Stochastic State Space Models for Mortality

Dr. Gareth W. Peters

Chair Prof. of Statistics for Risk and Insurance
Department of Actuarial Mathematics and Statistics

Heriot-Watt University, Edinburgh, UK

D. Toczydlowska, G.W. Peters, M.C. Fung, P.V. Shevchenko.

"Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors via Probabilistic Robust Principle Components"

Risks (to appear 2017)

September 1, 2017

Table of contents

Table of contents

Mortality Modelling Context

- Ageing populations are a major challenge that many countries are facing today.
  - **Fertility rates are declining while life expectancy is increasing.**
- **longevity risk:** *the adverse financial outcome of people living longer than expected, and hence the 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 to understand the costing of different economic assumptions and regulations regarding the age of retirement of a given population.

Mortality Modelling Context

The modelling and management of systematic mortality risk are two of the main concerns of large life insurers and pension plans:

Modelling:

- What is the best way to *forecast future mortality rates* and to model the *uncertainty surrounding these forecasts*?
- How do we value risky future cashflows that depend on future mortality rates?

Management:

- How can this risk be actively managed and reduced as part of an overall strategy of efficient risk management?
- What hedging instruments are easier to price than others?

Mortality Modelling Context

Stylized Facts of Mortality Data: Enhancing mortality models requires an understanding of common features of mortality behaviour (see discussion in [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.
  - *For example, for English and Welsh males, the age 25 rate improved dramatically before 1960 and then levelled off; conversely at age 65 the opposite was true*
- Absolute decreases have varied by age group.
  - *For example, for English and Welsh males, the age 45 improvements have been much higher than the age 85 improvements.*
- Aggregate mortality rates have significant volatility year on year.

Mortality Modelling Context

Consequences on Insurance Sector and Governments:

- **Prior to 2000's:** *the UK and Australia defined-benefit pension plans had limited exposure to effects of longevity risk since high equity returns on pension fund wealth management portfolios were masking the impact of longevity risk*
- **Post 2000:** *declining equity returns coupled with record low interest rate financial environments has demonstrated the significance of decades of longevity improvements, posing a very real problem for pension schemes.*

- Furthermore, by regulation, insurers who offer retirement income products are required to hold additional reserving capital to cover longevity risk.
- A **key input to address longevity risk is the development of advanced mortality modelling methodology**.

Table of contents

Stochastic Mortality Models

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, $D(t, x)$, will be random.
  - larger population $\Rightarrow$ 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.**

Background on Stochastic Mortality Modelling

- Pricing of retirement income products depends crucially on the accuracy of the predicted death or survival probabilities.

- *It has been widely documented that* **survival probability is consistently underestimated** *especially in the last few decades ([IMF, 2012]).*

- To capture the stochastic nature of mortality trends, [Lee and Carter, 92] proposed a stochastic mortality model to forecast the trend of age-specific mortality rates.

- **Since the introduction of the Lee-Carter model, a range of stochastic mortality models have been proposed in the literature.**

Stochastic Mortality Models

Single age group models:

- Model the individual age group mortality evolution either: force of mortality [1] or annual death counts.
- Typically such models include:
  - *temporal smoothing splines*;
  - *demographic factors*;
  - can be *count processes or functional regressions* (or both);
  - *ARIMA* type structures.

Term structure of mortality (multiple age group) models:

- Typically model the log mortality rate across the term structure of mortality.
- Typically such models include:
  - *temporal smoothing splines*;
  - *period effects*; and
  - *cohort effects*.

[1] force of mortality represents the instantaneous rate of mortality at a certain age measured on an annualized basis. It is identical in concept to failure rate, also called hazard function, in reliability theory.

Stochastic Mortality Models: Regression Formulations

Generalized Linear Model Type: have been widely adopted in mortality modelling (eg.[Forfar, 1988], [Renshaw, 2003,2000,1991], [Currie,2016]).

Modelling target was either:

- the probability of death q_x, based on initial exposures; or
- the force of morality μ_x, based on central exposures.

When targeting q_x it was common to use a **binomial observation distribution**

When targeting the force of mortality it was common practice to use a **Poisson observation distribution**

Stochastic Mortality Models: Time-series Regression

Stochastic Period Effect Models: influential stochastic factor model for mortality modelling given by [Lee and Carter, 1992]

- Dynamics of the log crude death rates, $y_{x,t} = \ln \hat{m}_{x,t}$, follow:

$$y_{x,t} = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t}, \quad \varepsilon_{x,t} \overset{iid}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2),$$

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

- The **period effect**, κ_t, *for forecasting*, is typically set as

$$\kappa_t = \kappa_{t-1} + \theta + \omega_t, \quad \omega_t \overset{iid}{\sim} \mathrm{N}(0, \sigma_\omega^2),$$

where $\varepsilon_{x,t}$ and ω_t are independent.

Stochastic Mortality Models: Time-series Regression

Stochastic Period Effect Models:

- [Renshaw and Haberman,2003,2006] and [Cairns,2009] extend Lee-Carter model to include: multiple period effects and cohort effect to capture the change of mortality with respect to year and year-of-birth, respectively:
  - multi-period $(\sum_{i=1}^{k} \beta_x^{(i)} \kappa_t^{(i)})$;
  - cohort factor (ζ_{t-x}).
- [Cairns et.al., 06] proposed a two-factor period effect mortality model, known as the Cairns-Blake-Dowd (CBD) model, for pensioner ages by modelling probability of death via logit:

$$\text{logit}(q_{x,t}) := \ln\left(q_{x,t}/(1-q_{x,t})\right).$$

- [Plat, 09] combines the desirable features of the previous models and includes a *term for infant mortality.*

Stochastic Mortality Models: Time-series Regression

Extensions to the LC model:

Model	Dynamics
[Lee and Carter,92]	$\ln(\hat{m}_{x,t}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{x,t}$
[Renshaw et al, 03]	$\ln(\hat{m}_{x,t}) = \alpha_x + \sum_{i=1}^{k} \beta_x^{(i)} \kappa_t^{(i)} + \varepsilon_{x,t}$
[Renshaw et al, 06]	$\ln(\hat{m}_{x,t}) = \alpha_x + \beta_x^{(1)} \kappa_t + \beta_x^{(2)} \zeta_{t-x} + \varepsilon_{x,t}$
[Currie,06]	$\ln(\hat{m}_{x,t}) = \alpha_x + \kappa_t + \zeta_{t-x} + \varepsilon_{x,t}$
[Cairns et al.,06]	$\mathrm{logit}(q_{x,t}) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x})$
[Cairns et al., 09]	$\mathrm{logit}(q_{x,t}) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \zeta_{t-x}$
[Plat, 09]	$\ln(\hat{m}_{x,t}) = \alpha_x + \kappa_t^{(1)} + \kappa_t^{(2)}(\bar{x} - x) + \kappa_t^{(3)}(\bar{x} - x)^+ + \zeta_{t-x} + \varepsilon_{x,t}$

Recently, [Fung et al, 16] and [Fung et al, 17] have proposed new extensions based on stochastic volatility structures in the latent processes as well as non-homoscedasticity in the mortality term structures, long-memory persistance and demographic factor model distributed lags.

