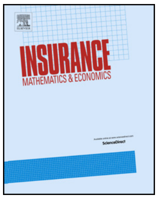
Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime



Cause-of-death mortality forecasting using adaptive penalized tensor decompositions

Xuanming Zhang [a], Fei Huang [b,*], Francis K.C. Hui [a], Steven Haberman [c]

[a] Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Australia
[b] School of Risk and Actuarial Studies, UNSW Sydney, Australia
[c] Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, United Kingdom

ARTICLE INFO

Article history:
Received October 2021
Received in revised form March 2023
Accepted 9 May 2023
Available online 16 May 2023

Dataset link: https://github.com/lullabies777/Adaptive-Penalized-Tensor-Decomposition.git

Dataset link: https://xuanmingzhang.shinyapps.io/rshiny_final/

JEL classification:
G22

Keywords:
Adaptive weights
Causes of death
Generalized lasso penalty
Model selection
Tensor decomposition

ABSTRACT

Cause-of-death mortality modeling and forecasting is an important topic in demography and actuarial science, as it can provide valuable insights into the risks and factors determining future mortality rates. In this paper, we propose a novel predictive approach for cause-of-death mortality forecasting, based on an adaptive penalized tensor decomposition (ADAPT). The new method jointly models the three dimensions (cause, age, and year) of the data, and uses adaptively weighted penalty matrices to overcome the computational burden of having to select a large number of tuning parameters when multiple factors are involved. ADAPT can be coupled with a variety of methods (e.g., linear extrapolation, and smoothing) for extrapolating the estimated year factors and hence for mortality forecasting. Based on an application to United States (US) male cause-of-death mortality data, we demonstrate that tensor decomposition methods such as ADAPT can offer strong out-of-sample predictive performance compared to several existing models, especially when it comes to mid- and long-term forecasting.

1. Introduction

Mortality modeling and forecasting is an important research topic in demography and actuarial science. Governments rely on mortality forecasts to make population projections and various aging and retirement-related policy decisions. It also plays a fundamental role in pricing, reserving, and capital modeling for life insurers and pension systems (He et al., 2021). Additionally, mortality rates can be regarded as an indicator of living quality, indirectly reflecting the socio-economic conditions of a country. An analysis of historic trends by cause of death can provide insight into some of the underlying drivers of mortality over time (Wilmoth, 1995; Hanewald, 2011), noting that there are often direct and delayed links between some causes of death and underlying risk factors e.g., lung cancer and smoking. Certain cause-of-death mortality rates can also decrease dramatically if there is a medical breakthrough for a widespread disease, or conversely increase by a non-negligible amount due to man-made or natural disasters such as wars and the outbreak of infectious diseases such as COVID-19. Cause-of-death mortality modeling thus plays a crucial role in quantifying and understanding these impacts.

Despite its importance, cause-of-death mortality forecasting continues to be a challenging task. Many early studies of cause-of-death mortality modeling often assumed that causes were independent of each other, and applied univariate models to forecast each cause separately. For example, McNown and Rogers (1992) used univariate Auto Regressive Integrated Moving Average (ARIMA) models to forecast parameters of an age-pattern mortality mode, while Caselli and Marsili (2006) applied linear, least squares, Lee-Carter (Lee and Carter, 1992), and APC (age, period, and cohort) models to extrapolate cause-of-death mortality rates. Both McNown and Rogers (1992) and Caselli

[*] Corresponding author at: School of Risk and Actuarial Studies, UNSW Business School, UNSW Sydney, NSW 2052, Australia.
E-mail address: feihuang@unsw.edu.au (F. Huang).

https://doi.org/10.1016/j.insmatheco.2023.05.003

and Marsili (2006) found that there was no advantage in taking into account the different causes of death when it came to forecasting all-cause or total mortality rates using their models. On the other hand, and as recognized by McNown and Rogers (1992), forecasting cause-specific mortality can be of interest in its own right, and this in turn has fueled growing research over the past decade into models specifically designed for cause-specific mortality data.

Gaille and Sherris (2011) and Arnold and Sherris (2013, 2015) modeled the age structure of cause-specific mortality rates using a Heligman-Pollard mortality model (Heligman and Pollard, 1980), and applied vector error correction models to the parameters of the Heligman-Pollard function to produce long-run trends and forecasting. Alternatively, Alai et al. (2015) used a multinomial logistic regression to model cause-of-death mortality, which allowed for a range of cause-specific mortality dependencies based on the mean structure used for each cause. As we shall see later on in our empirical analysis though, this approach is arguably more descriptive with its forecasting performance being relatively poor compared to (say) applying a Lee-Carter model to each cause individually. Other cause-of-death mortality models include forecast reconciliation techniques (Li and Lu, 2019), employing a copula framework (Li et al., 2019), and stochastic APC models to analyze cause-of-death specific cohort effects (Redondo Lourés and Cairns, 2021). Recently, Arnold and Glushko (2021) also developed a two-level cointegration analysis as an extension of Arnold-Gaille and Sherris (2016) to examine a large system of data variables (5 cause-specific mortality rates for 5 countries and 2 sexes). To the best of our knowledge however, such existing cause-of-death mortality approaches have primarily focused on interpretation and inference, rather than necessarily aiming to improve out-of-sample forecasting performance.

In this paper, we propose a novel approach for forecasting cause-of-death mortality based on tensor decomposition techniques, which models the three dimensions (age, year, and cause) jointly and accounts for the dependencies between causes in a data-driven manner. A tensor simply refers to the generalization of a matrix to more than two dimensions, and so tensor decompositions can be viewed as extensions of well-known matrix decompositions e.g., the singular value decomposition, for performing dimension reduction on data with three or more dimensions (Kolda and Bader, 2009). Tensor decompositions have been used in mortality modeling before, although not for cause-of-death mortality specifically. Russolillo et al. (2011) used a rank-2 Tucker decomposition, which can be regarded as a natural extension of the Lee-Carter model, to analyze mortality across ten European countries. More recently, Dong et al. (2020) generalized this approach by modifying both the canonical polyadic decomposition (CPD) and the Tucker decomposition to allow for different factors across age, year and country/gender, demonstrating substantial improvements in out-of-sample forecasting performance both for individual populations and the aggregate population (across ten European countries and two genders) compared with using single-population mortality models. We also acknowledge the burgeoning and important literature on coherent mortality modeling for multiple populations, such as that of Hyndman et al. (2013) among others. However, as many of these models developed in this literature are not specifically designed for cause-of-death mortality modeling, being applied for studying both genders or sub-populations (say) based on a coherence assumption, then they are not suitable to the current scope of this manuscript. In particular, the coherence assumption requires the long term convergence of mortality rates of the different sub-populations which, in this setting, are the causes of death. But such an assumption is not biologically reasonable given the different evolving patterns of causes.

Motivated by the above works, we apply tensor decomposition techniques to forecast cause-of-death mortality rates. We examine the penalized tensor decomposition method of Madrid-Padilla and Scott (2017), which couples the tensor decomposition with penalty matrices to capture underlying smooth trends (and associated dependencies) inherent in cause-of-death mortality rates. However, while Madrid-Padilla and Scott (2017) focused primarily on one-factor penalized tensor decompositions for epidemiological and motion-capture data, here we are interested in the multiple factor setting commonly required for cause-of-death mortality modeling. To overcome the subsequent, heavy computational burden resulting from (tuning parameter selection in) this setting, we propose a novel adaptive penalized tensor decomposition (ADAPT) method, which employs adaptively weighted penalty matrices instead. Aside from producing a far more computationally efficient approach, the use of adaptively weighted penalty matrices allows for varying degrees of smoothness for each factor and dimension in a flexible manner, which in turns leads to stronger forecasting performance compared to conventional unpenalized techniques such as CPD. ADAPT can be coupled with a variety of prediction methods, and we consider three examples (random walk with drift, linear extrapolation, and smoothing) to extrapolate the estimated year factors for forecasting. We illustrate an application of ADAPT to US male cause-of-death mortality data, demonstrating that it can provide better out-of-sample forecasting performance compared to existing cause-of-death mortality methods such as multinomial logistic regression, cause-specific standard Lee-Carter models, and product-ratio model (for comparison purposes, despite the unreasonable coherence assumption associated with this method for cause-of-death modeling as discussed earlier), especially for mid-term and long-term forecasting. Furthermore, we illustrate how the results of ADAPT can be visualized to interpret the dominant patterns of variation across causes of death, age group, and year.

This rest of the paper is organized as follows. Section 2 presents the US cause-of-death mortality data used and details of the pre-processing procedures. Section 3 reviews the existing methodology on generalized lasso penalties and penalized tensor decompositions. Section 4 discusses the new ADAPT approach including hyperparameter selection and forecasting. Section 5 applies ADAPT to the US cause-of-death mortality data as an illustration, and compares its performance with some existing methods. Section 6 concludes and provides potential directions for future research.

2. Data

This paper focuses on cause-of-death mortality data in the United States male population from ages 0-85+ and years 1950-2007. The data are obtained from the World Health Organization (WHO), which maintains a comprehensive database for cause-of-death mortality and population for more than one hundred countries.

We followed a similar data pre-processing procedure as in Gaille and Sherris (2011), who used the same data source but not exact same dataset (we utilized a larger time period).

We first binned the data into 5-year age groups, as is commonly done in cause-of-death mortality forecasting to ease interpretation. However, considering that there is a large difference in the nature of infant mortality, we chose to further split the first 5-year age group into two parts, corresponding to before and after 1-year old. Our data thus consisted of 19 age groups: infants less than 1-year old, children aged 1-4 years old, and thereafter 5 year age-groups ending with the group aged 85+. Next, the numbers of deaths for unknown ages were distributed proportionally across the entire age range, as recommended by the Human Mortality Database (University

Table 1
Coding system used to group the International Classification of Diseases (ICD) into five primary causes of death. The ICD changed three times 1950 to 2007, which covers the years of interest in the motivating US male cause-of-death mortality data.

Cause of death	ICD7 (1950-1967)	ICD8 (1968-1978)	ICD9 (1979-1998)	ICD10 (1999-2007)
Circulatory system	A079-A086	A080-A088	B25-B30	I00-I99
Cancer	A044-A060	A045-4061	B08-B17	C00-D48
Respiratory system	A087-A097	A089-A096	B31-B32	J00-J99
External causes	A138-A150	A138-A150	B47-B56	V00-Y89
Infectious and parasitic diseases	A001-A043	A001-A044	B01-B07	A00-B99

of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), 2020). Since the proportion of these deaths are usually negligible i.e., less than 1% for each cause, then distributing proportionally is not expected to have a major impact on the subsequent analysis. As the WHO database only provides mid-year populations, we calculated mortality based on the central death rates.

Two adjustments were introduced in regards to the criterion of classification of cause of death. First, note that the WHO classified the causes according to the international classification diseases (ICD, World Health Organization, 2004), which has changed three times between 1950 and 2016 to account for advances in science and technology as well as a gradually more refined cause-of-death descriptions (Table 1). Each ICD is also made up of thousands of causes, although for the purposes of analysis one is generally not interested in the cause-of-death mortality of each individual cause. Therefore, we opted to classify the causes into five primary causes of death: circulatory diseases, cancer, respiratory system, external causes and infectious and parasitic diseases. Details about how we performed this classification are provided in Table 1, noting each ICD has its own coding system. From the historical data, we find that these five primary causes on average account for 80% of deaths over the past sixty years. We assigned the remaining 20% to be a sixth cause of death, and denote it as "Others".

The second adjustment we made was motivated by the aforementioned change in ICD over time. Specifically, as the ICD changed three times between 1950 and 2007 from ICD7 to ICD10, the raw data were not directly comparable over time. To overcome this, we followed Gaille and Sherris (2011) and introduced comparability ratios to eliminate the gap between two consecutive different ICDs. The comparability ratio was obtained by requiring that the average of death rates over the last two years of a classification is equal to the average of the death rates over the first two years of the newly adopted classification. For the purpose of prediction and interpretation, the comparability ratios were introduced in a backward direction i.e., we fixed the latest mortality rates and changed the past data. An example of comparability ratios and the resulting adjustment in provided Appendix A. More importantly, the main trends in cause specific mortality are still preserved after adjustment using comparability ratios. For the remainder of this article, we will let $m(i, j, t)$ denote the central mortality rates for cause i, age group j, at year t.

Fig. 1 presents plots of the log age-specific death rates i.e., $\log(m(i, j, t))$ for each of the six causes as a function of age group. Results showed that all six causes of death share a similar pattern in that as age increased, the log mortality rate decreased at first, reaching its minimal level around ages 5-14, before increasingly at different rates for different causes. This pattern is consistent with the intuition that infants and elderly are more vulnerable to mortality risk. We also note the sharp increase in mortality rates for External causes around age 14, after which the mortality rate remained approximately the same until around age 60. Critically, Fig. 1 suggests underlying common patterns among the causes and across years, and for the purposes of forecasting it may be beneficial to jointly model these causes and any potential dependencies between them.

Next, to better understand the trends of cause-specific mortality for each age group, we plotted the log mortality rates as a function of year, where each curve represented the mortality dynamics of a single age group over time. Based on these plots (Fig. 2), we observe that most of the age groups changed gradually over time, while the direction of the change varies with the age group. Note also the sudden fallen in Respiratory system mortality rates between 1970 and 1980 for infants: this is a result of many premature babies dying due to their lack the ability to breathe by themselves, while reliable mechanical ventilators since the 1970s have become readily available (Whelan and Buhler-Wilkerson, 2020). Additionally, we point out that the mode of infections and parasitic diseases for many ages in 1950s was likely caused by tuberculosis and acute poliomyelitis (e.g., see Quah, 2016). For tuberculosis, isoniaid opened the modern era of treatment in 1952, and the first vaccine for acute poliomyelitis was developed in the 1950s. This in turn is reflected in the sudden drop in mortality rates for infectious and parasitic diseases for many age groups post 1950. Finally, within each cause we observe similar patterns across many age groups. This again suggests underlying common smoothness in the changes of mortality rates, and motivates methods which recover and exploit such joint smoothness for the purposes of forecasting.

3. Review of existing methodology

In this section, we first review the concept of generalized lasso penalties, before introducing the penalized tensor decomposition proposed by Madrid-Padilla and Scott (2017) for encouraging smoothing and/or sparsity specifically in tensor decompositions.

