Article

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

Dorota Toczydlowska [1,*], Gareth W. Peters [1,2], Man Chung Fung [3] and Pavel V. Shevchenko [4]

[1] Department of Statistical Science, University College London, 1-19 Torrington Place,
     London WC1E 7HB, UK; garethpeters78@gmail.com
[2] Man Institute of Quantitative Finance, University of Oxford, Oxford OX1 3BD, UK
[3] Data61, CSIRO, Sydney, NSW 2022, Australia; simon.fung@csiro.au
[4] Department of Applied Finance and Actuarial Studies, Macquarie University, Sydney, NSW 2109, Australia;
     pavel.shevchenko@mq.edu.au
* Correspondence: dtoczydlowska@gmail.com; Tel.: +44-207-679-1238

Abstract: In this study we develop a multi-factor extension of the family of Lee-Carter stochastic mortality models. We build upon the time, period and cohort stochastic model structure to extend it to include exogenous observable demographic features that can be used as additional factors to improve model fit and forecasting accuracy. We develop a dimension reduction feature extraction framework which (a) employs projection based techniques of dimensionality reduction; in doing this we also develop (b) a robust feature extraction framework that is amenable to different structures of demographic data; (c) we analyse demographic data sets from the patterns of missingness and the impact of such missingness on the feature extraction, and (d) introduce a class of multi-factor stochastic mortality models incorporating time, period, cohort and demographic features, which we develop within a Bayesian state-space estimation framework; finally (e) we develop an efficient combined Markov chain and filtering framework for sampling the posterior and forecasting. We undertake a detailed case study on the Human Mortality Database demographic data from European countries and we use the extracted features to better explain the term structure of mortality in the UK over time for male and female populations when compared to a pure Lee-Carter stochastic mortality model, demonstrating our feature extraction framework and consequent multi-factor mortality model improves both in sample fit and importantly out-off sample mortality forecasts by a non-trivial gain in performance.

1. Introduction

Modelling the "term-structure" of age specific mortality rates by gender and country has enjoyed a growing resurgence in the actuarial and statistics literature. This is primarily driven by the importance of better understanding and forecasting age specific mortality rates for purposes of understanding longevity risk, pension design, annuities pricing and population studies.

The most widely utilised class of stochastic mortality models in actuarial science and statistics arise from the class of regression or state-space models that incorporate explanatory factors which correspond to stylised latent stochastic factors representing structural features in the evolution of the age specific mortality rates. Typically these latent stochastic features are interpreted as either temporal

effects, period effects and cohort effects. The most famous class of such models is the Lee-Carter type models, see a summary recently in Fung et al. (2017) and references therein.

In this paper we aim to combine these classes of stochastic mortality model with other observable exogenous features obtained from a range of demographic data sets. The purpose being that they offer two advantages to standard Lee-Carter models, firstly they may improve predictive power of the models, secondly they may improve the interpretation of behaviour of the dynamic of the "term-structure" of age specific mortality rates.

We expect the mortality experience and demographic data to be characterised by a strong causal and time-varying interaction. There is an existing literature on incorporation of demographic data in stochastic mortality models. However, unlike the state space mortality age-term structure dynamic factor model approach we develop in this manuscript, the existing works have been primarily focused on regression type structures that consider single age group models. Furthermore, there is limited work on feature extraction methods in this space. We highlight a few related approaches that have considered demographic data to study single age group mortality. We comment on some of the widely used exogenous factors in such studies, which include for instance macroeconomic variables, as well as demographic variables. In (Hanewald 2011) and (Niu and Melenberg 2014), the authors investigate the links between the economic growth and morality trends through a class of single age group regression models which are estimated in a frequentist estimation framework. In addition to the period effect in the standard Lee-Carter mode, the authors incorporate gross domestic product (GDP) as an observable factor what improves the in-sample and out-of-sample performance of the model.

Other classes of factors that have been explored in such settings also include cause-of-death categorical variables, what has been also partly investigated in (Hanewald 2011). The relation between causes of death and their influence on mortality has started to be more detailed explore since the accessibility of the data improved. In Murray and Lopez (1997), the authors develop the scenarios of future mortality based on a multi factor linear regression model where 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 predictors. The Bayesian inference has been adopted in (Girosi and King 2008) to build a regression framework for forecasting mortality rates which are age, sex, country and case of death specific. The work is mostly focused on the methodological side of the forecasting but uses as examples the applications of demography and macro-epidemiology data as explanatory variables for the regression-type model of mortality. Moreover, the dependency structures between cause-specific death rates are studied in (Gaille and Sherris 2015). The authors use Vector Error Correction Models to examine such causal relations within the countries.

The usage of principal components of the mortality curves as linear regressors has been examined in (Hyndman and Yasmeen 2012). The authors explored the common features of the data applying the functional version of Principal Component Analysis. The concept is further developed in (Erbas et al. 2010), where the cause-of-death-specific smoothed mortality curves are treated as functional data. The obtained principal components serve as basis functions in functional data analysis.

In the following study, we aim to broaden these concepts and investigate the impact of global mortality trends given by various sets of international demographic data, and their potential influence on the mortality experience in one country, in our case study the United Kingdom. To achieve this in a manner suitable for incorporation in multi-age stochastic mortality models we need to perform a parsimonious feature extraction method in order to reduce the large dimensional sets of data to a form suitable for inclusion in such a mortality model. Therefore, we introduce a methodology which is not exclusive to one type of demographic data and is capable of handling the analysis jointly over many different exogenous variables.

To achieve this we must undertake several tasks: the first is to explain a canonical and principled approach to combining of such demographic time series data into the stochastic mortality models, for which there is a number of structural approaches we develop and present.

The second aspect is that large demographic data sets are now available, but a naive incorporation of such features into a stochastic mortality model would result in far too many parameters to perform estimation, the models would be overfit and would not provide good generalisation properties for out-off-sample forecast performance. Therefore, we introduce a class of probabilistic, statistically robust feature extraction approaches to reduce dimensionality and capture core information present in the demographic data that can be more parsimoniously included in the stochastic mortality models. The standard concept of robust Principal Component Analysis by means of M- and S-estimators cannot be easily utilized since the demographic data is not equal length and contains missing values across different age groups. Hence, we adopt a probabilistic formulation of Principal Component Analysis which additionally allows to model the hidden process of missing values.

Another challenge is the issue of parameters uncertainty in mortality modelling which we address by adopting the Bayesian Inference framework based on efficient Markov Chain Monte Carlo as in Fung et al. (2017). The estimation of the model is achieved via a Rao-Blackwellised Gibbs sampler. We sample the static parameters via conjugate Gibbs sampling steps which are followed by Forward Backward filtering sampler for state variables to inference from the resulting posteriors.

The contributions of each part of the paper are as follows. Firstly, we briefly overview the concept of the mortality modelling and discuss the state-space formulation of the Lee-Carter model with cohort effects and impose identification contains. Section 3 provides with several illustrations of how to incorporate observable factors into Lee- Carter models. The discussion is followed by an introduction to features extraction by means of Principal Component Analysis. Section 5 extends the standard Principal Component Analysis terminology to the probabilistic setting and derives the steps of its estimation via Expectation-Maximisation Algorithm in order to combine Principal Component with missingness. An overview of the data is given in Section 7 whereas numerical illustrations of empirical studies are presented in Section 8. Finally, Section 11 concludes.

2. Period and Cohort Effect Stochastic Mortality Models: State-Space Formulation

We begin this section by briefly recalling the classical two factor period effect and cohort effect models that have been proposed in Renshaw and Haberman (2006); Pedroza (2006) and Kogure and Kurachi (2010). This includes in particular the state-space formulation of such models which was developed in Fung et al. (2016) and Fung et al. (2017).

Extension of Lee-Carter model to cohort features proposed in (Renshaw and Haberman 2006) introduces the concept of the stochastic cohort factor, denoted by γ_{t-x}, is incorporated into the one factor stochastic Lee-Carter (Lee and Carter 1992) stochastic period effect, denoted κ_t to produce a two factor stochastic model. This second cohort factor, like the period effect factor, can also have an age-modulating coefficient, denoted by β_x^γ. In this work we adopt the recommendations discussed in (Cairns et al. 2009); (Haberman and Renshaw 2011) and (Hunt and Villegas 2015), where it is proposed to simplify this feature to be a constant age-modulating coefficient accross all age groups, given by:

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x}, \tag{1}$$

where $x \in \{x_1, \ldots, x_p\}$ and $t \in \{1, \ldots, T\}$ represent age and year respectively. $m_{x,t}$ denotes the mortality rate in age group x and time t. α_x and β_x are the age specific static parameters of the model. This simplifying assumption that $\beta_x^\gamma = 1$ (or generally any constant other than one) is supposed to improve estimation performance when fitting the model in practice. Furthermore, there is discussion in the literature to argue that it may also be justified based on empirical findings. By studying the mortality experience of England and Wales males, the study of (Willets 2004) finds that the cohort effect is not "wearing off" with increasing ages and the mortality improvement rates, defined as $1 - m_{x,t}/m_{x,t-1}$, of different cohorts seem to be rather stable. Together with the consideration of the convergence problem, one may argue that it is indeed reasonable to assume that $\beta_x^\gamma = 1$ to ensure estimation can be successfully performed for a range of mortality data while the explanatory power of the simplified model is comparable to the full model.

Next we recall the two-factor state-space formulation of the Lee-Carter type period-cohort models for stochastic mortality, see derivations and properties in Fung et al. (2016) and Fung et al. (2017). Note, we adopt the same standard notation as proposed in these papers to present the models in this manuscript.

The formulation of stochastic period-cohort models in state-space form is given by specification of both an observation equation and a state equation. Let $y_{x,t} = \log \hat{m}_{x,t}$ where $x = x_1, \ldots, x_p$ and $t = 1, \ldots, T$. The general form of the observation equation (when β_x^γ is flexible) of the cohort model Equation (1) is given in matrix form by (recall that $\gamma_t^x := \gamma_{t-x}$):

$$
\begin{pmatrix} y_{x_1,t} \\ y_{x_2,t} \\ \vdots \\ y_{x_p,t} \end{pmatrix} = \begin{pmatrix} \alpha_{x_1} \\ \alpha_{x_2} \\ \vdots \\ \alpha_{x_p} \end{pmatrix} + \begin{pmatrix} \beta_{x_1} & \beta_{x_1}^\gamma & 0 & \cdots & 0 \\ \beta_{x_2} & 0 & \beta_{x_2}^\gamma & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{x_p} & 0 & 0 & \cdots & \beta_{x_p}^\gamma \end{pmatrix} \begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{x_1,t} \\ \varepsilon_{x_2,t} \\ \vdots \\ \varepsilon_{x_p,t} \end{pmatrix}, \tag{2}
$$

where iid noise terms $\varepsilon_{x,t}$ are included as we aim to model the crude death rates. In a given year t, we identify in this model the state vector as $(\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^\top$ which represents the $p+1$ dimensional latent unobserved stochastic factor driving the observed log-mortality rates. The dynamic of this stochastic state vector is then specified in the state equation. In this work, we consider the state-space model given in matrix form as follows:

$$
\begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_{p-1}} \\ \gamma_t^{x_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \kappa_{t-1} \\ \gamma_{t-1}^{x_1} \\ \gamma_{t-1}^{x_2} \\ \vdots \\ \gamma_{t-1}^{x_{p-1}} \\ \gamma_{t-1}^{x_p} \end{pmatrix} + \begin{pmatrix} \theta \\ \eta \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \omega_t^\kappa \\ \omega_t^\gamma \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{3}
$$

In this particular instance, we assume κ_t is a random walk with drift process (ARIMA(0,1,0) with a constant) and the dynamics of $\gamma_t^{x_1}$ are described by a stationary AR(1) process with drift (ARIMA(1,0,0) with a constant) where $|\lambda| < 1$. One may consider other dynamics for $\gamma_t^{x_1}$ by specifying the second row of the $p+1$ by $p+1$ matrix in Equation (3). For example, an ARIMA(2,0,0) process for $\gamma_t^{x_1}$ can be assumed if one fixes

$$
\begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_{p-1}} \\ \gamma_t^{x_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_1 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \kappa_{t-1} \\ \gamma_{t-1}^{x_1} \\ \gamma_{t-1}^{x_2} \\ \vdots \\ \gamma_{t-1}^{x_{p-1}} \\ \gamma_{t-1}^{x_p} \end{pmatrix} + \begin{pmatrix} \theta \\ \eta \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \omega_t^\kappa \\ \omega_t^\gamma \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \tag{4}
$$

The matrix form of Equations (2) and (3) can be expressed succinctly as

$$
\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\varphi}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{I}_p), \tag{5a}
$$

$$
\boldsymbol{\varphi}_t = \boldsymbol{\Lambda}\boldsymbol{\varphi}_{t-1} + \boldsymbol{\Theta} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \overset{iid}{\sim} \mathcal{N}(0, Y), \tag{5b}
$$

where $\boldsymbol{\varphi}_t = (\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^\top$, $\mathbb{I}_p$ the p-dimensional identity matrix and Y is a $p+1$ by $p+1$ diagonal matrix with diagonal $(\sigma_\kappa^2, \sigma_\gamma^2, 0, \ldots, 0)$. The matrices $\boldsymbol{\alpha}$, $\mathbf{B}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ can be easily identified in Equations (2) and (3). For simplicity we assume homoscedasticity in the observation equation; heteroscedasticity can be incorporated straightforwardly as developed in Fung et al. (2016).

We adopt the identification constraints which are based on Hunt and Villegas (2015) and are broadly discussed and examined in Fung et al. (2017). These are given by

$$\sum_x \beta_x = 1, \ \sum_x \beta_x^\gamma = 1, \ \sum_t \kappa_t = 0, \ \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0. \tag{6}$$

3. Demographic Factor Model Extension to the Period-Cohort Stochastic Mortality State-Space Models

In this section we demonstrate several approaches one may adopt to extend the state-space model formulations presented previously for the Period-Cohort stochastic mortality models to allow for incorporation of additional observable covariate factors. The form of factor model we develop here is generic and in future sections we will develop a framework for factor extraction from demographic data that can be used in the models developed in this section.

We note that there are two fundamental ways to develop a factor time series based regression structure for incorporation of demographic data to a stochastic mortality model. We advocate in this paper an approach which is specifically developed to work with data which may be high dimensional in nature, structured but be represented by 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. The resultant features extracted are then added to the stochastic mortality model in a static form but with dynamic latent state processes for the factor loadings over time. That is the effect of the factor incorporated will be allowed to time vary through the factor loading. This approach has the advantage of not having to model explicitly the demographic data which may have a complex structure and furthermore, only requires forecasting components of the latent factor loading process. This is often significantly easier to perform since one may use for instance a standard parametric time series model such as a VAR model for their temporal evolution. Note, such approaches as this are also utilised in other financial term-structure state-space models for instance in the context of yield curve modelling, such as the dynamic Nelson-Siegel model of Diebold and Li (2006). However, we believe to the best of our knowledge we are the first to propose this type of factor model framework for incorporating demographic factors into stochastic mortality models.

We assume that we have an available set of factors that can be country, population, gender or age specific features. We wish to incorporate these age-specific and country-specific demographic/population information into the cohort state-space model described by Equations (5a) and (5b).

We will denote by $\mathbf{F}_t$ the $p \times k$ factors matrix where p may represent number of age groups and k may represent number of age specific factors. As in any feature based regression factor analysis such as PCA Regression (Jolliffe (2002)), we treat $\mathbf{F}_t$ as extracted via feature extraction methods from an exogenous observable input that is believed to have potential influence on the age specific mortality rates under study in the responses, over time. Then for each feature vector regressor, extracted from the exogenous demographic data, we will add this feature to the state-space model.

There are numerous structural ways to achieve this in a state-space model. For instance, the factor may either influence all age groups equally by entering the factor into the state equation, or it may influence each age specific mortality rate differently by adding it in the observation equation. Of course, there may also be a combination of such approaches, dependent on which demographic data the feature was extracted from the context of the model construction.

The influence of the feature on the log mortality is reasonable to assume it varies over time, so to achieve this we will specify a time dynamic for the regression factor loading. This requires that we

specify an additional latent variable, a *pk* dimensional vector ϱ_t which denotes the vector of factor loading for year *t*. We assume ϱ_t to be modelled by VAR(1) process given by

$$\varrho_t = \mathbf{\Omega}\varrho_{t-1} + \mathbf{\Psi} + \boldsymbol{\omega}_t^\varrho, \quad \boldsymbol{\omega}_t^\varrho \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varrho^2 \mathbb{I}_{pk}) \tag{7}$$

with homogeneous variant for covariance matrix of error term $\boldsymbol{\omega}_t^\varrho$. ϱ_t is a dynamic regression parameter for the 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.

As noted, depending on the interpretability of the desired model, one may incorporate $\mathbf{F}_t$ into observation Equation (5a) (**Case 1**) or into the latent dynamic of either calendar year factor period effect κ_t (**Case 2**) or cohort factors vector γ_t (**Case 3**) from the state Equation (5b).

Next we develop the extended model of Equations (5a) and (5b), which incorporates information $\mathbf{F}_t$. The general notation of the model is as follows

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \tilde{\mathbf{B}}_t \tilde{\boldsymbol{\varphi}}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{I}_p), \tag{8a}$$

$$\tilde{\boldsymbol{\varphi}}_t = \tilde{\mathbf{\Lambda}} \tilde{\boldsymbol{\varphi}}_{t-1} + \tilde{\mathbf{\Theta}} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\kappa}}_t \overset{iid}{\sim} \mathcal{N}(0, \tilde{\mathbf{Y}}) \tag{8b}$$

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

$$\tilde{\mathbf{\Theta}} = \begin{pmatrix} \mathbf{\Theta}_{(p+1)\times 1} \\ \mathbf{\Psi}_{pk \times 1} \end{pmatrix}_{(p+pk+1)\times 1} \tag{9}$$

is a vector of drift parameters for state equations, where $\mathbf{\Psi}$ corresponds to the model Equation (7) of ϱ_t. We assume independence of error terms in latent variables what gives the following structure of a covariance matrix for the state equation error term $\tilde{\boldsymbol{\omega}}_t$

$$\tilde{\mathbf{Y}} = \left(\begin{array}{c|c} \mathbf{Y}_{(p+1)\times(p+1)} & 0 \\ \hline 0 & \sigma_\varrho^2 \mathbb{I}_{pk} \end{array} \right)_{(p+pk+1)\times(p+pk+1)} \tag{10}$$

Let us specify the following two objects, $\tilde{\mathbf{F}}_t = \bigoplus_{j=1}^k [\mathbf{F}_t]_{j,\cdot}$, for $\bigoplus$ being a direct sum operator, and $\tilde{\mathbf{f}}_t = vec\left(\mathbf{F}_t^T\right)$, that is

$$\tilde{\mathbf{F}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,\cdot} & 0 & 0 & \cdots & 0 \\ 0 & [\mathbf{F}_t]_{2,\cdot} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & [\mathbf{F}_t]_{p,\cdot} \end{pmatrix}_{p \times pk} \quad \text{and} \quad \tilde{\mathbf{f}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,1} \\ [\mathbf{F}_t]_{1,2} \\ \vdots \\ [\mathbf{F}_t]_{p,k} \end{pmatrix}_{pk \times 1} \tag{11}$$

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. The structures of the other matrices and vectors for extended model Equations (8a) and (8b) depend on the introduced cases, that is

$$
\tilde{\mathbf{B}}_{t\ p\times(p+pk+1)} =
\begin{cases}
\left(\mathbf{B}_{p\times(p+1)} \;\middle|\; \tilde{\mathbf{F}}_t \right) & \text{for } \textbf{Case 1}, \\[2ex]
\left(\mathbf{B}_{p\times(p+1)} \;\middle|\; \mathbf{0}_{p\times pk} \right) & \text{otherwise},
\end{cases}
$$

$$
\tilde{\mathbf{\Lambda}}_{(p+pk+1)\times(p+pk+1)} =
\begin{cases}
\left(\begin{array}{c|c} \mathbf{\Lambda}_{(p+1)\times(p+1)} & \mathbf{0}_{(p+1)\times pk} \\ \hline \mathbf{0}_{pk\times(p+1)} & \mathbf{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \textbf{Case 1}, \\[4ex]
\left(\begin{array}{c|c} \mathbf{\Lambda}_{(p+1)\times(p+1)} & \begin{array}{c}\tilde{\mathbf{f}}_t^T \\ \mathbf{0}_{p\times pk}\end{array} \\ \hline \mathbf{0}_{pk\times(p+1)} & \mathbf{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \textbf{Case 2}, \\[4ex]
\left(\begin{array}{c|c} \mathbf{\Lambda}_{(p+1)\times(p+1)} & \begin{array}{c}\mathbf{0}_{1\times pk} \\ \tilde{\mathbf{F}}_t\end{array} \\ \hline \mathbf{0}_{pk\times(p+1)} & \mathbf{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \textbf{Case 3}.
\end{cases}
\tag{12}
$$

Given the formulated demographic factor model of the period-cohort stochastic mortality state-space models, we will denote this class of model by notation (DFM-PC). The estimation of this model will be achieved through a Bayesian model formulation and a specialised Markov chain Monte Carlo sampling framework based on Forward-Backward sampler for the latent state components and block-Gibbs conjugate sampling for the static model parameters. This follows closely the detailed framework developed extensively in Fung et al. (2016) and Fung et al. (2017), therefore, we repeat this only in relevant details in the Appendix A.

Remark 1. *The matrix $\tilde{\mathbf{F}}_t$ contains the observed (feature extracted) with exogenous factors and is a component of the model which is conditioned. Therefore it does not require to be estimated and, at this point, represents a known deterministic constant. As such $\tilde{\mathbf{F}}_t$ is fixed and held constant for the time period which the factor model is run. Hence, it is not a parameter but a covariate which is observed and deterministic. Therefore, the identifications constraints given by Equation (6) are valid for the new model.*

Remark 2. *Time series in the study undertaken are typically of the length less than 100. Compared to the number of available samples, the classical Lee-Carter model with cohort effect requires large number of parameters and has been already reported to be prone to overfitting as noted in (Cairns et al. 2009); (Haberman and Renshaw 2011) and (Hunt and Villegas 2015). The new model with the latent process introduced in Equation (7) provides with additional number of parameters. Therefore to decrease the risk of overfitting we choose to stay with the assumption of VAR(1) process for ϱ_t. We believe that the parsimony argument to keep the autoregressive structure without too many additional parameters was more important that adding more lags in this study.*

4. Approaches to Demographic Feature Extraction via Robust Probabilistic Principal Components

As we will demonstrate in the application in this paper there is a variety of issues to consider when undertaking feature extraction in demographic population data of relevance to modelling in stochastic mortality models such as those presented previously. In addition, we would like to make sure we achieve a parsimonious model presentation, where we extract features from the demographic data that are the most informative.

For instance, if we have d countries demographic data to consider where p denotes the number of different demographic attributes observed that can be considered, then the $p \times d$ matrix of this

data in year t will be denoted by $\mathbf{Y}_t$. 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 $\mathbf{Y}_t$ as in general it will produce a model with too many parameters, therefore we resort to feature extraction methods based on minimizing some pre-specified projection pursuit index.

Our attention is placed to linear methods of dimensionality reduction, more precisely, those expressible as linear projections as defined in (Friedman and Tukey 1974) which includes Principal Component Analysis (PCA) and its extensions and robust alternatives. In this paper, we have focussed on the use of Principal Component Analysis, hence, we incorporate the basis vectors of the projected lower rank space as the most meaningful factors in terms of variation.

In this regard we consider obtaining the column wise pre-whitened $\mathbf{Y}_t$ which we can then estimate the sample mean and sample covariance matrix for demographic time series data $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, which will be achieved robustly in this paper. This approach then produces a lower rank matrix which is obtained by projection according to

$$\mathbf{X}_t = \mathbf{Y}_t \mathbf{F}$$

for $\mathbf{F}$ being the first k selected eigenvectors of the covariance matrix robustly estimated from sample demographic date $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$. These factors are then entered into the state-space model as presented previously.

In the following subsections we introduce progressively the feature extraction methods that should be considered for demographic data, which have been developed to deal with real data issues such as missing data and outliers which may effect the feature extraction process.

4.1. Non-Stochastic Principal Component Analysis

Let us denote $N \times d$ matrix $\mathbf{Y}$ as original data set, where a row of the matrix is a single d-dimensional observation in a given moment of time. 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

$$\mathbf{Y}_{N \times d} \mathbf{W}_{d \times d} = \mathbf{X}_{N \times d} \tag{13}$$

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 $\mathbf{X}$.

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

$$[\mathbf{X}]_{\cdot,i}^T [\mathbf{X}]_{\cdot,i} = [\mathbf{W}]_{\cdot,i}^T \mathbf{C}_Y [\mathbf{W}]_{\cdot,i}, \tag{14}$$

and

$$[\mathbf{X}]_{\cdot,i}^T [\mathbf{X}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^T \mathbf{C}_Y [\mathbf{W}]_{\cdot,j} = 0, \tag{15}$$

where $\mathbf{C}_Y = \mathbf{Y}^T \mathbf{Y}$. We seek a linear combination given by Equation (13) 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.

$$Q(\mathbf{W}) = \mathbf{W}^T \mathbf{C}_Y \mathbf{W} - \mathbf{\Lambda} \left(\mathbf{W}^T \mathbf{W} - \mathbb{I}_d \right). \tag{16}$$

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

$$\frac{\partial Q}{\partial \mathbf{W}} = 2\mathbf{C}_Y \mathbf{W} - 2\mathbf{\Lambda} \mathbf{W} = 0 \Rightarrow \mathbf{C}_Y \mathbf{W} = \mathbf{\Lambda} \mathbf{W} \tag{17}$$

We see that $\mathbf{W}$ is a matrix which columns are eigenvectors of $\mathbf{C}_Y$ whereas $\mathbf{\Lambda}$ 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

$$[\mathbf{X}]^T_{.,i}[\mathbf{X}]_{.,j} = [\mathbf{W}]^T_{.,i}\mathbf{C}_Y[\mathbf{W}]_{.,j} = [\mathbf{W}]^T_{.,i}\lambda_j[\mathbf{W}]_{.,j} = \lambda_j[\mathbf{W}]^T_{.,i}[\mathbf{W}]_{.,j} = \mathbf{0} \tag{18}$$

and correspond to unequal eigenvalues. It is easily proven that $\mathbf{X}$, defined by $\mathbf{W}$ - the eigenvectors of $\mathbf{C}_Y$, maximizes the total trace of $\mathbf{C}_X$, its determinant and maximizes the Euclidean distance between the columns of $\mathbf{X}$, see (Basilevsky 1994). Also, the representation minimizes the mean square error between the observation and its projection as it is equivalent problem to maximizing the variance of $\mathbf{X}$.

We wish to find estimates of $\mathbf{W}$ and $\mathbf{X}$ which minimizes sum of squares, $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\mathbf{W}^T$, both of $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T$. Assuming that the residuals have homogeneous covariance matrix, that is $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sigma^2\mathbb{I}_d$ we have

$$\begin{aligned} Q(\mathbf{W},\mathbf{X}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sigma^2\mathbb{I}_d &= \left(\mathbf{Y} - \mathbf{X}\mathbf{W}^T\right)^T\left(\mathbf{Y} - \mathbf{X}\mathbf{W}^T\right) \\ &= \mathbf{Y}^T\mathbf{Y} + \mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W}^T - \mathbf{W}\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{W}^T. \end{aligned} \tag{19}$$

Since both $\mathbf{W}$ and $\mathbf{X}$ are treated as parameters to be estimated, we minimize Equation (19) by computing partial derivatives of function Q with respect to them and setting them to zero

$$\frac{\partial Q}{\partial \mathbf{W}} = -2\mathbf{Y}^T\mathbf{X} + 2\mathbf{W}\mathbf{X}^T\mathbf{X} = 0 \tag{20}$$

and since $\mathbf{YW} = \mathbf{X}$

$$\begin{aligned} \mathbf{Y}\mathbf{Y}^T\mathbf{X} &= \mathbf{Y}\mathbf{W}\mathbf{X}^T\mathbf{X} \\ \mathbf{Y}\mathbf{Y}^T\mathbf{X} &= \mathbf{X}\mathbf{X}^T\mathbf{X} \end{aligned} \tag{21}$$

As we are looking for uncorrelated explanatory variables, for $\mathbf{\Lambda} = \mathbf{X}^T\mathbf{X}$ we get

$$\mathbf{Y}\mathbf{Y}^T\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \tag{22}$$

which shows that $\mathbf{X}$ and $\mathbf{\Lambda}$ are eigenvectors and eigenvalues of the $N \times N$ matrix $\mathbf{Y}\mathbf{Y}^T$. What is more, differentiating Q with respect to $\mathbf{X}$ gives

$$\frac{\partial Q}{\partial \mathbf{X}} = -2\mathbf{Y}\mathbf{W} + 2\mathbf{X}\mathbf{W}^T\mathbf{W} = 0, \tag{23}$$

what using similar arguments as above provides with

$$\mathbf{Y}^T\mathbf{Y}\mathbf{W} = \mathbf{W}\mathbf{\Lambda}, \tag{24}$$

showing that $\mathbf{W}$ and $\mathbf{\Lambda}$ are eigenvectors and eigenvalues of the $d \times d$ covariance matrix of $\mathbf{Y}$, $\mathbf{C}_Y = \mathbf{Y}^T\mathbf{Y}$.

4.2. Stochastic Principal Component Analysis

In the following part we consider PCA from the population distribution point of view, i.e., instead of the matrix $\mathbf{Y}$, we have a d-dimensional random variable $\mathbf{y}_t$ which is linearly transformed into uncorrelated d-dimensional random variable $\mathbf{x}_t$. **At this stage the Principal Component Analysis does not require any assumption about the distribution of a random vector $\mathbf{y}_t$.** The only assumption we make refers to the projection matrix $\mathbf{W}$ and demands its orthonormality. If the random vector $\mathbf{y}_t$ has a known mean equal to zero and covariance matrix $\mathbf{C}_Y$, the model transforms to

$$\mathbf{y}_t\mathbf{W} = \mathbf{x}_t \tag{25}$$

and implies $\mathbf{x}_t$ is a d-dimensional multivariate random variable with a diagonal covariance matrix $\mathbf{\Lambda}$. If in addition we assume $\mathbf{y}_t$ to be normally distributed, the lack of correlation imposes independence.

If we consider N realisations of the random variable $\mathbf{y}_t$ which are placed in rows of the $N \times d$ matrix $\mathbf{Y}$, we have an algebraic problem as introduced in Section 4.1. The conceptual difference is that in the case of stochastic PCA we work with an estimator of covariance matrix, e.g., a sample estimator $\mathbf{S}_Y = \frac{1}{N}\mathbf{Y}^T\mathbf{Y}$.

4.2.1. Extending Stochastic Principal Component Analysis to Factor Analysis

In this section we no longer assume the underlying process to be perfectly observed as would be the assumption typically made in the stochastic version of PCA above. The implication of this can be interpreted as follows: we no longer assume that the underlying time series of demographic data is perfectly observed with no observation error. Instead there is 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. This is practically important to consider in feature extraction in practice. In this section we briefly introduce this relaxation and show its relationship to stochastic PCA above.

