertz life-times[2]. The WaFT model is then employed to fit the histogram of the warped life-times and to estimate the warping function. We assume a Gompertz distribution as target function and $\boldsymbol{\theta} = (0.005, 0.12)$, specifically the known true Gompertz parameters. In Section 5.5.2, we see how these parameters can be estimated from actual data. For the example in Figure 5.3, the value of λ selected by BIC is equal to 22.9, with the effective dimension equal to 7.43. The BIC profile is shown in the right panel of Figure 5.3.

Figure 5.3 shows also the histogram of the warped counts, along with the fitted values with an evident good fit. Figure 5.4 (left panel) shows the true warping function used in this setting, as well as the fitted $w(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$ for the given example. The grey dotted bisector in the right panel represents the identity transformation of the x-axis. It is easily seen that the WaFT model is able to reproduce the true warping function remarkably well for central x_i. Conversely, the model cannot properly capture the warping function at the edges of the distribution. This is mainly due to the presence of few counts at the boundary of the x-axis.

As mentioned in Section 5.2.1, the representation of the warping function allows us to easily compute $\frac{\partial w(\cdot)}{\partial x} = v(\cdot)$. Figure 5.4 (right panel) presents a both true and fitted derivative of the warping function for the given example. The derivative shows again the misfit of the WaFT model at the edges of the distribution. On the other hand, the model is flexible enough to capture the non-linearity of the warping function clearly portrayed by the derivative.

To assess the variability of the results, in this simulation study, this procedure was replicated 1,000 times, leading to 1,000 estimated distributions, warping functions and associated derivatives. Figure 5.5 (upper panel) presents the true Gompertz target function along with the true warped distribution and the 99% pointwise confidence interval from the 1,000 estimated distributions obtained in the simulation study.

In a similar fashion, the central and lower panels in Figure 5.5, the true warping function and

[2]Here the true warping function is given by $w(x_i; c, d, \omega) = x_i + c\,x_i^2 + d \cdot [\cos(x_i\,\omega) - 1]$ with $[c, d, \omega] = \left[0.003, 10, \frac{2\pi}{100}\right]$.

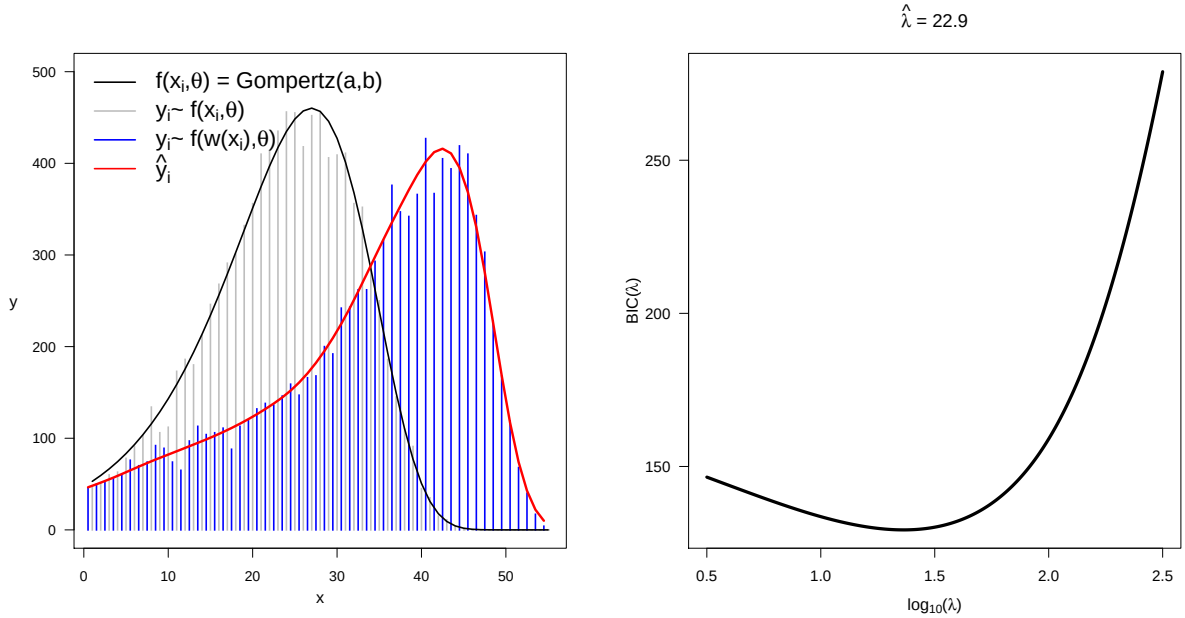

Figure 5.3: Left panel: an example of simulated data (grey) from a Gompertz distribution (black). Warped histogram and related fitted values from the WaFT model are depicted in blue and red, respectively. Right panel: BIC profile.

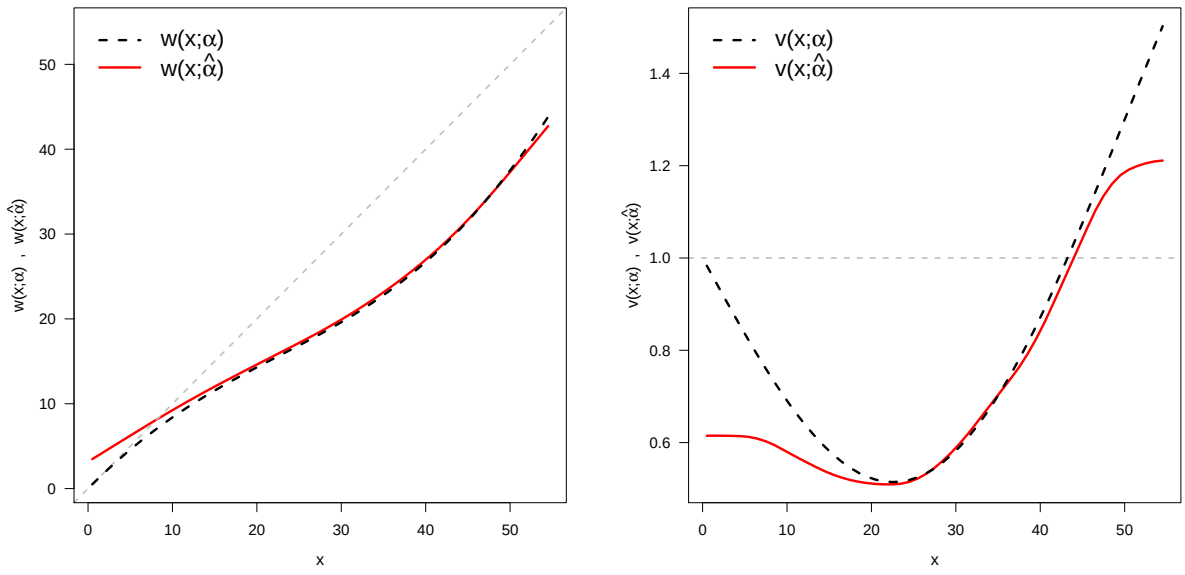

Figure 5.4: Outcomes from an example of the *parametric* simulation setting. Left panel: true and fitted warping function. The grey dotted bisector represents the identity transformation of the x-axis. Right panel: true and fitted derivative of warping function. The grey dotted line represents the derivative of any simple shifted transformation of the x-axis.

its derivative are accompanied by the 99% confidence interval from the simulation study.

While the true warping function is captured in the central part of the distribution, we see a moderate bias at the boundaries, where much less data are available. This feature translates into a corresponding bias of the density in Figure 5.5 (top). The estimates of the derivative $v(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$ are obtained by using the smoothing parameter determined from a criterion based on the warping function $w(\cdot)$. The always linear boundaries behavior of $w(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$ is translated into a basically flat derivative $v(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$. Simultaneous estimation of a smooth function and its derivatives by using a single smoothing parameter has been noted in the literature as being cumbersome (see discussion in Section 5.6). Here the problem is aggravated by the data sparseness in the tails of the distribution.

Non-parametric simulation setting

Often, parametric distributions cannot properly describe a particular target distribution, and they do not allow enough flexibility for comparing different density functions. In order to deal with this issue, the WaFT model also allows us to use a non-parametric target distribution.

Figure 5.6 (left panel) presents an example of a non-parametric target distribution (in black). It is represented by a linear combination of B-splines as explained in Section 2.1. We choose $K = 15$ B-splines of $q = 3$, and the associated coefficients are selected for leading to a fairly smooth true target distribution[3]. We can write the target density distribution as follows:

$$f(x_i) = \sum_{k=1}^{K=15} B_k^3(x_i)\, a_k \,.$$

Since we are now dealing with a non-parametric density, the cumulative hazard function does not allow a closed-form inversion formula. Therefore, life-times cannot be simulated analytically. A possible solution is the specification of piecewise-constant hazards within each interval $[x_{i-1}, x_i]$. Given this assumption, inverting the cumulative hazard function is relatively simple via numerical approximations. This approach has been used to simulate $n = 10,000$ life-times from the non-parametric target density and an example is shown in Figure 5.6 (left panel in grey).

A known warping function has been used to warp the simulated life-times[4]. In this way, we obtained the histogram of the warped life-times which are fitted using the WaFT model. Specifically, the parameters $\boldsymbol{\theta}$ of the target distribution are replaced by the known B-spline coefficients of the target distribution and the remaining components of the model are unchanged since we assumed fixed $\boldsymbol{\theta}$ during the estimation procedure (see Section 5.3).

Figure 5.6 (left panel) shows the warped data along with the fitted distribution for the given simulated example. The associated BIC profile is presented on the right panel in Figure 5.6 and the selected λ is equal to 13.8, with effective dimension equal to 7.63.

For the presented example of the *non-parametric* simulation setting, warping function and its derivative are presented in Figure 5.7. Also in this case, the WaFT cannot accurately capture

[3]The smooth target density $f(\boldsymbol{x}; \boldsymbol{\alpha})$ follows a mixture of a Gompertz and a negative exponential distribution.

[4]For this simulation study the true warping function is given by $w(x_i; c, d, \omega) = 2 + x_i + c\, x_i^2 + d \cdot [\cos(x_i\, \omega) - 1]$ with $[c, d, \omega] = \left[0.001, 5, \frac{2\pi}{100}\right]$.

