Advances in Stochastic Mortality Modelling

Robust Probabilistic Feature Extraction

Prof. Gareth W. Peters (FIOR, YAS-RSE)

Chair of Statistics for Risk and Insurance,
Department of Actuarial Mathematics and Statistics,
Heriot-Watt, Edinburgh, UK

Toczydlowska D., Peters G.W., Fung S. and Shevchenko P. (2018)
"**Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors via Probabilistic Robust Principle Components**"
Available at SSRN: `http://dx.doi.org/10.2139/ssrn.2977306`

July 15, 2018

Table of contents

Table of contents

Mortality Modelling Context

- Ageing populations are a major challenge for many countries.
  - **Fertility rates are declining while life expectancy is increasing.**

- **longevity risk:** *the adverse financial outcome of people living longer than expected* $\Rightarrow$ *possibility of outliving their retirement savings.*
  - long term demographic risk has significant implications for societies and manifests as a systematic risk for pension plans and annuity providers.

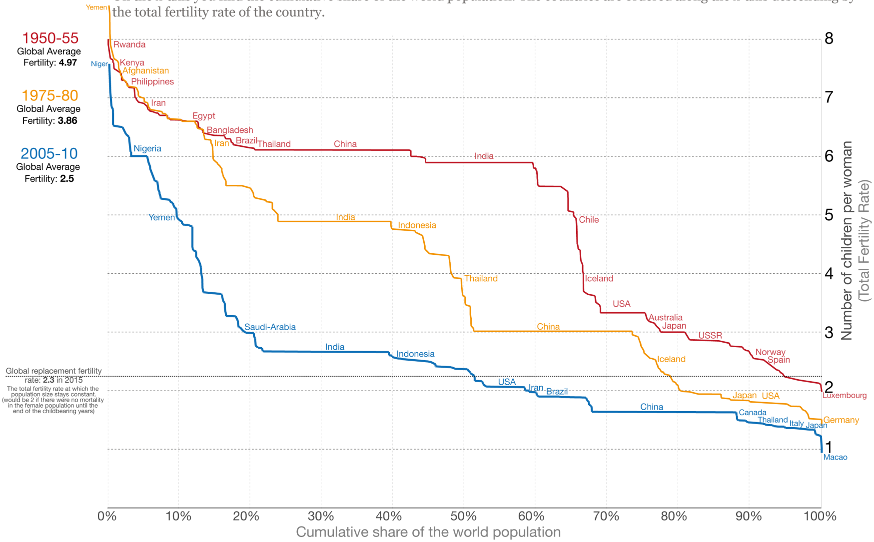
- Policymakers rely on **mortality projection** to determine appropriate pension benefits and regulations regarding retirement.

HERIOT WATT UNIVERSITY

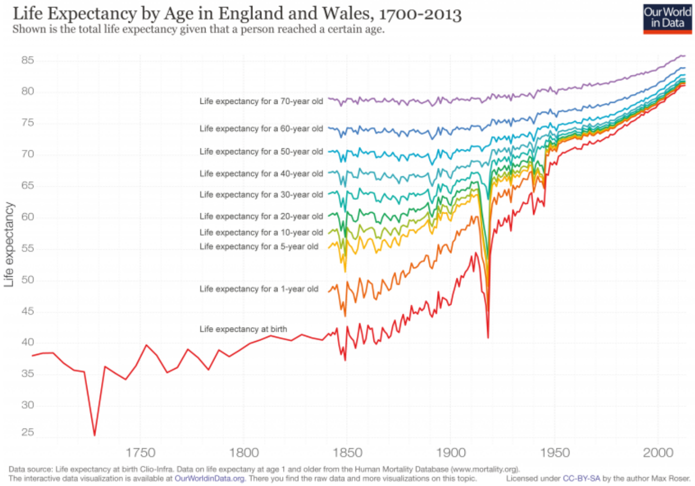
Our World in Data

World population by level of fertility over time (1950-2010)

On the x-axis you find the cumulative share of the world population. The countries are ordered along the x-axis descending by the total fertility rate of the country.



Data source: United Nations Population Division (2012 revision).
The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic.


Life Expectancy by Age in England and Wales, 1700-2013
Shown is the total life expectancy given that a person reached a certain age.
Our World in Data
Life expectancy
85
80
75
70
65
60
55
50
45
40
35
30
25
1750
1800
1850
1900
1950
2000
Life expectancy for a 70-year old
Life expectancy for a 60-year old
Life expectancy for a 50-year old
Life expectancy for a 40-year old
Life expectancy for a 30-year old
Life expectancy for a 20-year old
Life expectancy for a 10-year old
Life expectancy for a 5-year old
Life expectancy for a 1-year old
Life expectancy at birth
Data source: Life expectancy at birth Clio-Infra. Data on life expectany at age 1 and older from the Human Mortality Database (www.mortality.org).
The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic.
Licensed under CC-BY-SA by the author Max Roser.