To discuss the PCA by means of Factor Analysis we need to introduce an additional notation and variable to our model, that is, an d-dimensional error term, $\boldsymbol{\epsilon}_t$, and rewrite Equation (25) as

$$\mathbf{y}_t = \mathbf{x}_t\mathbf{W}^T + \boldsymbol{\epsilon}_t, \tag{26}$$

where $\mathbf{y}_t$, $\mathbf{x}_t$ and $\boldsymbol{\epsilon}_t$ are d-dimensional random vectors. Given N realisations of the random vector, which are placed in the rows of the matrices $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\epsilon}$ respectively, the above problem has the following matrix form

$$\mathbf{Y}_{n \times d} = \mathbf{X}_{n \times d}\mathbf{W}^T_{d \times d} + \boldsymbol{\epsilon}_{n \times d}. \tag{27}$$

Factor analysis assumes the diagonal covariance structure of $\boldsymbol{\epsilon}_t$. It 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

$$\mathbb{E}\mathbf{y}_t^T\mathbf{y}_t = \mathbb{E}\left[\left(\mathbf{x}_t\mathbf{W}^T + \boldsymbol{\epsilon}_t\right)^T\left(\mathbf{x}_t\mathbf{W}^T + \boldsymbol{\epsilon}_t\right)\right] = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T + \mathbf{\Psi}. \tag{28}$$

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

$$\mathbf{y}_t | \mathbf{x}_t, \mathbf{W}, \mathbf{\Psi} \sim \mathcal{N}\left(\mathbf{x}_t\mathbf{W}^T, \mathbf{\Psi}\right). \tag{29}$$

as $\mathbf{\Psi}$ is diagonal. Recall that the variable $\mathbf{x}_t$ reproduces all correlations between components of $\mathbf{y}_t$. Imposing normality assumptions on $\mathbf{y}_t$ and $\mathbf{x}_t$ enables performing ML estimation of $\mathbf{x}_t$, $\mathbf{W}$ and $\mathbf{\Psi}$ 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 is given via chain rule)

$$
\begin{aligned}
\pi(\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \mathbf{\Psi}) &= \pi(\mathbf{y}_t | \mathbf{x}_t, \mathbf{W}, \mathbf{\Psi})\pi(\mathbf{x}_t | \mathbf{W}, \mathbf{\Psi}) \\
&= (2\pi|\mathbf{\Psi}|)^{-\frac{d}{2}}\exp\left\{-\frac{1}{2}\left[\mathbf{y}_t - \mathbf{x}_t\mathbf{W}^T\right]\mathbf{\Psi}^{-1}\left[\mathbf{y}_t - \mathbf{x}_t\mathbf{W}^T\right]^T\right\}(2\pi)^{-\frac{d}{2}}\exp\left\{-\frac{1}{2}\mathbf{x}_t\mathbf{x}_t^T\right\}
\end{aligned}
\tag{30}
$$

with respect to the random variable $\mathbf{x}_t$, that is

$$\pi(\mathbf{y}_t | \mathbf{W}, \mathbf{\Psi}) = \int_{\mathbb{R}^d} \pi(\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \mathbf{\Psi})d\mathbf{x}_t = (2\pi)^{-\frac{d}{2}}|\mathbf{C}|^{-1}\exp\left\{-\frac{1}{2}\mathbf{y}_t\mathbf{C}^{-1}\mathbf{y}_t^T\right\} \tag{31}$$

for $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}$ where $|\mathbf{C}|$ denotes the determinant of the matrix. Hence, $\mathbf{y}_t|\mathbf{W}, \mathbf{\Psi} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^T + \mathbf{\Psi})$. Notice that since $\mathbf{\Psi}$ is diagonal, the correlation structure between components $\mathbf{y}_t$ is specified by the matrix $\mathbf{W}$.

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 (see further discussions in (Tipping and Bishop 1999))

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

$$\mathbf{0} = (\mathbf{C} - \mathbf{L})\mathbf{U} = \left(\mathbf{W}^T\mathbf{W} + \sigma^2 \mathbb{I}_d - \mathbf{L}\right)\mathbf{U} = \left(\mathbf{W}\mathbf{W}^T - \left(\mathbf{L} - \sigma^2 \mathbb{I}_d\right)\right)\mathbf{U}. \tag{32}$$

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 $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}$.

PCA as a Limiting Case of Factor Analysis

The assumption of the isotropic error term is crucial in order to establish the link between Factor Analysis and PCA. Standard derivation of PCA does not account for any error term. However, we can perceive PCA as a limiting case for $\sigma^2 \to 0$. Then, as noted in (Roweis 1998), PCA is a limiting case of the linear Gaussian model as the covariance matrix becomes infinitesimally small and equal in all directions. This has an effect that the likelihood of a point $\mathbf{y}_t$ is dominated by the squared residuals between the observation and its projection $\mathbf{x}_t \mathbf{W}^T$. As the σ^2 tends to zero, the posteriori over states $\mathbf{x}_t$ collapses to a single point and its covariance becomes zero since

$$\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2 \sim \mathcal{N}\left(\mathbf{y}_t\mathbf{W}\left(\mathbf{W}^T\mathbf{W} + \sigma^2 \mathbb{I}_d\right)^{-1}, \sigma^2 \mathbf{M}^{-1}\right)$$

$$\xrightarrow{\sigma^2 \to 0} \delta\left(\mathbf{x}_t - \mathbf{y}_t\mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\right). \tag{33}$$

The form of conditional probability $\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2$ is justified in Section 5.1.

4.2.2. Missing Values

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. Recall, that a single observation consists of the pair $(\mathbf{y}_t^o, \mathbf{r}_t)$ with distribution parameters (Θ, Θ^r) respectively. We assume the parameters to be distinct. The likelihood of parameters is proportional to the conditional probability $\mathbf{y}_t^o, \mathbf{r}_t|\Theta, \Theta^r$ that is

$$\pi\left(\mathbf{y}_t^o, \mathbf{r}_t|\Theta, \Theta^r\right) = \int \pi\left(\mathbf{y}_t^o, \mathbf{y}_t^m, \mathbf{r}_t|\Theta, \Theta^r\right) d\mathbf{y}_t^m = \int \pi\left(\mathbf{r}_t|\mathbf{y}_t, \Theta, \Theta^r\right) \pi\left(\mathbf{y}_t|\Theta, \Theta^r\right) d\mathbf{y}_t^m \tag{34}$$

In our study, we assume the pattern of missing data to be MAR-missing at random as defined in (Little and Rubin 2002). The assumptions imposes the indicator variable $\mathbf{r}_t$ to be independent of of the value of missing data. Then the vector $\mathbf{y}_t$ which is MAR satisfies

$$\pi(\mathbf{r}_t|\mathbf{y}_t, \Theta) = \pi(\mathbf{r}_t|\mathbf{y}_t^o, \Theta). \tag{35}$$

what results in

$$\pi\left(\mathbf{y}_t^o, \mathbf{r}_t|\Theta, \Theta^r\right) = \pi\left(\mathbf{r}_t|\mathbf{y}_t^o, \Theta^r\right) \int \pi\left(\mathbf{y}_t|\Theta\right) d\mathbf{y}_t^m = \pi\left(\mathbf{r}_t|\mathbf{y}_t^o, \Theta^r\right) \pi\left(\mathbf{y}_t^o|\Theta\right) \tag{36}$$

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 $\mathbf{y}_t^o|\Theta$. Hence, we do not worry about the distribution of the indicator random variable $\mathbf{r}_t$ and the joint distribution of $\mathbf{y}_t^o$ and $\mathbf{r}_t$. If the assumption about MAR does not hold, one needs to solve the integral from Equation (34) in order to maximize the joint likelihood.

5. Efficient Probabilisitic PCA Feature Extraction in the Presence of Missingness via EM Algorithm

The combination of PCA with missingness and the Factor Analysis leads us to the "Probabilistic PCA" which can be estimated via Expectation Maximisation framework as described below. This is an exceptionally efficient and numerically stable approach to apply in practice.

Let us consider $d \times 1$ vector of observable demographic data that we wish to extract features from, where it is denoted at time t by vector $\mathbf{y}_t$. We seek k-dimensional uncorrelated latent vector $\mathbf{x}_t$ which provides the most meaningful model of $\mathbf{y}_t$,

$$\mathbf{y}_t = \mathbf{x}_t \mathbf{W}_{d \times k}^T + \boldsymbol{\epsilon}_t \tag{37}$$

We aim for $\mathbf{W}^T\mathbf{W} = \mathbb{I}_k$ (i.e., orthonormality of the projection matrix), however it is not assumed in the estimation process. We assume the multivariate normal priori distributions of the k dimensional latent variable $\mathbf{x}_t \sim \mathcal{N}\left(0, \mathbb{I}_k\right)$ and the error term $\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \sigma^2 \mathbb{I}_d\right)$. Given N realisations of observable variable $\mathbf{y}_t$, the sample model has a form

$$\mathbf{Y}_{N \times d} = \mathbf{X}_{N \times k} \mathbf{W}_{d \times k}^T + \boldsymbol{\epsilon}_{N \times d} \tag{38}$$

where single realisations are placed in rows of $\mathbf{Y}, \mathbf{X}$ and $\boldsymbol{\epsilon}$, respectively.

Our goal is to estimate coefficient matrix $\mathbf{W}$, scalar σ^2 and filter realisations of latent variable $\mathbf{x}_t$ employing Expectation-Maximisation (EM) algorithm. The steps and derivation of the algorithm have been described in (Rubin and Thayer 1982) where no missigness in assumed. The authors use the results introduced by (Dempster et al. 1977) for factors treated as missing data. The EM algorithm uses the complete data logliklehood, i.e., the logarithm of the likelihood of $\mathbf{y}_t, \mathbf{x}_t|\mathbf{W}, \sigma^2$ Equation (30) given by

$$\mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t|\mathbf{W}, \sigma^2}(\sigma^2, \mathbf{W}; \mathbf{y}_{1:N}, \mathbf{x}_{1:N}) = \prod_{n=1}^{N} \pi\left(\mathbf{y}_n, \mathbf{x}_n|\mathbf{W}, \sigma^2\right) \tag{39}$$

for $\mathbf{y}_n = [\mathbf{Y}]_{n,\cdot}$, and maximizes the expression Equation (39) which is integrated with respect to the unobserved values of $\mathbf{x}_t$. The algorithm is summarized by the following two steps

1. Expectation step: Expectation of the loglikelihood function of the join distribution of $\mathbf{y}_t, \mathbf{x}_t$ given by Equation (30) with respect to the conditional distribution $\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2$

$$Q\left(\mathbf{W}, \sigma^2|\mathbf{W}^*, \sigma^{*2}\right) = \mathbb{E}_{\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2} \log\left[\mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t|\mathbf{W}, \sigma^2}(\sigma^{*2}, \mathbf{W}^*; \mathbf{y}_{1:n}, \mathbf{x}_{1:n})\right] \tag{40}$$

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

$$\left(\mathbf{W}^*, \sigma^{*2}\right) = \underset{\mathbf{W}^* \in \mathbb{R}^{d \times k}, \sigma^{*2} > 0}{\operatorname{argmax}} Q\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) \tag{41}$$

The Expectation step (E-step) provides with the expectation Equation (40) of complete data likelihood Equation (39) based on $\mathbf{y}_{1:N}$ and assumes $\mathbf{W}$ and σ^2 to be known. It uses the observed data, current estimates of parameters and the distribution of missing values conditioned on these elements. The Maximisation step (M-step) maximizes the expectation Equation (40) with respect to $\mathbf{W}^*$ and σ^{*2} as if it was based on complete data information. In the paper (Dempster et al. 1977), the author proofs that the loglikelihood Equation (39) is non-decreasing on each iteration of the algorithm and provides with conditions that ensure its convergence (Theorem 1 and Theorem 2 in paper (Dempster et al. 1977), respectively).

We derive the steps of the algorithm using the assumptions of the normality of $\mathbf{y}_t$ and $\mathbf{x}_t$ given at the beginning of the section. As mentioned in (Dempster et al. 1977), the convexity of regular exponential families (where normal distribution belongs to) ensures the uniqueness of the maximizers computed in the M-step. Also, the normal distribution provides us with the closed forms of the moments used in subsequent steps of the algorithm. Hence, it simplifies the computations.

5.1. Expectation Step and Its Maximum

Finding the expectation Equation (40) requires specifying the conditional distribution of $\mathbf{x}_t$ given observations $\mathbf{y}_t$ and parameters. It is given via Bayes' rule as

$$\pi(\mathbf{x}_t | \mathbf{y}_t, \mathbf{W}, \sigma^2) = \frac{\pi(\mathbf{y}_t | \mathbf{x}_t, \mathbf{W}, \sigma^2) \pi(\mathbf{x}_t | \mathbf{W}, \sigma^2)}{\pi(\mathbf{y}_t | \mathbf{W}, \sigma^2)} \tag{42}$$

and results in $\mathbf{x}_t | \mathbf{y}_t, \mathbf{W}, \sigma^2 \sim \mathcal{N}\left(\mathbf{y}_t \mathbf{W} \mathbf{M}^{-1}, \sigma^2 \mathbf{M}^{-1}\right)$ for $\mathbf{M} = \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbb{I}_k$. Given N realisations of $\mathbf{y}_t$, the expectation of the logliklihood with respect to the conditional distribution of $\mathbf{x}_t$ is equal to

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

$$
\begin{aligned}
Q\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) &= \int_{\mathbb{R}^k} \pi(\mathbf{x}_t | \mathbf{y}_t, \mathbf{W}, \sigma^2) \log\left[\prod_{n=1}^N \pi\left(\mathbf{y}_n, \mathbf{x}_n | \mathbf{W}^*, \sigma^{*2}\right)\right] d\mathbf{x}_t \\
&= -\sum_{n=1}^N \left\{ \frac{d}{2} \log \sigma^2 + \frac{1}{2} tr\left(\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]\right) + \frac{1}{2\sigma^2} \mathbf{y}_n \mathbf{y}_n^T \right. \\
&\quad \left. - \frac{1}{\sigma^2} \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right] \mathbf{W}^{*T} \mathbf{y}_n^T + \frac{1}{2\sigma^2} tr\left(\mathbf{W}^{*T} \mathbf{W}^* \mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]\right) \right\}
\end{aligned} \tag{43}
$$

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

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]_{1 \times k} &= \mathbf{y}_n \mathbf{W} \mathbf{M}^{-1} \\
\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]_{k \times k} &= \sigma^2 \mathbf{M}^{-1} + \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]^T \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]
\end{aligned} \tag{44}
$$

Proof. Please refer to the results of (Rubin and Thayer 1982). □

The M-step of EM algorithm uses the computed expectation and maximizes it with respect to the static parameters $\mathbf{W}^*$ and σ^{*2}. The maximizes are given by

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

$$
\begin{aligned}
\mathbf{W}^*_{d \times k} &= \left[\sum_{n=1}^{N} \mathbf{y}_n^T \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right] \right] \left[\sum_{n=1}^{N} \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]^T \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right] \right]^{-1} \\
\sigma^{*2} &= \frac{1}{Nd} \sum_{n=1}^{N} \left\{ \mathbf{y}_n \mathbf{y}_n^T - 2\mathbf{y}_t \mathbf{W}^* \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]^T \right. \\
&\quad \left. + tr\left(\mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right]^T \mathbb{E}\left[\mathbf{x}_n | \mathbf{y}_t, \mathbf{W}, \sigma^2\right] \mathbf{W}^{*T} \mathbf{W}^* \right) \right\}.
\end{aligned}
\tag{45}
$$

The iteration over E-step and M-step provided by Theorem 1 and 2 can be replaced by iterations over one combined step, as noted in (Tipping and Bishop 1999), which for iterations i and $i + 1$ is given by

$$
\begin{aligned}
&1: \mathbf{W}^{(i+1)} = \mathbf{S}\mathbf{W}^{(i)} \left(\sigma^{(i)2}\mathbb{I}_k + \mathbf{M}^{-1}\mathbf{W}^{(i)T}\mathbf{S}\mathbf{W}^{(i)} \right)^{-1} \\
&2: \sigma^{(i+1)2} = \frac{1}{d} tr\left(\mathbf{S} - \mathbf{S}\mathbf{W}^{(i)}\mathbf{M}^{-1}\mathbf{W}^{(i+1)T} \right)
\end{aligned}
\tag{46}
$$

until the convergence of $Q\left(\mathbf{W}^{(i)}, \sigma^{(i)2} | \mathbf{W}^{(i+1)}, \sigma^{(i+1)2}\right)$ when $\mathbf{M} = \mathbf{W}^{(i)T}\mathbf{W}^{(i)} + \sigma^{(i)2}\mathbb{I}_k$ and $\mathbf{S} = \frac{1}{N}\mathbf{Y}^T\mathbf{Y}$.

5.2. The Maximum Likelihood Estimation—The Convergence of EM Algorithm

The work of (Dempster et al. 1977) proves that EM algorithm for the normal distribution always converges to local maximum. Recall that the solution provided by EM algorithm converges to the solution obtained by maximizing the likelihood of marginal distribution $\mathbf{y}_t | \mathbf{W}, \sigma^2$, given by

$$
\mathcal{L}_{\mathbf{y}_t | \mathbf{W}, \sigma^2}\left(\mathbf{W}, \sigma^2; \mathbf{y}_{1:n}\right) = -\frac{N}{2}\left[d\log(2\pi) + \log|\mathbf{C}| + tr\left(\mathbf{C}^{-1}\mathbf{S}\right) \right]
\tag{47}
$$

where $\mathbf{S}_{d \times d} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n^T \mathbf{y}_n$ and $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$. The MLE estimator of $\mathbf{W}$ given by the likelihood Equation (47) is a solution to the following fix point equation

$$
\frac{\partial \mathcal{L}_{\mathbf{y}_t | \mathbf{W}, \sigma^2}}{\partial \mathbf{W}} = N\left(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W} \right) = 0 \Rightarrow \mathbf{W} = \mathbf{S}\mathbf{C}^{-1}\mathbf{W}
\tag{48}
$$

We may distinguish tree possible cases to the above solution:

Case 1: $\mathbf{W} = 0$

The solution to this case is treated as a minimum of the log-likelihood.

Case 2: $\mathbf{S} = \mathbf{C}$

The equality implies that $d - k$ smallest eigenvalues of $\mathbf{S}$ be equal to σ^2 and the problem is indefinable since $\mathbf{W}\mathbf{W}^T = \mathbf{S} - \sigma^2\mathbb{I}_d$. Given the eigendecomposition of $\mathbf{S}$, that is

$$
\mathbf{S} = \mathbf{U}_{d \times d}\mathbf{\Lambda}_{d \times d}\mathbf{U}_{d \times d}^T
\tag{49}
$$

for orthonormal matrix $\mathbf{U}$ such that $\mathbf{U}^T\mathbf{U} = \mathbb{I}_d$ and diagonal matrix $\mathbf{\Lambda}$ with non-negative entries, the matrix $\mathbf{W}$ is equal to $\mathbf{W} = \mathbf{U}\left(\mathbf{\Lambda} - \sigma^2\mathbb{I}_d\right)^{\frac{1}{2}}\mathbf{R}^T$ where $\mathbf{R}_{k \times d}$ is an arbitrary rotation matrix. The case is proven in Section 4.2.1.

Case 3: $\mathbf{W} \neq 0$ and $\mathbf{S} \neq \mathbf{C}$

In order to compute the solution to Equation (48) we use the singular value decomposition of $\mathbf{W}$, that is

$$
\mathbf{W} = \mathbf{V}_{d \times k}\mathbf{L}_{k \times k}\mathbf{R}_{k \times k}^T
\tag{50}
$$

where $\mathbf{V}$ and $\mathbf{R}$ are real valued orthogonal matrix by columns, and $\mathbf{L}$ is non-negative diagonal matrix. Using $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbb{I}_d$ we apply the above facts to the problem defined by the fix point equation

$$
\begin{aligned}
\mathbf{W}_{d \times k} &= \mathbf{S}\mathbf{C}^{-1}\mathbf{W} \\
\mathbf{V}\mathbf{L}\mathbf{R}^T &= \mathbf{S}\left(\mathbf{V}\mathbf{L}\mathbf{R}^T\mathbf{R}\mathbf{L}\mathbf{V}^T + \sigma^2\mathbb{I}_d\right)^{-1}\mathbf{V}\mathbf{L}\mathbf{R}^T \\
\mathbf{V}\mathbf{L}\mathbf{R}^T &= \mathbf{S}\left(\mathbf{V}\mathbf{L}^2\mathbf{V}^T + \sigma^2\mathbb{I}_d\right)^{-1}\mathbf{V}\mathbf{L}\mathbf{R}^T \\
\mathbf{V}\mathbf{L}^2\mathbf{V}^T + \sigma^2\mathbb{I}_d &= \mathbf{S} \\
\mathbf{U}\mathbf{L}^2\mathbf{L} + \sigma^2\mathbb{I}_d\mathbf{V}\mathbf{L} &= \mathbf{S}\mathbf{V}\mathbf{L} \\
\mathbf{V}\left(\mathbf{L}^2 + \sigma^2\mathbb{I}_k\right)\mathbf{L} &= \mathbf{S}\mathbf{V}\mathbf{L}
\end{aligned}
\tag{51}
$$

Notice that

$$
\mathbf{S}\mathbf{v}_j = (\sigma^2 + l_j^2)\mathbf{v}_j
\tag{52}
$$

where $\mathbf{v}_j = [\mathbf{V}]_{,j}$ and $l_j = [\mathbf{L}]_{j,j}$. Hence, the vectors $\mathbf{v}_j$ are eigenvectors of the estimated covariance matrix $\mathbf{S}$. Using the eigendecomposition of $\mathbf{S}$ given by Equation (49), we see that $\mathbf{v}_j$ corresponds to the eigenvectors of $\mathbf{S}$, $\mathbf{u}_j$ with eigenvalues $\lambda_j = l_j^2 + \sigma^2$. Since $\mathbf{L}$ has a different dimension than $\mathbf{\Lambda}$, we express it as

$$
\mathbf{L} = (\mathbf{K} - \sigma^2\mathbb{I}_k)^{\frac{1}{2}}
\tag{53}
$$

where $[\mathbf{K}]_{j,j} = \lambda_j$ is an j eigenvalue of $\mathbf{S}$ and corresponds to the j eigenvector of $\mathbf{S}$, $\mathbf{u}_j$. In a case when $l_j = 0$, the eigenvector $\mathbf{v}_j$ is arbitrary and $[\mathbf{K}]_{j,j} = \sigma^2$. The scalar σ^2 is estimated as average of $i > k$ eigenvalues of $\mathbf{S}$.

The remaining question is if the EM algorithm converges to the global maximum. If the stationary points of the likelihood with respect to $\mathbf{W}$ which are spanned by minor eigenvectors (the eigenvectors with corresponding negligible eigenvalues), are stable, then the convergence is not guaranteed. However, we can show that any eigenvectors which does not correspond the highest eigenvalues of $\mathbf{S}$ is a saddle point of the logliklihood and does not provide stable solutions. For the detailed discussion please refer to (Tipping and Bishop 1999). The authors highlight the case when all $d - k$ discarded eigenvalues are equal to the smallest major principal eigenvalue(s). They show that such a situation provides with maximum spanned by principal eigenvectors and noise distribution (corresponding to the smallest principal eigenvalue(s), which become(s) zero).

5.3. The EM Algorithm for Incomplete Data

Until now, we developed the EM algorithm for Probabilistic PCA under the assumption that the data does not contain any missing values. In Section 4.2.2 we introduced the background of how we are going to treat missing entries of the observation. We assume them to be Missing-At-Random what allows us to ignore the existence of indicator variable $\mathbf{r}_t$ while estimating $\mathbf{W}$ and σ^2. Given the vector of an observation with missing entries $\mathbf{y}_t = (\mathbf{y}_t^o, \mathbf{y}_t^m)$, EM algorithms treats $\mathbf{y}_t^m$ as an additional latent variable to $\mathbf{x}_t$ in the model Equation (37). The expectation of the joint logliklihod Equation (30) is computed with respect to the conditional distribution of $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ and provides with the two following steps:

1. Expectation step: Expectation of loglikelihood function of join distribution of $\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \sigma^2$ given by Equation (30) with respect to conditional distribution $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$

$$
Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) = \mathbb{E}_{\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2}\left\{\log\left[\mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \sigma^2}(\sigma^{*2}, \mathbf{W}^*; \mathbf{y}_{1:n}, \mathbf{x}_{1:n})\right]\right\}
\tag{54}
$$

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

$$\left(\mathbf{W}^*, \sigma^{*2}\right) = \underset{\mathbf{W}^* \in \mathbb{R}^{d \times k}, \sigma^{*2} > 0}{\operatorname{argmax}} Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) \tag{55}$$

We need to specify the moments of a conditional distribution of latent variables given the observation vector, when we include the latent variable $\mathbf{y}_t^m$. The conditional distribution $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ is obtained via Bayes' rule as

$$\pi\left(\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right) = \pi\left(\mathbf{x}_t | \mathbf{y}_t, \mathbf{W}, \sigma^2\right) \pi\left(\mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right) \tag{56}$$

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

$$
\begin{aligned}
Q^m\left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2}\right) &= \mathbb{E}_{\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2}\left\{\log\left[\mathcal{L}_{\mathbf{y}_t, \mathbf{x}_t | \mathbf{W}, \sigma^2}(\sigma^{*2}, \mathbf{W}^*; \mathbf{y}_{1:N}, \mathbf{x}_{1:N})\right]\right\} \\
&= \int_{\mathbb{R}^k \times \mathbb{R}^d} \pi(\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2) \log\left[\prod_{n=1}^{N} \pi\left(\mathbf{y}_n, \mathbf{x}_n | \mathbf{W}^*, \sigma^{*2}\right)\right] d\mathbf{x}_t d\mathbf{y}_t \\
&= -\sum_{n=1}^{N}\left\{\frac{d}{2}\log\sigma^{*2} + \frac{1}{2}tr\left(\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right) + \frac{1}{2\sigma^{*2}}tr\left(\mathbb{E}\left[\mathbf{y}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right)\right. \\
&\quad \left. - \frac{1}{\sigma^{*2}}tr\left(\mathbf{W}^* \mathbb{E}\left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right) + \frac{1}{2\sigma^{*2}}tr\left(\mathbf{W}^{*T}\mathbf{W}^* \mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]\right)\right\}
\end{aligned}
\tag{57}
$$

where $\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2\right]$ are derived in Equation (44) and need adjustment for 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.

5.3.1. The Moments of Joint Distribution $\mathbf{x}_t, \mathbf{y}_t^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$.

The first component of the conditional probability Equation (56) is given by Equation (42). 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

$$\mathbf{C}_{d \times d} = \begin{bmatrix} \mathbf{C}_{oo} & \mathbf{C}_{om} \\ \mathbf{C}_{mo} & \mathbf{C}_{mm} \end{bmatrix} \tag{58}$$

where indexes o and m correspond to the locations of observed and missing entries of the random vector $\mathbf{y}_t$. As shown in (Little and Rubin 2002), the joint distribution $\mathbf{y}_t | \mathbf{y}_t^o$ under MAR assumption is multivariate normal, that is

$$\mathbf{y}_t | \mathbf{y}_t^o \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{y}_t^o \\ \mathbf{y}_t^o \mathbf{C}_{oo}^{-1} \mathbf{C}_{om} \end{bmatrix}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{C}_{mo} \mathbf{C}_{oo}^{-1} \mathbf{C}_{om} \end{bmatrix}\right). \tag{59}$$

since

$$\pi\left(\mathbf{y}_t^m | \mathbf{y}_t^o\right) = \frac{\pi\left(\mathbf{y}_t^m, \mathbf{y}_t^o\right)}{\pi\left(\mathbf{y}_t^o\right)} \tag{60}$$

As derived in (Jamshidian 1997), the covariance matrix of the marginal distribution $\mathbf{y}_t | \mathbf{W}, \sigma^2$ is equal to

$$\mathbf{C} = \begin{bmatrix} \mathbf{W}_o \mathbf{W}_o^T + \sigma^2 \mathbb{I}_{d_o} & \mathbf{W}_o \mathbf{W}_m^T \\ \mathbf{W}_m \mathbf{W}_o^T & \mathbf{W}_m \mathbf{W}_m^T + \sigma^2 \mathbb{I}_{d_m} \end{bmatrix} \tag{61}$$

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 $\mathbf{x}_t$).

Having above in mind, we can compute the following moments of the conditional distribution $\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ are given by an alternative theorem to the Theorem 1 which accounts for the incomplete data case.

Theorem 3. *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 given by*

$$
\begin{aligned}
Q^m \left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2} \right) &= \int_{\mathbb{R}^k \times \mathbb{R}^d} \pi(\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2) \log \left[\prod_{n=1}^{N} \pi \left(\mathbf{y}_n, \mathbf{x}_n | \mathbf{W}^*, \sigma^{*2} \right) \right] d\mathbf{x}_t d\mathbf{y}_t \\
&= - \sum_{n=1}^{N} \left\{ \frac{d}{2} \log \sigma^{*2} + \frac{1}{2} tr \left(\mathbb{E} \left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \right) + \frac{1}{2\sigma^{*2}} tr \left(\mathbb{E} \left[\mathbf{y}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \right) \right. \\
&\quad \left. - \frac{1}{\sigma^{*2}} tr \left(\mathbf{W}^* \mathbb{E} \left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \right) + \frac{1}{2\sigma^{*2}} tr \left(\mathbf{W}^{*T} \mathbf{W}^* \mathbb{E} \left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \right) \right\}
\end{aligned}
\tag{62}
$$

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

$$
\begin{aligned}
\mathbb{E} \left[\mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]_{1 \times d} &= \begin{bmatrix} \mathbf{y}_n^o \\ \mathbb{E} \left[\mathbf{y}_n^m | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \end{bmatrix} \\
\mathbb{E} \left[\mathbf{y}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]_{d \times d} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{W}_m \mathbf{W}_o^T \mathbf{C}_{oo}^{-1} \mathbf{W}_o \mathbf{W}_m^T \end{bmatrix} + \mathbb{E} \left[\mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]^T \mathbb{E} \left[\mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \\
\mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]_{1 \times k} &= \mathbb{E} \left[\mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \mathbf{W} \left(\mathbf{W}^T \mathbf{W} + \sigma^2 \mathbb{I}_d \right)^{-1} \\
\mathbb{E} \left[\mathbf{x}_n^T \mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]_{k \times k} &= \sigma^2 \left(\mathbf{W}^T \mathbf{W} + \sigma^2 \mathbb{I}_d \right)^{-1} + \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]^T \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \\
\mathbb{E} \left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]_{k \times d} &= \begin{bmatrix} \mathbf{0} \\ \mathbf{W}_m - \mathbf{W}_m \mathbf{W}_o^T \mathbf{C}_{oo}^{-1} \mathbf{W}_o \end{bmatrix} + \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]^T \mathbb{E} \left[\mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]
\end{aligned}
\tag{63}
$$

Proof. We can find the corresponding steps of calculation in (Jamshidian 1997). □

5.3.2. The Maximizers of $Q^m \left(\mathbf{W}, \sigma^2 | \mathbf{W}^*, \sigma^{*2} \right)$

The M-step of EM algorithm uses the computed expectation defined in Theorem 3 and maximizes it with respect to the static parameters $\mathbf{W}^*$ and σ^{*2}. The corresponding values of the maximizes are given by

Theorem 4. *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*

$$
\begin{aligned}
\mathbf{W}_{d \times k}^* &= \left(\sum_{n=1}^{N} \mathbb{E} \left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]^T \right) \left(\sum_{n=1}^{N} \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]^T \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \right)^{-1} \\
\sigma^{*2} &= \frac{1}{Nd} \sum_{n=1}^{N} tr \left(\mathbb{E} \left[\mathbf{y}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] - 2 \mathbf{W}^* \mathbb{E} \left[\mathbf{x}_n^T \mathbf{y}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \right. \\
&\quad \left. + \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right]^T \mathbb{E} \left[\mathbf{x}_n | \mathbf{y}_t^o, \mathbf{W}, \sigma^2 \right] \mathbf{W}^{*T} \mathbf{W}^* \right)
\end{aligned}
\tag{64}
$$

Proof. We need to replace the moments of conditional distribution $\mathbf{x}_t | \mathbf{y}_t, \mathbf{W}, \sigma^2$ in the proof of the Theorem 2 with the corresponding moments of $\mathbf{x}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ derived in Theorem 3. Also, we need to replace terms of $\mathbf{y}_t$ by its moments related to the joint distribution $\mathbf{x}_t, \mathbf{y}_t | \mathbf{y}_t^o, \mathbf{W}, \sigma^2$ also given in Theorem 3. □

5.4. The Algorithm

The steps of computing eigenvectors and corresponding loadings are summarized in the following algorithm. Firstly, we standardized data using information from observed values stored in $\mathbf{Y}_{N \times d}$. The estimator of location and variance, $\Theta_N = \left(\hat{\mu}, \hat{\sigma}^2 \right)$ is a function of the non missing values of observation vector $\mathbf{y}$ among its N realisations. We execute PPCA on standardized data following two steps: the expectation and maximisation step.