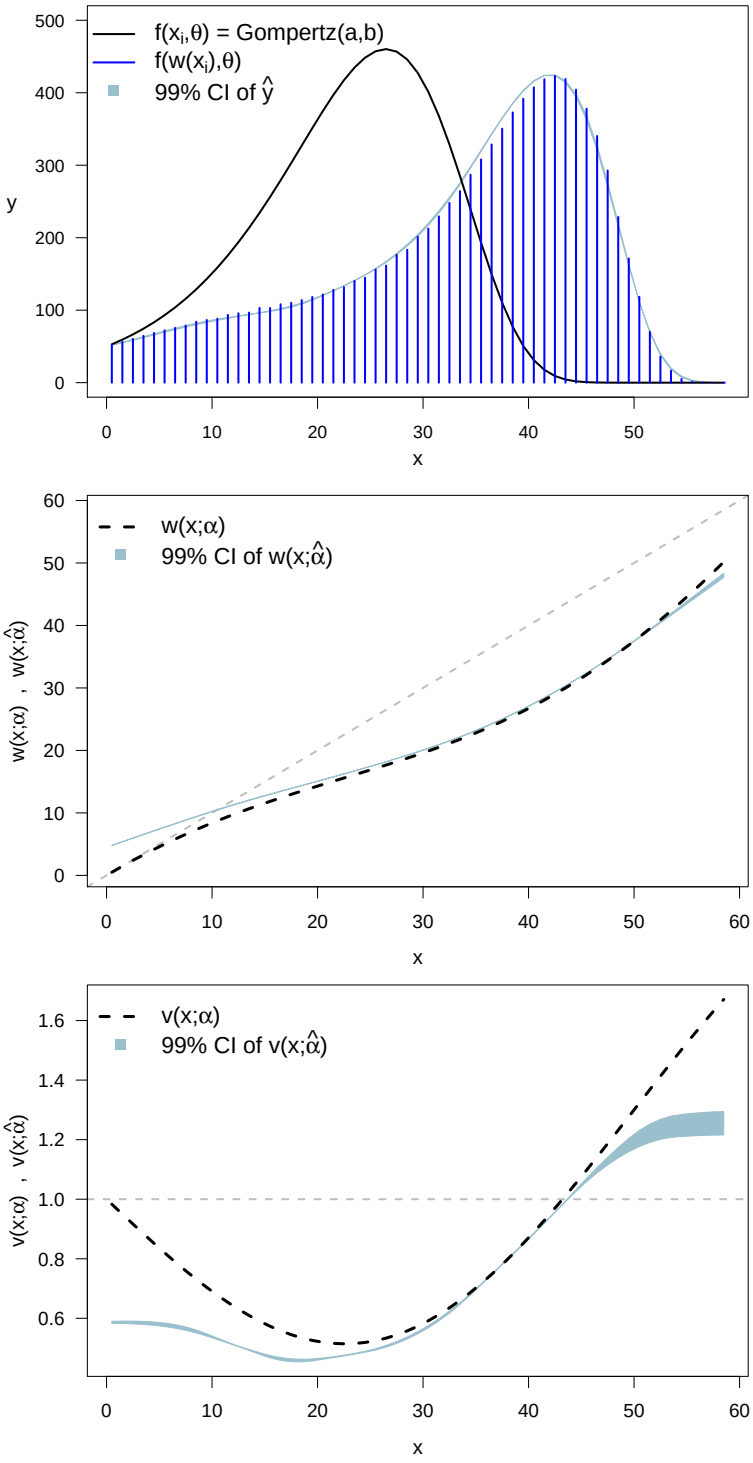

Figure 5.5: Outcomes from 1,000 replications of the *parametric* simulation setting. Upper panel: target Gompertz distribution (black) and true warped histogram (blue). The light-blue shadow depicts the 99% confidence interval for the fitted distributions. Central panel: true warping function and 99% confidence interval of the fitted warping functions. Lower panel: true derivative of the warping function and 99% confidence interval of the fitted derivatives of the warping functions.

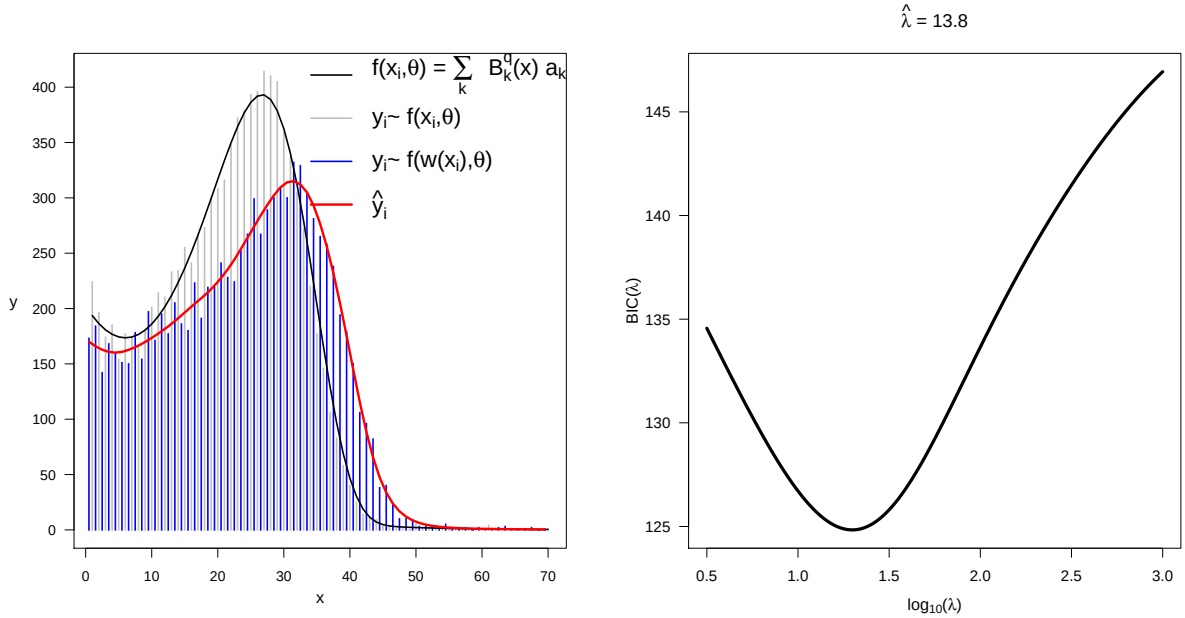

Figure 5.6: Left panel: an example of simulated data (grey) from a non-parametric distribution (black) represented as a linear combination of B-splines. Warped histogram and related fitted values from the WaFT model are depicted in blue and red, respectively. Right panel: BIC profile.

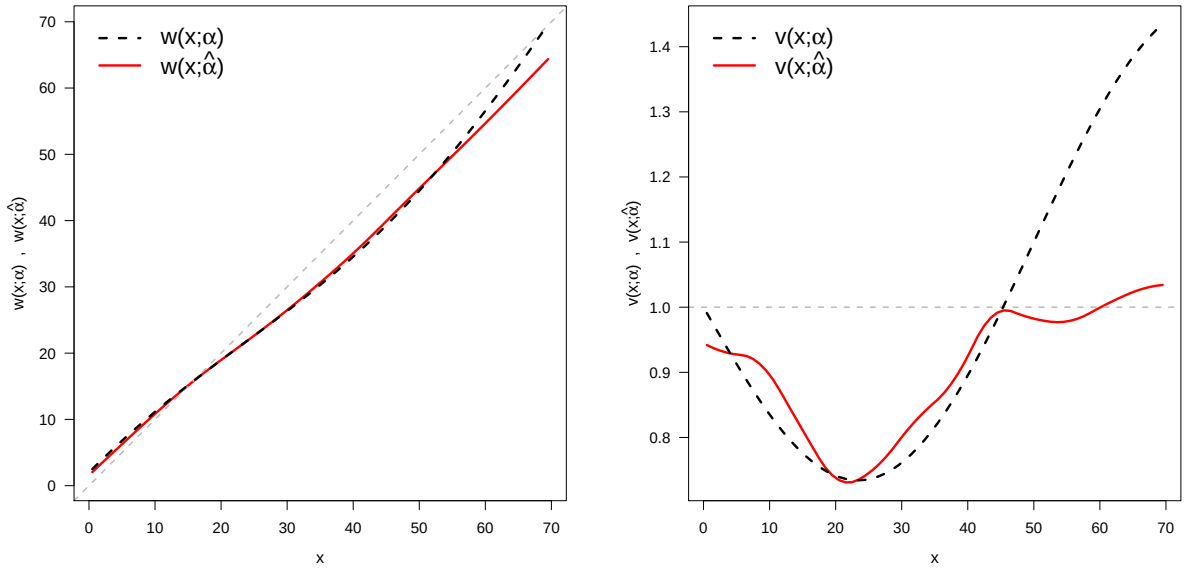

Figure 5.7: Outcomes from an example of the *non-parametric* simulation setting. Left panel: true and fitted warping function. The grey dotted bisector represents the identity transformation of the x-axis. Right panel: true and fitted derivative of the warping function. The grey dotted line represents the derivative of any simple shifted transformation of the x-axis.

the warping function and its derivative at the boundaries of the data. Therefore, outcomes about the edges of the distribution need to be interpreted with particular care.

As in the previous simulation setting, we replicated the procedure 1,000 times, leading again to 1,000 estimated distributions, warping functions and associated derivatives. Figure 5.8 (upper panel) presents the true non-parametric target function given by a linear combination of B-splines. The true warped distribution is plotted with the 99% pointwise confidence interval from the 1,000 fitted distributions.

The true warping function and the 99% confidence interval from the simulation study are in the central panel in Figure 5.8. It is worth pointing out that the WaFT model cannot capture only at the highest x_i the non-linearity of the warping function where we have few counts. Unlike the *parametric* setting, Figure 5.8 (central panel) shows a relatively good fit at the left tail of the distribution in which the density is not close to zero.

