Robust Parameter Estimation for the Lee-Carter Family: A Probabilistic Principal Component Approach

Speaker: Yiping Guo

University of Waterloo

2022-09-12

Motivation: Robustness in Mortality Projections

Model Robustness:

- What is model robustness?
- Why is model robustness important in actuarial science?
- The role of robustness in mortality modeling.

About this Project

- Our objective: Improving the model robustness of the **Lee-Carter model**.
- Methodology: Probabilistic principal component analysis (PPCA), a statistical machine learning technique.

Structure of this Talk

This talk will be covering the following:

- A review of the Lee-Carter model.
- A quick introduction to PPCA and its connection to the Lee-Carter model.
- The proposed method for robustifying the Lee-Carter model.
- Numerical results using the data from Human Mortality Database (HMD).
- Multi-population extensions and further remarks.

Review: The Lee-Carter Model

The Lee-Carter model (Lee and Carter, 1992):

$$\log(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}. \tag{1}$$

- a_x: Average log mortality rate for age x.
- b_x: Age-specific effect (sensitivity) for age x.
- k_t: Time trend. (t: time)

It involves two stages:

1. Estimating a_x, b_x and k_t; (Our focus)
2. Time series modeling on $\{k_t\}$ and forecasting.

Parameter Estimation for the Lee-Carter Model

Two main types of methods to estimate a_x, b_x and k_t:

$$y_{x,t} := \log(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}.$$

1. SVD or PCA (non-likelihood-based):

$$\min_{(\boldsymbol{a},\boldsymbol{b},\boldsymbol{k})} \sum_{x,t} \left(y_{x,t} - (a_x + b_x k_t)\right)^2. \tag{2}$$

2. Poisson regression (likelihood-based):

$$\max_{(\boldsymbol{a},\boldsymbol{b},\boldsymbol{k})} \sum_{x,t} \left(D_{x,t}(a_x + b_x k_t) - N_{x,t} e^{a_x + b_x k_t}\right) \tag{3}$$

Both methods are fragile to **outliers**. What is the consequence?

Why Probabilistic Principal Component Analysis (PPCA) for the Lee-Carter Model?

Our method is based on PPCA (Tipping and Bishop, 1999).

- PCA formulation of the Lee-Carter model:

$$\boldsymbol{y}_t := \log(\boldsymbol{m}_t) = \boldsymbol{a} + \boldsymbol{b}k_t + \boldsymbol{\varepsilon}_t, \tag{4}$$

  then the PCA (SVD) estimate of $\boldsymbol{b}$ is $\hat{\boldsymbol{b}} = \frac{\boldsymbol{u}_1}{\mathbf{1}^T \boldsymbol{u}_1}$.

- PPCA formulation of the Lee-Carter model:

$$\boldsymbol{y}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{a}, \boldsymbol{b}\boldsymbol{b}^T + \sigma^2 \boldsymbol{I}), \tag{5}$$

  then the maximum likelihood estimate (MLE) of $\boldsymbol{b}$ is $\hat{\boldsymbol{b}} = \boldsymbol{u}_1 \sqrt{\lambda_1 - \hat{\sigma}^2}$.

Question: How to interpret the *i.i.d.* condition in (5)?

Robustify the Lee-Carter Model via PPCA

One computationally efficient approach to robustify PPCA (Archambeau et al., 2006; Guo and Howard, 2022):

$$\underbrace{\boldsymbol{y}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{a}, \boldsymbol{b}\boldsymbol{b}^T + \sigma^2 \boldsymbol{I})}_{\text{Original Gaussian PPCA}} \overset{\text{Robustify}}{\Longrightarrow} \underbrace{\boldsymbol{y}_t \overset{iid}{\sim} t_\nu(\boldsymbol{a}, \boldsymbol{b}\boldsymbol{b}^T + \sigma^2 \boldsymbol{I})}_{\text{Robust } t\text{-PPCA}}. \quad (6)$$

- (Multivariate) t-distributions: Commonly adapted in robust statistics.
- MLE obtained from the t-PPCA: More robust against outliers.
- Implementation: Expectation-maximization (EM) algorithm.

