Multi-population Mortality Forecasting using Tensor Decomposition

Yumo Dong[1], Fei Huang*[1], Honglin Yu[1], and Steven Haberman[2]

[1]*Research School of Finance, Actuarial Studies and Applied Statistics , Australian National University, Australia*
[2]*Cass Business School, City, University of London*

Abstract

In this paper, we formulate the multi-population mortality forecasting problem based on 3-way (age, year, and country/gender) decompositions. By applying the canonical polyadic decomposition (CPD) and the different forms of the Tucker decomposition to multi-population mortality data (10 European countries and 2 genders), we find that the out-of-sample forecasting performance is significantly improved both for individual populations and the aggregate population compared with using the single-population mortality model based on rank-1 singular value decomposition (SVD), or the Lee-Carter model. The results also shed lights on the similarity and difference of mortality among different countries. Additionally, we compare the variance-explained method and the out-of-sample validation method for rank (hyper-parameter) selection. Results show that the out-of-sample validation method is preferred for forecasting purposes.

Keywords
Multi-population mortality forecasting, Tensor decomposition, CPD, Tucker, SVD

1 Introduction

Mortality modelling is an important topic in actuarial science and insurance practice. We divide the development of mortality modelling into four stages. In the first stage, the models are deterministic and 1 dimensional. Gompertz (1825) model suggests that mortality increases exponentially with age during the adult years of life. Makeham (1860) extends the Gompertz (1825) model by adding an age-independent component. The Gompertz-Makeham law states that the human death rate is the sum of an age-independent component (the Makeham term) and an age-dependent component (the Gompertz function) which increases exponentially with age.

*Correspondence to: Fei Huang, Research School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics, Australian National University, Canberra, ACT 2601, Australia. Email: fei.huang@anu.edu.au, Phone: (+61) 2 612 57390, Fax: (+61) 2 612 50087.

In the second stage, mortality models are deterministic and 2 dimensional. People began to realise at the start of the 20th century that time trends are important (especially for mortality of annuitants) and need to be incorporated in mortality models. Thus, the subject "Mortality tables for annuitants" was one of the main subjects discussed at the 5th International Congress of Actuaries in 1906 in Berlin, see Cramer & Wold (1935) for more information. Some models are based on the independent projection of age-specific mortality or hazard rates, including mortality reduction factor models (e.g. CMI 1990, CMI 1999, Willets 1999, Renshaw & Haberman 2000). The CMI first proposes an age-based innovation in 1924. This kind of models allows each age-specific rate to change at its own individual rate, though the projected age profile of mortality may depart from plausible, historically observed patterns (Keyfitz 1981). Some models are based on the projection of parameters of specific mortality laws, for example the Gompertz-based projection in Wetterstrand (1981) and the Heligman and Pollard-based projection in Forfar & Smith (1985). Comparing with the age-based models, the law-based models are usually more parsimonious. However, both models generate implausible projected mortality trends for age-specific mortality rates. A good reference on the history of mortality models is 'History of Actuarial Science' (Haberman & Sibbett 1995).

In the third stage, a number of new approaches have been developed for forecasting mortality using stochastic models, which focus on the age-sex-specific mortality data within a specific population. The seminal paper in this stage is the Lee & Carter (1992) model, which applies the singular value decomposition (SVD) method on the 2 dimensional mortality data. Although this model is an enlightening pioneer to subsequent researches, it also has drawbacks. For example, Lee & Miller (2001) find that the mortality forecast is biased when using the fitting period 1900 to 1989 to forecast the period 1990 to 1997 under the Lee & Carter (1992) framework. Also, it is pointed out that the pattern of changes in mortality is not fixed over time, contradicting an assumption of the Lee-Carter model. To fix the above issues, Lee & Miller (2001) changes the starting point of the fitting period to 1950 instead of 1933 and adjusts the mortality index k_t (year vector) by fitting to $e(0)$. Later, Renshaw & Haberman (2006) generalise the Lee & Carter (1992) model to include a cohort effect term. Currie et al. (2006) introduces the simpler Age-Period-Cohort (APC) model, in which mortality rates are influenced independently by age, period and cohort effects. The Cairns-Blake-Dowd (CBD) model, developed by Cairns et al. (2006), was primarily designed for older-age mortality predictions and has the advantage of a linear structure. There are two factors in this model. The first has a uniform over age effect on mortality rate dynamics, while the second has a differential effect on mortality rate dynamics. Since then, the CBD model has been further generalised by including additional terms, including those corresponding to cohort and quadratic age effects (Cairns et al. 2009). There have also been other developments in this stage, for example Acton et al. (2009), Yao et al. (2018).

In recent years, there have been research on multi-population mortality models, for example using a group of countries with similar social-economic status, or males and females in the same population. We call multi-population mortality modelling as the fourth stage. The seminal paper in this stage is the Li & Lee (2005) model, which extends the Lee & Carter (1992) model to allow for a group of populations. They use a common factor to describe the long-term mortality trend shared by all countries within a group and use a country-specific factor to describe the short-term country-specific mortality patterns. From then on, many other scholars develop their own models inspired by Li & Lee (2005). For example, Kleinow (2015) proposes a common-age-effect (CAE) multi-population model, which firstly uses a common set of age-response

parameters across the different populations. Li et al. (2015) consider two-population variants of seven of the models first considered by Cairns et al. (2009). Enchev et al. (2017) find that the Li & Lee (2005) potentially suffers from robustness problems when calibrated using maximum likelihood. An important application of the multi-population mortality models is to assess the demographic basis risk involved in an index-based longevity hedge by comparing and projecting the mortality experience for the reference and target populations. For example, Villegas et al. (2017) conducted a comparative study of two-population models for the assessment of basis risk in longevity hedges.

As discussed in Yao et al. (2018), the single-population mortality modelling problem can be solved based on matrix decomposition (or matrix factorization), which is to approximate the original data matrix using low-rank matrix representations. For example, Lee & Carter (1992) use the SVD method to forecast the US mortality data. However, the single-population models cannot be applied to multi-population mortality data with 3 or more dimensions. A natural extension of the matrix decomposition (2 dimensions) is tensor decomposition (3 or more dimensions), which can be used for multi-population mortality modelling. A tensor is a multidimensional or N-way array. There are mainly two commonly used tensor decomposition methods: the canonical polyadic decomposition (CPD) and the Tucker decomposition, see Rabanser et al. (2017). Decompositions of higher-order tensors (i.e., N-way arrays with $N \geq 3$) have been motivated and applied in many other fields. For examples, Appellof & Davidson (1981) first use tensor decompositions in chemometrics. Shashua & Levin (2001) apply tensor decompositions to image compression and classification. This approach has also been used in data mining, multilayer networks, signal processing and elsewhere, see Acar et al. (2005), Acar et al. (2006), Kivelä et al. (2014) and Comon (2009). In the field of mortality modelling, Russolillo et al. (2011) first use a rank-2 Tucker decomposition to analyse the mortality of 10 European countries, which can be regarded as a natural extension of the Lee-Carter model. They choose the rank-2 Tucker model, because it can explain 84.75% of the total variation. They also highlight that tensor decomposition allows researchers to obtain combined estimates for the three modes (age, time and different populations) and see relationships among them. In this paper, we generalise the model used in Russolillo et al. (2011) by analysing both the CPD method and 3 forms of the Tucker method to allow for different ranks in different dimensions (i.e. age, year and country/gender). By applying the CPD and Tucker methods to mortality data of 10 European countries and 2 genders, we evaluate the out-of-sample forecasting performance of the new methods. We also compare the variance-explained method and the out-of-sample validation method for selecting ranks from the perspective of prediction accuracy. The results of this paper indicate that the proposed tensor models can significantly and consistently improve out-of-sample predicting performance. We hope that the paper will motivate researchers and practitioners to apply tensor decomposition in multi-population mortality modelling and other related fields.

The paper is organised as follows. Section 2 formulates the multi-population mortality modelling problem based on tensor decompositions and describes the main steps to model and forecast mortality rates using the CPD method and the Tucker method. Section 3 discusses the modelling and prediction results using the mortality data of 10 European countries and 2 genders. Finally, Section 4 concludes this paper.

2 Model

Assume $M \in \mathbb{R}^{X \times T \times H}$ is the tensor data to be modelled and predicted, which has three dimensions: age, year and country. The element (x, t, h) of M is denoted by $m_{x,t,h}$. In this paper, we define $m_{x,t,h}$ to be the centered log mortality for age x, year t and country h. Let $\mu_{x,t,h}$ be the central death rate for age x, year t and country h, then we have

$$m_{x,t,h} = \log\mu_{x,t,h} - \alpha_{x,h} \tag{1}$$

where $\alpha_{x,h}$ is estimated by averaging the log-mortality $\log\mu_{x,t,h}$ over time. We formulate the multi-population mortality forecasting problem based on two commonly used tensor decompositions: the canonical polyadic decomposition (CPD) and Tucker decomposition, see Rabanser et al. (2017).