Background on Stochastic Mortality Modelling

Frequentist Models with demographic and economic data.

- In [Hanewald,2011] and [Niu,2014] investigate links between *economic growth and morality trends* via regression model:
  - **Single Age Group Period effect Lee-Carter model + covariate (GDP).**

- [Hanewald,2011] included *cause-of-death* categorical variables.

- [Murray and Lopez, 1997] developed a multi factor linear regression model: the *logarithm of the rate of mortality per age group, sex and clustered cause of death* is regressed against the *socio-economic, educational, technological and cause-of-death* related regressors.

- [Hyndman and Yasmeen, 2012] and [Erbas, 2010] considered dimension reduction based feature extraction methods for regressors: *functional PCA covariates from mortality curves*

Table of contents

State Space Based Stochastic Mortality Models

- A state space model is basically specification of two model components:
  - a stochastic observation equation; and
  - a stochastic latent Markov state process.

A key advantage of state space modelling is that the typical two-stage estimation and forecasting procedure under the SVD or Poisson regression maximum likelihood approaches can be combined in a single setting. This has the following advantages:

- more numerically and statistically robust than standard two stage regression modelling;
- can remove awkward identification specifications;
- is computationally more efficient; and
- can produce more accurate in-sample and out-of-sample forecasts;
- can be optimal from an efficiency and unbiased estimation perspective;
- easily adapted to Bayesian inference!

State Space Based Stochastic Mortality Models

Cohort effects: state-space formulation

Observation equation: modelling dynamics of crude death rate:

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

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

State equation for latent cohort:

- Consider a matrix of cells where the row and column corresponds to age (x) and year (t) respectively. (Assume $x = 1, \ldots, 3$ and $t = 1, \ldots, 4$ for illustration)
- The cohort factor γ_{t-x} is indexed by the year-of-birth $t - x$ and its value on each cell is displayed in the table.
- **We first notice that the value γ_{t-x} is constant on the "cohort direction", that is on the cells (x, t), $(x + 1, t + 1)$ and so on.**

Cohort effects: state-space formulation

age/year	$t = 1$	$t = 2$	$t = 3$	$t = 4$
$x = 1$	γ_0	γ_1	γ_2	γ_3
$x = 2$	γ_{-1}	γ_0	γ_1	γ_2
$x = 3$	γ_{-2}	γ_{-1}	γ_0	γ_1

Table: Values of the cohort factor γ_{t-x} on a matrix of cells (x, t).

Cohort effects: state-space formulation

Observation Equation is expressed in matrix form by letting $\gamma_t^x := \gamma_{t-x}$ to obtain:

$$\begin{pmatrix} \ln \widetilde{m}_{1,t} \\ \ln \widetilde{m}_{2,t} \\ \ln \widetilde{m}_{3,t} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \kappa_t + \begin{pmatrix} \beta_1^\gamma & 0 & 0 \\ 0 & \beta_2^\gamma & 0 \\ 0 & 0 & \beta_3^\gamma \end{pmatrix} \begin{pmatrix} \gamma_t^1 \\ \gamma_t^2 \\ \gamma_t^3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{pmatrix}.$$

As time flows from $t = 1$ to $t = 4$, the cohort vector $(\gamma_t^1, \gamma_t^2, \gamma_t^3)^\top$, which represents the cohort factor in matrix form, proceeds as

$$\begin{pmatrix} \gamma_1^1(=\gamma_0) \\ \gamma_1^2(=\gamma_{-1}) \\ \gamma_1^3(=\gamma_{-2}) \end{pmatrix} \to \begin{pmatrix} \gamma_2^1(=\gamma_1) \\ \gamma_2^2(=\gamma_0) \\ \gamma_2^3(=\gamma_{-1}) \end{pmatrix} \to \begin{pmatrix} \gamma_3^1(=\gamma_2) \\ \gamma_3^2(=\gamma_1) \\ \gamma_3^3(=\gamma_0) \end{pmatrix} \to \begin{pmatrix} \gamma_4^1(=\gamma_3) \\ \gamma_4^2(=\gamma_2) \\ \gamma_4^3(=\gamma_1) \end{pmatrix}.$$

State Space Based Stochastic Mortality Models

State-Space Formulation: Observation Process

Let $y_x = \ln \widetilde{m}_{x,t}$, in matrix notation we have (recall that $\gamma_t^x := \gamma_{t-x}$)

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

It is clear that, we have for $i \in \{1, \ldots, p\}$:

$$y_{x_i,t} = \alpha_{x_i} + \beta_{x_i}\kappa_t + \beta^{\gamma}_{x_i}\gamma_t^{x_i} + \varepsilon_{x_i,t}$$

Here $(\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^{\top}$ is the $p+1$ dimensional latent state vector.

State Space Based Stochastic Mortality Models

Cohort effects: state-space formulation

To obtain the **Cohort Latent State Equation** the key observation is that the first two elements of the cohort vector at time $t-1$ will appear as the bottom two elements of the cohort vector at time t.

Therefore, the evolution of the cohort vector must satisfy:

$$\begin{pmatrix} \gamma_t^1 \\ \gamma_t^2 \\ \gamma_t^3 \end{pmatrix} = \begin{pmatrix} * & * & * \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \gamma_{t-1}^1 \\ \gamma_{t-1}^2 \\ \gamma_{t-1}^3 \end{pmatrix} + \dots,$$

which is the **defining property of "cohort":** $\gamma_{t-x} = \gamma_{(t-i)-(x-i)}$.

Furthermore, it is therefore obvious that one only needs to model the dynamics of γ_t^1 but not γ_t^2 and γ_t^3.

State Space Based Stochastic Mortality Models

State-Space Formulation: State Process

A parsimonious state equation formulation is given in matrix form:

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

Here we assume κ_t is a random walk with drift process

$$\kappa_t = \kappa_{t-1} + \theta + \omega_t^{\kappa}, \quad \omega_t^{\kappa} \overset{iid}{\sim} \mathrm{N}(0, \sigma_{\omega}^2),$$

Dynamics of $\gamma_t^{x_1}$ is described by a stationary AR(1) process

$$\gamma_t^{x_1} = \lambda \gamma_{t-1}^{x_1} + \eta + \omega_t^{\gamma}, \quad \omega_t^{\gamma} \overset{iid}{\sim} \mathrm{N}(0, \sigma_{\gamma}^2),$$

where $|\lambda| < 1$.

