Smooth Constrained Mortality Forecasting

Carlo G. Camarda

Institut National d'Études Démographiques

carlo-giovanni.camarda@ined.fr

October 18, 2018

Abstract

Mortality can be forecast by means of parametric models, principal component methods and smoothing approaches. All of these methods either impose rigid modeling structures or produce implausible outcomes. In this paper, we propose a novel approach that combines a well-established smoothing model and demographic prior information. Specifically, we constrain future smooth mortality patterns to lie within a range of valid age profiles and time trends, both computed from observed patterns. We enforce these shape constraints through an asymmetric penalty approach on forecast mortality. Moreover, we properly integrate infant mortality in a smoothing framework so that the mortality forecast covers the whole age range. We illustrate the proposed approach to mortality data for Danish females and US males. The proposed model outperforms the plain smoothing approach as well as commonly used methodologies while retaining all the desirable properties that demographers expect from a forecasting method, e.g. smooth and plausible age profiles and time trends.

Keywords: Mortality forecast, Smoothing, Demographic Constraints, Age-Time patterns, Asymmetric penalty

1 Introduction

Mortality modeling and forecasting is crucial in epidemiology and population studies, as well as in the insurance and pensions industries. In recent decades, several methodologies have been proposed and many demographers, actuaries and statisticians have suggested approaches for projecting mortality. See Booth and Tickle (2008) and Cairns et al. (2008) and the references therein.

In addition to scenario- and expert-based approaches, in this growing body of methods, we can broadly distinguish three classes based on their assumptions. A traditional procedure relies on parametric models to describe mortality age patterns and then extrapolating the estimated parameters to reconstruct future mortality (see, among others, Tabeau et al. (2002)). When dealing with adult mortality, parametric models are extremely parsimonious, but a large number of parameters are often necessary when the whole age range is considered. Though these models present the advantage of clear-cut interpretation of their parameters, this feature is not particularly useful in a forecasting approach. Moreover, when dealing with the whole age range, it is generally hard to disentangle the meaning attached to each parameter and simultaneously forecast plausible mortality patterns based on their time-series.

Over-parametric models such as the Lee and Carter (1992) and its variants have been widely used and have become a benchmark for many newly proposed methodologies. They describe mortality development over age and time using their principal components. They reduce a two-dimensional problem to a fixed age effect with a univariate time index. The time index condenses the mortality changes in past years and is thus used to forecast future mortality. This class of models presents several drawbacks, however. A simple univariate time series results in mortality improvements at all ages being perfectly correlated. Due to a fixed age effect over time, lack of smoothness in the estimated mortality pattern is evident, especially in the forecast years. A large number of parameters are implicitly assumed in this class of model. Given the regular structure of human mortality development, this overparametrization may seem unnecessary. Various solutions for these issues have been proposed in successive papers, including a generalization with multiple times indices. (Currie, 2011, 2013; D'Amato et al., 2011; Delwarde et al., 2007; Hyndman and Ullah, 2007; Li et al., 2013; Renshaw and Haberman, 2003).

A combination of Lee-Carter and parametric approaches could be found within recently developed Bayesian methods for probabilistic population projections (Raftery et al., 2013; Ševčíková et al., 2016). First, they performed a Bayesian hierarchical models for forecasting life expectancy at birth considering available data for all countries in the world. A second step consists in the conversion of overall levels by means of a variant of the Lee-Carter model (Li et al., 2013) and an adjusted parametric model (Thatcher et al., 1998). Additionally, coherence in forecasting more populations simultaneously is accounted. These methods have been used to produce recent World Population Prospects released by the United Nations (Gerland et al., 2014).

However, both parametric and Lee-Carter approaches are based on rigid modeling structures which are often unable to capture certain features of mortality change.

An alternative compromise method was proposed by Currie et al. (2004). They employed a two-dimensional penalized B-splines approach (P-splines) to smooth mortality over age and time without any specific model structure, allowing a parsimonious description of mortality development. They treated forecasting of future values as a missing value problem and estimate the fitted and forecast values simultaneously. Moreover, routines for estimating and forecasting mortality based on this approach are freely available (Camarda, 2012).

P-splines have been extensively used for smoothing observed mortality developments in demographic, ecological and epidemiological studies. See, for instance, Colchero et al. (2016); Goicoa et al. (2012); Jones et al. (2014); Lindahl-Jacobsen et al. (2016); Minton et al. (2017); Ouellette and Bourbeau (2011); Trias-Llimós et al. (2016); Ugarte et al. (2010). Few studies produced mortality forecasts using this methodology, however (Bohk-Ewald and Rau, 2017; Carfora et al., 2017; Ribeiro, 2015; Ugarte et al., 2012). More extensive literature can be found in actuarial science where smoothness is relevant to future mortality trends (Barrieu et al., 2012; Blake et al., 2006; Cairns et al., 2006; Currie, 2016; Djeundje and Currie, 2011; Huang and Browne, 2017; Lu et al., 2014; Pitacco et al., 2009; Richards et al., 2014, 2006; Wang et al., 2016).

The reason for this mixed recognition of two-dimensional P-splines lies in their lack of robustness for forecasting mortality (Cairns et al., 2009). Even though P-splines outperforming all competitors in modeling mortality, this approach suffers from all the issues that encumber a purely data-driven approach when employed for forecasting purposes. Forecast mortality simply follows estimated trends with a blind adherence to extrapolation, and mortality structure over age is not fully considered in the forecast values. Moreover, the penalty structure, which ensures smoothness in the fitted values, critically affects future mortality. Unreasonable trends from a demographic perspective could then emerge: increasing mortality over time for specific ages and hence crossover of mortality trends for adjacent ages in future years (cf. Section 2.1).

This paper aims to enhance two-dimensional P-splines through incorporating demographic knowledge into the model allowing for a better performance in forecasting mortality trends. We retain all features of two-dimensional penalized B-splines and, additionally, we ensure that future mortality over the age range follows a known and well-behaved profile, estimated from past years. This prior knowledge is incorporated by means of asymmetric penalties into the P-spline system (Bollaerts et al., 2006; Eilers, 2005). Since we constrain as well as penalize splines, we call the proposed approach a CP-spline model.

Finally, in the original paper by Currie et al. (2004), no solution was proposed for modeling and forecasting mortality for the whole age range. In the following we will also propose a solution for smoothing and forecasting mortality from infancy to oldest-old ages.

The combination of powerful statistical methodology and prior demographic information makes the CP-spline model suitable for forecasting mortality in most demographic scenarios. Moreover, we will show how the suggested approach is an ideal balance between pure statistical methodology and traditional demographic models allowing large flexibility in the inclusion of prior knowledge about mortality development.

The remainder of this paper is structured as follows. Section 2 presents basic assumptions and describes the original P-spline methodology which lays the groundwork for further steps. Section 3 is then devoted to the proposed CP-spline approach for incorporating demographic knowledge into the model. Inferences on estimated patterns are also provided. Trends in age-specific death rates and in summary measures such as life expectancy at birth and lifespan variations are presented in Section 4. A critical discussion of the methodology and possible extensions conclude the paper. Throughout the paper and solely for illustrative purposes two datasets are used: Danish females and USA males. Additional supplementary materials will serve as means to further assess the performance of the proposed method by out-of-sample forecast using 8 populations and in comparison with 5 alternative forecasting methods. We also validate the influence of the time-window and robustness with respect to the parameters used in the model.

2 *P*-splines for mortality data

The proposed model requires two simple datasets as input data: deaths, and exposures to the risk of death, arranged in two $m \times n_1$ matrices, $\boldsymbol{Y} = (y_{ij})$ and $\boldsymbol{E} = (e_{ij})$:

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_1} \\ y_{21} & y_{22} & \cdots & y_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn_1} \end{bmatrix} \qquad \boldsymbol{E} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n_1} \\ e_{21} & e_{22} & \cdots & e_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mn_1} \end{bmatrix} . \tag{1}$$

Rows and columns are classified by single age at death, $\boldsymbol{a}$, $m \times 1$, and single year of death, $\boldsymbol{t}_1$, $n_1 \times 1$, respectively.

We assume that the number of deaths y_{ij} at age i in year j is Poisson distributed with mean $\mu_{ij}\, e_{ij}$ (Keiding, 1990):

$$y_{ij} \sim \mathcal{P}(e_{ij}\, \mu_{ij}) . \tag{2}$$

The value of μ_{ij} is commonly named force of mortality and its estimation is the object of all mortality models. For instance, the matrix of the empirical mortality rates, which are the fully non-parametric estimations of the force of mortality, can be easily computed as $\mu_{ij} \approx m_{ij} = y_{ij}/e_{ij}$. Forecasting approaches aim to reconstruct trends in μ_{ij} for n_2 future years, $\boldsymbol{t}_2, n_2 \times 1$.

In the following we will illustrate the proposed method on two populations - Danish females and US males - which differ in terms of their epidemiological history. A more extensive application can be found in the supplementary materials. Among populations with highest longevity in 1960, Danish females life expectancy shows a non-linear pattern with a stagnation in the 1970s, followed by rising in the last two decades. Cohort effects and high smoking prevalence have been shown to be the main driven factors behind this peculiar pattern (among others, Jacobsen et al., 2004; Lindahl-Jacobsen et al., 2016). US males have shown a constant stagnation of mortality starting from the 1970s and several determinants stood out in the literature: smoking behaviour (Thun et al., 2013), obesity (Masters et al., 2013), the performance of the health care system (Muennig and Glied, 2010) as well as the lately observed drug epidemic among young-adults (Dowell et al., 2017). A good performance on these two populations for a newly proposed forecasting method is a clear sign of the robustness and flexibility of the approach. Moreover, using data for both females and males serves as a challenge for the proposed methodology on diverse mortality age-patterns. For both populations, we use data from the Human Mortality Database (2018), from ages 0 to 105 over the period 1960-2014, forecasting up to 2050.

We will now give an overview of the *P*-spline approach for Poisson distributed data in both one- and two-dimensional settings. A more extensive description of the method can be found in the seminal paper of Eilers and Marx (1996) as well as in the review article by Eilers et al. (2015). A demographic perspective is provided in Camarda (2008).

In a simple one-dimensional setting, we extract either a column or a row of the original matrices of death counts and exposures, i.e. $\boldsymbol{y}$ and $\boldsymbol{e}$. In modelling mortality, one aims to portray the expected values of the Poisson distribution as follows:

$$\ln[\mathbb{E}(\boldsymbol{y})] = \ln(\boldsymbol{e}) + \ln(\boldsymbol{\mu}) = \ln(\boldsymbol{e}) + \boldsymbol{\eta} \tag{3}$$

where $\boldsymbol{\eta}$ is the linear predictor and, dealing with Poisson data, a logarithm is used as link-function. The logarithm of the exposures, $\ln(\boldsymbol{e})$, is commonly called offset.

In a parametric setting we would model the linear predictor by a simple structure. For instance, a Gompertz law over age can be written as follows:

$$\boldsymbol{\eta} = \boldsymbol{X}\,\boldsymbol{\alpha} \tag{4}$$

where $\boldsymbol{X} = [\boldsymbol{1} : \boldsymbol{a}]$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$. Commonly, these two parameters are used to describe the starting level of mortality and rate-of-aging, respectively.

