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

“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

May 16, 2019
Table of contents

1. Mortality Modelling Context
2. State Space Stochastic Mortality Models
3. State-Space Hybrid Factor Models: Regression Extensions
4. Robust Probabilistic Feature Extraction Methods
5. Application to Mortality Modelling and Demographic Data
6. Appendix
Joy of Flying with KLM!

**LOST LUGGAGE**

"I can't find my luggage, can you help me?"

"I'll try. Have you first verified that your flight has arrived?"

"Huh ?!?!"
Table of contents

1 Mortality Modelling Context

2 State Space Stochastic Mortality Models

3 State-Space Hybrid Factor Models: Regression Extensions

4 Robust Probabilistic Feature Extraction Methods

5 Application to Mortality Modelling and Demographic Data

6 Appendix
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 ⇒ 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.
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?
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.
Children per woman

Shown is the 'total fertility rate' (TFR). The TFR is the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.

Source: UN Population Division (2017 Revision)
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.

1950-55
Global Average Fertility: 4.97

1975-80
Global Average Fertility: 3.86

2005-10
Global Average Fertility: 2.5

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.
Life Expectancy by Age in England and Wales, 1700-2013

Shown is the total life expectancy given that a person reached a certain age.

Data source: Life expectancy at birth Clio-Infra. Data on life expectancy 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

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

1. Mortality Modelling Context

2. State Space Stochastic Mortality Models

3. State-Space Hybrid Factor Models: Regression Extensions

4. Robust Probabilistic Feature Extraction Methods

5. Application to Mortality Modelling and Demographic Data

6. Appendix
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.
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.
**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 $\mu_x$, 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**
Regression Formulations

**Generalized Linear Model Type:**

In the context of such models, in which age effects were deemed to be the only predictors affecting death the predictors in the GLM for an age $x$, denoted by $\eta_x$ can be linear in the regressor or non-linear, popular examples include:

<table>
<thead>
<tr>
<th>Model</th>
<th>Dynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Renshaw, 91]</td>
<td>$\eta_x = \alpha_0 + \sum_{s=1}^{k} \beta_s x^s$</td>
</tr>
<tr>
<td>[Currie, 04]</td>
<td>$\ln(m_x) = \alpha_0 + \sum_{s=1}^{k} \beta_s B_{s,3}(x^s)$</td>
</tr>
<tr>
<td>[Renshaw, 91] (M)</td>
<td>$\ln(m_x) = \alpha_0 + \alpha_1 c^x$</td>
</tr>
<tr>
<td>[Forfar, 88] (GM)</td>
<td>$\ln(m_x) = \sum_{m=0}^{r-1} \alpha_m x^m + \exp\left(\sum_{n=0}^{s-1} \beta_n x^n\right)$</td>
</tr>
</tbody>
</table>

| Table: Examples of GLM with linear and non-linear mean functions in Poisson death count models with respect to age factors. Note: $B_{s,3}(x^s)$ represents the cubic B-spline and the models discussed in [Forfar, 88] are natural applications of the models of Gompertz and Makeham denoted by M for Makeham and GM for Gompertz-Makeham. |
Generalized Linear Model Type:

- Note in the GM model it is convention that:
  - $r = 0$ implies the exponential polynomial term only; and
  - $s = 0$ implies the polynomial term only.

- Here we presented the models based on the linear predictor $\eta_x$

- $\eta_x$ is then linked to the force of moratilility $\mu_x$ or the probability of death $q_x$ ⇒
  
  *we have model with a mean function $m_x$ that we relate through a bijective response transform function (i.e inverse link function) to the linear predictor i.e. $m_x = g(\eta_x)$.

- Typically, a log-link function is considered, but other link functions are of course possible, see Section 3.3 of [Forfar, 88] for discussion on different choices in the age mortality model context.
Regression Formulations: Time Series

**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:

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

- $\alpha = \alpha_{x_1:x_p} := [\alpha_{x_1}, \ldots, \alpha_{x_p}]$ represents the **age-profile of the log death rates**

- $\beta = \beta_{x_1:x_p}$ measures the **sensitivity of of death rates for different age group** to a change of period effect $\kappa_t$.

- The **period effect**, $\kappa_t$, **for forecasting**, is typically set as

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

where $\varepsilon_{x,t}$ and $\omega_t$ 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 \((\zeta_{t-x})\).

- [Cairns et al., 2006] 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:

  \[
  \text{logit}(q_{x,t}) := \ln \left( \frac{q_{x,t}}{1 - q_{x,t}} \right).
  \]

- [Plat, 2009] 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:

<table>
<thead>
<tr>
<th>Model</th>
<th>Dynamics</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Lee and Carter,92]</td>
<td>(\ln(\hat{m}<em>x,t) = \alpha_x + \beta_x \kappa_t + \varepsilon</em>{x,t})</td>
</tr>
<tr>
<td>[Renshaw et al, 03]</td>
<td>(\ln(\hat{m}<em>x,t) = \alpha_x + \sum</em>{i=1}^{k} \beta_x^{(i)} \kappa_t^{(i)} + \varepsilon_{x,t})</td>
</tr>
<tr>
<td>[Renshaw et al, 06]</td>
<td>(\ln(\hat{m}<em>x,t) = \alpha_x + \beta_x^{(1)} \kappa_t + \beta_x^{(2)} \zeta</em>{t-x} + \varepsilon_{x,t})</td>
</tr>
<tr>
<td>[Currie,06]</td>
<td>(\ln(\hat{m}<em>x,t) = \alpha_x + \kappa_t + \zeta</em>{t-x} + \varepsilon_{x,t})</td>
</tr>
<tr>
<td>[Cairns et al.,06]</td>
<td>(\text{logit}(q_{x,t}) = \kappa_t^{(1)} + \kappa_t^{(2)} (x - \bar{x}))</td>
</tr>
<tr>
<td>[Cairns et al., 09]</td>
<td>(\text{logit}(q_{x,t}) = \kappa_t^{(1)} + \kappa_t^{(2)} (x - \bar{x}) + \zeta_{t-x})</td>
</tr>
<tr>
<td>[Plat, 09]</td>
<td>(\ln(\hat{m}<em>x,t) = \alpha_x + \kappa_t^{(1)} + \kappa_t^{(2)} (\bar{x} - x) + \kappa_t^{(3)} (\bar{x} - x)^+ + \zeta</em>{t-x} + \varepsilon_{x,t})</td>
</tr>
</tbody>
</table>

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.
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).


[Hyndman and Yasmeen, 2012] and [Erbas, 2010] considered dimension reduction based feature extraction methods for regressors: functional PCA covariates from mortality curves
Background on Stochastic Mortality Modelling

Bayesian Models with demographic and economic data.

- [Girosi and King, 2008] developed a Bayesian inference approach to build a regression framework for forecasting mortality rates which are \textit{age, sex, country and case of death specific}.
  - Used as examples applications of \textit{demography and macro-epidemiology} data as explanatory variables for the regression-type model of mortality.

- [Gaille and Sherris, 2015] studied dependency structures between \textit{cause-specific death rates} via a Vector Error Correction Model used to examine causal relations \textit{within the countries}. 
AFTER ALL THIS WORK:  

Survival probability is still consistently underestimated especially in the last few decades (IMF, 2012). This talk considers models to 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).
AFTER ALL THIS WORK:

*Survival probability is still consistently underestimated*

especially in the last few decades ([IMF, 2012]).
AFTER ALL THIS WORK: 
*Survival probability is still consistently underestimated* especially in the last few decades ([IMF, 2012]).

This talk considers models to 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])
Stochastic Mortality Modelling

- 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!
Cohort effects: state-space formulation

Observation equation: modelling dynamics of crude death rate:

\[ \ln \tilde{m}_{x,t} = \alpha_x + \beta_x \kappa_t + \beta_x^\gamma \gamma_{t-x} + \varepsilon_{x,t}, \]

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 \( \gamma_{t-x} \) 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 \( \gamma_{t-x} \) 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

<table>
<thead>
<tr>
<th>age/year</th>
<th>$t = 1$</th>
<th>$t = 2$</th>
<th>$t = 3$</th>
<th>$t = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 1$</td>
<td>$\gamma_0$</td>
<td>$\gamma_1$</td>
<td>$\gamma_2$</td>
<td>$\gamma_3$</td>
</tr>
<tr>
<td>$x = 2$</td>
<td>$\gamma_{-1}$</td>
<td>$\gamma_0$</td>
<td>$\gamma_1$</td>
<td>$\gamma_2$</td>
</tr>
<tr>
<td>$x = 3$</td>
<td>$\gamma_{-2}$</td>
<td>$\gamma_{-1}$</td>
<td>$\gamma_0$</td>
<td>$\gamma_1$</td>
</tr>
</tbody>
</table>

Table: Values of the cohort factor $\gamma_{t-x}$ on a matrix of cells $(x, t)$. 
Cohort effects: state-space formulation

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

$$
\begin{pmatrix}
\ln \tilde{m}_{1,t} \\
\ln \tilde{m}_{2,t} \\
\ln \tilde{m}_{3,t}
\end{pmatrix} =
\begin{pmatrix}
\alpha_1 \\
\alpha_2 \\
\alpha_3
\end{pmatrix} +
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3
\end{pmatrix} \kappa_t +
\begin{pmatrix}
\beta_1^\gamma & 0 & 0 \\
0 & \beta_2^\gamma & 0 \\
0 & 0 & \beta_3^\gamma
\end{pmatrix}
\begin{pmatrix}
\gamma_1^t \\
\gamma_2^t \\
\gamma_3^t
\end{pmatrix} +
\begin{pmatrix}
\varepsilon_{1,t} \\
\varepsilon_{2,t} \\
\varepsilon_{3,t}
\end{pmatrix}.
$$