State Space Based Stochastic Mortality Models

State-Space Formulation: State Process

An extended latent cohort dynamic for $\gamma_t^{x_1}$ is obtained by specifying the second row of the $p+1$ by $p+1$ matrix.

For example, one can consider generally the state equation as

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

where

$$\gamma_t^{x_1} = \lambda_1 \gamma_{t-1}^{x_1} + \lambda_2 \gamma_{t-1}^{x_2} + \cdots + \lambda_{p-1} \gamma_{t-1}^{x_{p-1}} + \lambda_p \gamma_{t-1}^{x_p} + \eta + \omega_t^{\gamma}$$

which is an ARIMA(p,0,0) process since $\gamma_{t-1}^{x_i} = \gamma_{t-i}^{x_1}$, $i = 2, \ldots, p$.

State Space Based Stochastic Mortality Models

Cohort effects: state-space formulation

We can express the matrix form succinctly as

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + B\boldsymbol{\varphi}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \stackrel{iid}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2 1_p),$$

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

where

$$B = \begin{pmatrix} \beta_{x_1} & \beta_{x_1}^{\gamma} & 0 & \cdots & 0 \\ \beta_{x_2} & 0 & \beta_{x_2}^{\gamma} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{x_p} & 0 & 0 & \cdots & \beta_{x_p}^{\gamma} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta \\ \eta \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

and $\boldsymbol{\varphi}_t = (\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^\top$, 1_p the p-dimensional identity matrix and Υ is a $p+1$ by $p+1$ diagonal matrix with diagonal $(\sigma_\kappa^2, \sigma_\gamma^2, 0, \ldots, 0)$.

Table of contents

State-Space Hybrid Factor Models for Mortality

- Extending stochastic mortality models with observable exogenous features/covariates from demographic data.

- This offers two advantages to standard Lee-Carter models:
  - *firstly they may improve predictive power of the models*
  - *secondly they may improve interpretation of the dynamic of the "term-structure" of age specific mortality rates.*

We address four new and important aspects in practice previously ignored:

1. **missing data** in time-series and panel (matrix) structured real demographic data;
2. **noisy observations and outliers** (in real data);
3. **parsimonious model** creation via dimension reduction; and
4. **optimal estimation** via **computational efficient** state-space filtering methods.

State-Space Hybrid Factor Models for Mortality

"Hybrid": a mix of observable stochastic features and latents stochastic factors

- 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*).
- Approach 2 is more appropriate for data which is **high dimensional** in nature, **time series / panel structured** but represented by relatively "short time series" lengths.
  - *This type of data is particularly prevalent in demographic studies!*
- The main concept here is that the **feature extraction** is performed over the entire available time series of observable demographic data.
- Features extracted are added to the stochastic mortality model in a static form with **dynamic latent state processes** for the factor loadings over time.

State-Space Hybrid Factor Models for Mortality

- There are numerous ways to achieve this in a state-space model:
  - the *factor may influence all age groups equally* by entering the factor into the **state equation**; or
  - the *factor may influence each age specific mortality rate differently* by adding it in the **observation equation**.
- Denote generically $\mathbf{F}_t$ as the $p \times k$ factors matrix where p may represent number of age groups and k may represent number of age specific factors.
- We then specify an additional latent pk dimensional vector $\boldsymbol{\varrho}_t$ for the factor loading for year t.
  - $\boldsymbol{\varrho}_t$ is a dynamic regression parameter for factors matrix $\mathbf{F}_t$ which specifies the impact of $x_i \in \{x_1, \dots, x_p\}$ age group and $m \in \{1, \dots, k\}$ component corresponding to $[\mathbf{F}_t]_{i,m}$ by $\varrho_t^{i,m}$ element.
- Assume $\boldsymbol{\varrho}_t$ is modelled by VAR(1) process given by

$$\boldsymbol{\varrho}_t = \boldsymbol{\Omega}\boldsymbol{\varrho}_{t-1} + \boldsymbol{\Psi} + \boldsymbol{\omega}_t^{\varrho}, \quad \boldsymbol{\omega}_t^{\varrho} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\varrho}^2 \mathbb{I}_{pk})$$

with homogeneous variant for covariance matrix of error term $\boldsymbol{\omega}_t^{\varrho}$.

State-Space Hybrid Factor Models for Mortality

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

The general notation of the model is as follows

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \tilde{\mathbf{B}}_t \tilde{\boldsymbol{\varphi}}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{I}_p),$$
$$\tilde{\boldsymbol{\varphi}}_t = \tilde{\boldsymbol{\Lambda}} \tilde{\boldsymbol{\varphi}}_{t-1} + \tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\kappa}}_t \overset{iid}{\sim} \mathcal{N}(0, \tilde{\Upsilon})$$

where $\tilde{\boldsymbol{\varphi}}_t = (\boldsymbol{\varphi}_t, \boldsymbol{\varrho}_t)$ is a $(p + pk + 1) \times 1$ latent process vector and

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

is a vector of drift parameters for state equations, where $\boldsymbol{\Psi}$ corresponds to the model of $\boldsymbol{\varrho}_t$.

We assume independence of error terms in latent variables to give a covariance matrix for the state error $\tilde{\boldsymbol{\omega}}_t$:

$$\tilde{\Upsilon} = \left(\begin{array}{c|c} \Upsilon_{(p+1)\times(p+1)} & 0 \\ \hline 0 & \sigma_\varrho^2 \mathbb{I}_{pk} \end{array} \right)_{(p+pk+1)\times(p+pk+1)}$$

State-Space Hybrid Factor Models for Mortality

Define the following two objects: $\tilde{\mathbf{F}}_t = \bigoplus_{j=1}^{k} [\mathbf{F}_t]_{j,\cdot}$ and $\tilde{\mathbf{f}}_t = vec\left(\mathbf{F}_t^T\right)$ giving:

$$\tilde{\mathbf{F}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,\cdot} & 0 & 0 & \cdots & 0 \\ 0 & [\mathbf{F}_t]_{2,\cdot} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & [\mathbf{F}_t]_{p,\cdot} \end{pmatrix}_{p \times pk} \quad \text{and } \tilde{\mathbf{f}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,1} \\ [\mathbf{F}_t]_{1,2} \\ \vdots \\ [\mathbf{F}_t]_{p,k} \end{pmatrix}_{pk \times 1}$$

where $[\mathbf{F}_t]_{j,\cdot}$ and $[\mathbf{F}_t]_{j,m}$ represent the vector of the jth row of the matrix $\mathbf{F}_t$ and the element corresponding to jth row and mth column, respectively.

Consider three cases of model:

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*.