In a smoothing context, instead of deciding a prior mortality shape, we describe the log-mortality as a linear combination of B-splines and associated coefficients:

$$\boldsymbol{\eta} = \boldsymbol{B}\,\boldsymbol{\alpha}\,, \tag{5}$$

where $\boldsymbol{B}$ are k equally spaced knots of B-spline bases: bell-shaped curves composed of smoothly joined polynomial pieces of degree q. In the following we will use cubic B-splines, $q = 3$. The positions on the horizontal axis, where the pieces come together, are called “knots”. Details on B-splines and related algorithms can be found in de Boor (1978).

The basic idea of the P-splines is to combine (fixed-knot) B-splines with a roughness penalty. A relatively large number of B-splines ensures enough flexibility to capture trends in the mortality patterns, and a roughness penalty acting on the associated coefficients enforces the desirable amount of smoothness. Specifically, the number of B-splines as well as their degree is irrelevant on final results (Eilers et al., 2015).

A penalized version of the iteratively re-weighted least squares (IRLS) algorithm (McCullagh and Nelder, 1989) is sufficient for estimating coefficients $\boldsymbol{\alpha} \in \mathbb{R}^k$:

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})\tilde{\boldsymbol{\alpha}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \tag{6}$$

where $\tilde{\boldsymbol{z}} = (\boldsymbol{y} - \boldsymbol{e} * \tilde{\boldsymbol{\mu}})/\boldsymbol{e} * \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\eta}}$ is the working dependent variable. The tilde-symbol and $*$ denote current approximations to the solution and element-wise product, respectively. $\tilde{\boldsymbol{W}}$ is a diagonal matrix of weights, $\tilde{\boldsymbol{W}} = \mathrm{diag}(\boldsymbol{e} * \tilde{\boldsymbol{\mu}})$.

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 a penalty factor given by

$$\boldsymbol{P} = \lambda\,\boldsymbol{D}'\boldsymbol{D}\,, \tag{7}$$

where the matrix $\boldsymbol{D}$ constructs differences in the coefficients over either ages or years.

When using a standard P-spline approach, the choice of the order of difference is crucial only for forecasting (cf. Section 2.1). Second order difference will be used in the following. The smoothing parameter λ regulates the trade-off between goodness-of-fit and effective dimension used in the model. On the one hand, higher values will lead to a higher penalty term and, consequently, smoother fitted values. On the other, $\lambda = 0$ results in a straightforward GLM estimation with B-splines as regressors.

Our aim is to model and forecast mortality both over age and time, so we need to set up a P-splines model in a two-dimensional setting. For the purpose of regression, we arrange the complete matrices as a column vector, that is, $\boldsymbol{y} = \mathtt{vec}(\boldsymbol{Y})$ and $\boldsymbol{e} = \mathtt{vec}(\boldsymbol{E})$. Then we can directly use Eq. (6) to estimate coefficients over age and years by generalizing both basis and penalty term.

Let $\boldsymbol{B}_a$, $m \times k_a$ and $\boldsymbol{B}_{t_1}$, $n_1 \times k_{t_1}$ be the B-splines over ages and years, respectively. The regression matrix for our two-dimensional model is given by

$$\boldsymbol{B} = \boldsymbol{B}_{t_1} \otimes \boldsymbol{B}_a \,, \tag{8}$$

where $\otimes$ denotes the Kronecker product of two matrices. Following the same idea as for the one-dimensional case, we will use a relatively large number of equally spaced B-splines over both domains ($k_a = 24$, $k_{t_1} = 14$).

Concerning the 2D generalization of the penalty term, from the definition of Kronecker product, Currie et al. (2004) show that $\boldsymbol{\alpha}$ can be independently penalized over ages and years. Let $\boldsymbol{D}_a$ and $\boldsymbol{D}_{t_1}$ be the difference matrices acting on the two domains. A two-dimensional penalty is given by

$$\boldsymbol{P} = \lambda_a(\boldsymbol{I}_{k_{t_1}} \otimes \boldsymbol{D}'_a\boldsymbol{D}_a) + \lambda_{t_1}(\boldsymbol{D}'_{t_1}\boldsymbol{D}_{t_1} \otimes \boldsymbol{I}_{k_a})\,, \tag{9}$$

where λ_a and λ_{t_1} are the smoothing parameters used for age and year, respectively. $\boldsymbol{I}_{k_a}$ and $\boldsymbol{I}_{k_{t_1}}$ are identity matrices of dimension k_a and k_{t_1}, respectively.

By changing λ_a and λ_{t_1}, smoothness can be tuned to balance smoothness and model fidelity. In the following λ_a and λ_{t_1} will be selected by minimizing the Bayesian Information Criterion (BIC, Schwarz, 1978) which penalizes model complexity more heavily and is more suitable for mortality data (Camarda, 2008; Currie et al., 2004).

As in the one-dimensional setting, B-splines provide enough flexibility to capture surface trends. The additional penalty reduces the number of parameters leading to a *wisely* 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 considerable model flexibility. Furthermore, P-splines in 2D can be embedded in the class of Generalized Linear Array Model, saving computational time and reducing storage problems in the estimation of the model (Currie et al., 2006).

2.1 Forecasting with P-splines

In the original paper by Currie et al. (2004), forecasting is treated as a missing value problem and the smooth surface is simply extrapolated into future years. Keeping the same age range and forecasting over the years, we augment data and B-spline bases as follows:

$$\breve{\boldsymbol{Y}} = [\boldsymbol{Y} : \boldsymbol{Y}_2]\,, \qquad \breve{\boldsymbol{E}} = [\boldsymbol{E} : \boldsymbol{E}_2]\,, \qquad \breve{\boldsymbol{B}} = [\boldsymbol{B}_{t_1} : \boldsymbol{B}_{t_2}] \otimes \boldsymbol{B}_x\,,$$

where $\boldsymbol{D}_2$ and $\boldsymbol{E}_2$ are $m \times n_2$ matrices filled with arbitrary future values. In this paper, the complete B-spline basis over years will extend the original basis with additional 8 B-splines.

Finally, let us denote by $\mathbf{1}_{m\times n_1}$ an all-ones matrix over ages and first n_1 observed years. Similarly $\mathbf{0}_{m\times n_2}$ is a zero matrix over ages and future n_2 years. If we define a weight matrix $\boldsymbol{V}$:

$$\boldsymbol{V} = \mathrm{diag}(\mathtt{vec}(\mathbf{1}_{m\times n_1} : \mathbf{0}_{m\times n_2}))\,, \tag{10}$$

we can adapted the algorithm in (6) as follows:

$$(\breve{\boldsymbol{B}}'\boldsymbol{V}\tilde{\boldsymbol{W}}\breve{\boldsymbol{B}} + \boldsymbol{P})\tilde{\boldsymbol{\alpha}} = \breve{\boldsymbol{B}}'\boldsymbol{V}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \tag{11}$$

with $\tilde{\boldsymbol{z}} = \boldsymbol{V}(\breve{\boldsymbol{y}} - \breve{\boldsymbol{e}} * \tilde{\boldsymbol{\mu}})/\breve{\boldsymbol{e}} * \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\eta}}$. The penalty is also augmented to account for future years by difference matrices $\boldsymbol{D}_a$ and $\boldsymbol{D}_{t_1+t_2}$. This unified structure allows us to simultaneously model and forecast mortality. Moreover, the structure of $\boldsymbol{V}$ makes it clear that first we use all observed data and second we assume nothing about the future, i.e. weights equal to one for the observed years and zero for the forecast horizon.

It is noteworthy that the form of the penalty determines the form of the forecast. Whereas the order of the penalty is negligible in the model section, it has a crucial role in the forecasting section. As suggested by Currie et al. (2004), second-order differences in $\boldsymbol{D}_{t_1+t_2}$ are preferable because the forecast coefficients will be approximately linear over the future years and this choice is more appropriate when working with mortality data. However, this aspect will lose its relevance when prior demographic knowledge is incorporated into the model (cf. Section 3).

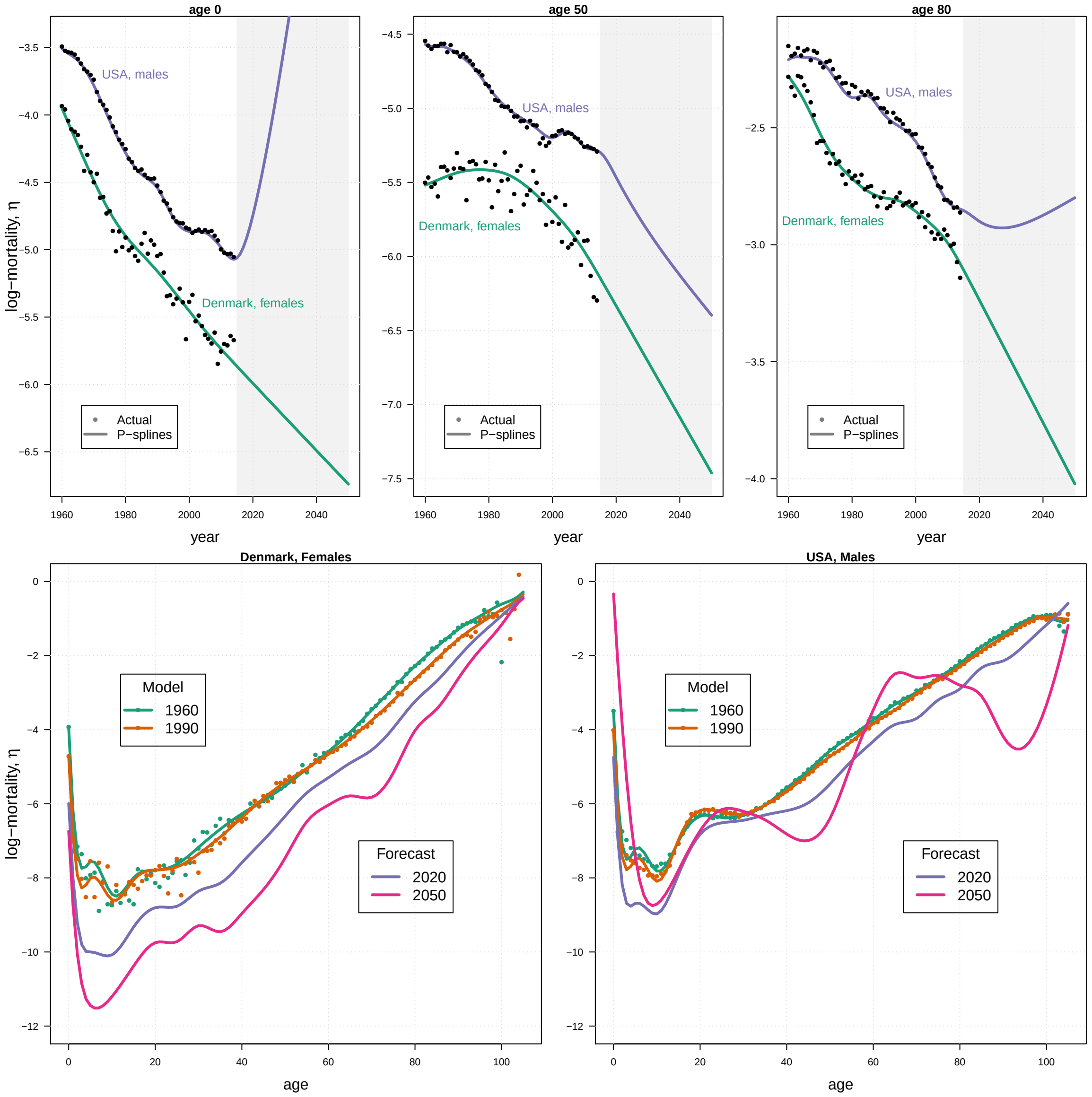

Figure 1: Actual, model and forecast mortality. Two-dimensional P-spline approach. Selected ages over years (top panels) and selected years over ages (bottom panels). Danish females and US males, ages 0-105, years 1960-2014, forecast up to 2050.