Algorithm 1 Probabilistic Principal Component Analysis with missing values

---

1: **for** $j = 1, \ldots, d$ **do**

2:      Compute $\Theta_N \left([\mathbf{Y}^o]_{\cdot,j} \right) = (\hat{\mu}_j, \hat{\sigma}_j^2)$

3:      Standardize data $\left[\tilde{\mathbf{Y}}^o \right]_{\cdot,j} = \frac{[\mathbf{Y}^o]_{\cdot,j} - \hat{\mu}_j}{\sigma_j}$

4: **end for**

5: $\tilde{\mathbf{Y}}^m = \mathbf{0}$ and $\tilde{\mathbf{Y}} = \left(\tilde{\mathbf{Y}}^o, \tilde{\mathbf{Y}}^m \right)$

6: Initialise: ε, $i = 0$, $\mathbf{W}^{(0)} = \mathbf{W}_0$, $\sigma^{2(0)} = \sigma_0^2$,

7: **repeat**

8:      E-step: Compute corresponding moments from Equation (62) for

$$Q^m \left(\mathbf{W}^{(i)}, \sigma^{2(i)} | \mathbf{W}^{(i)}, \sigma^{2(i)} \right)$$

9:      M-step: Compute maxima of $Q^m \left(\mathbf{W}^{(i)}, \sigma^{2(i)} | \mathbf{W}, \sigma^2 \right)$ from Equation (64):

$$\mathbf{W}^{(i+1)}, \sigma^{2(i+1)} = \text{argmax}_{\mathbf{W} \in \mathbb{R}^{d \times k}, \sigma^2 > 0} \, Q^m \left(\mathbf{W}^{(i)}, \sigma^{2(i)} | \mathbf{W}, \sigma^2 \right)$$

10:      $i = i + 1$

11: **until** a convergence criterion is satisfied

---

6. Statistically Robust Feature Extraction for Stochastic Principal Component Analysis

Until this point, we have assumed that any stochastic noise or observation errors in the demographic data is in some sense "well behaved", for instance: additive, light tailed, symmetric and zero mean. In this section we relax this inherent assumption by developing a class of robust estimators that can withstand violations of such assumptions which routinely arise in real data observations, especially as we will demonstrate in demographic data. Furthermore, we have assumed that the data is generally temporally stationary over the time period of study. If any of these assumptions does not hold then this has an influence from a statistical perspective for the real data analysis. In such cases we recommend to resort to implementation of feature extraction methods which are more robust (in a statistical sense) to violations of such features.

When non-robust feature extraction methods are naively utilised in the presence of violations of these implicit statistical assumptions it can lead to misleading feature extraction and falsify the information content of these features, leading to bias or variance enhancements in the forecast from stochastic mortality models incorporating such features.

Therefore, it is critical to ensure that the feature extraction is appropriately performed. To avoid or to robustify the feature extraction techniques presented previously against violations of such statistical features such as non-stationarity, heavy tails, hetroskedascity, non-Gaussianity one can turn to robust statistical methods. This can strongly influence the findings based on such demographic feature extractions. Therefore, in this section we demonstrate a statistically rigorous approach to perform feature extractions as detailed previously in a robust estimation and feature extraction extended framework.

To achieve this, we first recall some basics of robust statistical inference, primarily targeting the robust estimation of location and scale or mean and covariance, as will be directly relevant in the stochastic PCA based methods proposed above.

Regardless if we work with standard or probabilistic PCA, the most straightforward method to improve its statistical robustness, is to employ estimators of the covariance matrix which are less sensitive to outlying data points.

6.1. Robust Estimators of Mean and Covariance Matrix

To introduce the concept of statistically robust estimation for feature extraction in demographic data, we first introduce what exactly we mean by statistical robustness of feature extraction. This requires a short set of formal definitions.

Let us define an estimator Θ as a functional on the domain of distribution functions. We exchangeable use the definition of an estimator as a function of the d-dimensional sample $\mathbf{y}_1, \ldots, \mathbf{y}_N$, denoted by Θ_N. In the following part we drop the time related index of the random variable and denote it by the $\mathbf{y}$. The empirical distribution defined by sample is denoted by F_N. The true population distribution and density functions of $\mathbf{y}$ are denoted by capital letter F and f.

6.1.1. Concept of Robustness

We consider robustness according to two measures: a measure of local robustness and secondly a measure of global robustness. The two measures are defined by a ϵ-contamination set of a distribution functions, that is

Definition 1. *$\mathcal{F}_\epsilon$ is a contamination neighbourhood of distribution F defined as*

$$\mathcal{F}_\epsilon = \{G : G = (1 - \epsilon)F + \epsilon H, \text{ for } H \text{ any distribution } \} \tag{65}$$

given fraction of contamination $0 \leq \epsilon \leq 1$.

One can then define the local robustness of an estimator Θ as measured by influence function given in Definition 2.

Definition 2 (Influence function). *The influence function of an estimator Θ on the domain of distribution function is defined as*

$$IF(\mathbf{x}_0, \Theta, F) = \lim_{\epsilon \to 0} \frac{\Theta\left((1 - \epsilon)F + \epsilon\delta_{\mathbf{x}_0}\right) - \Theta(F)}{\epsilon} \tag{66}$$

where $\delta_{\mathbf{x}_0}$ is a probability measure which puts mass 1 at the point $\mathbf{x}_0$ if $(1 - \epsilon)F + \epsilon\delta_{\mathbf{x}_0}$ is included into the domain of Θ.

The influence function is a crucial information to calculate asymptotic variance and efficiency of an estimator as

$$\sqrt{n}\left(\Theta(F_N) - \Theta(F)\right) = \sqrt{N} \int_{\mathbb{R}^d} IF\left(\mathbf{y}, \Theta, F\right) dF_N(\mathbf{y}) + \cdots \tag{67}$$

where we used Tylor expansion of the empirical distribution function F_N around true population distribution function F.

The influence function provides us with the knowledge how ϵ contamination on a point $\mathbf{x}_0$ changes the information about the true distribution of the random variable $\mathbf{y}$ which is given by Θ. Thus, it is perceived as a local measure of robustness. To measure global robustness, one can examine a breakdown point ϵ^* of estimator Θ_N at the true population distribution F, given in Definition 3.

Definition 3 (Breakdown point)**.** *The finite-sample breakdown point ϵ^* of an estimator Θ_N at the true population distribution F is defined as*

$$\epsilon^* := \sup\{\epsilon \le 1 : \sup_{G \in \mathcal{F}_\epsilon} |\Theta_N(G)| < \infty\} \tag{68}$$

Intuitively, it is understood as the maximal contamination which does not cause the estimator to loose valid information about the true distribution *F*. We may define a finite sample definitions of the breakdown point as

Definition 4. *The breakdown point ϵ_N^* of estimator Θ at the empirical population distribution F_N is defined as*

$$\epsilon_N^*(\Theta, \mathbf{y}_1, \ldots, \mathbf{y}_N) := \max\{n_1 \in \mathbb{Z}, \max_{i_1, \ldots, i_{n_1}} \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_{n_1}} |\Theta(\tilde{F}_N)| < \infty\} \tag{69}$$

where points $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{n_1}$ were replaced with arbitrary points $\mathbf{z}_1, \ldots, \mathbf{z}_{n_1}$.

Having introduced these formal definitions of what exactly is meant by robust estimators, we overview the most frequently used estimators of a covariance matrix with respect to their robust characteristics according to introduced measures. Let us denote μ, $\mathbf{C}$ as a mean and covariance matrix of $\mathbf{y}$.

6.1.2. M-Estimators

In the study of ([Maronna 1976](#)) and ([Huber and Ronchetti 2009](#)) on the robust estimation of covariance matrix, the authors define one of the first classes of robust estimators, called M-estimators, which are a generalized version of Maximum Log-liklihood Estimators (MLE) where

$$\Theta_N = arg \max_{\Theta \in \Omega} \prod_{n=1}^{N} f(\mathbf{y}_n) \Leftrightarrow \Theta_N = arg \min_{\Theta \in \Omega} \sum_{n=1}^{N} -\log f(\mathbf{y}_n). \tag{70}$$

The idea behind M-estimators, is to replace the density function in Equation ([70](#)), *f*, with function $\rho : \mathbb{R}_+ \to \mathbb{R}$ which down weights outliers, that is

Definition 5 (M-estimators)**.** *The M-estimator of a parameter Θ is defined as a solution to the problem*

$$\Theta_N = \underset{\Theta \in \Omega}{\operatorname{argmin}} \sum_{n=1}^{N} \rho(d_n). \tag{71}$$

for a function $\rho : \mathbb{R}_+ \to \mathbb{R}$ where $d_n^2 = (\mathbf{y}_n - \mu)\,\mathbf{C}^{-1}\,(\mathbf{y}_n - \mu)^T$ is a Mahalanobias distance of $\mathbf{y}_n$.

Remark 3. *If ρ is a continuous function, denoting its derivative $\rho' = \psi$, the estimator $\hat{\Theta}_N$ satisfies*

$$\sum_{n=1}^{N} \frac{\psi(d_n)}{d_n} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_n - \mu) = 0$$

$$\sum_{n=1}^{N} \frac{\psi(d_n)}{d_n} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_n - \mu)^T (\mathbf{y}_n - \mu)\,\mathbf{C}^{-\frac{1}{2}} = \mathbf{0} \tag{72}$$

which are the robust analogue of the typical normal type equations one would solve in MLE estimations in regression for instance.

Remark 4. *If we additionally assume that* **y** *is a random vector from elliptical family with density of the form*

$$f(\mathbf{y}; \boldsymbol{\mu}, \mathbf{C}) = c^{-1} |\mathbf{C}|^{-\frac{1}{2}} g\left(d^2(\mathbf{y}, \Theta)\right). \tag{73}$$

with $g : \mathbb{R}_+ \to \mathbb{R}_+$ *being a density generator of random variable* **y**, *the solution to the problem Equation (71) is equivalent to*

$$\sum_{n=1}^{N} \frac{\psi(d_n)}{d_n} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_n - \boldsymbol{\mu}) = 0$$

$$\sum_{n=1}^{N} \frac{\psi(d_n)}{d_n} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu}) \mathbf{C}^{-\frac{1}{2}} = \mathbb{I}_d \tag{74}$$

More generalized notation is used in (Maronna 1976) and (Huber and Ronchetti 2009) who introduce functions $u_1, u_2 : \mathbb{R}_+ \to \mathbb{R}$ to rewrite Equation (74) as follows in Definition 6.

Definition 6. *The M-estimator* Θ_N *of the parameter* Θ *of a location and scatter of random variable* **y** *from elliptical family, is defined as a solution to*

$$\sum_{n=1}^{N} u_1(d_n)(\mathbf{y}_n - \boldsymbol{\mu}) = 0$$

$$\sum_{n=1}^{N} u_2\left(d_n^2\right) \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu}) \mathbf{C}^{-\frac{1}{2}} = \mathbb{I}_d \tag{75}$$

for functions $u_1, u_2 : \mathbb{R}_+ \to \mathbb{R}$, *given its N realisations.*

The authors provide conditions for u_1 and u_2 to ensure existence and uniqueness of Equation (75) and its normal asymptotic distribution with $-\frac{1}{2}$ convergence. (Maronna 1976) proves the breakdown point of Equation (75) to be very sensitive to dimensionality of the data as $\epsilon^* = \frac{1}{d+1}$.

Recall that in this study, we want to compute a variance of every variable rather than a covariance matrix (Algorithm 1) what makes a M-estimator a very suitable tool. What is more, the robustness of M-estimators does not depend on the sample size. It is another advantage of an M-estimator when working with population data as has a very limited number of observation available.

Remark 5. *As an example of function* u_1 *and* u_2, *(Huber 1964) gives the following*

$$u_1(s) = \frac{\psi_H(s, k)}{s}, \; u_2(s) = \frac{\psi_H(s, k^2)}{s}, \; \rho_H(s) = \begin{cases} \frac{1}{2}s^2 & |s| < k \\ \frac{1}{2}|s| - \frac{1}{2}k^2 & |s| \geq k \end{cases} \tag{76}$$

for $\psi_H(s, k) = \rho'_H(s, k)$ *and the tuning constant k.*

Defining estimator with ρ_H is proven to have the minimal maximal asymptotic variance of all affine invariant estimators under the assumption that the sample $\mathbf{y}_1, \ldots, \mathbf{y}_N$ is normally distributed with zero mean and identity covariance matrix.

For the problem of only scatter estimation, (Tyler 1987) introduced a function $u_2(s) = \frac{d}{s}$, which is investigated in details by (Frahm and Jaekel 2010).

Definition 7. *Tyler's estimator of an covariance matrix for unbounded function* $u_2(s) = \frac{d}{s}$ *is obtained by solving*

$$\frac{d}{N} \sum_{n=1}^{N} \frac{(\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu})}{(\mathbf{y}_n - \boldsymbol{\mu}) \mathbf{C}^{-1} (\mathbf{y}_n - \boldsymbol{\mu})^T} = \mathbf{C}. \tag{77}$$

Lemma 1. *Under the assumption of generalized elliptical distribution of* **y**, *the solution Equation* (77) *exists and is unique up to the scalar parameter.*

Recall that u_2 is not bounded and hence does not satisfy the conditions described in (Maronna 1976) or (Huber and Ronchetti 2009). Therefore the uniqueness can be obtain up to the scaling parameter.

Lemma 2. *if* **y** *belongs to generalized elliptical distribution, Tyler's estimator from Definition 7 is strongly consistent with true covariance matrix, if it exists, up to the scaling factor and has normal asymptotic distribution with convergence rate* $-\frac{1}{2}$.

The asymptotic variance of Tyler's M-estimator, which was analysed i.e., by the author of (Tyler 1987), is proven to have the lowest maximum bias among all normally distributed estimators for symmetric elliptical family random variables.

For a point contamination, the maximal breakdown point of the estimator is equal to $\epsilon^* = \frac{1}{d}$ as proved in (Tyler 1987). The later study shows that for any other contamination, the maximal breakdown point is between

$$\frac{1}{d+1} \le \epsilon^* \le \frac{1}{d} \tag{78}$$

Furthermore, Tyler's M-estimator is an MLE estimator of angular Gaussian distribution. To show this property we use the following Lemma

Lemma 3. *If* **y** *has generalized elliptical distribution then* $\mathbf{y} =_d \boldsymbol{\mu} + R\mathbf{C}^{\frac{1}{2}}\mathbf{u}$ *for* **u** *being uniformly distributed random variable on the unit d dimensional sphere and R being a scalar variable.*

Recall that R is a component which is responsible for generating tails of **y** distribution.

Remark 6. *If R is absolutely continuous and independent from* **u**, *the generator of probability density function of centred* **y**, $\tilde{\mathbf{y}}$, *is given by* $f(\tilde{\mathbf{y}}) := \frac{\Gamma\left(\frac{r}{2}\right)}{2\pi^{\frac{r}{2}}} \tilde{\mathbf{y}}^{-\frac{r-1}{2}} f_R(\sqrt{\tilde{\mathbf{y}}})$ *for* $r = rank(\mathbf{C})$ *and* f_R *being probability density function of R. Then* **y** *is symmetrically distributed.*

Remark 7. *If the variable R is allowed to be negative and has a dependence structure with uniformly distributed* **u**, *the distribution of* **y** *is called generalized elliptical distribution. The generalised elliptical family allows to model asymmetric and tail dependent distribution of* **y**.

Theorem 5 (Distribution-free Tyler estimator). *The estimator introduced by Definition 7 is distribution-free, i.e., it does not depend on values of R.*

Proof. Following Lemma 3 we can notice that

$$\mathbf{z} := \frac{\mathbf{y} - \boldsymbol{\mu}}{||\mathbf{y} - \boldsymbol{\mu}||_F} = \frac{R\mathbf{u}\mathbf{C}^{\frac{1}{2}}}{||R\mathbf{u}\mathbf{C}^{\frac{1}{2}}||_F} \stackrel{d}{=} sign(R)\frac{\mathbf{u}\mathbf{C}^{\frac{1}{2}}}{||\mathbf{u}\mathbf{C}^{\frac{1}{2}}||_F}. \tag{79}$$

and is robust against extreme realisations of R as does not depend on the values of R. $||\cdot||_F$ denotes the Frobenius norm. Rewriting Equation (77) using N realisations of $\mathbf{z}$, $\mathbf{z}_1, \ldots \mathbf{z}_N$, we obtain

$$\mathbf{C} = \frac{d}{N}\sum_{n=1}^{N}\frac{\mathbf{z}_n^T\mathbf{z}_n}{\mathbf{z}_n\mathbf{C}^{-1}\mathbf{z}_n^T} = \frac{d}{N}\sum_{n=1}^{N}\frac{\mathbf{C}^{-\frac{1}{2}}\mathbf{u}_n^T\mathbf{u}_n\mathbf{C}^{-\frac{1}{2}}}{\mathbf{u}_n^T\mathbf{u}_n} \tag{80}$$

□

Tyler's estimator is invariant under any change of distribution of R i.e., it is distribution - free.

Theorem 6. *The estimator introduced by Definition 7 is MLE estimator of Angular Gaussian Distribution.*

Proof. Since the probability density function of $\mathbf{v} = \frac{\mathbf{u}\mathbf{C}^{\frac{1}{2}}}{||\mathbf{u}\mathbf{C}^{\frac{1}{2}}||_F}$ is given by

$$f_v(\mathbf{v}) = \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{\frac{d}{2}}} \frac{|\mathbf{C}^{-1}|^{\frac{1}{2}}}{||\mathbf{v}\mathbf{C}^{-\frac{1}{2}}||_F^d}. \tag{81}$$

Having N realisations of $\mathbf{v}$, $\mathbf{v}_1, \ldots, \mathbf{v}_N$, the logliklihood function is given by

$$\log \mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_N; \mathbf{C}) = \sum_{n=1}^{N} \log f_v(\mathbf{v}_n) = \text{const} + \frac{N}{2} + \log|\mathbf{C}^{-1}| - \frac{d}{2}\sum_{n=1}^{N} \log\left(\mathbf{v}_n \mathbf{C}^{-1} \mathbf{v}_n^T\right)$$

$$= \text{const} + \frac{N}{2} + \log|\mathbf{C}^{-1}| - \frac{d}{2}\sum_{n=1}^{N} \log\left(\mathbf{z}_n \mathbf{C}^{-1} \mathbf{z}_n^T\right) \tag{82}$$

Since $\mathcal{P}^d$ is an open set, the maximizer of $\log\mathcal{L}$ with respect to $\mathbf{C}$ is the stationary point. By setting

$$\frac{\partial \log\mathcal{L}}{\partial \mathbf{C}^{-1}} = \mathbf{0} \Leftrightarrow$$

$$\frac{N}{2}(2\mathbf{C} - diag\mathbf{C}) - \frac{d}{2}\sum_{n=1}^{N} \frac{2\mathbf{s}_n^T \mathbf{s}_n - diag(\tilde{\mathbf{y}}_n^T \tilde{\mathbf{y}}_n)}{\tilde{\mathbf{y}}_n \mathbf{C}^{-1} \tilde{\mathbf{y}}_n^T} = \mathbf{0} \Leftrightarrow \tag{83}$$

$$N\mathbf{C} - d\sum_{n=1}^{N} \frac{(\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu})}{(\mathbf{y}_n - \boldsymbol{\mu}) \mathbf{C}^{-1} (\mathbf{y}_n - \boldsymbol{\mu})^T} = \mathbf{0}$$

what is precisely equal to Equation (77). □

The solution to the estimation problem of Tyler's estimator is in a form $x = f(x)$ what allows us to use the fix - point iteration scheme with $i + 1$ step

$$\mathbf{C}^{(i+1)} = \frac{d}{N}\sum_{n=1}^{N} \frac{(\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu})}{(\mathbf{y}_n - \boldsymbol{\mu}) \mathbf{C}^{(i)(-1)} (\mathbf{y}_n - \boldsymbol{\mu})^T}. \tag{84}$$

Lemma 4. *The fix-point algorithm Equation (84) converges to $a\mathbf{C}$ for a scalar $a > 0$.*

Proof. Since $f(\mathbf{C}) = \frac{d}{N}\sum_{n=1}^{N} \frac{(\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu})}{(\mathbf{y}_n - \boldsymbol{\mu}) \mathbf{C}^{-1} (\mathbf{y}_n - \boldsymbol{\mu})^T}$ being continuous on $\mathcal{P}^d$, the algorithm converges to $\mathbf{C}$ for a scalar $a > 0$. □

Following (Tyler 1987), let us define the following function $M : \mathcal{P}^d \to \mathcal{P}^d$ such that

$$M(\Gamma) = \frac{d}{N}\sum_{n=1}^{N} \frac{\Gamma^{\frac{1}{2}} (\mathbf{y}_n - \boldsymbol{\mu})^T (\mathbf{y}_n - \boldsymbol{\mu}) \Gamma^{\frac{1}{2}}}{(\mathbf{y}_n - \boldsymbol{\mu}) \Gamma (\mathbf{y}_n - \boldsymbol{\mu})^T} \tag{85}$$

The fixed point iteration step Equation (84) can be rewritten as

$$\mathbf{C}^{(i+1)} = \mathbf{C}^{(i)\frac{1}{2}} M\left(\mathbf{C}^{(i)(-1)}\right) \mathbf{C}^{(i)\frac{1}{2}} \tag{86}$$

Denoting $\Gamma^{(i)} = \mathbf{C}^{(i)(-1)}$ and $\mathbf{M}^{(k)} = M\left(\Gamma^{(i)}\right)$ we get

$$\Gamma^{(i+1)} = \Gamma^{(i)\frac{1}{2}} \mathbf{M}^{(i)(-1)} \Gamma^{(i)\frac{1}{2}} \tag{87}$$

In order to deal with the lack of uniqueness of the solution to Equation (84), (Tyler 1987) restricts the search space to the positive definite symmetric matrices with trace equal to 1 by

$$\Gamma^{(i+1)} = \frac{\Gamma^{(i)\frac{1}{2}}\mathbf{M}^{(i)(-1)}\Gamma^{(i)\frac{1}{2}}}{tr\left(\Gamma^{(i)\frac{1}{2}}\mathbf{M}^{(i)(-1)}\Gamma^{(i)\frac{1}{2}}\right)} = \frac{\Gamma^{(i)\frac{1}{2}}\mathbf{M}^{(i)(-1)}\Gamma^{(i)\frac{1}{2}}}{tr\left(\Gamma^{(i)}\mathbf{M}^{(i)(-1)}\right)}. \tag{88}$$

That way, it is ensured that $tr(\Gamma^{(i+1)}) = 1$. In his study, Tyler proves that the convergence of the sequence $\Gamma^{(i+1)}$ to a non singular matrix Γ implies $M(\Gamma) = \mathbb{I}_d$

Theorem 7. *(Tyler 1987) If the following conditions hold:*

1. *the sample $\mathbf{y}_1, \dots, \mathbf{y}_N$ does not contains values equal to $\boldsymbol{\mu}$*
2. *the empirical distribution measure F_N of the sample satisfies $F_N(S) < \frac{1}{d}$, for S being a any proper subspace of $\mathbb{R}^d$*
3. *for some the mth smallest eigenvalue of $\Gamma^{(i)}$, it is holds that $\lambda_{m,d} > \frac{d}{rank(S)} F_N(S)$ where S is any proper subspace of $\mathbb{R}^d$*

then $\Gamma^{(i)} \to \Gamma$ and $M(\Gamma) = \mathbb{I}_d$.

Remark 8. *Generally, to deal with lack of uniqueness of Equation (84), a common practice is to impose additional constraints such as $|\mathbf{C}| = 1$ as in (Frahm and Jaekel 2010) or $tr\mathbf{C} = 1$ as in (Tyler 1987) or (Sun et al. 2016).*

6.1.3. S-Estimators

As the robustness of the class of M-estimators is highly influenced by dimensionality of the data, we introduce an extension to the problem Equation (71) and define class of S-estimator. The class was firstly introduced by (Rousseeuw and Yohai 1984) and extended to multivariate setting by (Davies 1987)

Definition 8 (S-estimator). *S-estimator of a parameter Θ is defined as a solution to*

$$\Theta_N = \min_{\Theta \in \Omega} |\mathbf{C}| \text{ subject to}$$
$$\int \rho\left(d(\mathbf{y}, \Theta)\right) dF_N(\mathbf{y}) \geq b_0 \tag{89}$$

where the constant b_0 is a mean of $\rho(d_y)$, where d_y is a Mahalanobias distance of the random vector $\mathbf{y}$ under an assumption of the distribution of $\mathbf{y}$, that is

$$b_0 = \mathbb{E}\left[\rho\left(d_y\right)\right] = \mathbb{E}\left[\rho\left(||\mathbf{y} - \mu||_F\right)\right] = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \rho(r) f(r) r^{d-1} dr. \tag{90}$$

Function ρ is defined as in Definition 5.

Remark 9. *If ρ is continuous then the problem Equation (89) is an equality.*

In the study of (Davies 1987), the author investigates the properties of S-estimators under the assumption of elliptically distributed $\mathbf{y}$. (Davies 1987) gives the general assumptions on function $\rho : \mathbb{R}_+ \to [0, 1]$, e.g., being continuous on its domain and zero c such that $0 < c < \infty$. The author proves that under these assumptions and if $nb_0 \geq d + 1$, Equation (89) has at least one solution for non singular estimate $\Theta = (\mu, \mathbf{C})$.

The work of (Davies 1987) proves consistency of the estimator Θ_N and its uniqueness under these assumptions. When additionally ρ has a continuous third derivative, the solution to Equation (89), is asymptotically normal with $-\frac{1}{2}$ convergence.

Let us denote $\epsilon = 1 - b_0$. Following (Davies 1987), if $N(1 - 2\epsilon) \geq N + 1$, and the sample is in general position, that is no more that d points of the sample lies on $(d - 1)$-dimensional hyperplane, then the finite-sample breakdown point is equal to

$$\epsilon_N^* = \frac{[N\epsilon] + 1}{N} \Rightarrow \lim_{n \to \infty} \epsilon_n^* = \epsilon = \epsilon^*. \tag{91}$$

In the paper (Lopuhaa 1989), there is introduced an alternative definition of S-estimators, considering function $\rho : \mathbb{R}_+ \to [0, \infty)$ and investigates the relation between S-estimators and M-estimators. He imposes stronger assumptions on the function ρ than are required to make his definition of S-estimators equivalent to Davis's definition (Davies 1987) by transforming $\rho \to 1 - \frac{\rho}{\sup \rho}$. Under the assumption of elliptical distribution of **y**, the author shows that any solution to the problem Equation (89) satisfies Equation (71). He rewrites the problem Equation (89) as the similar to Equation (74) using Lagrangian

$$\log \mathcal{L}_N(\Theta, \lambda) = \log |\mathbf{C}| - \lambda \left[\frac{1}{N} \sum_{n=1}^{N} \rho(d_n) - b_0 \right] \tag{92}$$

to obtain the equivalent set of equations

$$(\Theta_N, \lambda) = \underset{\Theta \in \Omega, \lambda \in \mathbb{R}}{\operatorname{argmin}} \log \mathcal{L}_N(\Theta, \lambda) \Rightarrow \begin{cases} \frac{1}{N} \sum_{n=1}^{N} u(d_n)(\mathbf{y}_n - \boldsymbol{\mu}) = 0 \\ \frac{d}{N} \sum_{n=1}^{N} \frac{u(d_n)}{v(d_n)}(\mathbf{y}_n - \boldsymbol{\mu})^T(\mathbf{y}_n - \boldsymbol{\mu}) = \mathbf{C}, \end{cases} \tag{93}$$

for $u(s) = \frac{\psi(s)}{s}$ and $v(s) = \psi(s)s - \rho(s) + b_0$. The term $-\rho(s) + b_0$ substitutes Lagrangian multiplier λ. Hence, every solution to Equation (89) is a solution to Equation (93) what is M-estimator problem defined by Equation (75).

The author of (Lopuhaa 1989) argues that S-estimators achieve the same asymptotic variance as corresponding M-estimators but as dimensionality of data increases, they have higher breakdown point than M-estimators.

Remark 10. *The example of functions which satisfies the conditions of uniqueness and existence of Equation* (89) *are Tuckey's biweight functions, that is*

$$\rho_B(s) = \begin{cases} \frac{s^2}{2} - \frac{s^4}{2k^2} + \frac{s^6}{6k^4} & |s| \leq k \\ \frac{k^2}{6} & |s| \geq k \end{cases} \Rightarrow \psi_B(s) = s \left(1 - \left(\frac{s}{k}\right)^2 \right)^2 \mathbf{1}_{[-k,k]}(s) \tag{94}$$

The breakdown point of the S-estimator for $\rho = 1 - \frac{\rho_B}{\sup \rho_B}$ is equal to

$$\epsilon^* = \frac{[N(1 - b_0)] + 1}{N} = \frac{\left[\frac{N}{\sup \rho_B} \mathbb{E}\{\rho_B(\mathbf{y} - \boldsymbol{\mu})\} \right] + 1}{N} \tag{95}$$

7. Data

The examined data consists of male and female mortality and demographic data obtained from Human Mortality Database (http://www.mortality.org) for European countries. The Table 1 summarizes the availability of the data for all the countries included into the study.

We use four different sets of mortality data, raw data: Birth counts and Death counts, and life tables: Life Expectancy at Birth and Death Rates. We conduct separate analysis for female and male populations.

The time series vary in terms of the number of available observations. 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.

With regards to Birth counts and Life expectancy at Birth, the information per country in time point is one dimensional, i.e., annual counts of live births by sex in year *t* (Birth counts) or the expected life span of a person born in year *t* (Life Expectancy at Birth). Hence, a single observation in these cases consists of the number of entries equal to the number of countries included into the study, that is 31 listed in Table 1, per gender.

Table 1. The availability of the demographic data per country (Human Mortality Database).

Country	Life Expectancy ($E0$)	No. Births	Death Rate (m_x)	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
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

The 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). The availability of time series is different among age groups. Usually, the shortest time series are collected for age groups above 100 years.

Since the Lee-Carter model with the cohort effect has been already reported to be prone to over-fitting when fitted to short time series (the currently available mortality data is classified as a short time series), we decide to work with a data aggregated in the format "5 × 1", i.e., by 5-year age group per calendar year. The ages are grouped into following stratifications: 0, 1–4, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 5–9, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, 85–89, 90–94, 95–99, 100–104, 105–110, 110+ as additing additional latent processes increases the number of parameters to estimate.

A single observation in Deaths or Death Rates consist of the number of entries equal to the number of countries included into the study (i.e., 31) times the number of age groups, that is 24. It accounts for the information in time *t* available for all 31 countries among all 24 age groups.

7.1. Preprocessing of Data

Human Mortality Database (HMD) team applies several preprocessing procedures that aim to "clean" Death counts and population sizes before using them in order to calculate and distribute death rates and life tables. The subsequent steps are discussed in the technical report (Wilmoth et al. 2007). The adjustments are applied in order to distribute people of unknown age across age groups and splitting data into age categories, i.e., from age stratification "5 × 1" to "1 × 1" and from "1 × 1" to Lexis triangles. The common practice is to use a regression model for splitting deaths counts in format "1 × 1" to Lexis triangles and apply cubic splines to split "5 × 1" to "1 × 1". Additionally to the adjustments applied to Deaths counts, the age specific population size is estimated using four methods: linear interpolation, intercensal survival, extinct cohorts and survivor ratios.

The life tables are calculated using Lexis triangles and population sizes. Before death rates are converted into the probabilities of death, the rates at older ages (80 and above) are smoothed using logistic regression. The "abridged" life tables are calculated based on the Lexis triangles tables rather than the raw data. It ensures the both sets of tables to contain identical values of life expectancy and other quantities.