State-Space Hybrid Factor Models for Mortality

We can now define the different sub-models:

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

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

Table of contents

Robust Probabilistic Feature Extraction Methods

- For instance, if we have d countries demographic data, where p denotes the number of different demographic attributes observed, then the $p \times d$ matrix of data in year t is $\mathbf{Y}_t$.

- We assume that $\mathbf{Y}_t$ is observed (or partially observed) over periods $t \in \{1, \ldots, T\}$.

- We do not wish to utilise the raw demographic data $\mathbf{Y}_t$: *in general it will produce a model with too many parameters* $\Rightarrow$ feature extraction methods based on minimizing some pre-specified projection pursuit index.

- We concentrate on linear methods of dimensionality reduction, more precisely, those expressible as linear projections [Friedman, 1973] which includes **Principal Component Analysis (PCA)** and its extensions and robust alternatives.

Robust Probabilistic Feature Extraction Methods

Deterministic vs. Probabilisitic PCA Method Types:

- Deterministic (observed sample based projections);
- Probabilistic population based projections;
- Partial Probabilistic PCA based projections via Factor Analysis;
- Missing Data Probabilistic PCA via Factor Analysis and Augmented Data
  $\Rightarrow$ (ideal for demographic data).
- Statistically Robust variations....

(due to time) JUMP SLIDES TO ANALYSIS

Robust Probabilistic Feature Extraction Methods

Deterministic PCA

- Simple case of $\mathbf{Y} \in \mathbb{R}^{N \times d}$ of original data: i.e. a single d - dimensional observation in a given moment of time (no missingness)
- The goal of Principal Component Analysis is to identify the most meaningful unit length basis to re-express a data set $\mathbf{Y}$.
- The purpose of a new basis is to better filter out the noise and reveal hidden structure.

Therefore, PCA looks for the given projection of the observation data

$$\mathbf{Y}_{N \times d}\mathbf{W}_{d \times d} = \mathbf{X}_{N \times d}$$

where $\mathbf{W}$ is a $d \times d$ matrix denotes a linear projection.

- The columns of $\mathbf{W}$ are the new basis vectors, that is $\mathbf{W}^T\mathbf{W} = \mathbb{I}_d$, and express rows of $\mathbf{X}$.
- Re-expressing $\mathbf{Y}$ in meaningful way means that PCA aims to lower a redundancy in data set, i.e. leads to removing the linear dependencies which provide measurements with additional noise.

Robust Probabilistic Feature Extraction Methods

Deterministic PCA

In mathematical terms, the goal can be written for i, j columns of $\mathbf{X}$

$$[\mathbf{X}]_{\cdot,i}^{T} [\mathbf{X}]_{\cdot,i} = [\mathbf{W}]_{\cdot,i}^{T} \mathbf{C}_Y [\mathbf{W}]_{\cdot,i} \,,$$

and

$$[\mathbf{X}]_{\cdot,i}^{T} [\mathbf{X}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^{T} \mathbf{C}_Y [\mathbf{W}]_{\cdot,j} = 0,$$

where $\mathbf{C}_Y = \mathbf{Y}^T\mathbf{Y}$.

We seek such a linear combination that maximizes the overall variance of $\mathbf{X}$, $\mathbf{C}_X = \mathbf{X}^T\mathbf{X}$.

The solution to the problem is found by a maximiser of the following Lagrangian expression.

$$Q(\mathbf{W}) = \mathbf{W}^T\mathbf{C}_Y\mathbf{W} - \boldsymbol{\Lambda}\left(\mathbf{W}^T\mathbf{W} - \mathbb{I}_d\right).$$

for $\boldsymbol{\Lambda}_{d\times d}$ being a diagonal $d \times d$ matrix with Lagrangian coefficients.

Robust Probabilistic Feature Extraction Methods

The roots of a quadratic form are found by setting partial derivatives to zero

$$\frac{\partial Q}{\partial \mathbf{W}} = 2\mathbf{C}_Y\mathbf{W} - 2\mathbf{\Lambda}\mathbf{W} = 0 \Rightarrow \mathbf{C}_Y\mathbf{W} = \mathbf{\Lambda}\mathbf{W}$$

- $\mathbf{W}$ is a matrix with columns as the eigenvectors of $\mathbf{C}_Y$; and
- $\mathbf{\Lambda}$ is a matrix of corresponding eigenvalues with the number of non-zero elements equal to the rank of $\mathbf{C}_Y$.

The columns of $\mathbf{X}$ indeed are orthogonal since

$$[\mathbf{X}]_{\cdot,i}^T [\mathbf{X}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^T \mathbf{C}_Y [\mathbf{W}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^T \lambda_j [\mathbf{W}]_{\cdot,j} = \lambda_j [\mathbf{W}]_{\cdot,i}^T [\mathbf{W}]_{\cdot,j} = \mathbf{0}$$

and correspond to unequal eigenvalues.
It is easily proven that $\mathbf{X}$, defined by $\mathbf{W}$ - the eigenvectors of $\mathbf{C}_Y$ satisfies that it:

- maximizes the total trace of $\mathbf{C}_X$
- maximises the determinant of $\mathbf{C}_X$ and
- maximizes the Euclidean distance between the columns of $\mathbf{X}$
- miminizes the mean square error between the observation and its projection.

Robust Probabilistic Feature Extraction Methods

Extending PCA to Stochastic Factor Analysis

- Relax the assumption that the underlying process is perfectly observed (typically assumed in PCA above)
- Assume an observation error present and the covariance matrix used in the PCA (deterministic or stochastic-population estimator based analysis) no longer explains all variation in the response or the time series demographic data.

Presenting PCA by means of Factor Analysis: with N realisations of the d-dimensional random vector, placed in the rows of the random matrices $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\epsilon}$ giving:

$$\mathbf{Y}_{n\times d} = \mathbf{X}_{n\times d}\mathbf{W}^{T}_{d\times d} + \boldsymbol{\epsilon}_{n\times d}.$$

Factor analysis assumes the diagonal covariance structure of $\boldsymbol{\epsilon}_t$.
Stochastic Factor PCA: *differs from the PCA model discussion from the previous subsections as the components given by* $\mathbf{x}_t$ *and* $\mathbf{W}$ *accounts for correlation between elements of* $\mathbf{y}_t$ *and only part of the variation* (in standard PCA $\mathbf{x}_t$ and $\mathbf{W}$ account for the entire variance) since

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

Robust Probabilistic Feature Extraction Methods

If we assume multivariate distribution of $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$ and $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ we obtain conditional independence of $\mathbf{y}_t$ given latent variable $\mathbf{x}_t$, i.e.

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

as $\boldsymbol{\Psi}$ is diagonal.

- Imposing normality assumptions on $\mathbf{y}_t$ and $\mathbf{x}_t$ enables performing ML estimation of $\mathbf{x}_t$, $\mathbf{W}$ and $\boldsymbol{\Psi}$ with optimality properties.