As time flows from $t = 1$ to $t = 4$, the cohort vector $(\gamma_1^t, \gamma_2^t, \gamma_3^t)^\top$, which represents the cohort factor in matrix form, proceeds as

$$
\begin{pmatrix}
\gamma_1^1 (= \gamma_0) \\
\gamma_2^1 (= \gamma_{-1}) \\
\gamma_3^1 (= \gamma_{-2})
\end{pmatrix} \rightarrow
\begin{pmatrix}
\gamma_2^1 (= \gamma_1) \\
\gamma_3^2 (= \gamma_0) \\
\gamma_2^3 (= \gamma_{-1})
\end{pmatrix} \rightarrow
\begin{pmatrix}
\gamma_3^1 (= \gamma_2) \\
\gamma_3^2 (= \gamma_1) \\
\gamma_3^3 (= \gamma_0)
\end{pmatrix} \rightarrow
\begin{pmatrix}
\gamma_4^1 (= \gamma_3) \\
\gamma_4^2 (= \gamma_2) \\
\gamma_4^3 (= \gamma_1)
\end{pmatrix}.
$$
State-Space Formulation: Observation Process

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

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

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

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

Here $(\kappa_t, \gamma^x_t, \ldots, \gamma^x_t)^\top$ is the $p + 1$ dimensional latent state vector.
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:

$$
\begin{pmatrix}
\gamma^1_t \\
\gamma^2_t \\
\gamma^3_t \\
\gamma^4_t
\end{pmatrix}
= 
\begin{pmatrix}
* & * & * \\
1 & 0 & 0 \\
0 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
\gamma^1_{t-1} \\
\gamma^2_{t-1} \\
\gamma^3_{t-1}
\end{pmatrix}
+ \ldots,
$$

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 $\gamma^1_t$ but not $\gamma^2_t$ and $\gamma^3_t$. 
State Space Based Stochastic Mortality Models

State-Space Formulation: State Process

A parsimonious state equation formulation is given in matrix form:

\[
\begin{pmatrix}
\kappa_t \\
\gamma_{t1} \\
\gamma_{t2} \\
\vdots \\
\gamma_{tp-1} \\
\gamma_{tp}
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & \lambda & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
0 & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{pmatrix}
\begin{pmatrix}
\kappa_{t-1} \\
\gamma_{t-1} \\
\gamma_{t-1} \\
\vdots \\
\gamma_{tp-1} \\
\gamma_{tp-1}
\end{pmatrix}
+ \begin{pmatrix}
\theta \\
\eta \\
0 \\
\vdots \\
0 \\
0
\end{pmatrix}
+ \begin{pmatrix}
\omega_{t1} \\
\omega_{t2} \\
\omega_{t3} \\
\vdots \\
\omega_{tp-1} \\
\omega_{tp}
\end{pmatrix}.
\]

Here we assume \( \kappa_t \) is a random walk with drift process

\[
\kappa_t = \kappa_{t-1} + \theta + \omega_{t1}^{\kappa}, \quad \omega_{t1}^{\kappa} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\omega}^2),
\]

Dynamics of \( \gamma_{t1} \) is described by a stationary AR(1) process

\[
\gamma_{t1} = \lambda \gamma_{t-1} + \eta + \omega_{t1}^{\gamma}, \quad \omega_{t1}^{\gamma} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\gamma}^2),
\]

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

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

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 $\gamma_{t-1}^{x_i} = \gamma_{t-i}^{x_1}$,

$i = 2, \ldots, p$. 
Cohort effects: state-space formulation

We can express the matrix form succinctly as

\[ y_t = \alpha + B\varphi_t + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 I_p), \]
\[ \varphi_t = \Lambda \varphi_{t-1} + \Theta + \omega_t, \quad \omega_t \overset{iid}{\sim} \mathcal{N}(0, \Upsilon), \]

where

\[
\begin{pmatrix}
\beta_{x1} & \beta_{x1} & 0 & \cdots & 0 \\
0 & \beta_{x2} & \beta_{x2} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \beta_{xp}
\end{pmatrix}, \quad \Lambda =
\begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & \lambda & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{pmatrix}, \quad \Theta =
\begin{pmatrix}
\theta \\
\eta \\
0 \\
\vdots \\
0 \\
0
\end{pmatrix},
\]

and \( \varphi_t = (\kappa_t, \gamma_{t1}^{x1}, \ldots, \gamma_{tp}^{xp})^\top \), \( I_p \) the \( p \)-dimensional identity matrix and \( \Upsilon \) is a \( p + 1 \) by \( p + 1 \) diagonal matrix with diagonal \((\sigma_{\kappa}^2, \sigma_{\gamma}^2, 0, \ldots, 0)\).
Table of contents

1 Mortality Modelling Context

2 State Space Stochastic Mortality Models

3 State-Space Hybrid Factor Models: Regression Extensions

4 Robust Probabilistic Feature Extraction Methods

5 Application to Mortality Modelling and Demographic Data

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

- **Hybrid** := Stochastic Latent Factors + Observable Covariate Features

- **observable features** extracted 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.
[Toczydlowska and Peters, 2017] 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.
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!
Consider the State-Space Hybrid Period-Cohort-Demographic Model

\[ y_t = \alpha + \tilde{B}_t \tilde{\phi}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_p), \]

\[ \tilde{\phi}_t = \tilde{\Lambda} \tilde{\phi}_{t-1} + \tilde{\Theta} + \tilde{\omega}_t, \quad \tilde{\omega}_t \sim \mathcal{N}(0, \tilde{\Gamma}) \]