2.1 The CPD method

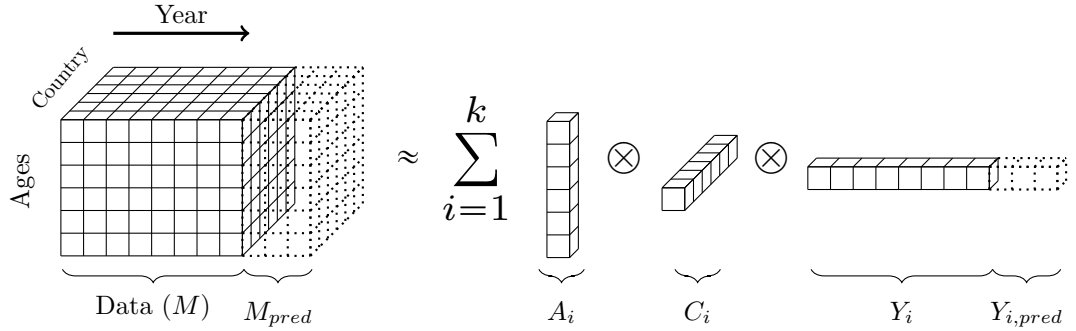

Figure 1: CPD method

Using $A_i \in \mathbb{R}^X$, $C_i \in \mathbb{R}^H$ and $Y_i \in \mathbb{R}^T$ to represent the ith latent factors of age, country and year respectively, the CPD model can be written as,

$$M \approx \hat{M} = \sum_{i=1}^{k} A_i \circ C_i \circ Y_i, \tag{2}$$

where k is the rank of the tensor decomposition, which indicates the number of latent factors. It is often useful to assume that the columns of A, C, and Y are normalized to length one with the weights absorbed into the vector $\lambda \in \mathbb{R}^k$.

$$M \approx \hat{M} = \sum_{i=1}^{k} \lambda_i A_i \circ C_i \circ Y_i \tag{3}$$

The $\circ$ symbol denotes the vector outer product, which means that each element of the tensor is the product of the corresponding vector elements. Let the jth entry of vectors A_i, C_i and Y_i be $a_j^{(i)}$, $c_j^{(i)}$ and $y_j^{(i)}$, respectively. Elementwise, we have

$$m_{x,t,h} \approx \hat{m}_{x,t,h} = \sum_{i=1}^{k} \lambda_i a_x^{(i)} c_h^{(i)} y_t^{(i)} \tag{4}$$

The predicted value can be calculated by,

$$M_{pred} = \sum_{i=1}^{k} \lambda_i A_i \circ C_i \circ Y_{i,pred} \tag{5}$$

In this paper, we apply the canonical polyadic decomposition (CPD) method for mortality modelling and forecasting, which factorises the tensor into a sum of rank-1 tensors. We use the alternating least squares (ALS) algorithm to compute the CPD of a tensor. The key idea behind this algorithm is to fix all factor matrices except for one in order to optimise the non-fixed matrix and then repeat this step for every matrix repeatedly until some stopping criteria is satisfied (Rabanser et al. 2017). The details of this algorithm can be found in the supplementary materials. The solution of the CPD method is usually not unique[1]. To overcome this problem, we randomly initialise 300 times for the CPD method, and select the solution with the optimal fit using the root mean square error (RMSE) as the selection criterion. Assume $\hat{\log\mu}_{x,t,h}$ is the fitted value and $\log\mu_{x,t,h}$ is the actual observation in the data set, we have

$$RMSE = \sqrt{\frac{\sum_{x=1}^{X}\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\hat{\log\mu}_{x,t,h} - \log\mu_{x,t,h}\right)^2}{XTH}}.$$

The reason for using RMSE is that the cost function in the algorithm of the tensor decomposition is the square root of the sum of the square errors, see Kolda & Bader (2009). Hence, it is reasonable to apply a consistent criterion. We find that the RMSE becomes stable when the number of random initializations is over 200 times, so 300 times is good enough to make sure the solution converges. Through this process, we can get the unique solution for a given rank.

After getting the decomposition result, we fit random walk with drift (RMD) models to the estimated year vectors Y_i to capture their dynamic patterns while keeping the number of parameters consistent and low for each time-series model[2]. Combining the predicted year vectors $Y_{i,pred}$ with the fitted country vectors C_i and age vectors A_i, we can obtain the predicted M_{pred} following Equation (5).

[1]As tensor decomposition is a non-convex optimization problem, we are unable to find the global optimum. Hence, each time that we implement the tensor decomposition, we will obtain a local optimum.

[2]If we fit a different automatic autoregressive integrated moving average (ARIMA) to each latent year vector, the final model will be very complex. However, we have considered the same ARIMA(p,d,q) model structure to each latent year vector to reduce the complexity. The out-of-sample forecasting result shows no significant improvement. So we choose to use the more parsimonious random walk with drift model.

2.2 *The Tucker method*

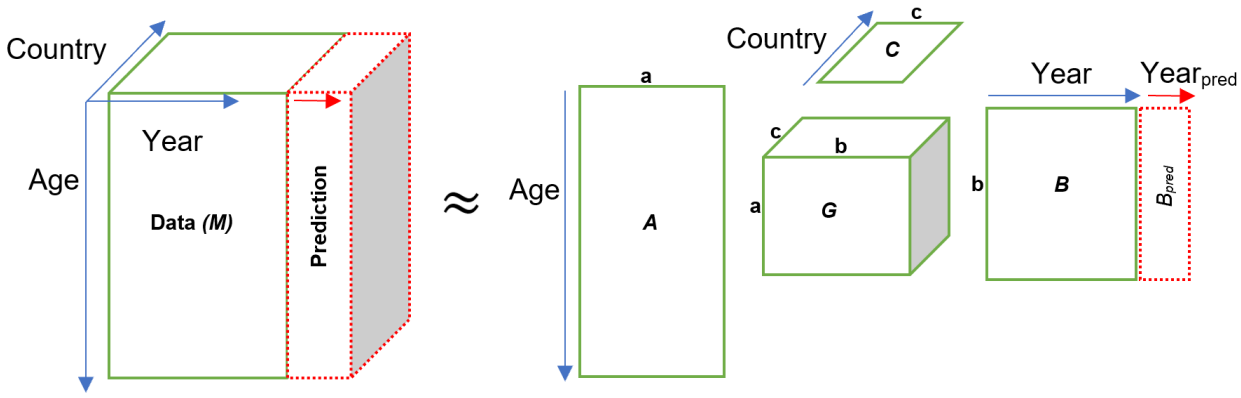

Figure 2: Tucker method

The Tucker decomposition is a form of higher-order principal component analysis (PCA). It decomposes a tensor into a core tensor (G) multiplied by a matrix (A, B, C) along each mode. A, B and C correspond to the age, year and country matrices, respectively. The algorithm that we use to compute the Tucker decomposition is the higher order orthogonal iteration (HOOI), which is essentially an ALS algorithm as discussed in Section 2.1. For more details of this algorithm, please refer to the supplementary materials. In the three-way case, the Tucker decomposition can be written as,

$$M \approx \hat{M} = G \times_1 A \times_2 B \times_3 C \tag{6}$$

In this setting, $G \in \mathbb{R}^{a \times b \times c}$ is the core tensor, which expresses how the principal components interact with each other. The factor matrices $A \in \mathbb{R}^{X \times a}$, $B \in \mathbb{R}^{T \times b}$, and $C \in \mathbb{R}^{H \times c}$ are often referred to as the principal component in the respective tensor mode. The vectors within each factor matrices are orthogonal. The a, b, c are the ranks of A (age), B (year) and C (country) respectively. The n-mode product of a tensor $G \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$ and a matrix $X \in \mathbb{R}^{J \times I_n}$ is denoted by $Y = G \times_n X$ with $Y \in \mathbb{R}^{I_1 \times \ldots I_{n-1} \times J \times I_{n+1} \ldots \times I_N}$. Elementwise, we have

$$m_{x,t,h} \approx \hat{m}_{x,t,h} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{q=1}^{c} g_{i,j,q} a_{x,i} b_{t,j} c_{h,q}, \tag{7}$$

where $g_{i,j,q}$, $a_{x,i}$, $b_{t,j}$ and $c_{h,q}$ are elements of G, A, B and C, respectively. If a, b, c (ranks) are smaller than X, H and T, the core tensor G can be regarded as a compressed version of M. In this paper, we find the core tensor G significantly smaller than the original tensor M.

The predicted values can be calculated by,

$$M_{pred} = G \times_1 A \times_2 B \times_3 C_{pred} \tag{8}$$

Tucker is more flexible than CPD as there are 3 hyper-parameters (a, b, c) to consider. In other words, the rank of each dimension (a, b, and c) can be different from each other. In fact,

the CPD can be viewed as a special case of Tucker where the core tensor is superdiagonal and $a = b = c$. In order to investigate the best Tucker model for mortality forecasting, we consider 3 forms of the Tucker models:

- The first form is called the Tucker (aaa), which is similar to the CPD because the constraint is $a = b = c$. Russolillo et al. (2011) considered Tucker (222) in their paper.
- The second form is called the Tucker (aab), whose constraint is $a = b$. We consider this form because the lengths of the Age and the Year are comparable in this research.
- The last form is the Tucker (abc), which is the most flexible form among the three.

We will investigate how the flexibility of the Tucker models affects their predictive ability.

Similar to the CPD method, the Tucker method usually does not provide unique solutions. So we apply the same process mentioned in the Section 2.1 with 50 random initializations to select the optimal fit[3]. Then we fit a random walk with drift model to each latent year vector of the matrix B in Figure 2. Combining each predicted year vectors with the core tensor G, age matrix A and country matrix C, we can obtain the predicted M_{pred} following Equation (8).

The latent year vectors in the Tucker models are orthogonal and uncorrelated, which, however, is not the case for CPD models. To make the model structures consistent and parsimonious, we apply the univariate RMD models for all latent year vectors considered in the paper to capture the dynamic patterns. Additionally, fitting a multivariate time series model for the latent year vectors, such as vector autoregressive moving average (VARMA), is not appealing in this case, as there are usually more than 3 latent factors. VARMA models are seldom used directly in practice when the dimension is more than 3. This is partially due to the lack of identifiability for VARMA models in general (Chang et al. 2018). On the other hand, we did not fit different ARIMA models on different latent year vectors, as the final model would be too complex.

2.3 Modelling steps

We summarise the modelling process by the following 5 steps which are feasible for both the CPD method and the Tucker method. For brevity, we only use the CPD method as an example to illustrate the modelling steps. We use the Matlab Tensor Toolbox by Bader et al. (2017) to do the tensor decomposition.

Step 1: Centering the Data We conduct the log transformation of the central death rates and organise the data in a 3-dimensional way. Then we center the log-mortalities by subtracting the mean of each row from the 3-dimensional data, see Equation (1). After centering the data, we name the cube as the target tensor, see Figure 1. The main objective of the transformation is to reduce the data complexity. We divide the target tensor chronologically into three parts, which are the training set, the validation set and the testing set, see Figure 3.

[3]The required number of random initializations of Tucker method is much smaller than that of the CPD method, which leads to faster computation applying the Tucker method.

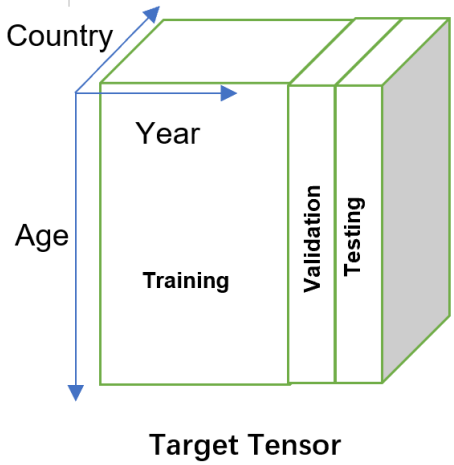

Figure 3: Organising the mean centered log-mortality in 3-dimension

Step 2: Selecting the Rank The rank of a tensor decomposition is the only hyper-parameter in the model, which is selected using the out-of-sample validation. Out-of-sample validation is a widespread strategy because of its simplicity and its (apparent) universality, see Arlot et al. (2010). The goal of out-of-sample validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like over-fitting or selection bias (Cawley & Talbot 2010); and to give insight into how the model will generalise to an independent dataset.