The marginal distribution of $\mathbf{y}_t$ is then calculated by the integration of the joint distribution of $\mathbf{y}_t$ and $\mathbf{x}_t$ which gives:

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

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 $\mathbf{W}$.

Robust Probabilistic Feature Extraction Methods

Link to Principal Component Analysis:

- If we assume that the error term ϵ_t is homogeneous, that is $\mathbf{\Psi} = \sigma^2 \mathbb{I}_d$ for $\sigma^2 > 0$, then the problem of finding $\mathbf{W}$ by means of PCA given $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbb{I}_d$ is identifiable.

Having the eigendecomposition of the covariance matrix, $\mathbf{C} = \mathbf{U}_{d\times d}\mathbf{L}_{d\times d}\mathbf{U}^T$, for diagonal matrix $\mathbf{L}$ and orthonormal matrix $\mathbf{U}$, we have

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

- Thus, the matrix $\mathbf{\Lambda} = \left(\mathbf{L} - \sigma^2 \mathbb{I}_d\right)$ 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 $\mathbf{\Lambda}$.
- Then the factors loadings are given by $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$.

Robust Probabilistic Feature Extraction Methods

Probabilistic PCA with Missing Data:

- Until now, we assumed the data did not contain any missing observations.
- However, in many demographic time series there are numerous types of missing data.
- This is therefore an important aspect to address in the feature extraction.
- When considering missing values we need to incorporate additional variables which describe a distribution of missing observations.
- Let us denote $\mathbf{y}_t = (\mathbf{y}_t^o, \mathbf{y}_t^m)$ to be a real valued d - dimensional random vector, where $\mathbf{y}_t^o$ is a sub-vector of observed entries of $\mathbf{y}_t$ and $\mathbf{y}_t^m$ is a sub-vector of unobserved entries, i.e. missing.
- The indicator random variable $\mathbf{r}_t$ decides which entries of $\mathbf{y}_t$ are missing denoting them by 1, otherwise 0.

Robust Probabilistic Feature Extraction Methods

Probabilistic PCA with Missing Data:

- Recall, that a single observation consists of the pair $(\mathbf{y}_t^o, \mathbf{r}_t)$ with distribution parameters (Θ, Θ^r) respectively.

The likelihood of parameters is proportional to the conditional probability $\mathbf{y}_t^o, \mathbf{r}_t | \Theta, \Theta^r$ that is

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

- In our study, we assume the pattern of missing data to be MAR - missing at random as defined in [Little, 2002].
- This assumptions imposes the indicator variable $\mathbf{r}_t$ to be independent of the value of missing data.

Robust Probabilistic Feature Extraction Methods

Probabilistic PCA with Missing Data:
If $\mathbf{y}_t$ is MAR it satisfies

$$\pi(\mathbf{r}_t|\mathbf{y}_t, \Theta) = \pi(\mathbf{r}_t|\mathbf{y}_t^o, \Theta).$$

which results in

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

Under the MAR assumption, the estimation of Θ via maximum likelihood of the joint distribution $\mathbf{y}_t^o, \mathbf{r}_t|\Theta, \Theta^r$ is equivalent to the maximisation of the likelihood of the marginal distribution $\mathbf{y}_t^o|\Theta$.

Robust Probabilistic Feature Extraction Methods

Efficient Probabilistic PCA with Missing Data: EM Algorithm

The algorithm is summarized by the following two steps

- **Expectation step:** Expectation of the loglikelihood function of joint distribution of $\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \sigma^2$ given by

$$\pi(\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \boldsymbol{\Psi}) = \pi(\mathbf{y}_t | \mathbf{x}_t, \mathbf{W}, \boldsymbol{\Psi}) \pi(\mathbf{x}_t | \mathbf{W}, \boldsymbol{\Psi})$$

$$= (2\pi|\boldsymbol{\Psi}|)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2}\left[\mathbf{y}_t - \mathbf{x}_t\mathbf{W}^T\right] \boldsymbol{\Psi}^{-1} \left[\mathbf{y}_t - \mathbf{x}_t\mathbf{W}^T\right]^T\right\} (2\pi)^{-\frac{d}{2}} \exp\{-\frac{1}{2}\mathbf{x}_t\mathbf{x}_t^T\}$$

  is taken with respect to conditional distribution $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$

$$Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) = \mathbb{E}_{\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2}\left\{\log\left[\mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \sigma^2}(\sigma^{*2}, \mathbf{W}^*; \mathbf{y}_{1:n}, \mathbf{x}_{1:n})\right]\right\}$$

- **Maximisation step:** Finding $\mathbf{W}^*$ and σ^{*2} that maximize $Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right)$

$$\left(\mathbf{W}^*, \sigma^{*2}\right) = \text{argmax}_{\mathbf{W}^* \in \mathbb{R}^{d\times k}, \sigma^{*2} > 0} Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right)$$

- In the non-missing data case, the previous EM steps can be solved in closed from, see [Todzwolska et al, 2017]
- In the missing data case, to proceed with the EM algorithm, we need to specify the moments of a conditional distribution of latent variables given the observation vector, when we include the latent variable $\mathbf{y}_t^m$.

Robust Probabilistic Feature Extraction Methods

The conditional distribution $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ is obtained via Bayes' rule as

$$\pi\left(\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right) = \pi\left(\mathbf{x}_t | \mathbf{y}_t, \mathbf{W}, \sigma^2\right) \pi\left(\mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right)$$

Given N realisation of $\mathbf{y}_t$ with arbitrary missing entries, the expectation step has a form

$$Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) = \mathbb{E}_{\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2}\left\{\log\left[\mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \sigma^2}(\sigma^{*2}, \mathbf{W}^*; \mathbf{y}_{1:N}, \mathbf{x}_{1:N})\right]\right\}$$

$$= \int_{\mathbb{R}^k \times \mathbb{R}^d} \pi(\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2) \log\left[\prod_{n=1}^{N} \pi\left(\mathbf{y}_n, \mathbf{x}_n | \mathbf{W}^*, \sigma^{*2}\right)\right] d\mathbf{x}_t d\mathbf{y}_t$$

$$= -\sum_{n=1}^{N}\left\{\frac{d}{2}\log\sigma^{*2} + \frac{1}{2} tr\left(\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right) + \frac{1}{2\sigma^{*2}} tr\left(\mathbb{E}\left[\mathbf{y}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right)\right.$$

$$\left. - \frac{1}{\sigma^{*2}} tr\left(\mathbf{W}^* \mathbb{E}\left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right) + \frac{1}{2\sigma^{*2}} tr\left(\mathbf{W}^{*T} \mathbf{W}^* \mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right)\right\}$$

- $\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]$ are derived as in the complete data case with an adjustment for the missing data.
- The other moments of the conditional distribution $\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ need to calculated.