Summary of the Proposed t-PPCA Lee-Carter Method

Denote $\boldsymbol{y}_t := \log(\boldsymbol{m}_t)$,

$$\boldsymbol{y}_t = \boldsymbol{a} + \boldsymbol{b}k_t + \boldsymbol{\varepsilon}_t, \qquad \text{(SVD/PCA)}$$

$$\overset{\text{Equivalent}}{\Longleftrightarrow} \boldsymbol{y}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{a}, \boldsymbol{b}\boldsymbol{b}^T + \sigma^2 \boldsymbol{I}), \qquad \text{(PPCA)}$$

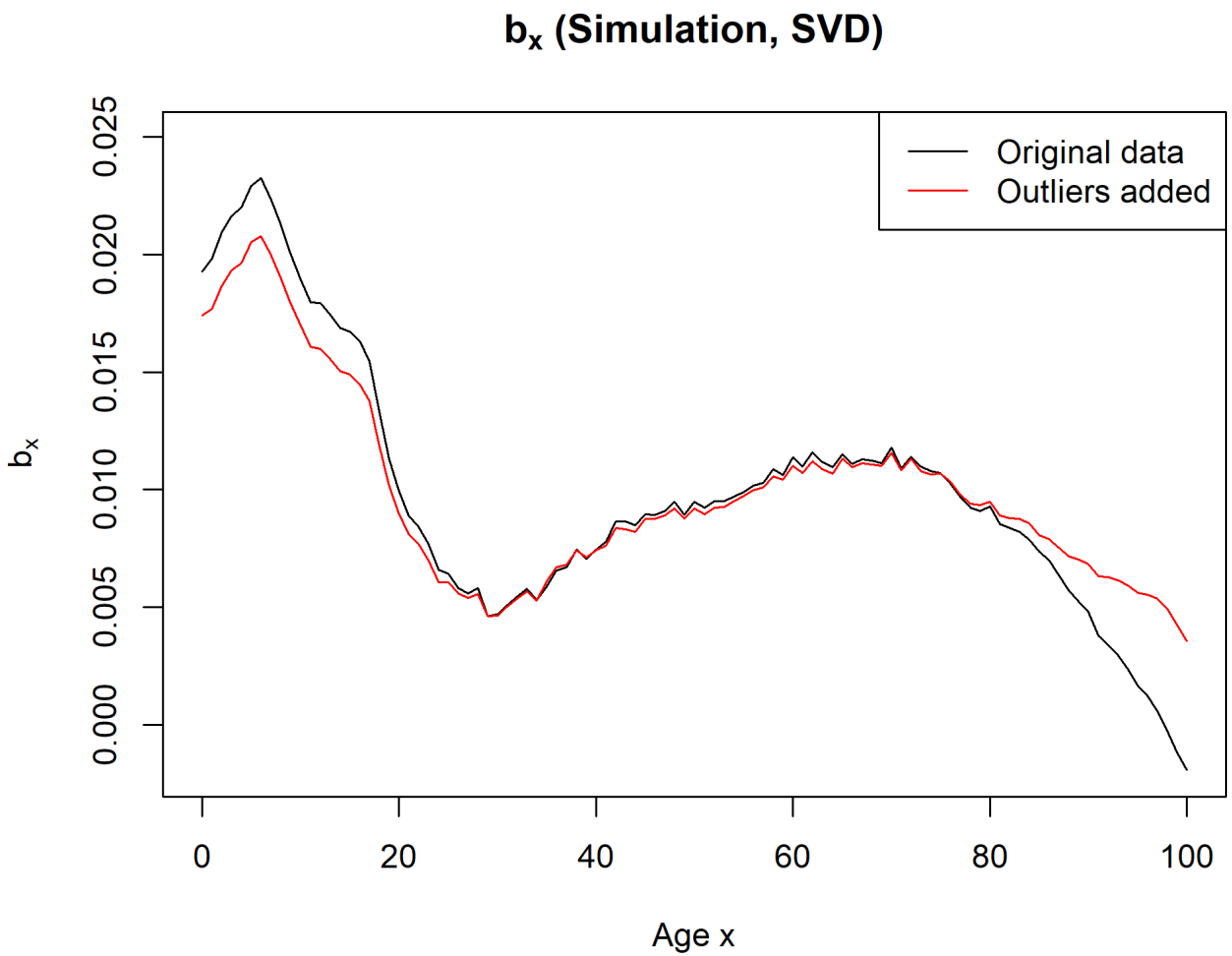
$$\overset{\text{Robustify}}{\Longrightarrow} \boldsymbol{y}_t \overset{iid}{\sim} t_\nu(\boldsymbol{a}, \boldsymbol{b}\boldsymbol{b}^T + \sigma^2 \boldsymbol{I}). \qquad (t\text{-PPCA})$$