Recall that the smoothing applied to the mortality data can influence the feature extraction and the diminish the effect of robust versus non-robust versions of feature extraction methodology which we study in this paper. The topic is further discussed in Section 8.

Additionally to the briefly discussed procedures which have been already applied to the data in HMD, we needed to adjust the data to provide reliable information about missing values. We notice the ambiguity in labelling unavailable data which is either denoted by "NA" value or "0". Death Rates, Birth counts and Life Expectancy at Birth are unlikely to produce values equal to "0". Hence we replace all "0" which appeared in these data sets by "NA". The zeros which appear in Deaths counts in older age groups are more difficult to handle as there is no certainty whether there was no person in particular age group who died or the record has missing values. Due to this fact we decided to limit our analysis to age groups up to 90 and again replace any "0" by "NA".

7.2. Missing Data

The following subsection is a summary of different types of missingness across countries, age groups and sexes which occurs in the demographic data set analysed in our study. The findings of the subsection are the following: among the Birth related data where a total observation is of vector-type (one dimensional information per country), the incompleteness of data is due to a general unavailability of the information per country. However, for Death related data where a total observation is of matrix-type (a age specific vector of information per country), we notice a patter of single missing values in time point which fits definition of MAR. In the subsequent section with the empirical analysis of the data set under the derived framework of Probabilistic Principal Component Analysis, we assumed MAR type of missingness. Extending this assumption requires calculating the integral from Equation (34) and incorporating it to PPCA framework.

7.2.1. Observed Patterns of Missingness

Missing data appear when no value is available for a component of the observation vector. The Probabilistic Principal Component Analysis with missing data discussed in Section 5.3 handles missing data by filling them with projection using principal components that are calculated from available information. Thus, the more missing data is present for a given variable, the less impact this variable has on the specification of the projection over the assumptions. The results for variables with missing data are more influenced by our assumption of the distributions from Section 5, as will be discussed in detail below.

In order to handle missing values we need to understand and study how their pattern. In the following part we demonstrate the three patterns of missingness which are present in the analysed data sets:

Type 1: no information about a variable in a few observations for a given country;
Type 2: no information about a variable in all observations for a given country;
Type 3: general unavailability of information about all variables for a country except for a limited set of observations.

The analysis is conducted separately for four data sets and sexes. We show the results for four cases where we segment the data among four proportions of missing entries per total observation with maximal percentage of missing entries equal to: 0%, 25%, 50% and 75%. For instance, a single observation of Birth counts is 31 dimensional. When we analyse the data set with respect to the case 50%, we exclude all observation, when the number of missing entries per observation is greater than $0.5 \times 31 \approx 15$.

In the case of no missing data, that is 0% for Death counts and Death Rates in Females, the number of rows without missing entries is too small for any significant analysis. Due to this fact we drop the minimal number of columns which have the highest number of missing entries in order to collect significant sample for our analysis.

The results of the analysis are similar among Birth counts and Life Expectancy at Birth and Death counts and Death Rates. Hence we discuss the patterns of missingness in Number of Births and Number of Deaths.

7.2.2. Births

The left plot of Figure 1 shows the percentage of missing entries per observation vector of total births over all countries considered versus calendar years for Births counts disaggregated for Females and Males. The sample starts in 1751 (Swedish data) and spans to 2014. Until 1946, the sample has a percentage of missing entries above 50% (the middle vertical red line on the plot). The sample of case 25% starts in 1950. We observe the same missing values pattern between female and male population (the corresponding lines for populations overlap and only the black line, corresponding to male population, is visible).

Figure 2 indicates the availability of data per country (y axis) versus the calendar year (x axis). The black colour denotes points in time when the data is not available for a given country. The Swedish data is not labelled by any black entry except in 2014 as it is the longest time series. Recall that the missing data pattern which is characteristics for this data set is a limited availability of data for a given country. From the empirical analysis of the data we learn that the missing entries do not appear randomly. However, it does not violate the assumed behaviour of missingness and we can still proceed with the methodology of feature extraction described in the previous parts of the paper.

The red vertical lines correspond to the starting points of the subsamples when the maximum number of missing entries per row (an observation in time) is equal to (from the left side on corresponding plots) 75%, 50% and 25%. The subsamples for the cases 25% and 0% provide with principal components which are determined by almost equal sizes of information from every country included into the analysis for the examined period of time. However, the calculation of the components

for the subsample in the case 75% is dominated by the time series of countries which are not available before 1950. Therefore, the calculated features are more prone to be impacted by the distribution assumptions and convey less information about dynamics present in the examined period of time that the components obtained from longer time series.
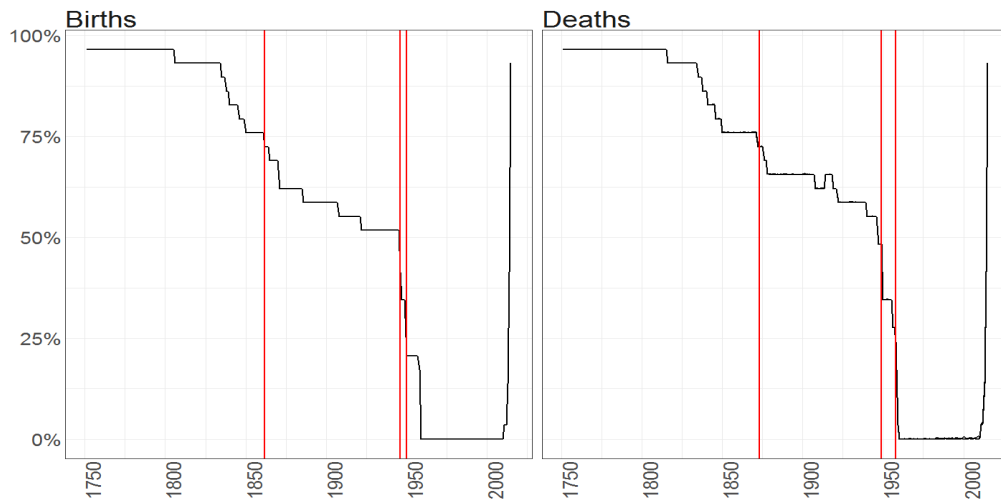

Figure 1. The percentage of missing entries (y axis) per observation vector over time (x axis) for the Births counts (left plot) and Deaths counts (right plot) for female (blue line) and male (blacke line) population. Red vertical lines correspond to the starting points in time of samples when maximum of missing entries is equal to (from the left side on corresponding plots) 75%, 50% and 25%.
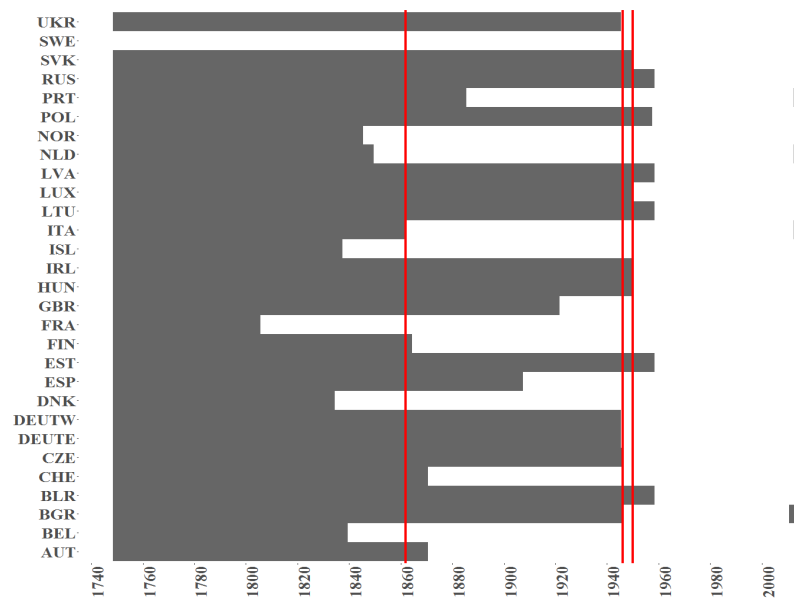
Figure 2. The indicator of a missing value (black colour) per country (y axis) over time (x axis) for the Births counts for female population. Red vertical lines correspond to the starting points in time when samples with maximum of missing entries is equal to (from the left side on corresponding plots) 75%, 50% and 25%.

7.2.3. Deaths

The single observation of Death counts is a 504 dimensional vector which reflect the numbers of death per country and age group in time. The proportion of non missing entries per calendar year, again disaggregated by gender for Death counts is shown on the right plot in Figure 1. The percentage of missing entries decreases slower than for Births counts (the slope of the curve is flattener) which indicates the longer distance between the cases 50% and 25%. We notice the discrepancies in the patterns of missing values between female and male population. Interestingly, the female population data has no observations without missing values.

It is also informative to demonstrate the pattern of missing values per age group, as displayed in Figures 3 and 4 below. Red vertical lines correspond to the starting points in calendar time for the total proportions of missing data corresponding to the cases 50% and 25%. We observe a new pattern of missing values: particular variables have a few missing observations within the subsample. The pattern occurs mainly after 1950 in age groups between 1 and 25 (darker shade of blue for single observations) or in age group 95–100. The other interesting analysis can be done when we display the patterns of missigness which are present by age and per country, which is shown in Figures 5 and 6, again disaggregated by gender. The pattern is present only for the subset of countries. For instance, data related to the Irish population has higher percentage of missing values among only four age groups. Still, the dominant pattern of missingness is the availability of the information for a country which is limited to some period of time.
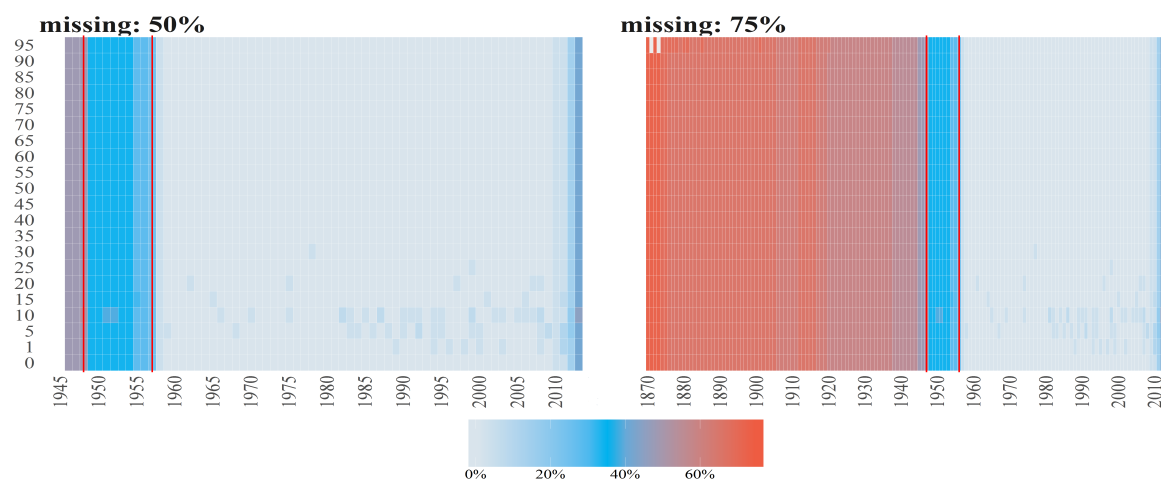
Figure 3. Percentages of missing values (denoted by diffrent colours) per observation for Number of Deaths for Females per age groups (y axis) over time (x axis). The titles of the subplots indicate the case of missing values (50%, 75%). The percentage for a given country and given age group is computed dividing number of missing values by number of countries. Red vertical lines correspond to the starting points in time when the cases 50% and 25% start (from the left to right side on corresponding plots).
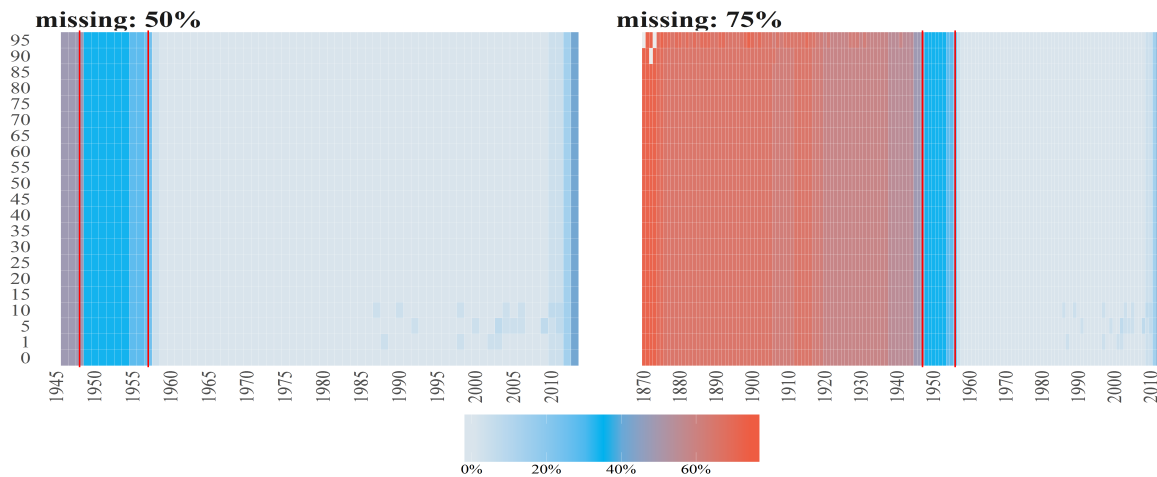
Figure 4. Percentages of missing values (denoted by diffrent colours) per observation for Number of Deaths for Males per age groups (y axis) over time (x axis). The titles of the subplots indicate the case of missing values (50%, 75%). The percentage for a given country and given age group is computed dividing number of missing values by number of countries. Red vertical lines correspond to the starting points in time when the cases 50% and 25% starts (from the left to right side on corresponding plots).
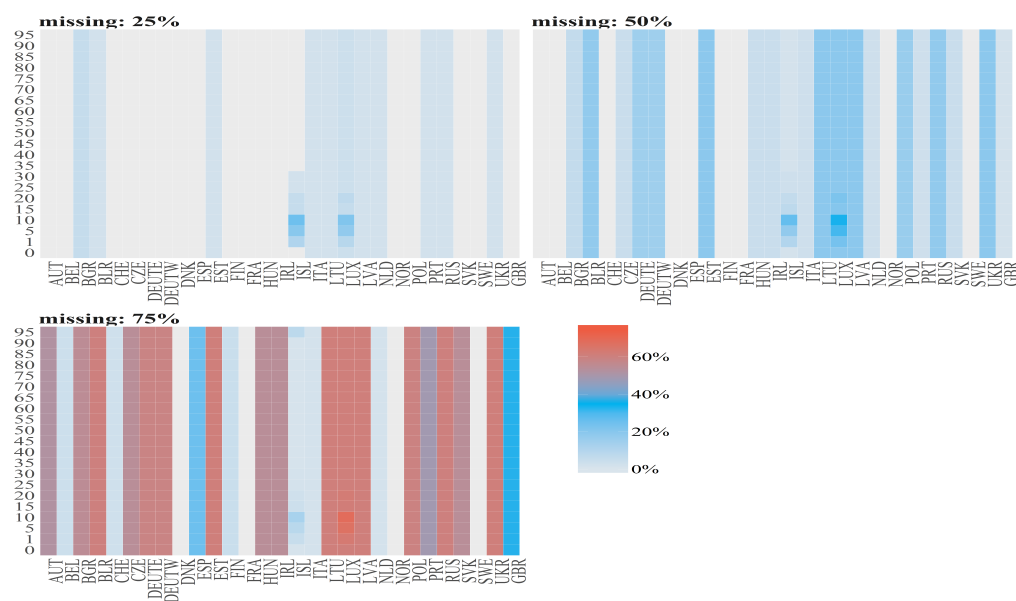
Figure 5. Percentages of missing values (denoted by diffrent colours) for Number of Deaths for Females per country (x axis) and age group (y axis). The titles of the subplots indicate the case of missing values (25%, 50%, 75%). The percentage of missing values for a given country and an age group is calculated dividing number of missing values by length of subsample which is different for different cases.
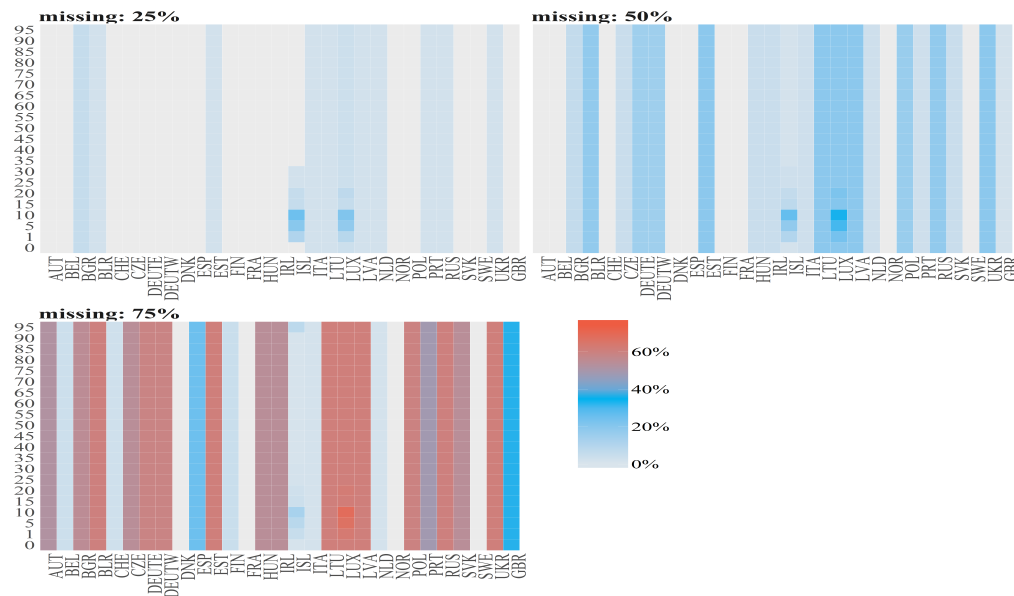
Figure 6. Percentages of missing values (denoted by diffrent colours) for Number of Deaths for Males per country (x axis) and age group (y axis). The titles of the subplots indicate the case of missing values (25%, 50%, 75%). The percentage of missing values for a given country and an age group is calculated dividing number of missing values by length of subsample which is different for different cases.

8. Feature Extraction from European Demographic Data via Probabilistic Principal Component Analysis

The following section provides results for the feature extraction using the methodology introduced in Section 5 and applied to different types of European demographic data sets. The attention is drawn to the effect of simple and straightforward robustification overviewed in the previous sections which application we demonstrate on overviewed data sets. The main observation from this study is the difference in the consistency of the features over time and proportion of missigness for two frameworks, robust and non-robust. The features calculated using the means of robust estimators are more consistent over time and over different proportions of missigness that their non robust alternatives. It is especially visible for the features extracted from data that has not been previously preprocessed, e.g., Birth counts. The effect of robustification is smaller if a data set has been smoothed as Life Expectancy at Birth.

8.1. The Assessment of the Methodology

We conduct a comparison between robust and non-robust Probabilistic Principal Component Analysis (PPCA) in a stochastic setting. We undertake this exercise in order to incorporate the most meaningful eigenvectors as exogenous factors to the model in Equation (8). Each of the datasets discussed in the section is treated separately, that is, we compute the eigenvectors for Births, Life Expectancy at Birth, Deaths and Death Rates. Recall, an observation in time t, $\mathbf{y}_t$, conveys the information about a given data set in calendar year t from the 31 countries listed in Table 1. Hence, an observation of Births or Life Expectancy is 31 dimensional. Since Deaths and Death Rates carry the information which is age group specific, a single observation in these data sets is equal to the number of countries times the number of age groups.

The following section summarizes the results of PCA according to

- Population: Females and Males;
- Subsamples referring to maximum allowed proportion of missing values per an observation: 0% (no missing), 25%, 50% and 75% as discussed in Section 7.2;

- Type of the standardisation procedure: robust and non-robust, which are used for the estimators of location and covariance in the PPCA framework.

We use M-estimators of the covariance and mean as a robust alternative to the sample estimators. As discussed in Section 6.1.2, the class of M-estimator has a very good performance in small dimensions as its robustness is a function of the dimensionality and not the sample size. Since the data we use is not a long time series and we standardize every variable marginally (Algorithm 1), this simple estimator should provide us with reliable outcomes. In particular, we consider the Huber type M-estimators as discussed in Remark 5 which are characterized by normal asymptotic distribution with convergence rate of $\frac{1}{2}$ and several characteristics ensuring both uniqueness and optimality of the estimation Additionally, it is the estimator of covariance which has the minimal asymptotic variance among all estimators for Gaussian data. Thus, the choice is consistent with our assumption of normally distributed data for the treatment of missingness in the PPCA framework which we outlined in Section 5.

Among the objects which we analyse are eigenvectors, eigenvalues, scores and Mahalanobias distances which use estimated covariance matrix $\hat{\mathbf{C}} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ calculated by iteratively evaluated σ^2 and the projection matrix $\mathbf{W}$. The Mahalanobias distance is measured around vector $\mathbf{0}$ (the data has been centred), and is therefore given by

$$d_n^2 := d^2\left(\mathbf{y}_n, \hat{\mathbf{C}}\right) = \mathbf{y}_n\hat{\mathbf{C}}^{-1}\mathbf{y}_n^T. \tag{96}$$

The results show how distant from the assumed long term mean is a single observation.

The EM algorithm described in Section 5 provides the eigenvectors corresponding to the *k* largest eigenvalues. Due to the specifics of the data (small number of observations), it would be difficult to incorporate many factors into the model from Equation (8) and achieve reliable estimation results. This is primarily a result of the curse of dimensionality in the parameter space that would lead in this case to diffusivity in the resulting likelihood utilised in the estimation. The latent states and static parameters will become difficult to filter and estimate. Thus, we limit our analysis to $k = 3$ main eigenvectors which explain most of the variability as shows the standard PCA with non missing data conducted separately for each country showed.

8.2. Births

The results of PPCA for Births counts among Females and Males are similar. It is the outcome which follows general intuition as there is no external factor which influences differently the births of woman and man in European countries. Also, recall that the Birth counts are the least pre-processed data set in out analysis.

Figure 7 shows the Mahalanobias distances (x axis) of Number of Births over the time (y axis) for female (a) and male (b) population. Each sub-panel consists of two plots, which present results for data being standardized by robust (lower plots) and non-robust (upper plots) estimators of the mean and covariance matrix. Different colours of lines depict distances for subsamples where maximum allowed proportion of missigness per an observation is 0%, 25%, 50% and 75%. Recall, that the subsamples starts in different times and therefore the corresponding results are the outcome of the estimation on different data with different impacts of distribution assumptions.
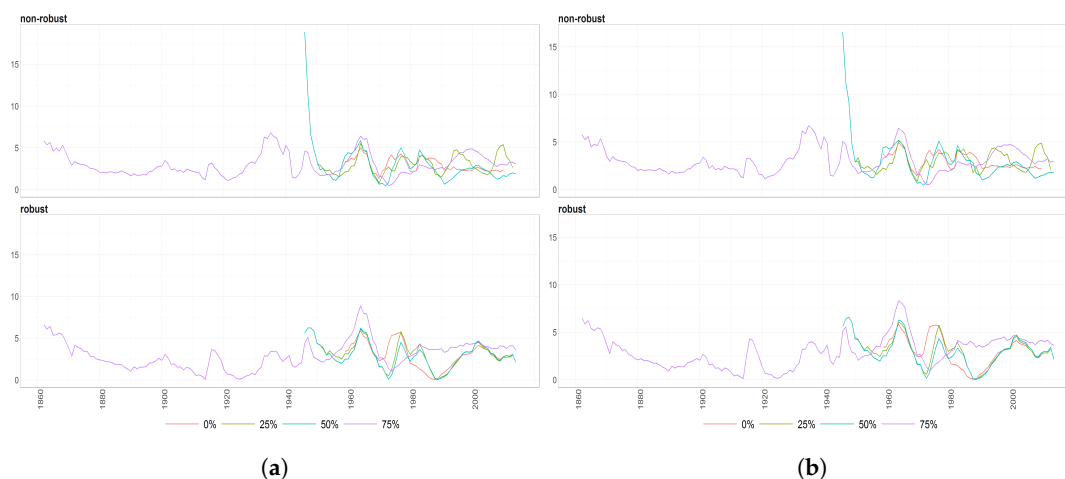
(**a**)	(**b**)

Figure 7. The Mahalanobias distances obtained using Probabilistic Principal Component Analysis (PPCA) for Females (**a**) and Males (**b**) Births over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing values in a signle observation (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

As expected, the effect of robust standardisation is substantial since the framework produces features by down-weights outliers in data. The distances which correspond to the robust standardisation, are more aligned historically, that is, are more 'robust' when the missingness increases. It indicates that robust standardisation of data captures more efficiently the characteristics of the population distribution. Recall the earliest non-robust Mahalanobias distance of 50% case which is very distant from the statistic in the same calendar year but for the 75% missingness case. For the robust case, the corresponding distances are more aligned what demonstrates the effect of robustness.

Since the subsample of the 75% missingness case is substantially longer, we expect the PPCA results to be different as the sample captures more regimes present in demographic data. Also, the sample has higher number of missing values which are estimated using the projection based on assumption of normal distribution. It also impacts the outcomes of PPCA.

The eigenvalues of estimated covariance matrices are shown in Figure 8. Different colours of lines highlight eigenvalues which correspond to different cases of missingness. Upper panels show the results for non-robust framework whereas bottom plots for the robust one. The magnitude of eigenvalues as well as the spreads between them over different levels of accepted missigness are higher for robust case.

The 75% case of missigness results in smaller discrepancies between eigenvalues for robust and non-robust frameworks. The corresponding eigenvectors exhibit similar behaviour what is shown in Figure 9. The discrepancies between the robust and non-robust eigenvectors are more significant for the cases with smaller proportions of missing values per observation. This outcome can be justified by the fact that the case of 75% is more affected by the priori assumptions of the normal distributions. The discrepancies between two methods of standardisation got smaller as the projection of missing values starts to dominate the estimation of the principal components. The robust and non-robust estimators similarly capture the information about the normal distribution.

The blue dotted vertical lines on the plots of eigenvectors disaggregate the outcomes into developed and developing countries listed in Table 1. Regardless of the case of missingness (except 75% case) and type of the standardisation, we notice resembling features for countries from each of the groups.

We would expect the alignment of the robust scores as it has been observed for the corresponding Mahalanobias distances. However, the described PPCA methodology does not re-estimate the mean. The data is centred one during the initialisation. In the presence of missing values, their projection changes the mean and the data is centred only at the start of the procedure. This, the mean is only static when there is no missing data. This simplification of the framework results in different levels of the scores in Figure 10.
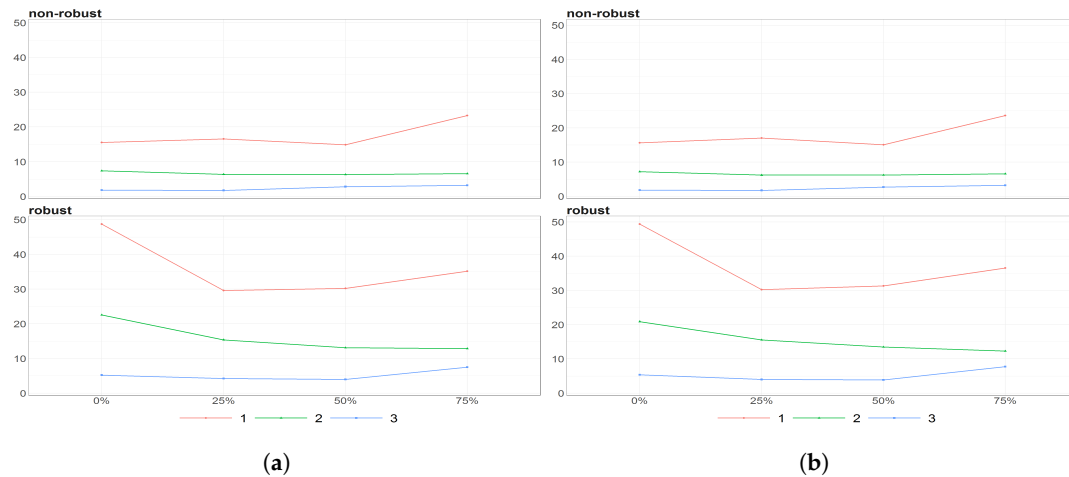

(a)   **(b)**

Figure 8. The eigenvalues obtained using PPCA for Females (**a**) and Males (**b**) population of Number of Birthsfor different percentages of maximal missing entries in rows (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).
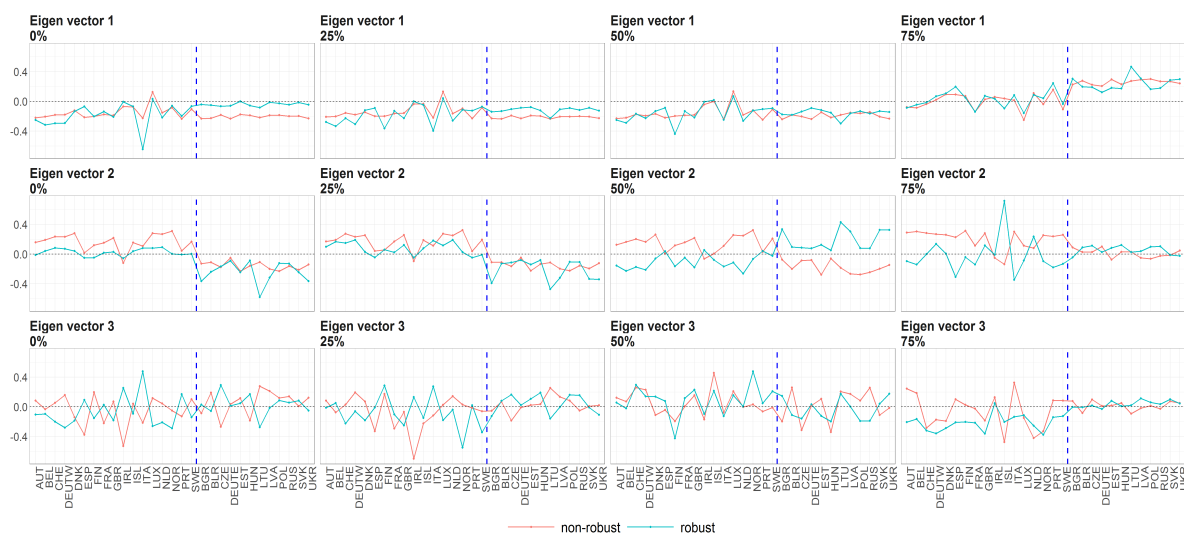
Figure 9. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for Females in Births. Every row of subfigure corresponds to a different eigenvector. Every column corresponds to different case of missing values (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.
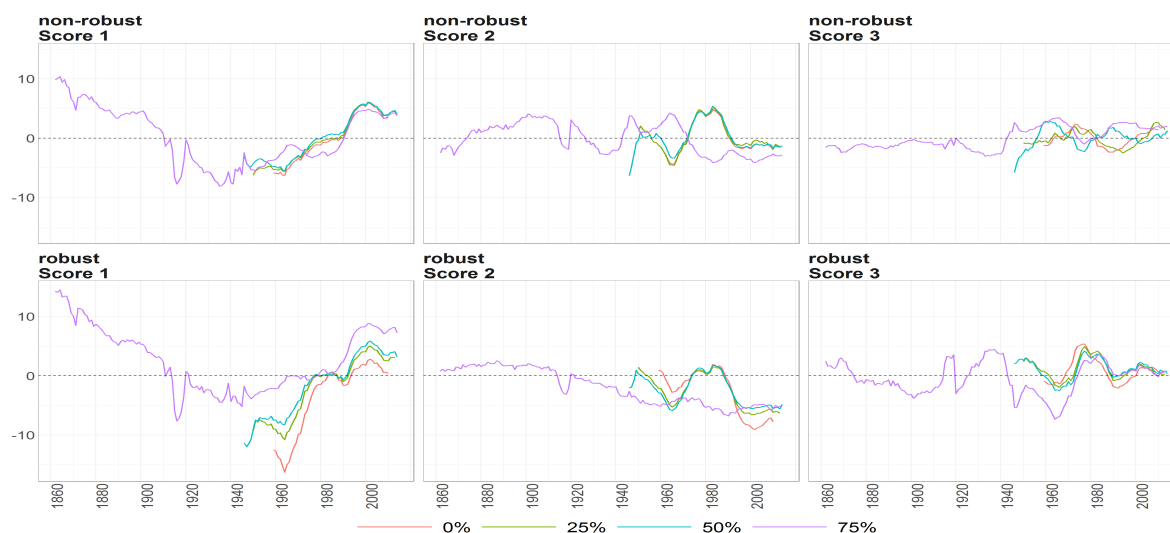
Figure 10. The scores (y axis) over time (x axis) obtained using PPCA for Females in Births. Colours of lines correspond to the scores calculated on subsample different cases of missing values (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots scores, first, second and thirds are ordered by columns.