The lower panel in Figure 5.8 shows the derivative of the true warping function along with the 99% confidence interval from the simulation study. The problem that had been described in the *parametric* simulation setting is present in the case of a non-parametric target density as well. Again data sparseness in the tail is an issue here, which will have to be addresses in the extended version of the model.

5.5.2 Applications to the Danish data

Gompertz target distribution

For the Danish age-at-death distributions introduced in Section 5.1, we use the year 2006 as the target density. Since we are dealing with mortality over age 30, we choose a Gompertz distribution as a parametric model for the age-at-death distribution in 2006. A maximum likelihood estimator has been employed to estimate the parameters of the Gompertz. And a Newton-Raphson method was needed due to the non-linearity of the Gompertz model (Deuflhard, 2004). We obtained the parameters $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b})' = (1.14e^{-5}, 0.11)'$ for the Japanese females in 2006.

Given the fixed target Gompertz distribution for year 2006, the WaFT model has been used to warp the age-axis and fit the Danish age-at-death distribution in 1930. Figure 5.9 (left panel) shows the target distribution with its Gompertz estimates as well as the fitted values from the WaFT model. The BIC profile for this Danish example is presented in the right panel of Figure 5.9, and the smoothing parameter λ was selected equal to 47.9. Figure 5.10 shows both the resulting transformation function $w(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$ and its derivative. The identity transformation is indicated by a dashed line. The warping function is clearly nonlinear, that is, neither a simple shift nor a uniform scaling can map one density on to the other.

The impact of the nonlinearity in the fitted warping function is evident in fitting a WaFT model with a very large smoothing parameter. With this approach, since the order of the penalty in $w(\cdot)$ is set to $d = 2$, the "ultimate smooth" warping function is a simple linear fit (see Section 2.1.3). Figure 5.11 shows outcomes for a linear transformation of the age-axis for the Danish population over age 30. The left panel of Figure 5.11 clearly shows a misfit of the model for almost the full age-range. In this example, we considered $\lambda = 10^8$ which leads to effective dimension of about three, i.e. two parameters for the warping function and the normalizing parameter γ. Conversely,

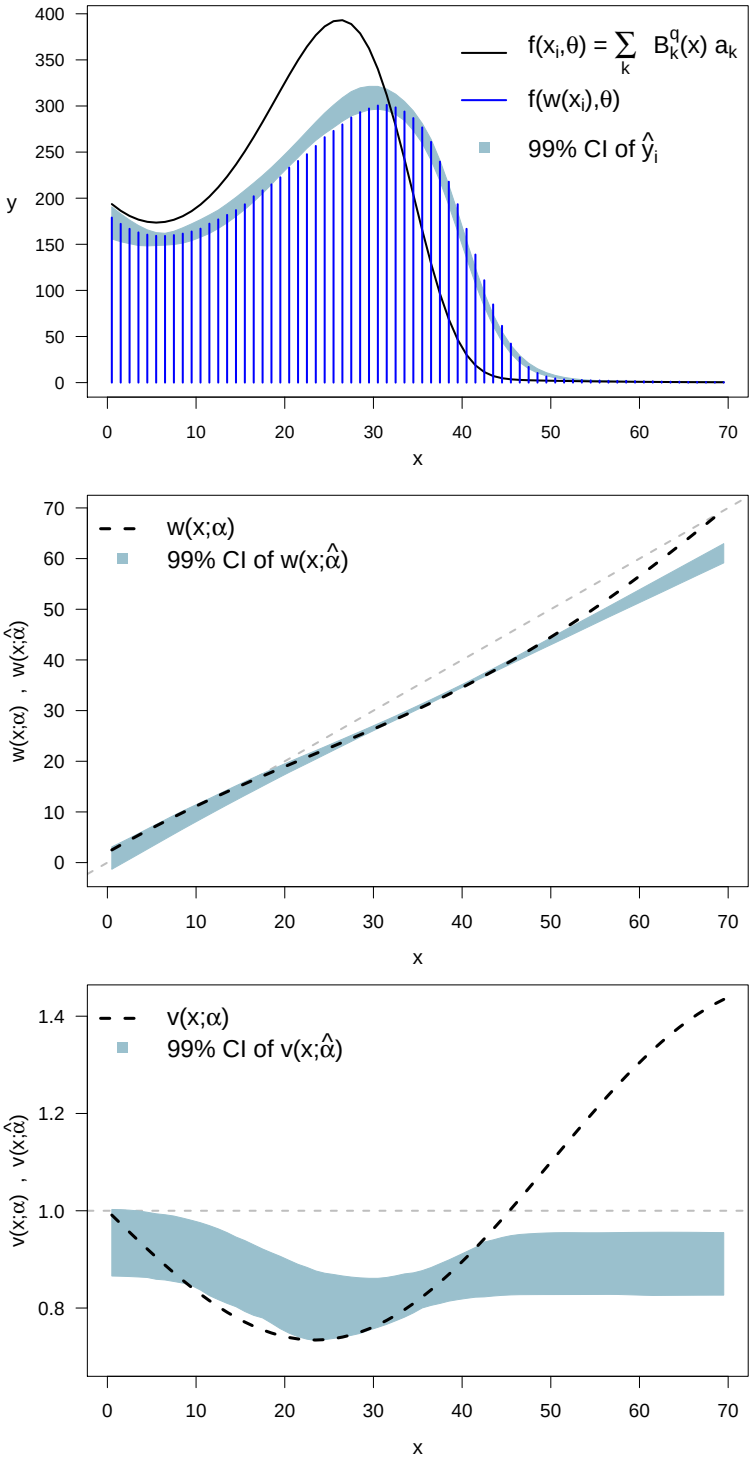

Figure 5.8: Outcomes from 1,000 replications of the *non-parametric* simulation setting. Upper panel: target non-parametric distribution (black) and true warped histogram (blue). The light-blue shadow depicts the 99% confidence interval for the fitted distributions. Central panel: true warping function and 99% confidence interval of the fitted warping functions. Lower panel: true derivative of the warping function and 99% confidence interval of the fitted derivatives of the warping functions.

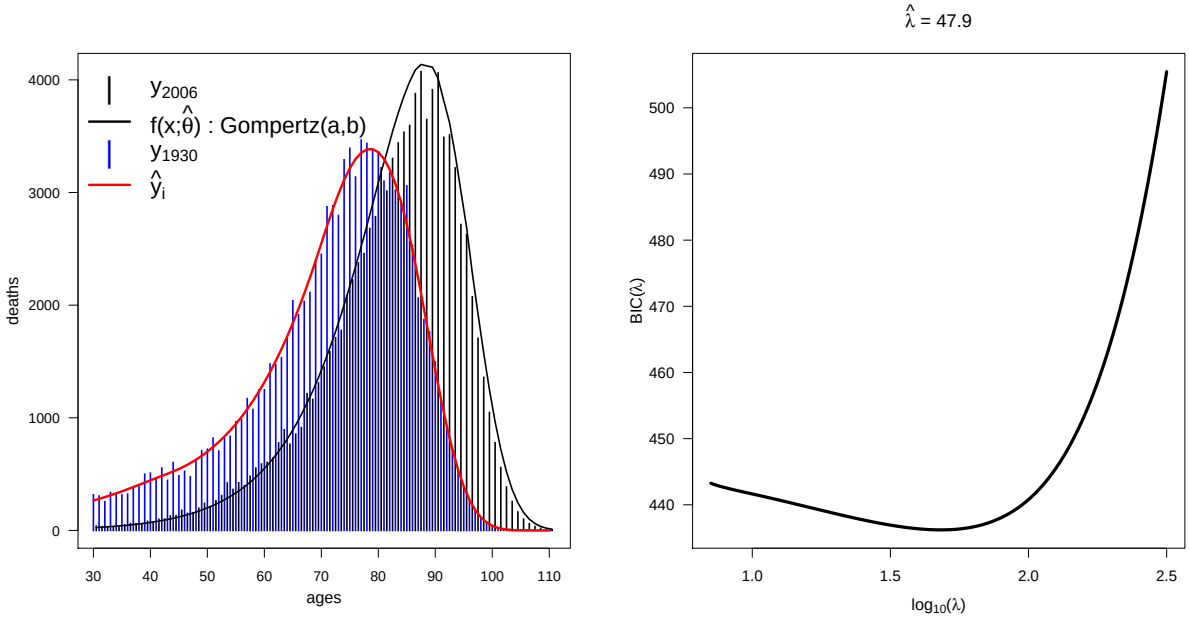

Figure 5.9: Left panel: Life-table age-at-death distributions for the Danish data over age 30. Data from 2006 are fitted with a Gompertz function and used as target distribution. Data from 1930 are estimated with the WaFT model. Right panel: BIC profile.

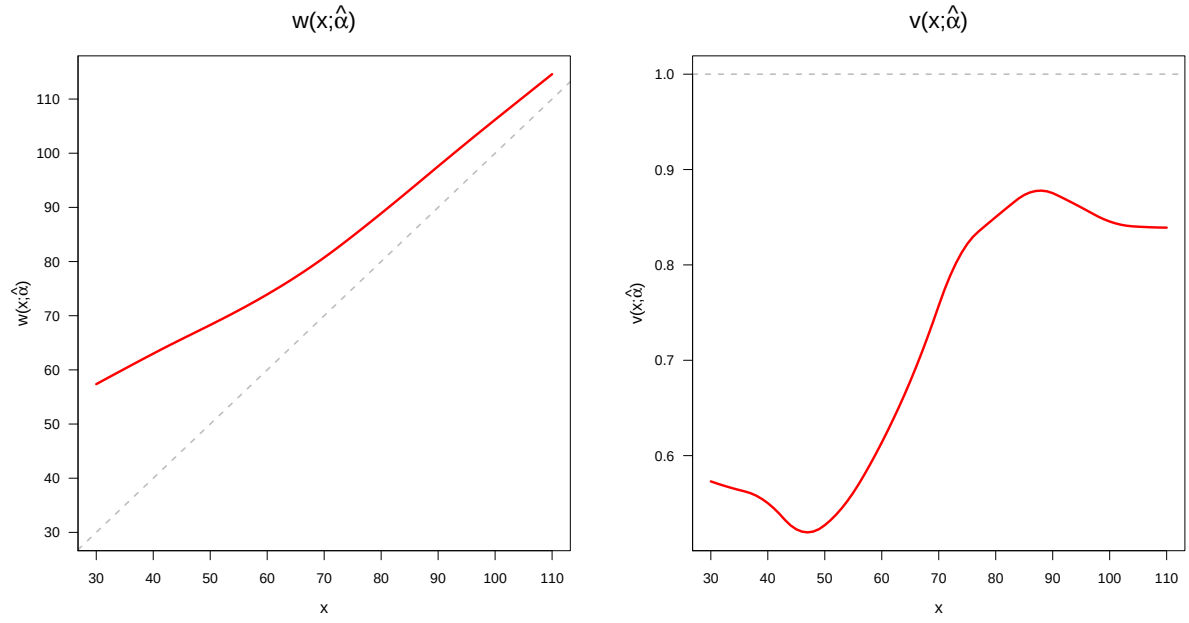

Figure 5.10: Outcomes from the Danish female population over age 30. Left panel: estimated warping function $w(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$. The identity transformation is indicated by a dashed grey line. Right panel: estimated derivative of the warping function $v(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$. The grey dotted lines represents any simple shift transformation of the x-axis.

the effective dimension of the WaFT model selected by BIC is equal to 8.19.