- After obtaining the estimates $\hat{\boldsymbol{b}}$, the time trend $\hat{k}_t$ can be easily derived.
- A huge advantage: Flexible, can be combined into any other time series models for $\{k_t\}$.

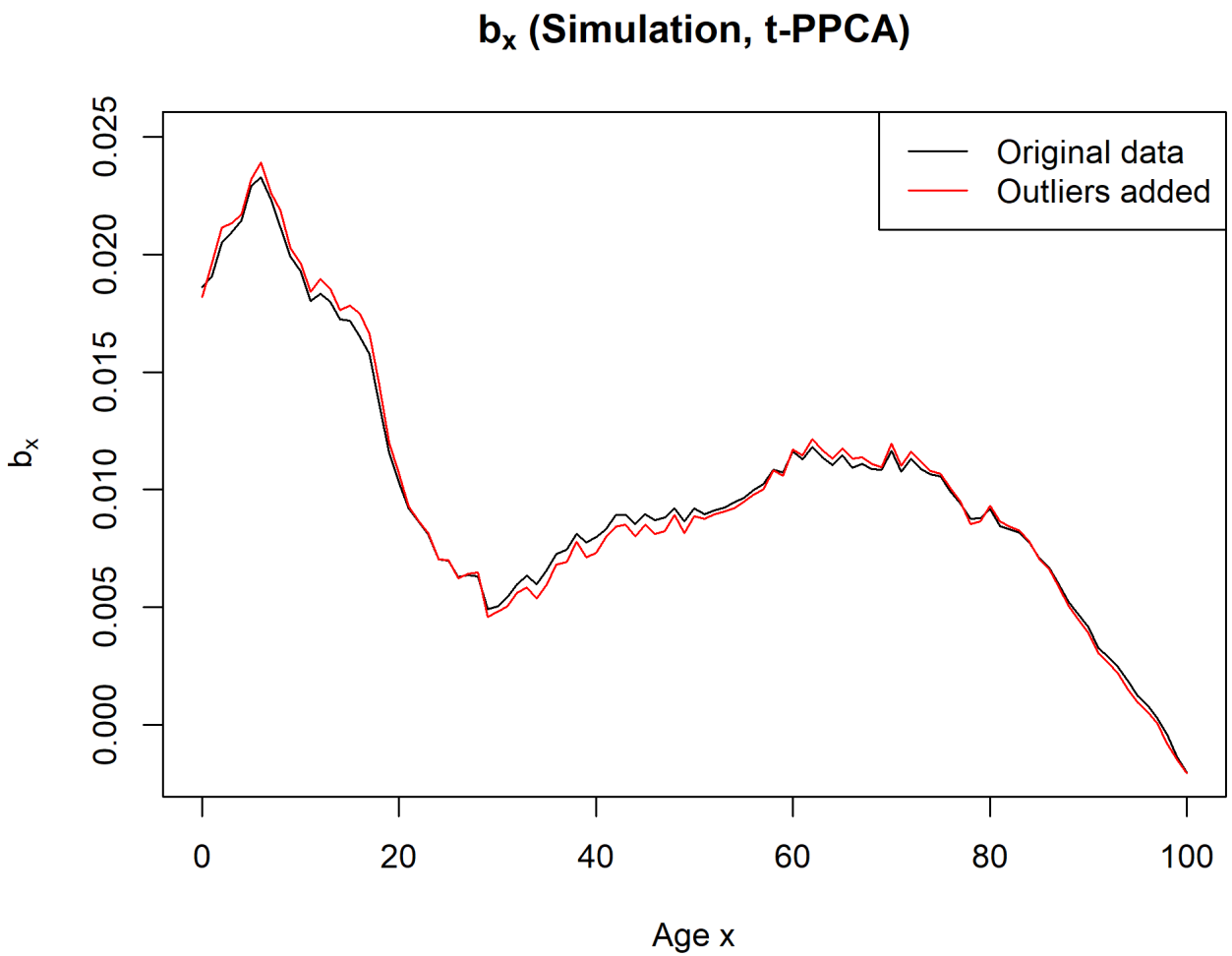
Experimental Design

- Objective: Examining the robustness of the t-PPCA Lee-Carter model.
- Key idea: Inserting **hypothetical pandemics** into past year(s).