Mortality Modelling Context

Enhancing mortality models requires an understanding of common features of mortality behaviour [Cairns, Blake and Dowd, 2008]

- Mortality rates have fallen dramatically at all ages.
- Rate of decrease in mortality has varied over time and by age group.
- Absolute decreases have varied by age group.
- Aggregate mortality rates have significant volatility year on year.

Table of contents

The uncertainty in future death rates can be divided into two components:

- **Unsystematic mortality risk**. Even if the true mortality rate is known, the number of deaths, will be random.
  - larger population ⇒ smaller unsystematic mortality risk (due to pooling of offsetting risks - diversification).
- **Systematic mortality risk**. This is the undiversifiable component of mortality risk that affects all individuals in the same way.
  - **Forecasts of mortality rates in future years are uncertain.**

- [Lee and Carter, 92] proposed a stochastic mortality model (LC) to forecast the trend of age-specific mortality rates.
- Several extensions to Lee-Carter model have been proposed, overview in [Fung et al. 2017].

- [Lee and Carter, 92] proposed a stochastic mortality model (LC) to forecast the trend of age-specific mortality rates.
- Several extensions to Lee-Carter model have been proposed, overview in [Fung et al. 2017].

Survival probability is still consistently underestimated especially in the last few decades ([IMF, 2012]).

- [Lee and Carter, 92] proposed a stochastic mortality model (LC) to forecast the trend of age-specific mortality rates.
- Several extensions to Lee-Carter model have been proposed, overview in [Fung et al. 2017].

Survival probability is still consistently underestimated especially in the last few decades ([IMF, 2012]).

This talk considers models aiming to help resolve this issue via

- **Stochastic State-Space Mortality Models** with **Period** and **Cohort** stochastic latent effects (LCC).
- + Extensions to State-Space Hybrid Regression Structures!

(see [Fung et al. 2017] and [Fung et al. 2018])

A state space model has two model components:

- **a stochastic observation equation**; and
- **a stochastic latent Markov state process**.

A state space model has two model components:

- **a stochastic observation equation**; and
- **a stochastic latent Markov state process**.

Key advantages of state space modelling approach:

- remove awkward identification specifications;
- computational efficiency and numerical robustness;
- accurate in-sample and out-of-sample forecasts;
- optimal statistical efficiency and unbiased estimation;

HERIOT WATT UNIVERSITY

Period-Cohort effect state-space formulation

Observation equation: log crude death rates, $y_{x,t} = \ln \hat{m}_{x,t}$, follow:

$$\ln \widehat{m}_{x,t} = \alpha_x + \beta_x \kappa_t + \beta_x^{\gamma} \gamma_{t-x} + \varepsilon_{x,t},$$

where $\varepsilon_{x,t}$ is a regression noise term.

- $\boldsymbol{\alpha} = \alpha_{x_1:x_p} := [\alpha_{x_1}, \ldots, \alpha_{x_p}]$ represents the *age-profile of the log death rates*
- $\boldsymbol{\beta} = \beta_{x_1:x_p}$ measures the *sensitivity of death rates for different age group* to a change of period effect κ_t.
- $\boldsymbol{\beta}^{\gamma} = \beta^{\gamma}_{x_1:x_p}$ measures the *sensitivity of death rates for different age group* to a change of cohort effect γ_{t-x}.

Observation Process: in matrix form.

$$\begin{pmatrix} y_{x_1,t} \\ y_{x_2,t} \\ \vdots \\ y_{x_p,t} \end{pmatrix} = \begin{pmatrix} \alpha_{x_1} \\ \alpha_{x_2} \\ \vdots \\ \alpha_{x_p} \end{pmatrix} + \begin{pmatrix} \beta_{x_1} & \beta^{\gamma}_{x_1} & 0 & \cdots & 0 \\ \beta_{x_2} & 0 & \beta^{\gamma}_{x_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{x_p} & 0 & 0 & \cdots & \beta^{\gamma}_{x_p} \end{pmatrix} \begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{x_1,t} \\ \varepsilon_{x_2,t} \\ \vdots \\ \varepsilon_{x_p,t} \end{pmatrix}.$$

Here $(\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^\top$ *is the* $p+1$ *dimensional latent state vector.* $\gamma_t^x := \gamma_{t-x}$