The procedures for using out-of-sample validation to select ranks are as follows. We firstly apply tensor decomposition on the training set with rank numbers from 1 to 10 [4], see Equation (3). We then forecast the 10 models with different ranks in the time period of the validation set, see Figure 4 as an example of rank n and Equation (5). By comparing the root mean square forecasting error (RMSFE [5]) of the 10 models in the validation set, we select the model with the optimal rank, which has the lowest RMSFE. The selected rank will be used in later steps.

Step 3: Model Fitting In this step, we combine the training set and the validation set together as a new training set for model fitting. Assume we have selected the rank n of the tensor decomposition in Step 2. We then apply the tensor decomposition model to the new training set with the selected rank, see Figure 5 as an example of the rank n.

[4] In this paper, we set the largest possible rank to be 10 for the CPD, which is the length of the country vector. For the Tucker method, the largest theoretical ranks (a, b and c) are the lengths of the age vector, year vector and country vector respectively. To avoid the Tucker model being too complex, we set the maximum values of a and b to be 15.

[5] The formula for RMSFE is the same as that for RMSE, except that for RMSFE, $\hat{\log}\mu_{x,t,h}$ denotes the predicted log-mortality.

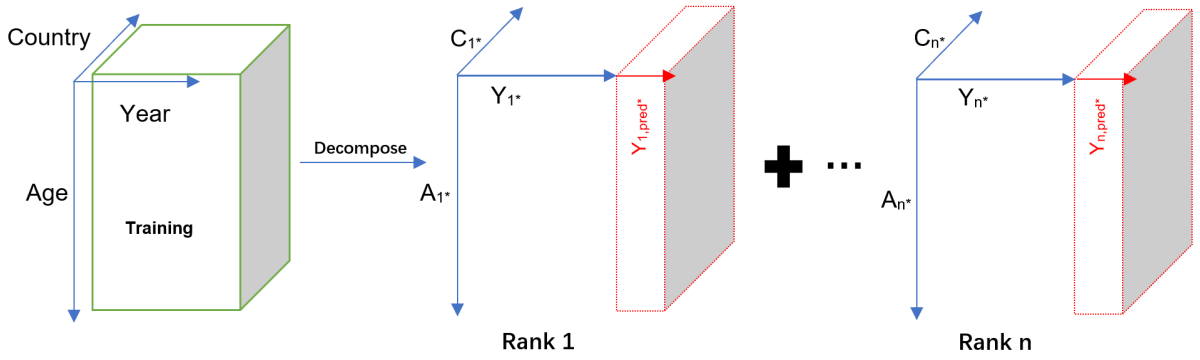

Figure 4: Constructing the rank-n forecasts in the validation set

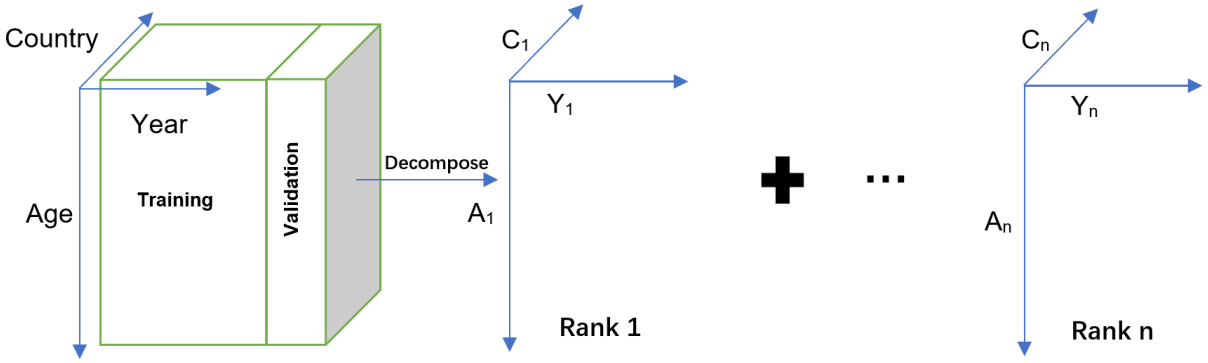

Figure 5: Decomposing the new training set into the determined rank decomposition

Step 4: Model Forecasting Given the model fitting results obtained in Step 3, we apply the random walk with drift to the fitted year vectors Y_i and get the predicted values $Y_{i,pred}$ in the forecasting time horizon. Combining the predicted year vectors $Y_{i,pred}$ with the fitted country vectors C_i and age vectors A_i, we obtain the predicted M_{pred} following Equation (5) in the forecasting time horizon, see Figure 6 as an example of the rank n. We can also add uncertainties using the random walk with drift forecasts.

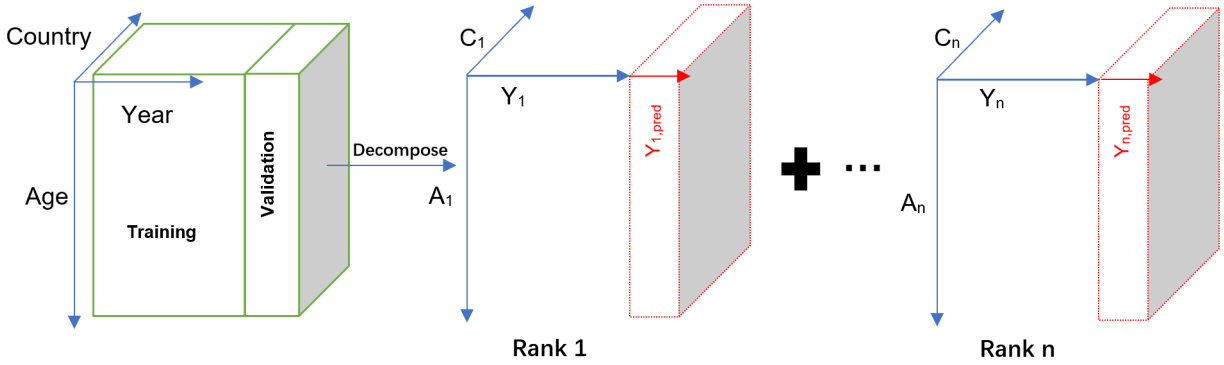

Figure 6: Constructing the rank-n model forecasts

Step 5: Model Testing By forecasting the model in the time period of the testing set, we can obtain the RMSFE of the testing set, which can be used to compare the out-of-sample forecasting performance of different models.

We can obtain unique solutions of all parameters through the above 5 steps. Since the target of this paper is to optimise the final forecasting performance rather than uncovering the "true" data generating process, we choose not to add constraints in the fitting process to get identifiable solutions, as those constraints could potentially lead to unoptimised forecasting results. However, this decision leads to the difficulty in interpreting the latent factors of the model, which is a common issue in machine learning practice: the trade-off between model flexibility and interpretability.

3 Results

We perform our analysis using data obtained from the Human Mortality Database, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at `www.mortality.org` (Human Mortality Database 2018).

3.1 10 European countries

Russolillo et al. (2011) applied Tucker (2,2,2) model for 10 European countries, which are geographically closer and share similar social-economic status. Similarly, we also consider 10 European countries[6] as an example for modelling and comparison purposes. They are Denmark, United Kingdom, Finland, France, Italy, the Netherlands, Norway, Spain, Sweden, Switzerland, which are summarised in Table 1. We consider the ages from 0 to 90 for modelling purposes to avoid the sparse and noisy data of the older ages. However, the tensor decomposition models do not require similar mortality experience among different populations, which can be applied to other more different populations too.

[6]Note that 8 out of the 10 countries are the same as those used in Russolillo et al. (2011).

Table 1: Data

	Starting time	Ending time
Denmark	1835	2016
United Kingdom	1922	2016
Finland	1878	2015
France	1816	2016
Italy	1872	2014
Netherlands	1850	2016
Norway	1840	2014
Spain	1908	2014
Sweden	1751	2016
Switzerland	1876	2016

The allocation of countries to positions along the Country axis were determined according to the order of countries listed in the first column of Table 1 in the paper. A change of the allocation would affect the decomposition results of country vectors for each rank. However, it would only affect the prediction result a little. From the theoretical and mathematical perspective, if we could find the global optimum (which may be difficult to find practically), then the decomposed results would not change if we changed the order of countries [7]. However, in reality tensor decomposition is a non-convex problem, which can only achieve a local optimum at each run. If we changed the order of countries, the decomposition results would change (getting a different local optimum). However, as explained in Sections 2.1 and 2.2, by running the model multiple times with random initialisations, we can obtain a stable "optimal" fit, which can be regarded as very close to the global optimum. So the final forecasting results (RMSFE) would only change a little if there were a change of the country order, which should not be a big concern nor a reasonable way to improve the out-of-sample forecasting performance.

Three target tensors are analysed in this paper, which are the total population tensor, the male tensor and the female tensor. We consider four forecasting horizons: short-term (5 years), mid-term (10 yeas), long-term (20 years) and extra-long-term (30 years) forecasts. We also consider two time frames. Time frame A (Table 2) includes the Second World War period (1939 to 1945) and Time frame B excludes the Second World War period. The length of validation sets and testing sets are the same in all cases. Due to the shorter period of Time frame B, we will not do an extra-long-term forecast in this frame.

[7] For example, if we changed the order of the first and the second country of a CPD model, then the i-th rank country vector would be changed to $C_i = (c_2^{(i)}, c_1^{(i)}, ..., c_h^{(i)})$. However, the prediction for the first country and the second country would remain the same (they just changed their seats). Similarly, for the Tucker method, changing the order or countries would only change the order of the country vectors in matrix C accordingly, see Figure 2 and Equation (7).