- Base mortality data: U.S. data (1970-2019) from Human Mortality Database (HMD).
- Pandemic data: U.S. Covid-19 death data in 2020 from Centers of Disease Control and Prevention (CDC).
- Part 1: Illustrative example (plots).
- Part 2: Full experiment (tables).

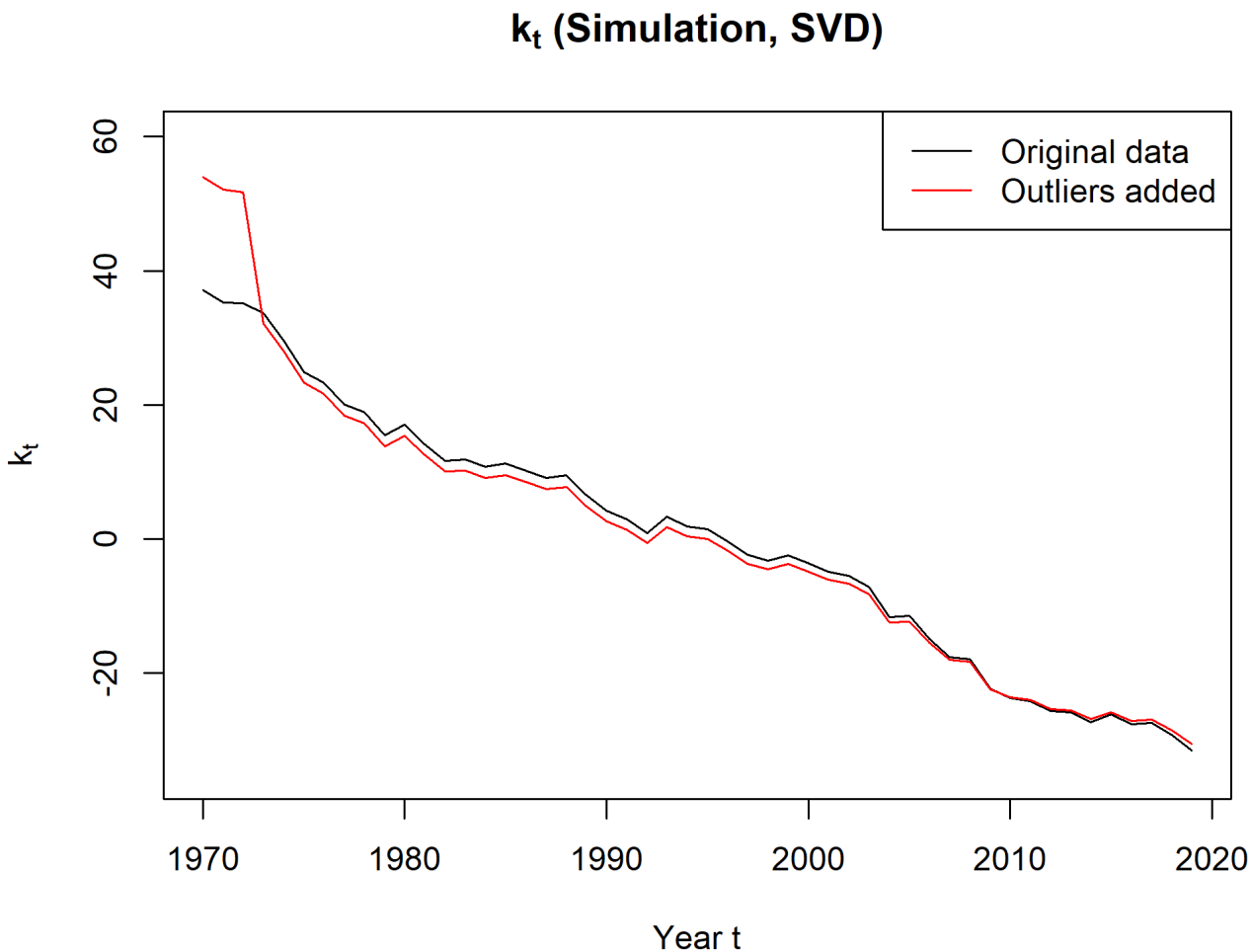
Hypothetical Pandemics in 1970-1972, estimates of b_x by SVD