State Equation in matrix form:

$$\begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_{p-1}} \\ \gamma_t^{x_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_1 & \lambda_2 & \cdots & \lambda_{p-1} & \lambda_p \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \kappa_{t-1} \\ \gamma_{t-1}^{x_1} \\ \gamma_{t-1}^{x_2} \\ \vdots \\ \gamma_{t-1}^{x_{p-1}} \\ \gamma_{t-1}^{x_p} \end{pmatrix} + \begin{pmatrix} \theta \\ \eta \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \omega_t^{\kappa} \\ \omega_t^{\gamma} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

Period effect κ_t is a random walk with drift process with $\omega_t^{\kappa} \overset{iid}{\sim} \mathrm{N}(0, \sigma_\omega^2)$ and Cohort effect $\gamma_t^{x_1}$ is a stationary AR(p) process with $\omega_t^{\gamma} \overset{iid}{\sim} \mathrm{N}(0, \sigma_\gamma^2)$

Table of contents

GOAL: develop *stochastic mortality state-space hybrid factor models*.

- **Hybrid** := Stochastic Latent Factors $+$ Observable Covariate Features
- **observable features extracted** from demographic data
- Feature extraction should aim for dimension reduction $\Rightarrow$ model parsimony.

[Toczydlowska and Peters, 2017] address important aspects of feature extraction:

1. **missing data** in time-series and panel (matrix) structured real demographic data;
2. **noisy observations and outliers** (in real data);

State-Space Hybrid Factor Models

Two fundamental approaches to develop Hybrid Factor Models:

1. time varying factor with **static loading coefficient** (*classical distributed lag regressions such as ARDL models*);

2. static factor with **time varying stochastic loading coefficients.** (*state space models e.g. dynamic Nelson-Siegel yield curves*).

- Option 2: suitable for **high dimensional** data, **time series** / **panel structured** but represented by relatively "short time series" lengths.
  - ⇒ *particularly prevalent in demographic studies!*

State-Space Hybrid Factor Models

Consider the State-Space Hybrid Period-Cohort-Demographic Model

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \tilde{\mathbf{B}}_t \tilde{\varphi}_t + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{I}_p),$$

$$\tilde{\varphi}_t = \tilde{\boldsymbol{\Lambda}} \tilde{\varphi}_{t-1} + \tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\omega}}_t \overset{iid}{\sim} \mathcal{N}(0, \tilde{\Upsilon})$$

where $\tilde{\varphi}_t = (\varphi_t, \varrho_t)$ is a $(p + pk + 1) \times 1$ latent process vector of φ_t **stochastic mortality factors** (period-cohort) and ϱ_t **dynamic factor loadings**, with

$$\tilde{\boldsymbol{\Theta}} = \begin{pmatrix} \boldsymbol{\Theta}_{(p+1)\times 1} \\ \boldsymbol{\Psi}_{pk\times 1} \end{pmatrix}_{(p+pk+1)\times 1}$$

a vector of drift parameters for state equations.

Consider three models:

Case 1: *Factors in Observation Equation Only*;

Case 2: *Factors in Period Effect State Equation Only*;

Case 3: *Factors in Cohort Effect State Equation Only*.

$$\tilde{\mathbf{B}}_{t\ p\times(p+pk+1)} = \begin{cases} \left(\mathbf{B}_{p\times(p+1)} \mid \tilde{\mathbf{F}}_t \right) & \text{for } \mathbf{Case\ 1}, \\ \left(\mathbf{B}_{p\times(p+1)} \mid \mathbf{0}_{p\times pk} \right) & \text{otherwise}, \end{cases}$$

$$\tilde{\boldsymbol{\Lambda}}_{(p+pk+1)\times(p+pk+1)} = \begin{cases} \left(\begin{array}{c|c} \boldsymbol{\Lambda}_{(p+1)\times(p+1)} & \mathbf{0}_{(p+1)\times pk} \\ \hline \mathbf{0}_{pk\times(p+1)} & \boldsymbol{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \mathbf{Case\ 1}, \\ \left(\begin{array}{c|c} \boldsymbol{\Lambda}_{(p+1)\times(p+1)} & \begin{array}{c} \tilde{\mathbf{f}}_t^T \\ \mathbf{0}_{p\times pk} \end{array} \\ \hline \mathbf{0}_{pk\times(p+1)} & \boldsymbol{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \mathbf{Case\ 2}, \\ \left(\begin{array}{c|c} \boldsymbol{\Lambda}_{(p+1)\times(p+1)} & \begin{array}{c} \mathbf{0}_{1\times pk} \\ \tilde{\mathbf{F}}_t \end{array} \\ \hline \mathbf{0}_{pk\times(p+1)} & \boldsymbol{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \mathbf{Case\ 3}. \end{cases}$$

Table of contents

Probabilistic Feature Extraction

HERIOT WATT UNIVERSITY

- Data $\mathbf{Y}_t$ is observed (or partially observed) over periods $t \in \{1, \ldots, n\}$ and will be reduced to factors $\widetilde{\mathbf{F}}_t$

  Example: d countries demographic data and p denotes the number of different demographic attributes observed
  $\Rightarrow$ then $p \times d$ matrix of data in year t is $\mathbf{Y}_t$.