3.1. Generalized lasso penalties

The generalized lasso penalty (Tibshirani et al., 2011; Arnold and Tibshirani, 2016; Ali and Tibshirani, 2019) is an extension of the well-known least absolute shrinkage and selection operator (lasso, Tibshirani, 1996) to a larger class of penalties that can induce both sparsity and/or differing levels of smoothness based on ℓ_1 regularization. The latter will be particularly useful for mortality forecasting to recover and exploit any underlying smooth trends in the data (see the discussion in the preceding section). Mathematically, the generalized lasso penalty can be formulated as a constrained maximization problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{GL}} = \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}), \quad \text{subject to } \|\mathbf{D}\boldsymbol{\beta}\|_1 \leqslant C, \tag{1}$$

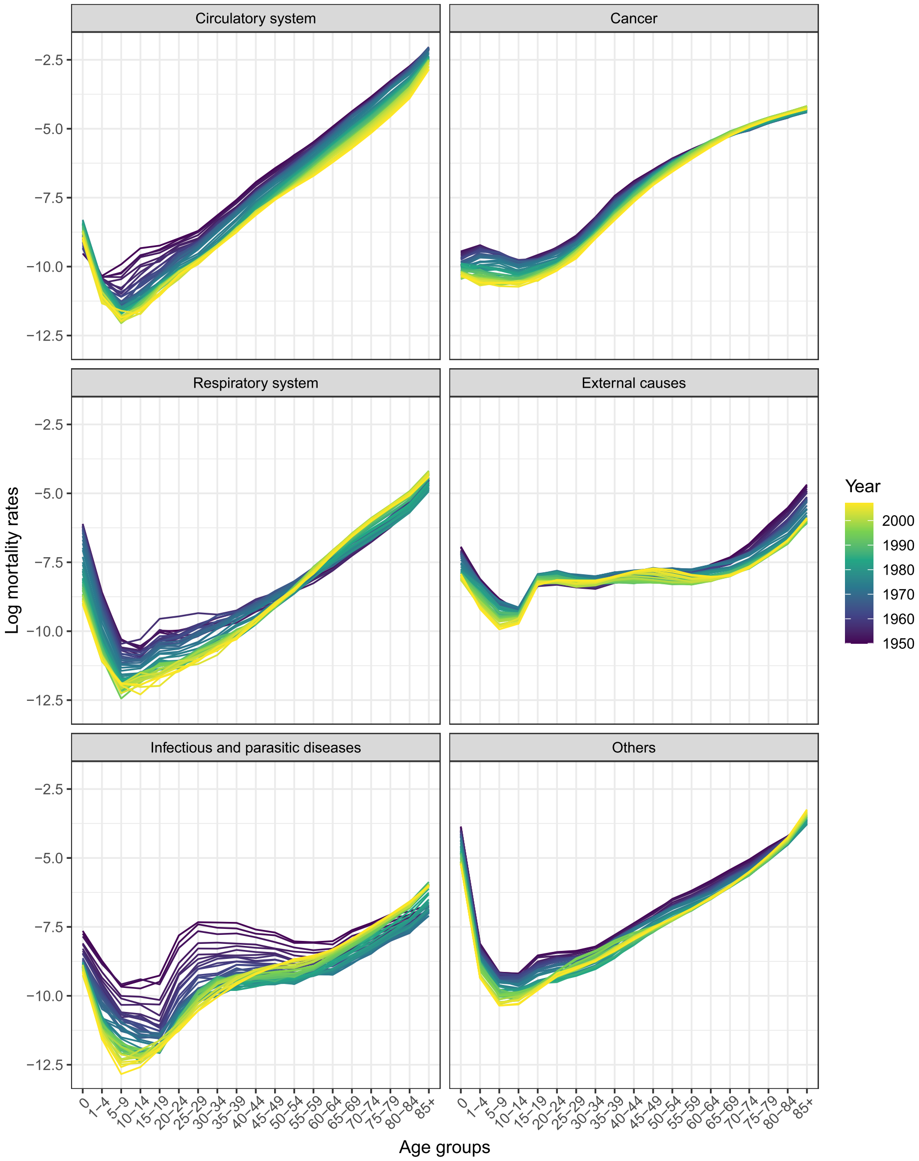


Fig. 1. US male log mortality rates as a function of age group across six causes of death, colored by years from 1950 to 2007. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

where $f(\boldsymbol{\beta})$ generically denotes an objective function involving a p-vector of parameters $\boldsymbol{\beta}$, $C > 0$ is a tuning parameter, $\mathbf{D}$ is the penalty matrix designed to achieve smoothness or sparsity in the structure of $\boldsymbol{\beta}$, and $\|\cdot\|_1$ denotes the ℓ_1 norm. An equivalent way of writing the generalized lasso problem is in the unconstrained form $\hat{\boldsymbol{\beta}}_{\mathrm{GL}} = \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \|\mathbf{D}\boldsymbol{\beta}\|_1$, where $\lambda > 0$ is a tuning parameter related to C.

In practice, there are many choices for the penalty matrix $\mathbf{D}$ depending on the level of smoothness or sparsity we want in the estimated $\hat{\boldsymbol{\beta}}_{\mathrm{GL}}$. For example, setting $\mathbf{D}$ to a (first-order) difference matrix leads to the fused lasso penalty of Tibshirani et al. (2005), which takes the form $\sum_{k=1}^{p-1} \left\|\beta_{k+1} - \beta_k\right\|_1$ and induces a piecewise constant structure. Also, if $\mathbf{D}$ is set an identity matrix then $\|\mathbf{D}\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|$, and the problem reduces back to the standard lasso penalty. We provide more details on the choice of $\mathbf{D}$ in Section 4.2.

3.2. Penalized tensor decomposition

The concept of penalized tensor decomposition applied in this paper is based on coupling the CPD with the aforementioned generalized lasso penalties (Madrid-Padilla and Scott, 2017). Consider the input tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times T}$, where the element $x_{ijt} = \log(m(i, j, t))$ is the log mortality rate observed for cause $i = 1, \dots, I$, age group $j = 1, \dots, J$, and year $t = 1, \dots, T$. A rank-R (unpenalized) CPD aims to decompose $\mathcal{X}$ as a sum of R rank-1 tensors based on solving the problem

$$\underset{\mathbf{u}_r \in \mathbb{R}^I, \mathbf{v}_r \in \mathbb{R}^J, \mathbf{w}_r \in \mathbb{R}^T}{\text{minimize}} \quad \left\| \mathcal{X} - \sum_{r=1}^{R} d_r \cdot \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \right\|_{\mathcal{T}},$$

where $d_r > 0$ is a scalar, the vectors $\mathbf{u}_r, \mathbf{v}_r, \mathbf{w}_r$ correspond to the estimated effects of cause, age group, and year dimension respectively, the operations $\circ$ and $\cdot$ denotes the outer product and standard multiplication, and $\|\cdot\|_{\mathcal{T}}$ denotes the tensor norm (a generalization of the

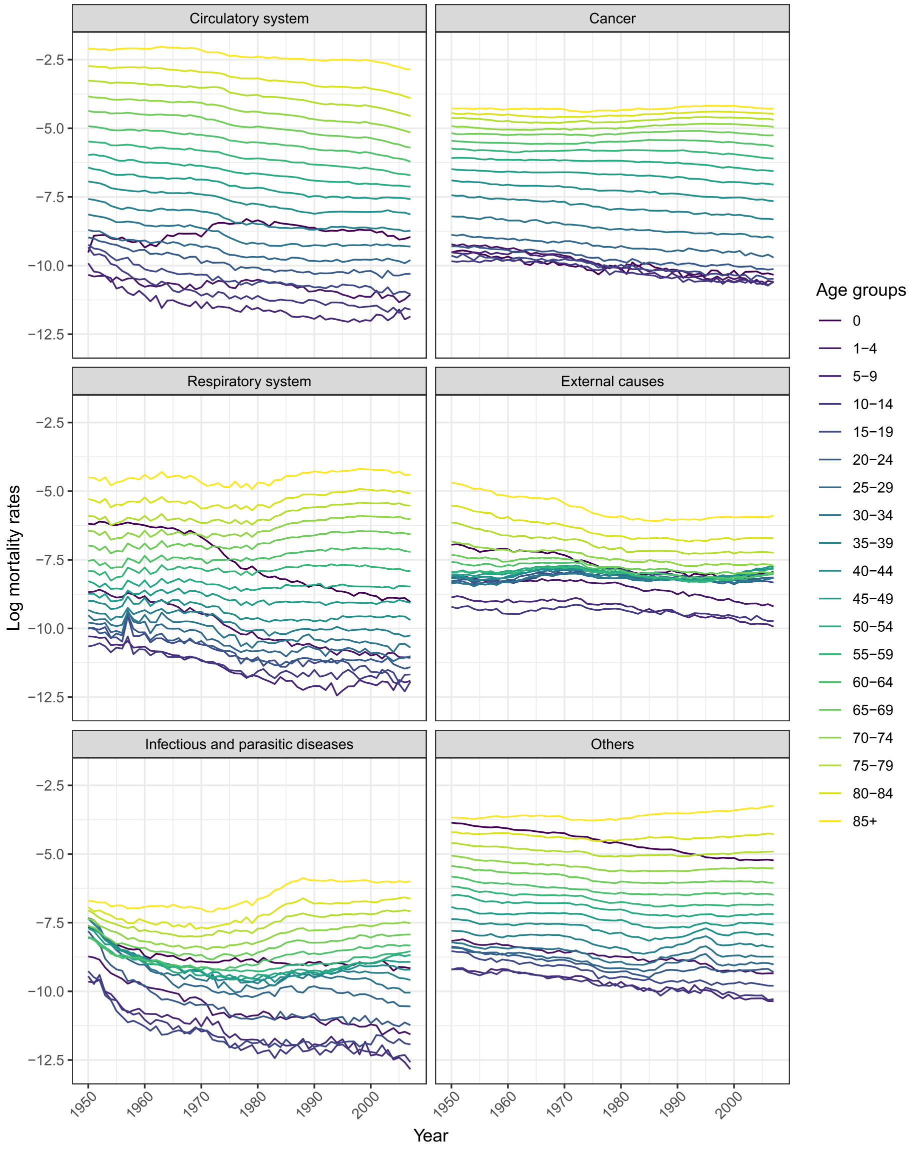


Fig. 2. US male log mortality rates as a function of year across six causes of death, colored by age group from age 0 to ages 85 and over.

Frobenius norm to more than two dimensions). Fig. 3 graphical illustrates this decomposition. Note that for identifiability reasons, all the vectors $\mathbf{u}_r, \mathbf{v}_r, \mathbf{w}_r$ are normalized to have unit length e.g. $\|\mathbf{u}_r\| = 1$ for all $r = 1, \ldots, R$, such that the $d_r > 0$ governs the scale.

After fitting the CPD, and indeed for other tensor decomposition methods including the ADAPT method proposed later on, we obtain the estimated log mortality rates via the decomposition $\log(\hat{m}(i, j, t)) = \sum_{r=1}^{R} \hat{d}_r \hat{u}_{r,i} \hat{v}_{r,j} \hat{w}_{r,t}$. When viewed this way, the CPD bears some similarity to, but is not a special case of, the general Lee-Carter model of Renshaw and Haberman (2003), which we also describe and use in application in Section 5.3 as a benchmark. In particular, the decomposition underlying a (general) Lee-Carter model is performed separately for each cause of death, while in the CPD and the proposed ADAPT method the decomposition occurs across all three dimensions of the tensor simultaneously e.g., the estimated effects of cause $u_{r,i}$'s are shared across all age groups and time. In doing so, tensor decomposition methods seek to leverage potential underlying similarities in covariation patterns between the different causes of death. We refer to Madrid-Padilla and Scott (2017) and references therein for a more detailed introduction to tensors and the CPD, along with connections to the singular value decomposition for matrices. It is important to highlight though that, in contrast to the eigenvalue or singular value decomposition for matrices, each rank-1 tensor component in the tensor decomposition has equal status, i.e., the d_r's are not ordered in any particular way At the same time, the CPD uses $R(I + J + T + 1)$ parameters to approximate the input tensor, which presents a substantial dimension reduction particularly if one or more of I, J and R is non-negligible.

Following Madrid-Padilla and Scott (2017), the penalized tensor decomposition augments the CPD with one or more generalized lasso penalties on the vectors $\mathbf{u}_r, \mathbf{v}_r$ and $\mathbf{w}_r$. That is,

$$\begin{aligned} \underset{\mathbf{u}_r \in \mathbb{R}^I, \mathbf{v}_r \in \mathbb{R}^J, \mathbf{w}_r \in \mathbb{R}^T}{\text{minimize}} \quad & \left\| \mathcal{X} - \sum_{r=1}^{R} d_r \cdot \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \right\|_{\mathcal{T}} + \sum_{r=1}^{R} (\lambda_u, r \left\| \mathbf{D}_r^u \mathbf{u}_r \right\|_1 \\ & + \lambda_v, r \left\| \mathbf{D}_r^v \mathbf{v}_r \right\|_1 + \lambda_w, r \left\| \mathbf{D}_r^w \mathbf{w}_r \right\|_1) \\ \text{subject to} \quad & \mathbf{u}_r^T \mathbf{u}_r \le 1, \quad \mathbf{v}_r^T \mathbf{v}_r \le 1, \quad \mathbf{w}_r^T \mathbf{w}_r \le 1, \quad r = 1,\ 2,\ \ldots\ R, \end{aligned} \tag{2}$$

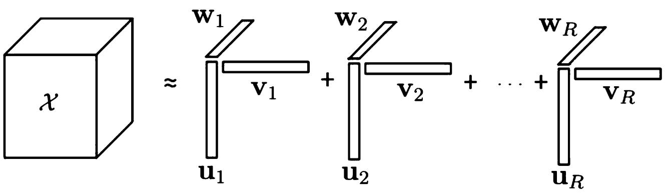


Fig. 3. CPD of a third-order Tensor, with the number of factors set to R.

where $\boldsymbol{\lambda}_u = (\lambda_{u,1}, \dots, \lambda_{u,R}), \boldsymbol{\lambda}_v = (\lambda_{v,1}, \dots, \lambda_{v,R})$ and $\boldsymbol{\lambda}_w = (\lambda_{w,1}, \dots, \lambda_{w,R})$ represent a total of $3R$ tuning parameters, $\mathbf{D}_r^u, \mathbf{D}_r^v$, and $\mathbf{D}_r^w$ are the corresponding penalty matrices. Note the unit norm constraints are now relaxed to convex constraints, as they are easier to optimize while not affecting the performance of the decomposition (Madrid-Padilla and Scott, 2017). In Section 4.2, we will discuss how for cause-of-death mortality modeling, we opt not to penalize the $\mathbf{u}_r$ vectors, since there is no reason *a-priori* to believe that the cause of death dimension exhibits any underlying smoothness or sparsity. However, for the current generality of presentation we will focus on the general formulation in (2).

To solve the penalized tensor decomposition in (2), we can employ block coordinate descent type algorithms (Friedman et al., 2007), where the basic idea is to optimize the objective function by successively fixing most parameters at their current iteration value and updating the remaining parameters(s). We provide details and a formal algorithm to solve the rank-1 CPD and penalized tensor decomposition in Appendix B. For the rank-R penalized tensor decomposition formulated above, we can apply block coordinate descent to estimate each rank-1 tensor in turn, cycling over $r = 1, \dots, R$. That is, we update each of the R rank-1 tensors based on holding the other $(R-1)$ tensors fixed. Algorithm 1 formalizes the details of this procedure. As starting values for $\mathbf{u}_r, \mathbf{v}_r$ and $\mathbf{w}_r$, we use the estimates obtained from a rank-R CPD. This approach has two advantages: 1) we will be using the estimates from the CPD to construct adaptively weighted penalty matrices later on anyway; 2) by starting from the CPD, the number of iterations required for both the penalized tensor decomposition and adaptive penalized tensor decomposition to converge is typically much less than, say, using random vectors as starting values.