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

\[ \tilde{\Theta} = \begin{pmatrix} \Theta_{(p+1) \times 1} \\ \Psi_{pk \times 1} \end{pmatrix}_{(p+pk+1) \times 1} \]

a vector of drift parameters for state equations.
Define the following two objects: $\tilde{F}_t = \bigoplus_{j=1}^{k} [F_t]_j$, and $\tilde{f}_t = \text{vec} \left( F^T_t \right)$ giving:

$$\tilde{F}_t = \begin{pmatrix}
[F_t]_1, & 0 & 0 & \cdots & 0 \\
0 & [F_t]_2, & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & [F_t]_{p,} \\
0 & \cdots & \cdots & \cdots & [F_t]_{p,}
\end{pmatrix}_{p \times pk}$$

and

$$\tilde{f}_t = \begin{pmatrix}
[F_t]_{1,1} \\
[F_t]_{1,2} \\
\vdots \\
[F_t]_{p,k}
\end{pmatrix}_{pk \times 1}$$

where $F_t$ is the $p \times k$ factors matrix where $p$ may represent number of age groups and $k$ may represent number of age specific factors.
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.

\[
\begin{align*}
\tilde{B}_t \times (p+pk+1) & = \begin{cases} 
\begin{pmatrix} B_{p \times (p+1)} & \tilde{F}_t \\ B_{p \times (p+1)} & 0_{p \times pk} \end{pmatrix} & \text{for Case 1,} \\
\end{cases} \\
\tilde{\Lambda}_{(p+pk+1) \times (p+pk+1)} & = \begin{cases} 
\begin{pmatrix} \Lambda_{(p+1) \times (p+1)} & 0_{(p+1) \times pk} \\ 0_{pk \times (p+1)} & \Omega_{pk \times pk} \end{pmatrix} & \text{for Case 1,} \\
\begin{pmatrix} \Lambda_{(p+1) \times (p+1)} & \tilde{f}_t^T \\ 0_{pk \times (p+1)} & \Omega_{pk \times pk} \end{pmatrix} & \text{for Case 2,} \\
\begin{pmatrix} \Lambda_{(p+1) \times (p+1)} & 0_{1 \times pk} \\ 0_{pk \times (p+1)} & \Omega_{pk \times pk} \end{pmatrix} & \text{for Case 3.} 
\end{cases}
\end{align*}
\]
Probabilistic Feature Extraction

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

  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 $Y_t$.

- We do not wish to utilise the raw demographic data $\tilde{F}_t \neq Y_t$:

  *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.
Deterministic vs. Probabilistic 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
  ⇒ (ideal for demographic data).
- Statistically Robust variations....
Robust Probabilistic Feature Extraction Methods

Deterministic PCA

- Simple case of $Y \in \mathbb{R}^{N \times d}$ 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 $Y$.

- 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

$$Y_{N \times d} W_{d \times d} = X_{N \times d}$$

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

- The columns of $W$ are the new basis vectors, that is $W^T W = \mathbb{I}_d$, and express rows of $X$.

- Re-expressing $Y$ 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.
Deterministic PCA

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

$$[X]_{.,i} [X]_{.,i} = [W]_{.,i} C_Y [W]_{.,i},$$

and

$$[X]_{.,i} [X]_{.,j} = [W]_{.,i} C_Y [W]_{.,j} = 0,$$

where $C_Y = Y^T Y$.

*We seek such a linear combination that maximizes the overall variance of $X$, $C_X = X^T X$.*

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

$$Q(W) = W^T C_Y W - \Lambda \left(W^T W - I_d\right).$$

for $\Lambda_{d \times d}$ being a diagonal $d \times d$ matrix with Lagrangian coefficients.
The roots of a quadratic form are found by setting partial derivatives to zero

$$\frac{\partial Q}{\partial W} = 2C_YW - 2\Lambda W = 0 \Rightarrow C_YW = \Lambda W$$

- \(W\) is a matrix with columns as the eigenvectors of \(C_Y\); and
- \(\Lambda\) is a matrix of corresponding eigenvalues with the number of non-zero elements equal to the rank of \(C_Y\).

The columns of \(X\) indeed are orthogonal since

$$[X]_{.,i}^T[X]_{.,j} = [W]_{.,i}^T C_Y [W]_{.,j} = [W]_{.,i}^T \lambda_j [W]_{.,j} = \lambda_j [W]_{.,i} [W]_{.,j} = 0$$

and correspond to unequal eigenvalues.
It is easily proven that $X$, defined by $W$ - the eigenvectors of $C_Y$ satisfies that it:

- maximizes the total trace of $C_X$
- maximises the determinant of $C_X$ and
- maximizes the Euclidean distance between the columns of $X$
- minimizes the mean square error between the observation and its projection.
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.
PCA by means of Factor Analysis: with \( n \) realisations of the \((p \times d')\)-dimensional observed demographic data, vectorized into columns \( Y \).
Consider linear decompositions:

\[
Y_{n \times pd} = X_{n \times pd} W^T_{pd \times pd} + \epsilon_{n \times pd}.
\]

Factor analysis assumes diagonal covariance for \( \epsilon_t \).