- **We do not wish to utilise the raw demographic data $\widetilde{\mathbf{F}}_t \neq \mathbf{Y}_t$:**

  in general it will produce a model with too many parameters

- [Toczydlowska and Peters, 2017] considered stochastic projection methods of dimensionality reduction

  $\Rightarrow$ **Probabilistic Principal Component Analysis (PPCA)** and Robust extensions.

PCA by means of Factor Analysis: with n realisations of the $(p \times d)$-dimensional observed demographic data, vectorized into columns $\mathbf{Y}$.

Consider linear decompositions:

$$\mathbf{Y}_{n\times pd} = \mathbf{X}_{n\times pd}\mathbf{W}^{T}_{pd\times pd} + \epsilon_{n\times pd}.$$

Factor analysis assumes diagonal covariance for ϵ_t.

Stochastic Factor PCA: *differs from deterministic PCA as components $\mathbf{x}_t$ and factor loading matrix $\mathbf{W}$ account for correlation between elements of $\mathbf{y}_t$ and only part of the variation*:

$$\mathbb{E}\mathbf{y}_t^T\mathbf{y}_t = \mathbb{E}\left[\left(\mathbf{x}_t\mathbf{W}^T + \epsilon_t\right)^T\left(\mathbf{x}_t\mathbf{W}^T + \epsilon_t\right)\right] = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^T + \boldsymbol{\Psi}.$$

In standard PCA $\mathbf{x}_t$ and W account for the entire covariance.

HERIOT WATT UNIVERSITY

Show $\mathbf{x}_t$ and $\mathbf{W}$ account for correlation!

Example: assume $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$ and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ to obtain,

$$\mathbf{y}_t | \mathbf{x}_t, \mathbf{W}, \boldsymbol{\Psi} \sim \mathcal{N}\left(\mathbf{x}_t \mathbf{W}^T, \boldsymbol{\Psi}\right),$$

$$\pi(\mathbf{y}_t | \mathbf{W}, \boldsymbol{\Psi}) = \int_{\mathbb{R}^d} \pi(\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \boldsymbol{\Psi}) d\mathbf{x}_t = (2\pi)^{-\frac{d}{2}} |\mathbf{C}|^{-1} \exp\{-\frac{1}{2}\mathbf{y}_t \mathbf{C}^{-1} \mathbf{y}_t^T\}$$

for $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$ where $|\mathbf{C}|$ denotes the determinant of the matrix.

- Notice that since $\boldsymbol{\Psi}$ is diagonal, the correlation structure between components $\mathbf{y}_t$ is specified by the matrix $\mathbf{W}$.

Show $\mathbf{x}_t$ and $\mathbf{W}$ account for correlation cont.

Eigen decomposition of covariance $\mathbf{C} = \mathbf{U}_{d\times d}\mathbf{L}_{d\times d}\mathbf{U}^T$, for diagonal $\mathbf{L}$ and orthonormal $\mathbf{U}$, gives

$$\mathbf{0} = (\mathbf{C} - \mathbf{L})\mathbf{U} = \left(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_d - \mathbf{L}\right)\mathbf{U} = \left(\mathbf{W}\mathbf{W}^T - \left(\mathbf{L} - \sigma^2\mathbb{I}_d\right)\right)\mathbf{U}.$$

- Thus, the matrix $\boldsymbol{\Lambda} = \left(\mathbf{L} - \sigma^2\mathbb{I}_d\right)$ and $\mathbf{U}$ are matrices of eigenvalues and corresponding eigenvectors of $\mathbf{W}\mathbf{W}^T$.
- Since $\lambda_i = l_i - \sigma^2 \geq 0$, the scalar σ^2 can be chosen as the smallest diagonal element of $\boldsymbol{\Lambda}$.
- **Factor loadings are given by $\mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$.**

  Assuming the error term ϵ_t is homogeneous s.t. $\boldsymbol{\Psi} = \sigma^2\mathbb{I}_d$, then estimating $\mathbf{W}$ via PCA given $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ is identifiable.