Algorithm 1 Block coordinate descent method for a rank-R penalized tensor decomposition. Note if all the tuning parameters are set to zero, then it corresponds to a rank-R CPD.

```
Input: Given tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times T}$; initial values $\{(\mathbf{u}_r^{(0)}, \mathbf{v}_r^{(0)}, \mathbf{w}_r^{(0)}); r = 1, \dots, R\}$ e.g., obtained using a rank-$R$ CPD; tuning parameters $(\boldsymbol{\lambda}_u, \boldsymbol{\lambda}_v, \boldsymbol{\lambda}_w)$; penalty matrices
    $\{\mathbf{D}_r^u, \mathbf{D}_r^v, \mathbf{D}_r^w; r = 1, \dots, R\}$.
for each $r = 1, \dots, R$ do
    Obtain values $d_r^{(0)} = \mathcal{X} \times_1 \mathbf{u}_r^{(0)} \times_2 \mathbf{v}_r^{(0)} \times_3 \mathbf{w}_r^{(0)}$, where $\times_n$ denotes the mode-$n$ product operator; see Appendix B for details.
end for
Construct $\mathcal{Y} = \sum_{r=1}^{R} d_r^{(0)} \cdot \mathbf{u}_r^{(0)} \circ \mathbf{v}_r^{(0)} \circ \mathbf{w}_r^{(0)} \triangleq [\![\mathbf{d}^{(0)}; \mathbf{U}^{(0)}, \mathbf{V}^{(0)}\mathbf{W}^{(0)}]\!]$.
repeat  $k = 1, 2, 3 \dots$
    for each $r = 1 \dots, R$ do
        Set $\mathcal{Y} \Leftarrow \mathcal{Y} - \left( d_r^{(k-1)} \cdot \mathbf{u}_r^{(k-1)} \circ \mathbf{v}_r^{(k-1)} \circ \mathbf{w}_r^{(k-1)} \right)$.
        Calculate $\mathcal{Z} = \mathcal{X} - \mathcal{Y}$.
        Apply a rank-1 penalized tensor decomposition to tensor $\mathcal{Z}$, with initial values $(d_r^{(k-1)}, \mathbf{u}_r^{(k-1)}, \mathbf{v}_r^{(k-1)}, \mathbf{w}_r^{(k-1)})$ and the input tuning parameters and penalty matrices; see Appendix B for details.
        Store output as $(d_r^{(k)}, \mathbf{u}_r^{(k)}, \mathbf{v}_r^{(k)}, \mathbf{w}_r^{(k)})$, and set $\mathcal{Y} \Leftarrow \mathcal{Y} + \left( d_r^{(k)} \cdot \mathbf{u}_r^{(k)} \circ \mathbf{v}_r^{(k)} \circ \mathbf{w}_r^{(k)} \right)$.
    end for
until Convergence e.g., difference in $(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)})$ between successive iteration is less than some tolerance value $\epsilon > 0$.
Output: Estimates $\{(\hat{d}_r, \hat{\mathbf{u}}_r, \hat{\mathbf{v}}_r, \hat{\mathbf{w}}_r); r = 1, \dots, R\}$.
```

When applying the rank-R penalized tensor decomposition in (2), we need to specify both the penalty matrices and tuning parameters. In the former, it is often appropriate to assume the same penalty matrix form for a specific dimension, to ensure that the same type of smoothness or sparsity is adopted across all factors of the same dimension e.g., with the year dimension we set $\mathbf{D}^w = \mathbf{D}_1^w = \dots = \mathbf{D}_R^w$. Subsequently, it means we have to select the forms for (at most) three penalty matrices $\mathbf{D}^u, \mathbf{D}^v$ and $\mathbf{D}^w$, and we discuss how to choose this in more detail in Section 4.2.

By contrast, for any given dimension the tuning parameters are expected to differ greatly across the various factors to accommodate varying degrees of regularization. Consequently, it means we need to select the $3R$ tuning parameters i.e., the elements in each of the vectors $(\boldsymbol{\lambda}_u, \boldsymbol{\lambda}_v, \boldsymbol{\lambda}_w)$, individually, which poses an immense computational burden. For instance, suppose we wish to select the best combination of tuning parameters for a rank-R penalized tensor decomposition applied to the motivating US male cause-of-death mortality data, based on a grid-search with 20 candidate values for tuning parameter. Then, even if only age group and year were penalized we would still need to apply Algorithm (1) on a total combination of 20^{2R} possible sets of tuning parameter values. For $R \geq 3$, this rapidly becomes computationally infeasible, and motivates the development of an alternative approach to penalized tensor decomposition in the multi-factor setting.

4. Adaptive penalized tensor decomposition

To overcome the computational challenge of penalized tensor decomposition with multiple factors, we propose a new *ADAptive Penalized Tensor decomposition (ADAPT)* method, inspired by the idea of adaptive lasso regression (e.g., Zou, 2006; Hui et al., 2020). The idea is to replace the standard penalty matrices with a set of pre-specified weighted penalty matrices instead, where the weights automatically allow for varying levels of smoothness and regularization across both the dimensions and factors. In doing so, ADAPT allows the same tuning parameter to used for all factors of a given dimension. Mathematically, the rank-R ADAPT is formulated as

Algorithm 2 Constructing adaptively weighted penalty matrices for rank-R ADAPT.

```
Input: Estimates from a rank-$R$ CPD, denoted as $\{(\tilde{d}_r, \tilde{\mathbf{u}}_r, \tilde{\mathbf{v}}_r, \tilde{\mathbf{w}}_r); r = 1, 2, \dots, R\}$; penalty matrices $\{\mathbf{D}^u, \ \mathbf{D}^v, \ \mathbf{D}^w\}$, which are assumed to be the same across of factors of the same dimension.
1: Calculate $\tilde{\boldsymbol{\tau}}_r^u = \mathbf{D}^u \tilde{\mathbf{u}}_r$, $\tilde{\boldsymbol{\tau}}_r^v = \mathbf{D}^v \tilde{\mathbf{v}}_r$, $\tilde{\boldsymbol{\tau}}_r^w = \mathbf{D}^w \tilde{\mathbf{w}}_r$. Let $\tilde{\tau}_{rk}^u$ denote as the $k^{th}$ element in $\tilde{\boldsymbol{\tau}}_r^u$, and similarly for $\tilde{\tau}_{rk}^v$ and $\tilde{\tau}_{rk}^w$.
2: Construct $\tilde{\mathbf{D}}_r$ element-wise as $[\tilde{\mathbf{D}}_r^u]_{kl} = [\mathbf{D}^u]_{kl} / |\tilde{\tau}_{rk}^u|$. That is, we divide all elements in the $k^{th}$ row of $\mathbf{D}^u$ by $|\tilde{\tau}_{rk}^u|$. Similarly, we can construct $\tilde{\mathbf{D}}_r^v$ and $\tilde{\mathbf{D}}_r^w$ for $r = 1, 2, \dots, R$ in an analogous manner.
Output: Adaptively weighted penalty matrices $\tilde{\mathbf{D}_r^u}, \tilde{\mathbf{D}_r^v}, \tilde{\mathbf{D}_r^w}; r = 1, \dots, R$.
```

$$
\begin{aligned}
\underset{\mathbf{u}_r \in \mathbb{R}^I, \mathbf{v}_r \in \mathbb{R}^J, \mathbf{w}_r \in \mathbb{R}^T}{\text{minimize}} \quad & \left\| \mathcal{X} - \sum_{r=1}^{R} d_r \cdot \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \right\|_{\mathcal{T}} + \sum_{r=1}^{R} (\lambda_u \left\| \tilde{\mathbf{D}}_r^u \mathbf{u}_r \right\|_1 \\
& + \lambda_v \left\| \tilde{\mathbf{D}}_r^v \mathbf{v}_r \right\|_1 + \lambda_w \left\| \tilde{\mathbf{D}}_r^w \mathbf{w}_r \right\|_1) \\
\text{subject to} \quad & \mathbf{u}_r^T \mathbf{u}_r \le 1, \quad \mathbf{v}_r^T \mathbf{v}_r \le 1, \quad \mathbf{w}_r^T \mathbf{w}_r \le 1, \quad r = 1, \ 2, \ \dots \ R,
\end{aligned}
\tag{3}
$$

where compared to the standard penalized tensor decomposition in (2), ADAPT uses at almost three tuning parameters $(\lambda_u, \lambda_v, \lambda_w)$. Put another way, the number of tuning parameters in ADAPT is only determined by the number of dimensions to be penalized, and no longer depends on the rank. Returning to the example discussed at the end of Section 3.2, the grid search now involves a total combination of $20^2 = 400$ candidate tuning parameters, presenting a substantial computation reduction compared to CPD and thus making ADAPT feasible as a procedure for cause-of-death mortality modeling.

Regarding the form of the adaptively weighted penalty matrices $\tilde{\mathbf{D}}_r^u, \tilde{\mathbf{D}}_r^v, \tilde{\mathbf{D}}_r^w$ in (3), note that the adaptive weights often take the form of the inverse absolute value of the estimates from an unpenalized fit (e.g., Zou, 2006; Hui et al., 2017). Such a form induces a relatively heavier degree of penalization for parameters expected to be close or equal to zero (as based on how close the estimates of these parameters from the unpenalized fit were to zero). We adopt a similar approach here, and construct the adaptively weighted penalty matrices based on the (unpenalized) CPD fit. Algorithm 2 outlines their construction in detail. In particular, the form $|\tilde{\tau}_{rk}|^{-1}$ in step 2 the algorithm serves as an added weight to differentially penalize parameters in the tensor decomposition. A larger weight leads to more regularization for a particular factor and dimension of a tensor, and consequently induces more sparsity or smoothness. It is also important to emphasize that computing the adaptive weighted penalty matrices involves minimal computational cost, since as discussed previously, we use the estimates from the CPD as initial values in the penalized tensor decomposition and ADAPT.

After constructing the adaptive weighted penalty matrices, we can employ the same block coordinate algorithm as presented in Algorithm 1 to solve for the estimates for ADAPT in (3). Indeed, if we compare equations (2) and (3), we observe that the two differ only in terms of the penalty matrices, with the latter involving at most three tuning parameters. Also, note that even though we assume the same type of penalty matrix for all factors of a given dimension e.g., $\mathbf{D}^w = \mathbf{D}_1^w = \ldots = \mathbf{D}_R^w$ as discussed in the preceding section, the adaptively weighted penalty matrices in ADAPT will no longer be exactly the same across a given dimension. For instance, while the $\tilde{\boldsymbol{D}}_r^w$'s for $r = 1, \dots, R$ will share the same structure in terms of what elements are zero or not, they will differ in the actual values of the non-zero elements.

4.1. Forecasting

To forecast cause-of-death mortality rates using a tensor decomposition such as CPD or ADAPT, we use the estimated vectors for cause $\hat{\mathbf{u}}_r$ and age group $\hat{\mathbf{v}}_r$, and extrapolate the values of the year vector $\hat{\mathbf{w}}_r$. Afterwards, we can construct a predicted tensor as $\hat{\mathcal{X}}_{pred} = \sum_{r=1}^{R} \hat{d}_r \cdot \hat{\mathbf{u}}_r \circ \hat{\mathbf{v}}_r \circ \hat{\mathbf{w}}_{r,\text{pred}}$, where $\hat{\mathbf{w}}_{r,\text{pred}}$ denotes the extrapolated year vector for factor r whose length is equal to the number of years ahead to forecast.

As an approach, tensor decomposition methods do not themselves offer a method of prediction. That is, analogous to how a Lee-Carter model uses an ARIMA model (or some variation thereof) to forecast mortality, the tensor decomposition needs to be coupled with a method of extrapolating the time vectors in order to forecast cause-of-death mortality rates. While there are a number of approaches to accomplishing this, here we discuss three univariate approaches based on predicting the year vectors separately for each factor in the decomposition. In recent work, Dong et al. (2020) showed that fitting a multivariate model across all year vectors e.g., a vector ARMA model, typically does not produce any improvement in terms of forecasting performance, and can sometimes be worse than using R separate univariate models due to potential instabilities in their estimation. Also, note that the estimated year vectors when using ADAPT are, by construction, generally quite smooth. That is, plots of $\hat{\mathbf{w}}_r$ against year tend to resemble an "error-free" smooth curve (see for example the right column of Fig. 4 later on). As a result, the use of time-series techniques tends to be of secondary importance for forecasting here, compared to using a reliable extrapolation technique for the (mean) curve itself. With this mind, we consider three univariate prediction approaches for $\mathbf{w}_r$.

1. **Random walk with drift:** Also known as an ARIMA(0,1,0) model plus a constant, such models are commonly used for forecasting time-series data (see also Dong et al., 2020). For element t in $\hat{\mathbf{w}}_{r,\text{pred}}$, which we denote here as $\hat{w}_{r,t,\text{pred}}$, the random walk with drift model results in the prediction

$$
\hat{w}_{r,\text{pred},t} = \hat{w}_{r,T} + t \cdot \hat{\mu}_r; \quad t = 1, 2, \dots; \ r = 1, 2, \dots, R
$$

where $\hat{w}_{r,T}$ denotes the last element in the estimated year vector for factor, r and $\hat{\mu}_r$ is the estimated constant term reflecting the average year-to-year change.

2. **Linear Extrapolation:** This simple non-model-based approach is often used when extrapolations not far outside from the given data range are sought. For extrapolating the year vector in particular, it results in the prediction

$$\hat{w}_{r,\text{pred},t} = \hat{w}_{r,T} + t \cdot (\hat{w}_{r,T} - \hat{w}_{r,T-1}) \quad \text{for } t = 1,2,\ldots;\ r = 1,2,\ldots,R,$$

from which we see that linear interpolation relies on the two most recent estimated values of the time vector. It is also possible to construct higher-order polynomial extrapolations, however we choose not to explore these alternatives given the third prediction method below.

3. **Smoothing:** This approach is based on regressing $\hat{\mathbf{w}}_r$ against a smooth function of time i.e., for the factor $r = 1, \ldots, R$, we fit the model $\hat{w}_{r,t} = \beta_{0r} + f_r(t) + \epsilon_{r,t}; t = 1, \ldots, T$ where β_{0r} is an intercept term, $f_r(\cdot)$ is a univariate smoother, and $\epsilon_{r,t}$ is an additive error term which we assume to follow a normal distribution with mean zero and a common error variance. For $f_r(\cdot)$, we follow Wood (2017) and use a thin plate regression spline, although other choices such as P-splines and B-splines (e.g., Currie et al., 2004) could be used. Unlike the linear extrapolation method above, the smoothing approach uses the entire year vector to estimate the smoothing function and intercept. Subsequently, predictions can be constructed as

$$\hat{w}_{r,\text{pred},t} = \hat{\beta}_{0r} + \hat{f}_r(t) \quad t = 1,2,\ldots;\ r = 1,2,\ldots,R.$$

In the above, we have focused on point predictions. However, it is also possible to construct probabilistic forecasts of cause of death mortality using any of the above approaches. Specifically, we first construct probabilistic forecasts of the estimated time vectors using one of the above techniques e.g., the random walk with drift or its generalization to any ARIMA model. Afterward, we then use these probabilistic forecasts of $\hat{\mathbf{w}}_{r,\text{pred}}$ to reconstruct the entire predicted tensor with uncertainty. Such an approach to is analogous to the methods used for constructing probabilistic forecasts for Lee-Carter type models; this is not surprising given, as discussed in Section 3.2 ADAPT shares similarities with the general Lee-Carter model of Renshaw and Haberman (2003) but applied to all three dimensions of the tensor simultaneously. However, to further account for parameter and/or model uncertainty, future work could examine approaches for constructing prediction intervals such as bootstrapping or jackknifing (e.g., Wang et al., 2019).

It is important to note that the forecasting approach and the proposed ADAPT model can be chosen separately. For the latter, we examine three approaches in the form of linear extrapolation, random walk with drift, and smoothing via generalized additive models. Each of these forecasting techniques involve some level of tuning e.g., with linear extrapolation, the user needs to select which observations (in this case estimated values in each of the time vectors) are used to construct the extrapolation, with a default of using the observations from the most recent two time points as we do in this paper. This need for tuning is well-known and has arisen in the literature on stochastic mortality models, following the introduction of the Lee-Carter model in 1992. The issue is often referred to as "jump off error". In discussing this point, the recent review paper of Basellini et al. (2022) notes that "in order to avoid reproducing in the forecast any idiosyncrasies in the observed data, jump-off rates may be smoothed either independently or as part of the estimation procedure". This tuning can have a noticeable effect on the forecasts. In practice, practitioners should always choose and tune the forecasting methods according to appropriate background knowledge and sensitivity analysis.

To conclude this section, we highlight (once more) that practitioners have the flexibility to couple whichever prediction technique they like with ADAPT (and other tensor decomposition methods). The above three techniques are by no means the only choices, and future research could investigate a wide variety of other approaches.

4.2. Hyperparameter selection

To use ADAPT, there are a number of choices we need to make regarding the form of the penalty matrices, the number of factors or rank R, and the tuning parameters. Collectively, we will refer to these as the hyperparameters in ADAPT, and discuss each in detail below.