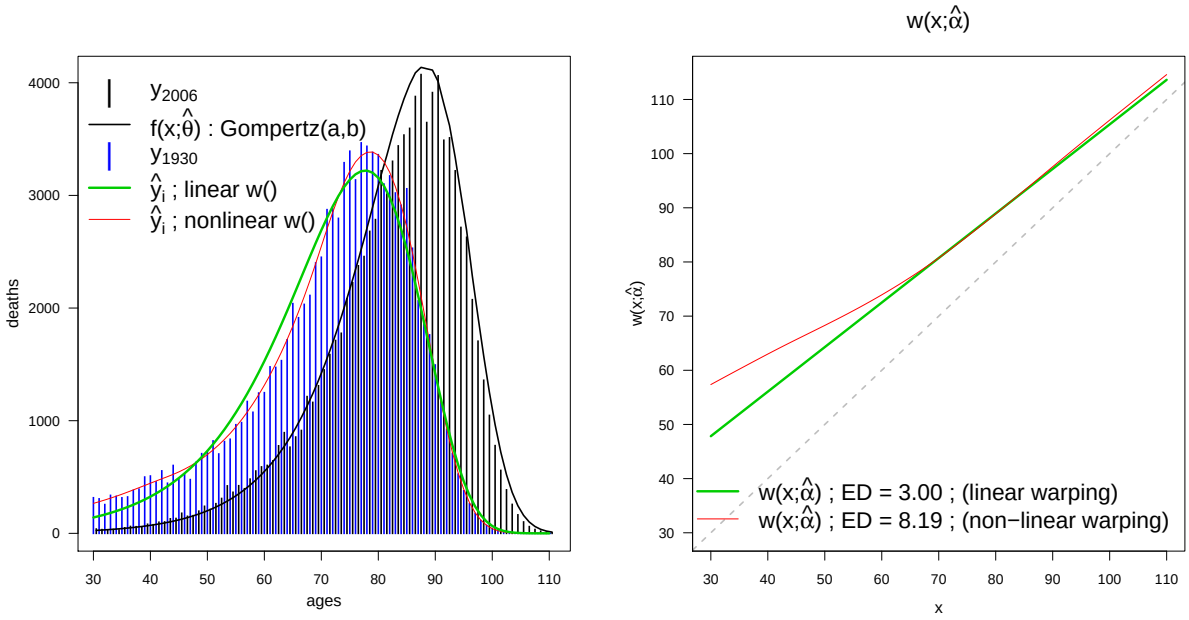

Figure 5.11: Comparison between linear and non-linear transformation of the age-axis. Left panel: Life-table age-at-death distributions for the Danish data over age 30. Data from 2006 are fitted with a Gompertz function and used as target distribution, data from 1930 are estimated with the WaFT model with λ equal to 10^8 (green) and 47.9 (red). Right panel: estimated death warping functions $w(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$. ED stands for the effective dimension of the full WaFT model.

Non-parametric target distribution

The Gompertz distribution plays a prominent role in the study of adult human mortality, but sometimes such parametric distribution cannot properly describe more complex patterns of adult mortality (see Section 1.4). Instead of searching alternative parametric distributions for portraying the target density, we can free the WaFT from any parametric assumption even regarding the estimation of the target distribution.

We estimate a target distribution using a P-spline approach for Poisson death counts, as described in Section 2.1.2. In this way, the target age-at-death distribution can be described as a linear combination of B-splines in which the associated coefficients are penalized following the methodology introduced by Eilers and Marx (1996). Once a target distribution is fitted, we follow the approach described in Section 5.5.1 for the *non-parametric* simulated example.

Figure 5.12 shows outcomes from a P-spline approach for the Japanese women above age 10 over which Gompertz distribution with only two parameters is likely inappropriate. For fitting the age-at-death distribution for the target density (year 2006), we used 25 equally-spaced B-splines with degree $q = 3$ and order of difference $d = 2$. Given these specifications, the value of the smoothing parameter selected by minimizing the BIC is equal to 100.

Figure 5.12 presents also the fitted values from the WaFT model. Since we do not assume

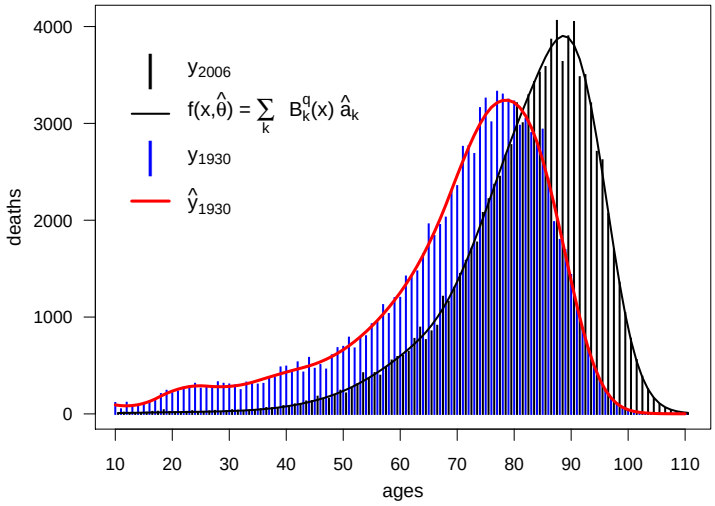

Figure 5.12: Life-table age-at-death distributions for the Danish data over age 10. Non-parametric P-splines estimate for the target distribution (year 2006). Data from 1930 are estimated with the WaFT model.

any parametric distribution, the WaFT model actually warps the age-axis such that the Danish age-at-death distribution in 1930 conforms the age-at-death distribution in 2006. Figure 5.13 shows both the fitted transformation function and its derivative. Also in this case, the derivative clearly shows that a simple linear warping of the age-axis would not be enough to account for the differences in the age-at-death distributions between these two years.

5.6 Further extensions

In this chapter, we present a new approach for dealing with the estimation of a nonlinear transformation to align densities (Camarda et al., 2008a). The proposed WaFT model is a rather general tool and brings together the ideas of warping and smoothing. Starting from a specific target distribution, the model allows estimation of the warping function of the age-axis that can map one density onto the other.

The only assumption that is made about the warping function is smoothness and, implicitly, monotonicity. By using a P-spline approach, not only can the warping function be estimated, but we may also directly express its derivative via B-splines. A penalized Poisson likelihood approach is then employed to estimate the model. The target function can be estimated either with parametric or non-parametric approaches which provides to great flexibility.

Simulation studies have shown that the WaFT model can properly capture nonlinear transformations of the x-axis. The derivatives of the warping function are accurate in the central part of the distribution where higher death counts are available. We also noticed that fitted derivatives of the warping function are inaccurate at the boundaries of the distribution. This problem can

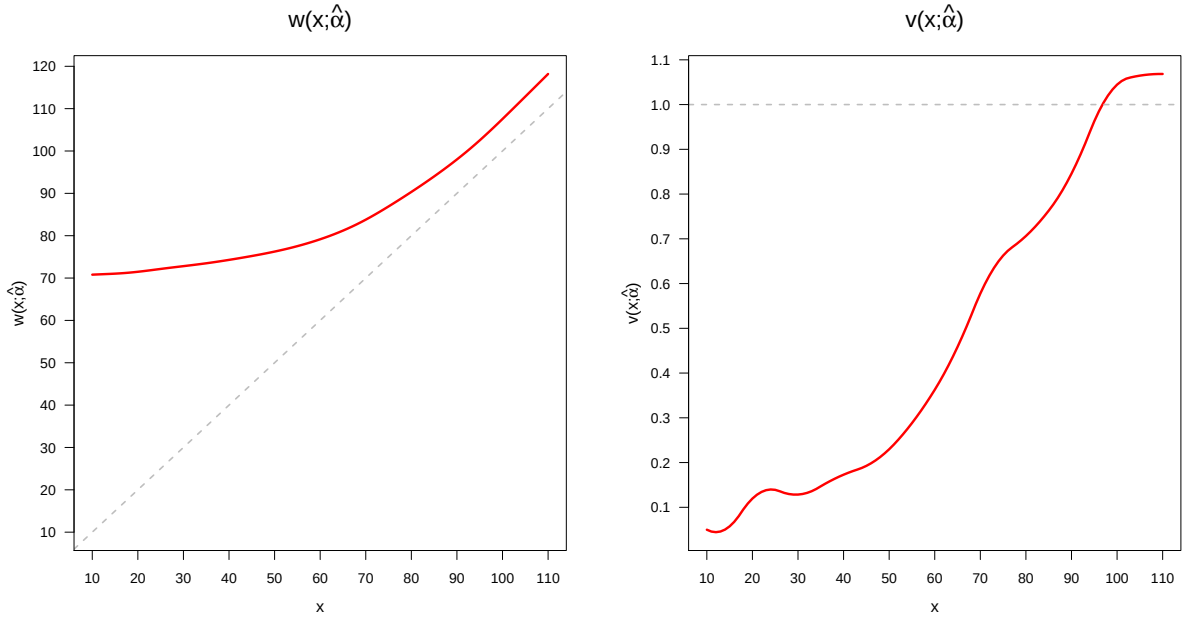

Figure 5.13: Outcomes from the Danish female population over age 10. Left panel: estimated warping function $w(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$. The identity transformation is indicated by a dashed grey line. Right panel: estimated derivative of the warping function $v(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$. The grey dotted lines represent any simple shift transformation of the x-axis.

be generally observed when derivatives are to be estimated. Several authors have pursued the idea of finding better smoothing estimators for the derivative itself. Among others, see Erickson et al. (1995); Härdle et al. (1992); Härdle and Stoker (1989); Song et al. (1995); Stoker (1993). Nevertheless, Ramsay (1998a) pointed out that "typically one sees derivatives go wild at the extremes, and the higher the derivative, the wilder the behavior" and that further problems arise when it comes to smoothing parameter selection. We expect that using splines explicitly designed for estimating derivatives of a function, as suggested by Heckman and Ramsay (2002) with the L-splines, can further enhance the WaFT model. Alternatively, a different weight system within the penalized IRWLS can help increase the leverage of the death counts close to the boundaries.