Feature Extraction via EM Algorithm Estimation!

Goal is to estimate:

- projection matrix $\mathbf{W}$,
- vector $\boldsymbol{\mu}$ and
- scalar σ^2

given marginal distribution of $\mathbf{Y}_t$

$$\mathbf{Y}_t | \Psi \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbb{I}_d\right)$$

for the vector of static parameters $\Psi = \left[\mathbf{W}, \boldsymbol{\mu}, \sigma^2\right]$ of the model.

The EM algorithm uses logarithm of the the complete data likelihood, i.e. the joint distribution of $\mathbf{Y}_{1:N}, \mathbf{X}_{1:N} | \Psi$ given by

$$\pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N} | \Psi}\left(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}\right) = \prod_{t=1}^{N} \pi_{\mathbf{Y}_t | \mathbf{X}_t, \Psi}\left(\mathbf{y}_t\right) \pi_{\mathbf{X}_t | \Psi}\left(\mathbf{x}_t\right)$$

$$= (2\pi)^{-N\frac{d+k}{2}} \left(\sigma^2\right)^{-N\frac{d}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^{N} \left(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{x}_t \mathbf{W}^T\right) \left(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{x}_t \mathbf{W}^T\right)^T - \frac{1}{2} \sum_{t=1}^{N} \mathbf{x}_t \mathbf{x}_t^T \right\}.$$

HERIOT WATT UNIVERSITY

Feature Extraction via EM Algorithm Estimation!

1. Expectation step: Expectation of the loglikelihood function of the join distribution of $\mathbf{Y}_{1:N}, \mathbf{X}_{1:N} | \Psi$ for a fixed vector of static parameters Ψ^* with respect to the conditional distribution $\mathbf{X}_{1:N} | \mathbf{Y}_{1:N}, \Psi$

$$Q(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N} | \mathbf{Y}_{1:N}, \Psi} \left[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N} | \Psi^*} \left(\mathbf{y}_{1:N}, \mathbf{x}_{1:N} \right) \right]$$

2. Maximisation step: Finding $\mathbf{W}^*, \boldsymbol{\mu}^*$ and σ^{*2} that maximize $Q(\Psi | \Psi^*)$

$$\left(\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2} \right) = \underset{\mathbf{W}^* \in \mathbb{R}^{d \times k}, \boldsymbol{\mu}^* \in \mathbb{R}^d, \sigma^{*2} > 0}{\operatorname{argmax}} Q(\Psi, \Psi^*)$$

Probabilistic Feature Extraction

Theorem

The E-step of the EM algorithm for Gaussian Probabilistic Principal Component Analysis given N realisations of the observation vector $\mathbf{Y}_t$ *denoted by* $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ *is obtained in the closed form as follows*

$$Q(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\right]$$

$$= -\frac{N(d+k)}{2}\log 2\pi - \frac{Nd}{2}\log \sigma^{*2} - \frac{1}{2}\sum_{t=1}^{N}\Bigg\{\frac{1}{\sigma^{*2}} Tr\{\mathbf{y}_t^T \mathbf{y}_t\}$$

$$- \frac{2}{\sigma^{*2}}\mathbf{y}_t \boldsymbol{\mu}^{*T} + \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} - \frac{2}{\sigma^{*2}} Tr\left\{\mathbf{W}^* \mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\right]\mathbf{y}_t\right\}$$

$$+ \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + Tr\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T \mathbf{X}_t\right]\right\}\Bigg\}$$

see details of expectations and proof in [Toczydlowska and Peters, 2017].

Theorem

The maximizers of the function $Q(\Psi, \Psi^*)$ *are given by*

$$\boldsymbol{\mu}^* = \bar{\mu}(\mathbf{y}_{1:N}; \Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}$$

$$\mathbf{W}^* = \bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}; \Psi, \Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}(\mathbf{y}_{1:N}; \Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1}$$

$$\sigma^{*2} = \frac{1}{d}\, Tr\Big\{\bar{\mathbf{C}}_{\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}; \Psi, \Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}; \Psi, \Psi^*)$$
$$+ \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}(\mathbf{y}_{1:N}; \Psi)\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\Big\}$$

see details of components and proof in [Toczydlowska and Peters, 2017].

Probabilistic PCA with Missing Data:
Until now, we assumed the data did not contain any missing observations!

- Real demographic time series data have numerous types of missingness.
- $\Rightarrow$ *missingness is an important aspect to address in the feature extraction!*

[Toczydlowska and Peters, (2017), (2018)] *address different components of* ***PPCA in missing data*** *estimation settings via* ***robust*** *versions of Expectation-Maximisation.*

- Distributional Extensions: **Student-t, Skewed and Grouped Student-t cases.**

HERIOT WATT UNIVERSITY

Feature Extraction via EM Algorithm with **MISSING DATA**!