Table 2: Time frames A & B

	Time frame A		
	Training	**Validation**	**Testing**
Short-term (5 Year)	1922-2004	2005-2009	2010-2014
Mid-term (10 Year)	1922-1994	1995-2004	2005-2014
Long-term (20 Year)	1922-1974	1975-1994	1995-2014
Extra-long (30 Year)	1922-1954	1955-1984	1985-2014
	Time frame B		
Short-term (5 Year)	1950-2004	2005-2009	2010-2014
Mid-term (10 Year)	1950-1994	1995-2004	2005-2014
Long-term (20 Year)	1950-1974	1975-1994	1995-2014

Table 3 shows the results of Step 2 (Selecting the Rank). We can see that, holding other conditions the same, ranks determined in Frame A are larger than those in Frame B upon most occasions. It suggests that a simpler data frame often requires a simpler model. For the Tucker (abc) method, a (rank of Age) is always smaller or equal to b (rank of Year). It suggests that the data structure of the Age dimension is easier to be captured and fitted than that of the Year dimension. c (rank of Country) is close to its upper bound (10) for most cases. It suggests that the data structures among 10 countries are so different that almost all ranks are needed to fit the model. The a (rank of Age) of Tucker(aaa) determined by out-of-sample validation method in this paper is much larger than that determined by variance-explained method used by Russolillo et al. (2011), who selected Tucker(222) in their paper. We treat the Tucker(222) model as one of the benchmarks and we will show how differently models perform in the testing set.

Table 3: The ranks determined in the validation set (Frame A & B)

	Time frame A			
	CPD	**Tucker (aaa)**	**Tucker (aab)**	**Tucker (abc)**
Short-term(5 Year)	9	10*10*10	11*11*10	10*14*10
Mid-term(10 Year)	7	10*10*10	8*8*10	6*8*10
Long-term(20 Year)	6	10*10*10	14*14*10	10*14*10
Extra-long(30 Year)	4	9*9*9	11*11*9	8*14*9
	Time frame B			
Short-term(5 Year)	10	8*8*8	8*8*10	6*8*9
Mid-term(10 Year)	4	4*4*4	3*3*2	3*3*2
Long-term(20 Year)	8	5*5*5	5*5*4	5*9*8

Table 4 shows the total number of parameters of each fitted model in Frames A and B respectively. We compare all the tensor decomposition models with independent mortality modelling of each population using the SVD (Lee-Carter) model. The number of parameters of the CPD model is always less than that of the SVD (Lee-Carter) model except for short-term (5 years) forecast in Frame B. Tucker models usually have more parameters to estimate in Frame A due

to the core tensor needed in the Tucker decomposition, see Figure 2. However, the number of parameters of Tucker models becomes smaller in Frame B compared with that in Frame A. The reason is that the ranks (a, b ,c) determined in Frame B are much smaller due to the less complicated dataset.

Table 4: The total number of parameters for each determined model in Frame A & B.

	Time frame A				
	SVD	**CPD (n)**	**Tucker (aaa)**	**Tucker (aab)**	**Tucker (abc)**
Short-term(5 Year)	2650	2566	3750	4134	4482
Mid-term(10 Year)	2550	2128	3650	2962	2620
Long-term(20 Year)	2350	1834	3450	4986	4062
Extra-long(30 Year)	2150	1446	2845	3453	3198
	Time frame B				
Short-term(5 Year)	2370	2470	2670	2818	2418
Mid-term(10 Year)	2270	1494	1558	1356	1356
Long-term(20 Year)	2070	1918	1665	1630	2030

When forecasting the fitted year vectors Y_i, we apply the random walk with drift process. Table 5 to Table 6 show the out-of-sample forecasting performance of the CPD method and 3 forms of the Tucker method (measured by RMSFE) on the testing data set of Frame A and Frame B. Table 5 also shows the performance of SVD (Lee-Carter model) on individual populations and the improvement ratios of the multi-population Tensor method (CPD & Tucker) compared with the single-population SVD method. Table 6 summarizes the overall forecasting performance for 10 countries in different time frames. The improvement ratio is defined as follows:

$$Improvement\ ratio = \frac{RMSFE_{SVD} - RMSFE_{Tensor}}{RMSFE_{SVD}} \times 100\%$$

Positive improvement ratios indicate a better forecasting performance of the Tensor method compared with SVD (Lee-Carter) method. Generally speaking, the tensor decomposition method works better when the data set is bigger and has a more complex structure. The results show that the 3-way decomposition (Tensor) on multi-population mortality outperforms the 2-way decomposition (SVD) on single-population mortality significantly both for individual countries and the aggregate populations. We consider three aggregate populations: males, females and total population (including both genders). And the results are consistent for all three cases. In this paper, we show the results of the total population (including both genders) as an example. The improvement in Frame A is more significant than that in Frame B, as shown by Table 6. The reason is that Frame A includes the World War II data and has a more complicated mortality structure which is difficult for the SVD model to fit.

In Table 6, we also compare the forecasting performance of the Tensor methods, namely the CPD method and the 3 forms of the Tucker method. In Frame A, the forecasting performance of these 4 tensor models are comparable while none of them shows significantly better performance than the others. However, the Tucker(222) model performs less well than others consistently. It suggests that selecting the ranks for Tensor model based on the variance-explained

method can be improved by the out-of-sample validation method. All the total improvement ratios of the Tensor model (excluding Tucker(222)) are in the range of 20% to 30% and the country-specific improvement ratios vary in each scenario. In Frame B, the CPD method shows more stable forecasting performance than the 3 forms of the Tucker methods especially in the mid-term and long-term forecasts. More results for each time frame and forecasting horizon can be found in the supplementary materials.

From Table 5, we find that Switzerland, Denmark, Finland and Norway are the countries with relatively high RMSFEs when using the Tensor method, which could be caused by both the individual effect (measured by the individual SVD model) and the group effect (not suitable for grouping)[8]. The forecast results of the single-population SVD model show that the RMSFEs of Denmark, Finland and Norway are relatively high, which can be used to approximate the individual effect. However, the improvement ratios of Switzerland and Denmark are relatively low and are negative for some cases (which indicates the SVD model performs better in those cases); while those of Finland and Norway are relatively high. This finding indicates that Switzerland and Denmark are not benefiting from the grouping as much as Finland and Norway in Frame A, which sheds lights on the similarity of different populations. It also provides insights into the clustering of mortality among different populations. Zhu et al. (2017) also find similar results when using a new classification method for 12 European countries. According to their results, Belgium and Switzerland show the most similar mortality patterns with each other and are different from the other European countries. However, Belgium is not considered in this research.

Table 5: Out-of-sample Short-term forecasting performance, Europe (Frame A)

	SVD	**CPD**	**Imp.%**	**T(222)**	**Imp.%**	**T(aaa)**	**Imp.%**	**T(aab)**	**Imp.%**	**T(abc)**	**Imp.%**
Swe.	0.273	0.201	26.5%	0.212	22.4%	0.205	24.9%	0.199	27.0%	0.200	26.7%
U.K.	0.247	0.141	42.9%	0.162	34.5%	0.103	58.4%	0.102	58.5%	0.094	61.8%
Swi.	0.229	0.234	-2.5%	0.305	-33.3%	0.219	4.0%	0.225	1.7%	0.220	3.6%
Den.	0.282	0.291	-3.1%	0.330	-17.0%	0.269	4.6%	0.285	-0.8%	0.267	5.2%
Spa.	0.186	0.151	19.2%	0.231	-24.3%	0.134	28.1%	0.116	37.5%	0.123	33.9%
Fin.	0.356	0.249	30.2%	0.270	24.2%	0.244	31.5%	0.247	30.6%	0.263	26.2%
Ita.	0.193	0.112	42.0%	0.157	-18.6%	0.110	42.9%	0.108	43.9%	0.114	40.9%
Net.	0.172	0.131	23.7%	0.154	10.7%	0.126	26.6%	0.124	27.8%	0.122	29.0%
Nor.	0.334	0.247	26.1%	0.251	24.9%	0.248	26.0%	0.246	26.5%	0.297	11.1%
Fra.	0.146	0.107	26.6%	0.174	-18.8%	0.097	33.8%	0.083	42.9%	0.087	40.4%
Total	0.251	0.197	21.6%	0.233	7.3%	0.187	25.5%	0.187	25.3%	0.194	22.6%

[8]Switzerland, Denmark, Finland and Norway are the four countries with the smallest sizes of populations among the 10 European countries considered for modelling, which could also be one of the reasons of the differences shown in the results.

Table 6: Out-of-sample overall forecasting performance, Europe (Frame A & B)

	Time frame A										
	SVD	**CPD**	**Imp.%**	**T(222)**	**Imp.%**	**T(aaa)**	**Imp.%**	**T(aab)**	**Imp.%**	**T(abc)**	**Imp.%**
Short	0.251	0.197	21.6%	0.233	7.3%	0.187	25.5%	0.187	25.3%	0.194	22.6%
Mid	0.269	0.214	20.4%	0.233	13.5%	0.197	26.7%	0.201	25.3%	0.199	26.1%
Long	0.305	0.226	25.9%	0.233	23.8%	0.232	23.9%	0.239	21.8%	0.236	22.6%
Extra	0.353	0.267	24.3%	0.298	15.6%	0.256	27.4%	0.257	27.0%	0.252	28.6%
	Time frame B										
Short	0.232	0.193	17.0%	0.239	-3.0%	0.201	13.4%	0.202	13.0%	0.192	17.4%
Mid	0.247	0.217	12.2%	0.244	1.2%	0.214	13.6%	0.227	8.4%	0.227	8.4%
Long	0.268	0.259	3.4%	0.252	6.0%	0.255	4.9%	0.251	6.6%	0.259	3.6%

Next, we show some plots in Frame A to illustrate the forecasting performance of the tensor decomposition method. Figures 7 and 8 illustrate the age-specific RMSFEs of 5-year and 30-year forecasting horizons for the Lee-Carter (SVD) model and the tensor decomposition (CPD & Tucker) models. The gap between the red line (SVD) and other lines (CPD & Tucker) in each plot can be explained by the "improvement ratio", which we have displayed earlier.

- The RMSFE of all models at young ages below 20 are relatively high. Tensor decomposition models (CPD & Tucker) are not very stable at younger ages below 10. For example, we can see peaks of the Tucker(abc) in the Figure 7 at age 8 and the Tucker(222) in the Figure 8 at age 1. The reason is that the model determined by the out-of-sample method minimizes the forecasting error in the validation set rather than the testing set. The algorithm of the tensor decomposition could reach to a scenario-specific local optimum for the validation set and causes the overfitting. To fix the "peak" problem for specific young ages, we can change the age range or time frame to be modelled. For example, we show the results without the "peak" problem with different age ranges and time frames in the supplementary materials.