8.3. Life Expectancy at Birth

As mentioned in Section 7.1, Birth counts are the only data set in our analysis, which has been not modified or preprocessed before being available in HMD. The stages of several adjustments which are applied to the Death counts and population sizes result in a smaller number of outliers in Life Expectancy at Birth. The outcomes of PPCA for the data which is standardized using robust and non-robust estimators of mean and covariance matrix do not vary so significantly as in Birth counts over different proportions of missigness.

The robust Mahalanobias distances in Figure 11 are slightly more distant than their non-robust equivalents. However, the distances for different cases of missing values are similarly aligned in both cases in contrast to Birth counts. We may observe the same pattern among the scores of three principal components in Figures 15 and 16.

Figures 12, 13 and 14 show eigenvalues and eigenvectors for Females and Males respectively. Only the results for the 75% case of missingness exhibit more variation among the standardisation procedures. Since the subsample corresponding to 75% case is significantly longer, the discrepancies can be again rationalized by the two reasons: effect of the priori assumption on distributions in Section 5 and higher number of captured regimes in the data. Moreover, anakysing the outcomes from the aggregation among the European countries, we notice that the second eigenvector has opposite signs for the countries of the two countries groups. Recall that its values are more volatile for the male population of developing countries (except for the case of 75% of missing values).

The obtained scores for Life Expectancy at Birth are shown in Figures 15 and 16. The levels of scores are very close to zero. It is an expected outcome as we differenced the data since it exhibits polynomial trend. The distribution of differences is expected to have zero mean. Hence, the scores are aligned for all the cases of missingness and distributed around zero what is negligibly affected by the simplification of the discussed framework. However, the two methodologies of the standardisation result in different magnitudes of scores. Since the corresponding eigenvectors are very similar for two standardisations, the magnitude is a reliable indicator whether an observation is outlying.
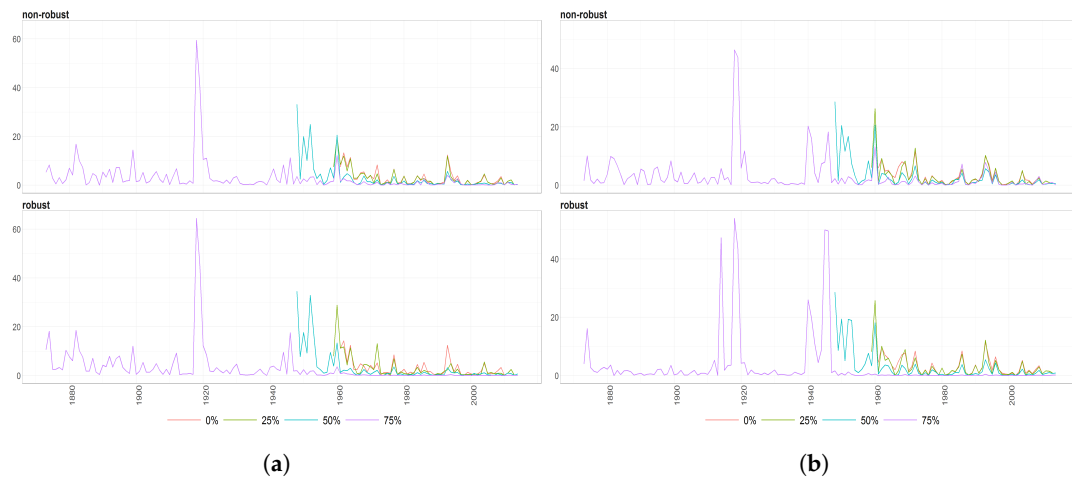
Figure 11. The Mahalanobias distances obtained using PPCA for Females (**a**) and Males (**b**) of Life Expectancy at Birth over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing entries in rows (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to the robust estimation of standard divinations (upper plot) and sample one (bottom plot).
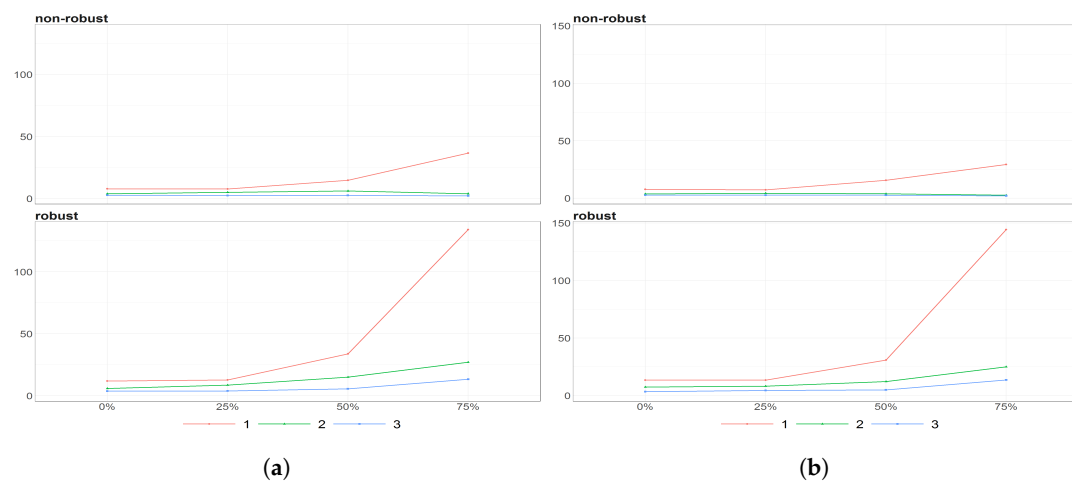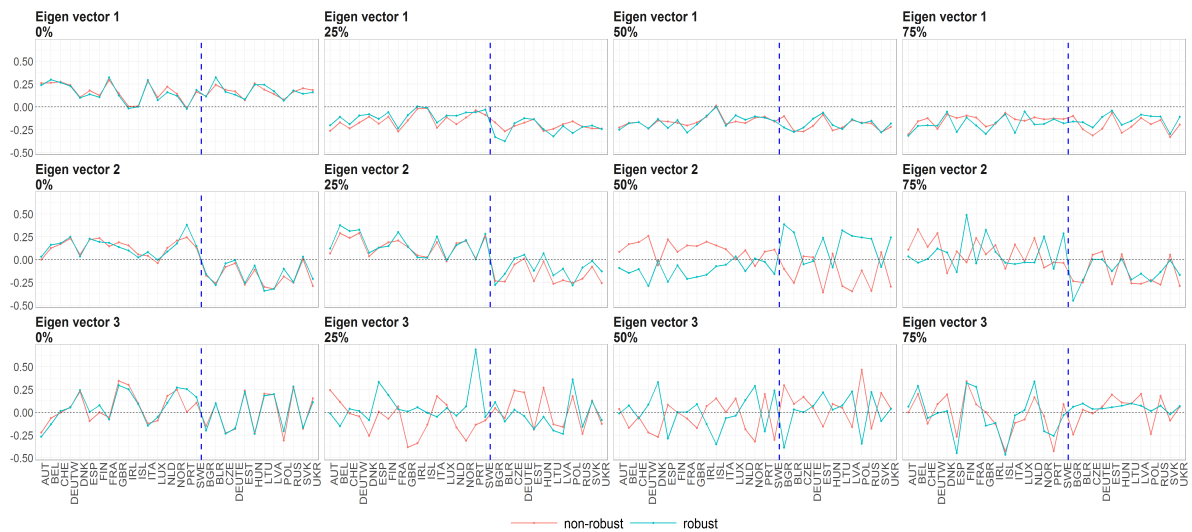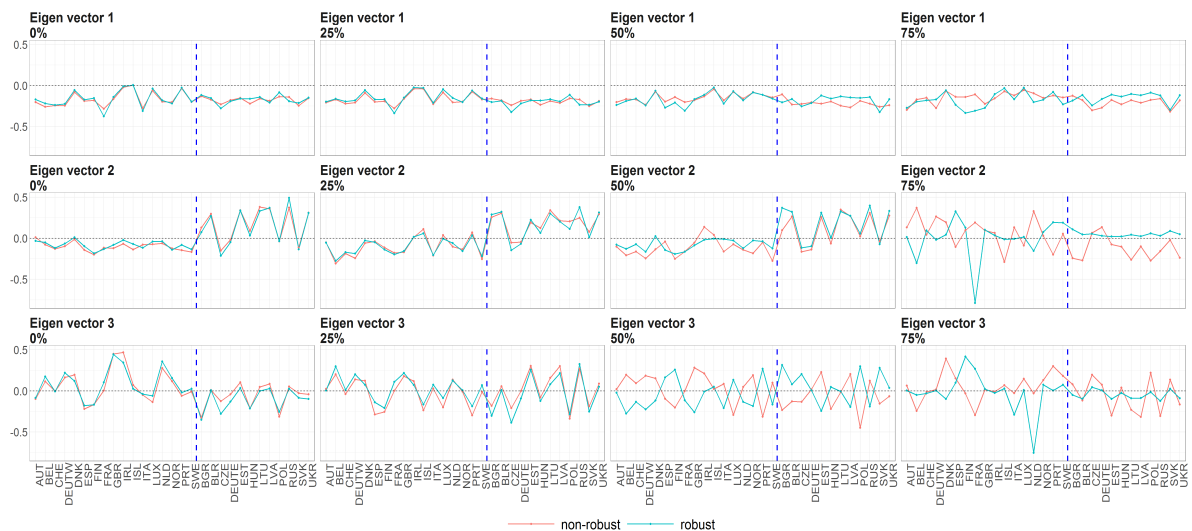

Figure 12. The eigenvalues obtained using PPCA for Females (**a**) and Males (**b**) of Life Expectancy at Birth data for different percentages of maximal missing entries in rows (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

Figure 13. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for female population of Life Expectancy at Birth. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.


Figure 14. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for male population of Life Expectancy at Birth. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.
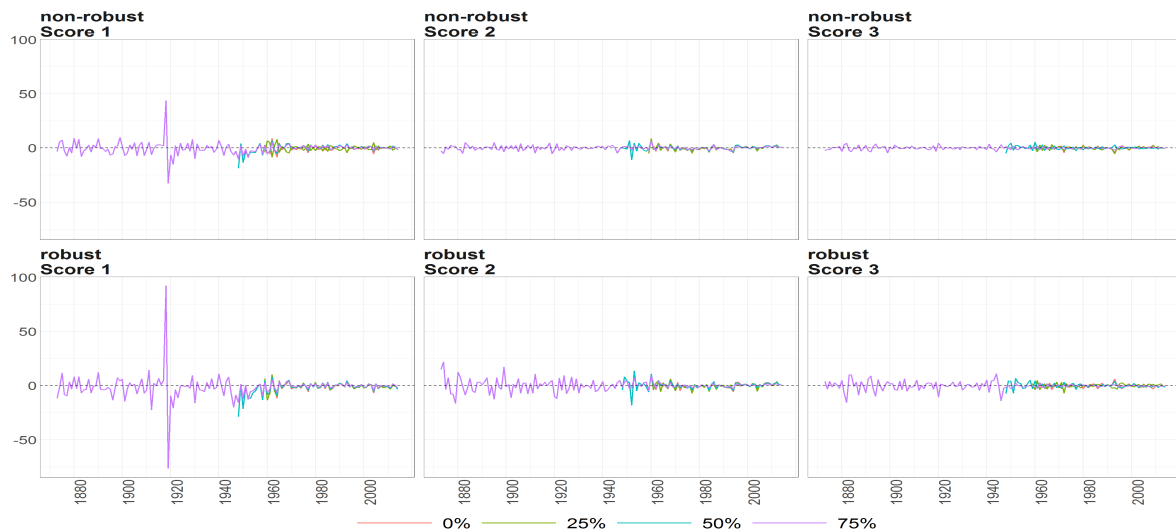
Figure 15. The scores (y axis) over time (x axis) obtained using PPCA for female population of Life Expectancy at Birth. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.
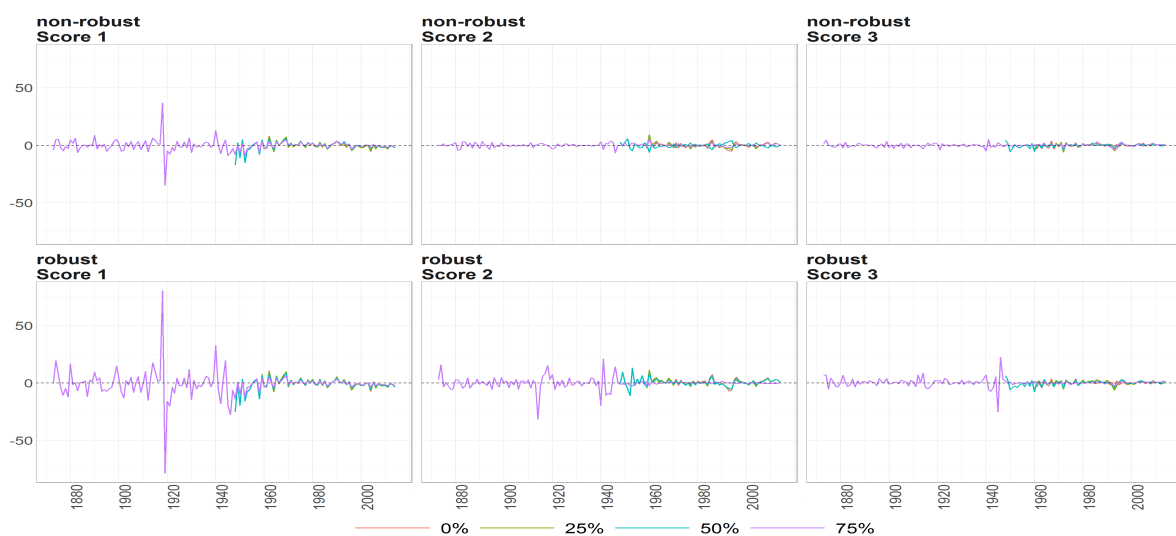
Figure 16. The scores (y axis) over the time (x axis) obtained using PPCA for Males of Life Expectancy at Birth. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

8.4. Deaths

The Mahalanobias distances for the Deaths counts exhibit resembling behaviour for Females regardless of the standardisation procedure. The discrepancies between the statistics are more substantial for the male population what is highlighted in Figure 17.

The eigenvalues of the non-robust estimator of the covariance matrix does not vary between the cases of missigness up to 50%. The plots are shown in Figure 18. For the case 75% the first eigenvalue starts to dominate more significantly than for other missigness cases, especially for Females. On the other hand, the robust eigenvalues are more volatile. Especially the results for Females provide unexpected outcomes for the case 50% in comparison to the case 25% even though the subsamples for these cases differ only by a few calendar years in mid 1940. With regards to the Males, we observe that the dominance of first eigenvalue increases with the number of missing values.

The corresponding eigenvectors are shown in Figures 19 and 20 for Females and Males respectively. The colours of the heatmaps correspond to the magnitude of components of eigenvector which are country (x axis) and age group (y axis) specific. The non-robust estimation results in the eigenvectors with smaller magnitude and smoother within the age groups and countries. Recall, that the distribution of colours for the robust case has bigger spreads between values (so called "bumps") what is highlighted by more intense colours of blue and red. The exception is made for the case of 75%.

The vertical black dotted lines on the heatmaps divides the countries listed on x axis as developed (left side) and developing (right side). Again, this order of results stresses the differences between the eigenvectors within these two groups of countries. The first eigenvector for developing countries has the break point around age group of 40 for all cases of missingness for the male population and all cases except 50% for the female population. The eigenvector for developed countries has a break point for age group in 80 for female population and 75 for male population with additional break in 35 for 0% and 25% cases.

The case of 75 is analysed separately. The first eigenvectors do not differ within two types of standardisation but exhibit the structure which is country group specific. The developed countries are characterized by the uniformed values around zero for all age groups. The vectors of developing countries are more volatile with breaks around 20–30 for Males and even more volatile for Females. The second eigenvector differs within both types of standardisation and among two groups of countries. It is almost constant around zero for Males in developing countries and more volatile for Males in developed countries with breaks in age groups between 50 and 60. Second eigenvector for Females resembles third eigenvector of Males .
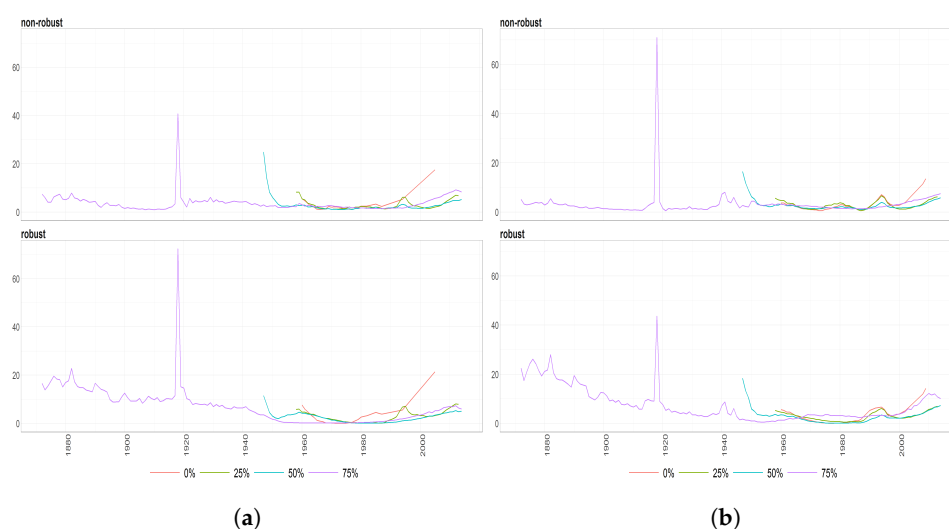

(a)　　　　　　　　　　　　　　　　　　　　　　(b)

Figure 17. The Mahalanobias distances obtained using PPCA for Females (**a**) and Males (**b**) of Number of Death over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing entries in rows (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).
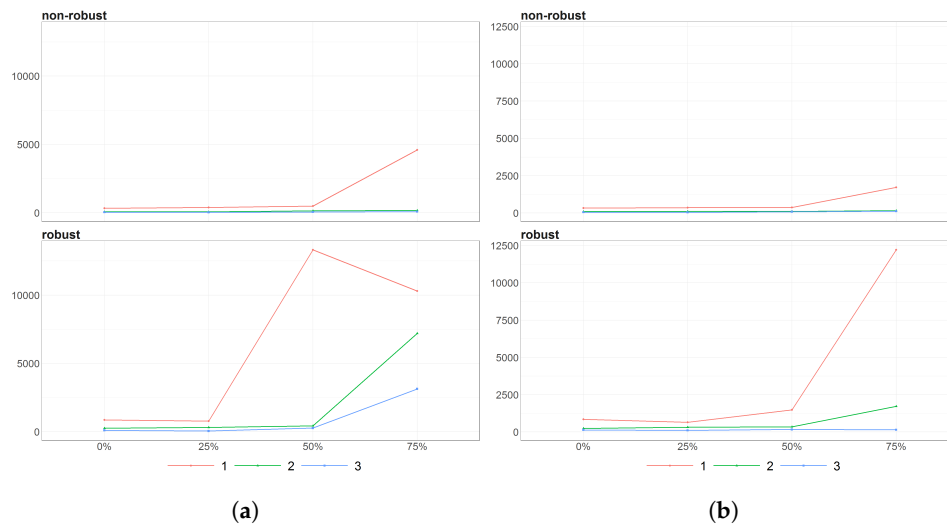
Figure 18. The eigenvalues of Deaths counts obtained using PPCA for Females (**a**) and Males (**b**) over diferent cases of missing entries (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard deviations (upper plot) and sample one (bottom plot).

The scores are presented in Figures 21 and 22. The second and third scores of Males are smoother in contrary to results for Females. The exception is made for the case 75% which results vary both among sexes and standardization procedures.
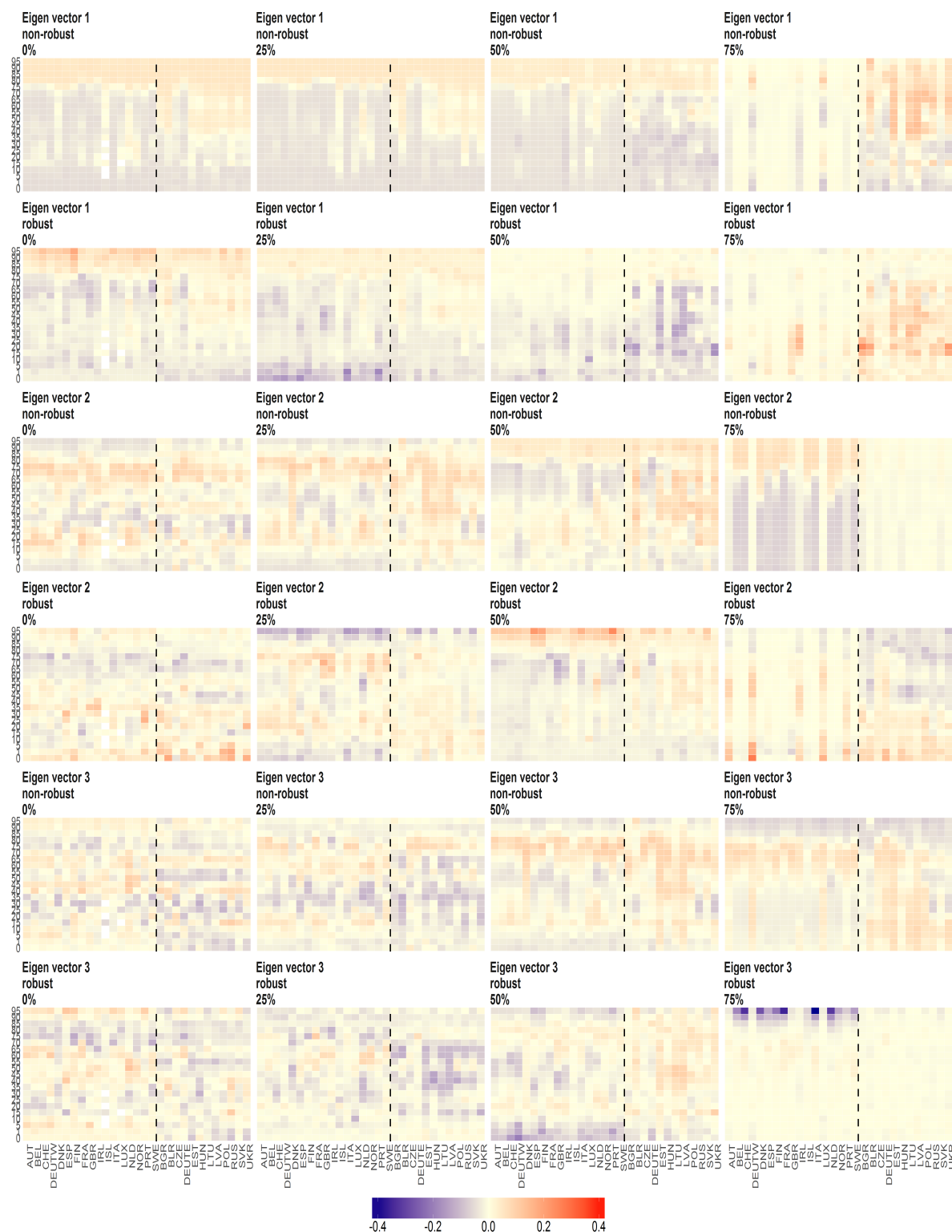
Figure 19. The eigenvectors of Death counts (y axis) over age groups (y axis) and countries (x axis) obtained using PPCA for Females. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.
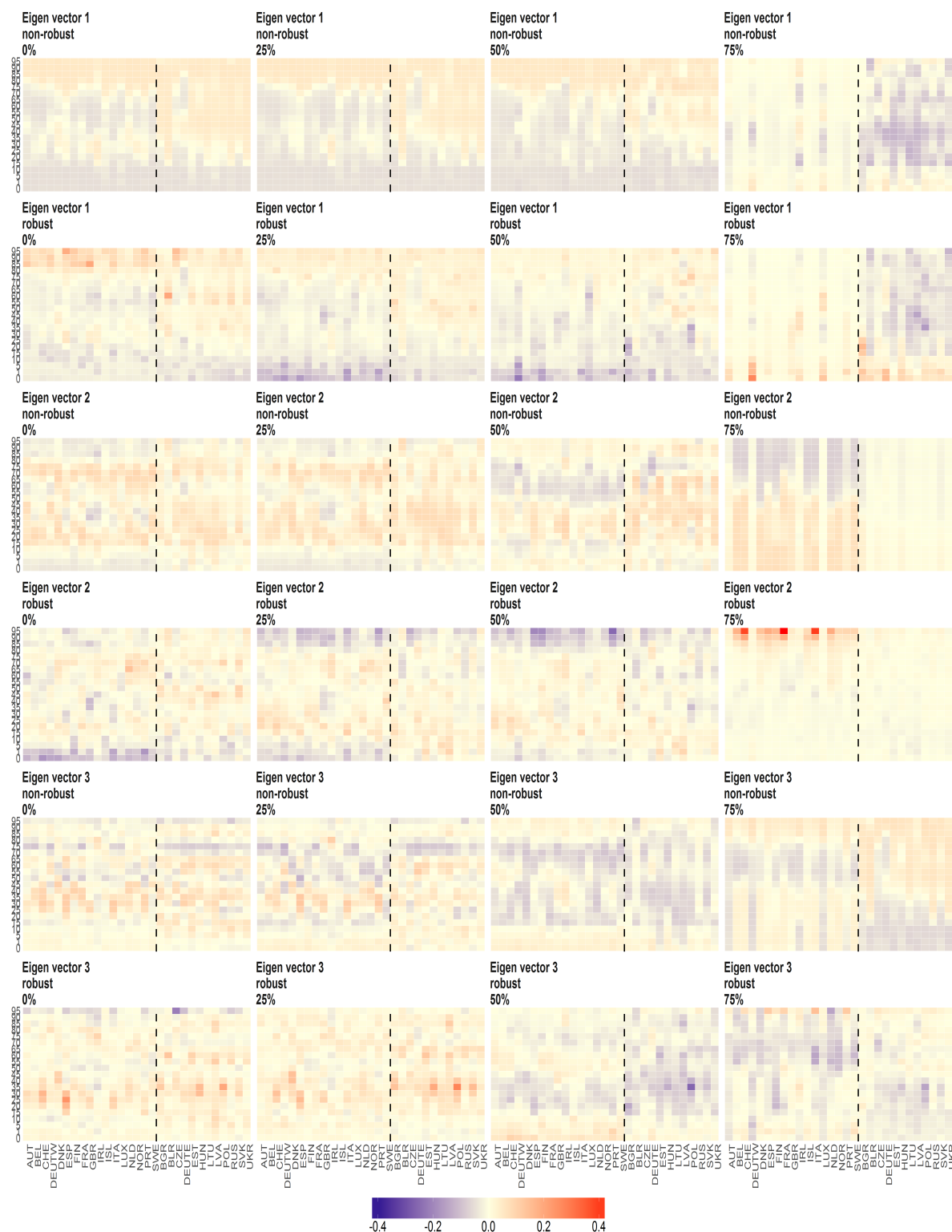
Figure 20. The eigenvectors of Death counts (y axis) over age groups (y axis) and countries (x axis) obtained using PPCA for Males . Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.
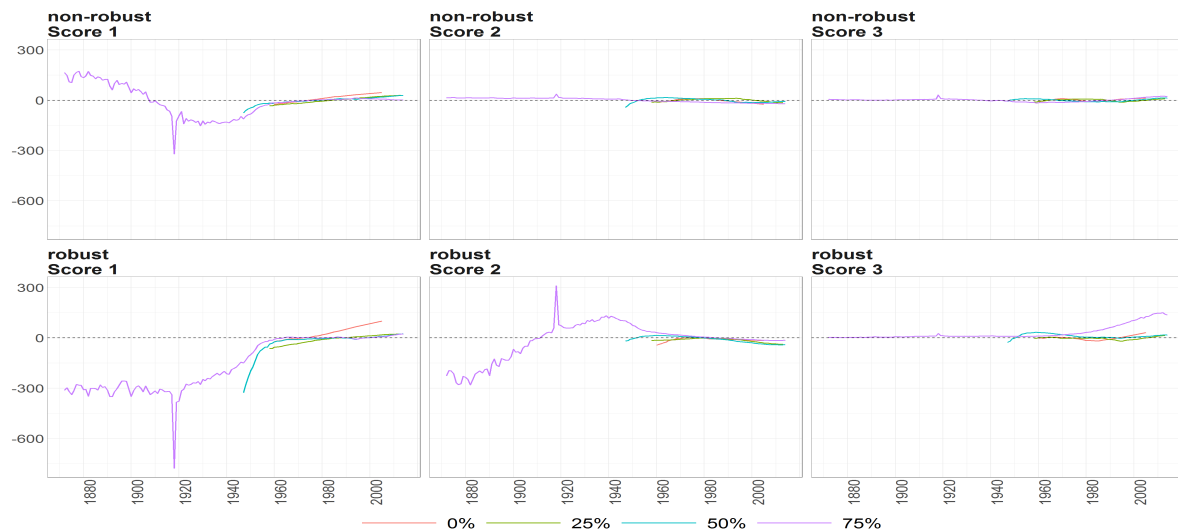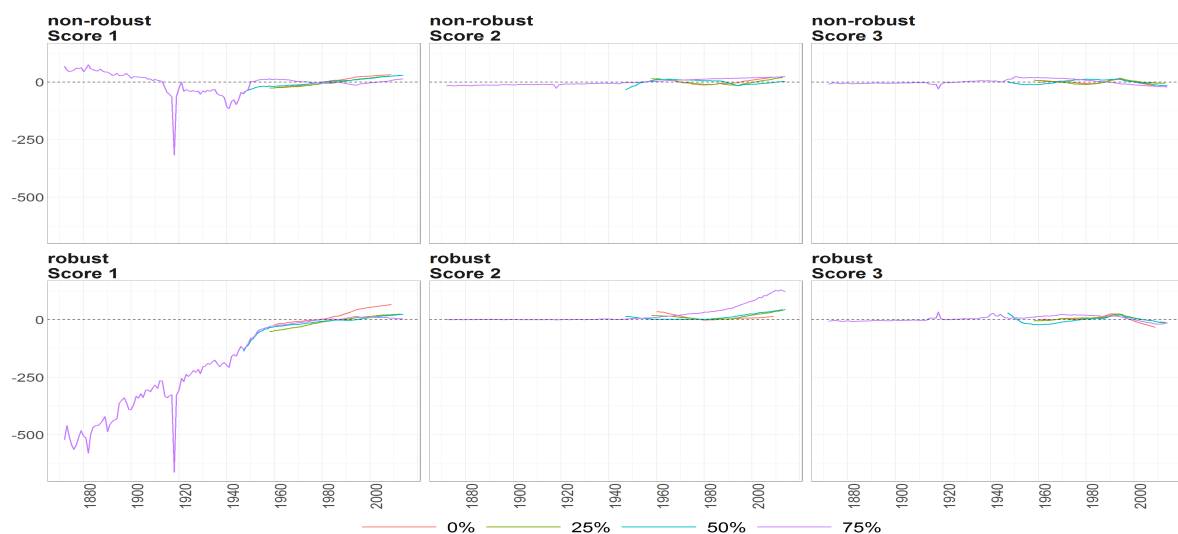
Figure 21. The scores (y axis) over time (x axis) obtained using PPCA for female population of Number of Deaths. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

Figure 22. The scores (y axis) over time (x axis) obtained using PPCA for male population of Number of Deaths. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

8.5. Death Rates

The analysis for Death Rates provides with similar conclusions as for Death counts. The corresponding Figures are 23, 24, 25 and 26 respectively. The only discrepancies are exhibited by the eigenvalues for Males for the 75% case as they are more aligned and higher in terms of magnitude from other cases than the corresponding eigenvalues in the Deaths counts analysis.