Define the indicator random variable $\mathbf{R}_t$ which decides which entries of $\mathbf{Y}_t$ are missing denoting them by 1, otherwise 0.

- Each observation consists of the pair $[\mathbf{Y}_t^o, \mathbf{R}_t]$ with distribution parameterized according to parameters $[\Psi, \Theta]$ respectively.

Likelihood is given by conditional probability $\mathbf{Y}_t^o, \mathbf{R}_t|\Psi, \Theta$:

$$\pi_{\mathbf{Y}_t^o, \mathbf{R}_t|\Psi,\Theta}\left(\mathbf{y}_t^o, \mathbf{r}_t\right) = \int \pi_{\mathbf{Y}_t^o, \mathbf{Y}_t^m, \mathbf{R}_t|\Psi,\Theta}\left(\mathbf{y}_t^o, \mathbf{y}_t^m, \mathbf{r}_t\right) d\mathbf{y}_t^m$$

$$= \int \pi_{\mathbf{R}_t|\mathbf{Y}_t,\Psi,\Theta}\left(\mathbf{r}_t\right) \pi_{\mathbf{Y}_t|\Psi,\Theta}\left(\mathbf{y}_t\right) d\mathbf{y}_t^m$$

We assume for simplicity a pattern of missing data according to MAR - missing at random

- The assumptions imposes the indicator variable $\mathbf{R}_t$ to be independent of of the value of missing data.

Then the vector $\mathbf{Y}_t$ which is MAR satisfies

$$\pi_{\mathbf{R}_t|\mathbf{Y}_t,\Psi}(\mathbf{r}_t) = \pi_{\mathbf{R}_t|\mathbf{Y}_t^o,\Psi}(\mathbf{r}_t)$$

resulting in

$$\begin{aligned}\pi_{\mathbf{Y}_t^o,\mathbf{R}_t|\Psi,\Theta}\left(\mathbf{y}_t^o\right) &= \pi_{\mathbf{R}_t|\mathbf{Y}_t^o,\Theta}\left(\mathbf{r}_t\right)\int \pi_{\mathbf{Y}_t|\Psi}\left(\mathbf{y}_t\right) d\mathbf{y}_t^m \\ &= \pi_{\mathbf{R}_t|\mathbf{Y}_t^o,\Theta}\left(\mathbf{r}_t\right)\pi_{\mathbf{Y}_t^o|\Psi}\left(\mathbf{y}_t^o\right).\end{aligned}$$

$\Rightarrow$ Under the MAR assumption, the estimation of Ψ via maximum likelihood of the joint distribution $\mathbf{Y}_t^o, \mathbf{R}_t|\Psi, \Theta$ is equivalent to the maximisation of the likelihood of the marginal distribution $\mathbf{Y}_t^o|\Psi$.

Table of contents

Demographic data that we extract "Observable" covariate regression Features from:

- Data from Human Mortality Database (`http://www.mortality.org`).

- We use four different data sets:
  - Birth counts;
  - Death counts;
  - Life tables: Life Expectancy at Birth and Death Rates.

- The time series vary in terms of **data structure, the number of available observations and the missingness attributes** of the records.

HERIOT WATT UNIVERSITY

TYPES OF DATA:

- **One dimensional time series data per country per gender**
  (31 countries, M and F, gives 124 time series):
  - Birth counts and
  - Life expectancy at Birth.
- **Multivariate cross sectional time series data per country & gender**: age specific data for Death counts and Death Rates.
- A single observation per country in time t describes:
  - number of deaths of people with ages from 0 to 110+ (Death counts) or;
  - number of deaths for ages from 0 to 110+ scaled to the size of that population, per unit of time (Death Rates).

HERIOT WATT UNIVERSITY

Model estimation performed by Forward-Backward Kalman Filter within Rao-Blackwellised Adaptive Gibbs Sampler (MCMC).

The state space models we considered in our studies were of type:

1. [LCC:] Lee-Carter model with the stochastic period + cohort effect.
2. [DFM-PC:] demographic factor model versions of Lee-Carter (Period-Cohort).

The factors are obtained by performing **Probabilistic Principle Component Analaysis PPCA** jointly on the set of data for all countries listed, excluding:
United Kingdom (response variable)

LC State Space Model - only a Period Effect κ_t included.