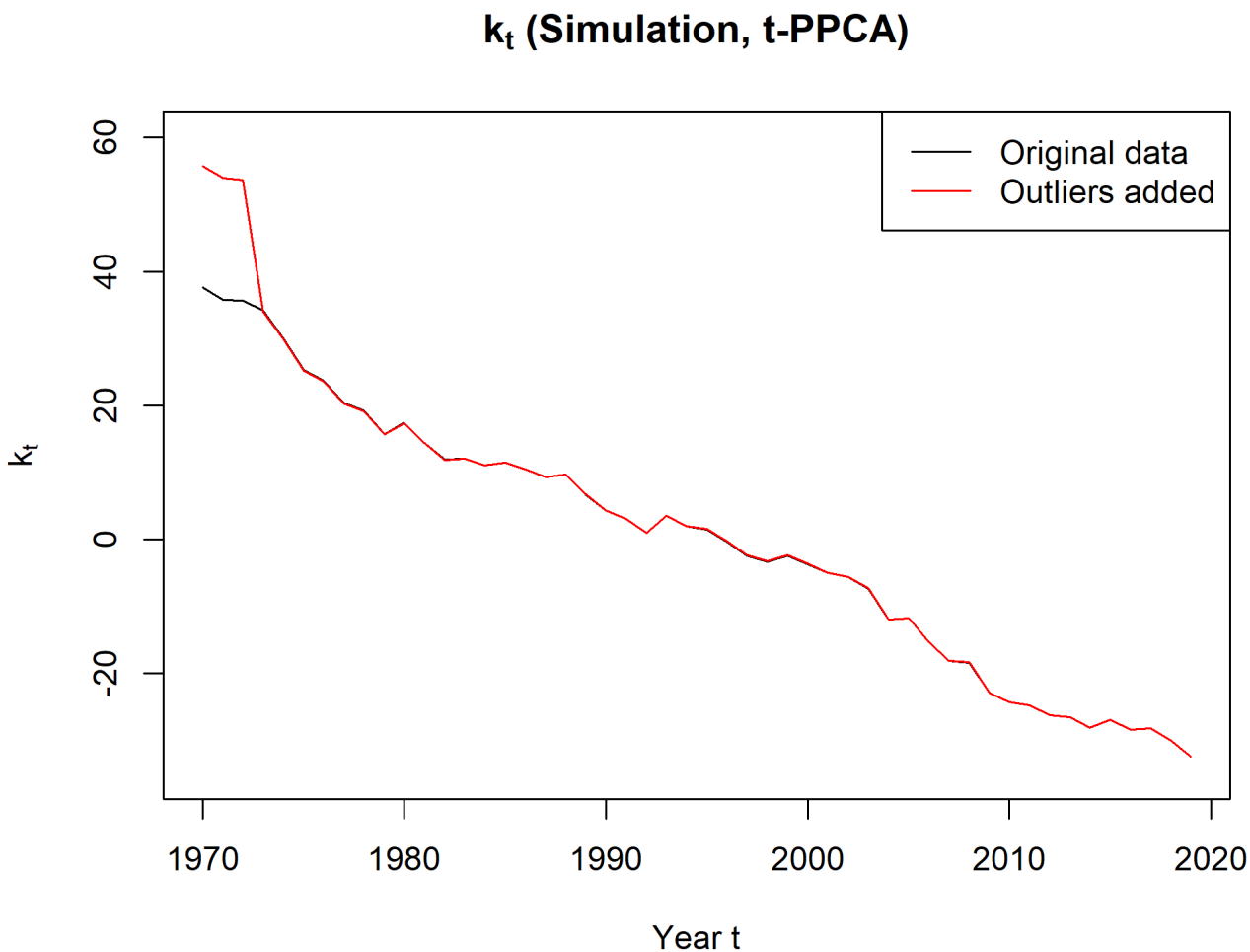
Hypothetical Pandemics in 1970-1972, estimates of b_x by t-PPCA