- The tensor decomposition models (CPD & Tucker) perform significantly better than the Lee-Carter model for ages between 20 and 30.

- All models show similar forecasting performances and have relatively low RMSFE for middle ages between 35 and 45.

- The tensor decomposition models (CPD & Tucker) significantly outperform the Lee-Carter model again for older ages after 50.

- The RMSFE curves of 3 Tucker models share similar patterns which are different from those of the CPD to some extent.

- The Tucker(222) performs less well than the CPD and the 3 forms of Tucker model in most cases. The plots intuitively show how forecasting performance is improved by using the out-of-sample method instead of the varianced-explained method to select the rank parameter.

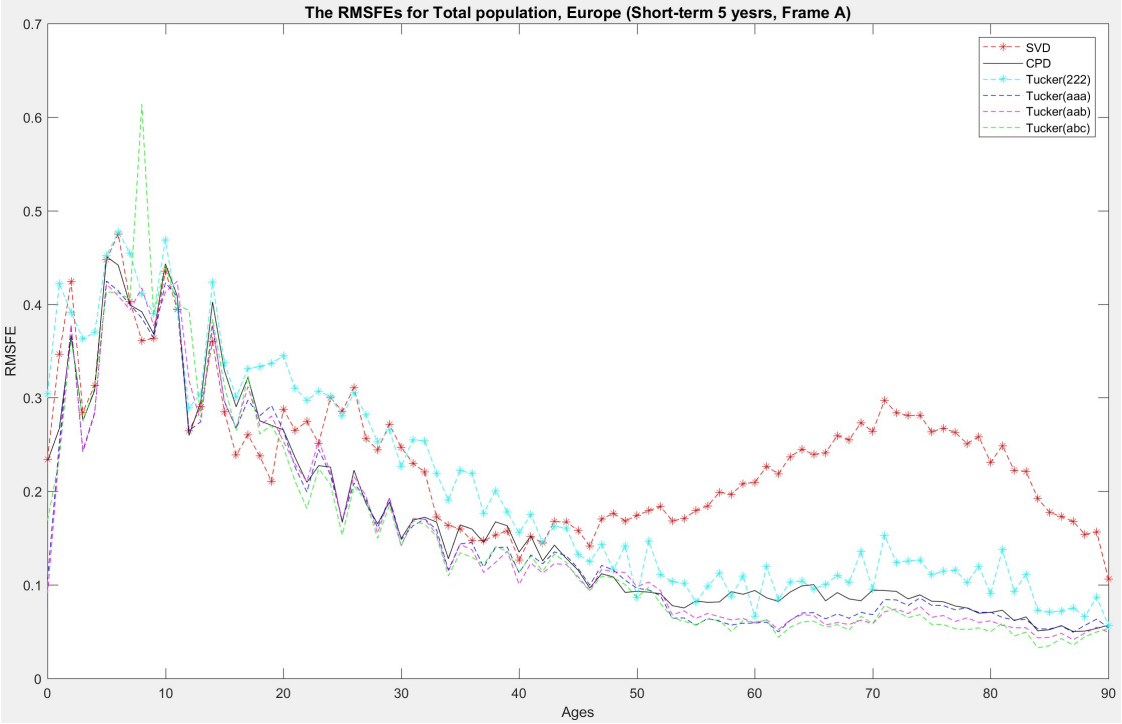

Figure 7: Age-specific RMSFE for short-term forecasting: 5 years

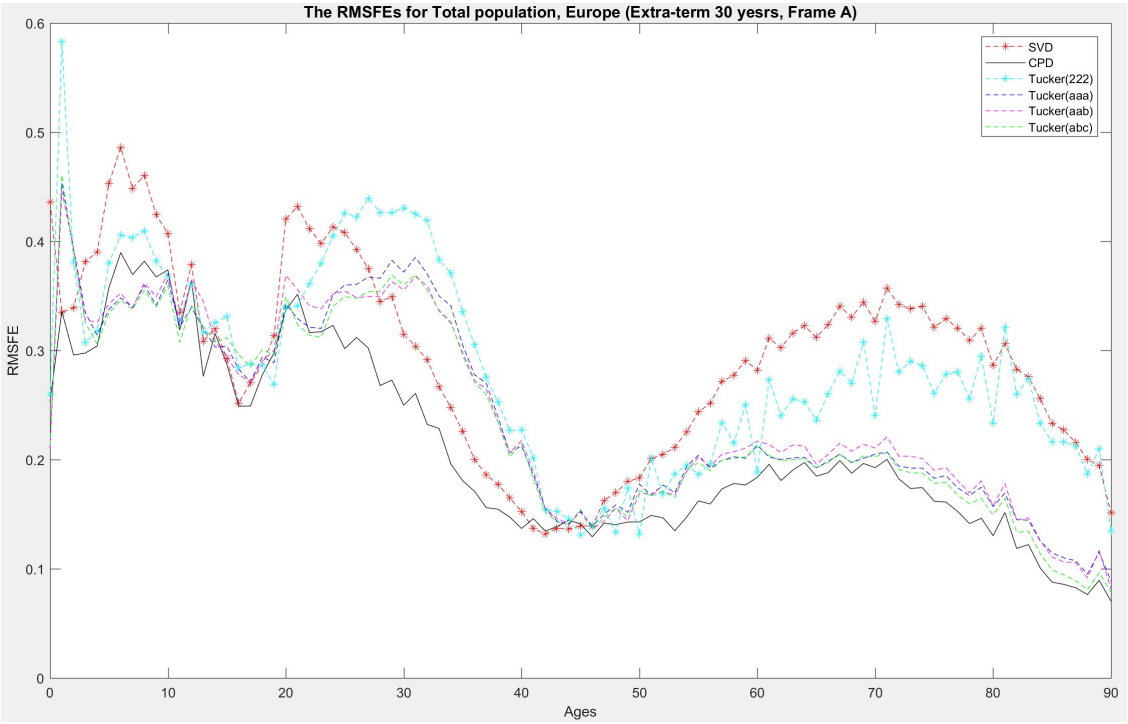

Figure 8: Age-specific RMSFE for extra long-term forecasting: 30 years

3.2 Two genders

In this section, we apply the 3-way tensor decomposition to a population of two genders. The 3 dimensions are Age, Year, and Gender. The mortality data of the USA and UK are considered in this section. We apply ages 20 to 90 in this section to avoid the sparse and noisy data of the younger and older ages.

3.2.1 USA

The time periods considered are 1933 to 2016 for Frame A (including the WWII) and 1950 to 2016 for Frame B (excluding the WWII). Similar with the previous analysis, we divide the entire data set into training, validation and testing sets, which are summarised in Table 7 for Frame A and Frame B respectively.

Table 7: Time frames A & B

	Time frame A		
	Training	**Validation**	**Testing**
Short-term (5 Year)	1933-2006	2007-2011	2012-2016
Mid-term (10 Year)	1933-1996	1997-2006	2007-2016
Long-term (20 Year)	1933-1976	1977-1996	1997-2016
	Time frame B		
Short-term (5 Year)	1950-2006	2007-2011	2012-2016
Mid-term (10 Year)	1950-1996	1997-2006	2007-2016
Long-term (20 Year)	1950-1976	1977-1996	1997-2016

The ranks selected using out-of-sample validation for each forecasting horizon are presented in Table 8.

Table 8: The ranks determined in the validation set, USA (Frame A & B)

	Time frame A			
	CPD	**Tucker (aaa)**	**Tucker (aab)**	**Tucker (abc)**
Short-term(5 Year)	8	2*2*2	14*14*2	12*15*2
Mid-term(10 Year)	9	2*2*2	9*9*2	5*9*2
Long-term(20 Year)	10	2*2*2	11*11*1	7*15*1
	Time frame B			
Short-term(5 Year)	10	2*2*2	11*11*2	10*12*2
Mid-term(10 Year)	7	1*1*1	7*7*2	6*10*2
Long-term(20 Year)	5	2*2*2	3*3*2	10*7*2

Table 9 shows the out-of-sample short-term forecasting performance of the tensor decomposition method compared with the SVD method in Frame A. Table 10 summarizes the overall

forecasting performance in Frame A and Frame B. Both the SVD model and Tensor models have lower RMSFE in Frame B than in Frame A. The two-population CPD model consistently outperforms the SVD model for both individual genders and the total population. Tucker (abc) outperforms the other two Tucker models. And the Tucker methods do not perform well over the long-term horizon, especially in Frame A. It also shows that Tucker(aaa) is not a suitable model in this case since the upper bound of the c (rank of gender) is 2. However, rank 2 is not enough to capture the Year information (b) or the Age information (a). Due to the constraint of rank in Tucker(aaa) model, it is only suitable to use when the rank of the 3 dimensions are in the similar scale.

Table 9: Out-of-sample Short-term forecasting performance, USA (Frame A)

	SVD	**CPD**	**Imp.%**	**Tucker(aaa)**	**Imp.%**	**Tucker(aab)**	**Imp.%**	**Tucker(abc)**	**Imp.%**
Female	0.166	0.097	41.7%	0.129	22.5%	0.094	43.0%	0.094	43.4%
Male	0.148	0.088	40.5%	0.081	45.0%	0.087	40.9%	0.088	40.8%
Total	0.157	0.092	41.2%	0.108	31.5%	0.091	42.1%	0.091	42.2%

Table 10: Out-of-sample overall forecasting performance, USA (Frame A & B)

	Time frame A								
	SVD	**CPD**	**Imp.%**	**Tucker(aaa)**	**Imp.%**	**Tucker(aab)**	**Imp.%**	**Tucker(abc)**	**Imp.%**
Short	0.157	0.092	41.2%	0.108	31.5%	0.091	42.1%	0.091	42.2%
Mid	0.158	0.096	39.7%	0.117	26.4%	0.093	41.6%	0.096	39.6%
Long	0.187	0.151	19.5%	0.181	3.1%	0.194	-3.7%	0.194	-3.7%
	Time frame B								
Short	0.107	0.077	28.2%	0.103	4.3%	0.079	26.4%	0.079	26.6%
Mid	0.105	0.080	24.2%	0.117	-11.4%	0.081	22.9%	0.079	25.4%
Long	0.136	0.121	10.5%	0.154	-13.3%	0.136	-0.5%	0.127	6.3%

Figure 9 illustrates the age-specific root mean square forecasting errors over 10-year forecasting horizon of the SVD model and the tensor decomposition (CPD & Tucker) models in Frame A. The comparative performance of the Tensor methods on different ages is similar with the previous analysis for 10 European countries, except for Tucker(aaa) due to its constraints on rank selection.