Figure 1 presents the outcomes of a two-dimensional P-splines approach in modeling and forecasting mortality data for Danish females and US males. The top panels show actual and fitted as well as forecast trends for selected ages (0, 50, 80). Concerning estimation on observed data, P-splines show a rather good fit, though unsmooth patterns are visible around age 10 in both populations (bottom panels in Figure 1). Nevertheless, an odd mortality increase in future years is visible for some ages in the US male data and likely too fast mortality improvement is forecast at age 80 for Danish females. These outcomes are obviously implausible given the observed mortality trends of the past years. The bottom panels in Figure 1 mirror the above-mentioned results from a different perspective. For the future years mortality age-profiles are obviously improbable given the knowledge we already have on the phenomenon: unreasonable wiggling behavior is evident from age 20 onward in both populations. This outcome can be seen as a consequence of a low smoothing parameter selected by the BIC for the age domain. A small λ_a is necessary in any case to allow flexibility in describing properly the whole age-pattern with its notable peak at age 0.

3 The CP-spline model

In the last section, we noted that several issues may occur when two-dimensional P-splines are applied without any demographic consideration in a plain data-driven approach. Specifically, we need to deal with infant mortality preserving smoothness in the following ages. Moreover, previous outcomes call for inclusion in forecast years of prior knowledge on typical mortality age-profiles and time-trends.

3.1 Addressing infant mortality

The first year of life is usually treated in a different manner when life tables are constructed (Chiang, 1984). Therefore, we decided to follow this practice by modifying the basis related to the age domain. The new basis will then be:

$$\boldsymbol{B}_a^0 = \begin{bmatrix} 1 & \boldsymbol{0}_{1\times k_a} \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{B}_a \end{bmatrix}, \tag{12}$$

where $\boldsymbol{B}_a$ is now a $(m-1)\times k_a$ matrix of B-splines.

Moreover we replace the first cell in $\boldsymbol{D}_a$ by a zero which implies that infant mortality is not connected with variation in subsequent coefficients. In other words, we separate development of infant mortality from the remaining ages.

Using this new basis and a new penalty ensures that a single and specialized coefficient will be attached to infant mortality values. In a one-dimensional setting this additional coefficient will be exactly the log of death rates at age 0. In a two-dimensional framework, we allow for a smooth change in infant mortality over time.

On the one hand, we increase the number of coefficients in the model. On the other, we allow a certain freedom in describing mortality at age 0 via its specific series of coefficients over years. This ensures that the smoothness of the surface from age 1 onward will not be affected by the evident disruption due to infant mortality.

Figure 2 shows the outcomes for our Danish and US datasets when mortality levels at age 0 are explicitly considered. Although the forecast trends are still unreasonable, especially for US males, considering infant mortality as a peculiar phenomenon helps to improve goodness-of-fit during the observed period: BIC reduces from 80,131 to 36,353 for US males and from 7,949 to 7,517 for Danish females. The specifically optimized

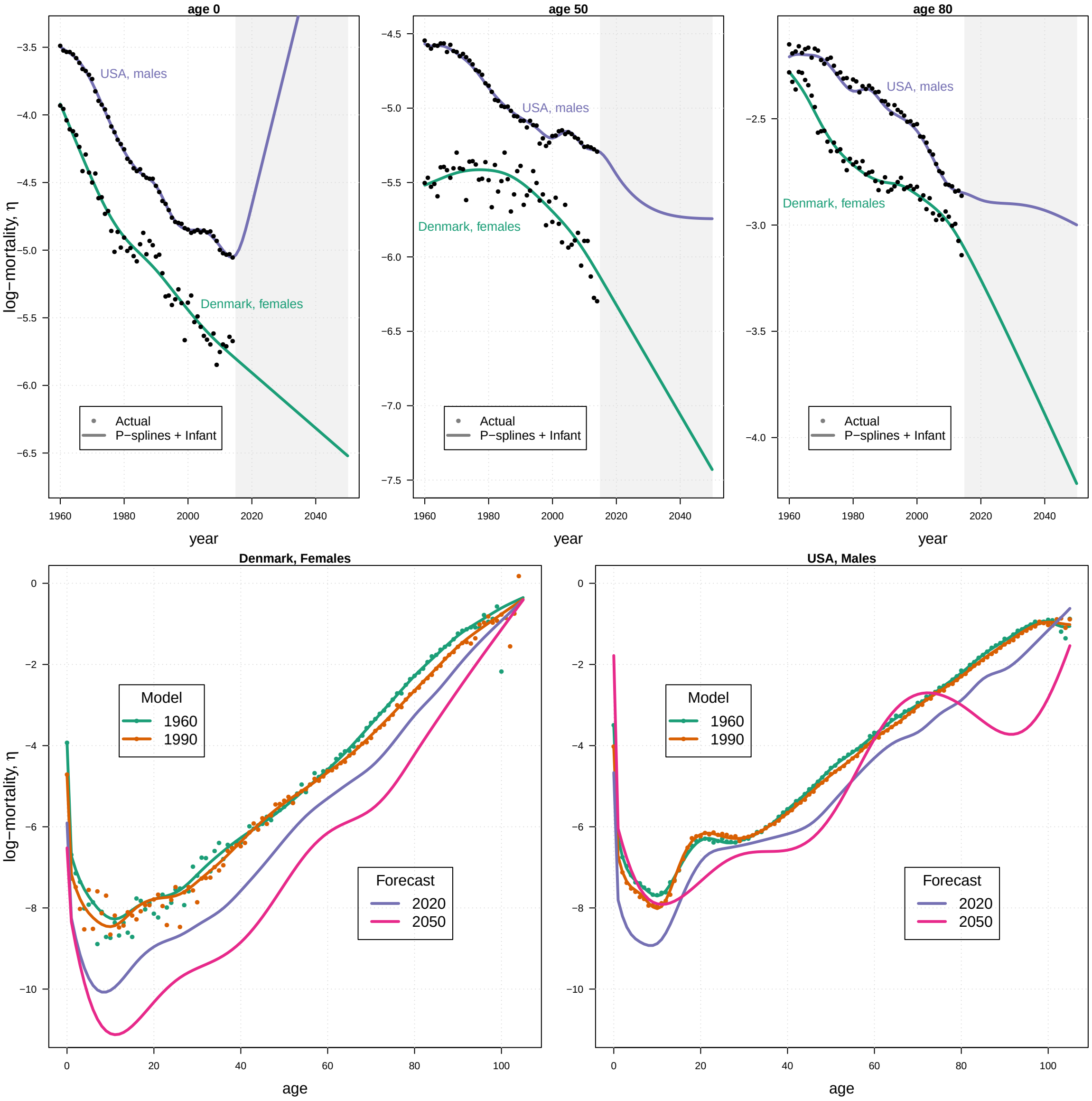

Figure 2: Actual, model and forecast mortality. Two-dimensional P-spline approach including a specialized coefficient for infant mortality. Selected ages over years (top panels) and selected years over ages (bottom panels). Danish females and US males, ages 0-105, years 1960-2014, forecast up to 2050.

smoothing parameter for the age domain (λ_a) becomes larger with respect to the original approach. As a result, we achieve smoother outcomes over ages, avoiding wiggling behavior around age 10 in both populations.

3.2 Enforcing mortality patterns over age and time

As we have seen in Section 2.1, the original P-spline approach is purely data-driven and the extrapolated trends are based on the last estimated coefficients solely constrained by a certain amount of smoothness. However, following Gompertz (1825), demographers started observing well defined regularities in the shape of mortality over ages. Moreover, past mortality trends present a certain predictability which must guide any model. It would be unreasonable to disregard information on mortality patterns over ages and time in a forecasting method.

Regarding the age dimension, Figure 3 shows the mean along with the 95% confidence interval over age based on the mortality surface, smoothed by P-splines with the additional specialized coefficient for infant mortality (cf. Section 3.1). A rather stable mortality behavior over ages is evident. In general, at infant ages, mortality decreases steeply, dropping rapidly within the first few years (Levitis, 2011). A minimum is commonly reached at about ages 10-15. Afterwards, especially for men, mortality rates show a hump at young-adult ages (Goldstein, 2011; Remund, 2015). Mortality then rises exponentially after approximately age 30 and levels off at ages above 80 (Preston, 1976; Thatcher et al., 1998).

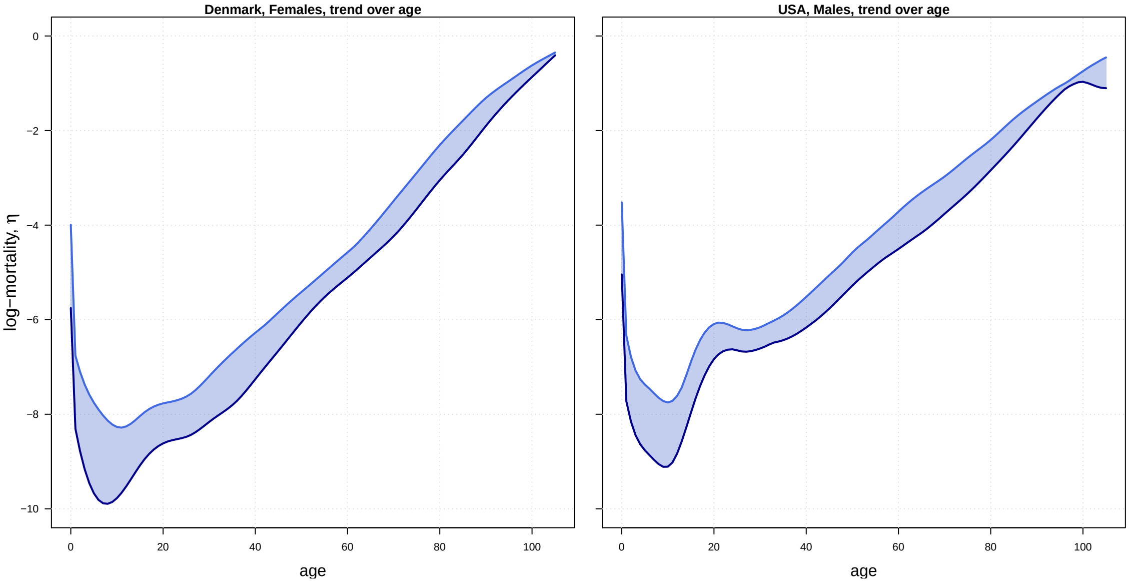

Figure 3: 95% point-wise confidence intervals of mortality age-profiles from fitted two-dimensional P-splines approach. Model includes a specialized coefficient for infant mortality. Danish females (left panel) and US males (right panel), ages 0-105, years 1960-2014.

We aim to incorporate the information about this stable profile within the forecast period without modifying fitted values which are based on observed and past data. Instead of borrowing mortality profiles from model life tables or parametric models, we constrain forecast age profiles lie within the 95% confidence interval of the fitted age profiles.