Outcomes from actual data revealed that simple linear transformations of the age-axis are inappropriate for portraying the mortality development over time and therefore a flexible model such as the WaFT is a suitable tool for comparing age-at-death distributions.

In this chapter, we present only applications from human mortality, hence stressing the generalization of the more simple accelerated failure time models. However, the WaFT model is appropriate for comparison of any two densities. We therefore envision alternative applications of the WaFT model in which nonlinear transformation of the x-axis is a suitable and reasonable idea. Such wider applications, however, are beyond the aims of this thesis.

Analyzing life-table age-at-death distributions can be misleading because life-table construction itself already involves estimation procedures and adjustments (see ch. 3 in Preston et al., 2001). We expect that using actual death counts and exposures in the estimation procedure will

improve the WaFT model and will allow for embedding it in a more straightforward setting, which is closer to the raw data.

In this chapter, we fitted actual mortality data either over age 30 or over age 10. The former restriction was due to the using of the Gompertz model for estimation of the target distribution. A P-spline methodology is employed to estimate the target distribution just over age 10, since, as mentioned in Section 1.3, infant mortality presents features which P-splines cannot cope with. If we would consider age-at-death distribution for all age, i.e. $x \geq 0$, then the warping function would warp the non-negative axis $\mathbb{R}_0^+$ onto itself. If we restrict our attention to a limited age-range, e.g. ages beyond 30, than the warped age-axis will have a different domain. As warping of the age-axis can be interpreted as the gain (or loss) in longevity, ages at death that formerly correspond to younger ages in case of mortality decline can now be found at much higher ages. Consequently, we should adjust the support for the two densities ($f(\cdot)$, which is the target, and $g(\cdot)$, which operates on the original age-axis) accordingly.

If we use data from life-table distribution, as we currently do, then the corresponding age-intervals can be easily obtained. If we intend to model only a particular age-range on the original axis, then the warped age-range should correspond to the same proportion of deaths as does the age-range for x (relative to the full age-axis $\mathbb{R}_0^+$).

Finally, in case of mortality data, a generalization of the WaFT can account for a two-dimensional setting. Warping functions between two subsequent years are expected to change smoothly. Therefore, one can cope with a sequence of warping functions over time by an additional penalty that controls the temporal pattern in the age-axis transformation. In Section 2.2, we presented a two-dimensional smoothing methodology which generalized the unidimensional P-spline approach. Such concepts can be used to generalize the WaFT model to two dimensions, as well.

Chapter 6

Conclusions

Mortality, i.e. the risk of death, changes with age and also steadily changes over time, at least in the developed world for more than the last 150 years. Understanding mortality dynamics over age and time is crucial for demography, as they are one of the driving forces of population change. An ongoing mortality decline and hence increasing longevity has considerable consequences both for the individual, as well as for society as a whole. Changing family structures, health care provision, saving for old age, and the financing of future pensions are but a few important fields.

Both the trajectory of death risks over age and over time, in general, change gradually, apart from certain "crisis" years of epidemics, wars, or political and social turmoil. It is this smooth behavior that is the focus of this thesis.

Analysis of mortality started with the first life-table about 350 years ago, which is not only is the birth date of modern demography, but can also be seen as a milestone in statistics. While starting from individuals' ages at death, contemporary national vital statistics provide aggregated data on life-spans, classified by sex, age at death, year of death, and, mostly also, year of birth. Together with counts of the corresponding population at risk, these data are the basis for the study of mortality.

The data used throughout this dissertation are taken from the Human Mortality Database (HMD). This database contains information from official sources in different countries and provides them for scientific analyses in a uniform format. Data contained in the HMD are known for their reliability and accuracy and offer a valuable resource for mortality research. In this thesis, we focus on methodology and not on the analysis of particular countries, so any actual data analyses, which are mostly performed on data taken from Denmark and Portugal, are provided as examples and could be replaced by different countries represented in the HMD.

Despite the regular, that is, smooth pattern of mortality, life (and death) still is, at least to a certain extent, a stochastic process. The statistical tools used to analyze life span distributions are commonly summarized under the notion of survival analysis, which predominantly deals with data on individuals. In the case of aggregated data, the concepts of survival analysis of course still are valid, however, the aggregation process leads to changes in the statistical models. The connection between individual and aggregated data is nicely represented by the so-called Lexis diagram, one of the key graphical displays in demography. The hazard of dying is the central concept in both the analysis of individual life spans as well as aggregated mortality data. Approximating the hazard

by a piecewise constant function, the observed numbers of deaths in an age-year square of the Lexis surface can be modeled as Poisson variates with means equal to the product of exposures multiplies by the age-year specific hazard. In this way, generalized linear models and all their extensions are available as analysis tools for mortality research over age and time. They are also the basis for the models proposed in this thesis.

Given the wealth of data, more traditional demographic methods for analyzing mortality surfaces, i.e. data on deaths and exposures cross-classified by age and year of occurrence, tend to apply a high number of parameters leading to all but parsimonious models. Nevertheless, to gain a better understanding of mortality dynamics, a mere descriptive representation of the empirical death rates is not informative. Furthermore, with increasing interest in the upper tail of the life span distribution, to gain insights in the prospects of mortality at high ages, data sparseness can become an issue even when working with aggregated mortality data. Hence, using models that explicitly address the smooth nature of mortality dynamics, both over age and time, is an obvious step.

Several models for capturing changes in mortality have been suggested in the literature, and we have provided a comprehensive overview. These models range from the classic parametric distributions of Gompertz and Makeham, which only target mortality changes over adult ages, to age-period-cohort models, which, in their overparameterized version, suffer from a fundamental identification problem. As an alternative, a bilinear model proposed by Lee and Carter about 25 years ago is used in many fields of current demography as a kind of standard for capturing age-time dynamics in mortality surfaces. Still the Lee-Carter model is a highly parameterized model. Given its importance in the demographic community, new models will have to be compared to the Lee-Carter model. Another approach to model changes in mortality are so-called relational models, which were among the first to attempt to model mortality in both its dimensions. They use a standard mortality function and a simple linear transformation for relating this standard to different distributions, i.e. across time or across countries.

The starting point of this thesis are two-dimensional smoothing methods for Poisson-distributed count data. P-splines are our method of choice. In one dimension, this approach combines a relatively large number of B-splines with a roughness penalty. On one hand, B-splines provide enough flexibility to capture trends in the data. On the other hand, an additional penalty on neighboring coefficients is used to ensure smoothness and reduces the number of parameters. This fact avoids the need for backfitting and knot selection schemes. P-splines can be seen as a generalization of a linear regression model, in which the B-splines act as regressors. The least squares algorithm (in the Normal case) or the iteratively reweighted least-squares algorithm (in the generalized linear model case) can then be employed and the only change is an additional penalty on the coefficients weighted by a positive regularization parameter. This allows for a continuous turning over different amount of smoothness. In order to measure the roughness of the coefficients, a matrix of dth order differences is constructed and included in the penalty. For a fixed smoothing parameter, the parameters can be easily estimated, and, as in the classic linear regression setting, we can specify the hat matrix for the fitted P-splines model.

For smoothing mortality data, both over age and time P-splines can be generalized to a two-dimensional setting. Kronecker and tensor products of simple unidimensional B-splines basis

are used for constructing a two-dimensional basis with local support. We showed how to apply a roughness penalty over the two dimensions of the mortality surface.

We illustrated several criteria for selecting the smoothing parameter, such as the Akaike's Information Criterion or the Bayesian Information Criterion. Since we model Poisson data, the deviance of the fitted model has been used as a measure of discrepancy. We showed that the hat-matrix contains information about the effective dimensions of the model and we use this feature in the selection of the smoothing parameter. Different amounts of smoothing can be employed over the two mortality dimensions. Since two-dimensional P-splines are still in a regression setting, the hat-matrix can be computed and consequently effective dimension and information criteria can be rearranged in a two-dimensional setting for selecting the two smoothing parameters.

Additional features of P-splines are an easy computation of standard errors and residual analysis, as they can be directly borrowed from classic regression methodology, as well. We reviewed the most common residuals for Poisson-distributed data, Pearson, Anscombe and deviance residuals and we describe their relations and features. Specifically, deviance residuals have been demonstrated to perform better in the case of Poisson-distributed data.

We also showed how shaded contour maps of the deviance residuals from fitted models over ages and years can be used to locate where the model cannot properly capture mortality trends and to understand additional demographic insights. Using this graphical technique, we demonstrated how the P-splines can capture mortality development more accurately than the Lee-Carter model does, despite the fact that the P-spline smoothing model uses remarkably fewer degrees of freedom.