Robust Probabilistic Feature Extraction Methods

The moments of joint distribution $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$.

For simplicity assume for a moment $\mathbf{y}_t = (\mathbf{y}_t^o, \mathbf{y}_t^m) \sim \mathcal{N}\left(\mathbf{0}_d, \mathbf{C}_{d\times d}\right)$ for a covariance matrix

$$\mathbf{C}_{d\times d} = \begin{bmatrix} \mathbf{C}_{oo} & \mathbf{C}_{om} \\ \mathbf{C}_{mo} & \mathbf{C}_{mm} \end{bmatrix}$$

where indexes o and m correspond to the locations of observed and missing entries of the random vector $\mathbf{y}_t$.

The joint distribution $\mathbf{y}_t | \mathbf{y}_t^o$ under MAR assumption is multivariate normal, that is

$$\mathbf{y}_t | \mathbf{y}_t^o \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{y}_t^o \\ \mathbf{y}_t^o \mathbf{C}_{oo}^{-1} \mathbf{C}_{om} \end{bmatrix}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{C}_{mo} \mathbf{C}_{oo}^{-1} \mathbf{C}_{om} \end{bmatrix} \right).$$

since

$$\pi\left(\mathbf{y}_t^m | \mathbf{y}_t^o\right) = \frac{\pi\left(\mathbf{y}_t^m, \mathbf{y}_t^o\right)}{\pi\left(\mathbf{y}_t^o\right)}$$

Robust Probabilistic Feature Extraction Methods

The covariance matrix of the marginal distribution $\mathbf{y}_t|\mathbf{W},\ \sigma^2$ can be derived as

$$\mathbf{C} = \begin{bmatrix} \mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o} & \mathbf{W}_o\mathbf{W}_m^T \\ \mathbf{W}_m\mathbf{W}_o^T & \mathbf{W}_m\mathbf{W}_m^T + \sigma^2\mathbb{I}_{d_m} \end{bmatrix} \quad (3)$$

where d_o and d_m such that $d_o + d_m = d$ are numbers of elements observed and missing (which can be zero) respectively, m and o are the indexes of matrices denote sets of rows which correspond to missing and observed values of $\mathbf{y}_t$, respectively (recall that columns of matrix $\mathbf{W}$ correspond to values of $\mathbf{x}_t$).

Robust Probabilistic Feature Extraction Methods

Theorem *The expectation of the E-step,* $\mathbb{E}_{\mathbf{x}_t|\mathbf{y}_t,\mathbf{W},\sigma^2} \log\left[\mathcal{L}_{\mathbf{y}_t,\mathbf{x}_t|\mathbf{W},\sigma^2}(\sigma^{*2},\mathbf{W}^*;\mathbf{y}_{1:n},\mathbf{x}_{1:n})\right]$, *where* $\mathbf{y}_t = (\mathbf{y}_t^o, \mathbf{y}_t^m)$, *is*

$$Q^m\left(\mathbf{W},\sigma^2|\mathbf{W}^*,\sigma^{*2}\right) = \int_{\mathbb{R}^k\times\mathbb{R}^d} \pi(\mathbf{x}_t,\mathbf{y}_t|\mathbf{y}_t^o,\mathbf{W},\sigma^2) \log\left[\prod_{n=1}^N \pi\left(\mathbf{y}_n,\mathbf{x}_n|\mathbf{W}^*,\sigma^{*2}\right)\right] d\mathbf{x}_t d\mathbf{y}_t$$

$$= -\sum_{n=1}^N \left\{ \frac{d}{2}\log\sigma^{*2} + \frac{1}{2} tr\left(\mathbb{E}\left[\mathbf{x}_n^T\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]\right) + \frac{1}{2\sigma^{*2}} tr\left(\mathbb{E}\left[\mathbf{y}_n^T\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]\right) - \frac{1}{\sigma^{*2}} tr\left(\mathbf{W}^*\mathbb{E}\left[\mathbf{x}_n^T\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]\right) + \frac{1}{2\sigma^{*2}} tr\left(\mathbf{W}^{*T}\mathbf{W}^*\mathbb{E}\left[\mathbf{x}_n^T\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]\right)\right\}$$

for the corresponding moments of the conditional distribution $\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2$

$$\mathbb{E}\left[\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]_{1\times d} = \begin{bmatrix} \mathbf{y}_n^o \\ \mathbb{E}\left[\mathbf{y}_n^m|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]\end{bmatrix}$$

$$\mathbb{E}\left[\mathbf{y}_n^T\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]_{d\times d} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{W}_m\mathbf{W}_o^T\mathbf{C}_{oo}^{-1}\mathbf{W}_o\mathbf{W}_m^T\end{bmatrix} + \mathbb{E}\left[\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]^T\mathbb{E}\left[\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]$$

$$\mathbb{E}\left[\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]_{1\times k} = \mathbb{E}\left[\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]\mathbf{W}\left(\mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_d\right)^{-1}$$

$$\mathbb{E}\left[\mathbf{x}_n^T\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]_{k\times k} = \sigma^2\left(\mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_d\right)^{-1} + \mathbb{E}\left[\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]^T\mathbb{E}\left[\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]$$

$$\mathbb{E}\left[\mathbf{x}_n^T\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]_{k\times d} = \begin{bmatrix} \mathbf{0} \\ \mathbf{W}_m - \mathbf{W}_m\mathbf{W}_o^T\mathbf{C}_{oo}^{-1}\mathbf{W}_o\end{bmatrix} + \mathbb{E}\left[\mathbf{x}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]^T\mathbb{E}\left[\mathbf{y}_n|\mathbf{y}_t^o,\mathbf{W},\sigma^2\right]$$

Robust Probabilistic Feature Extraction Methods

Theorem *The maximizers of* $Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right)$ *are the solution to the set of the problems* $\frac{\partial Q^m}{\partial \mathbf{W}^*} = 0$ *and* $\frac{\partial Q^m}{\partial \sigma^{*2}} =$ *and are given by*

$$\mathbf{W}^*_{d\times k} = \left(\sum_{n=1}^{N} \mathbb{E}\left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]^T\right)\left(\sum_{n=1}^{N} \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]^T \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right)$$

$$\sigma^{*2} = \frac{1}{Nd}\sum_{n=1}^{N} tr\left(\mathbb{E}\left[\mathbf{y}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right] - 2\mathbf{W}^*\mathbb{E}\left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right] + \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]^T \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right] \mathbf{W}^{*T}\mathbf{W}^*\right)$$

Table of contents

Mortality Data and Demographic Data Description

- The examined data consists of male and female mortality and demographic data obtained from Human Mortality Database (`http://www.mortality.org`) for European countries.

- 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.
  - The longest time series is provided by Swedish and French mortality data, starting from 1751 and 1816, respectively.
  - The shortest time series are given for Greece and Slovenia, 1983-2014 and 1981-2013, respectively.