Stochastic Factor PCA: differs from deterministic PCA as components \( x_t \) and factor loading matrix \( W \) account for correlation between elements of \( y_t \) and only part of the variation:

\[
\mathbb{E} y_t^T y_t = \mathbb{E} \left[ \left( x_t W^T + \epsilon_t \right)^T \left( x_t W^T + \epsilon_t \right) \right] = W \Lambda W^T + \Psi.
\]

In standard PCA \( x_t \) and \( W \) account for the entire covariance.
Show $x_t$ and $W$ account for correlation!

Example: assume $x_t \sim \mathcal{N}(\mathbf{0}, I_d)$ and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \Psi)$ to obtain,

$$y_t|x_t, W, \Psi \sim \mathcal{N}\left(x_tW^T, \Psi\right),$$

$$\pi(y_t|W, \Psi) = \int_{\mathbb{R}^d} \pi(y_t, x_t|W, \Psi)dx_t = (2\pi)^{-\frac{d}{2}} |C|^{-1} \exp\left\{-\frac{1}{2}y_tC^{-1}y_t\right\}$$

for $C = WW^T + \Psi$ where $|C|$ denotes the determinant of the matrix.

- Notice that since $\Psi$ is diagonal, the correlation structure between components $y_t$ is specified by the matrix $W$. 

Show $x_t$ and $W$ account for correlation cont.

Eigen decomposition of covariance $C = U_d \times d L_d \times d U^T$, for diagonal $L$ and orthonormal $U$, gives

$$0 = (C - L)U = \left(W^T W + \sigma^2 I_d - L\right)U = \left(WW^T - (L - \sigma^2 I_d)\right)U.$$

- Thus, the matrix $\Lambda = (L - \sigma^2 I_d)$ and $U$ are matrices of eigenvalues and corresponding eigenvectors of $WW^T$.
- Since $\lambda_i = l_i - \sigma^2 \geq 0$, the scalar $\sigma^2$ can be chosen as the smallest diagonal element of $\Lambda$.
- **Factor loadings are given by $U\Lambda^{1/2}$.**

Assuming the error term $\epsilon_t$ is homogeneous s.t. $\Psi = \sigma^2 I_d$, then estimating $W$ via PCA given $C = WW^T + \sigma^2 I_d$ is identifiable.
Probabilistic Feature Extraction

Feature Extraction via EM Algorithm Estimation!

Goal is to estimate:
- projection matrix \( W \),
- vector \( \mu \) and
- scalar \( \sigma^2 \)

given marginal distribution of \( Y_t \)

\[
Y_t | \psi \sim \mathcal{N} \left( \mu, WW^T + \sigma^2 I_d \right)
\]

for the vector of static parameters \( \psi = [W, \mu, \sigma^2] \) of the model.

The EM algorithm uses logarithm of the complete data likelihood, i.e. the joint distribution of \( Y_{1:N}, X_{1:N} | \psi \) given by

\[
\pi_{Y_{1:N}, X_{1:N} | \psi} (y_{1:N}, x_{1:N}) = \prod_{t=1}^{N} \pi_{Y_t | x_t, \psi} (y_t) \pi_{x_t | \psi} (x_t)
\]

\[
= (2\pi)^{-\frac{N(d+k)}{2}} (\sigma^2)^{-\frac{Nd}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^{N} \left( y_t - \mu - x_t W^T \right) \left( y_t - \mu - x_t W^T \right)^T - \frac{1}{2} \sum_{t=1}^{N} x_t x_t^T \right\}.
\]
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 $\psi^*$ with respect to the conditional distribution $\mathbf{X}_{1:N} | \mathbf{Y}_{1:N}, \psi$

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

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

\[
\left( \mathbf{W}^*, \mu^*, \sigma^2 \right) = \arg \max_{\mathbf{W}^* \in \mathbb{R}^{d \times k}, \mu^* \in \mathbb{R}^d, \sigma^2 > 0} Q(\psi, '\psi^* )
\]
Theorem

The E-step of the EM algorithm for Gaussian Probabilistic Principal Component Analysis given $N$ realisations of the observation vector $Y_t$ denoted by $y_{1:N} = \{y_1, \ldots, y_N\}$ is obtained in the closed form as follows

$$Q(\psi, \psi^*) = \mathbb{E}_{X_{1:N} | Y_{1:N}, \psi} \left[ \log \pi_{Y_{1:N}, X_{1:N}} | \psi^* (y_{1:N}, x_{1:N}) \right]$$

$$= - \frac{N(d + k)}{2} \log 2\pi - \frac{Nd}{2} \log \sigma^2 - \frac{1}{2} \sum_{t=1}^{N} \left\{ \frac{1}{\sigma^2} \text{Tr} \left\{ y_t^T y_t \right\} \right\}$$

$$- \frac{2}{\sigma^2} y_t \mu^* T + \frac{1}{\sigma^2} \mu^* \mu^* T - \frac{2}{\sigma^2} \text{Tr} \left\{ W^* \mathbb{E}_{X_t | Y_t, \psi} [X_t] y_t \right\}$$

$$+ \frac{2}{\sigma^2} \mathbb{E}_{X_t | Y_t, \psi} [X_t] W^* T \mu^* T + \text{Tr} \left\{ \left( \frac{1}{\sigma^2} W^* T W^* + I_k \right) \mathbb{E}_{X_t | Y_t, \psi} [X_t^T X_t] \right\}$$

see proof in [Toczydlowska and Peters, 2017].
[Toczydlowska and Peters, 2017] show that the corresponding moments of the conditional distribution $X_t|Y,\psi$ are given by the following

$$E_{X_t|Y_t,\psi} [X_t]_{1 \times k} = (y_t - \mu)WM^{-1},$$

$$E_{X_t|Y_t,\psi} [X_t^T X_t]_{k \times k} = \sigma^2 M^{-1} + M^{-1}W^T(y_t - \mu)^T(y_t - \mu)WM^{-1}$$

where $M = W^T W + \sigma^2 I_k$. 