Hypothetical Pandemics in 1970-1972, estimates of k_t by SVD



Hypothetical Pandemics in 1970-1972, estimates of k_t by t-PPCA



Full Experiment: Design

- Randomly inserting 1-year or 3-year pandemics into past years.
- Comparing the robustness of SVD, Poisson GLM and t-PPCA.
- Error metric: Root mean square percentage error (RMSPE):

$$\text{RMSPE}(\hat{a}) = \sqrt{\frac{1}{p}\sum_{x=0}^{100}\left(\frac{a_x - \hat{a}_x}{a_x}\right)^2}, \tag{7}$$

$$\text{RMSPE}(\hat{b}) = \sqrt{\frac{1}{p}\sum_{x=0}^{100}\left(\frac{b_x - \hat{b}_x}{b_x}\right)^2}, \tag{8}$$

$$\text{RMSPE}(\hat{k}) = \sqrt{\frac{1}{n}\sum_{t\in T}\left(\frac{k_t - \hat{k}_t}{k_t}\right)^2}, \tag{9}$$

$T = \{1970, \cdots, 2019\} \backslash \{t | \boldsymbol{y}_t \text{ is a hypothetical outlier}\}$

Full Experiment: Robustness Analysis for $\hat{a}_x$, $\hat{b}_x$ and $\hat{k}_t$

Outliers	Method	RMSPE($\hat{a}$)	RMSPE($\hat{b}$)	RMSPE($\hat{k}$)
1-year	SVD	0.18%	18.90%	16.47%
	Poisson GLM	0.54%	30.01%	10.27%
	t-PPCA	0.08%	4.72%	9.05%
3-year	SVD	0.54%	51.20%	48.78%
	Poisson GLM	0.79%	84.33%	31.49%
	t-PPCA	0.23%	13.26%	29.34%

Table 1: Sample average of RMSPE over all samples

Multi-Population Extensions

Natural extensions to multi-population models, for example:

- Augmented common factor (ACF) model (Li and Lee, 2005):

$$y_{x,t,i} := \log(m_{x,t,i}) = a_{x,i} + \underbrace{b_x k_t}_{\text{common factor}} + \underbrace{b_{x,i} k_{t,i}}_{\text{specific factor}} + \varepsilon_{x,t,i}, \tag{10}$$

- Common age-effect (CAE) model (Kleinow, 2015):

$$y_{x,t,i} := \log(m_{x,t,i}) = a_{x,i} + b_x k_{t,i} + \varepsilon_{x,t,i}, \tag{11}$$

Conclusion

The advantages of the proposed t-PPCA Lee-Carter model are threefold:

1. Yielding more robust estimates for a_x, b_x and k_t, particularly for b_x.
2. Can be naturally extended to a large family of Lee-Carter type models.
3. Is Flexible to use with other existing time series models for $\{k_t\}$.

Reference I

Archambeau, C., Delannay, N., & Verleysen, M. (2006, June). Robust probabilistic projections. *In Proceedings of the 23rd International conference on machine learning*, (pp. 33-40).

Guo, Y., & Bondell, H. (2022). On robust probabilistic principal component analysis using multivariate t-distributions. *Communications in Statistics-Theory and Methods*, 1-19.

Kleinow, T. (2015). A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, 63, 147-152.

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, 87(419), 659-671.

Reference II

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

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611-622.