1. **Penalty Matrices:** As mentioned previously below (1), the choice of $\mathbf{D}^u, \mathbf{D}^v$ and $\mathbf{D}^w$ depends on the level of sparsity or smoothness we want to induce in the corresponding estimated vectors $\hat{\mathbf{u}}_r, \hat{\mathbf{v}}_r$ and $\hat{\mathbf{w}}_r$. That is, the chosen penalty matrix should reflect any *a-priori* information or expectation we have about the corresponding dimensions of the data. For cause-of-death mortality specifically, there is no *a-priori* reason to expect that the effects of specific causes of death should be exactly zero or be necessarily smooth (although we acknowledge this may be a function of the pre-processing steps used in constructing the causes of death for analysis). Therefore, we choose not to impose any penalty on this dimension and set $\mathbf{D}^u = \mathbf{0}$ (or equivalently, set $\lambda_u = 0$). By contrast, for both the age group and year dimensions, we expect there to be underlying smooth trends (see Figs. 1 and 2), which we aim to exploit for the purposes of prediction. With this in mind, we consider a small set of related choices for $\mathbf{D}^v$ and $\mathbf{D}^w$. The first of these is the fused lasso (Tibshirani et al., 2005), which encourages the estimated vectors to have a piecewise constant structure. Mathematically, the fused lasso is equivalent to setting the penalty matrix in the generalized lasso penalty in (1) to take the form of a first order difference matrix

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

The fused lasso sits within a larger class of so-called trend filtering penalties (see Kim et al., 2009; Tibshirani et al., 2011, among others), such that a trend filtering penalty of order k regularizes the $(k+1)^{th}$ derivative, and is similar to enforcing the resulting smooth factor to be $(k+1)$ times differentiable. The fused lasso penalty introduced above is equivalent to a trend filtering penalty of order 0, while an order 1 penalty encourages a piecewise linear structure, an order 2 penalty encourages a piecewise quadratic structure, and so on. We refer the reader to Tibshirani et al. (2011) for more details on constructing $\boldsymbol{D}$ for these trend filtering penalties, noting their connection to penalty matrices used in, say, the P-splines method of Currie et al. (2004). Following the recommendation of Tibshirani et al. (2011), for both the age group and year dimensions we choose their penalty matrices to come from the set of trend filtering penalties of order $k = 0, 1, 2$. This means that are a total of nine possible combinations of $\mathbf{D}^v$ and $\mathbf{D}^w$ to consider.

Table 2
The optimal set of hyperparameters chosen (excluding the tuning parameters (λ_v, λ_w)), along with the best prediction method, for ADAPT under each of the three forecast horizons. Order refers to the order of the trend filtering penalty matrix chosen.

	Order for $(\mathbf{D}^v, \mathbf{D}^w)$	Number of factors (R)	Forecasting method
Short-term (5 years)	(1,1)	11	Smoothing
Mid-term (10 years)	(2,1)	12	Linear Extrap.
Long-term (15 years)	(2,1)	11	Linear Extrap.

2. **Tuning Parameters:** With only the age group and year dimensions having non-zero penalties, we need to select two tuning parameters (λ_v, λ_w). For both, we considered a sequence of 20 tuning parameters evenly spaced on the log scale from 10^{-6} to 2×10^{-2}, resulting in a grid of 400 candidate tuning parameters in total. The range of the sequence of tuning parameters was determined based on preliminary testing i.e., the smallest value effectively led to no penalization and returned the same results as CPD, while the largest value effectively produced a constant vector for most of the estimated $\hat{\mathbf{v}}_r$'s and $\hat{\mathbf{w}}_r$'s.
3. **Number of factors:** For cause-of-death mortality modeling, we can expect a non-negligible number of factors is required for strong forecasting performance. Therefore, we considered values of the rank R ranging from 3 to 14 in our application to the US male cause-of-death mortality data. We found that ADAPT with ranks 1 and 2 tended to performed poorly in general, while predictive performance plateaued around 14 factors and little was gained for $R > 14$. We also emphasize that these choices are specific to our motivating US male mortality data, and practitioners may consider a broader or narrower range as appropriate for their analysis at hand; see also our discussion regarding model interpretation in Section 5.2.

To summarize, we have nine combinations to choose for the forms of the penalty matrices $\mathbf{D}_v, \mathbf{D}_w$, a grid of 400 candidate tuning parameters for (λ_v, λ_w), and twelve possibles values of the number of factors R. To select all of these hyperparameters, along with which prediction method to use with ADAPT (see Section 4.1), we apply five-fold rolling origin cross-validation (Liu and Yang, 2022), which takes into account the temporal nature of the data. In detail, we first hold out a sample of data that for testing and comparison with other methods later on e.g., the last 5, 10, or 15 years, depending on the forecasting horizon (see our application in Section 5.3 later on). Next, we split the remaining data into a training and validation set, where the validation set consists of a subset of data containing the same number of years as the test set, and the training set consists of data occurring temporally before the validation set. We then fit ADAPT to the training set, and predict using one of the approaches in Section 4.1 to the validation set. The splitting procedure is then repeated by rolling the validation set forward, one year at a time, and we repeat this five times i.e., the forecasting origin is successively updated. Note the size of training set increases with each fold, as more years are included. A schematic of the general K-fold rolling origin cross-validation approach is provided in Appendix C, and we refer to Liu and Yang (2022) and references therein for more details of the approach as a whole. As a metric for assessing performance, we use the tensor norm between the validation set tensor and the predicted tensor, $\|\mathcal{X}_{\text{validation}} - \hat{\mathcal{X}}_{\text{pred}}\|_{\mathcal{T}}$. We then select the set of hyperparameters and prediction method based on minimizing the mean tensor norm, averaged across the five folds.

5. Application to US cause-of-death mortality data

We illustrate an application of ADAPT to the motivating US male cause-of-death mortality data. We first present results from hyperparameter selection for ADAPT based on Section 4.2. Afterward, we assess the performance of ADAPT, presenting examples of how to visualize and interpret the results, and comparing it against some other available methods for cause-of-death mortality modeling. We emphasize that our empirical analysis of the US male cause-of-death mortality data should be viewed as an illustration of how the proposed ADAPT procedure can be employed in practice. We leave a more in-depth application to this data as an avenue of future research.

We evaluated performance under three different forecasting horizons, namely short-term (5 years), mid-term (10 years), and long term (15 years), where all validation and test sets in the rolling origin cross-validation were set to be one of these lengths. Under the short-term forecasting horizon, the test set consisted of years 2003-2007, while the training and validations sets were formed from years 1950-2002. Under the mid-term forecasting horizon, the test set consisted of years 1998-2007, while the training and validations sets were formed from years 1950-1997. Finally, under the long-term forecasting horizon, the test set consisted of years 1993-2007, while the training and validations sets were formed from years 1950-1992.

5.1. Results for hyperparameter selection in ADAPT

Based on the five-fold rolling origin cross-validation approach, Table 2 summarizes the optimal set of hyperparameters chosen (excluding the tuning parameters (λ_v, λ_w)), along with the best prediction method, when applying ADAPT to each of the three forecast horizons; see Appendix D.1 for the complete results. The order of the trend filtering penalty for the age group dimension was selected to be two for both mid-term and long-term forecasting horizons, but was selected to be one for short-term forecasting. This suggests that a smoother structure was preferred for longer term forecasting using ADAPT. On the other hand, for the year dimension an order one trend filtering matrix was chosen for all three forecasting horizons. Turning to the choice of R, cross-validation selected a relatively large number of factors for all forecasting horizons. That is, while smooth underlying trends were observed in the overall data, it suggests that for prediction purposes it is necessary to include a larger of factors to ensure (say) the finer scale trends were successfully captured. Furthermore, these choices for R also offered further justification for our development of the ADAPT method: the number of tuning parameters that would need to be selected when $R = 11$ or 12 using the standard penalized tensor decomposition approach in Section 3.2 would have been computationally infeasible. Finally, we see that for all three forecasting horizon periods, a non-time-series based prediction method was chosen.

In Appendix D.2, we present some examples of the estimated vectors for ADAPT versus the (unpenalized) CPD in the case of the short-term forecasting horizon; see also Appendix E for a link to the full set of estimated vectors. As expected, there was relatively little difference in the estimated vectors between ADAPT and CPD for the cause dimension. On the other hand, for both the age group and year dimensions, ADAPT produced much smoother estimated vectors compared to CPD.

5.2. Model interpretation

Although we focus on and select the hyperparameters for ADAPT based on out-of-sample predictive performance, it is important to recognize that model interpretability is a key factor in many practical applications of cause-of-death mortality models (Cairns et al., 2009). With heavily data-driven approaches such as tensor decompositions and ADAPT though, it is challenging to interpret all the factors estimated, especially when the chosen of R is large as is the case in Table 2. In the paragraphs below, we offer one approach to interpreting the results obtained from applying ADAPT, based on examining the estimated factor/s which explain the most variation.

For each of the three forecasting horizons, the factor with the largest corresponding value of $\hat{d}_r$ (we will refer to this as simply the first factor) accounted for approximately 85% of the total variation explained by the chosen ADAPT fit. For example, in Table 2 we see that while for the mid-term forecasting horizon rolling origin cross-validation selected a rank of $R = 12$, of the total variation explained by these 12 factors the first factor accounted for 84.6%. With this in mind, we decided to visualize and interpret the first estimated factor in these of these fits i.e., we plotted the $\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{v}}_1$ and $\hat{\boldsymbol{w}}_1$'s. Fig. 4 presents the results of this, from which we see that across all three forecasting horizons, the highest cause of mortality was Circulatory System, followed by Cancer and Others. The estimated factors for age groups generally exhibited an increasing pattern over the adult ages and a decreasing pattern over the younger ages, which is similar both to what was observed in Fig. 1, and to the aggregate mortality age pattern as observed in the original work of Lee and Carter (1992), among others. For the year dimension, in the short-term mortality generally increased over time for the first factor. By contrast, in the mid-term and long-term fits mortality decreased over the years. Note also a sharp decreasing pattern of mortality around the year 1960 for the mid-term forecasting horizon: as explained in Section 2, this could be caused by the modern treatment of tuberculosis and acute poliomyelitis, which led to a sudden drop in mortality rates for infectious and parasitic diseases. On the other hand, relative to the cause and age group dimension, the values of the estimated first factor for year varied much less across its domain, and so we caution against over-interpreting this result.

More generally, Fig. 4 encapsulates how, in the proposed ADAPT method as well as tensor decompositions more generally, the dependencies across different causes are taken into account via shared common factors. That is, by sharing the (estimated) vectors for age group and year, ADAPT aims to improve the predictive accuracy of the tensor (mortality data) as a whole by borrowing strength across different causes. Penalties which encourage either smoothness or sparsity in these estimated vectors (in a computationally efficient manner) further helps to reflect knowledge of and exploit any smoothly varying trends shared across different causes of mortality in both age group and time. At the same time, it is important to acknowledge that, as a purely data-driven approach, there are limits to the proposed ADAPT method for uncovering biological reasonableness (in contrast to say parametric and less flexible methods e.g., Alai et al., 2015). We view this as typical of the trade-off which has to be made between prediction performance and model interpretability, and it means practitioners should be careful when attempting to attribute biological/epidemiological interpretations to results from a tensor decomposition fit.

Finally, as an alternative approach to the above, especially for practitioners who prioritize interpretability over forecasting performance, we can consider employing ADAPT with a deliberately smaller number of factors. For example, rather than choosing R based on a predictive performance criterion, we can instead select the minimum number of factors for which the explained variation (as quantified based on Frobenius norm i.e., the tensor version of R-squared commonly used in linear regression) exceeds 95%. Such an approach is analogous to what was done in the original work of Lee and Carter (1992). In Appendix D.3, we present results based on applying this rule for choosing R. In particular, for all three forecasting horizons this leads to ADAPT selecting three factors for the tensor decomposition. This smaller number of factors (also) makes visualization, interpretation, and potential correlation structure analysis more tenable, and we offer interpretations of the results in Appendix D.3.

5.3. Comparison with other methods

We compared ADAPT with three available methods for cause-of-death mortality modeling, which we discuss in more detail below. Note that, although other methods aside from these have been proposed in the literature (as reviewed in Section 1), many do not have publicly available software and/or were not straightforward to implement, and so we were unable to include them for comparison. This includes the approach of Gaille and Sherris (2011), which uses the same data source as we do but applies the smoothing spline methods of Currie et al. (2004) in a non-cause-of-death mortality context.

- **Cause-specific Lee-Carter models:** One currently available approach is to fit separate Lee-Carter models (Lee and Carter, 1992) to each cause of death (recall there are $I = 6$ total causes in the US male mortality data), for which we consider two versions of. First, the original Lee-Carter model is defined as $\log(m(i, j, t)) = a_{i,j} + b_{i,j}k_{i,t} + \epsilon_{i,j,t}$, where for the i-th cause, $a_{i,j}$ represents the average mortality pattern for age group j, $k_{i,t}$ is a time-varying parameter reflecting the overall mortality trend, $b_{i,j}$ represents the sensitivity of log mortality at age group j to changes in $k_{i,t}$, and $\epsilon_{i,j,t}$ is an additive error term specific to each age group and year. The constraints $\sum_{t=1}^{T} k_{i,t} = 0$ and $\sum_{j=1}^{J} b_{i,j} = 1$ are imposed on the model to ensure that all parameters are identifiable. Estimation of the original Lee-Carter model is based on applying a singular value decomposition and then taking the vectors corresponding to the first singular value, again emphasizing that a separate model is fitted to each of the causes. For each cause, mortality rates forecasts are then constructed by fitting an ARIMA model (in most cases, a random walk with drift similar to that discussed in Section 4.1) to the $k_{i,t}$'s, and then forecasting this.