Theorem

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

$$
\mu^* = \tilde{\mu}(y_{1:N}; \psi) \left( \mathbb{I}_d - WM^{-1}W^*T \right) + \mu WM^{-1}W^*T
$$

$$
W^* = \tilde{C}_{\mu, \mu^*}(y_{1:N}; \psi, \psi^*)WM^{-1} \left( \sigma^2M^{-1} + M^{-1}W^T\tilde{C}_{\mu}(y_{1:N}; \psi)WM^{-1} \right)^{-1}
$$

$$
\sigma^*2 = \frac{1}{d} Tr \left\{ \tilde{C}_{\mu^*}(y_{1:N}; \psi, \psi^*) - 2W^*M^{-1}W^T\tilde{C}_{\mu, \mu^*}(y_{1:N}; \psi, \psi^*) + W^* \left( \sigma^2M^{-1} + M^{-1}W^T\tilde{C}_{\mu}(y_{1:N}; \psi)WM^{-1} \right) W^*T \right\}
$$

where

$$
\tilde{\mu}(y_{1:N}; \psi)_{1 \times d} = \frac{1}{N} \sum_{t=1}^{N} y_t, \quad \tilde{S}(y_{1:N}; \psi)_{d \times d} = \frac{1}{N} \sum_{t=1}^{N} y_t^T y_t,
$$

$$
\tilde{C}_{\mu}(y_{1:N}; \psi)_{d \times d} = \tilde{S}(y_{1:N}; \psi) - 2\mu^T\tilde{\mu}(y_{1:N}; \psi) + \mu^T\mu,
$$

$$
\tilde{C}_{\mu^*}(y_{1:N}; \psi, \psi^*)_{d \times d} = \tilde{S}(y_{1:N}; \psi) - 2\mu^{*T}\tilde{\mu}(y_{1:N}; \psi) + \mu^{*T}\mu^{*},
$$

$$
\tilde{C}_{\mu, \mu^*}(y_{1:N}; \psi, \psi^*)_{d \times d} = \tilde{S}(y_{1:N}; \psi) - (\mu + \mu^*)^T\tilde{\mu}(y_{1:N}; \psi) + \mu^T\mu.
$$
[Toczydlowska and Peters, (2018a),(2018b)] extend this PPCA framework with multiple extensions:

- Missing data in Observations: **Missing at Random framework**
- Distributional Extensions: **Student-t, Skewed and Grouped Student-t cases.**

We briefly mention the extension to 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 $y_t = (y^o_t, y^m_t)$ to be a real valued $d$-dimensional random vector, where $y^o_t$ is a sub-vector of observed entries of $y_t$ and $y^m_t$ is a sub-vector of unobserved entries, i.e. missing.

The indicator random variable $r_t$ decides which entries of $y_t$ are missing denoting them by 1, otherwise 0.
Probabilistic PCA with Missing Data:

- Recall, that a single observation consists of the pair \((y_t^o, r_t)\) with distribution parameters \((\Theta, \Theta')\) respectively.

The likelihood of parameters is proportional to the conditional probability \(y_t^o, r_t|\Theta, \Theta'\) that is

\[
\pi (y_t^o, r_t|\Theta, \Theta') = \int \pi (y_t^o, y_t^m, r_t|\Theta, \Theta') \, dy_t^m
\]

\[
= \int \pi (r_t|y_t, \Theta, \Theta') \pi (y_t|\Theta, \Theta') \, dy_t^m
\]

- 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 \(r_t\) to be independent of the value of missing data.
Probabilistic PCA with Missing Data:
If $y_t$ is MAR it satisfies

$$\pi(r_t|y_t, \Theta) = \pi(r_t|y^o_t, \Theta).$$

which results in

$$\pi(y^o_t, r_t|\Theta, \Theta^r) = \pi(r_t|y^o_t, \Theta^r) \int \pi(y_t|\Theta) \, dy^m_t$$

$$= \pi(r_t|y^o_t, \Theta^r) \pi(y^o_t|\Theta)$$

Under the MAR assumption, the estimation of $\Theta$ via maximum likelihood of the joint distribution $y^o_t, r_t|\Theta, \Theta^r$ is equivalent to the maximisation of the likelihood of the marginal distribution $y^o_t|\Theta$. 
Efficient Probabilistic PCA with Missing Data: EM