As mentioned, demography enjoys a wealth of data. Large sample sizes are the rule rather that the exception and this fact has implications for statistical inference. Not only are the confidence limits of the fitted P-spline models extremely narrow, but other measures of goodness-of-fit also become basically uninformative and are not able to properly discriminate between models of different complexity. If we want to compare different models for mortality surfaces, we therefore need new measures. The particular emphasis here was to assess how much of the mortality dynamics is captured by a model. Statistically speaking, this is a question of explained variation.

The classic measure of explained variation is the R^2 from the Normal linear model. For Poisson-distributed data, such as death counts, adaptations have to be made. These are available for the more general case of exponential families. Proportional reduction in uncertainty, due to the inclusion of regressors is based on the Kullback-Leibler divergence. Practically, we presented R^2 measures which make use of the different forms of residuals presented in the previous chapter, namely Pearson and deviance residuals.

In the statistical and demographic models previously introduced, we usually employ a moderate to large number of parameters. In these cases, R^2 measures may be inflated and need to be adjusted. That means that the number of parameters needs to be incorporated into such measures, too. Hence, we presented several adjustments for goodness-of-fit measures for models that employed relatively large number of parameters.

Since P-splines can be incorporated into a generalized linear model framework, R^2 measures, based on deviance residuals can be applied for this smoothing approach as well. Nevertheless, P-splines, and smoothing methods in general, require further adjustments. Specifically, we showed the relation between the number of parameters and effective dimensions. Using this relationship,

we derived general R^2 measures for smoothing methods.

We also showed that R^2 measures, when applied to real data, and even when adjusted, are always close to 1, regardless of the applied mortality model and the specific actual data set. Such outcomes are essentially uninformative, and this is mainly due to two reasons. One is the large sample size. The other is that the classic goodness-of-fit measures compare fitted models to a null model, which is a simple overall mean of the data.

We proposed a new measure of explained variation, which we called $R^2_{\text{(bi)lin}}$, that overcomes these issues and can be used for comparing models for mortality data. The basic idea is to consider a different null model, which is particularly appropriate for mortality data. This model is linear or bilinear for unidimensional or two-dimensional data, respectively. We showed how the bilinear model is nested within a P-spline model but also a Lee-Carter model, and therefore it is natural to use it as the null model if these models are to be compared.

Specifically, we presented a new representation for P-splines as mixed models. This allowed us to separate the fixed and the random part of the model. Using the Singular Value decomposition and a specific penalty for the coefficients of the B-splines, we showed how the fixed part of the mixed models representation is exactly the linear or bilinear model which is then used as null model in the proposed $R^2_{\text{(bi)lin}}$. On the other hand, we showed that a bilinear model over age and time is a special case of the Lee-Carter model, in which its vectors of parameters vary linearly over ages and years. As a consequence of these findings, the variation explained by such models is now compared to the bilinear model.

The relations between $R^2_{\text{(bi)lin}}$ and information criteria such as Akaike's and Bayesian Information Criteria have also been derived. Specifically, we showed that the proposed new measure of explained variation can be relatively close to the Akaike's Information Criterion due to the presence of the Deviance of the null model in the denominator of $R^2_{\text{(bi)lin}}$. Nevertheless, if we would use $R^2_{\text{(bi)lin}}$ as a criterion for smoothing parameter selection, we would obtain an optimal value that is different from the one obtained by minimizing the other criteria.

In order to evaluate the performance of the proposed measure, we carried out different simulation studies, in both one and two dimensions. The simulation settings closely resembled real mortality data. $R^2_{\text{(bi)lin}}$ was able to capture differences between mortality models. Both in simulated and actual examples, the two-dimensional P-spline approach gave a better fit than the Lee-Carter model did.

For instance, $R^2_{\text{(bi)lin}}$ for the Danish female population from 1930 to 2006 and from age 30 to 100 turned out to be 0.828 and 0.705 for the two-dimensional P-spline regression and Lee-Carter model, respectively. Therefore, if we compare this to the limit model that is contained in both the two-dimensional P-spline regression and Lee-Carter model, i.e. the bilinear model, then we see that the P-spline approach captures more of the additional variability than the Lee-Carter approach does.

We have emphasized the high quality of the data contained in the Human Mortality Database, however, if we go back in time, then data quality may become an issue. The particular problem we may have to deal with in historical mortality data, or in countries with relatively poor data, is

age heaping or, in a more general context, digit preference. This is the tendency to round numbers to pleasant digits. In particular, age-at-death distributions can present systematic peaks at ages ending in 0 and, less prominently, 5. Such misreported data can generate misleading outcomes and adjustments are often required before any further statistical analysis. Also in this context, smoothing approaches prove to be valuable.

We suggested a general model that combines the concept of penalized likelihood with the composite link model. The composite link model allows us to describe how a latent true age-at-death distribution is mixed by the digit preference mechanism, by partly redistributing certain death counts to the preferred ages, so that the actual age-at-death distribution is observed. The mixing can be embodied in a matrix whose elements basically contain the probabilities of redistribution. The only assumption that is made about the underlying true latent distribution is smoothness.

For estimating this model, a generalization of the iteratively reweighted least squares algorithm can be used, which also includes the composition matrix for the misreporting pattern. The model matrix in the algorithm has to be adapted and needs to be updated in each iteration, since it depends on the misreporting pattern. Smoothness of the latent distribution is enforced with a difference penalty, analogous to the penalty employed in the P-spline approach. Also here, the penalty is weighted by the smoothing parameter, which needs to be optimized.

The misreporting pattern can be quite general, i.e. we allow a partial redistribution of the observations from any digit to its adjacent neighbors. In this way, the tendency to misreport need not be the same for identical end-digits, but may vary over the age range, which is often seen in real demographic data. We also proposed a more general approach, in which exchanges between digits that are more than one category apart can be made.

Consequently, this rather flexible preference pattern leads to a huge number of misreporting probabilities that need to be estimated. To estimate this second model component, we use a weighted least-squares regression within the iteratively reweighted least-squares procedure. Again a penalty, in this case a L_1-ridge penalty, restrains the problem and makes estimation feasible. This constrained weighted least-squares regression depends on an additional smoothing parameter. We showed how the Akaike's Information Criterion can properly select both smoothing parameters via a two-dimensional grid-search. Specifically we showed how the effective dimension of the model can be denoted by the sum of the two model components, i.e. the penalized composite link model and the penalized weighted least-squares regression.

Simulation studies and applications on actual data demonstrated that this new approach gives remarkably accurate results. It directly addresses the process that leads to heaping of certain ages, and the model goes beyond the mere quantification of the digit preference as provided by many commonly used indices. Extracting the latent distribution will be most important in many applications, however, the pattern of misclassification may also be of interest in itself.

Changes in mortality over time, which in recent times mostly have been reductions of mortality, can also be viewed as gains in life spans. Deaths that would have occurred at younger ages in the past now happen at older ages. Such a way of describing mortality improvements address the age-at-death distribution (the density) rather than the hazard. So we may ask the question:

Which transformation of the age-axis would have to be applied to transform one age-at-death distribution into another. Again, smoothness of the transformation is the only assumption we are willing to make.

In the simplest case, the transformation is linear, leading to a simple accelerated failure time model. A uniform rescaling of the age-axis, however, for the most part is too rigid to capture human mortality dynamics. Therefore, we considered nonlinear transformations and, consequently, we introduced a model which brings together the ideas of warping the time-axis and smoothing methodologies. We called it Warped Failure Time (WaFT) model.

In the WaFT model, we first choose a target distribution which is assumed fixed during the estimation procedure. For an observed life-table age-at-death distribution, the model estimates the warping function, so that after transforming the age-axis, the density of the observed distribution matches the specified target. We showed that a B-splines representation of the warping function naturally allows for controlling smoothness and incorporating derivative of the warping function, which is needed in this transformation approach.

We presented an algorithm to estimate the coefficients via a penalized Poisson likelihood. In particular, we illustrated a generalization of the iteratively reweighted least-squares algorithm. For the WaFT model, the model matrix depends on the B-spline coefficients and needs to be updated with each iteration. A normalizing constant is also incorporated into the model and into the estimation procedure in order to adjust the correct sample size.

Following again a P-spline approach, the number of B-splines in the warping function is purposely chosen high and the spline coefficients are restrained by a roughness penalty. Also in this case, a smoothing parameter controls the trade-off between smoothness of the warping function and model fidelity. In choosing the optimal smoothing parameter, we minimized the Bayesian Information Criterion.

The WaFT model is computationally rather intensive, therefore proper starting values for the B-spline coefficients are needed. We showed how to first estimate a warping function that only shifts the distribution, so that the modes of the two densities coincide. Routines for fitting the adjusted iteratively reweighted least-squares algorithm are also given.

The target density can be given by a parametric distribution, such as the Gompertz distribution, but additional flexibility can be gained if the target function is estimated non-parametrically as well. A unidimensional P-splines density estimator can be used in this case.

Simulation studies showed that the WaFT model can properly capture nonlinear transformations of the age-axis. Moreover, analyses of actual data revealed that simple linear transformations of the age-axis are inappropriate for portraying the mortality development over time and therefore, a flexible model such as the WaFT is a suitable tool for comparing age-at-death distributions.

This thesis demonstrated the usefulness of smoothing methods, in particular P-spline approaches, for the analysis of several aspects of mortality development. A new measure of explained variation for comparing different models for bivariate mortality surfaces has been proposed. Two new models have been suggested. One addresses the smoothness of death counts across the age-range, which can be distorted by age-misreporting. The other one offers an alternative way of exploring

mortality changes by looking at gains (or losses) in lifespans rather than by addressing decreases (or increases) in hazards. As a common theme, no rigid parametric restrictions are made and estimates are derived based on smoothness assumptions only. While motivated by a mortality context, both the model for digit preference and the warped failure time model offer opportunities outside the field of demography.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Cźaki (Eds.), *Second International Symposium on Information Theory*, Budapest, Hungary, pp. 267–281. Akademiai Kiadó.

Alho, J. M. and B. D. Spencer (2005). *Statistical Demography and Forecasting*. Springer Series in Statistics. Springer.

Anscombe, F. J. (1953). Contribution to the discussion of H. Hotelling's paper. *Journal of the Royal Statistical Society 15*, 229–230.