Since we aim to carry out our analysis referring to the mortality shape and regardless of its level, our constraints must be based on the relative derivatives of the age mortality profile, commonly named rate-of-aging. Given the estimated linear predictor $\hat{\boldsymbol{\eta}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}}$, the rate-of-aging for each year can be computed by a linear combination of a modified version of the B-splines and the estimated coefficients:

$$\frac{\frac{\partial}{\partial \boldsymbol{a}}\hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}} = \frac{\partial}{\partial \boldsymbol{a}} \ln(\hat{\boldsymbol{\mu}}) = \frac{\partial}{\partial \boldsymbol{a}}\hat{\boldsymbol{\eta}} = \boldsymbol{D}_a^{t_1}\,\hat{\boldsymbol{\alpha}}\,, \tag{13}$$

where the matrix $\boldsymbol{D}_a^{t_1}$ computes the first difference of the estimated coefficients for each year and simultaneously multiplies them by B-splines of lower degree. In formula:

$$\boldsymbol{D}_a^{t_1} = \boldsymbol{B}_{t_1} \otimes \boldsymbol{C}_a\,, \tag{14}$$

where

$$\boldsymbol{C}_a = \frac{1}{h}\left[{}^{q-1}\boldsymbol{B}_a^k - {}^{q-1}\boldsymbol{B}_a^{k-1}\right] \tag{15}$$

with h, q and k being knot-distance, degree and positions of the original B-spline basis, $\boldsymbol{B}_a$.

In this way, using directly estimated coefficients, we can compute instantaneous rate-of-aging over all ages and for each year. This allows us to modify the algorithm in (11) for incorporating possible constraints. Moreover, working on smooth mortality surfaces, estimated relative derivatives over age will not show the wiggling behavior produced by simple differentiation of observed death rates.

Figure 4 presents 95% confidence intervals of the instantaneous rate-of-aging over all ages above 0 for Danish females and US males. We denote by $\boldsymbol{\delta}_L^a$ and $\boldsymbol{\delta}_U^a$ the lower and upper bounds of these confidence intervals, respectively.

Relative derivatives for infant mortality are not displayed for a better readability of the graph: a steep decrease in mortality at age 0 will enormously expand the limits of the ordinate of the associated rate-of-aging. Specifically, the 95% confidence interval of the relative derivatives with respect to age 0 is $[-2.76, -2.39]$ for Danish females and $[-2.83, -2.55]$ for US males.

In Figure 4 we can read the mortality age-patterns of our data without referring to their level. In general, values above zero correspond to mortality increase and, conversely, ages with mortality reduction coincide with negative values of rate-of-aging.

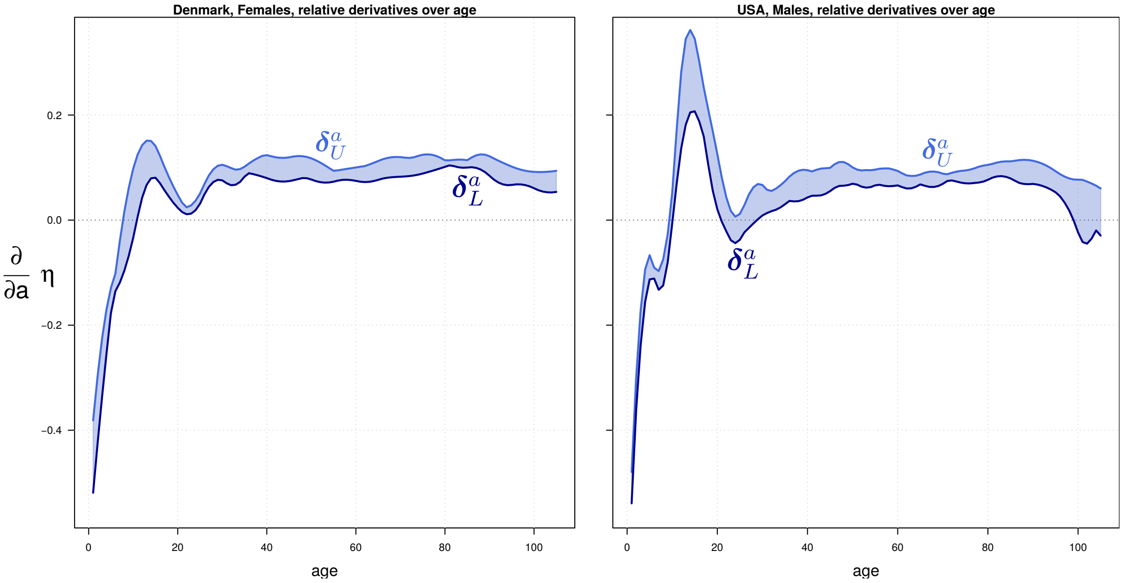

Figure 4: 95% point-wise confidence intervals of relative derivatives of force of mortality with respect to age, i.e. rate-of-aging, $(\boldsymbol{\delta}_L^a, \boldsymbol{\delta}_U^a)$. Shown only for ages 1-105. Fitted values computed from fitted two-dimensional P-splines approach. Model included a specialized coefficient for infant mortality. Danish females (left panel) and US males (right panel), ages 0-105, years 1960-2014.

The rather constant values about 0.1 after age 30 in both datasets correspond to the exponential mortality increase in adult and old ages, commonly described by a single parameter in the Gompertz model, α_2 in Eq. (4). However, we do not predefine constant rate-of-aging for adult mortality and we are also capable of capturing a possible levelling-off of mortality at oldest-old ages: see the declining relative derivatives above age 80.

The rate-of-aging also shows clear sex differences in young-adult mortality: Danish females present a U-shape pattern about age 20 which is much less pronounced than for US males. This latter population also reaches relative derivatives equal to zero around age

25, which means a corresponding constant mortality. In both populations the decreasing mortality trend during childhood is clear: the associated rate-of-aging shows a steeply increasing pattern with negative values.

Concerning time trends, we can compute relative derivatives with respect to time as we did for the age domain:

$$\frac{\frac{\partial}{\partial t_1}\hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}} = \frac{\partial}{\partial \boldsymbol{t}_1}\ln(\hat{\boldsymbol{\mu}}) = \frac{\partial}{\partial \boldsymbol{t}_1}\hat{\boldsymbol{\eta}} = \boldsymbol{D}_{t_1}^{t_1}\,\hat{\boldsymbol{\alpha}}\,. \tag{16}$$

Again matrix $\boldsymbol{D}_{t_1}^{t_1}$ computes first difference of estimated coefficients by differentiating the associated B-spline basis over years for each age. Unlike the age profile, the rate-of-change over time is more fluctuating, as one can see by looking at a particular age (65) for both Danish females and US males on Figure 5. Whereas the trend is generally smooth and downward, there are periods of mortality increase: during the 1980s for Danish females and the most recent years for US males. Mortality stagnation in US males is also evident during the 1960s.

These features are immediately visible in the associated relative derivatives with respect to time for mortality at age 65 in both populations (bottom panels in Figure 5) which express mortality improvement over time regardless of the actual level. Likewise for the rate-of-aging, negative values correspond to downward mortality trends which represents the majority of the observed relative derivatives for this specific age. Positive values are observed when mortality stagnates and/or deteriorates.

Though values for mortality rate-of-change over time are smaller than the observed rate-of-aging, variation is wider: every mortality fluctuation over time - even minor - is amplified in the associated relative derivatives. Note, for example, the time-trend in US males aged 65 in the most recent years: a rather small estimated increase translates into a disproportionate jump in the rate-of-change (right panels in Figure 5). This is a well-known issue in statistics: derivatives will always show undersmooth behavior with respect to the associated estimated function (Erickson et al., 1995).

On the one hand, we aim to constrain future mortality developments to lie within observed mortality rates-of-change. On the other, we intend to avoid in future years the peculiar past mortality trends that we assume to be solely due to specific and unlikely events. For this reason we decide to use only 50% confidence intervals of the observed mortality rate-of-change, i.e. the interquartile range of past experienced mortality development.

We recommend the interquartile range based on several experiments conducted on numerous populations from the Human Mortality Database (2018) (not shown here). However, it is important to note that this value (50%) is simply a way to express the forecaster's prior knowledge of how past mortality should inform future development. As a result, different attitudes toward the future or the peculiar mortality history of a specific population may guide forecasters to different values in computing $\boldsymbol{\delta}_L^{t_1}$ and $\boldsymbol{\delta}_U^{t_1}$. In the Supplementary Materials C we evaluate this choice for Danish females and USA males: outcomes do not change markedly, as long as extreme mortality fluctuations over time are not considered.

The horizontal green stripes in the bottom panels of Figure 5 depict the 50% confidence intervals of the observed mortality rate-of-change for age 65 in both populations.

Figure 6 shows the 50% confidence intervals of the relative derivatives of the force of mortality with respect to time for each age. The lower and upper levels of these confidence intervals are denoted by $\boldsymbol{\delta}_L^{t_1}$ and $\boldsymbol{\delta}_U^{t_1}$, respectively.

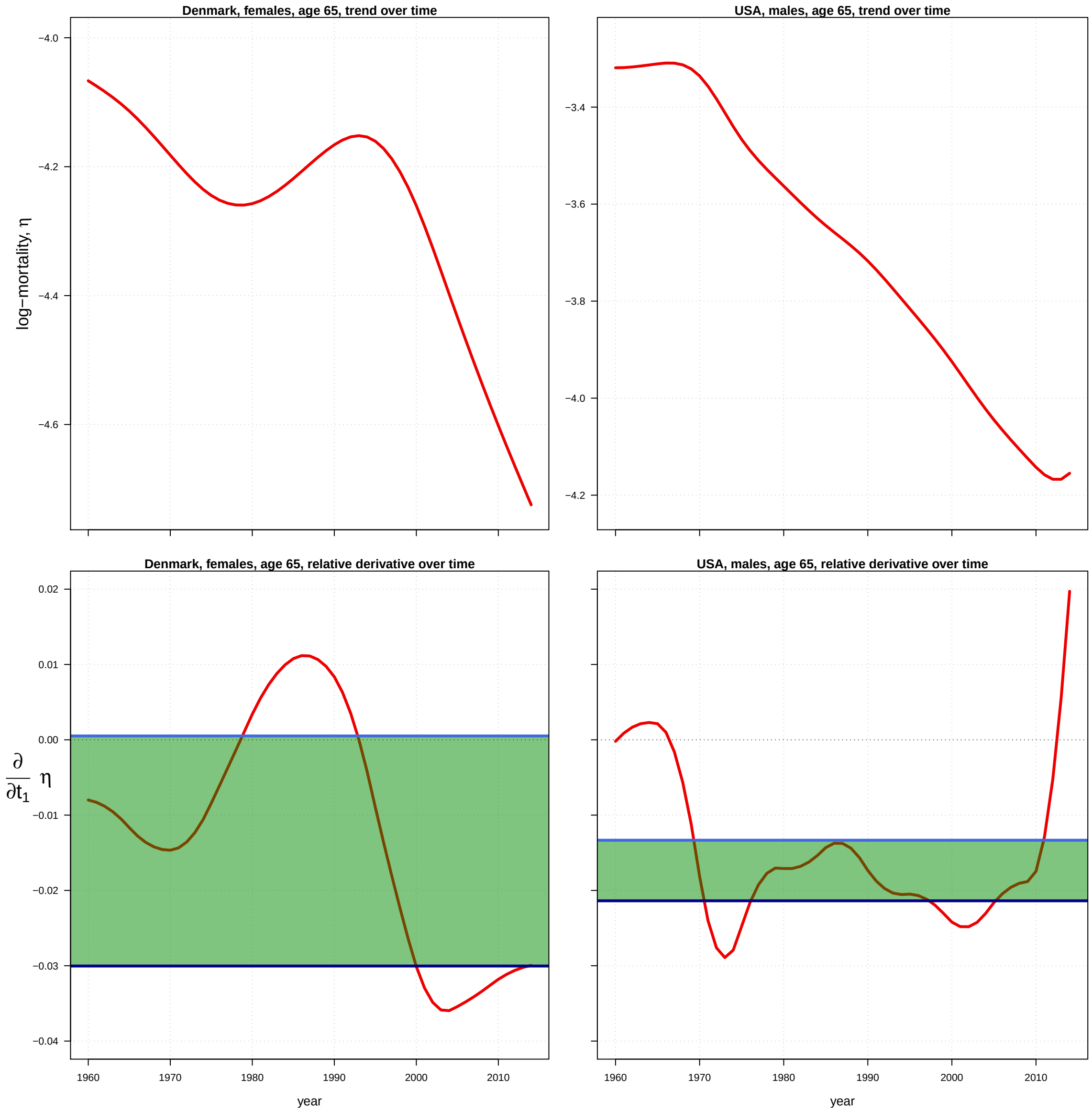

Figure 5: Actual and smooth log-mortality for age 65 over time (top panels) and associated rate-of-change (bottom panels). Horizontal blue lines depict 50% confidence intervals of the rate-of-change. Fitted values computed by two-dimensional P-splines approach. Model includes a specialized coefficient for infant mortality. Danish females (left panel) and US males (right panel), ages 0-105, years 1960-2014.