In conclusion, for the US tensor data with age, year and gender as the three dimensions, we recommend the use of the CPD model due to its consistent good performance in two time frames (A and B). For mortality forecasting with short-term (5 years) and mid-term (10 years) horizons, Tucker(aab) and Tucker(abc) can also be considered.

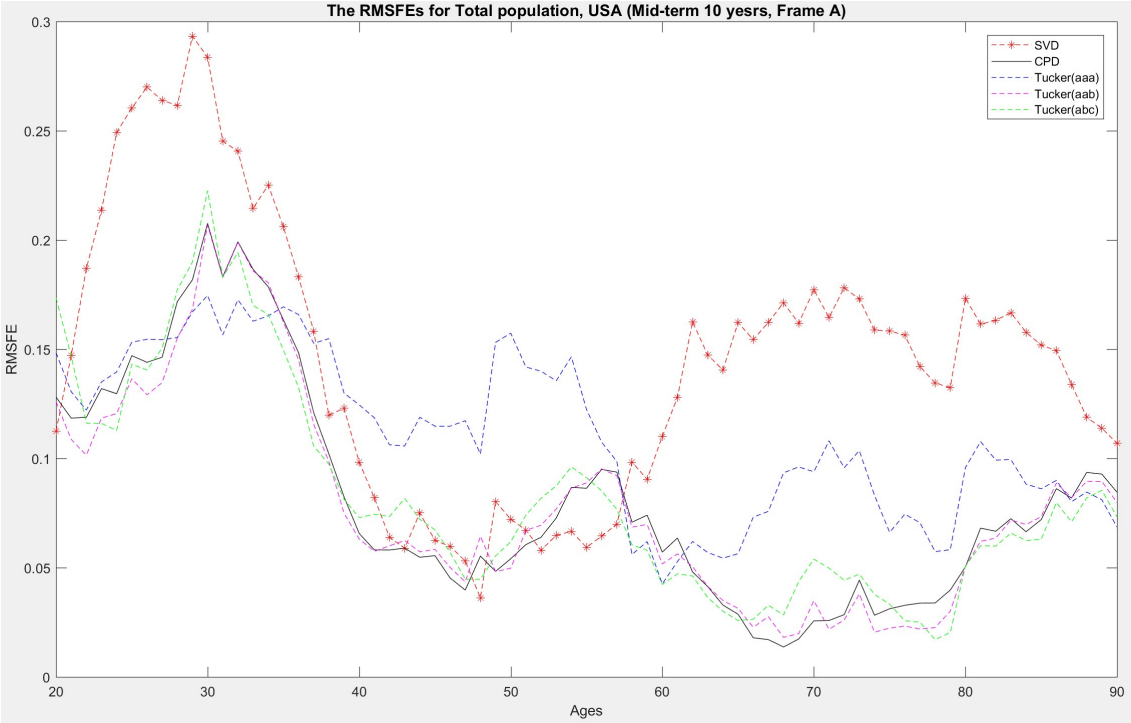

Figure 9: Age-specific RMSE for mid-term forecasting: 10 years

3.2.2 UK

The time periods considered using the UK data are 1922 to 2016 for Frame A (including the WWII) and 1950 to 2016 for Frame B (excluding the WWII). Similar with the previous analysis, we divide the entire data set into training, validation and testing sets, which are summarised in Table 11 for Frame A and Frame B.

Table 11: Time frames A & B

	Time frame A		
	Training	**Validation**	**Testing**
Short-term (5 Year)	1922-2006	2007-2011	2012-2016
Mid-term (10 Year)	1922-1996	1997-2006	2007-2016
Long-term (20 Year)	1922-1976	1977-1996	1997-2016
	Time frame B		
Short-term (5 Year)	1950-2006	2007-2011	2012-2016
Mid-term (10 Year)	1950-1996	1997-2006	2007-2016
Long-term (20 Year)	1950-1976	1977-1996	1997-2016

The ranks selected using out-of-sample validation for each forecasting horizon are presented in Table 12.

Table 12: The ranks determined in the validation set, UK (Frame A & B)

	Time frame A			
	CPD	**Tucker (aaa)**	**Tucker (aab)**	**Tucker (abc)**
Short-term(5 Year)	7	2*2*2	9*9*2	7*12*2
Mid-term(10 Year)	10	2*2*2	11*11*2	4*7*2
Long-term(20 Year)	4	2*2*2	3*3*1	3*7*1
	Time frame B			
Short-term(5 Year)	7	2*2*2	6*6*2	6*7*2
Mid-term(10 Year)	5	2*2*2	5*5*2	5*5*2
Long-term(20 Year)	7	2*2*2	7*7*2	10*5*2

Table 13 shows the out-of-sample short-term forecasting performance of the tensor decomposition method compared with the SVD method in Frame A. Table 14 summarizes the overall forecasting performance in Frame A and Frame B. The tensor decomposition models (CPD & Tucker) consistently outperform the SVD model significantly for both individual genders and the total population, except for Tucker(aaa), whose improvement ratios are relatively smaller in some cases.

Table 13: Out-of-sample Short-term forecasting performance, UK (Frame A)

	SVD	**CPD**	**Imp.%**	**Tucker(aaa)**	**Imp.%**	**Tucker(aab)**	**Imp.%**	**Tucker(abc)**	**Imp.%**
Female	0.232	0.070	69.6%	0.120	48.4%	0.068	70.6%	0.068	70.8%
Male	0.311	0.060	80.6%	0.114	63.4%	0.064	79.3%	0.060	80.8%
Total	0.274	0.066	76.0%	0.117	57.4%	0.066	75.9%	0.064	76.7%

Table 14: Out-of-sample overall forecasting performance, UK (Frame A & B)

	Time frame A								
	SVD	**CPD**	**Imp.%**	**Tucker(aaa)**	**Imp.%**	**Tucker(aab)**	**Imp.%**	**Tucker(abc)**	**Imp.%**
Short	0.274	0.066	76.0%	0.117	57.4%	0.066	75.9%	0.064	76.7%
Mid	0.320	0.093	70.9%	0.137	57.0%	0.093	70.8%	0.091	71.5%
Long	0.358	0.181	49.5%	0.193	46.1%	0.196	45.1%	0.196	45.1%
	Time frame B								
Short	0.117	0.062	46.9%	0.111	4.4%	0.063	45.7%	0.062	47.0%
Mid	0.146	0.092	36.8%	0.129	11.6%	0.089	39.1%	0.089	39.1%
Long	0.209	0.163	22.3%	0.174	17.1%	0.163	22.2%	0.171	18.5%

Similar to the conclusion in Section 3.2.1 for the USA data, we recommend the use of the CPD model for UK data too. For the short-term (5 years) and mid-term (10 years) forecasting horizons, Tucker(aab) and Tucker(abc) can also be considered.

Figure 10 illustrates the age-specific root mean square forecasting errors over 10-year forecasting horizon of the SVD model and the tensor decomposition (CPD & Tucker) models. The

comparative performance of the tensor methods on different ages is similar with the previous analysis for the US. The gaps between the SVD curve and Tensor curves are big in all forecasting periods (except for Tucker(aaa) due to its rank constraints), which implies high improvement ratios.

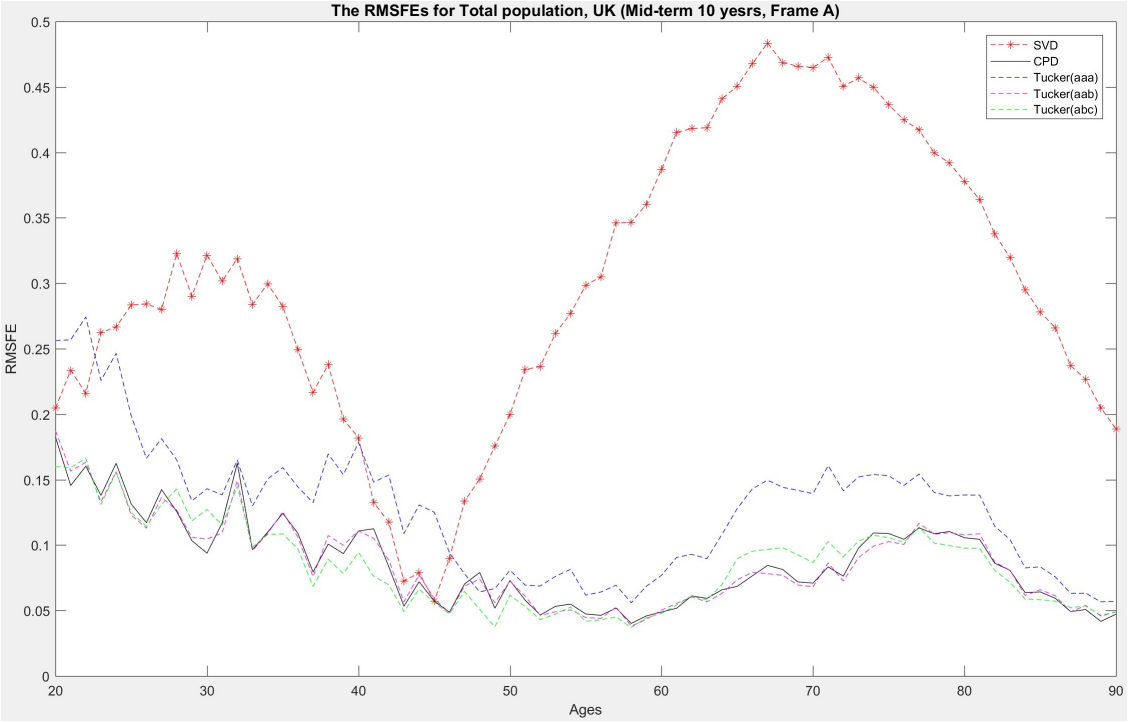

Figure 10: Age-specific RMSE for mid-term forecasting: 10 years

Comparing the result of UK with that of USA, the improvement ratios in the UK are consistently and significantly higher than those in the USA. The most obvious reason is that the SVD (Lee-Carter) model performs worse on the UK data while the Tensor models perform consistently well on the data of both countries. This phenomenon possibly reflects the well-known cohort effect that has been identified in the trends, see Renshaw & Haberman (2006).