The algorithm is summarized by the following two steps

- **Expectation step**: Expectation of the loglikelihood function of joint distribution of $y_t, x_t|W, \sigma^2$ given by

  $$\pi(y_t, x_t|W, \Psi) = \pi(y_t|x_t, W, \Psi)\pi(x_t|W, \Psi)$$

  $$= (2\pi|\Psi|)^{-\frac{d}{2}} \exp \left\{ -\frac{1}{2} \left[ y_t - x_tW^T \right] \Psi^{-1} \left[ y_t - x_tW^T \right]^T \right\} (2\pi)^{-\frac{d}{2}} \exp \left\{ -\frac{1}{2} x_t x_t^T \right\}$$

  is taken with respect to conditional distribution $x_t, y_t^m|y_t^o, W, \sigma^2$

  $$Q^m(W, \sigma^2|W^*, \sigma^*2) = E_{x_t,y_t^m|y_t^o,w,\sigma^2} \left\{ \log \left[ L_{y_t,x_t|W,\sigma^2}(\sigma^*2, W^*, y_1:n, x_1:n) \right] \right\}$$

- **Maximisation step**: Finding $W^*$ and $\sigma^*2$ that maximize

  $$Q^m(W, \sigma^2|W^*, \sigma^*2)$$

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

- In the non-missing data case, the previous EM steps can be solved in closed form, 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 $y_t^m$. 
The conditional distribution $\mathbf{x}_t, \mathbf{y}^m_t|\mathbf{y}_t^o, \mathbf{W}, \sigma^2$ is obtained via Bayes’ rule as

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

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

$$
Q^m (\mathbf{W}, \sigma^2|\mathbf{W}^*, \sigma^{*2}) = \mathbb{E}_{\mathbf{x}_t, \mathbf{y}^m_t|\mathbf{y}_t^o, \mathbf{W}, \sigma^2} \left\{ \log \left[ \mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t|\mathbf{W}, \sigma^2}(\sigma^{*2}, \mathbf{W}^*, \mathbf{y}_1:N, \mathbf{x}_1:N) \right] \right\}
$$

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

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

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

- $\mathbb{E} \left[ \mathbf{x}_n^T \mathbf{x}_n|\mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]$ 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

$\mathbf{y}_t = (\mathbf{y}_t^o, \mathbf{y}_t^m) \sim \mathcal{N}(\mathbf{0}_d, \mathbf{C}_{d \times d})$ 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

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

since

$$\pi (\mathbf{y}_t^m | \mathbf{y}_t^o) = \frac{\pi (\mathbf{y}_t^m, \mathbf{y}_t^o)}{\pi (\mathbf{y}_t^o)}$$
The covariance matrix of the marginal distribution \( y_t|W \), \( \sigma^2 \) can be derived as

\[
C = \begin{bmatrix}
W_o W_o^T + \sigma^2 I_{d_o} & W_o W_m^T \\
W_m W_o^T & W_m W_m^T + \sigma^2 I_{d_m}
\end{bmatrix}
\]  \hspace{1cm} (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 \( y_t \), respectively.
Theorem The expectation of the E-step, 

\[ E_{x_t \mid y_t, w, \sigma^2} \log \left[ \mathcal{L}_{y_t, x_t \mid w, \sigma^2} (\sigma^2, W^*, y_1:n, x_1:n) \right], \] where \( y_t = (y^o_t, y^m_t) \), is

\[
Q^m \left( W, \sigma^2 \mid W^*, \sigma^* \right) = \int_{\mathbb{R}^k \times \mathbb{R}^d} \pi(x_t, y_t \mid y^o_t, W, \sigma^2) \log \left[ \prod_{n=1}^N \pi \left( y_n, x_n \mid W^*, \sigma^2 \right) \right] \, dx_t \, dy_t
\]

\[
= - \sum_{n=1}^N \left\{ \frac{d}{2} \log \sigma^* + \frac{1}{2} \text{tr} \left( \mathbb{E} \left[ x_n^T x_n \mid y^o_t, W, \sigma^2 \right] \right) + \frac{1}{2\sigma^*} \text{tr} \left( \mathbb{E} \left[ y_n^T y_n \mid y^o_t, W, \sigma^2 \right] \right) \right. \\
- \frac{1}{\sigma^2} \text{tr} \left( W^* \mathbb{E} \left[ x_n^T y_n \mid y^o_t, W, \sigma^2 \right] \right) + \frac{1}{2\sigma^*} \text{tr} \left( W^* W^* \mathbb{E} \left[ x_n^T x_n \mid y^o_t, W, \sigma^2 \right] \right) \right\}
\]

for the corresponding moments of the conditional distribution

\( x_n \mid y^o_t, W, \sigma^2 \)
Moments of the conditional distribution $x_n|y_o, W, \sigma^2$

\[
\mathbb{E}[y_n|y_o, W, \sigma^2]_{1 \times d} = \mathbb{E}[y_n^m|y_o, W, \sigma^2]
\]

\[
\mathbb{E}[y_n^T y_n|y_o, W, \sigma^2]_{d \times d} = \begin{bmatrix} 0 & 0 \\ 0 & c_{mm} - w_m w_o^T c_{oo}^{-1} w_o w_m^T \end{bmatrix} + \mathbb{E}[y_n|y_o, W, \sigma^2]^T \mathbb{E}[y_n|y_o, W, \sigma^2] + \mathbb{E}[y_n|y_o, W, \sigma^2] \mathbb{E}[y_n|y_o, W, \sigma^2]^T
\]