In Figure 6 we can easily see which ages have experienced greater improvements (larger negative values) as well as with greater variations in terms of observed mortality changes (broader confidence bounds).

The idea is to inform our model about rate-of-aging and mortality changes over time, i.e. constrain future mortality to lie within a range of plausible age patterns and time trends expressed by estimated values and portrayed in Figures 4 and 6. Hence, we propose the Constrained Penalized spline (CP-spline) model.

Though specific values are suggested for computing $\boldsymbol{\delta}$, forecasters can adapt the model to their needs and to prior knowledge about future mortality development. In general, the lower the percentages in estimating the vectors of $\boldsymbol{\delta}$, the closer future mortality will be to mean age patterns and time trends, as observed in past years. In other words, extremely

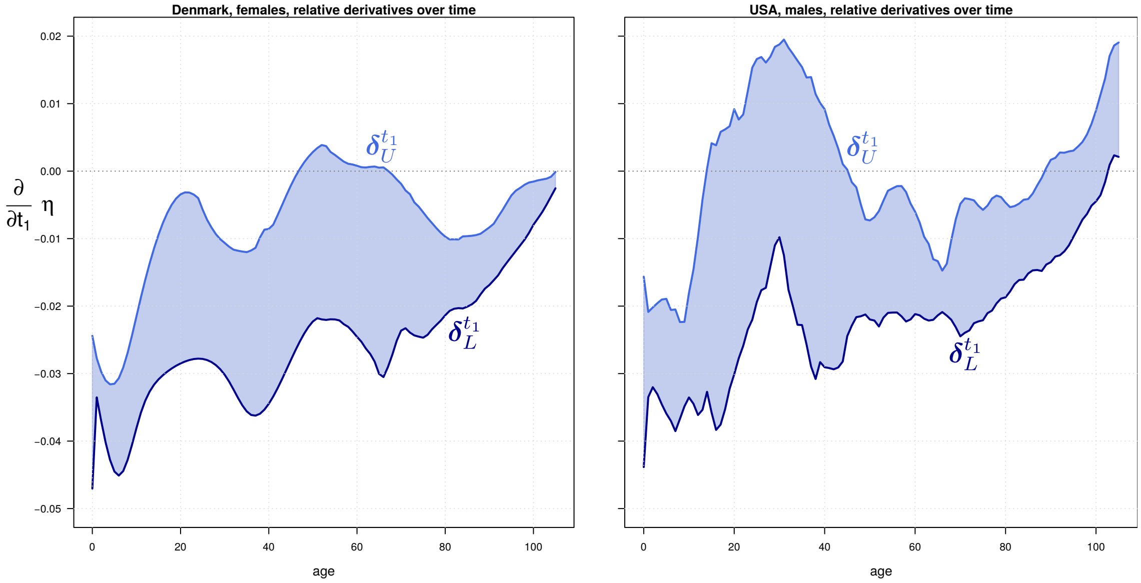

Figure 6: 50% point-wise confidence intervals of relative derivatives of force of mortality with respect to time for each age, i.e. age-specific rate-of-change. Fitted values computed from fitted two-dimensional P-splines approach. Model included a specialized coefficient for infant mortality. Danish females (left panel) and US males (right panel), ages 0-105, years 1960-2014.

low percentages for $\boldsymbol{\delta}$ lead to a fixed age profile along with an invariant age specific rate of mortality improvement, i.e. something similar to the Lee-Carter model, but with a non-linear time index based on the amount of smoothness. Conversely, large confidence levels leads to an extremely flexible CP-spline model without much prior knowledge about future mortality, i.e. a plain P-splines approach. Additionally, different levels of mortality improvement can be set for different ages, making the model highly flexible.

We must warn the forecaster on two important issues. On the one hand, allowing high flexibility may lead to unreasonable outcomes (see Figure 1). On the other, a rigid model will forecast patterns which reproduce slavishly the structure of the model in future years. Secondly, recommended levels for computing $\boldsymbol{\delta}$ in CP-splines should be adopted with care: although they have been tested for many datasets in Human Mortality Database (2018), specific populations might need distinct confidence levels to either account or neglect unique patterns over age and/or years.

3.2.1 Incorporating prior knowledge into the model

Once the constraints are set, we retain them for all years by augmenting the values in $\boldsymbol{\delta}$ over both dimensions:

$$\begin{array}{rcl} \boldsymbol{g}_L^a & = & \mathbf{1}_{n_1+n_2} \otimes \boldsymbol{\delta}_L^a \\ \boldsymbol{g}_U^a & = & \mathbf{1}_{n_1+n_2} \otimes \boldsymbol{\delta}_U^a \end{array} \quad \text{and} \quad \begin{array}{rcl} \boldsymbol{g}_L^t & = & \mathbf{1}_{n_1+n_2} \otimes \boldsymbol{\delta}_L^{t_1} \\ \boldsymbol{g}_U^t & = & \mathbf{1}_{n_1+n_2} \otimes \boldsymbol{\delta}_U^{t_1} \end{array} \tag{17}$$

We enforce our shape constraints by adding asymmetric penalties within the system introduced in (11):

$$(\breve{\boldsymbol{B}}'\boldsymbol{V}\tilde{\boldsymbol{W}}\breve{\boldsymbol{B}} + \boldsymbol{P} + \boldsymbol{P}^a + \boldsymbol{P}^t)\tilde{\boldsymbol{\alpha}} = \breve{\boldsymbol{B}}'\boldsymbol{V}\tilde{\boldsymbol{W}}\breve{\boldsymbol{z}} + \boldsymbol{p}^a + \boldsymbol{p}^t \,. \tag{18}$$

where

$$\begin{array}{rcl} \boldsymbol{P}^a & = & \boldsymbol{P}_L^a + \boldsymbol{P}_U^a \\ \boldsymbol{P}^t & = & \boldsymbol{P}_L^t + \boldsymbol{P}_U^t \end{array} \quad \text{and} \quad \begin{array}{rcl} \boldsymbol{p}^a & = & \boldsymbol{p}_L^a + \boldsymbol{p}_U^a \\ \boldsymbol{p}^t & = & \boldsymbol{p}_L^t + \boldsymbol{p}_U^t \end{array} . \tag{19}$$

As an example, we present the penalty terms for the lower bounds over ages. The other terms are constructed in a similar fashion. In formulas

$$\begin{array}{rcl} \boldsymbol{P}_L^a & = & \kappa\, \boldsymbol{D}_a^{t_1+t_2}\, \texttt{diag}(\boldsymbol{s}\, \tilde{\boldsymbol{v}}_L^a)\, \boldsymbol{D}_a^{t_1+t_2} \\ \boldsymbol{p}_L^a & = & \kappa\, \boldsymbol{D}_a^{t_1+t_2}\, \texttt{diag}(\boldsymbol{s}\, \tilde{\boldsymbol{v}}_L^a)\, \boldsymbol{g}_L^a \end{array} \quad \text{with} \quad \boldsymbol{v}_L^a = \begin{cases} 0 & \text{if} \quad \boldsymbol{D}_a^{t_1+t_2}\tilde{\boldsymbol{\alpha}} \geqslant \boldsymbol{g}_L^a \\ 1 & \text{if} \quad \boldsymbol{D}_a^{t_1+t_2}\tilde{\boldsymbol{\alpha}} < \boldsymbol{g}_L^a \end{cases} \tag{20}$$

and $\boldsymbol{s}$ is a 0/1 vector equal to 1 when the constraint is to be applied (future years).

Note that, being asymmetric, these penalties act on both the left- and right-hand sides of the system of equations and values in $\boldsymbol{v}_L^a$, $\boldsymbol{v}_U^a$, $\boldsymbol{v}_L^t$ and $\boldsymbol{v}_U^t$ are computed iteratively. In other words, for each new value of $\tilde{\boldsymbol{\alpha}}$ during the iteration (18), the shape penalties exert an influence only when the shape constraint is violated. The size of κ regulates how strictly the constraints are enforced. In this paper, we chose $\kappa = 10^4$ which is an intermediate value in this setting: we inform the model about our shape constraints, but meanwhile their effects should not overwhelm the smoothness behavior of the fitted values as expressed in the roughness penalty $\boldsymbol{P}$. More details on asymmetric penalties and their statistical properties can be found in Eilers (2005) and Bollaerts et al. (2006).

3.2.2 Confidence Intervals by Bootstrap

No forecasting method can be completely satisfactory without a good estimation of the uncertainty affecting the forecast quantities. Estimation of confidence intervals is thus particularly necessary.

Plain two-dimensional P-splines are a straightforward extension of a regression model and methods for computing the covariance matrix and associated standard errors can be borrowed from regression theory. However, with our CP-spline model, we depart from this setting with the inclusion of asymmetric penalties in the system of equations.

In the absence of analytical solutions for the estimation of uncertainty in our model, we opt for confidence intervals constructed via a bootstrap approach. We are thus able to combine all sources of uncertainty in the model and simultaneously compute confidence intervals for summary measures which are complicated non-linear functions of the estimated coefficients. Details on bootstrap methods in general can be found in Efron and Tibshirani (1993). While bootstraps have been used by Koissi et al. (2006) and Brouhns et al. (2005) in the Lee-Carter context, Ouellette et al. (2012) have adapted this methodology to a two-dimensional P-spline model.