We also compare our results with the findings discovered by Tsai & Lin (2017). They propose a new 2-dimensional model to forecast the gender-specific mortality rates of U.K. and U.S. for the mid-term (10 years), long-term (20 years) and extra-long-term (30 years). The improvement ratios generated by their model differ significantly between male population and female population in both the US and UK data (see Table 6 and Table 7 in their paper). However, the CPD model and Tucker(abc) model show similar forecasting performance on both genders (The improvement ratios of males and females are close to each other.). Additionaly, the CPD model and Tucker (abc) model also show much better forecasting performance compared with the models in Tsai & Lin (2017). This comparison result again implies that tensor decomposition models can utilise the common information between different populations and improve the forecast of individual populations.

3.3 Variance-explained method v.s. Out-of-sample validation for rank selection

In this section, we focus on the comparison of using the varianced explained method and out-of-sample validation method for rank selection. The Lee & Carter (1992) model is in fact a form of rank-1 SVD model (SVD(1)). And they used the variance-explained method to select the rank, as the first rank is able to explain 97.5% of variance over time in all the age groups lower than 85 years old. We can also see similar arguments in many other papers when selecting the rank. Following the same method, we calculate the variance-explained ratios using the rank-1 SVD model on both US and UK data sets under the two frames in this paper. In Table 15, the SVD(1) model is able to explain 94.04% variance on average in all scenarios. As we have shown in the previous section, although they have similar ability in explaining the variance of the US and UK data, the forecasting performance is significantly different.

Table 15: Variance-explained ratios by SVD(1) model on UK and US, Frame A, Training set

	US(female)	**US(male)**	**UK(female)**	**UK(male)**
Short-term(5 Year)	97.10%	89.21%	96.65%	87.95%
Mid-term(10 Year)	97.02%	85.67%	97.44%	92.15%
Long-term(20 Year)	98.03%	94.22%	97.59%	95.42%

Dimension-reduction techniques such as SVD and PCA usually require researchers to determine the number of components or factors to retain (Osborne & Costello 2004). One of the most widely used methods for this purpose is Cattell's scree test (Cattell 1966). The scree test is a heuristic graphic method that consists of :

- Plotting the eigenvalues (y-axis) against the components (x-axis)
- Inspecting the shape of the resulting curve in order to detect the point at which the curve changes drastically (and the "scree on the hill slope" begins). This point on the curve indicates the maximum number of components to retain.

The scree plot is a decreasing function showing the variance explained by each rank, which is consistent with the variance-explained method. While the approach is simple and generally useful, such an intuitive but also fuzzy procedure has been criticized as subjective (Zwick & Velicer 1982). Figure 11 is the scree plot for the female population of US (Frame A, Short-term) and we will use it as an visual example in this section. Following the logic of scree plot, we may choose the rank of SVD model to be 1 or 2 since the slope after rank 3 is relatively flat. However, it is difficult to see the relationship between the percentage of variance explained by ranks and their forecasting ability. Is 90% or 95% good enough for forecasting?

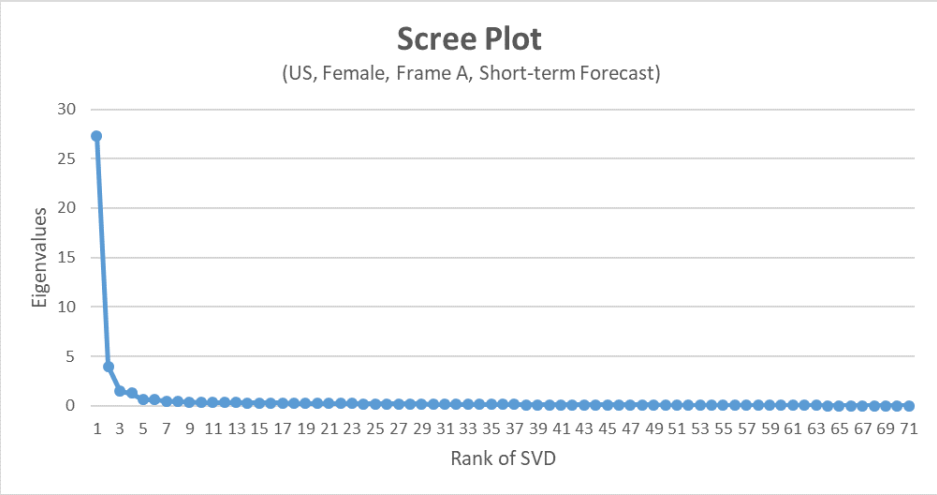

Figure 11: Scree Plot for US, Female, Frame A, Short-term Forecast

In order to compare the scree plot method (variance-explained method) with the out-of-sample validation method, we apply the out-of-sample validation method to select the rank for the SVD model in each scenario. For the same data set, we select the rank to be 16 since it provides the lowest RMSFE in the validation set for the female population of US (Frame A, Short-term), see Figure 12. The RMSFEs become relatively stable when the Rank is larger than 8. However, rank 16 gives us the lowest RMSE in the validation set. If forecasting performance is the only objective to consider, then 16 should be selected according to the result. However, in practice, if people have other considerations and prefer a more parsimonious model, then the selected rank should be determined by the trade-off between the model flexibility and the forecasting performance.

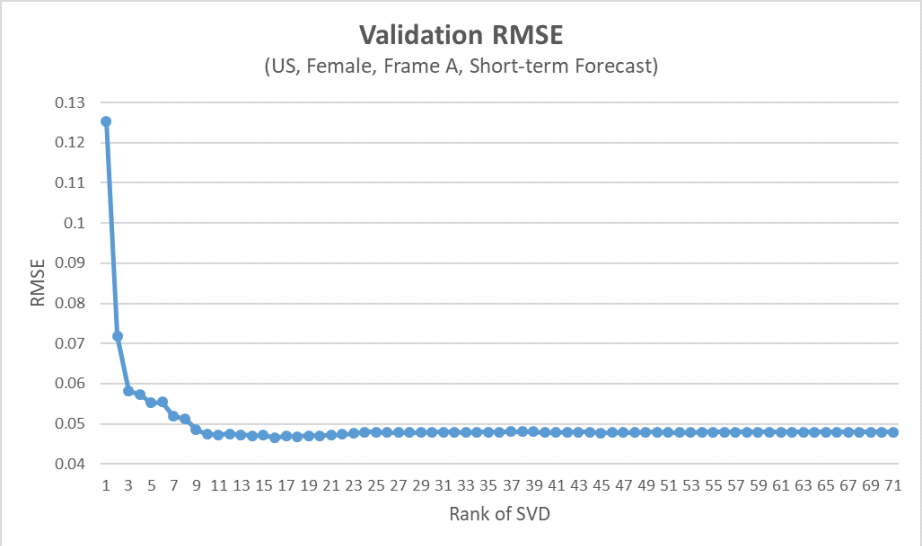

Figure 12: Validation RMSE for US, Female, Frame A, Short-term Forecast

In Table 16, the rank selected by out-of-sample validation in all scenarios are shown and the corresponding variance explained ratios are in brackets. There are two main differences that we observe. First, the rank selected by out-of-sample method is far larger than 1 in most cases. In another words, the rank determined by the out-of-sample validation method is far larger than that determined by the scree plot method. Second, the average variance explained ratio is 99.58 % gained by SVD(k) model, which is also larger than 94.04% gained by SVD(1) model as the rank increases.

Table 16: The rank selected and the variance explained, Frame A

	US(female)	**US(male)**	**UK(female)**	**UK(male)**
Short-term(5 Year)	16 (99.89%)	20 (99.88%)	4 (99.39%)	20 (99.78%)
Mid-term(10 Year)	13 (99.86%)	20 (99.85%)	9 (99.66%)	7 (99.36%)
Long-term(20 Year)	16 (99.92%)	8 (99.38%)	3 (99.31%)	3 (98.73%)

Next, we show the forecasting performance of the SVD(k) model in the testing set in which k is the rank determined by out-of-sample validation. For the purpose of facilitating the comparison of the forecasting performance between the SVD(k) and SVD(1) models, Table 17 copies the corresponding RMSFEs of the SVD(1) model from previous sections. Clearly, the forecasting performance of SVD(k) model is consistently and significantly better than that of SVD(1) model in all cases. Comparing the forecasting performance of SVD(k) to that of CPD model and Tucker(abc) model in the twelve cases listed in Table 17, we find that both CPD and Tucker(abc) outperform the SVD(k) in most cases, though the improvement is much smaller than that of the tensor models over SVD(1). The improvement of SVD(k) over SVD(1) is significant, but it is only suggested for single-population mortality modelling, as the number of parameters will explode if we apply it to multiple populations. More results applying the tensor decomposition models to USA and UK data with different time frames and forecasting horizons can be found in the supplementary materials.

Table 17: Out-of-sample forecasting performances by SVD(k) and SVD(1), Frame A

	US(female)		**US(male)**		**UK(female)**		**UK(male)**	
	SVD(k)	**SVD(1)**	**SVD(k)**	**SVD(1)**	**SVD(k)**	**SVD(1)**	**SVD(k)**	**SVD(1)**
Short-term(5 Year)	0.095	0.166	0.085	0.148	0.070	0.232	0.068	0.311
Mid-term(10 Year)	0.106	0.165	0.080	0.152	0.092	0.253	0.100	0.374
Long-term(20 Year)	0.167	0.195	0.133	0.179	0.163	0.272	0.201	0.427

In this section, we show how the out-of-sample validation method can be used to improve the forecasting ability of the SVD model. It also reveals the weaknesses of the variance-explained method (e.g. scree plot method). We summarise the two weaknesses of the variance-explained as below:

- In the case of one population, the relationship between variance explained and forecasting performance is ambiguous. One way to increase the variance explained is to increase the rank of the SVD model, but it may not improve the forecasting ability of the model. Hayton et al. (2004) remark that: "Although the scree test may work well with strong factors, it suffers from subjectivity and ambiguity, especially when there are either no clear breaks or two or more apparent breaks".