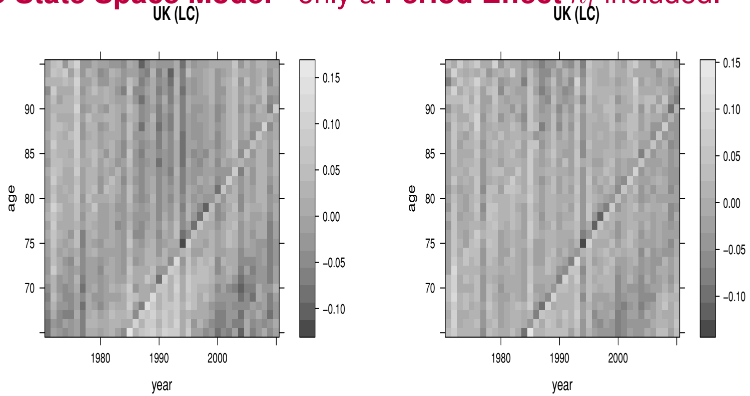
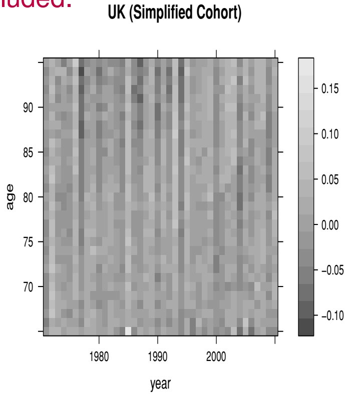


Figure: In sample analysis residuals (left Female, right Male).

LCC State Space Model - with **Period + Cohort Effects** κ_t, γ_{t-x} included.



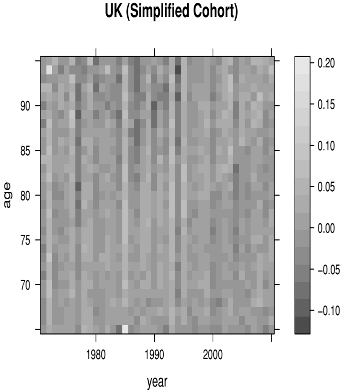


Figure: In sample analysis residuals (left Female, right Male).

Application

- [DFM-PC-B:] the mean of first principal component of Birth counts as a static parameter, age specific element of ϱ_t;
- [DFM-PC-D-r/s:] the first principal component of Death counts (which is age and country specific) as an exogenous factor, one element of ϱ_t corresponds to a country specific subvector of the component.;
- [DFM-PC-Mx-r/s:] the first principal component of Death Rates (which is age and country specific) as an exogenous factor, one element of ϱ_t corresponds to a country specific subvector of the component.

r/s - is robust vs standard

- **Out-of-Sample Study:** Model calibration period is 1922 – 2002
  ⇒ forecast performance analysis for 2003 – 2013

Model	MSE	DIC	$MSEP_{MCMC}$	$MSEP_{Kalman}$
LCC	0.0097	-3627	0.1778	0.1774
DFM-PC-B	**0.0072**	**-6500**	**0.0057**	**0.0062**
DFM-PC-D-r	0.0182	-6380	0.0177	0.0251
DFM-PC-D-s	**0.0065**	**-5996**	**0.0185**	**0.0156**
DFM-PC-Mx-r	**0.0081**	**-8225**	**0.0111**	**0.0129**
DFM-PC-Mx-s	0.0174	-3951	0.0692	0.0285

- **The results confirm that adding demographic features, as additional explanatory variables to the LCC model, improves both in-sample fit out-of-sample fit and therefore the predictability of log death rates.**

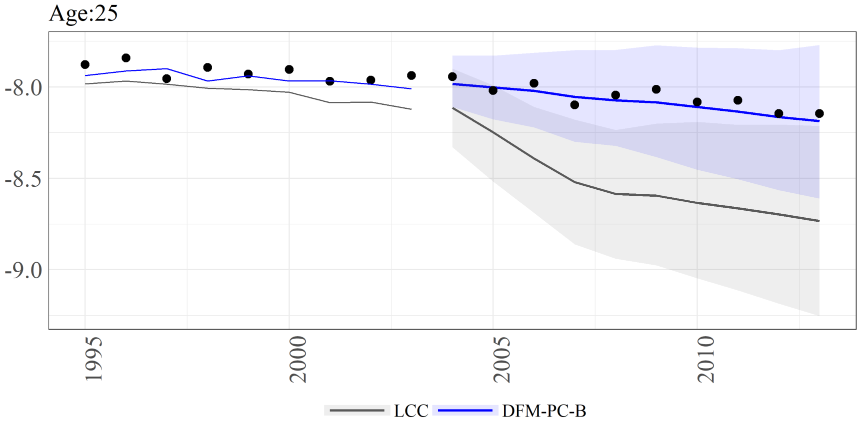


Figure: 10-year out-of-sample forecasted log death (y axis) rates by age with corresponding prediction intervals.