Mortality Data and Demographic Data Description

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 per gender**: age specific information is provided for Death counts and Death Rates.
- A single observation per country in time t describes
  - a 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).

Mortality Data and Demographic Data Description

Country	**Life Expectancy** (E0)	**No. Births**	**Death Rate** (mx)	**No. Deaths**
Austria	1947 - 2014	1871 - 2014	1947 - 2014	1947 - 2014
Belarus	1959 - 2014	1959 - 2014	1959 - 2014	1959 - 2014
Belgium	1841 - 2015	1840 - 2015	1841 - 2015	1841 - 2015
Czech Republic	1950 - 2010	1947 - 2014	1950 - 2014	1950 - 2014
Denmark	1835 - 2014	1835 - 2014	1835 - 2014	1835 - 2014
Estonia	1959 - 2013	1959 - 2013	1959 - 2013	1959 - 2013
Finland	1878 - 2012	1865 - 2012	1878 - 2012	1878 - 2012
France	1816 - 2014	1806 - 2014	1816 - 2014	1816 - 2014
East Germany	1956 - 2013	1946 - 2013	1956 - 2013	1956 - 2013
West Germany	1956 - 2013	1946 - 2013	1956 - 2013	1956 - 2013
Greece	1981 - 2013	1981 - 2013	1981 - 2013	1981 - 2013
Estonia	1959 - 2013	1959 - 2013	1959 - 2013	1959 - 2013
Hungary	1950 - 2014	1950 - 2014	1950 - 2014	1950 - 2014
Iceland	1838 - 2013	1838 - 2013	1838 - 2013	1838 - 2013
Ireland	1950 - 2014	1950 - 2014	1950 - 2014	1950 - 2014

Table: Demographic data available per country (HM Database).

Mortality Data and Demographic Data Description

Country	Life Expectancy (E0)	No. Births	Death Rate (mx)	No. Deaths
Italy	1872 - 2012	1862 - 2012	1872 - 2012	1872 - 2012
Latvia	1959 - 2013	1959 - 2013	1959 - 2013	1959 - 2013
Lithuania	1959 - 2013	1959 - 2013	1959 - 2013	1959 - 2013
Luxembourg	1960 - 2014	1950 - 2014	1960 - 2014	1960 - 2014
Netherlands	1850 - 2012	1850 - 2012	1850 - 2012	1850 - 2012
Norway	1846 - 2014	1846 - 2014	1846 - 2014	1846 - 2014
Poland	1958 - 2014	1958 - 2014	1958 - 2014	1958 - 2014
Portugal	1940 - 2012	1886 - 2012	1940 - 2012	1940 - 2012
Russia	1959 - 2014	1959 - 2014	1959 - 2014	1959 - 2014
Slovakia	1950 - 2014	1950 - 2014	1950 - 2014	1950 - 2014
Slovenia	1983 - 2014	1983 - 2014	1983 - 2014	1983 - 2014
Spain	1908 - 2014	1908 - 2014	1908 - 2014	1908 - 2014
Sweden	1751 - 2014	1747 - 2014	1751 - 2014	1751 - 2014
Switzerland	1876 - 2014	1871 - 2014	1876 - 2014	1876 - 2014
United Kingdom	1922 - 2013	1922 - 2013	1922 - 2013	1922 - 2013
Ukraine	1959 - 2013	1946 - 2013	1959 - 2013	1959 - 2013

Table: Demographic data available per country (HH Database).

Table of contents

Results and Analysis

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

The models we considered in our studies were of type:

1. [LCC:] Lee-Carter model with the stochastic cohort effect.
2. [DFM-PC:] demographic factor model version 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)
Greece and Slovakia (due to short time series).

Results and Analysis

[DFM-PC:] demographic factor model version of LCC sub-models:

- [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, robust standardisation (s = non-robust standardisation);
- [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, robust standardisation (s = non-robust standardisation);

Results and Analysis

Detailed Analysis of Results and Studies can be found in:

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

Detailed Description of Estimation Method and Properties for Mortality State Space Models can be found in:

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

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

Results and Analysis

We will demonstrate a sub-set of results that illustrate some interesting properties:

- Posterior estimates of the latent state processes for $\mathbb{E}\left[\kappa_t|\mathcal{F}_t\right]$ and $\mathbb{E}\left[\gamma_{t-x}|\mathcal{F}_t\right]$ vs. time t and 95% posterior credible intervals.
- Posterior estimates of the latent factor loading processes $\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left[\boldsymbol{\rho}_s|\mathcal{F}_t\right]$ by year and age.
- Estimates of the latent state processes for the cohort effect $\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left[\gamma_{t-x}|\mathcal{F}_t\right]$ by year and age
- Posterior estimates of the factor loadings $\mathbb{E}\left[\boldsymbol{\rho}_t|\mathcal{F}_t\right]$ vs. year and country and 95% posterior credible intervals.
- **In-sample Model Selection:** MSE, BIC and DIC.
- **Out-of-sample forecast performance:** posterior predictive analysis MSEP

Results and Analysis

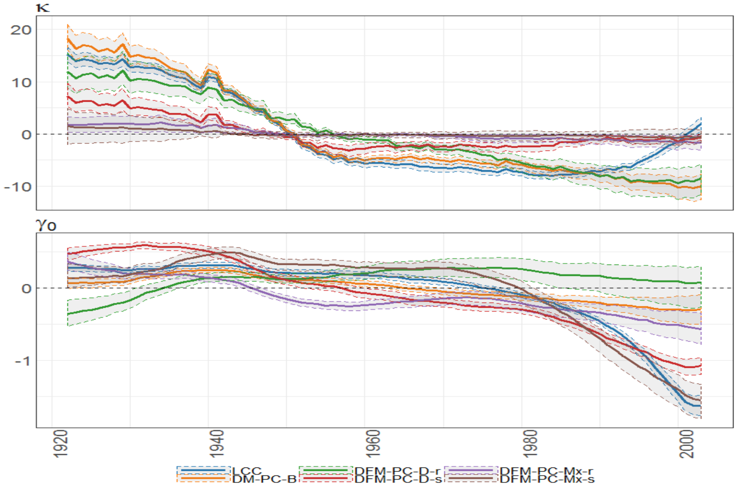

Figure: The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (upper panel) and cohort effect state process γ_t^0 (lower panel) under different models (colours of lines) for British female log death rates during 1922-2002.