Anscombe, F. J. (1961). Examination of Residuals. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

Anson, J. (1988). The Parameters of Death: a Consideration of the Quantity of Information in a Life Table using a Polynomial Representation of the Survivorship Curve. *Statistics in Medicine 7*, 895–912.

Arthur, W. B. and J. W. Vaupel (1984). Some general relationship in population dynamics. *Population Index 50*, 214–226.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families*. Chichester: Wiley.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practise*. MIT Press.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

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

Box, G. E. P. and D. R. Cox (1964). An Analysis of Transformation (with discussion). *Journal of the Royal Statistical Society. Serie B 26*, 211–252.

Box, G. E. P. and G. Jenkins (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Brass, W. (1971). On the Scale of Mortality. In W. Brass (Ed.), *Biological Aspect of Demography*. London: Taylor & Francis.

Breiman, L. and J. H. Friedman (1985). Estimating Optimal Trasformations for Multiple Regression and Correlation. *Journal of the American statistical Association 80*, 580–597.

Breslow, N. (1984). Extra-Poisson Variation in Log-Linear Models. *Applied Statistics 33*, 38–44.

Brouhns, N., M. Denuit, and J. K. Vermunt (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics & Economics 31*, 373–393.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear Smoother and Additive Models (with discussion). *The Annals of Statistics 17*, 453–555.

Buse, A. (1973). Goodness of Fit in Generalized Least Squares Estimation. *The American Statisticians 27*, 106–108.

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2008a). A Warped Failure Time Model for Human Mortality. In P. H. C. Eilers (Ed.), *Proceedings of the 23rd International Workshop of Statistical Modelling*.

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2008b). Modelling General Patterns of Digit Preference. *Statistical Modelling*. To appear.

Cameron, A. C. and P. K. Trivedi (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics 1*, 29–53.

Cameron, A. C. and F. A. G. Windmeijer (1996). R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization. *Journal of Buniness & Economic Statistics 14*, 209–220.

Cameron, A. C. and F. A. G. Windmeijer (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics 77*, 329–342.

Carstensen, B. (2007). Age-Period-Cohort models for the Lexis diagram. *Statistics in Medicine 26*, 3018–3045.

Carstensen, B. and N. Keiding (2005). Age-Period-Cohort models: Statistical Inference in the Lexis diagram. Unpublished manuscript available at `www.biostat.ku.dk/~bxc`.

Caswell, H. (2001). *Matrix Population Models. Construction, Analysis, and Interpretation* (2nd ed.). Sinauer Associates.

Chatfield, C. (2003). Model selection, data mining and model uncertainty. In *Proceedings of the 18th International Workshop of Statistical Modelling*.

Clayton, D. and E. Schifflers (1987). Models for temporal variation in cancer rates. II. Age-period-cohort models. *Statistics in Medicine 6*, 469–481.

Cleveland, W. S. and S. J. Devlin (1988). Locally Weighted Regression: an Approach to Regression Analysis by Local Fitting. *Journal of the Statistical American Association 83*, 597–610.

Coale, A. J. and S. Li (1991). The Effect of Age Misreporting in China on the Calculation of Mortality Rates at Very High Ages. *Demography 28*(2), 293–301.

Coull, B. A., D. Ruppert, and M. P. Wand (2001). Simple incorporation of interactions into additive models. *Biometrics 57*, 539–545.

Cox, D. R. and E. J. Snell (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society 30*(2), 248–275.

Cox, D. R. and E. J. Snell (1971). On Test Statistics Calculated from Residuals. *Biometrika 58*, 589–594.

Crawford, S. L., C. B. Johannes, and R. K. Stellato (2002). Assessment of Digit Preference in Self-Reported Year at Menopause: Choice of an Appropriate Reference Distribution. *American Journal of Epidemiology 156*(7), 676–683.

Currie, I. D. and M. Durban (2002). Flexible Smoothing with P-splines: a unified approach. *Statistical Modelling 2*, 333–349.

Currie, I. D., M. Durban, and P. H. C. Eilers (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling 4*, 279–298.

Currie, I. D., M. Durban, and P. H. C. Eilers (2006). Generalized linear array models with applications to multidimentional smoothing. *Journal of the Royal Statistical Society 68*, 259–280.

Currie, I. D., J. G. Kirkby, M. Durban, and P. H. C. Eilers (2007). Smoothing Age-Period-Cohort models with P-splines: a mixed model approach. Unpublished draft.

Czado, C., A. Delwarde, and M. Denuit (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics & Economics 36*, 260–284.

Das Gupta, P. (1975). A General Method of Correction for Age Misreporting in Census Populations. *Demography 12*(2), 303–312.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). New York: Springer.

de Jong, P. and L. Tickle (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies 13*, 1–18.

Delwarde, A., M. Denuit, and P. H. C. Eilers (2007). Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Statistical Modelling 7*, 29–48.

Deuflhard, P. (2004). *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*. Springer.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford: Clarendon Press.

Dijksterhuis, G. and P. H. C. Eilers (1997). Modelling Time-Intensity Curves using Prototype Curves. *Food Quality and Preference 8*, 131–140.

Durban, M., I. D. Currie, and P. H. C. Eilers (2006). Mixed models, array methods and multidimentional density estimation. In J. Hinde, J. Einbeck, and J. Newell (Eds.), *Proceedings of the 21st International Workshop of Statistical Modelling*.

Edouard, L. and A. Senthilselvan (1997). Observer error and birthweight: digit preference in recording. *Public Health 111*, 77–79.

Eilers, P. H. C. (2004a). Parametric Time Warping. *Analytical Chemistry 76*, 404–411.

Eilers, P. H. C. (2004b). The Shifted Warped Normal Model for Mortality. In A. Biggeri, E. Dreassi, C. Lagazio, and M. Marchi (Eds.), *Proceedings of the 19th International Workshop of Statistical Modelling*.

Eilers, P. H. C. (2007). Ill-posed Problems with Counts, the Composite Link Model and Penalized Likelihood. *Statistical Modelling 7*(3), 239–254.

Eilers, P. H. C. and M. W. Borgdorff (2004). Modeling and correction of digit preference in tuberculin surveys. *International Journal of Tuberculosis and Lung Diseases 8*, 232–239.

Eilers, P. H. C., I. D. Currie, and M. Durban (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis 50*, 61–76.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with B-splines and Penalties (with discussion). *Statistical Science 11*(2), 89–102.

Eilers, P. H. C. and B. D. Marx (2002a). Generalized Linear Additive Smooth Structures. *Journal of Computational and Graphical Statistics 11*(4), 758–783.

Eilers, P. H. C. and B. D. Marx (2002b). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems 66*, 159–174.

Erickson, R. V., V. Fabian, and J. Marik (1995). An Optimum Design for Estimating the First Derivative. *The Annals of Statistics 23*, 1234–1247.

Ewbank, D. C., J. Gomez De Leon, and M. A. Stoto (1983). A reducible Four-Parameter System of Model Life Tables. *Population Studies 37*(1), 105–127.

Forfar, D., J. McCutcheon, and A. Wilkie (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries 115*, 1–149.

Frank, I. E. and J. H. Friedman (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics 35*(2), 109–148.

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics 31*, 3–39.

Frome, E. L. (1983). The analysis of rates using Poisson regression models. *Biometrics 39*, 665–674.

Frome, E. L., M. H. Kutner, and J. J. Beauchamp (1973). Regression analysis of Poisson-distributed data. *Journal of American Statistical Association 68*, 935–940.

Gasser, T. and H. G. Müller (1979). Kernel estimation of regression functions. In *Lecture Notes in Mathematics*, Volume 757, pp. 23–68. New York: Springer.

Gavrilov, L. A. and N. S. Gavrilova (1991). *The Biology of Life Span: A Quantitative Approach*. New York: Harwood Academic Publisher.

Gompertz, B. (1825). *On the nature of the function expressive of the law of human mortality*. 115: 513–585. London, UK: Philosophical Transactions Royal Society.

Good, I. J. (1969). Some Applications of the Singular Decomposition of a Matrix. *Technometrics 11*(4), 823–831.

Green, P. J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika 72*, 527–537.

Gu, C. and G. Wahba (1993). Semiparametric Analysis of Variance with Tensor Product Thin Plate Splines. *Journal of Royal Statistical Society. Serie B. 55*, 353–368.

Haberman, S. and A. Renshaw (2008). Mortality, longevity and experiments with the LeeCarter model. *Lifetime Data Analysis 14*, 286–315.

Haberman, S. H. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Härdle, W., J. Hart, J. S. Marron, and A. B. Tsybakov (1992). Bandwidth Choice for Average Derivative Estimation. *Journal of the American Statistical Association 87*, 218–226.

Härdle, W. and T. M. Stoker (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association 84*, 986–995.

Hastie, T. (1987). A Closer Look at the Deviance. *The American Statisticians 41*, 16–20.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall.

Heckman, N. E. and J. O. Ramsay (2002). Penalized Regression with Model-Based Penalties. *The Canadian Journal of Statistics 28*, 241–258.

Heitjan, D. F. and D. B. Rubin (1990). Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping. *Journal of the American Statistical Association 85*(410), 304–314.

Heligman, L. and J. H. Pollard (1980). The Age Pattern of Mortality. *Journal of the Institute of Actuaries 107*(1), 49–80.

Heuer, C. (1997). Modeling of Time Trends and Interactions in Vital Rates using Restricted Regression Splines. *Biometrics 53*, 161–177.

Himes, C. L., S. H. Preston, and G. A. Condran (1994). A Relational Model of Mortality at Older Ages in Low Mortality Countries. *Population Studies 48*(2), 269–291.

Hoerl, A. and R. Kennard (1988). Ridge Regression. In *Encyclopedia of Statistical Sciences*, Volume 8, pp. 129–136. New York: Wiley.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics 39*, 311–324.

Human Mortality Database (2008). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at `www.mortality.org` (Data downloaded on April 2008).

Hyndman, R. J. and M. S. Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis 51*, 4942–4956.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). Wiley.

Keiding, N. (1990). Statistical Inference in the Lexis Diagram. *Philosophical Transactions: Physical Sciences and Engineering 332*, 487–509.