As shown in Figures 27 and 28 the robustification does not influenced greatly the estimation of eigenvectors. As the Deaths Rates are preprocessed and smoothed before being distributed by HMD, we again conclude that the preprocessing decreased number of outlying data points. In this particular case, the robust standardisation is similarly informative about the the true distribution as the non robust one.

Also, recall the similarity of results among different cases of missingness, especially for first eigenvector. The second and third ones are smoother for high levels of missigness. The colour map is affected by the scaling parameter $(1, -1)$ which may cause red to become blue, but except this fact, we see resembling outcomes.

The three most meaningful eigenvectors differ among two groups of analysed countries: developed and developing. The corresponding results are divided by the vertical black dotted lines. The first eigenvectors of developed countries is very smooth among age groups. For developing countries we observe U shape structure with peak in 45–60 age groups for Males and similarly for Females in the cases with small proportion of missigness. When we allow more missing values, the eigenvectors for Females in developed countries are closer to zero and flat among age groups whereas for developing countries are more distant and volatile.
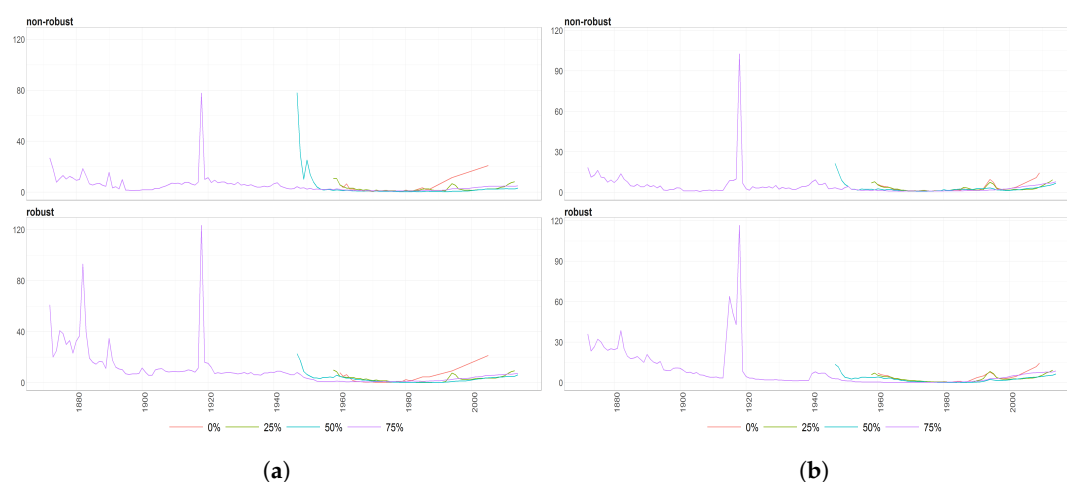

Figure 23. The Mahalanobias distances obtained using PPCA for female (**a**) and male (**b**) population of Death Rates over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing entries in rows (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).
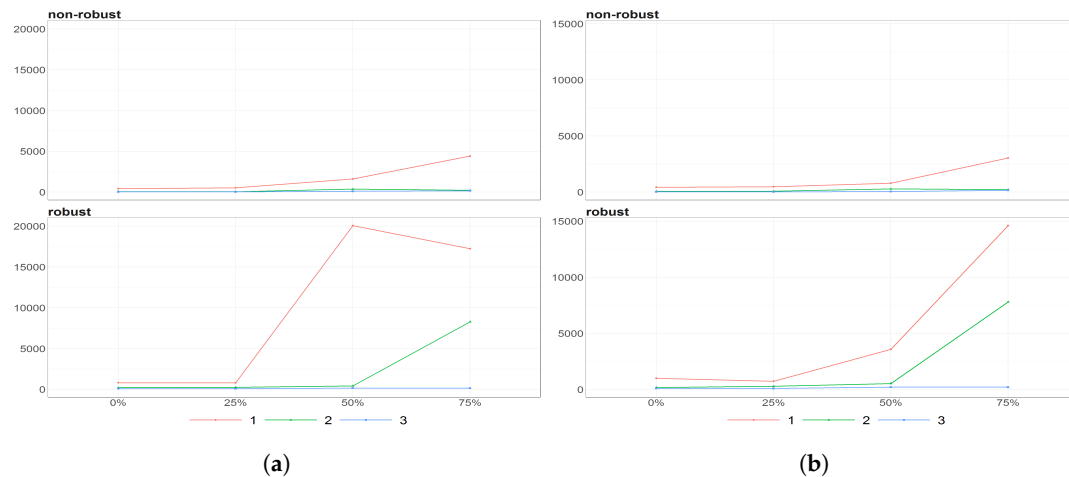
Figure 24. The eigenvalues obtained using PPCA for female (**a**) and male (**b**) population of Death Rates for different percentages of maximal missing entries in rows (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).
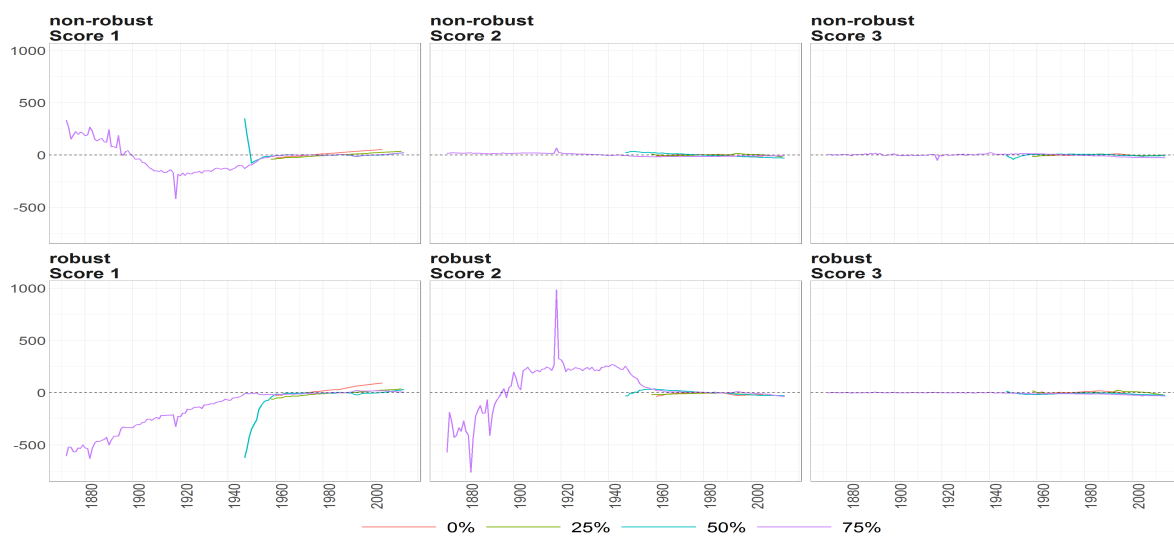

Figure 25. The scores (y axis) over time (x axis) obtained using PPCA for female population of Death Rates. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.
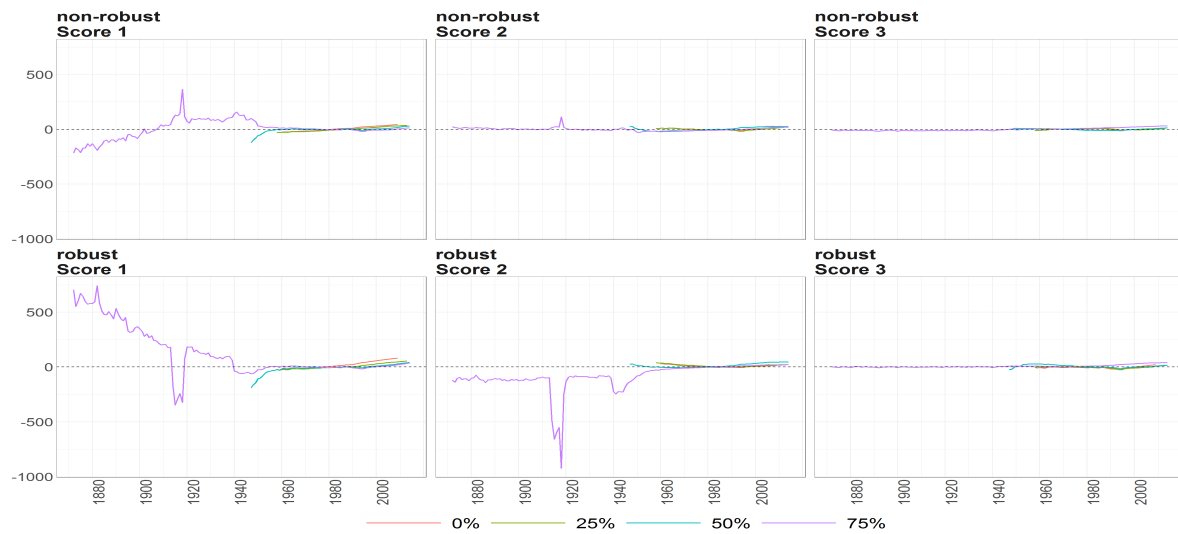
Figure 26. The scores (y axis) over time (x axis) obtained using PPCA for male population of Death Rates. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non-robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.
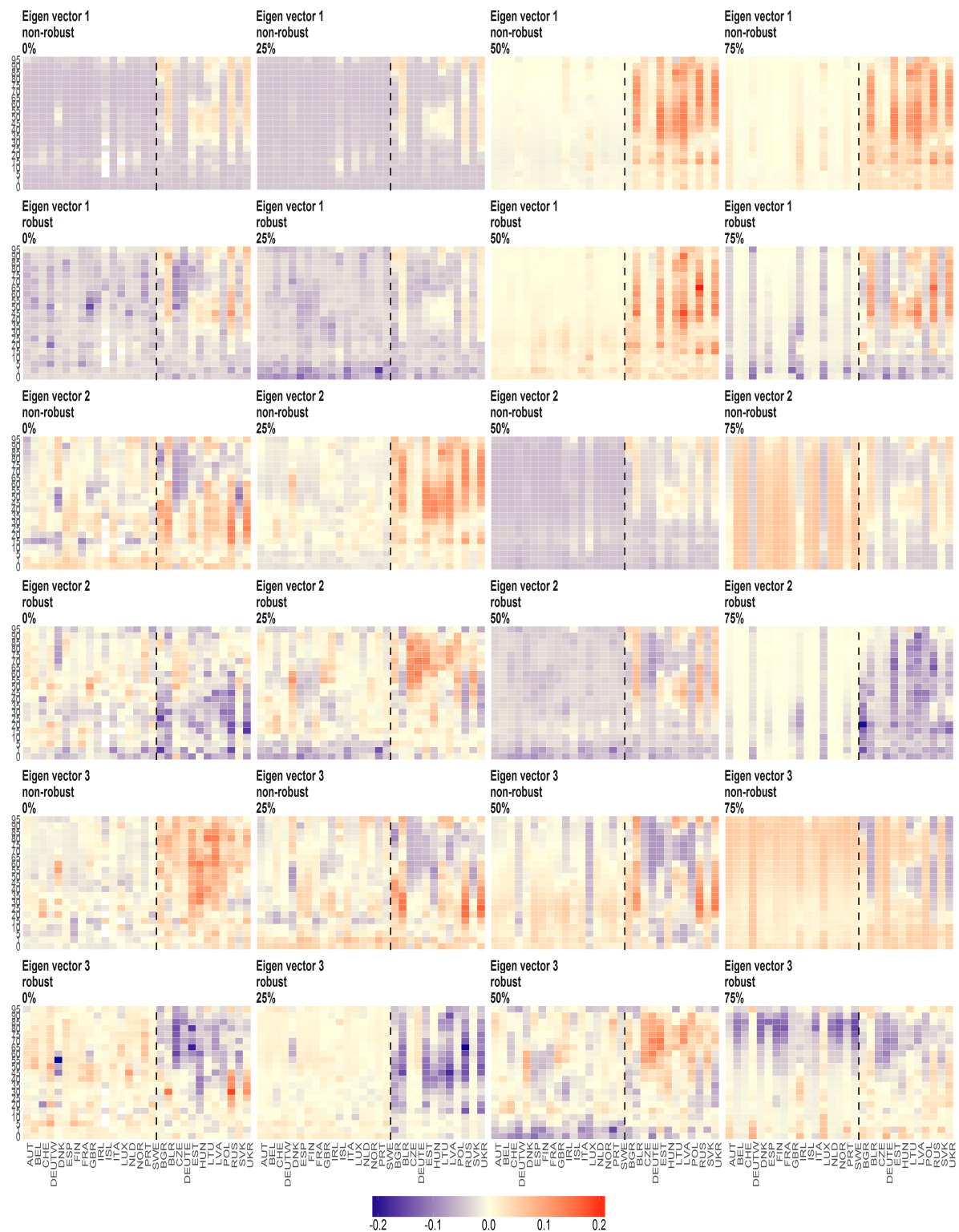
Figure 27. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for female population of Death Rates. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

Figure 28. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for male population of Death Rates. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.
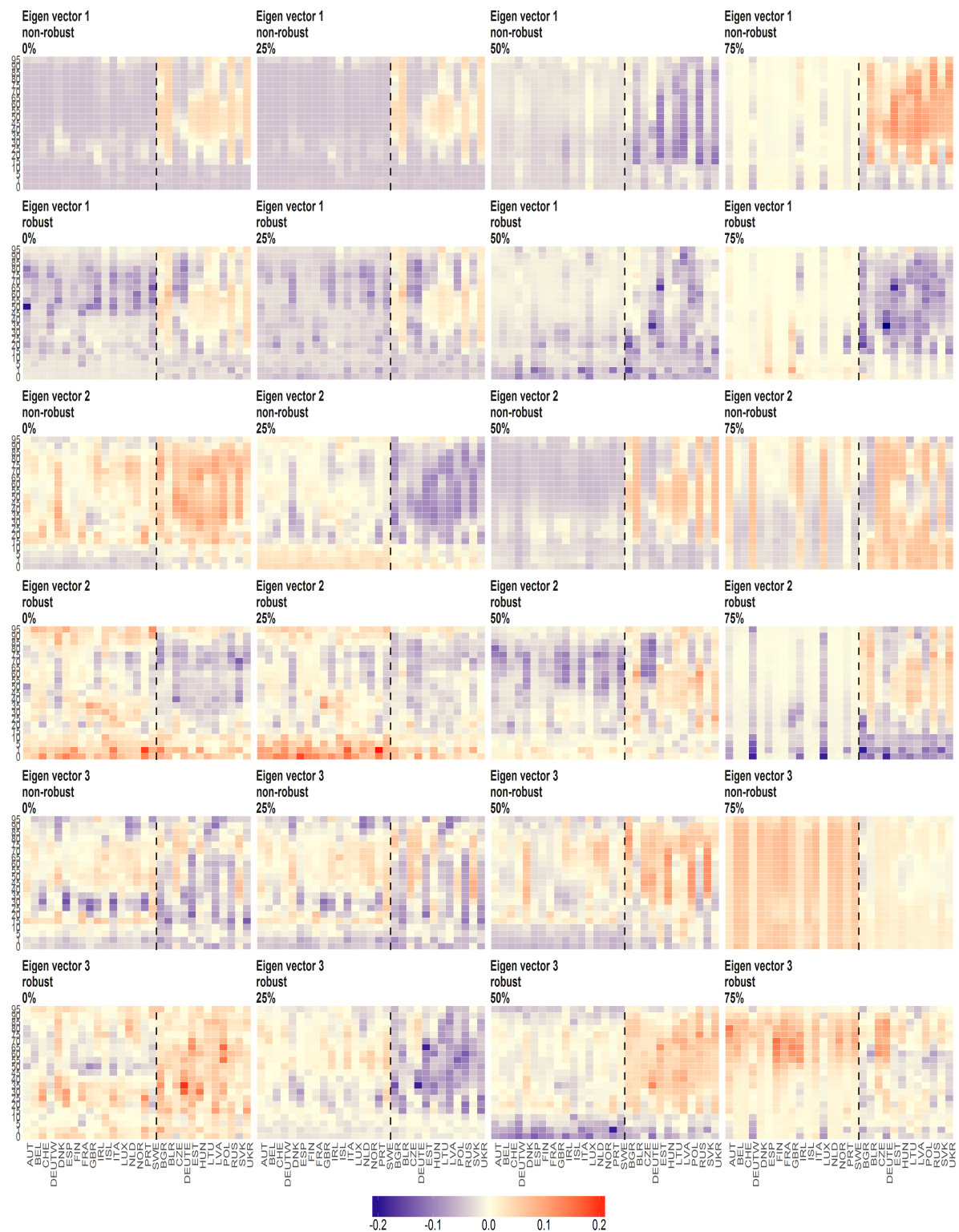
9. Stochastic Mortality Models for UK utilizing Factor Extraction from European Demographic Data

In this section we demonstrate the results of incorporating features extracted from European demographic data in the stochastic mortality model for British female mortality data over a study period of 1922 to 2014 with 10 years ahead forecasting validation.

The key findings are that the utilisation of demographic features improves the in-sample and out-of-sample predictive posterior mean Bayesian point estimation and forecasts for the log death rates. Additionally, the employed robustification methodology reduces the variances of the error terms in both observation and state equations and produces a better out-of-sample fit than its non-robust alternative. It indicates that the features extracted in the robust manner are more consistent over time and capture better the information about the true distribution of the demographic data.

The model which has the smallest mean square error of estimation and prediction adds the age-specific components to the latent process. It is later referred as DFM-PC-B. However, the other examined models, which also improve the predictability of the log death rates, are useful in terms of the interpretation as they reveal the individual country specific impacts of each of the European countries data on British female log death rates.

9.1. Description of the Models

Before presenting the real data example we note that all the Bayesian models developed and Markov chain samplers constructed were first tested on synthetic case studies in which the true parameters and state variables are known. The performance was found to be very good and this provides confidence in the accuracy and performance behaviour of the methods and models developed. The synthetic study results are provided in technical appendix and are not discussed in this paper. The use of synthetic data enables us to validate the estimation by Forward-Backward Kalman Filter with Gibbs Sampler. The models that we considered in our simulations and empirical studies are labelled by

LCC: Lee-Carter model with the stochastic cohort effect given by Equations (2) and (3);

DFM-PC: Demographic factor model which incorporates ϱ_t into LCC given by Equation (8). Please refer to Appendix B for illustration how the models from this class are created;

> **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:** 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;
>
> **DFM-PC-D-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, non-robust standardisation;
>
> **DFM-PC-Mx-r:** 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;
>
> **DFM-PC-Mx-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, non-robust standardisation;

The models of the class DFM-PC address **Case 1** from Section 3, where factors are incorporated into the observation Equation (2). The factors are obtained by performing PPCA jointly on the set of data for all countries listed in Table 1 excluding the following specific countries: United Kingdom (as it is our response variable), Greece and Slovakia (due to short time series).

DFM-PC-B incorporates the mean of the first principal component of the Birth counts which is a country specific vector. The matrix $\tilde{\mathbf{F}}$ is a 21×21 diagonal matrix with the mean on the diagonal.

Hence, ϱ_t which correspond to the model DFM-PC-B, is a 21 dimensional, age specific state process and attempts to capture an age-specific dynamic in addition to the cohort-period effects.

The models DFM-PC-D-r, DFM-PC-D-s and DFM-PC-Mx-r and DFM-PC-Mx-s incorporate the first component of Death counts and Death Rates, respectively. Recall that the components for these data sets can be presented as age specific and country specific matrices as shown in Section 8. Due to the high dimensionality of the problem, we want the one element of ϱ_t to correspond to the subvector of the first component which is specific only for one country. Such a subvector has 21 dimensions which correspond to the age groups. Hence, ϱ_t is a 28 dimensional country specific state process. The country specific subvectors of the first components are placed in the columns of the 21×28 matrix $\bar{\mathbf{F}}$. The last letter of the name of the models DFM-PC-D and DFM-PC-Mx denotes the type of the standardisation which is applied to the data before performing PPCA: robust (by M-estimator) or non-robust (sample estimator).

In the following part we analyse the population mortality from United Kingdom based on the models listed above and Bayesian methodology studied in this paper. We then examine the models in terms of the forecasting properties of death rates.

9.2. Setup

For the Bayesian estimation of models, we assume the priors given in the Appendix A.3 and A.4 to be

$$\kappa_0 \sim \mathcal{N}(0, 10^2), \quad , \gamma_0 \sim \mathcal{N}(0, 10^2), \quad \alpha_x \sim \mathcal{N}(0, 10^2), \quad \beta_x \sim \mathcal{N}(0, 10^2),$$
$$\sigma_\varepsilon^2 \sim \text{IG}(2.01, 0.01), \quad \theta \sim \mathcal{N}(0, 10^2), \eta \sim \mathcal{N}(0, 10^2), \quad \lambda \sim \mathcal{N}_{[-1,1]}(0, 10^2),$$
$$\sigma_\kappa^2 \sim \text{IG}(2.01, 0.01), \quad \sigma_\gamma^2 \sim \text{IG}(2.01, 0.01), \quad \varrho_0^i \sim \mathcal{N}(0, 10^2,)$$
$$[\mathbf{\Omega}]_{i,j} \sim \mathcal{N}(0, 10^2), \quad \Psi_j \sim \mathcal{N}(0, 10^2), \quad \sigma_\varrho^2 \sim \text{IG}(2.01, 0.01).$$

The number of iterations of the Markov chain is 50,000 for LCC model and 200,000 for other models with 90% burn-in. The chain is initialised at $\boldsymbol{\alpha} = \bar{\mathbf{y}}_{1:T}$, $\beta_x = \frac{1}{21}$, $\sigma_\varepsilon^2 = 0.0005$, $\theta = -0.005$, $\eta = -0.02$, $\sigma_\kappa^2 = 0.01$, $\sigma_\gamma^2 = 0.0005$, $\sigma_\varrho^2 = 1.0$, $[\mathbf{\Omega}]_{i,i} = \frac{1}{m}$ for m being either number of countries or number of age groups (depending on the model). The convergence of the sampler has been tested on synthetic data studies. The synthetic data study revealed that the estimation of the drift parameters corresponding to the factor state process model ϱ_t converges very slowly for shorter time series such as those found in mortality data. Thus, we decided to set these parameters to zero and do not sample them in this study.

9.3. Estimation of Static Parameters

Estimated values of the static parameters (except $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$) for the British female mortality data (1922–2003) are shown in Table 2. The rest of the estimated static parameters is displayed in Figures 29, 30 and 31. The results are shown under different models listed in the first columns of the table or indicated by the colour of lines on the plots.

The static parameters of the factor process under DFM-PC-B model are age specific. In addition to the cohort and period effect, they provide supplementary information related to the corresponding age groups. Figure 29 shows the estimated diagonal elements of the transition matrix $\boldsymbol{\Omega}$ under the model. The parameters with values close to unity indicate that the factor state process corresponding to these parameters have a slowly decreasing dynamic. The elements of the state processes which correspond to the values of parameters closer to zero, are characterized by higher decrease. The parameters which are negative and close to zero indicate that the corresponding latent state processes fluctuate around zero.

With regards to DFM-PC-B model which incorporates age specific latent processes (supplementary to cohort effect), the elements of $\boldsymbol{\Omega}$ which are positive and close to zero, describe the decreasing dynamic of the corresponding age specific processes. These latent processes are shown to have more significant impact on death rate modelling when the sample starts, however, this impact decreases over

time and the cohort effect becomes sufficient to model log death rates in these age groups. For instance, recall the age groups between 70–80 in Figure 29. Such process can be interpreted as a period effects which are specific for particular age groups. The elements of Ω which are close to unity describe the latent process which have consistent impact or its lack over the time. If they have an effect on log death rates (i.e., their domain is not close to zero), they covey age-specific information which is supplementary to cohort and period effect and consistently demanded by the model over the time.

With regards to Ω estimated under models DFM-PC-D-r, DFM-PC-D-s and DFM-PC-Mx-r and DFM-PC-Mx-s, it refers to the country specific features. Here the parameters are related to the influences of the specific countries on British female log death rates. The plots with the estimates and their confidence intervals are displayed in Figure 31. For instance, estimates of Ω under all models agrees on lack of effects of Austrian or Bulgarian demographic data on the log death rates over whole sample span. On the other hand, the estimate of the parameter corresponding to Belarusian eigenvector is close to unity under all models and therefore highlights the informative effect of the feature on British log mortality rates which is consistent over the times.

It is worth to point out that the values of estimated variances of the observation and state equations error terms are higher for DFM-PC-D-s and DFM-PC-Mx-s where the data has been non-robustly standardized. These models are examined to have a greater mean square error of in-sample and out-of-sample fit than their robust alternatives as shown in Table 4. Hence, the robustification procedure employed in this study improved the overall goodness of fit of the considered models. The features which have been extracted from European demographic data by means of robust estimators of mean and covariance are shown to provide the information which is consistent over the times and conveys the better knowledge about the true distribution of the demographic data sets.

Table 2. Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the static parameters $\left(\lambda,\ \theta,\ \eta, \sigma_\epsilon^2,\ \sigma_\gamma^2,\ \sigma_\kappa^2, \sigma_\varrho^2\right)$ of $\log m_{x,t}$.

Model	λ	θ	η	σ_ϵ^2	σ_γ^2	σ_κ^2	σ_ϱ^2
LCC	0.998 (0.994; 1)	−0.154 (−0.331; 0.026)	−0.024 (−0.034; −0.014)	6.4×10^{-3} $(6\times10^{-3}; 6.9\times10^{-3})$	2×10^{-3} $(1.4\times10^{-3}; 2.8\times10^{-3})$	0.663 (0.449; 0.96)	
DFM-PC-B	0.991 (0.968; 1)	−0.332 (−0.53; −0.137)	−0.005 (−0.01; 0.002)	8×10^{-4} $(6\times10^{-4}; 9\times10^{-4})$	5×10^{-4} $(4\times10^{-4}; 7\times10^{-4})$	0.753 (0.537; 1.055)	0.049 (0.042; 0.057)
DFM-PC-D-r	0.949 (0.913; 0.992)	−0.246 (−0.415; −0.101)	0.011 (−0.002; 0.021)	1×10^{-3} $(9\times10^{-4}; 1.1\times10^{-3})$	5×10^{-4} $(4\times10^{-4}; 8\times10^{-4})$	0.39 (0.227; 0.739)	0.092 (0.074; 0.113)
DFM-PC-D-s	0.998 (0.993; 1)	−0.093 (−0.221; 0.029)	−0.019 (−0.025; −0.013)	1.3×10^{-4} $(1.1\times10^{-4}; 1.4\times10^{-4})$	8×10^{-4} $(5\times10^{-4}; 1.1\times10^{-3})$	0.324 (0.152; 0.616)	0.144 (0.114; 0.178)
DFM-PC-Mx-r	0.985 (0.959; 0.999)	−0.042 (−0.115; 0)	−0.013 (−0.02; −0.007)	8×10^{-4} $(7\times10^{-4}; 1\times10^{-3})$	6×10^{-4} $(4\times10^{-4}; 8\times10^{-4})$	0.044 (0.002; 0.116)	0.08 (0.066; 0.094)
DFM-PC-Mx-s	0.999 (0.995; 1)	−0.024 (−0.111; 0.044)	−0.02 (−0.03; −0.01)	1.2×10^{-4} $(8\times10^{-4}; 2.1\times10^{-4})$	1.7×10^{-3} $(1.2\times10^{-3}; 2.3\times10^{-3})$	0.036 (0.001; 0.137)	0.834 (0.594; 0.994)

We did not choose to calculate the MLE estimates of the parameter's in our models as it has been documented that even for the standard period-cohort type Lee-Carter stochastic mortality models, the classical MLE estimation frameworks can produce convergence and estimation challenges due to gradient based and method of scoring recursive optimization methods getting stuck in local optima of the marginal likelihood surface. We refer the interested reader to the paper Fung et al. (2017) where we discuss such issues in more depth. Therefore, instead of resolving the known problems that may arise with classical MLE estimations of such models, which may be further compounded in the extended models we developed in the frequentist setting, we have chosen to stick with the Bayesian modelling paradigm and to report an analogue result to the MLE that may be obtained from Bayesian inference, in the case of uninformative priors. That is we have relative uninformative priors and so we can report the Maximum a-postiori (MAP) posterior mode estimator for the parameters as the Bayesian analogue of the MLE, defined as

$$\boldsymbol{\psi}^{MAP} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}} \, \pi(\boldsymbol{\varrho}_{0:T}, \boldsymbol{\psi}|\mathbf{y}_{1:T}) \tag{97}$$

for $\pi(\boldsymbol{\varrho}_{0:T}, \boldsymbol{\psi}|\mathbf{y}_{1:T})$ being a joint posterior density of the states $\boldsymbol{\varrho}_{0:T}$ and the vector of static parameters $\boldsymbol{\psi}$ given the observation $\mathbf{y}_{1:T}$ as introduced in Appendix A.

The MAP estimates and MLE should be similar in the case of uninformative priors, with the advantage that the MAP estimation is obtained via an MCMC sampler output, which is less prone to the types of estimation challenges experienced in gradient descent methods working directly with the Instead, as we used fairly uninformative priors, note that the MAP estimate of the posterior is a case of uninformative priors will correspond to MLE estimates. Please refer to Table 3 for the analogous of point estimates to the outcomes of Table 2.

Table 3. The MAP estimates of the static parameters $\left(\lambda, \theta, \eta, \sigma_\epsilon^2, \sigma_\gamma^2, \sigma_\kappa^2, \sigma_\varrho^2\right)$ of $\log m_{x,t}$.