Results and Analysis

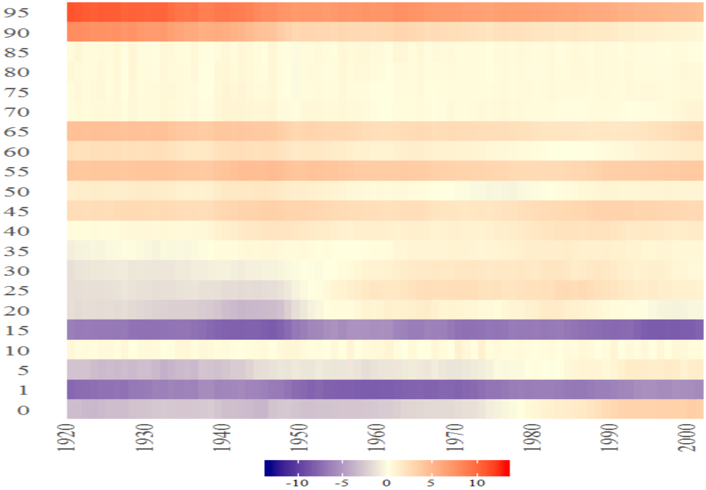

Figure: The Bayesian posterior mean estimates for ϱ_t across age groups (y axis) over time (x axis) under DFM-PC-B model for British female log death rates during 1922-2002.

Results and Analysis

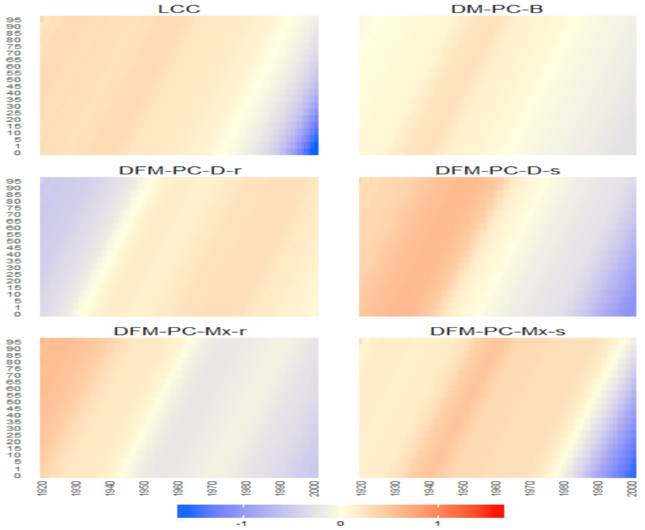

Figure: The Bayesian posterior mean estimates for the cohort effect latent processes vector γ_t across age groups (y axis) over time (x axis) under different models for British female log death rates during 1922-2002.

Results and Analysis

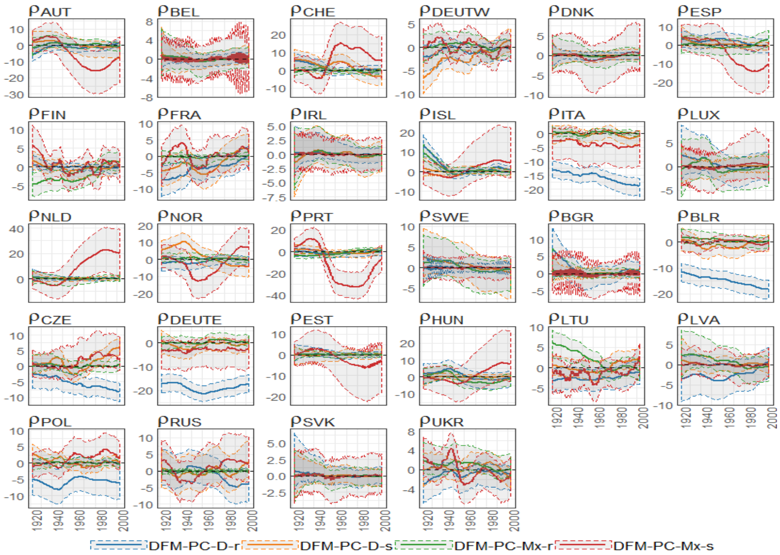

Figure: The Bayesian posterior mean estimates with 95% posterior credible intervals for ϱ_t

Results and Analysis

- Some select factors corresponding to the age specific vectors of features from European countries can have a significant influence in a causal manner on the UK morality data.
    - Specific countries have factor loadings ϱ_t indicating significant effects on UK log mortality rates at certain periods of time.
    - Whilst other countries consistently have a posterior mean loading at the origin, these countries maybe interpreted as not having an influence on the mortality experience of the UK.
- The models are more consistent about the set of countries which do not have an effect on the log death rates of United Kingdom Females.
- The models corresponding to the non-robust standardisation indicate bigger impact of western Europe countries whereas their robust alternatives indicate the significance of the patterns from Easter and Central Europe countries such as Lithuania, Poland or Russia.

Out-of-Sample Forecast Age Specific Log-Death Rates: Performance Analysis

- We choose for the out-of-sample study the last 10 years of the available sample for British Female death rates.
- Model calibration period is 1922 – 2002
  $\Rightarrow$ 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.**

Results and Analysis

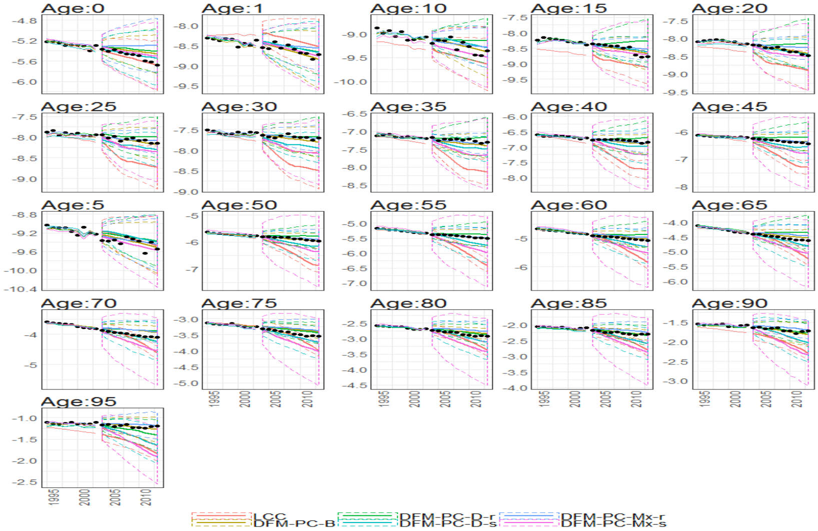

Figure: 10-year out-of-sample forecasted log death (y axis) rates by age with corresponding prediction intervals.

Conclusions

- 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.
- We see that the standard Lee-Carter stochastic mortality models consistenly under performs in-sampel and out-of-sample in a range of estimation criteria, compared to the new proposed models.
- Lee-Carter Period-Cohort model consistently under estimates log-death rates
- Extended models proposed improve significantly the forecast performance of log-death rates.

Recent Papers on Stochastic Mortality Modelling

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`