Following Koissi et al. (2006) and Ouellette et al. (2012), we carried out a residual bootstrap of our fitted model. Specifically, deviance residuals are the standard measures to assess the discrepancy between actual and fitted data (McCullagh and Nelder, 1989, p. 39-40):

$$\boldsymbol{r} = \text{sign}(\boldsymbol{y} - \hat{\boldsymbol{y}})\sqrt{2\left[\boldsymbol{y}\,\ln\left(\frac{\boldsymbol{y}}{\hat{\boldsymbol{y}}}\right) - \boldsymbol{y} + \hat{\boldsymbol{y}}\right]}\,. \tag{21}$$

These residuals should be independent and identically distributed (provided the model is well specified). Hence we sample from them (with replacement) an entire new set of residuals $\boldsymbol{r}_{(b)}$ called the bootstrapped residuals. Replacing deviance residuals $\boldsymbol{r}$ by bootstrapped residuals $\boldsymbol{r}_{(b)}$ in (21) and rearranging the equation leads to

$$\hat{\boldsymbol{y}} - \boldsymbol{y}\ln(\hat{\boldsymbol{y}}) + \boldsymbol{r}_{(b)}^2 + \boldsymbol{y} - \boldsymbol{y}\ln(\boldsymbol{y}) = 0\,. \tag{22}$$

Given $\boldsymbol{r}_{(b)}$ and actual death $\boldsymbol{y}$, equation (22) can be solved numerically with respect to $\hat{\boldsymbol{y}}$, thus obtaining a new matrix of bootstrapped deaths $\hat{\boldsymbol{y}}_{(b)}$. Together with the original exposures $\boldsymbol{e}$, we can estimate the proposed CP-spline model on the bootstrapped deaths $\hat{\boldsymbol{y}}_{(b)}$ obtaining new bootstrapped coefficients $\hat{\boldsymbol{\alpha}}_{(b)}$.

The procedure described above, starting with the residual sampling step, was repeated 1,000 times. This led to a bootstrapped distribution of constrained and penalized coefficients. From this distribution, we extract empirical percentiles and compute confidence intervals for force of mortality $\boldsymbol{\mu}$, linear predictor $\boldsymbol{\eta}$ as well as for any desirable summary measure, e.g. life expectancy at birth.

Whereas common approaches focuses on the variability in the (univariate) time index (see both Lee-Carter and Hyndman-Ullah variants), residual bootstrap incorporates the variability from all model parameters. This approach is more suitable in a non-parametric framework and it allows us to account for uncertainty due to the Poisson stochastic process in (2), i.e. larger populations would tend to have narrower confidence intervals in the fitting periods.

It is noteworthy that assuming smoothness may result in narrowing uncertainties around estimations in the observed years. However, flexibility of a nonparametric approach like CP-splines produces larger confidence intervals in future years with respect to models with intrinsic rigid structures. Finally, confidence intervals will not account for uncertainty due to model misspecification: values of optimized smoothing parameters and confidence levels for computing $\boldsymbol{\delta}$ are kept fixed in the bootstrapping procedure.

To sum up, a forecaster that intends to estimate and forecast mortality by CP-splines needs to adopt the following procedure:

1. collect as in (1) deaths and exposures in two matrices over age and years;
2. estimate a two-dimensional P-spline model over the observed period with the algorithm in (6) and a specialized basis for infant mortality as in (12);
3. optimize smoothing parameters in the penalty term (9);
4. evaluate rate-of-aging and rate-of-change for each age using (13) and (16);
5. portray previous relative derivatives as in Figures 6 and 4 to decide on level of confidence for future rate-of-aging $(\boldsymbol{\delta}_L^a, \boldsymbol{\delta}_U^a)$ and rate-of-change $(\boldsymbol{\delta}_L^{t_1}, \boldsymbol{\delta}_U^{t_1})$. Here, eventual prior knowledge about mortality developments in a specific population could modify the recommended 95% and 50% levels;
6. solve the system of equations in (18) which adds in the forecasting algorithm (11) the asymmetric penalty terms
7. carry out the residual bootstrap presented above to obtain confidence intervals of the fitted model

Routines for running all previous steps and forecast mortality with CP-splines were implemented in `R` (R Development Core Team, 2018). Codes are available from the author [on the journal's web-site].

4 Application

Figure 7 shows the outcomes of the proposed smooth constrained forecasting model. In order to better visualize uncertainty around estimated values, we portray outcomes by mean of `fanplot` (Abel et al., 2013). Coloured bends are limited by 10% and 90% of the empirical percentiles.

Moreover, in this section, we compare the outcomes from the suggested CP-splines with a variant of the Lee-Carter model. Specifically, we apply the Lee-Carter model estimated in a smooth setting as described in Delwarde et al. (2007). Both CP-spline model and this Lee-Carter variant are embedded in a Poisson framework and smoothing of the coefficients is ensured in both settings. Hence differences between models will be solely due to differences in model structure.

It is evident in Figure 7 how fitted values from CP-splines follow actual patterns adequately and forecast values from CP-splines present reasonable trends both over ages and years. Specifically, there is no wiggling behavior since we enforce a specific shape via the asymmetric penalties and this also ensures no crossover of adjacent ages in the long term. Moreover, future mortality of US males no longer shows increasing trends.

Additionally to an accurate modelling performance due to its flexibility, the proposed approach ensures smooth mortality age-patterns for future years. This is extremely important in both demographic and actuarial studies since it means that forecast mortality can be treated in a continuous setting.

At first glance, the boundary of the confidence intervals in Figure 7 is extremely close to the fitted values in the observed period (1960-2014). This is mainly due to the large expected values in the Poisson distribution (2) which are the products of force of mortality and population exposure, remarkably large for US males. Consequently, we see relatively larger confidence bends at oldest ages over the years. In this area, high force of mortality is compensated by exposures with moderate counts. Similarly, slightly wider variability is observed at young ages (see pattern at age 20). This results from low force of mortality along with a large exposure population. The relatively larger uncertainty at age 0 is due to its peculiar treatment within the model: smoothness is enforced only over time and therefore larger variability is expected. Moreover, variability increases for all ages the further we move toward future years.

In Figure 7 it is evident how the proposed CP-spline approach clearly outperforms the Lee-Carter model in both fitting past trends: for US males, the Deviance which measures goodness-of-fit in a Poisson setting, is equal to 30,977 for CP-splines and 158,590 for the smooth Lee-Carter variant. For Danish females, the Deviance is equal to 6,605 (11,059) for CP-splines (smooth Lee-Carter). A larger comparison with other forecasting methods is available in the supplementary materials to this paper.

Concerning patterns for future years, the Lee-Carter model is not able to capture all observed mortality changes of the past decades and its forecast trends are often unreasonable, i.e. a simple linear extrapolation of fitted values which are a poor description of actual trends. It is noteworthy that, though embedded in a smooth setting, the rigidity of the Lee-Carter structure leads to unsmooth fitted values. On the contrary, the proposed model does not suffer from this drawback and is able to accommodate diverse trends.

A direct consequence of these differences is visible in the age-profiles in 2050 (bottom panels of Figure 7). Whereas the CP-spline model yields smooth and plausible future mortality age-patterns, the Lee-Carter produces unrealistic wiggling behaviour of the age-profile in 2050 for both populations.

We decided to assess the performance of the proposed CP-spline model by means of two complementary summary measures. Life expectancy at birth is used as a classic measure of average lifespan. Lifespan variation describes differences in the length of life across members of a population and, among the large number of possible measures, we opt for the average number of life years lost at birth (Vaupel and Canudas-Romo, 2003; Zhang and Vaupel, 2009), commonly denoted by $e_0^\dagger$. Easy to interpret as a potential increase in life expectancy at birth, this measure has already been used to evaluate the performance of mortality forecasts (Bohk-Ewald et al., 2017).

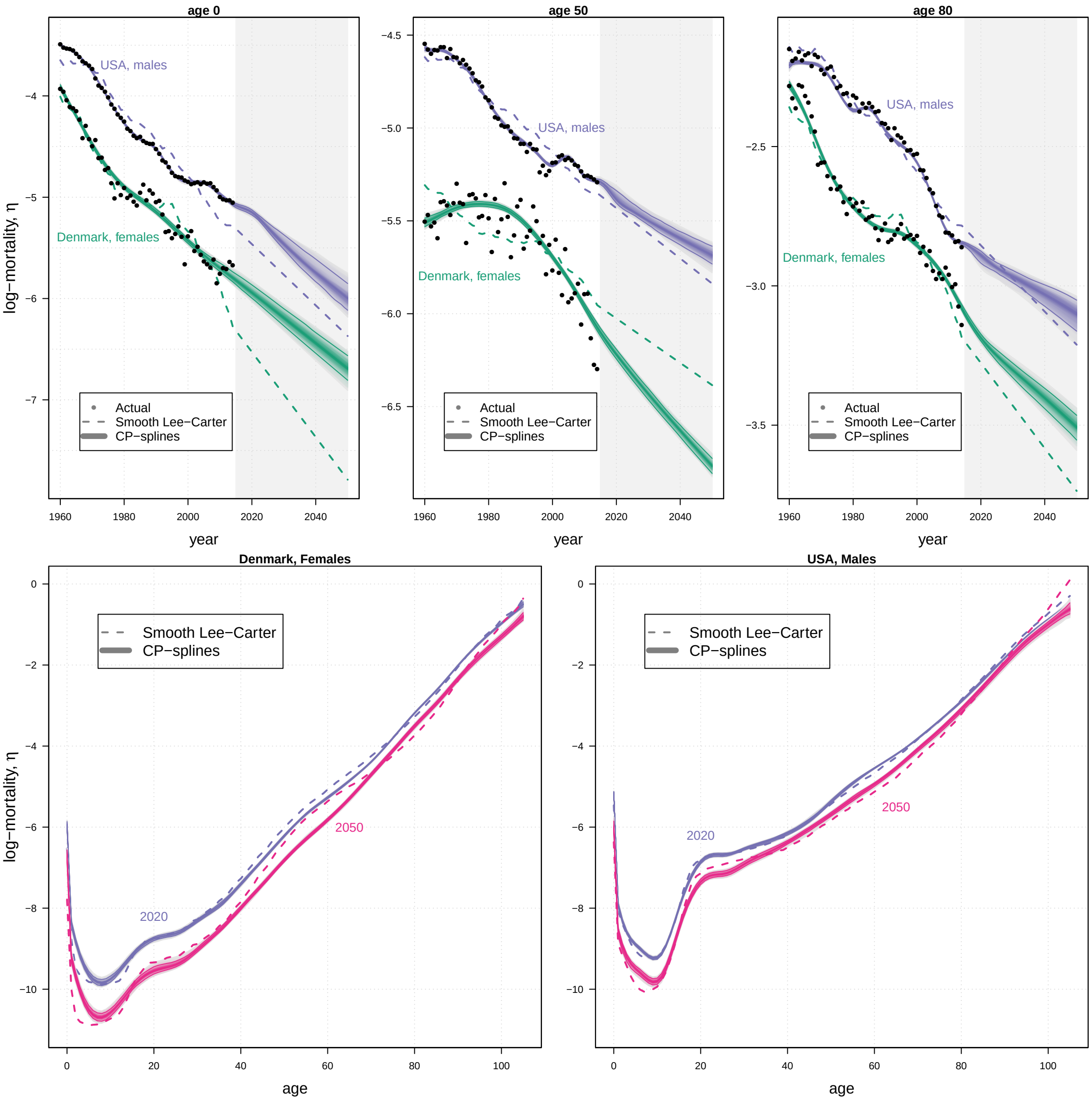

Figure 7: Actual, model and forecast mortality along with their bootstrapped distributions. Colored bends depict 80% of the empirical percentiles. Constrained Penalized spline model including a specialized coefficient for infant mortality. Estimates from a smooth Lee-Carter variant are given for comparative reasons. Selected ages over years (top panels) and selected years over ages (bottom panels). Danish females and US males, ages 0-105, years 1960-2014, forecast up to 2050.

The results in terms of these summary measures are presented in Figure 8. As in the case for the log-mortality, the CP-spline model fits rather well the development of e_0 over the past years in both populations: the largest difference in e_0 between actual and fitted values is equal to 0.26 and 0.66 years for US males and Danish females, respectively.