Model	λ	θ	η	σ_ϵ^2	σ_γ^2	σ_κ^2	σ_ϱ^2
LCC	0.999	−0.155	−0.024	0.0064	0.0019	0.6172	–
DFM-PC-B	0.998	−0.331	−0.005	8.00×10^{-4}	5.00×10^{-4}	0.7279	0.0487
DFM-PC-D-r	0.948	−0.253	0.012	0.001	5.00×10^{-4}	0.3234	0.0895
DFM-PC-D-s	0.999	−0.095	−0.02	0.0013	7.00×10^{-4}	0.3005	0.1416
DFM-PC-Mx-r	0.995	−0.023	−0.012	7.00×10^{-4}	6.00×10^{-4}	0.029	0.0818
DFM-PC-Mx-s	1	−0.023	−0.021	9.00×10^{-4}	0.0016	0.0088	0.8583

9.4. Filtering of Latent Variables

The Bayesian posterior mean estimates of the latent stochastic mortality factors in the models for κ_t in the top panel and for γ_t^0 in the bottom panel of Figure 32. The colours of lines denote the filtered processes under different models. As expected, adding new state variables related to the factors significantly changes the dynamics of the period and cohort effect state processes. The blue line correspond to the cohort=period only LCC model. The increase of κ_t and decrease of γ_t^0 at the end of the sample is greater for this model in contrast to the the other examined models.
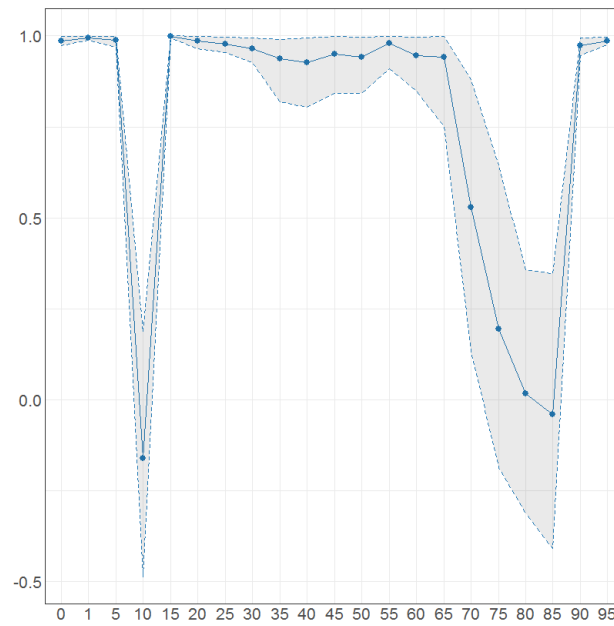
Figure 29. Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the age-specific diagonal elements of the transition matrix Ω (x axis) under DFM-PC-B.
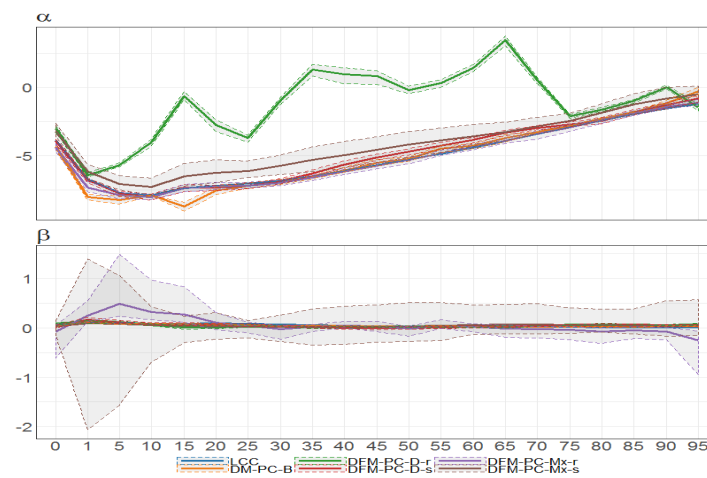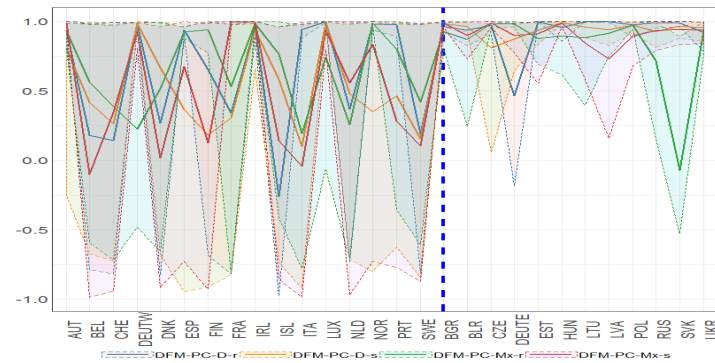

Figure 30. Bayesian posterior estimators with 95% posterior credible intervals for the estimation of α and β under different models (colours of lines) for British female mortality data (1922–2002).

Figure 31. Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the diagonal elements of the transition matrix Ω (x axis) under DFM-PC-D-r, DFM-PC-D-s, DFM-PC-Mx-r and DFM-PC-Mx-s models (colours of lines). The dashed blue line divides the set of countries into developed (on the left side) and developing (on the right hand side) European countries, respectively.

The dynamic of the cohort effect latent process vector γ_t over time, which is age group specific, is shown in Figure 34. The panels correspond to the Bayesian posterior mean estimates of the process under different models. The age group specific features, which has been utilised in DFM-PC-B model, clearly provide LCC model with supplementary information to the cohort effect state processes. The corresponding ϱ_t reduces the variability of the cohort effect process with comparison to the cohort effect estimated under LCC model (the colours of the surface on DFM-PC-B panel are plain and variance of the error term is smaller). The state processes corresponding to the factors under DFM-PC-B model are shown in Figure 33. Recall that in the contrast to the cohort and period effects processes, the latent process vector ϱ_t under the model DFM-PC-B has age group specific stochastic components (κ_t is calendar year specific and γ_t is '0' age group and calendar year specific). As κ_t latent process estimated under DFM-PC-B is not distant from the corresponding estimated under LCC model for majority of the sample, we can conclude that the additional state processes given by this model provide supplementary, age-specific information to the cohort effect process γ_t^0.
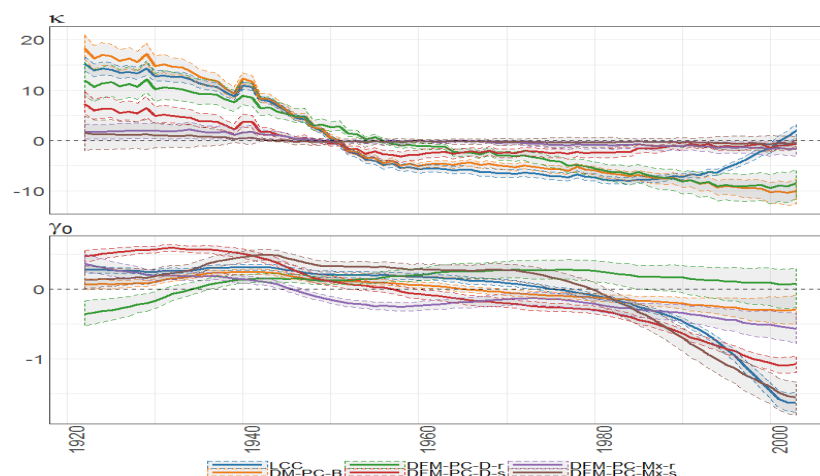

Figure 32. The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (upper panel) and cohort effect state process γ_t^0 (lower panel) under different models (colours of lines) for British female log death rates during 1922–2002.
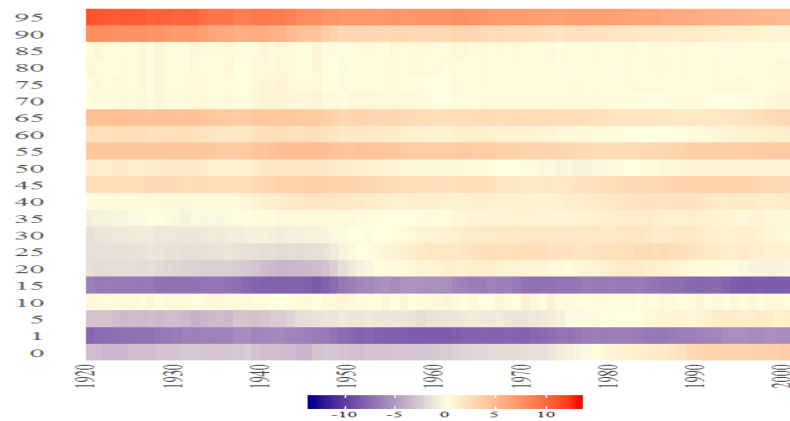
Figure 33. The Bayesian posterior mean estimates for ϱ_t across age groups (y axis) over time (x axis) under DFM-PC-B model for British female log death rates during 1922–2002.
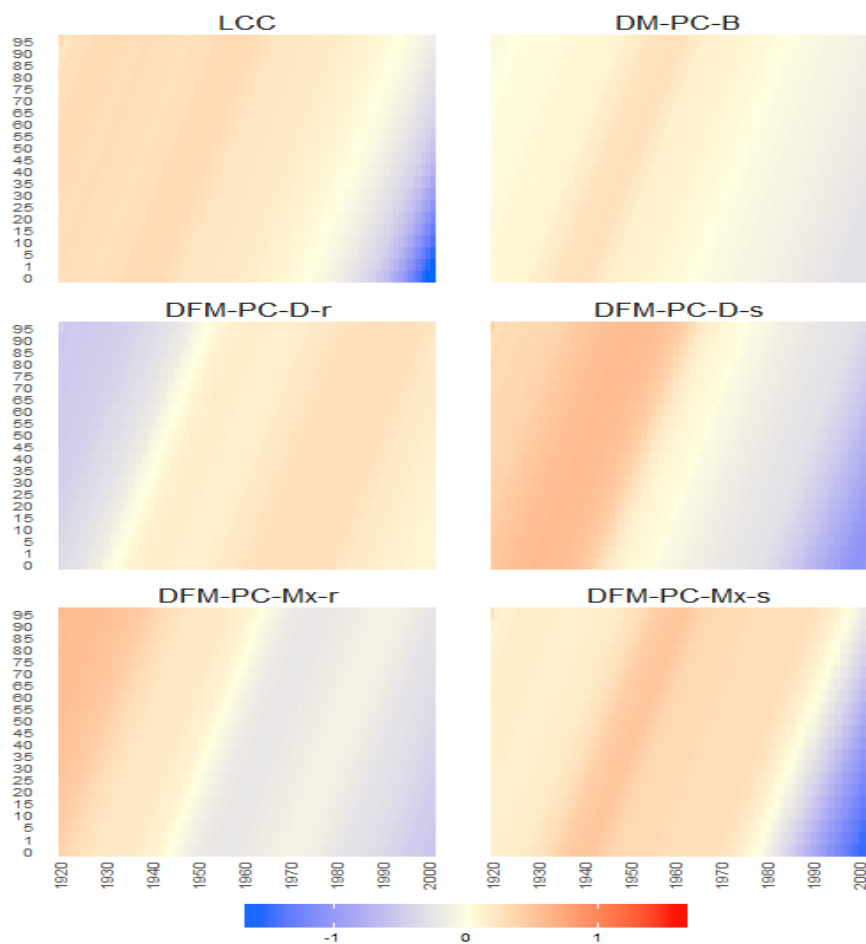

Figure 34. The Bayesian posterior mean estimates for the cohort effect latent processes vector γ_t across age groups (y axis) over time (x axis) under different models for British female log death rates during 1922–2002.

The models DFM-PC-D-r, DFM-PC-D-s, DFM-PC-Mx-r and DFM-PC-Mx-s are characterized by the processes which correspond to the country specific vectors with elements related to age groups.

The models attempt to address the question whether the structure of demographic data from European countries can be an efficient explanatory variable for British mortality. Figures 32 and 34 show that dynamics of period and cohort effect state processes are sensitive to different sets of features extracted from demographic data as well as a standardisation methodology. Figures 33 and 35 provide insight into how these variables change the information extracted from the models. The information highlight the impact of the first components of the considered countries on British mortality rates. As the components are orthonormal, $\varrho_t^{country}$ indicates the magnitude of this effect. To be consistent with the country set specific notation from Section 8, the blue vertical line divides the results into two categories: results for the developed countries (below the blue line) and developing countries (above the blue line). The processes which moves closely to zero give information that the data of country they correspond to has small influence on the mortality rates from United Kingdom. The developed models when incorporating the demographic features from the European countries, indicate the significance of the factor loading from a given country on the mortality of UK. Hence, we can specify the countries which has a positive effect on the mortality of United Kingdom (factor state processes are negative), neutral (factor state processes are fluctuate around zero) or negative (factor state processes are positive and enlarge log death rates).

For a given model it is not true that all factors which correspond to the age specific vectors of features from European countries, have an influence in the causal fashion on the UK morality data in the same way. As such, what we are showing in the plots Figure 31 (or Figure 29 for DFM-PC-B models) that some countries have very wide posterior credible intervals (the flat posterior) for $\boldsymbol{\Omega}$ which is in an alignment with the findings in Figure 36 (Figure 33 for DFM-PC-B model) where we see that indeed for those countries the ϱ_t upon a model indicate insignificant effect on UK log mortality rate. Let us consider the example of DFM-PC-Mx-s (the red colours of lines on both of the plots) for Belgium (BEL) and Austria (AUT). We see in Figure 36, that the effect of Austrian factor to the British mortality data, expressed by the dynamic of ϱ_{AUT}, is non-zero over the time. On the other has, the element of ϱ_t corresponding to Belgium, labelled by ϱ_{BEL} does not load significantly in any way on the UK mortality experience. As a consequence we see that the posterior for this country on $\boldsymbol{\Omega}$ is also very flat when the credit intervals of Ω_{AUT} on Austria are significantly narrower. This simply means that the factor loading of Belgium does not influence the UK mortality experience.

Hence, to conclude this discussion, we note that when we look at the results in Figure 36, they show the effect of each individual countries influence on the UK mortality experience. In fact, what we learn is that some countries have a mean of 0 with large uncertainty, these countries maybe interpreted as not having an influence on the mortality experience of the UK.

The four models are more consistent about the set of countries which does not have any 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 as Lithua, Poland or Russia.
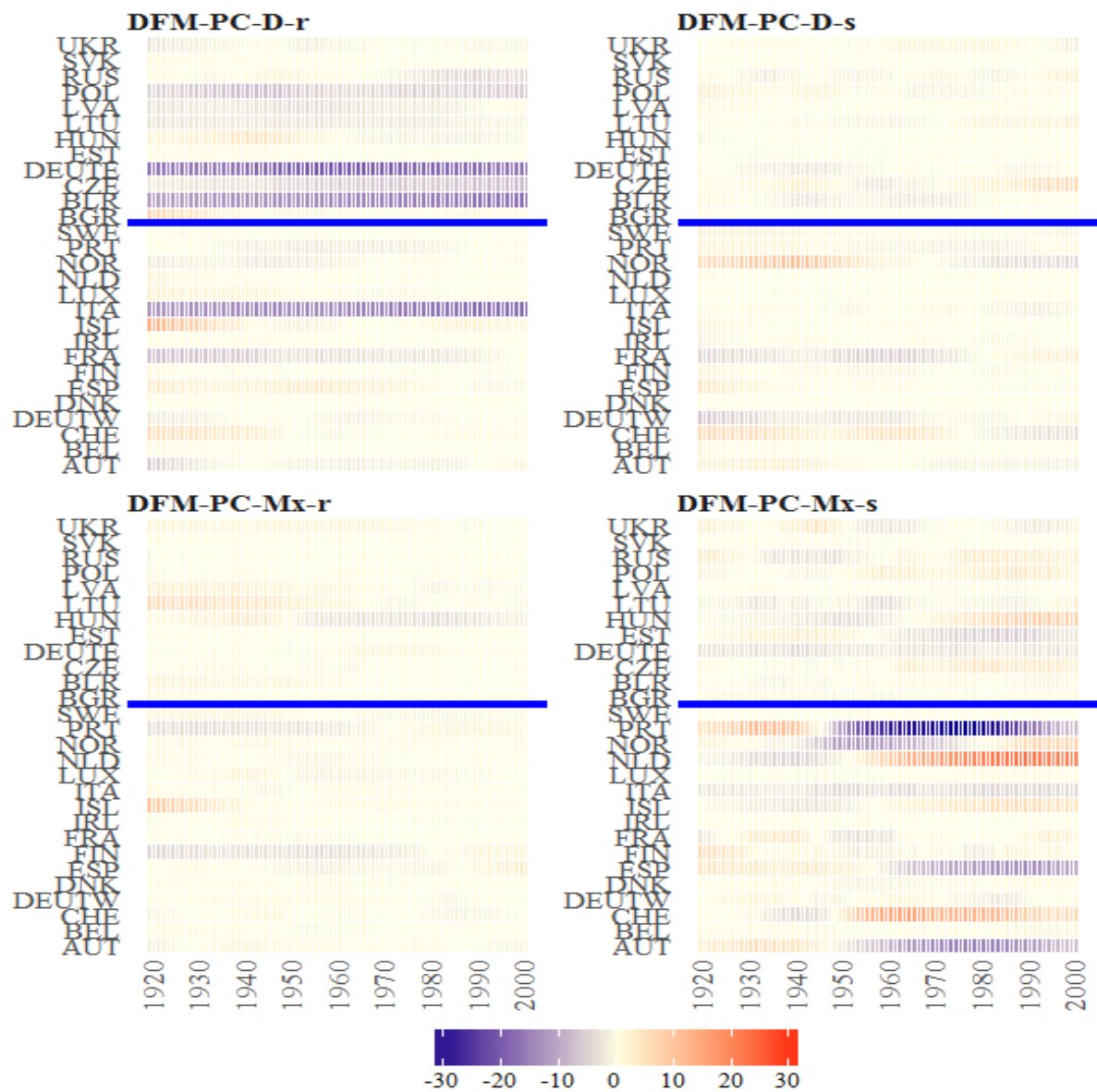
Figure 35. The Bayesian posterior mean estimates for ϱ_t across countries (y axis) over time (x axis) under the models from the classes DFM-PC-D and DFM-PC-Mx for British female log death rates during 1922–2002. The vertical blue line divides sets into developed (on the left side) and developing (on the right sie) European countries.
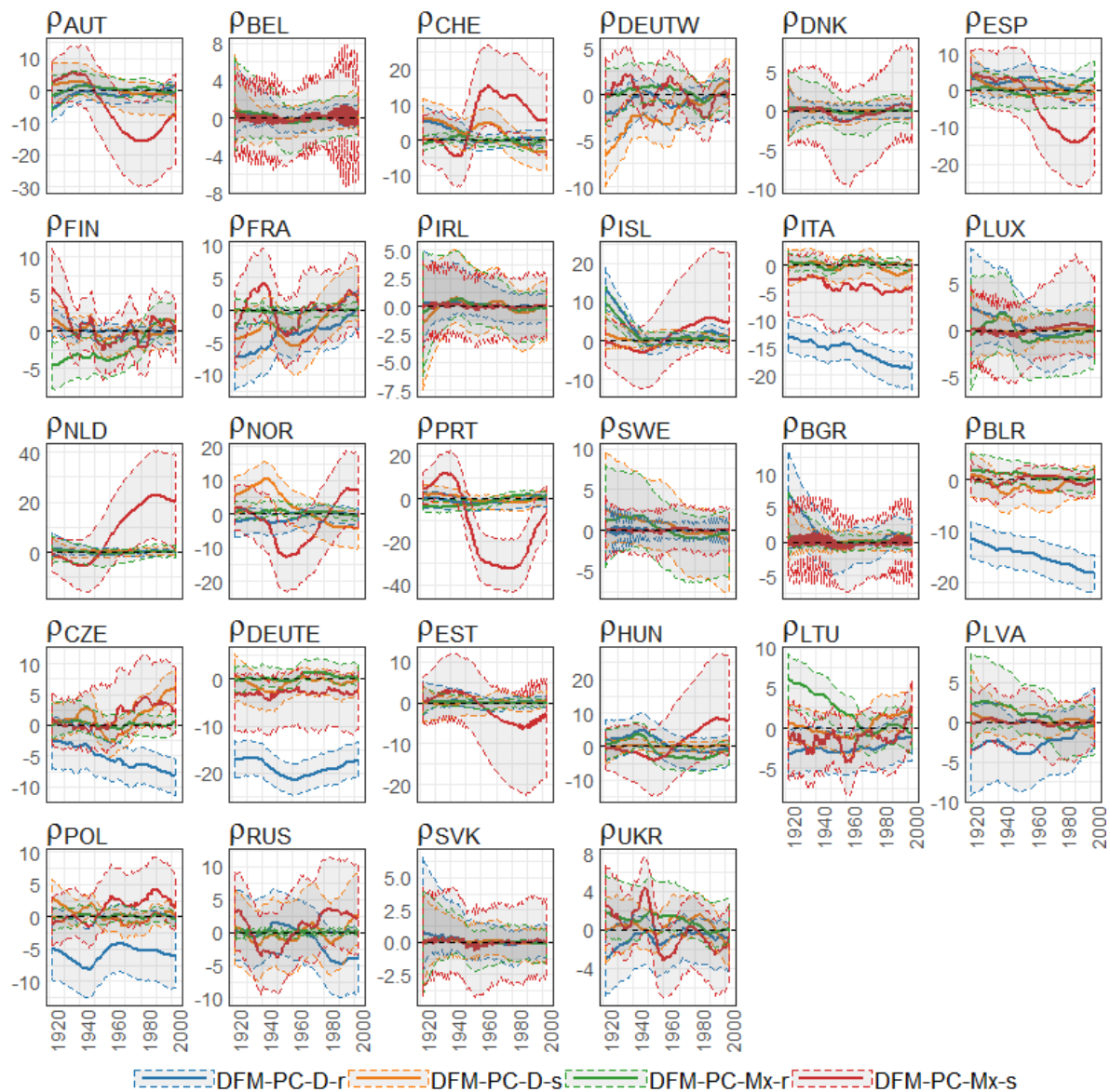
Figure 36. The Bayesian posterior mean estimates with 95% posterior credible intervals for ϱ_t across countries (different panels) overtime (x axis) under the models from the classes DFM-PC-D and DFM-PC–Mx (colours of lines) for British female log death rates during 1922–2002.

9.5. In-Sample and Out-of-Sample Performance

In this section, we investigate the in-sample and out-of-sample performance of the mortality models summarized in the introduction to this section. The model selection is based on two in-sample performance measures, MSE (Mean Square Error) and DIC (Deviance Information Criterion), whereas the forecasting performance examined based on MSEP (Mean Square Error of Prediction) using two forecasting distributions of log death rates, the one obtained by Gibbs sampler and the one provided by Kalman Filter.

9.5.1. Model Selection

The device information criterion is a popular measure in Bayesian setting which trades off model fit against its complexity (the effective of parameters) as introduced in (Spiegelhalter et al. 2002). Among the various versions of DIC we decided to use so-called conditional DIC which treats the

latent states as parameters when calculating the conditional loglikelihood, for details please refer to (Celeux et al. 2006). The conditional loglikelihood if given by

$$\log \mathcal{L}_{\mathbf{y}_t|\boldsymbol{\varphi}_t,\boldsymbol{\psi}}\left(\boldsymbol{\psi},\boldsymbol{\varphi}_{1:N};\mathbf{y}_{1:N}\right) = -\frac{1}{2}\sum_{x=x_1}^{x_p}\sum_{t=1}^{N}\left(\log 2\pi\sigma_{\epsilon}^2 + \frac{\left(y_{x,t}-\alpha_x-[\tilde{\boldsymbol{B}}]_{x,\cdot}\boldsymbol{\varphi}_t\right)^2}{\sigma_{\epsilon}^2}\right) \quad (98)$$

By denoting $\boldsymbol{\psi} = (\boldsymbol{\varphi}_t, \boldsymbol{\psi})$ we define the deviance of the model as follows

$$D(\boldsymbol{\psi}) := -2\log \mathcal{L}_{\mathbf{y}_t|\boldsymbol{\psi}}\left(\boldsymbol{\psi};\mathbf{y}_{1:N}\right). \quad (99)$$

The function h as it is independent of the model specifications, is usually considered to be equal to 1. Hence, the effective number of parameters is defined as

$$p_D := \bar{D}(\boldsymbol{\psi}) - D(\bar{\boldsymbol{\psi}}) \quad (100)$$

where $\bar{D}(\boldsymbol{\psi})$ is a mean of deviance over different samples of the vector $\boldsymbol{\psi}$ and $D(\bar{\boldsymbol{\psi}})$ is a deviance of the posteriori mean of the vector of parameters $bm\psi$. The DIC then is defined as

$$DIC := \bar{D}(\boldsymbol{\psi}) + p_D = 2\bar{D}(\boldsymbol{\psi}) - D(\bar{\boldsymbol{\psi}}) \quad (101)$$

which can be calculated using the MCMC samples.

In addition to DIC, we calculate the mean square errors (MSE) for considered models, defined as the mean of the difference between the observed data, $\mathbf{y}_t$, and the mean of the in-sample one-step ahead model forecast given by Kalman Filter, $\mathbf{f}_t = \mathbb{E}[\mathbf{y}_t|\boldsymbol{\psi},\mathbf{y}_{1:(t-1)}]$ given in Equation (A3b). Therefore, we define $\mathbf{e}_t := \mathbf{y}_t - \mathbf{f}_t$ and

$$\text{MSE}(\bar{\boldsymbol{\psi}}) := \mathbb{E}\left[\mathbf{e}_t\mathbf{e}_t^T\right] \quad (102)$$

where for the point estimator of the vector of static parameters $\boldsymbol{\psi}$ we use the vector of posterior means.

9.5.2. Forcasting Distribution

The Bayesian state-space framework allows us to obtain the forecasting distributions using MCMC samples given by

$$\pi\left(\mathbf{y}_{T+m}|\mathbf{y}_{1:T}\right) = \int \pi\left(\mathbf{y}_{T+m}|\boldsymbol{\varphi}_{T+m},\boldsymbol{\psi}\right)\pi\left(\boldsymbol{\varphi}_{T+m}|\boldsymbol{\varphi}_{T+m-1},\boldsymbol{\psi}\right)...\pi\left(\boldsymbol{\varphi}_T,\boldsymbol{\psi}|\mathbf{y}_{1:T}\right)d\boldsymbol{\psi}d\boldsymbol{\varphi}_{T:T+m} \quad (103)$$

where $\boldsymbol{\psi}$ is the static parameter vector and $\boldsymbol{\varphi}_t$ is a state process vector. Following [104 lcstatespace], by sampling recursively, we obtain the following forecasting distributions, when (i) denotes a sample

$$\begin{aligned}\boldsymbol{\varphi}_{T+k}^{(i)} &\sim \mathcal{N}\left(\tilde{\boldsymbol{\Lambda}}^{(i)}\boldsymbol{\varphi}_{T+k-1}^{(i)} + \tilde{\boldsymbol{\Theta}}^{(i)}, \tilde{\boldsymbol{\Psi}}^{(i)}\right) \\ \mathbf{y}_{T+k}^{(i)} &\sim \mathcal{N}\left(\boldsymbol{\alpha}^{(i)} + \tilde{\mathbf{B}}_t^{(i)}\boldsymbol{\varphi}_{T+k-1}^{(i)} + \tilde{\boldsymbol{\Theta}}^{(i)}, \sigma_{\epsilon}^{2(i)}\mathbb{I}_d\right).\end{aligned} \quad (104)$$

Alternatively, we can use the forecasting distribution given by the Kalman Filter, that is

$$\begin{aligned}\boldsymbol{\varphi}_{T+k} &\sim \mathcal{N}\left(\tilde{\boldsymbol{\Lambda}}\boldsymbol{\varphi}_{T+k-1} + \tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}\right) \\ \mathbf{y}_{T+k} &\sim \mathcal{N}\left(\boldsymbol{\alpha} + \tilde{\mathbf{B}}_t\boldsymbol{\varphi}_{T+k-1} + \tilde{\boldsymbol{\Theta}}, \sigma_{\epsilon}^2\mathbb{I}_d\right).\end{aligned} \quad (105)$$

for the static parameters which has been estimated by averages within sampled realisation provided by the Gibbs sampler. Let us define mean square error of prediction function as follow

$$\text{MSEP}(\boldsymbol{\psi}) := \mathbb{E}\left[\left(\mathbf{y}_{T+k} - \mathbb{E}\left[\mathbf{y}_{T+k} | \mathbf{y}_{1:T}, \boldsymbol{\psi} \right] \right) \left(\mathbf{y}_{T+k} - \mathbb{E}\left[\mathbf{y}_{T+k} | \mathbf{y}_{1:T}, \boldsymbol{\psi} \right] \right)^{T} \right]. \tag{106}$$

Therefore the mean square error of prediction using MCMC distribution is calculated as a mean of $\text{MSEP}(\boldsymbol{\psi})$ over different samples of the vector $\boldsymbol{\psi}$ and denoted by MSEP_{MCMC}. The mean square error of prediction using the distribution provided by Kalman Filter is calculated the posterior mean of the vector of parameters $\boldsymbol{\psi}$.

9.5.3. Comparison of the Models

We choose for out-of-sample study last 10 years of the available sample for British Female death rates. The calibration period is 1922–2002. The Table 4 summarizes the calculated mean squared errors of the estimated observations using Kalman Filter (MSE), deviance information criterion (DIC) and mean square errors of predictions using the MCMC distribution (MSEP_{MCMC}) and the Kalman Filter distribution (MSEP_{Kalman}). The results confirm that adding the features, which has been extracted from demographic data, as an additional explanatory variable to the LCC model improves both in-sample fit out-of-sample fit and therefore the predictability of log death rates. The plots with age group specific prediction results can be found in Figure 37.

Table 4. Mean square error of the fit of the models to the data (MSE), deviance information criterion (DIC) and mean square errors of predictions using forecasting distributions given by MCMC samples (MSEP_{MCMC}) and Kalman Filter (MSEP_{Kalman}).

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

For the in-sample performance, the MSE and DIC agree to the group of two best performing models, DFM-PC-B and DFM-PC-Mx-r, however are conflicted with regards to the group of two worse performing models. Due to assessing the performance of the model considering its complexity, DIC more successfully captures the models which result in poorest out-of-sample performance, LCC and DFM-PC-Ms respectively. Especially it is worth to notice significant over-fitting of the LCC model which is further investigated in Section 10. Recall, that using MSE as model selection criterion would not be sufficient to choose the model with good performance. In terms of MSE, the in-sample performance of LCC model is comparable to DFM-PC-B model, while DIC labels the model as one with the worse explanatory power.

10. Additional Remarks on Modelling and Forecasting Results

While conducting the study we encountered two issues which are worth separate discussion, the influence of the stratification on the class of Lee-Carter models and the intuition behind the vector $\boldsymbol{\alpha}$ in the model Equation (8) when we incorporate the demographic features.

10.1. The Affect of the Stratification and Identification Constraints on Estimation of Stochastic Lee-Carter Type Models

We draw the readers attention at this point to the substantially lower predictive accuracy of the Lee-Carter cohort (LCC) model in comparison to the models which employ demographic factors,

as shown in Table 4. This is especially important since such LCC models, without factors have been previously documented to have better out-of-sample performance for UK data when no age stratification is applied.

We explain the steps we have taken to explore this feature that the reviewer has pointed out. Firstly, we clarify this is not a problem with the sampler or the prior specification, rather it is related to the particular suitability of different model structures under particular assumptions in the model. We explore this, specifically with respect to age group stratification and its influence on the model fit and performance.

Please note that the stratification was adopted, where we looked at 20 sets of age groups in 5 year buckets to reduce the dimensionality of the model, of course this can influence the model fit and the assumptions made regarding model simplification and cohort interpretation. Therefore, we decided to include additional studies to investigate these effects more carefully and in the process we believe we may also address the question raised by the reviewer on this point. Our attention has been drawn to two points: rapid increase of κ_t in 2000 and the substantially lower predictive accuracy of Lee-Carter cohort (LCC) model in comparison to the models which employ demographic factors as shown in Table 3. Since the LCC model has been documented to has better out-of-sample explanatory power for the UK data when there is no stratification, we decided to undertake an additional investigation. In particular, in this new class of studies, we explore in more detail the effect of age stratification and the appropriate choice of adjustment of model assumptions and identification constraints to be performed in order to compare models in a meaningful manner. The details are described below and in the manuscript in Section 10. The notation of the models which we decided to analyse are the following

LC: Lee-Carter model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t,$$

with the constraints

$$\sum_x \beta_x = 1, \ \sum_t \kappa_t = 0,$$

LCC: Simplified Lee-Carter cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = 1 \ , \ \sum_t \kappa_t = 0, \ \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

LCCF: Lee-Carter full cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \beta_x^\gamma \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = 1, \ \sum_x \beta_x^\gamma = 1, \ \sum_t \kappa_t = 0, \ \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

In this study we consider two age group stratifications, the 1×1 study which has 100 age groups per year and the more parsimonious class of models given by the 5×1 age group stratification with 21 age groups. Since the number of ages groups differs among stratified and non-stratified mortality data, that is, we have 21 age groups for the data in "5×1" format, we expect that the parameters and latent variables, which are estimated using above constraints, may differ in magnitude. Therefore, in order to ensure comparability of the results when we examine the stratification effect on the family of Lee-Carter models, one must therefore to standardize the magnitude of parameters and variables of models in order to compare between the stratified and non-stratified case. We demonstrate in the

new studies that such a problem may be resolved via a simple scaling adjustment to the identification constraints in order to resolve this issue. Hence, we introduce the scaling parameter $a > 0$. The models with imposed adjustment are denoted by lower index *adj* as follows

LC$_{adj}$: Lee-Carter model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t,$$

with the constraints

$$\sum_x \beta_x = \frac{1}{a}, \ \sum_t \kappa_t = 0,$$

LCC$_{adj}$: Simplified Lee-Carter cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = \frac{1}{a}, \ \sum_t \kappa_t = 0, \ \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

LCCF$_{adj}$: Lee-Carter full cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \beta_x^\gamma \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = \frac{1}{a}, \ \sum_x \beta_x^\gamma = \frac{1}{a}, \ \sum_t \kappa_t = 0, \ \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

Lastly, in order to distinguish between the results of the models for stratified and non-stratified data, we denote the models for stratified data with the lower index "5 × 1" and for non-stratified data with "1 × 1", for instance, the results for the Lee-Carter model LC for stratified data are denoted by $LC_{5\times 1}$ and the results for the same model for a data without any stratification are denoted by $LC_{1\times 1}$.