\[
\mathbb{E}[x_n|y_o, W, \sigma^2]_{1 \times k} = \mathbb{E}[y_n|y_o, W, \sigma^2] W (W^T W + \sigma^2 I_d)^{-1}
\]

\[
\mathbb{E}[x_n^T x_n|y_o, W, \sigma^2]_{k \times k} = \sigma^2 (W^T W + \sigma^2 I_d)^{-1} + \mathbb{E}[x_n|y_o, W, \sigma^2]^T \mathbb{E}[x_n|y_o, W, \sigma^2] + \mathbb{E}[x_n|y_o, W, \sigma^2] \mathbb{E}[x_n|y_o, W, \sigma^2]^T
\]

\[
\mathbb{E}[x_n^T y_n|y_o, W, \sigma^2]_{k \times d} = \begin{bmatrix} 0 \\ w_m - w_m w_o^T c_{oo}^{-1} w_o \end{bmatrix} + \mathbb{E}[x_n|y_o, W, \sigma^2]^T \mathbb{E}[y_n|y_o, W, \sigma^2] + \mathbb{E}[x_n|y_o, W, \sigma^2] \mathbb{E}[y_n|y_o, W, \sigma^2]^T
\]
**Theorem** The maximizers of $Q^m (W, \sigma^2 | W^*, \sigma^*2)$ are the solution to the set of the problems $\frac{\partial Q^m}{\partial W^*} = 0$ and $\frac{\partial Q^m}{\partial \sigma^*2} = 0$ and are given by

$$W^*_{d \times k} = \left( \sum_{n=1}^{N} \mathbb{E} \left[ x_n^T y_n | y_o^t, W, \sigma^2 \right] \right) \left( \sum_{n=1}^{N} \mathbb{E} \left[ x_n | y_o^t, W, \sigma^2 \right] \mathbb{E} \left[ x_n | y_o^t, W, \sigma^2 \right]^T \right)$$

$$\sigma^*2 = \frac{1}{Nd} \sum_{n=1}^{N} tr \left( \mathbb{E} \left[ y_n^T y_n | y_o^t, W, \sigma^2 \right] - 2W^* \mathbb{E} \left[ x_n^T y_n | y_o^t, W, \sigma^2 \right] \right)$$

$$+ \mathbb{E} \left[ x_n | y_o^t, W, \sigma^2 \right]^T \mathbb{E} \left[ x_n | y_o^t, W, \sigma^2 \right] W^* \mathbb{E} \left[ y_n^T y_n | y_o^t, W, \sigma^2 \right] W^*$$
Table of contents

1 Mortality Modelling Context

2 State Space Stochastic Mortality Models

3 State-Space Hybrid Factor Models: Regression Extensions

4 Robust Probabilistic Feature Extraction Methods

5 Application to Mortality Modelling and Demographic Data

6 Appendix
Application

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.
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).
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 $\kappa_t$ included.

Figure: In sample analysis residuals (left Female, right Male).
LCC State Space Model - with Period + Cohort Effects $\kappa_t, \gamma_{t-x}$ 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 \( \rho_t \);
- [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 \( \rho_t \) 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 \( \rho_t \) corresponds to a country specific subvector of the component.

r/s - is robust vs standard
Application

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

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>DIC</th>
<th>$\text{MSEP}_{MCMC}$</th>
<th>$\text{MSEP}_{\text{Kalman}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC</td>
<td>0.0097</td>
<td>-3627</td>
<td>0.1778</td>
<td>0.1774</td>
</tr>
<tr>
<td>DFM-PC-B</td>
<td><strong>0.0072</strong></td>
<td><strong>-6500</strong></td>
<td><strong>0.0057</strong></td>
<td><strong>0.0062</strong></td>
</tr>
<tr>
<td>DFM-PC-D-r</td>
<td>0.0182</td>
<td>-6380</td>
<td>0.0177</td>
<td>0.0251</td>
</tr>
<tr>
<td>DFM-PC-D-s</td>
<td><strong>0.0065</strong></td>
<td><strong>-5996</strong></td>
<td><strong>0.0185</strong></td>
<td><strong>0.0156</strong></td>
</tr>
<tr>
<td>DFM-PC-Mx-r</td>
<td><strong>0.0081</strong></td>
<td><strong>-8225</strong></td>
<td><strong>0.0111</strong></td>
<td><strong>0.0129</strong></td>
</tr>
<tr>
<td>DFM-PC-Mx-s</td>
<td>0.0174</td>
<td>-3951</td>
<td>0.0692</td>
<td>0.0285</td>
</tr>
</tbody>
</table>

- **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

- 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

1 Mortality Modelling Context
2 State Space Stochastic Mortality Models
3 State-Space Hybrid Factor Models: Regression Extensions
4 Robust Probabilistic Feature Extraction Methods
5 Application to Mortality Modelling and Demographic Data
6 Appendix
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.*
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.
Available at SSRN: https://ssrn.com/abstract=2907868