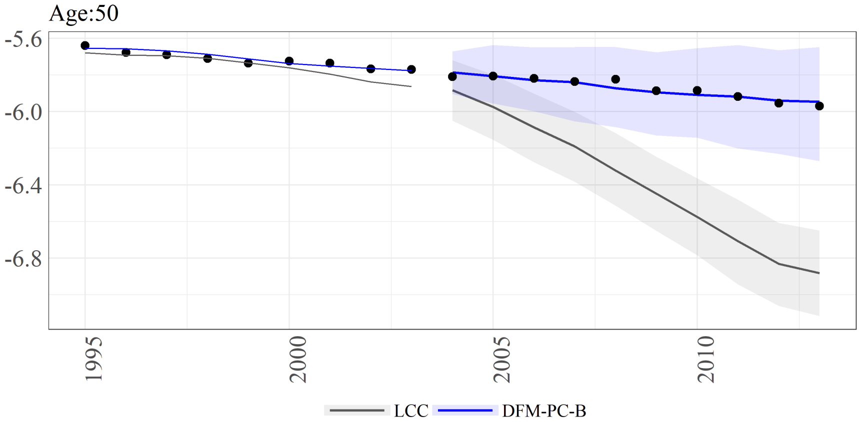


Figure: 10-year out-of-sample forecasted log death (y axis) rates by age with corresponding prediction intervals.

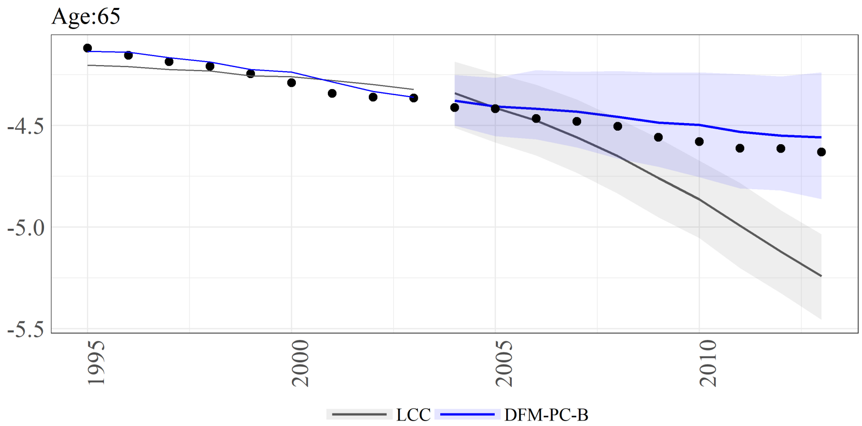


Figure: 10-year out-of-sample forecasted log death (y axis) rates by age with corresponding prediction intervals.

Conclusions

HERIOT WATT UNIVERSITY

- We explored how to construct a state space formulation of the stochastic mortality models for Period and Cohort factors
- We explored how to extend to Hybrid Multi-Factor Stochastic State-Space Mortality models with Period-Cohort factors as well as demographic regressors.
- We briefly learnt about feature/covariate extraction methods to extract the demographic factors used in the extended HMF Stochastic State-Space Mortality models.
- Standard Lee-Carter Period-Cohort model consistently under estimates forecast log-death rates
- Extended models proposed improve significantly the forecast performance of log-death rates.

Table of contents

References

1 Fung M.C., Peters G.W., Shevchenko P.V.
A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting.
Annals of Actuarial Science. 2017 May:1-47.
Available at SSRN: `https://ssrn.com/abstract=2786559`

2 Fung M.C. and Peters G.W. and Shevchenko P.V.
A State-Space Estimation of the Lee-Carter Mortality Model and Implications for Annuity Pricing
MODSIM Modelling and Simulation Society. 2015, July.
Available at SSRN: `https://ssrn.com/abstract=2699624`

3 Toczydlowska D., Peters G.W., Fung M.C. and Shevchenko P.V.
Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors via Probabilistic Robust Principle Components
Risks: Special Issue on “Aging Population Risks”.
Available at SSRN: `https://ssrn.com/abstract=2977306`

4 Fung M.C., Peters G.W. and Shevchenko P.V.
Cohort Effects in Mortality Modelling: A Bayesian State-Space Approach (March 24, 2017).
Available at SSRN: `https://ssrn.com/abstract=2907868`