In 2050 the proposed method results in a life expectancy at birth of 87.1 years with a 95% confidence interval (86.55-87.63) for Danish females, and of 80.54 years for US males (80.01-81.34). We compare our results to those released by the United Nations in the 2017 World Population Prospects, medium variant. For Danish females, for the periods 2045-50 and 2050-2055, they provide a life expectancy at birth equal to 85.82 and 86.40, respectively. These values are slightly lower than our forecast 95% confidence intervals.

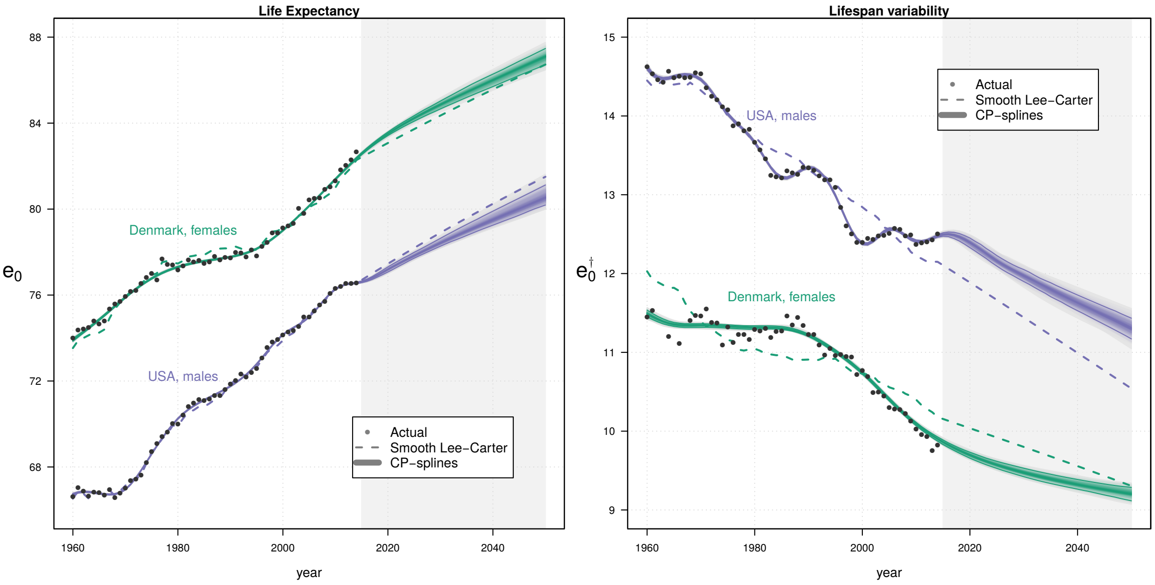

Figure 8: Actual, model and forecast values for life expectancy at birth (left panel) and a measure of lifespan variability ($e_0^\dagger$, right panel). Colored bends depict 80% of the empirical percentiles. Constrained Penalized spline model including a specialized coefficient for infant mortality. Estimates from a smooth Lee-Carter variant are given for comparative reasons. Danish females and US males, ages 0-105, years 1960-2014, forecast up to 2050.

On the contrary, we produce more pessimistic outcomes for US males: United Nations forecast e_0 equal to 81.91 and 82.76 for the two periods about 2050.

Lifespan variability measured by $e_0^\dagger$ is extremely well captured by the CP-spline model for the period 1960-2014: the largest errors in estimating average number of life years lost at birth is equal to 0.1 (0.24) years for US males (Danish females). Forecast trends also appear to be reasonable. For US males, we forecast an average of 11.31 life years lost at birth in 2050 with a 95% confidence interval (11.12-11.52). Danish females already start with a lower level of $e_0^\dagger$ and smoothly move to a value of 9.20 in 2050 (9.09-9.33) with a levelling-off trend. The relationships between both summary measures are also adequately described and forecast by the proposed model (not shown here).

Despite a similar final value in 2050, the development of future life expectancy at birth is different from that of the Lee-Carter model which simply extrapolates a linear trend from erroneously estimated values. The proposed CP-spline model, on the other hand, is able to capture curved trends over time and consequently to provide more reliable forecasts.

Unlike the proposed approach which is capable of reproducing all changes in lifespan variability measures, trends in $e_0^\dagger$ are completely misfitted by the Lee-Carter model which then linearly extrapolates its fitted values.

In the supplementary materials, we further assess the performance of the proposed model in additional ways. We compare CP-splines with alternative methods (Suppl. Mat. A) and we test stability with respect to the time-window over which the model is estimated, an important and often neglected choice in all forecasting methods (Suppl. Mat. B).

Specifically, we present a large and detailed comparison study in which CP-splines have been analyzed with 5 alternative forecasting methods: the mentioned smooth variant of the Lee-Carter model (Delwarde et al., 2007) as well as the Lee-Carter variants proposed Lee and Miller (2001) and Booth et al. (2002), and two of the approaches proposed within a functional data framework by Hyndman and Ullah (2007). We model 4 countries (USA, Denmark, France and Japan), both males and females, for the period 1960-2014 and forecast up to 2050. For the observed years, we assess goodness-of-fit (balanced with model complexity) using the Bayesian Information Criterion: in all datasets the proposed CP-splines outperformed the alternative approaches.

To evaluate the performance performance against observed mortality trends, we performed an out-of-sample forecast study. We estimate all models on the period 1960-2004, forecast up to 2014 and compare to actual values in 2005-2014. Models are compared using three different measures computed on e_0, $e_0^{\dagger}$ and the whole mortality surface ($\boldsymbol{\eta}$). Given 4 populations, 2 sexes, 3 accuracy measures and 3 demographic indicators, we compare six alternative approaches over 72 values: the proposed CP-splines outperformed its competitors 51 times (71%), followed by a variant of Hyndman-Ullah model with only 8 times. All details are showed in Supplementary Materials A (Table C.1 and C.2).

5 Conclusions

Our paper starts from a simple consideration about established forecasting methods for mortality. We recognized that widely used methodologies are either too rigid to properly capture mortality developments over age and time, or too flexible to impose certain well-known structures in absence of observed data.

Among the rigid models one can certainly list parametric models such as Gompertz and Heligman-Pollard: these models predefine a mortality law over age, and forecast estimated parameters will reproduce with a blind adherence these laws in future years. However, several Lee-Carter variants suffer from equivalent drawbacks describing the whole mortality developments within a bi-linear model and fixing age-specific rate of mortality improvement. In order to free mortality models from rigid structures, nonparametric methods have been suggested. Superior in describing observed patterns, these approaches do not account for demographic knowledge in guiding forecast mortality developments.

Our study bridges the gap between these seemingly distant approaches. We enhance a powerful nonparametric statistical methodology such as the P-splines, incorporating observed demographic information from past years into the model. Under a P-spline approach we obtain good fit, flexibility and smooth outcomes. Nevertheless, we show that a plain smoothing approach results in unreasonable outcomes when used to forecast mortality. This purely data-driven approach is not able to reproduce past mortality experience since it extrapolates last estimated trends with blind adherence. A certain amount of smoothness is the only restriction integrated into the model and it is not sufficient when no data are available, as in the case for future years.

Thus, our proposal accustoms the P-splines approach constraining future mortality to lie within trends estimated from observed data. We thus propose a Constrained Penalized spline model. Initially, we consider infant mortality as a specific age and attach to age 0 a specialized coefficient. This improves the estimation of past trends, but is insufficient to correct inconsistent trends in future years. Consequently, we additionally incorporate information about past mortality experience. Instead of working on observed log-mortality, we operate in terms of rate-of-aging and rate-of-change over time. In practice, we enforce shape constraints by asymmetric penalties based on observed relative derivatives of the force of mortality with respect to age and time. Uncertainty on estimated and forecast quantities is then computed by a residual bootstrap approach. In this way we are able to simultaneously smooth past trends and forecast mortality in a sensible manner.

Results on two distinct populations in terms of mortality development (Danish females and US males) show that the CP-spline method performs well. Low deviance indicates that we are able to accurately capture past trends. Moreover, forecast mortality patterns and age profiles are reasonable given the past observed development. Fitted and forecast summary measures are also presented and these results confirm that the proposed forecast model performs remarkably well with very small errors in the observed period. Moreover we show how the suggested CP-splines outperforms a smooth variant of the Lee-Carter model and, in the supplementary materials, a wider comparison shows the better performances of CP-splines with respect to other forecasting methods.

In the paper, we specified levels of confidence for rate-of-aging and rate-of-change aiming to constrain future mortality to observed shapes. These levels can be adapted to express diverse prior knowledge about future mortality for a distinct population and/or for specific ages. Specifically, our model might provide a means for guiding expert-based forecast approaches: whereas it is hard to explicitly foresee future values for age-specific mortality rates, experts in the field may have more precise opinions on (un-)feasible ranges of mortality improvements for ages, or groups of ages, looking to what has been observed in the past. This is an obviously population-specific procedure and it will help to forecast populations that experienced exceptional patterns such as HIV epidemics, wars and general mortality crises. We envisage a generalization of the proposed CP-splines for these peculiar situations.

Specific cohort effects can be observed in certain populations and, in these situations, forecasters search for a procedure to incorporate these effects in the forecast horizon. Shape constraints over the diagonal of the mortality surface can be employed for accommodating peculiar cohort behaviors without influencing neighboring cohorts. Similarly, the proposed model may provide a flexible and rigorous approach for deriving the age-pattern of mortality given a predicted level of life expectancy and/or lifespan variability measures. We plan to extend our model along both lines.

Finally, we think that the CP-spline model can be generalized to produce coherently forecasts for both sexes, or more populations, following a recent strand of research on mortality forecasting (Ahmadi and Li, 2014; Bergeron-Boucher et al., 2016; Hyndman et al., 2013; Li and Lee, 2005; Shang, 2016; Ševčíková et al., 2016). Constraining future mortality to observed age-profiles and time-trends can be also viewed as a way of constraining different sub-populations to behave analogously in terms of mortality shape and trend. This idea can be also adapted to estimate, and eventually forecast, mortality patterns for small areas where prior knowledge is often necessary to obtain reasonable outcomes. We shall pursue these ideas in future research.

References

Abel, G., J. Bijak, J. J. Forster, J. Raymer, P. W. Smith, and J. S. Wong (2013). Integrating uncertainty in time series population forecasts: An illustration using a simple projection model. *Demographic Research 29*, 1187–1226.

Ahmadi, S. S. and J. S.-H. Li (2014). Coherent mortality forecasting with generalized linear models: A modified time-transformation approach. *Insurance: Mathematics and Economics 59*, 194–221.

Barrieu, P., H. Bensusan, N. El Karoui, C. Hillairet, S. Loisel, C. Ravanelli, and Y. Salhi (2012). Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian actuarial journal 2012*(3), 203–231.

Bergeron-Boucher, M.-P., V. Canudas-Romo, J. Oeppen, and J. W. Vaupel (2016). Coherent forecasts of mortality with compositional data analysis. *Demographic Research 37*, 527–566.

Blake, D., A. J. G. Cairns, and K. Dowd (2006). Living with mortality: Longevity bonds and other mortality-linked securities. *British Actuarial Journal 12*(1), 153–197.

Bohk-Ewald, C., M. Ebeling, and R. Rau (2017). Lifespan Disparity as an Additional Indicator for Evaluating Mortality Forecasts. *Demography 54*(4), 1559–1577.

Bohk-Ewald, C. and R. Rau (2017). Probabilistic mortality forecasting with varying age-specific survival improvements. *Genus 73*(1), 1–37.