  While the above model is prevalent in mortality modeling, modifications can be made to enhance its predictive performance, and so we considered a second general version of the Lee-Carter model involving R factors $\log(m(i, j, t)) = a_{i,j} + \sum_{r=1}^{R} b_{r,i,j}k_{r,i,t} + \epsilon_{i,j,t}$ (see Renshaw and Haberman, 2003). This general Lee-Carter model is again estimated using a singular value decomposition, then extracting the relevant vectors and fitting and forecasting time-series models for each of the $k_{r,i,t}$'s. Again, note a model is fitted

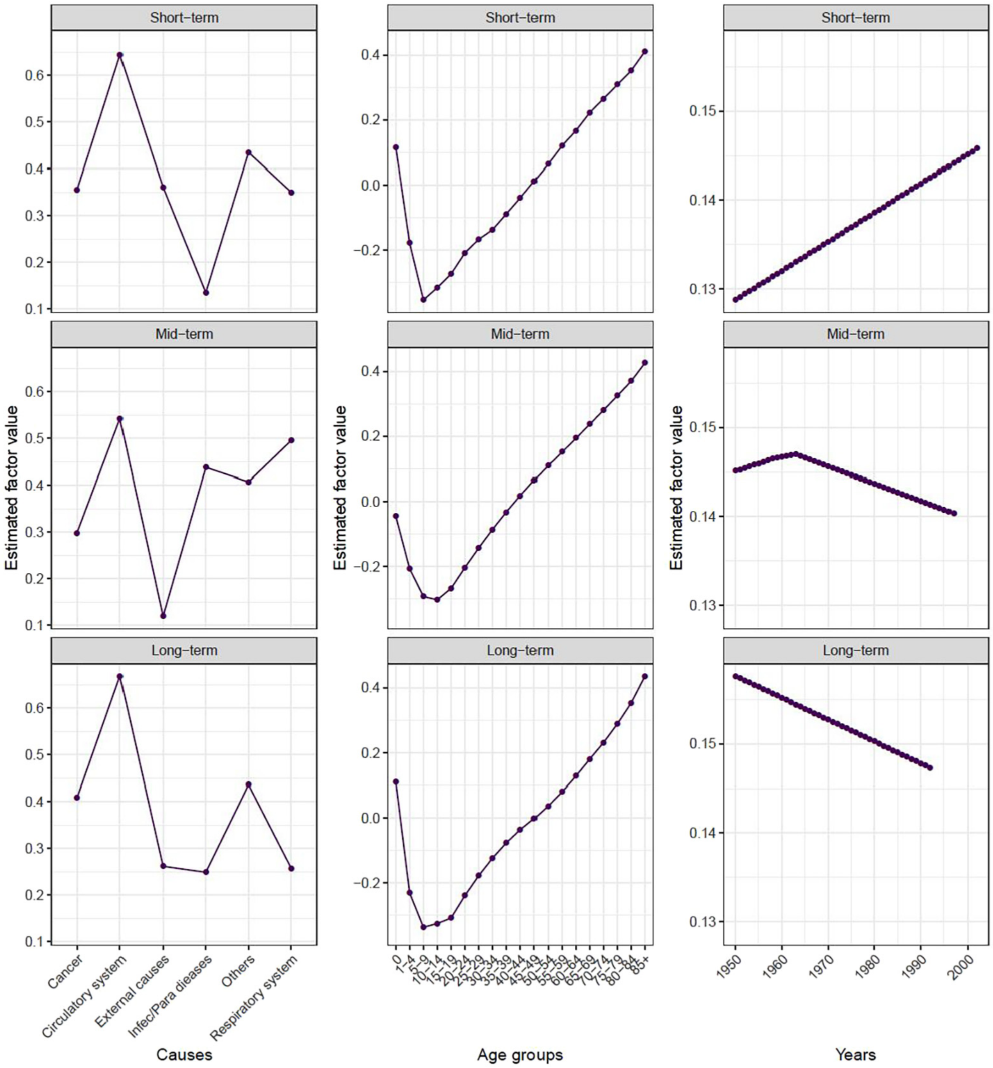


Fig. 4. Estimated vectors for cause (left), age group (middle), and year (right) for the first factor of the ADAPT fits (that is, the one with the largest corresponding value of $\hat{d}_r$) under each of the three forecasting horizons (top: short-term; middle: mid-term; bottom: long-term).

separately to each cause. To choose R in the general Lee-Carter model, we used the same five-fold rolling origin cross-validation method as described in Section 4.2. Note that since, for each cause, there are $J = 19$ age groups, then we chose R for the general Lee-Carter model in the integer range from 1 to 19, where $R = 1$ corresponds to the original Lee-Carter model. Based on this approach, we found that, perhaps not surprisingly and given the results seen for ADAPT in the previous section, a non-negligible number of factors was chosen for the general Lee-Carter model across all forecasting horizons: $R = 17$ factors for both short-term and mid-term, and $R = 8$ for long term. On the other hand, since this approach models each cause separately, then it involves many more parameters compared to the ADAPT approach, which models all causes together.

- **Multinomial logistic regression:** We examine the multinomial logistic regression (MLR) approach developed by Alai et al. (2015). Focusing on the US male mortality data, MLR assumes that for an age group j and year t, the vector containing the number of deaths attributable to cause augmented with the number of survivors follows a multinomial distribution with probabilities $(q_1(j,t), \ldots, q_6(j,t), p(j,t))$, where $q_i(j,t)$ denotes the probability of death due to cause i and age group j and year t, and $p(j,t)$ represents the probability of surviving that year. The probabilities in the MLR are then modeled as $\log(q_i(j,t)/p(j,t)) = \boldsymbol{z}_{j,t}^{\top}\boldsymbol{\beta}_i$, $i = 1, \ldots, 6$, where $\boldsymbol{z}_{j,t}$ denotes the vector of explanatory variables and $\boldsymbol{\beta}_k$ denotes the corresponding regression coefficients for cause i. After fitting using maximum likelihood estimation, we can forecast with the MLR by setting $\boldsymbol{z}_{j,t}$ to the appropriate future time period and obtaining the estimated probabilities of death for each cause.

  To apply MLR to the US male cause-of-death mortality data, we treated age group as a factor variable and year as a continuous variable. Following Arnold and Sherris (2015) and Alai et al. (2015), we allowed for and performed variable selection on models involving linear, quadratic, and cubic terms for year, along with interaction terms between age group and (polynomials of) year, with the idea that different age groups may respond to year differently. To select the best structure, we applied five-fold rolling origin cross-validation as described in Section 4.2. Results from this variable selection showed that, for all three forecasting horizons, a structure involving a linear term of year plus interaction terms for age group and year was deemed most appropriate; see also Sithole et al. (2000) who uses a similar modeling structure for all cause mortality.
- **Product-ratio model:** A third competing method we consider is the product-ratio model of Hyndman et al. (2013). This is a functional forecasting method which ensures coherent mortality forecasting i.e., requiring the long-term convergence of mortality rates of the

Table 3
Out-of-sample forecasting performance, as based on the norm between the predicted and test set tensor. Values marked with an asterisk indicate methods for which there was statistically clear evidence (at a 5% level using both two-sample t-test and Wilcoxon two-sample test for the difference in mean absolute errors between the predicted and test data) that ADAPT produced more accurate out-of-sample predictions than that method at that specific forecasting horizon.

	Short-term	Mid-term	Long-term
ADAPT	3.512	5.177	9.348
CPD	3.663	5.611*	10.838*
Original Lee-Carter Model	4.317*	8.526*	11.333*
General Lee-Carter Model	4.068*	5.009	8.415
MLR - Structure 2	7.517*	12.921*	19.425*
Product-ratio	3.198	5.865*	9.753*

different sub-populations (in our setting these are the different causes of death). Note that such an assumption is arguably not biologically reasonable given the different evolving patterns of causes, although we nevertheless include this model to examine its forecasting performance. To implement the product-ratio method, we model (and forecast) the geometric mean mortality rate of the five causes of death, along with the ratio of the mortality rates for each cause of death over the geometric mean. We impose the coherence constraint through the use of stationary time series models for each of ratio models; see Hyndman et al. (2013) and the `R` package `demography` (Hyndman, 2023) for further mathematical and implementation details.

To summarize, we compared out-of-sample forecasting performance of ADAPT, the unpenalized CPD, the two versions of the Lee-Carter model, MLR, and the product-ratio model on the US male cause-of-death mortality data. As discussed above, tuning for all methods (except for the original Lee-Carter model and the product-ratio model) was done through five-fold rolling origin cross-validation, using the same training and validation splits. After tuning, each method was then fitted to the entire training and validation period, and predictions made for the test set. Performance was assessed based on the tensor norm between the predicted and test set tensor. Additionally, we performed difference-in-means tests (using both two-sample t-test and Wilcoxon two-sample test, and at a nominal 5% significance level) to assess the null hypothesis that there was no difference in the mean absolute errors (where the error is the difference between the prediction mortality rate for a particular method and the true value in the test tensor at given combination of cause, age group and year, and the mean is then based on averaging across cause, age group, and years in the forecasting horizon) between ADAPT and another method.

Results showed that ADAPT performed strongly overall (Table 3). It produced lower forecasting error than the unpenalized CPD across all three forecasting horizons, thus providing evidence that regularizing to exploit the underlying smooth trends in the data helped to improve prediction for this dataset. There was statistically clear evidence that ADAPT performed better than the original Lee-Carter model and MLR for all forecasting periods. It was interesting to observe how poorly the MLR performed, and we conjecture that this was due to its parametric form, meaning that it lacked sufficient flexibility to produce reliable forecast in general. For the short-term forecasting horizon, the product-ratio method performed best, followed by ADAPT and the general Lee-Carter model. For the mid- and long- term forecasting horizons, there was statistically clear evidence that ADAPT produced stronger predictive performance than all other methods except for the general Lee-Carter model, for which the null hypothesis that the mean absolute error was different between these two methods was not rejected. Indeed, it worth pointing out that had we not included a general Lee-Carter model with many factors for comparison, then ADAPT would have outperformed all the other candidate methods in the mid- and long- term forecasting horizon. The inclusion of the general Lee-Carter model thus offers a strong benchmark in terms of forecasting performance with which to compare tensor decomposition methods. However, since the general Lee-Carter approach models each cause separately, it involves many more parameters compared to the ADAPT approach, which models all causes together. In Appendix D, we present cause-specific out-of-sample prediction performance results, from which we see that ADAPT is the top performer in a number of but not in every case across the three forecasting horizons.

Figs. 5-6 present two examples of the fitted/predicted cause-specific log mortality rates based on applying the various methods to the US male cause-of-death mortality data: for infants (age 0) under a long-term forecasting horizon, and for older ages (age group 60-64) under a short-term forecasting horizon. Plots for all age groups and forecasting horizons broken down by age and group, may be found in the link provided in Appendix E. We also constructed predictions of the period life expectancy at birth (age 0), adulthood age (age 20), and retirement age (age 65) over a 25-years forecasting horizon. Note in practice insurers may desire even longer term forecasts, albeit produced with a larger volume of training data. However in this article we restrict ourselves to 25-year forecast particularly given we have around (only) 50 years of training data. Also, because of the fact that we have a final age group 85+, then we have provided truncated (or fixed-term) period life expectancies based on the methods discussed in Haberman and Pitacco (2018) and Pitacco et al. (2009). The resulting life expectancy plots are provided in Appendix D.5, from we observe that all models predict an increasing life expectancy at birth and at adulthood, except for the MLR model. The ADAPT and CPD models produce lower life expectancy forecasts than the general and original Lee-Carter models. Focusing in ADAPT specifically, for the predicted life expectancy plot at retirement age (65) it produced a relatively flat pattern with a little bit decreasing towards the end of 25 year forecasting. This was not surprising considering the more recent falls in US life expectancy (Shmerling, 2022).

6. Conclusion

This paper proposes a new adaptive penalized tensor decomposition approach for modeling cause-of-death mortality data. As a method which jointly models all three dimensions of the data, rather analyzing each cause independently, ADAPT builds on the existing approach of Madrid-Padilla and Scott (2017) by incorporating adaptive weighted penalty matrices to ensure that regularization/smooth remains computationally viable in the multi-factor setting. ADAPT can be coupled with different forecasting methods for the year dimension,

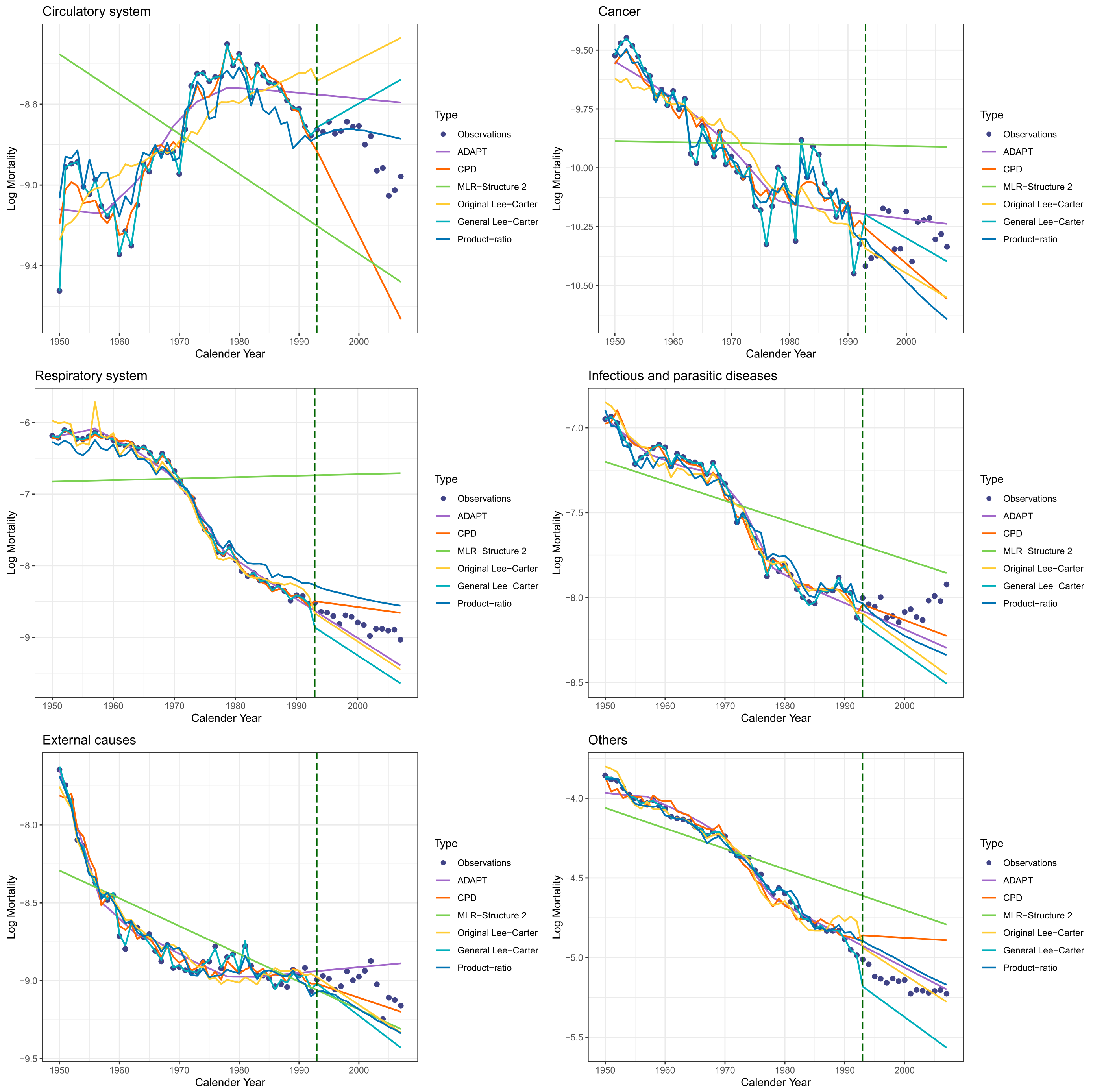


Fig. 5. Fitted/predicted cause-specific log mortality rates for infants (age 0) using a long-term forecasting horizon, based on applying various methods for mortality modeling. In each panel, the green vertical dashed line separates the figure into data used to fit the method (left) and the data used for assessing out-of-sample performance (right) parts.

including time-series methods such as random walk with drift as well as non-time-series approaches such as linear extrapolation and smoothing. We illustrate an application of ADAPT to US male cause-of-death mortality data over the period of 1950-2007, comparing it with several existing methods such as the unpenalized CPD, two versions of the Lee-Carter model, multinomial logistic regression, and product-ratio model. Results show that across six main causes of death and three forecasting horizons, ADAPT consistently outperformed the unpenalized CPD, the standard (rank-1) Lee-Carter model and multinomial logistic regression model in terms of out-of-sample forecasting performance. ADAPT was also competitive with and sometimes produced superior forecasting performance than the general Lee-Carter model (in the short-term, while requiring considerably fewer parameters) and product-ratio model (in the mid- and long-term, for which the underlying coherence assumption made not be biologically reasonable).