Keyfitz, N. (1966). Sampling Variance of Standardized Mortality Rates. *Human Biology 38*, 309–317.

Keyfitz, N. and H. Caswell (2005). *Applied Mathematical Demography*. Springer.

Kirkby, J. and I. D. Currie (2007). Smooth models of mortality with period shocks. In J. del Castillo, A. Espinal, and P. Puig (Eds.), *Proceedings of the 22nd International Workshop of Statistical Modelling*, pp. 374–379.

Klein, J. P. and M. L. Moeschberger (2003). *Survival Analysis. Techniques for Censored and Truncated Data* (2nd ed.). New York: Springer.

Kneip, A. and T. Gasser (1988). Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics 16*, 1266–1305.

Kneip, A. and T. Gasser (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics 20*, 1266–1305.

Kooperberg, C. and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis 12*, 327–347.

Kooperberg, C. and C. J. Stone (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics 1*, 301–328.

Kostaki, A. and V. Panousis (2001). Expanding an abridged life table. *Demographic Research 5*, 1–22.

Krivobokova, T., C. M. Crainiceanu, and G. Kauermann (2006). Computationally efficient spatially-adaptive penalized splines. In J. Hinde, J. Einbeck, and J. Newell (Eds.), *Proceedings of the 21st International Workshop of Statistical Modelling*, pp. 303–308.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Lang, S. and A. Brezger (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics 13*, 183–212.

Ledauphin, S., E. Vigneau, and E. M. Qannari (2006). A procedure for the analysis of time intensity curves. *Food Quality and Preference 17*, 290–295.

Lee, R. D. and L. R. Carter (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association 87*(419), 659–671.

Lee, R. D. and T. Miller (2001). Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality. *Demography 38*(4), 537–549.

Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Strassburg: Trübner.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society 61*, 381–400.

Makeham, W. M. (1860). On the law of mortality. *Journal of the Institute of Actuaries 13*, 283–287.

Mari Bhat, P. N. (1990). Estimating Transition Probabilities of Age Misstatement. *Demography 27*(1), 149–163.

Marron, J. S. and D. Ruppert (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of Royal Statistical Society. Serie B 56*, 653–671.

Marx, B. D. and P. H. C. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis 28*, 193–209.

Marx, B. D. and P. H. C. Eilers (1999). Generalized linear regression on sampled signals and curves: a *P*-spline approach. *Technometrics 41*, 193–209.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Model* (2nd ed.). Monographs on Statistics Applied Probability. Chapman & Hall.

Mittlböck, M. and T. Waldhör (2000). Adjustments for R^2-measures for Poisson regression models. *Computational Statistics & Data Analysis 34*, 461–472.

Myers, R. J. (1940). Errors and Bias in the Reporting of Ages in Census Data. *Transactions of the Actuarial Society of America 41*(Pt. 2 (104)), 395–415.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society 135*, 370–384.

Ogata, Y., K. Katsura, N. Keiding, C. Holst, and A. Green (2000). Empirical Bayes Age-Period-Cohort Analysis of Retrospective Incidence Data. *Scandinavian Journal of Statistics 27*, 415–432.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing 9*(2), 363–379.

Parise, H., L. Ryan, D. Ruppert, and M. Wand (2001). Incorporation of historical controls using semiparametric mixed models. *Applied Statistics 50*, 31–42.

Pedroza, C. (2006). A Bayesian forecasting model: predicting US male mortality. *Biostatistics 7*, 530–550.

Perks, W. (1932). On some experiments in the graduation of mortality statistics. *Journal of the Institute of Actuaries 63*, 12–40.

Pickering, R. (1992). Digit preference in estimated gestational age. *Statistics in Medicine 11*, 1225–1238.

Pierce, D. A. and D. W. Schafer (1986). Residuals in Generalized Linear Models. *Journal of the American Statistical Association 81*(396), 977–986.

Preston, S. H. (1976). *Mortality Patterns in National Populations. With special reference to recorded causes of death.* Academic Press.

Preston, S. H., P. Heuveline, and M. Guillot (2001). *Demography. Measuring and Modeling Population Processes.* Blackwell.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramsay, J. O. (1998a). Derivative estimation. StatLib. S-News Thu, 12 March 1998. Available at `www.math.yorku.ca/Who/Faculty/Monette/S-news/0556.html`.

Ramsay, J. O. (1998b). Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B 60*, 365–375.

Ramsay, J. O. and X. Li (1998). Curve Registration. *Journal of the Royal Statistical Society, Series B 60*, 351–363.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.). Springer.

Renshaw, A. and S. Haberman (2003a). Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections. *Applied Statistics 52*(1), 119–137.

Renshaw, A. and S. Haberman (2006). A cohort-based Extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics 38*, 556–570.

Renshaw, A. E. and S. Haberman (2003b). Lee-Carter Mortality Forecasting with Age-specific Enhancement. *Insurance: Mathematics and Economics 33*, 255–272.

Ridout, M. S. and B. J. T. Morgan (1991). Modelling Digit Preference in Fecundability Studies. *Biometrics 47*, 1423–1433.

Ruppert, D. and D. B. H. Cline (1994). Bias Reduction in Kernel Density Estimation by Smoothed Empirical Transformations. *The Annals of Statistics 22*, 185–210.

Ruppert, D. and M. P. Wand (1992). Correcting for Kurtosis in Density Estimation. *Australian Journal of Statistics 34*, 19–29.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Sachia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician 41*, 169–178.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa (1986). *Akaike Information Criterion Statistics*. D. Reidel.

Sakoe, H. and S. Chiba (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*, 43–49.

Schlossmacher, E. (1973). An Iterative Technique for Absolute Deviations Curve Fitting. *Journal of the American Statistical Association 68*, 857–865.

Schmid, V. J. and L. Held (2007). Bayesian Age-Period-Cohort Modeling and Prediction - BAMP. *Journal of Statistical Software 21*, 1–15.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Siegel, J. S. and D. A. Swanson (2004). *Methods and Materials of Demography*. Elsevier Academic Press.

Siler, W. (1983). Parameters of Mortality in Human Populations with Widely Varying Life Spans. *Statistics in Medicine 2*, 373–380.

Silverman, B. W. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, Series B 57*, 673–689.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer.

Smith, H. L. (2008). Advances in Age-Period-Cohort Analysis. *Sociological Methods & Research 36*, 287–296.

Song, K. S., H. G. Muller, A. J. Clifford, H. C. Furr, and J. A. Olson (1995). Estimating Derivatives of Pharmacokinetic Response Curves with Varying Bandwidths. *Biometrics 51*, 12–20.

Stoker, T. M. (1993). Smoothing Bias in Density Derivative Estimation. *Journal of the American Statistical Association 88*, 855–863.

Taubenberger, J. K. and D. M. Morens (2006). 1918 Influenza: the Mother of All Pandemics. *Emerging Infectious Diseases 12*(1), 15–22.

Thatcher, A. R. (1999). The long-term pattern of adult mortality and the highest attained age (with discussion). *Journal of Royal Statistical Society 127*, 5–43.

Thatcher, R., V. Kannisto, and J. W. Vaupel (1998). *The Force of Mortality at Ages 80 to 120*, Volume 5 of *Monographs on Population Aging*. Odense, DK: Odense University Press.

Thompson, R. and R. J. Baker (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics 30*(2), 125–131.

Tibshirani, R. (1988). Estimating Trasformations for Regression Via Additivity and Variance Stabilization. *Journal of the American Statistical Association 83*, 394–405.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society 58*(1), 267–288.

Tukey, J. W. (1957). The comparative anatomy of transformation. *Annals of Mathematical Statistics 28*, 602–632.

Tuljapurkar, S., N. Li, and C. Boe (2000). A universal pattern of mortality decline in the G7 countries. *Nature 405*, 789–792.

Vaupel, J. W. (1997). Trajectories of Mortality at Advanced Ages. In *Beween Zeus and the salmon: The biodemography of longevity.*, pp. 17–37. Washington, DC: National Academy Press.

Vaupel, J. W., B. A. Gambill, and A. I. Yashin (1986). Thousands of data at a glance: shaded contour maps of demographic surfaces. Research paper, International Institute for Applied Systems Analysis (IIASA).

Veall, M. R. and K. F. Zimmermann (1996). Pseudo-R^2 Measures for Some Common Limited Dependent Variable Models. *Journal of Economic Surveys 10*, 241–59.

Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham (1999). The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines (with discussion). *Applied Statistics 48*, 269–311.

Vos, P. W. (1991). A geometric approach to detecting influential cases. *The Annals of Statistics 19*, 1570–1581.

Waldhör, T., G. Haidinger, and E. Schober (1998). Comparison of R^2 measures for Poisson regression by simulation. *Journal of Epidemiology and Biostatistics 3*(2), 209–215.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics 18*, 233–249.

Wand, M. P., J. S. Marron, and D. Ruppert (1991). Transformations in Density Estimation. *Journal of the American Association 86*, 343–353.

Wang, D. and P. Lu (2005). Modelling and Forecasting Mortality Distributions in England and Wales using the LeeCarter Model. *Journal of Applied Statistics 32*, 873–885.

Wang, K. and T. Gasser (1997). Alignment of Curves by Dynamic Time Warping. *The Annals of Statistics 25*, 1251–1276.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics 18*, 293–297.

Weisberg, S. (1985). *Applied Linear Regression* (2nd ed.). Wiley series in probability and mathematical statistics. New York: John Wiley & Sons.

Wilmoth, J. R. (1993). Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change. Technical report, Department of Demography, University of California, Berkeley.

Windmeijer, F. A. G. (1995). Goodness-of-Fit Measures in Binary Choice Models. *Econometric Reviews 14*, 101–116.

Wood, S. (2003). Thin Plate Regression Splines. *Journal of Royal Statistical Society. Serie B. 65*, 95–114.

Wood, S. N. (2006). *Generalized Additive Models. An Introduction with R.* Chapman & Hall.

Zaba, B. (1979). The Four-Parameter Logit Life Table System. *Population Studies 33*(1), 79–100.