Bollaerts, K., P. H. C. Eilers, and I. van Mechelen (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology 59*, 451–469.

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

Booth, H. and L. Tickle (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science 3*(1-2), 3–43.

Brouhns, N., M. Denuit, and I. Van Keilegom (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal 3*, 212–224.

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

Cairns, A. J. G., D. Blake, and K. Dowd (2006). Pricing death: Frameworks for the valuation and securitization of mortality risk. *ASTIN Bulletin: The Journal of the IAA 36*(1), 79–120.

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, A. Ong, and I. Balevich (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal 13*(1), 1–35.

Camarda, C. G. (2008). *Smoothing Methods for the Analysis of Mortality Development.* Ph. D. thesis, Programa de Doctorado en Ingeniería Matemática. Universidad Carlos III, Departamento de Estadística, Madrid.

Camarda, C. G. (2012). MortalitySmooth: An `R` Package for Smoothing Poisson Counts with P-Splines. *Journal of Statistical Software 50*, 1–24. Available on `www.jstatsoft.org/v50/i01`.

Carfora, M. F., L. Cutillo, and A. Orlando (2017). A quantitative comparison of stochastic mortality models on Italian population data. *Computational Statistics and Data Analysis 112*, 198–214.

Chiang, C. (1984). *The Life Table and its Application.* Malabar, FL: Krieger.

Colchero, F., R. Rau, O. R. Jones, J. A. Barthold, D. A. Conde, A. Lenart, L. Nemeth, A. Scheuerlein, J. Schoeley, C. Torres, V. Zarulli, J. Altmann, D. K. Brockman, A. M. Bronikowski, L. M. Fedigan, A. E. Pusey, T. S. Stoinski, K. B. Strier, A. Baudisch, S. C. Alberts, and J. W. Vaupel (2016). The emergence of longevous populations. *Proceedings of the National Academy of Sciences 113*(48), E7681–E7690.

Currie, I. D. (2011). Modelling and forecasting the mortality of the very old. *ASTIN Bulletin: The Journal of the IAA 41*, 419–427.

Currie, I. D. (2013). Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling 13*, 69–93.

Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal 2016*(4), 356–383.

Currie, I. D., M. Durbán, and P. H. C. Eilers (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling 4*, 279–298.

Currie, I. D., M. Durbán, and P. H. C. Eilers (2006). Generalized Linear Array Models with Applications to Multidimensional Smoothing. *Journal of the Royal Statistical Society. Series B 68*, 259–280.

D'Amato, V., G. Piscopo, and M. Russolillo (2011). The mortality of the Italian population: Smoothing techniques on the LeeCarter model. *The Annals of Applied Statistics 5*(2A), 705–724.

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer.

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.

Djeundje, V. A. B. and I. D. Currie (2011). Smoothing Dispersed Counts with Applications to Mortality Data. *Annals of Actuarial Science 5*, 33–52.

Dowell, D., R. K. Noonan, and D. Houry (2017). Underlying Factors in Drug Overdose Deaths. *Journal of the American Medical Association 318*(23), 2295–2296.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

Eilers, P. H. C. (2005). Unimodal Smoothing. *Journal of Chemometrics 19*, 317–328.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with B-splines and Penalties (with discussion). *Statistical Science 11*, 89–102.

Eilers, P. H. C., B. D. Marx, and M. Durbán (2015). Twenty years of P-splines. *SORT. Statistics and Operations Research Transactions 39*(2), 149–186.

Erickson, R. V., V. Fabian, and J. Marik (1995). An Optimum Design for Estimating the First Derivative. *The Annals of Statistics 23*, 1234–1247.

Gerland, P., A. E. Raftery, H. Ševčíková, N. Li, D. Gu, T. Spoorenberg, L. Alkema, B. K. Fosdick, J. L. Chunn, N. Lalic, G. Bay, T. Buettner, G. K. Heilig, and J. Wilmoth (2014). World population stabilization unlikely this century. *Science 346*, 234–237.

Goicoa, T., M. D. Ugarte, J. Etxeberria, and A. F. Militino (2012). Comparing CAR and P-spline models in spatial disease mapping. *Environmental and ecological statistics 19*(4), 1–27.

Goldstein, J. R. (2011). A Secular Trend toward Earlier Male Sexual Maturity: Evidence from Shifting Ages of Male Young Adult Mortality. *PLoS ONE 6*(8), e14826. doi:10.1371/journal.pone.0014826.

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.

Huang, F. and B. Browne (2017). Mortality forecasting using a modified Continuous Mortality Investigation Mortality Projections Model for China I: methodology and country-level results. *Annals of Actuarial Science 11*(1), 20–45.

Human Mortality Database (2018). *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).* Available at `www.mortality.org`. Data downloaded on January 2018.

Hyndman, R. J., H. Booth, and F. Yasmeen (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography 50*, 261–283.

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.

Jacobsen, R., M. Von Euler, M. Osler, E. Lynge, and N. Keiding (2004). Women's death in scandinavia - what makes denmark different? *European Journal of Epidemiology 19*(2), 117–121.

Jones, O. R., A. Scheuerlein, R. Salguero-Gomez, C. G. Camarda, R. Schaible, B. B. Casper, J. P. Dahlgren, J. Ehrlen, M. B. Garcia, E. S. Menges, P. F. Quintana-Ascencio, H. Caswell, A. Baudisch, and J. W. Vaupel (2014). Diversity of ageing across the tree of life. *Nature 505*(7482), 169–173.

Keiding, N. (1990). Statistical Inference in the Lexis Diagram. *Philosophical Transactions: Physical Sciences and Engineering 332*, 487–509.

Koissi, M.-C., A. F. Shapiro, and G. Högnäs (2006). Evaluating and extending the Lee-Carter model for mortality forecasting:Bootstrap confidence interval. *Insurance: Mathematics and Economics 38*, 1–20.

Lee, R. D. and L. R. Carter (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association 87*, 659–671.

Lee, R. D. and T. Miller (2001). Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality. *Demography 38*, 537–549.

Levitis, D. A. (2011). Before senescence : the evolutionary demography of ontogenesis. *Proceedings of the Royal Society B : Biological Sciences 278*(1707), 801–809.

Li, N. and R. D. Lee (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography 42*, 575–594.

Li, N., R. D. Lee, and P. Gerland (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality-decline for long-term projection. *Demography 50*, 2037–2051.

Lindahl-Jacobsen, R., R. Rau, B. Jeune, V. Canudas-Romo, A. Lenart, K. Christensen, and J. W. Vaupel (2016). Rise, stagnation, and rise of Danish womens life expectancy. *Proceedings of the National Academy of Sciences 113*(15), 4015–4020.

Lu, J. L. C., W. Wong, and M. Bajekal (2014). Mortality improvement by socio-economic circumstances in England (1982 to 2006). *British Actuarial Journal 19*(1), 1–35.

Masters, R. K., E. N. Reither, D. A. Powers, Y. C. Yang, A. E. Burger, and B. G. Link (2013). The Impact of Obesity on US Mortality Levels: The Importance of Age and Cohort Factors in Population Estimates. *American Journal of Public Health 103*(10), 1895–1901.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). Monographs on Statistics Applied Probability. London: Chapman & Hall.

Minton, J., R. Shaw, M. A. Green, L. Vanderbloemen, F. Popham, and G. McCartney (2017). Visualising and quantifying 'excess deaths' in Scotland compared with the rest of the UK and the rest of Western Europe. *Journal of Epidemiology & Community Health 71*(5), 461–467.

Muennig, P. A. and S. A. Glied (2010). What Changes in Survival Rates Tell Us About US Health Care. *Health Affairs 29*(11), 2105–2113.

Ouellette, N. and R. Bourbeau (2011). Changes in the age-at-death distribution in four low mortality countries: a nonparametric approach. *Demographic Research 25*, 595–628.

Ouellette, N., R. Bourbeau, and C. G. Camarda (2012). Regional Disparities in Canadian Adult and Old-age Mortality: A Comparative Study Based on Smoothed Mortality Ratio Surfaces and Age-at-death Distributions. *Canadian Studies in Population 39*(3-4), 79–106.

Pitacco, E., M. Denuit, and S. Haberman (2009). *Modelling longevity dynamics for pensions and annuity business.* Oxford University Press.

Preston, S. H. (1976). *Mortality Patterns in National Populations. With special reference to recorded causes of death.* Academic Press.

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A. E., J. Chunn, P. Gerland, and H. Ševčíková (2013). Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography 50*, 777–801.

Remund, A. (2015). *Jeunesses vulnérables? Mesures, composantes et causes de la surmortalité des jeunes adultes.* Ph. D. thesis, University of Geneva.

Renshaw, A. E. and S. Haberman (2003). Lee-Carter Mortality Forecasting with Age-specific Enhancement. *Insurance: Mathematics and Economics 33*, 255–272.

Ribeiro, F. (2015). *Statistical analysis and forecasting of cause of death data: Novel approaches and insights.* Ph. D. thesis, Universidade de Evora (Portugal).

Richards, S. J., I. D. Currie, and G. P. Ritchie (2014). A value-at-risk framework for longevity trend risk. *British Actuarial Journal 19*(1), 116–139.

Richards, S. J., J. Kirkby, and I. D. Currie (2006). The Importance of Year of Birth in Two-Dimentional Mortality Data. *British Actuarial Journal 12*, 5–61.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*, 461–464.

Shang, H. L. (2016). Mortality and life expectancy forecasting for a group of populations in developed countries: A multilevel functional data method. *The Annals of Applied Statistics 10*, 1639–1672.

Tabeau, E., F. Willekens, and F. van Poppel (2002). Parameterisation as a tool in analysing age, period and cohort effects on mortality: a case study of the netherlands. In G. Wunsch, M. Mouchart, and J. Duchene (Eds.), *The life table: modelling survival and death*, pp. 141–169. Dordrecht etc.: Kluwer Academic Publishers.

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.

Thun, M. J., B. D. Carter, D. Feskanich, N. D. Freedman, R. Prentice, A. D. Lopez, P. Hartge, and S. M. Gapstur (2013). 50-Year Trends in Smoking-Related Mortality in the United States. *New England Journal of Medicine 368*(4), 351–364.

Trias-Llimós, S., M. J. Bijlsma, and F. Janssen (2016). The role of birth cohorts in long-term trends in liver cirrhosis mortality across eight European countries. *Addiction 112*, 250–258.

Ugarte, M. D., T. Goicoa, J. Etxeberria, and A. F. Militino (2012). Projections of cancer mortality risks using spatio-temporal P-spline models. *Statistical methods in medical research 21*(5), 545–560.

Ugarte, M. D., T. Goicoa, and A. F. Militino (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics 21*(3-4), 270–289.

Vaupel, J. W. and V. Canudas-Romo (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography 40*, 201–216.

Ševčíková, H., N. Li, V. Kantorova, P. Gerland, and A. E. Raftery (2016). Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections. In R. Schoen (Ed.), *The Springer series on demographic methods and population analysis: Vol. 39. Dynamic demographic analysis*, pp. 285–310. Springer.

Wang, H. C., J. C. Yue, and Y.-H. Tsai (2016). Marital Status as a Risk Factor in Life Insurance: An Empirical Study in Taiwan. *ASTIN Bulletin: The Journal of the IAA 46*(2), 487–505.

Zhang, Z. and J. W. Vaupel (2009). The age separating early deaths from late deaths. *Demographic Research 20*, 721–730.