To the best of our knowledge, this paper is the first to apply penalized tensor decompositions to forecast the cause-of-death mortality rates. As such, there are a variety of future research directions to explore the full capacity of this method. Methodologically, ADAPT could be used to simulate the effects on mortality (and morbidity) of policy changes affecting the risk factors, as well the impact of improvements or elimination of specific causes (similar to Alai et al., 2015). Whether this involves directly modifying the elements in

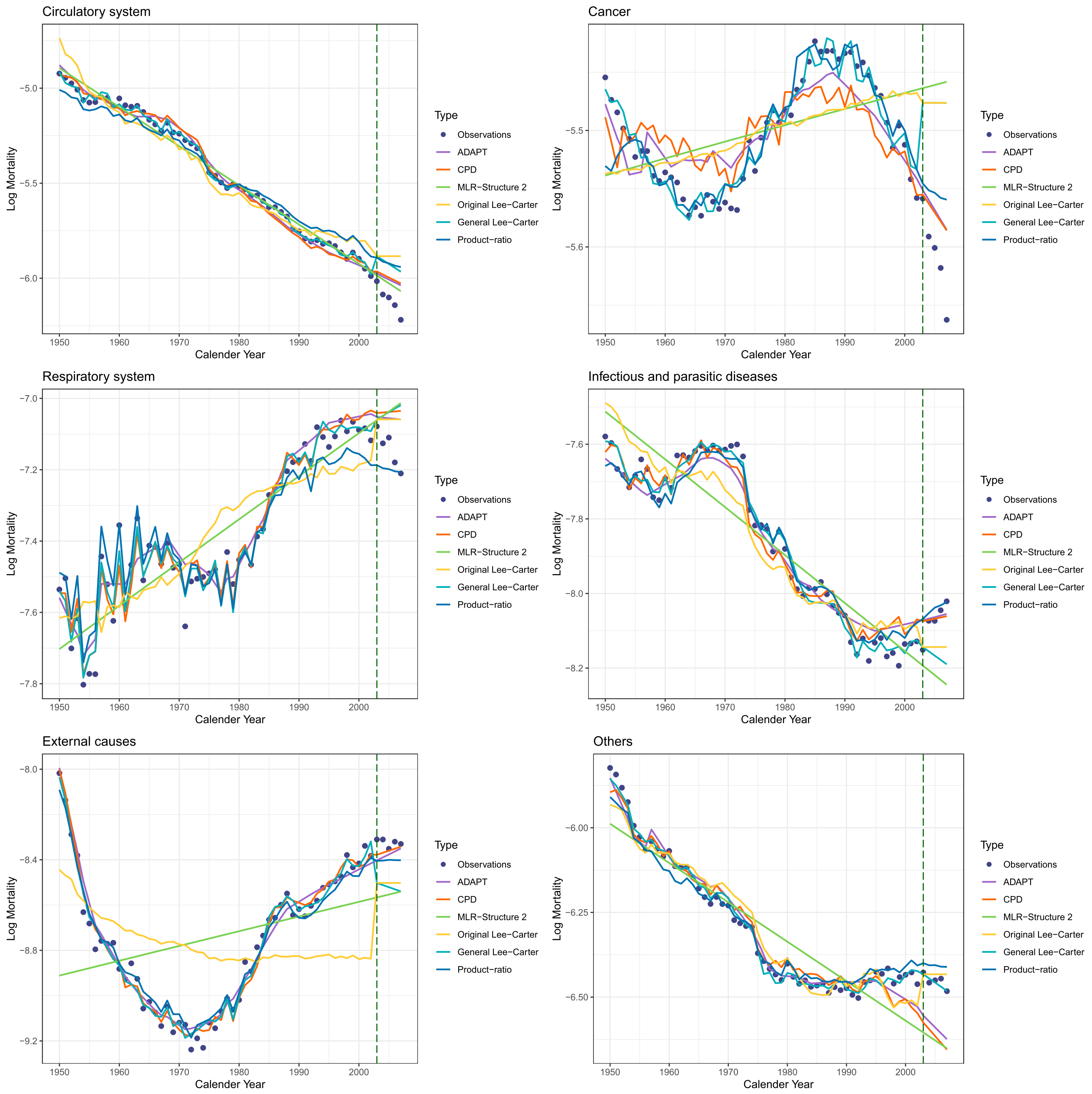


Fig. 6. Fitted/predicted cause-specific log mortality rates for age group 60-64 using a short-term forecasting horizon, based on applying various methods for mortality modeling. In each panel, the green vertical dashed line separates the figure into data used to fit the method (left) and the data used for assessing out-of-sample performance (right) parts.

the estimated vectors in ADAPT, or modifying the penalties to allow for such theoretical cause-elimination scenarios, is subject to further investigation. Extensions to the ADAPT framework could also be made to allow the vectors of non-temporal dimensions i.e., cause and age group, to vary over time. This could potentially make the forecasting more flexible, although the computation and prediction would necessarily become more complicated. Furthermore, instead of choosing the number of factors R separately, it may be possible to modify the adaptively weighted penalty matrices to directly incorporate selection of R (similar to Hui et al., 2018).

In terms of our application to the US male cause-of-death mortality data, note we adjusted for changes in the ICD codes using the comparability ratios approach of Gaille and Sherris (2011). However, another possibility would be compare forecasting performance across all methods when the same ICD code is used across the entire time span e.g., ICD7 all the way up to 2007. More generally, while ADAPT offers a novel method to jointly model cause, age group, and year dimensions, it is important to acknowledge that we do not always expect a method that explicitly accounts for dependencies among different causes of death to offer (substantial) improvements in predictive performance. Indeed, while there exists some literature showing that accounting for dependencies between causes of death is useful for a number of reasons, (e.g., see Alai et al., 2015; Arnold-Gaille and Sherris, 2016), from the particular illustration here the gains in out-of-sample forecasting performance are only marginal compared to the general Lee-Carter model. Had we fitted ADAPT to a different subset of the US cause-of-death mortality data, then perhaps it may not have performed as strongly as a well-tuned, cause-

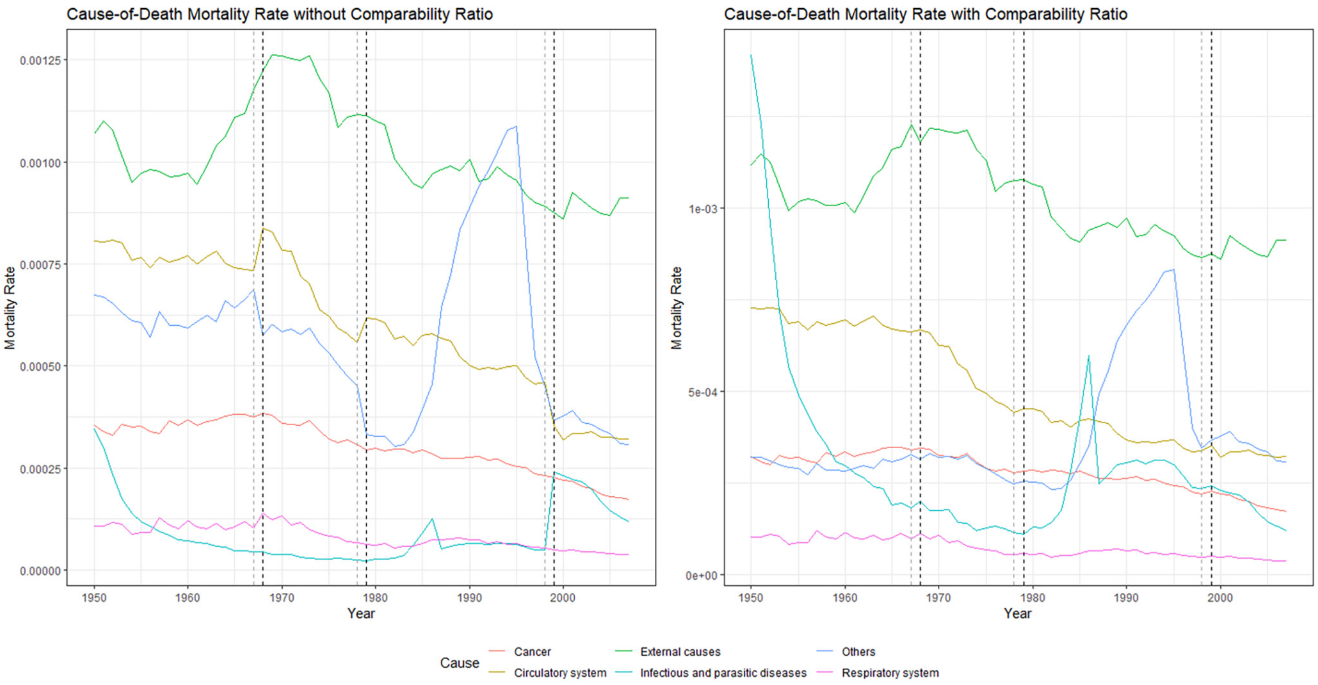


Fig. A1. The Cause-of-Death mortality rate for male between 35 and 39 in the US from 1950 to 2007 without the comparability ratio (left) and after the comparability ratio is used (right). Grey dash line indicates the last year of one ICD while black dash line indicates the first year of one newly adopted ICD. As ICD has changed three times over the past fifty years, in each plot, there are three grey dash lines and three black dash lines in total.

specific general Lee-Carter model (say) in terms of forecasting accuracy. We leave such applications to different versions of the data (including to more recent versions containing surges in deaths due to infectious diseases resulting from COVID-19, and examine/test for the possibility of structural changes in mortality patterns over time), along with broader investigations into comparing methods which account for dependencies among different causes versus those which treat each cause as separate, as a future line of research.

Declaration of competing interest

There is no competing interest.

Data availability

Codes published on GitHub already and the link has been provided in the paper.

Acknowledgement

We thank the reviewers for their helpful and insightful comments, which have enabled us to improve the paper substantially. FKCH was supported by an Australian Research Council Discovery Early Career Research Award DE200100435.

Appendix A. Additional details regarding the US male cause-of-death mortality data

Fig. A1 presents an example of the influence of comparability ratios on raw data for male mortality between 35 and 39 in the US. Notice that, after introducing comparability ratios, there was a dramatic increase in past mortality for “Infectious and parasitic diseases” as the result of a huge gap between 1998 and 1999 when the ICD10 was adopted, while for the “Others” cause a corresponding decrease could be observed. More importantly, while comparability ratios may have an outstanding impact on past mortality, the main trends in cause specific mortality are still preserved.

Appendix B. Algorithms for solving tensor decompositions

We begin by defining the mode **mode-n product** of a tensor $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ and a matrix $\mathbf{Y} \in R^{J \times I_n}$, which is denoted by $\mathcal{Z} = \mathcal{X} \times_n \mathbf{Y}$ where $\mathcal{Z} \in R^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N}$. Element-wise, we have $(\mathcal{X} \times_n \mathbf{Y})_{i_1 \ldots i_{n-1} j i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \ldots i_N} y_{j i_n}$.

By definition, the above operation can also be viewed such that each mode-n fiber (a fiber is the subarray of a tensor formed by fixing every index of a tensor but one) of the tensor $\mathcal{X}$ is multiplied by the matrix $\mathbf{Y}$. Therefore, the mode-n product of a tensor and a matrix can also be expressed via a matricized tensor as follows,

$$\mathcal{Z} = \mathcal{X} \times_n \mathbf{Y} \quad \Leftrightarrow \quad \mathbf{Z}_{(n)} = \mathbf{Y}\mathbf{X}_{(n)}.$$

Similarly, the mode-n product a tensor $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ and a vector $\mathbf{y} \in R^{I_n}$ is denoted by $\mathcal{Z} = \mathcal{X} \times_n \mathbf{y}$, where $\mathcal{Z} \in R^{I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N}$. Element-wise, we obtain $(\mathcal{X} \times_n \mathbf{y})_{i_1 \ldots i_{n-1} i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \ldots i_N} y_{i_n}$.

Algorithm B1 Block coordinate descent method for rank-1 CPD.

```
Input: given tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$; initial values of the vectors which are normalized to unit length $(\mathbf{u}^{(0)}, \mathbf{v}^{(0)}, \mathbf{w}^{(0)})$.
repeat  $k = 1, 2, 3 \ldots$
    $\mathbf{u}^{(k)}$ : $\tilde{\mathbf{u}}^{(k)} = \mathcal{X} \times_2 \mathbf{v}^{(k-1)} \times_3 \mathbf{w}^{(k-1)}$
        $d_u^{(k)} = \left\| \tilde{\mathbf{u}}^{(k)} \right\|$
        $\mathbf{u}^{(k)} = \tilde{\mathbf{u}}^{(k)} / d_u^{(k)}$
    $\mathbf{v}^{(k)}$ : $\tilde{\mathbf{v}}^{(k)} = \mathcal{X} \times_1 \mathbf{u}^{(k)} \times_3 \mathbf{w}^{(k-1)}$
        $d_v^{(k)} = \left\| \tilde{\mathbf{v}}^{(k)} \right\|$
        $\mathbf{v}^{(k)} = \tilde{\mathbf{v}}^{(k)} / d_v^{(k)}$
    $\mathbf{w}^{(k)}$ : $\tilde{\mathbf{w}}^{(k)} = \mathcal{X} \times_1 \mathbf{u}^{(k)} \times_2 \mathbf{v}^{(k)}$
        $d_w^{(k)} = \left\| \tilde{\mathbf{w}}^{(k)} \right\|$
        $\mathbf{w}^{(k)} = \tilde{\mathbf{w}}^{(k)} / d_w^{(k)}$
    $d^{(k)} = \mathcal{X} \times_1 \mathbf{u}^{(k)} \times_2 \mathbf{v}^{(k)} \times_3 \mathbf{w}^{(k)}$
until Convergence e.g., difference in $(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)})$ between successive iteration is less than some tolerance value $\epsilon > 0$.
Output: Estimates $(d^*, \mathbf{u}^*, \mathbf{v}^*, \mathbf{w}^*)$
```

Algorithm B2 Block coordinate descent method for rank-1 penalized tensor decomposition, using an ADMM algorithm.

```
Input: given tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$; initial values generated using the CPD $(\mathbf{u}^{(0)}, \mathbf{v}^{(0)}, \mathbf{w}^{(0)})$ and $\boldsymbol{\lambda}_u$, $\boldsymbol{\lambda}_v$, $\boldsymbol{\lambda}_w$; penalty matrices $\mathbf{D}_r^u$, $\mathbf{D}_r^v$, $\mathbf{D}_r^w$.
repeat  $k = 1, 2, 3 \ldots$
    $\mathbf{u}^{(k)}$ : Fix $\mathbf{v}^{(k-1)}$ and $\mathbf{w}^{(k-1)}$, and then update $\mathbf{u}$ using an ADMM algorithm (Madrid-Padilla and Scott, 2017).
    $\mathbf{v}^{(k)}$ : Fix $\mathbf{u}^{(k)}$ and $\mathbf{w}^{(k-1)}$, and then update $\mathbf{v}$ using an ADMM algorithm.
    $\mathbf{w}^{(k)}$ : Fix $\mathbf{u}^{(k)}$ and $\mathbf{v}^{(k)}$, and then update $\mathbf{w}$ using an ADMM algorithm.
    $d^{(k)} = \mathcal{X} \times_1 \mathbf{u}^{(k)} \times_2 \mathbf{v}^{(k)} \times_3 \mathbf{w}^{(k)}$.
until Convergence e.g., difference in $(\mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)})$ between successive iteration is less than some tolerance value $\epsilon > 0$.
Output: Estimates $(d^*, \mathbf{u}^*, \mathbf{v}^*, \mathbf{w}^*)$.
```

Below, we present the algorithms used for solving rank-1 CPD and rank-1 penalized tensor decomposition respectively. They are used as the basic building blocks for estimation as discussed in Sections 3.2 and 4, and are explicitly referred to in Algorithm 1. Note there are many existing algorithms to compute the CPD, and in this article we focus on the alternating least squares algorithm (ALS) approach of Kolda and Bader (2009) given that it can be formulated as a special case of the general class of block coordinate descent methods for optimization. Algorithm B1 summaries the procedures for using ALS to solve a rank-1 CPD i.e., when $R = 1$. An extension for solving the rank-R CPD is provided as a special case of Algorithm 1 in the main text, noting that it successively employs the above algorithm.

Turning to the estimation of a rank-1 penalized tensor decomposition, we can extend the ALS algorithm above by using the fast alternating direction method of multipliers (ADMM) algorithm of Arnold and Tibshirani (2016) to update vectors when a trend filtering penalty structure is augmented; we refer the reader to Madrid-Padilla and Scott (2017) and links to R code provided in Appendix E. Algorithm B2 summarizes the procedure. Note the details of convergence analysis for block coordinate update methods have been studied in Madrid-Padilla and Scott (2017).