- In the case of many populations, there is no uniform criterion for the variance that models should explain when using different datasets. For example, SVD(1) model explains on average 93.54% variance on US data and on 94.53% on UK data. However, the forecasting ability of SVD(1) in US data is much better than that in the UK data. On the other hand, the SVD(k) model determined by out-of-sample method could explain on average 99.80% variance on US data and 99.37% on UK data. Different data sets require different explained variance ratios to optimize their forecasting ability. Zwick & Velicer (1986) state that scree plot interpretations often lack consistency and depend heavily on the training received by the examiners and on the nature of the solution.

Based on these two weaknesses and the consistently better forecasting performance of the SVD(k) model and Tensor models, we suggest using the out-of-sample validation method to select the rank parameter for mortality forecasting if forecasting performance is the only focus (objective).

4 Conclusion

Comparing multi-population mortality modelling using tensor decomposition methods (CPD & Tucker) with single-population mortality modelling using SVD (Lee-Carter model), we see that tensor decomposition methods always achieve superior out-of-sample forecasting performance both for individual populations and the overall population. There are two reasons to explain this. The first reason is that, multi-population decomposition methods have far fewer parameters (latent factors) to estimate compared with using the single-population SVD method. The more parsimonious structure of the tensor decomposition methods generates lower estimation error and hence leads to better forecasting performance. This reason could expain why the forecasting performance of tensor decomposition methods outperforms SVD for the overall population. The second reason is that different populations with similar socio-economic status share similar mortality trends (common information) to some extent. That common information could be extracted and utilised in the multi-population framework to help forecast the mortality of individual populations. However, when using the single-population SVD method, this common information is ignored and each population is modeled independently. This reason could explain why the forecasting performance of tensor decomposition methods outperforms SVD for individual populations.

We also compared using the variance-explained method and out-of-sample validation method to select the rank parameter. We show that for predictive modelling (forecasting performance is the focus), out-of-sample validation is a better method to select the rank for mortaltiy forecasting.

Supplementary Material

Supplementary Materials: The sumplementary materials are provided to include more figures, tables and algorithms related to this paper.
Codes: The codes to reproduce results, including figures and tables, in this paper can be obtained via `https://github.com/YumoDong/Tensor-mortality-prediction`.

Acknowledgements

The authors thank the editor and two anonymous reviewers for many constructive comments.

References

Acar, E., Çamtepe, S. A., Krishnamoorthy, M. S. & Yener, B. (2005), Modeling and multiway analysis of chatroom tensors, *in* 'International Conference on Intelligence and Security Informatics', Springer, pp. 256–268.

Acar, E., Camtepe, S. A. & Yener, B. (2006), Collective sampling and analysis of high order tensors for chatroom communications, *in* 'International Conference on Intelligence and Security Informatics', Springer, pp. 213–224.

Acar, E., Dunlavy, D. M., Kolda, T. G. & MÃžrup, M. (2011), 'Scalable tensor factorizations for incomplete data', *Chemometrics and Intelligent Laboratory Systems* **106**(1), 41 – 56.

Acton, D., Plat-Sinnige, M. T., Van Wamel, W., De Groot, N. & Van Belkum, A. (2009), 'Intestinal carriage of staphylococcus aureus: how does its frequency compare with that of nasal carriage and what is its clinical impact?', *European Journal of Clinical Microbiology & Infectious Diseases* **28**(2), 115.

Appellof, C. J. & Davidson, E. R. (1981), 'Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents', *Analytical Chemistry* **53**(13), 2053–2056.

Arlot, S., Celisse, A. et al. (2010), 'A survey of cross-validation procedures for model selection', *Statistics Surveys* **4**, 40–79.

Bader, B. W. & Kolda, T. G. (2007), 'Efficient matlab computations with sparse and factored tensors', *SIAM Journal on Scientific Computing* **30**(1), 205–231.

Bader, B. W., Kolda, T. G. et al. (2017), 'Matlab tensor toolbox version 3.0-dev', Available online.
URL: *https://www.tensortoolbox.org*

Cairns, A. J., Blake, D. & Dowd, K. (2006), 'A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration', *Journal of Risk and Insurance* **73**(4), 687–718.

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), 1–35.

Cattell, R. B. (1966), 'The scree test for the number of factors', *Multivariate Behavioral Research* **1**(2), 245–276.

Cawley, G. C. & Talbot, N. L. (2010), 'On over-fitting in model selection and subsequent selection bias in performance evaluation', *Journal of Machine Learning Research* **11**(Jul), 2079–2107.

Chang, J., Guo, B., Yao, Q. et al. (2018), 'Principal component analysis for second-order stationary vector time series', *The Annals of Statistics* **46**(5), 2094–2124.

CMI (1990), 'Continuous mortality investigation reports no.10'.

CMI (1999), 'Continuous mortality investigation reports no.17'.

Comon, P. (2009), 'Tensor decompositions, state of the art and applications', *IMA Conf. Mathematics in Signal Processing* .

Cramer, H. & Wold, H. (1935), 'Mortality variations in sweden: a study in graduation and forecasting', *Scandinavian Actuarial Journal* **1935**(3-4), 161–241.

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

Currie, I. D., Durban, M. & Eilers, P. H. (2006), 'Generalized linear array models with applications to multidimensional smoothing', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(2), 259–280.

Enchev, V., Kleinow, T. & Cairns, A. J. (2017), 'Multi-population mortality models: fitting, forecasting and comparisons', *Scandinavian Actuarial Journal* **2017**(4), 319–342.

Forfar, D. & Smith, D. (1985), 'The changing shape of english life tables', *Transactions of the Faculty of Actuaries* **40**, 98–134.

Gompertz, B. (1825), 'Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c', *Philosophical Transactions of the Royal Society of London* **115**, 513–583.

Haberman, S. & Sibbett, T. A. (1995), *History of actuarial science*, Vol. 10, William Pickering.

Hayton, J. C., Allen, D. G. & Scarpello, V. (2004), 'Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis', *Organizational Research Methods* **7**(2), 191–205.

Human Mortality Database (2018). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), Available at www.mortality.org or www.humanmortality.de (data downloaded on [1 September, 2018]).

Kamangar, F., Dores, G. M. & Anderson, W. F. (2006), 'Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world', *Journal of Clinical Oncology* **24**(14), 2137–2150.

Keyfitz, N. (1981), 'The limits of population forecasting', *Population and Development Review* pp. 579–593.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y. & Porter, M. A. (2014), 'Multilayer networks', *Journal of Complex Networks* **2**(3), 203–271.

Kleinow, T. (2015), 'A common age effect model for the mortality of multiple populations', *Insurance: Mathematics and Economics* **63**, 147–152.

Kolda, T. G. & Bader, B. W. (2009), 'Tensor decompositions and applications', *SIAM Review* **51**(3), 455–500.

Lee, R. D. & Carter, L. R. (1992), 'Modeling and forecasting us mortality', *Journal of the American Statistical Association* **87**(419), 659–671.

Lee, R. & Miller, T. (2001), 'Evaluating the performance of the lee-carter method for forecasting mortality', *Demography* **38**(4), 537–549.

Li, J. S.-H., Zhou, R. & Hardy, M. (2015), 'A step-by-step guide to building two-population stochastic mortality models', *Insurance: Mathematics and Economics* **63**, 121–134.

Li, N. & Lee, R. (2005), 'Coherent mortality forecasts for a group of populations: An extension of the lee-carter method', *Demography* **42**(3), 575–594.

Makeham, W. M. (1860), 'On the law of mortality and construction of annuity tables', *Journal of the Institute of Actuaries* **8**(6), 301–310.

Osborne, J. W. & Costello, A. B. (2004), 'Sample size and subject to item ratio in principal components analysis', *Practical Assessment, Research & Evaluation* **9**(11), 8.

Rabanser, S., Shchur, O. & Günnemann, S. (2017), 'Introduction to tensor decompositions and their applications in machine learning', *arXiv preprint arXiv:1711.10781* .

Renshaw, A. E. & Haberman, S. (2000), 'Modelling for mortality reduction factors'.

Renshaw, A. E. & Haberman, S. (2006), 'A cohort-based extension to the lee–carter model for mortality reduction factors', *Insurance: Mathematics and Economics* **38**(3), 556–570.

Russolillo, M., Giordano, G. & Haberman, S. (2011), 'Extending the lee–carter model: a three-way decomposition', *Scandinavian Actuarial Journal* **2011**(2), 96–117.

Shashua, A. & Levin, A. (2001), Linear image coding for regression and classification using the tensor-rank principle, *in* 'Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on', Vol. 1, IEEE, pp. I–I.

Smith, G. D., Hart, C., Blane, D., Gillis, C. & Hawthorne, V. (1997), 'Lifetime socioeconomic position and mortality: prospective observational study', *British Medical Journal* **314**(7080), 547.

Smith, G. D., Hart, C., Blane, D. & Hole, D. (1998), 'Adverse socioeconomic conditions in childhood and cause specific adult mortality: prospective observational study', *British Medical Journal* **316**(7145), 1631–1635.

Tsai, C. C.-L. & Lin, T. (2017), 'A bühlmann credibility approach to modeling mortality rates', *North American Actuarial Journal* **21**(2), 204–227.

Villegas, A. M., Haberman, S., Kaishev, V. K. & Millossovich, P. (2017), 'A comparative study of two-population models for the assessment of basis risk in longevity hedges', *ASTIN Bulletin: The Journal of the IAA* **47**(3), 631–679.

Wetterstrand, W. H. (1981), 'Parametric models for life insurance mortality data: Gompertz's law over time', *Transactions of the Society of Actuaries* **33**, 159–175.

Willets, R. (1999), 'Mortality in the next millennium', *Staple Inn Actuarial Society* **7**.

Yao, Y., Yu, H., Zhang, X., Roberts, S. & Huang, F. (2018), 'Mortality forecasting using the regularized matrix factorization method'. Working Paper.

Yu, H.-F., Rao, N. & Dhillon, I. S. (2016), Temporal regularized matrix factorization for high-dimensional time series prediction, *in* 'Advances in Neural Information Processing Systems', pp. 847–855.

Zhu, W., Tan, K. S. & Wang, C.-W. (2017), 'Modeling multicountry longevity risk with mortality dependence: A levy subordinated hierarchical archimedean copulas approach', *Journal of Risk and Insurance* **84**(S1), 477–493.

Zwick, W. R. & Velicer, W. F. (1982), 'Factors influencing four rules for determining the number of components to retain', *Multivariate Behavioral Research* **17**(2), 253–269.

Zwick, W. R. & Velicer, W. F. (1986), 'Comparison of five rules for determining the number of components to retain.', *Psychological Bulletin* **99**(3), 432.