An extension for solving the rank-R penalized tensor decomposition is provided as Algorithm 1 in the main text, noting that it successively employs the above algorithm.

Appendix C. Additional details of the hyperparameter selection process

See Fig. C1.

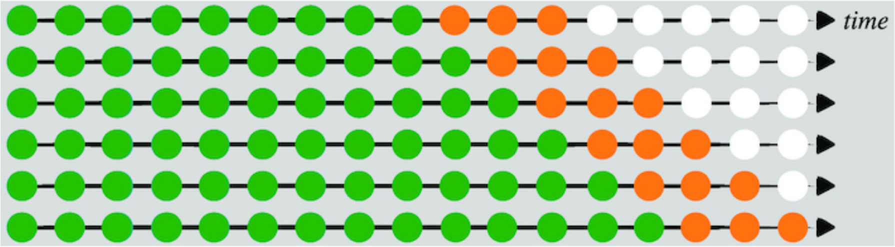


Fig. C1. A schematic of K-fold rolling origin cross-validation, in this case where the validation set consists of three time points and $K = 6$ folds. In the context of tensor analysis, each point in the schematic consists of a $I \times J$ matrix where the (i, j)-th element denotes the mortality rate of the i-th cause and j-th age group. Note the length of the training set increases with each fold by one time point per fold. The figure comes from Figure 2 of Li et al. (2021).

Appendix D. Additional results for application to the US cause-of-death mortality data

D.1. Full results of rolling origin cross-validation

Tables D1–D3 present examples results from applying five-fold rolling origin cross-validation results (based the mean tensor norm error between the predicted and validation set tensors, averaged across the folds) to ADAPT for three different forecasting horizons. As a baseline, we also provide the results based on applying the unpenalized CPD. Results show that in many of the settings, the decreasing trend of the mean tensor norm error starts to flatten out after 8 factors are included, indicating that the appropriate number of factor

Table D1
Results from five-fold rolling origin cross-validation for different number of factors R, choices of the form of the penalty matrices, and prediction method (RW = random walk with drift; LE = linear extrapolation; smooth = smoothing method via splines), using a 5-year (short-term) forecasting horizon. Note k_v denote the order of the trend filtering penalty used for $\mathbf{D}_v$, and similarly for k_w. The combination of hyperparameters that produces the lowest tensor norm for ADAPT is in bold.

	CPD			ADAPT			ADAPT			ADAPT			ADAPT		
				$(k_v = 1, k_w = 1)$			$(k_v = 1, k_w = 2)$			$(k_v = 2, k_w = 1)$			$(k_v = 2, k_w = 2)$		
R	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth
3	8.163	8.408	8.124	8.155	8.096	8.097	8.125	8.103	8.098	8.144	8.083	8.084	8.112	8.090	8.086
4	6.569	6.823	6.516	6.568	6.462	6.463	6.543	6.521	6.509	6.584	6.470	6.472	6.616	6.622	6.596
5	4.960	5.370	4.835	4.956	4.819	4.814	4.989	4.901	4.869	4.949	4.814	4.809	4.928	4.846	4.817
6	4.144	4.786	4.043	4.142	4.008	4.012	4.082	4.093	4.003	4.141	4.076	4.030	4.120	4.065	4.020
7	3.446	4.256	3.366	3.431	3.296	3.283	3.351	3.398	3.323	3.430	3.303	3.288	3.342	3.388	3.316
8	3.293	4.169	3.182	3.278	3.153	3.099	3.293	3.254	3.184	3.266	3.130	3.126	3.298	3.214	3.151
9	3.235	4.129	3.087	3.207	3.013	2.979	3.069	3.187	2.987	3.109	2.906	2.900	3.130	3.142	3.009
10	3.063	4.149	3.047	3.059	2.980	2.986	3.056	3.238	3.066	3.023	2.985	2.926	3.036	3.286	3.057
11	3.001	3.962	3.021	2.996	2.980	**2.765**	3.029	3.361	3.069	3.016	3.088	3.017	3.050	3.369	3.086
12	3.073	4.279	3.132	3.070	3.024	3.020	3.021	3.424	3.068	3.019	3.016	2.978	3.049	3.410	3.076
13	2.903	4.385	2.977	2.894	2.898	2.869	2.891	3.469	2.999	2.940	2.901	2.915	2.979	3.535	2.967
14	2.946	4.671	3.029	2.908	2.919	2.922	2.951	3.408	3.056	2.933	2.839	2.834	2.866	3.344	2.888

Table D2
Results from five-fold rolling origin cross-validation for different number of factors R, choices of the form of the penalty matrices, and prediction method (RW = random walk with drift; LE = linear extrapolation; smooth = smoothing method via splines), using a 10-year (mid-term) forecasting horizon. Note k_v denote the order of the trend filtering penalty used for $\mathbf{D}_v$, and similarly for k_w. The combination of hyperparameters that produces the lowest tensor norm for ADAPT is in bold.

	CPD			ADAPT			ADAPT			ADAPT			ADAPT		
				$(k_v = 1, k_w = 1)$			$(k_v = 1, k_w = 2)$			$(k_v = 2, k_w = 1)$			$(k_v = 2, k_w = 2)$		
R	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth
3	13.681	13.425	12.866	13.631	12.796	12.778	13.547	12.878	12.816	13.668	12.823	12.817	13.590	12.901	12.843
4	11.899	11.526	11.094	11.875	10.933	10.964	11.716	10.997	10.910	11.663	10.778	10.829	11.514	10.781	10.687
5	10.253	10.126	9.480	10.239	9.051	9.200	10.087	9.187	9.059	10.408	9.278	9.337	10.336	9.476	9.342
6	8.956	9.115	8.556	8.950	7.733	7.971	8.887	8.222	7.975	8.924	7.878	7.900	8.915	8.304	7.994
7	8.141	8.922	7.724	8.097	6.721	6.998	7.986	7.469	7.178	8.131	6.742	7.046	8.083	7.438	7.159
8	7.939	8.681	8.093	7.889	6.717	7.052	7.824	7.528	7.213	7.913	6.718	7.071	7.819	7.505	7.267
9	7.890	9.136	7.396	7.854	6.483	6.684	7.658	7.268	7.069	7.787	6.460	6.632	7.720	7.166	6.993
10	7.767	8.837	7.301	7.644	6.448	6.568	7.586	6.906	6.610	7.667	6.340	6.489	7.613	7.066	6.681
11	7.702	9.113	7.328	7.674	6.409	6.558	7.551	7.396	7.183	7.630	6.439	6.592	7.648	7.064	6.773
12	7.586	8.938	7.516	7.558	6.392	6.535	7.553	6.980	7.031	7.551	**5.968**	6.262	7.500	6.955	6.755
13	7.603	9.230	7.703	7.517	6.364	6.400	7.538	7.454	7.284	7.509	6.090	6.114	7.513	7.504	6.994
14	7.538	9.198	7.469	7.491	6.323	6.384	7.424	7.347	7.029	7.542	6.451	6.521	7.440	7.274	7.143

Table D3
Results from five-fold rolling origin cross-validation for different number of factors R, choices of the form of the penalty matrices, and prediction method (RW = random walk with drift; LE = linear extrapolation; smooth = smoothing method via splines), using a 15-year (long-term) forecasting horizon. Note k_v denote the order of the trend filtering penalty used for $\mathbf{D}_v$, and similarly for k_w. The combination of hyperparameters that produces the lowest tensor norm for ADAPT is in bold.

	CPD			ADAPT			ADAPT			ADAPT			ADAPT		
				$(k_v = 1, k_w = 1)$			$(k_v = 1, k_w = 2)$			$(k_v = 2, k_w = 1)$			$(k_v = 2, k_w = 2)$		
R	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth	RW	LE	Smooth
3	19.193	21.759	19.757	18.311	17.358	17.424	19.163	20.193	18.961	18.311	17.358	17.424	19.162	20.191	18.962
4	18.983	21.492	18.462	18.577	15.909	15.988	19.061	18.593	17.546	18.647	16.165	16.265	18.719	19.209	17.662
5	17.030	20.299	16.386	16.725	13.437	13.515	16.715	16.320	15.175	16.436	13.010	13.099	16.975	16.579	15.574
6	17.157	20.036	17.833	16.541	13.292	13.358	17.041	17.559	15.358	16.468	13.141	13.243	16.498	16.806	15.214
7	16.019	20.091	15.453	15.402	12.155	12.067	16.140	16.303	14.705	15.603	12.458	12.385	16.107	15.931	14.761
8	16.368	21.392	15.471	16.322	13.438	13.328	16.195	16.527	15.080	16.124	13.314	13.420	16.222	16.890	14.736
9	15.800	21.312	16.166	15.687	13.450	13.435	15.674	16.778	14.966	15.854	13.734	12.808	15.684	17.655	15.383
10	15.869	22.259	15.615	15.748	13.204	12.778	15.763	16.509	14.129	15.694	13.807	13.219	15.667	17.123	13.572
11	15.742	22.371	15.637	15.674	13.478	13.366	15.640	17.583	14.968	15.690	**12.706**	12.872	15.656	17.118	13.017
12	15.647	22.689	14.520	15.596	14.780	13.426	15.564	16.147	14.109	15.604	14.148	12.554	15.556	17.299	13.991
13	15.589	23.362	17.032	15.477	14.638	14.381	15.495	17.102	13.364	15.543	14.198	13.461	15.581	17.095	13.211
14	15.534	23.533	17.345	15.517	14.275	13.428	15.451	16.904	14.067	15.465	14.439	12.270	15.534	17.288	14.349

can vary between 8 and 14. Also, all things being equal, models with order 1 trend filtering tend to have errors, indicating that we do not need to impose relatively strong smooth constraints especially on the Year dimension. Arguably the most consistent and important trend though is that ADAPT consistently outperforms CPD, suggesting that the inclusion of penalty matrices and smoothing can lead to improved predictive performance. This is (ultimately) demonstrated when we considered out-of-sample performance in Section 5.3 in the main text.

D.2. Example results of estimated vectors from ADAPT versus CPD

Fig. D1 presents examples of the estimated vectors for ADAPT versus the (unpenalized) CPD in the case of the short-term forecasting horizon; see Appendix E for links to the full set of plots. Overall, there was relatively little difference in the estimated vectors between ADAPT and CPD for the cause dimension, which was not surprising given that this dimension was not penalized. On the other hand, for the age group dimension the underlying quadratic trends in the US male mortality data (especially for mid- and long-term forecasting

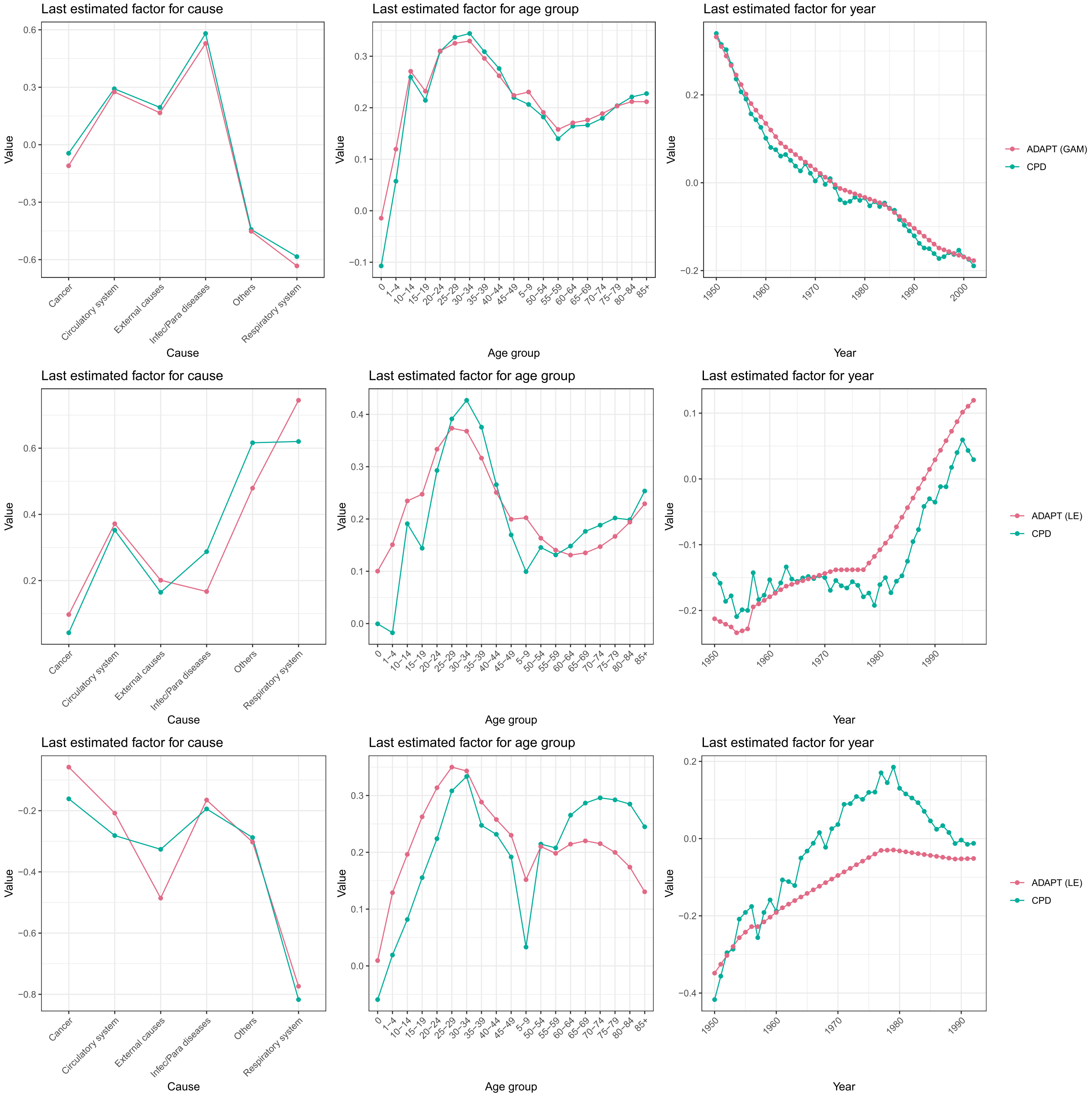


Fig. D1. Estimated vectors for cause (left), age group (middle), and year (right) for the last factor of ADAPT versus (unpenalized) CPD under the three forecasting horizons (top: short-term (5 years); middle: mid-term (10 years); bottom: long-term (15 years).

horizons) were identified by ADAPT, and as a result the estimated vectors were much smoother than those based on CPD. Turning to the year dimension, we again see the clear effects of trend filtering in ADAPT when considering mid- and long-term forecasting horizons, as it produces much smoother estimated vectors that we anticipate will lead to superior forecasting performance.

D.3. Results from a three-factor ADAPT model

Fig. D2 presents the resulting estimated vectors for cause, age group, and year under the three forecasting horizons, noting that the factors in ADAPT (as well as the tensor decomposition methods considered in this article) are only identifiable up to sign-flipping. It is interesting to observe, though perhaps not surprisingly, that the estimated factors for all three dimensions exhibit very similar patterns across the three forecasting horizons. As such, we focus on interpreting the long-term forecasting horizon case here as an example.

The estimated factors for age groups generally exhibit similar patterns to what we discussed earlier for Fig. 4. Note however that the third estimated factor for age group presents a different pattern compared to the first two factors, which is perhaps not too surprising as it aims to capture the remaining patterns after the first two factors have been accounted for. Turning to the cause-of-death dimension, we observe that the first two estimated factors fluctuate around similar levels across different causes of death. On the other hand, the third estimated factor presents a different pattern, indicating Cancer cause mortality as relatively the lowest and Others/Respiratory system

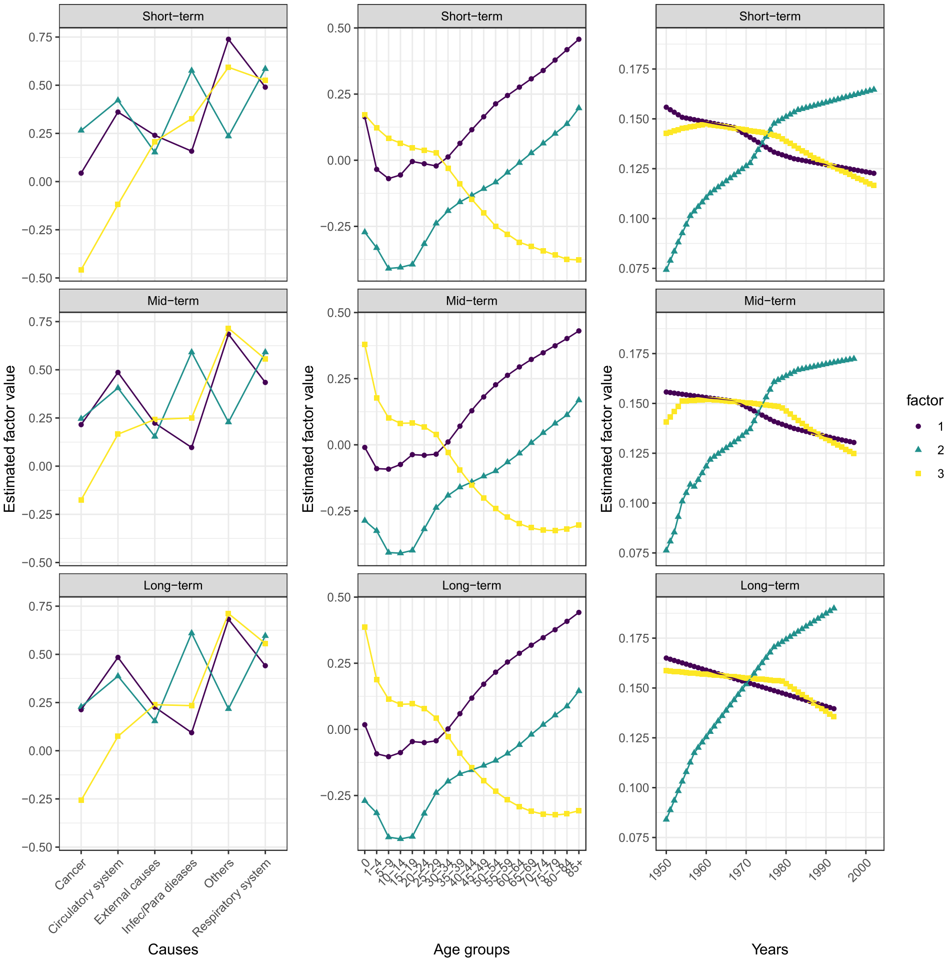


Fig. D2. Estimated vectors for cause (left), age group (middle), and year (right) for the first three factors of ADAPT under the three forecasting horizons (top: short-term; middle: mid-term; bottom: long-term).

cause mortality as relatively the highest, conditional on age group (mostly adult ages) and year. Finally, the patterns seen in the estimated factors for year can be interpreted conditional on the causes and age factors. That is, and generally speaking, the first and third estimated factors for year decreased over time conditional on cause of death and age group, while the second estimated factor for year exhibits an increasing pattern for most adult ages.

D.4. Cause-specific out-of-sample forecasting error

See Table D4.

Table D4
Cause-specific out-of-sample forecasting performance across the three forecasting horizons, based on the norm between the predicted and test set matrix for each cause. The methods compared include ADAPT, the unpenalized CPD, the two versions of the Lee-Carter (LC) model, MLR, and the product-ratio method (PR). The best performing method for each cause at each forecasting horizon is highlighted.

	5-year (short-term) forecasting horizon					
	ADAPT	CPD	Original LC	General LC	MLR	PR
Cancer	0.729	**0.615**	1.161	0.998	0.928	0.732
Circulatory system	**1.425**	1.629	1.759	1.591	2.728	1.532
External causes	2.089	2.088	2.778	2.569	5.895	**1.635**
Infectious/parasitic diseases	1.505	1.478	1.664	1.746	2.349	**1.370**
Others	1.096	1.373	0.880	0.844	1.640	**0.697**
Respiratory system	**1.394**	1.400	1.714	1.632	2.291	1.518
	10-year (mid-term) forecasting horizon					
	ADAPT	CPD	Original LC	General LC	MLR	PR
Cancer	1.126	0.936	1.297	**0.894**	1.421	1.407
Circulatory system	**1.889**	2.305	2.996	2.095	4.297	2.425
External causes	**2.712**	3.308	6.245	2.976	10.657	3.465
Infectious and parasitic diseases	2.416	2.440	3.004	**2.188**	3.581	2.419
Others	1.841	1.561	2.476	**1.319**	2.707	1.591
Respiratory system	2.319	2.440	2.806	**2.125**	3.571	2.480

(*continued on next page*)

Table D4 (*continued*)

	15-year (long-term) forecasting horizon					
	ADAPT	CPD	Original LC	General LC	MLR	PR
Cancer	**1.335**	3.330	1.698	1.415	6.387	1.949
Circulatory system	3.463	3.179	4.112	3.467	2.926	**2.851**
External causes	**4.879**	6.804	8.010	5.273	11.172	5.232
Infectious and parasitic diseases	4.801	4.529	3.733	**2.970**	3.905	4.834
Others	3.549	4.062	3.524	**2.602**	4.717	3.364
Respiratory system	3.761	**3.601**	4.260	3.661	13.025	4.598

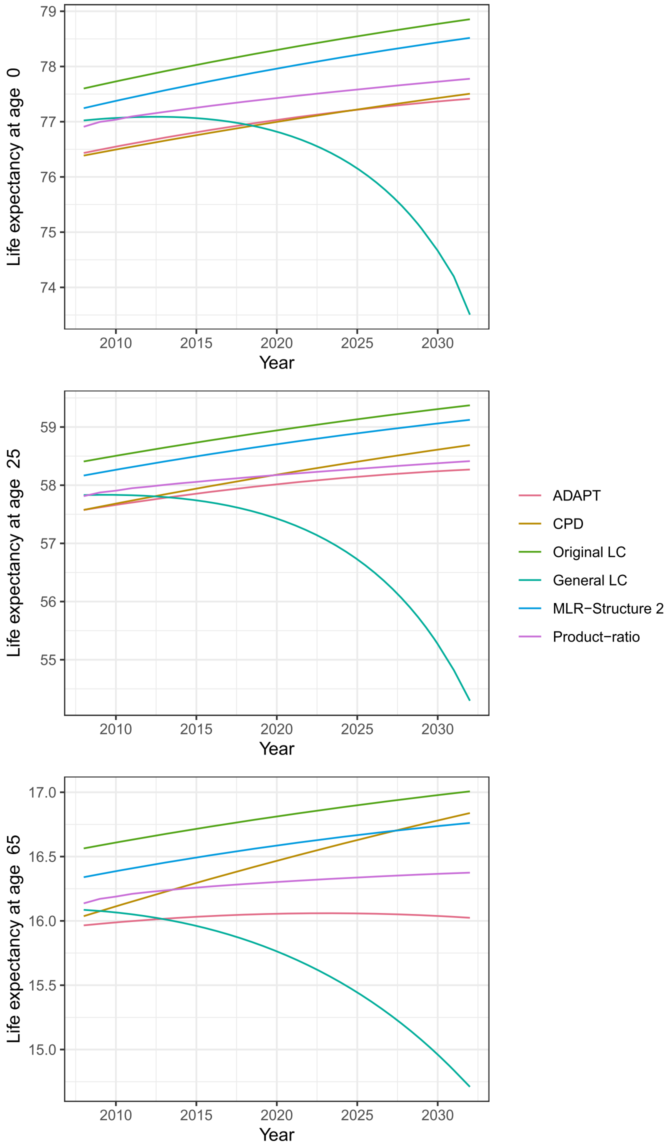


Fig. D3. Predicted truncated period life expectancy at birth (0), adulthood age (25), and retirement age (65) over a 25 years horizon, based on different methods.

D.5. Predicted life expectancy plots

See Fig. D3.

Appendix E. R codes and additional plots

The GitHub repository https://github.com/lullabies777/Adaptive-Penalized-Tensor-Decomposition.git includes R code for implementing CPD, ADAPT, both versions of the Lee-Carter model, multinomial logistic regression. An R shiny app providing further results from the application to the US male cause-of-death mortality data can be found at https://xuanmingzhang.shinyapps.io/rshiny_final/.

References

Alai, D.H., Arnold, S., Sherris, M., 2015. Modelling cause-of-death mortality and the impact of cause-elimination. Annals of Actuarial Science 9, 167–186.

Ali, A., Tibshirani, R.J., 2019. The generalized lasso problem and uniqueness. Electronic Journal of Statistics 13, 2307–2347.

Arnold, S., Glushko, V., 2021. Cause-specific mortality rates: common trends and differences. Insurance: Mathematics and Economics 99, 294–308.

Arnold, S., Sherris, M., 2013. Forecasting mortality trends allowing for cause-of-death mortality dependence. North American Actuarial Journal 17, 273–282.

Arnold, S., Sherris, M., 2015. Causes-of-death mortality: what do we know on their dependence? North American Actuarial Journal 19, 116–128.

Arnold, T.B., Tibshirani, R.J., 2016. Efficient implementations of the generalized lasso dual path algorithm. Journal of Computational and Graphical Statistics 25, 1–27.

Arnold-Gaille, S., Sherris, M., 2016. International cause-specific mortality rates: new insights from a cointegration analysis. ASTIN Bulletin 46, 9–38.

Basellini, U., Camarda, C.G., Booth, H., 2022. Thirty years on: a review of the Lee–Carter method for forecasting mortality. International Journal of Forecasting.

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

Caselli, G.J.V., Marsili, M., 2006. How useful are the causes of death when extrapolating mortality trends. In: Social Insurance Studies from the Swedish Social Insurance, pp. 9–36.

Currie, I.D., Durban, M., Eilers, P.H., 2004. Smoothing and forecasting mortality rates. Statistical Modelling 4, 279–298.

Dong, Y., Huang, F., Yu, H., Haberman, S., 2020. Multi-population mortality forecasting using tensor decomposition. Scandinavian Actuarial Journal 2020, 754–775.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Annals of Applied Statistics 1, 302–332.

Gaille, S., Sherris, M., 2011. Modelling mortality with common stochastic long-run trends. The Geneva Papers on Risk and Insurance. Issues and Practice 36, 595–621.

Haberman, S., Pitacco, E., 2018. Actuarial Models for Disability Insurance. Routledge.

Hanewald, K., 2011. Explaining mortality dynamics. North American Actuarial Journal 15 (2), 290–314.

He, L., Huang, F., Shi, J., Yang, Y., 2021. Mortality forecasting using factor models: time-varying or time-invariant factor loadings? Insurance: Mathematics and Economics 98, 14–34.

Heligman, L., Pollard, J.H., 1980. The age pattern of mortality. Journal of the Institute of Actuaries 107, 49–80.

Hui, F.K.C., Mueller, S., Welsh, A.H., 2017. Hierarchical selection of fixed and random effects in generalized linear mixed models. Statistica Sinica 27, 501–518.

Hui, F.K.C., Müller, S., Welsh, A.H., 2020. The LASSO on latent indices for regression modeling with ordinal categorical predictors. Computational Statistics & Data Analysis 149, 106951.

Hui, F.K.C., Tanaka, E., Warton, D.I., 2018. Order selection and sparsity in latent variable models via the ordered factor LASSO. Biometrics 74, 1311–1319.

Hyndman, R., 2023. demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 2.0.

Hyndman, R.J., Booth, H., Yasmeen, F., 2013. Coherent mortality forecasting: the product-ratio method with functional time series models. Demography 50 (1), 261–283.

Kim, S.-J., Koh, K., Boyd, S., Gorinevsky, D., 2009. l_1 trend filtering. SIAM Review 51, 339–360.

Kolda, T.G., Bader, B.W., 2009. Tensor decompositions and applications. SIAM Review 51, 455–500.

Lee, R.D., Carter, L.R., 1992. Modeling and forecasting U.S. mortality. Journal of the American Statistical Association 87, 659–671.

Li, H., Li, H., Lu, Y., Panagiotelis, A., 2019. A forecast reconciliation approach to cause-of-death mortality modeling. Insurance: Mathematics and Economics 86, 122–133.

Li, H., Lu, Y., 2019. Modeling cause-of-death mortality using hierarchical archimedean copula. Scandinavian Actuarial Journal 2019, 1–26.

Li, Q., Bedi, T., Lehmann, C.U., Xiao, G., Xie, Y., 2021. Evaluating short-term forecasting of covid-19 cases among different epidemiological models under a bayesian framework. GigaScience 10:giab009.

Liu, Z., Yang, X., 2022. Cross validation for uncertain autoregressive model. Communications in Statistics. Simulation and Computation 51 (8), 4715–4726.

Madrid-Padilla, O.H., Scott, J., 2017. Tensor decomposition with generalized lasso penalties. Journal of Computational and Graphical Statistics 26, 537–546.

McNown, R., Rogers, A., 1992. Forecasting cause-specific mortality using time series methods. International Journal of Forecasting 8, 413–432.

Pitacco, E., Denuit, M., Haberman, S., Olivieri, A., 2009. Modelling Longevity Dynamics for Pensions and Annuity Business. Oxford University Press.

Quah, S., 2016. International Encyclopedia of Public Health. Academic Press.

Redondo Lourés, C., Cairns, A.J., 2021. Cause of death specific cohort effects in U.S. mortality. Insurance. Mathematics & Economics 99, 190–199.

Renshaw, A., Haberman, S., 2003. Lee-Carter mortality forecasting with age-specific enhancement. Insurance. Mathematics & Economics 33, 255–272.

Russolillo, M., Giordano, G., Haberman, S., 2011. Extending the Lee-Carter model: a three-way decomposition. Scandinavian Actuarial Journal 2011, 96–117.

Shmerling, R.H., 2022. Why life expectancy in the us is falling. https://www.health.harvard.edu/blog/why-life-expectancy-in-the-us-is-falling-202210202835#:~:text=A%20dramatic%20fall%20in%20life,just%20over%2076%2C%20in%202021. (Accessed 22 January 2023).

Sithole, T.Z., Haberman, S., Verrall, R., 2000. An investigation into parametric models for mortality projections, with applications to immediate annuitants' and life office pensioners' data. Insurance. Mathematics & Economics 27, 285–312.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B) 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Series B, Statistical Methodology 67.

Tibshirani, R.J., Taylor, J., et al., 2011. The solution path of the generalized lasso. The Annals of Statistics 39, 1335–1371.

University of California, Berkeley (USA), Max Planck Institute for Demographic Research (Germany), 2020. Human mortality database. www.mortality.org. (Accessed 22 October 2020).

Wang, M., Fischer, J., Song, Y.S., 2019. Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. Annals of Applied Statistics 13, 1103–1127.

Whelan, J.C., Buhler-Wilkerson, K., 2020. Nursing, history, and health care. https://www.nursing.upenn.edu/nhhc/nurses-institutions-caring/care-of-premature-infants/. (Accessed 22 October 2020).

Wilmoth, J.R., 1995. Are mortality projections always more pessimistic when disaggregated by cause of death? Mathematical Population Studies 5, 293–319.

Wood, S., 2017. Generalized Additive Models: An Introduction with R, second edition. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

World Health Organization, 2004. ICD-10: international statistical classification of diseases and related health problems: tenth revision. https://www.cdc.gov/nchs/icd/icd10.htm.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.