10.1.1. The Estimates of the Static Parameters and Filtered Latent Variables

The list of models, which have been examined and are discussed in this subsection, is given in the first column of Table 5. In addition to LC, LCC and LCCF models, we include into the comparison also the models with adjusted constraints for the mortality data in format "5 × 1" with the adjustment parameter $a = \frac{\# \text{ age groups 1x1}}{\# \text{ age groups 5x1}} = \frac{100}{21} \approx 4.762$ being a proportion between the number of age groups in in the 1×1 stratification and the 5×5 stratification of the age groups in the mortality data.

The Bayesian posterior estimators of the static parameters α_x and β across age groups are shown in Figure 38 when the Bayesian posterior mean estimates of period effect κ_t and cohort effect state process γ_t^0 are shown in Figure 39.

The first straightforward remark on the investigation is to note that there is an inconsequential influence of both the stratification and the adjustment to the identification constraint for stratification, when investigating the estimation of the level vector of the model, as denoted by parameter vector α.

Also, the basic Lee-Carter models is not affected by the stratification as both its in-sample and out-of-sample quality of fit is comparable among data with format "5 × 1" and "1 × 1". However, the estimates of $\boldsymbol{\beta}$ for $LC_{5\times 1}$ are greater in the magnitude in comparison to $LC_{1\times 1}$ model which appears to be an offset by a smaller slope of the filtered κ_t. Importantly, we note that when the adjustment to the number of age groups is imposed, the $\boldsymbol{\beta}$ and κ_t for $LC_{5\times 1,adj}$ are in line to those of $LC_{1\times 1}$ and the model still keeps comparable explanatory power. This observation gives us the intuition that the stratification influences mainly the cohort effect. This is something that we intuitively can understand due to the interplay between age stratification and cohort effect.

Table 5. Mean square error of the fit of the models to the data (MSE) and mean square errors of predictions using forecasting distributions given by MCMC samples (MSEP$_{MCMC}$) and Kalman Filter (MSEP$_{Kalman}$). The models highlated by bold font exhibit significant levels of over-fitting.

Model	MSE	MSEP$_{MCMC}$	MSEP$_{Kalman}$
LC$_{1\times1}$	0.0128	0.0577	0.0568
LC$_{5\times1}$	0.0113	0.0457	0.0457
LC$_{5\times1,adj}$	0.0116	0.0512	0.0516
LCC$_{1\times1}$	0.0079	0.0249	0.0243
LCC$_{5\times1}$	**0.0097**	**0.1778**	**0.1774**
LCC$_{5\times1,adj}$	0.0099	0.1664	0.1625
LCCF$_{1\times1}$	**0.2588**	**0.4735**	**0.6150**
LCCF$_{5\times1}$	0.0107	0.0458	0.0464
LCCF$_{5\times1,adj}$	0.0131	0.0481	0.0508

The results in Table 5 shows that the out-of-sample quality of the fit for the LCC$_{5\times1}$ model is significantly lower than the corresponding result for LCC$_{1\times1}$. As the adjustment, that is LCC$_{5\times1,adj}$ model, produces similar outcomes, the discrepancy in the out-of-sample quality of fit between LCC model applied to "5 × 1" and "1 × 1" data is not caused by the smaller number of age groups and therefore different magnitude of estimates of β and κ_t (recall Figures 38 and 39). Since the discrepancy of the in-sample explanatory power is smaller between the models, we observe that the LCC model tends to over-fit when applied to stratified data.

We begin discussion of these results by noting the following finding. There appears to be an interplay present between the model parsimony and the bias and variance in the results for both in-sample fits and out-of-sample forecasts, as reflected by the Mean Squared Error (MSE) results, which is more largely affected by the model structure rather than stratification effects.

For instance, we see that the more parsimonious model choices, corresponding to say the three LC sub-family of models always had a larger MSE than the less parsimonious class of simplified LCC model. That is the in-sample MSE improved by around an order of magnitude when we incorporated extra structure corresponding to the cohort feature. This was not influenced by the age stratification reformed. We conjecture that although the LC models will have potentially lower variance, due to less model parameters to be estimated, the in-sample MSE is still worse generally due to increased bias that may arise from not capturing sufficiently the stochastic structure of the data.

Furthermore, we also see a pronounced effect of stratification on the out-of-sample forecast performance of the simplified class of LCC models in which no adjustment was made for the stratification effect. This indicates that the adjustment we propose to use when undertaking age-group stratification can substantially reduce the bias in the resulting model estimates when we compare between the simplified LCC$_{5\times1}$ model and the adjusted form.

Thirdly, we observe that the most flexible class of LCC model, the non-simplified LCCF class of models was significantly affected by removing the age stratification of 5 × 1 compared to the 1 × 1 case. To understand this, we have significantly increased the dimension of the model parameters to be estimated in the LCCF$_{1\times1}$ compared to the LCCF$_{5\times1}$. This we believe produces a poor in sample and out-of-sample MSE and MSEP due to the resulting over-fitting and increased variance in the model estimates, compared to the simplified LCC model equivalents. However, importantly the stratification effect is significant here, the dimension reduction in model parameters in the LCCF$_{5\times1}$ compared to the LCCF$_{1\times1}$ reduces the variance in the estimates of the mortality in sample and out-of-sample as well as providing additional degrees of freedom to also reduce the bias that arises from the constrained version of the LCC$_{5\times1}$ model, resulting in the optimal MSE and MSEP performance.
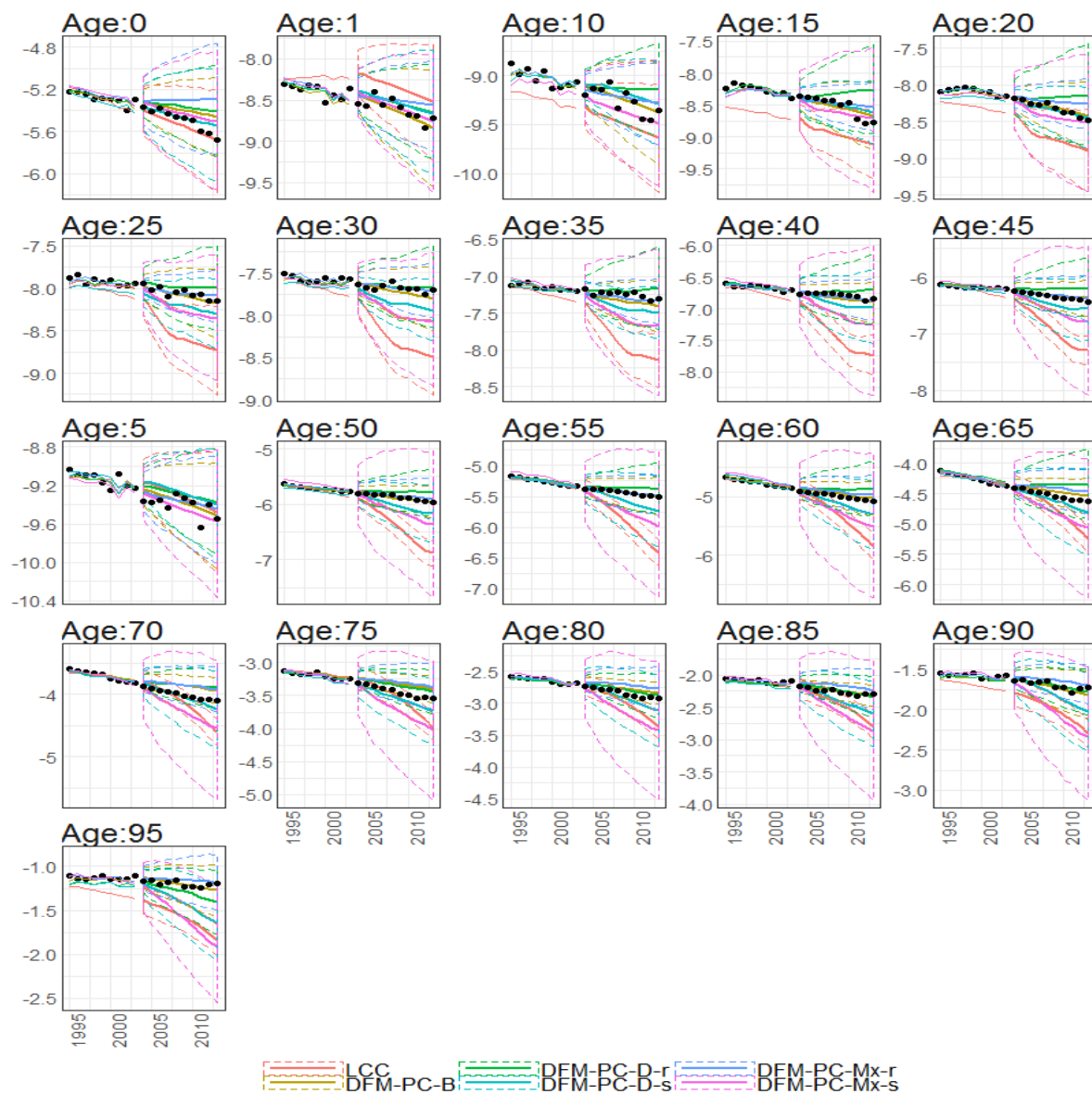
Figure 37. 10-year out-of-sample forecasted log death (y axis) rates of different age groups (different subplots) under different models (colours of lines) with corresponding prediction intervals. Calibration period: 1922–2002
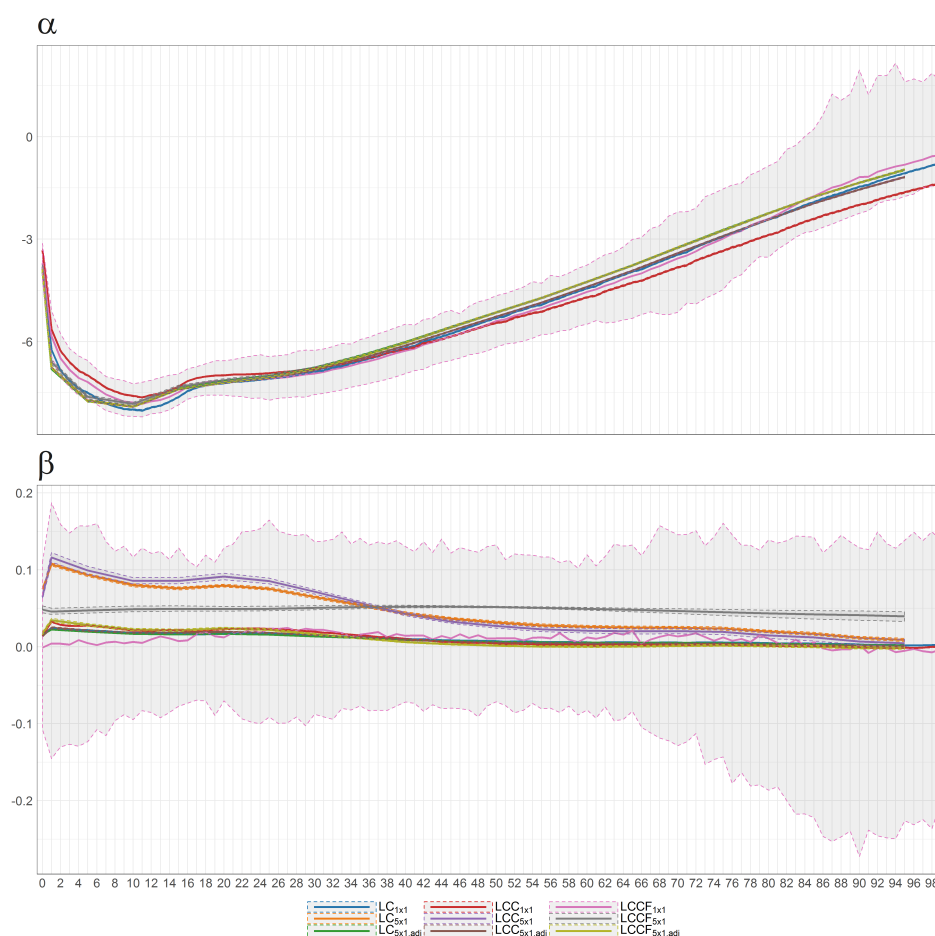
Figure 38. Bayesian posterior estimators with 95% posterior credible intervals for the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under different models (colours of lines) for British female mortality data (1922–2002).

10.2. The Estimates of the Intercept for the Factor Model DFM-PC-D-r

The following study addresses the interpretation of the estimates of $\boldsymbol{\alpha}$ under the the new class of stochastic mortality factor model in comparison to the standard Lee-Carter models without the matrix with demographic factors. Please refer to Section 10.2. The argument we proposed to interpret $\boldsymbol{\alpha}$ is adjusted to the fact that we have the exogenous factors incorporated into our model compared to standard Lee-Carter model. As such, we argue that the the interpretation of the intercept should now incorporate both $\boldsymbol{\alpha}$, the classical intercept, and the term $\tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ corresponding to the intercept which arises from the exogenous factors. In the time series context this is considered as a stochastic intercept. Hence the interpretation of $\boldsymbol{\alpha}$ typically adopted in the classical stochastic Lee-Carter type period-cohort models, does not hold under the new model, since we have now incorporated the additional structure corresponding to the regression term from the demographic factors. The expression $\tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ which is added to the observation equation provides with time-varying supplementary information to the static level given by $\boldsymbol{\alpha}$. To validate this claim we have undertaken the additional studies which demonstrate when we combine the $\boldsymbol{\alpha}$ with this component of the model, the posterior mean of this quantity behaves in analogues fashion to what you would expect on the posterior mean of $\boldsymbol{\alpha}$ in the standard, non-factor class of Lee-Carter models. This is interesting as it shows the factor influence and additional interpretation to the level contributed by the long-term demographic exogenous factors. The plot in Figure 40 shows the Bayesian posterior mean estimates with 95% creditable intervals of $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t^T$ averaged over time, whereas Figure 41 illustrates the posteriori mean over time. The level of

the expression on either of plots is below zero and behaves in a fashion we would have expected from α in standard Lee-Carter models. It confirms our interpretation as well as answers the question asked by the reviewer.
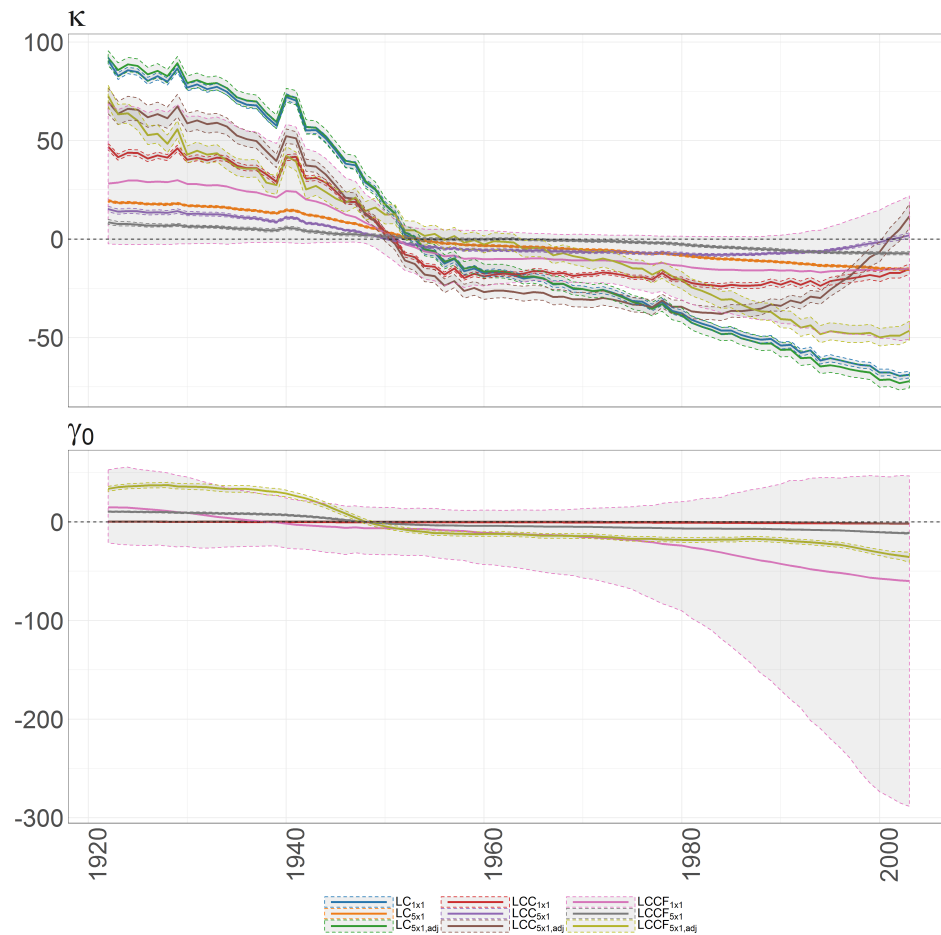

Figure 39. The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (**upper panel**) and cohort effect state process γ_t^0 (**lower panel**) under different models (colours of lines) for British female log death rates during 1922–2002.
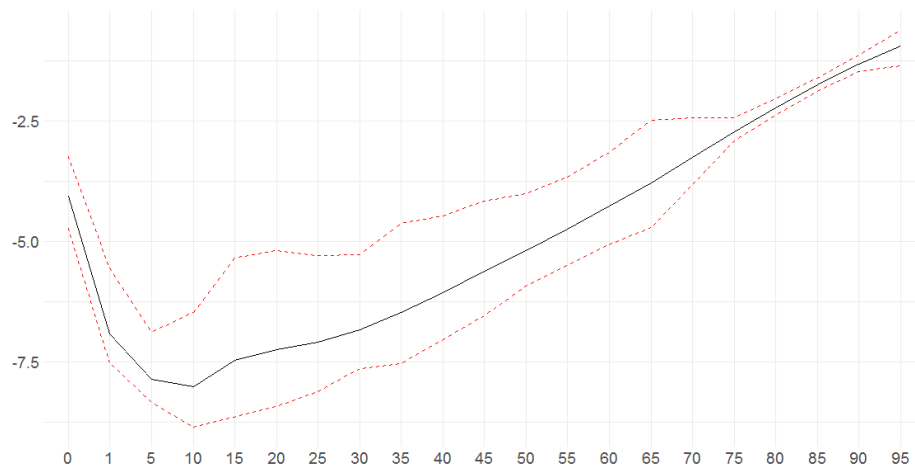
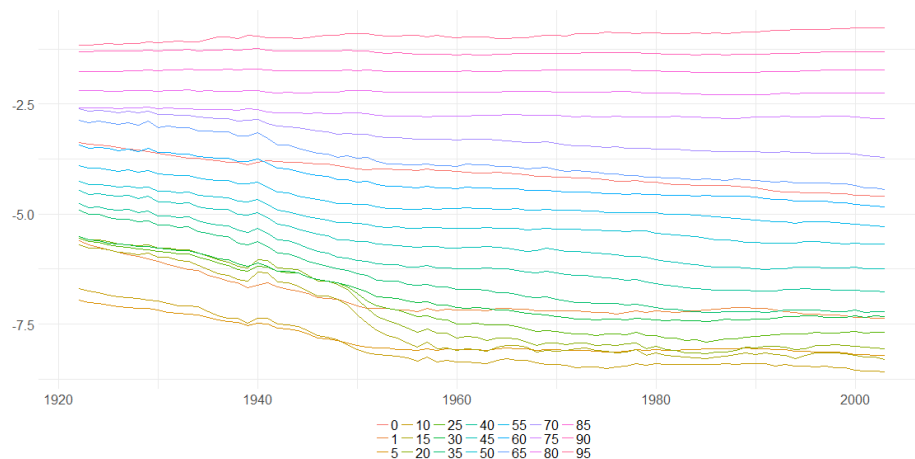**Figure 40.** The Bayesian posterior mean estimates with 95% posterior credible intervals average over time for $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ for DFM-PC-D-r.

Figure 41. The Bayesian posterior mean estimates of $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t^T$ over time for DFM-PC-D-r. Colours of lines are related to the age groups (the elements of the vector $\boldsymbol{\alpha}$).

11. Conclusions

We developed and presented a comprehensive study which focuses on the analysis and the incorporation of demographic data into state-space framework for stochastic mortality modelling. We have extended the well-known Lee-Carter model with stochastic cohort effect by introducing new state processes which correspond to the age-specific dynamic of European female log death rates.

We showed by means of Probabilistic Principal Component Analysis the ideas of extracting the meaningful features from demographic data of European countries and applying them as explanatory variables to the mortality estimation and forecasting. In the presence of short time series and different types of missingness, the suggested methodology aims to be as parsimonious as possible. By analysing of the extracted features, we found more evidences about region specific mortality structures. Also, the features exhibit significant sensitivity to the methodology of estimation of moments. As overviewed in Section 8, the robust alternatives to the sample estimators provide with more consistent results regardless to the number of missing entries, especially if the data has been not preprocessed or smoothed.

The results of applying studied models to the British female log mortality data showed that incorporating the features extracted from European demographic data provide valuable information about the mortality forces which affect British population. Also, the models with dynamic factors exhibit better in-sample and out-of-sample fit than the Lee-Carter model with the cohort effect.

As an additional outcome of the study, we analyse the effect of the stratification of the data on the family of Lee-Carter Model. In Section 10 we argue that the stratification influences mainly the cohort effect process and requires more flexibility of modelling than provided by simplified Lee-Carter cohort model. The investigation showed that the standard Lee-Carter model has similar performance for stratified and none-stratified data, whereas simplified Lee-Carter cohort model is prone to significant overfitting when fitted to stratified data. Also, the investigation shows that stratification helps to resolve the issues of overfitting related to Lee-Carter full cohort due to smaller dimensionality of the observation vector.

There are a few ways in which the paper can be further extended. First of all, with appropriate data it would be straightforward to conduct similar analysis for extended data set, as for gender and region specific disaggregation of mortality and demographic data with the inclusion of various population related factors such as cause-of-death Murray and Lopez (1997); (Girosi and King 2008) or more recently (Gaille and Sherris 2015), midlife conditions as in Gavrilov and Gavrilova (2015) or migration. Secondly, considering different distribution priori assumptions and introducing the methodology to handle more advanced patterns of missigness would benefit in better explanatory power and even more consistent interpretability. We can also improve the framework of feature extraction to account for fat tail distributions by means of different robustification methodologies or extensions to Principal Component analysis such as Independent Component Analysis and their functional alternatives as in Shang and Hyndman (2016).

Author Contributions: Dorota Toczydlowska was the main author of the manuscript and the methodological development and implementations. The co-authors Gareth W. Peters , Man Chung Fung and Pavel V. Shevchenko contributed to aspects of the methodological developments, the derivations, implementations and data analysis.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Bayesian Modelling and Sampling of Demographic Factor Model Extension to the Period-Cohort Stochastic Mortality State-Space Models

In this appendix section we explain the Bayesian models developed for the estimation of the demographic factor model extension to the period-cohort stochastic mortality state-space models that are applied in this paper. These are based on the frameworks detailed and developed in Fung et al. (2016) and Fung et al. (2017).

Appendix A.1. Bayesian Model Development and Inference for Stochastic Mortality Models

The observation and state equations Equations (5a) and (5b) imply that the cohort model that we have formulated here belongs to the linear-Gaussian class of state-space models. As a result one can perform efficient maximum-likelihood or Bayesian estimation on fitting the model to data as discussed in detail in Fung et al. (2016). In this paper we focus on Bayesian inference so that the forecasting distribution can take into account parameter uncertainty.

To achieve these inference goals we must first develop the Bayesian models. In this section we detail the Bayesian estimation of the cohort model Equations (5a) and (5b) and its extensions incorporating population information to state-space formulation Equations (8a) and (8b). We firstly note that the models belong to the linear and Gaussian class of state-space models. As a result one can apply an efficient MCMC estimation algorithm based on Gibbs sampling with conjugate priors combined with forward-backward filtering as described in Fung et al. (2016). We describe procedures

using the notation of general cohort model Equations (5a) and (5b) and indicate the differences to considered extensions.

We borrow the notation from the LC cohort model Equations (5a) and (5b) and indicate similarities and differences between Equations (5a) and (5b) and Equations (8a) and (8b) while developing the estimation algorithm. In the general setting, our target density is

$$\pi\left(\boldsymbol{\varphi}_{0:T}, \boldsymbol{\psi} | \boldsymbol{y}_{1:T}\right) \tag{A1}$$

where $\boldsymbol{\varphi}_{0:T}$ is a vector of latent variables and $\boldsymbol{\psi}$ is a vector of static parameters. In the case of the cohort LC model $\boldsymbol{\varphi}_{0:T} := (\kappa_{0:T}, \gamma_{0:T}^{x_1}, \dots, \gamma_{0:T}^{x_p})$ is the $p+1$ dimensional (for each t) latent state vector and $\boldsymbol{\psi} := (\alpha_{x_3:x_p}, \beta_{x_2:x_p}, \theta, \eta, \lambda, \sigma_\varepsilon^2, \sigma_\kappa^2, \sigma_\gamma^2)$ is the $2p-1$ dimensional static parameter vector. For extended models, we add vectors of global factors latent variable $\boldsymbol{\varrho}_t$ and its static model parameters. Recall that our proposed identification constraint is given by Equation (6); therefore only $\alpha_{x_3:x_p}$ and $\beta_{x_2:x_p}$ are required to be estimated. We perform block sampling for the latent state via the so-called forward-filtering-backward-sampling (FFBS) algorithm (Carter and Kohn 1994) and the posterior samples of the static parameters are obtained via conjugate priors. The sampling procedure is described in Algorithm 2, where N is the number of MCMC iterations performed.

Algorithm 2 MCMC sampling for $\pi(\boldsymbol{\varphi}_{0:T}, \boldsymbol{\psi}|\boldsymbol{y}_{1:T})$

1: Initialise: $\boldsymbol{\psi} = \boldsymbol{\psi}^{(0)}$.

2: **for** $i = 1, \dots, N$ **do**

3:　　Sample $\boldsymbol{\varphi}_{0:T}^{(i)}$ from $\pi(\boldsymbol{\varphi}_{0:T}|\boldsymbol{\psi}^{(i-1)}, \boldsymbol{y}_{1:T})$ via FFBS (Appendix A.2).

4:　　**for** $h = 1, \dots, 2p-1$ **do**

5:　　　　Sample $\psi_h^{(i)}$ from $\pi(\psi_h|\boldsymbol{\varphi}_{0:T}^{(i)}, \boldsymbol{\psi}_{-h}^{(i)}, \boldsymbol{y}_{1:T})$, (Appendix A.3)

6:　　　　where $\boldsymbol{\psi}_{-h}^{(i)} = (\psi_1^{(i)}, \dots, \psi_{h-1}^{(i)}, \psi_{h+1}^{(i-1)}, \dots, \psi_{2p-2}^{(i-1)})$.

7:　　**end for**

8: **end for**

Appendix A.2. Forward-Backward Filtering for Latent State Dynamics

The FFBS procedure requires to carry out multivariate Kalman filtering forward in time and then sample backwardly using the obtained filtering distributions. For the cohort model Equations (2) and (3) (or Equations (5a) and (5b)), the conditional distributions involved in the multivariate Kalman filtering recursions are given by

$$\boldsymbol{\varphi}_{t-1}|\boldsymbol{y}_{1:t-1} \sim \mathcal{N}(\boldsymbol{m}_{t-1}, C_{t-1}), \tag{A2a}$$

$$\boldsymbol{\varphi}_t|\boldsymbol{y}_{1:t-1} \sim \mathcal{N}(\boldsymbol{a}_t, R_t), \tag{A2b}$$

$$\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1} \sim \mathcal{N}(\boldsymbol{f}_t, Q_t), \tag{A2c}$$

$$\boldsymbol{\varphi}_t|\boldsymbol{y}_{1:t} \sim \mathcal{N}(\boldsymbol{m}_t, C_t) \tag{A2d}$$

where

$$\boldsymbol{a}_t = \boldsymbol{\Lambda}\boldsymbol{m}_{t-1} + \boldsymbol{\Theta}, \quad R_t = \boldsymbol{\Lambda}C_{t-1}\boldsymbol{\Lambda}^\top + Y, \tag{A3a}$$

$$\boldsymbol{f}_t = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{a}_t, \quad Q_t = \mathbf{B}R_t\mathbf{B}^\top + \sigma_\varepsilon^2\mathbb{I}_p, \tag{A3b}$$

$$\boldsymbol{m}_t = \boldsymbol{a}_t + R_t\mathbf{B}^\top Q_t^{-1}(\boldsymbol{y}_t - \boldsymbol{f}_t), \quad C_t = R_t - R_t\mathbf{B}^\top Q_t^{-1}\mathbf{B}R_t. \tag{A3c}$$

for $t = 1, \ldots, T$. Since

$$\pi(\boldsymbol{\varphi}_{0:T}|\boldsymbol{\psi}, \boldsymbol{y}_{1:T}) = \prod_{t=0}^{T} \pi(\boldsymbol{\varphi}_t|\boldsymbol{\varphi}_{t+1:T}, \boldsymbol{\psi}, \boldsymbol{y}_{1:T}) = \prod_{t=0}^{T} \pi(\boldsymbol{\varphi}_t|\boldsymbol{\varphi}_{t+1}, \boldsymbol{\psi}, \boldsymbol{y}_{1:t}), \tag{A4}$$

We see that for a block sampling of the latent state, one can first draw $\boldsymbol{\varphi}_T$ from $N(\boldsymbol{m}_T, C_T)$ and then, for $t = T - 1, \ldots, 1, 0$ (that is backward in time), draws a sample of $\boldsymbol{\varphi}_t|_{\boldsymbol{\varphi}_{t+1}, \boldsymbol{\psi}, \boldsymbol{y}_{1:T}}$ recursively given a sample of $\boldsymbol{\varphi}_{t+1}$. It turns out that $\boldsymbol{\varphi}_t|_{\boldsymbol{\varphi}_{t+1}, \boldsymbol{\psi}, \boldsymbol{y}_{1:T}} \sim \mathcal{N}(\boldsymbol{h}_t, H_t)$ where

$$\boldsymbol{h}_t = \boldsymbol{m}_t + C_t \boldsymbol{\Lambda}^\top R_{t+1}^{-1}(\boldsymbol{\varphi}_{t+1} - \boldsymbol{a}_{t+1}), \tag{A5a}$$

$$H_t = C_t - C_t \boldsymbol{\Lambda}^\top R_{t+1}^{-1} \boldsymbol{\Lambda} C_t, \tag{A5b}$$

based on Kalman smoothing (Carter and Kohn 1994). **For the extended models, we simply need to replace the vectors and matrices of the LC cohort model with objects stressed by tildes from Section 3.**

Appendix A.3. Posteriors for Static Parameters in the Cohort Model

To sample the posterior distribution of the static parameters, we assume the following independent conjugate priors:

$$\alpha_x \sim \mathcal{N}(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2), \quad \beta_x \sim \mathcal{N}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2), \quad \theta \sim \mathcal{N}(\tilde{\mu}_\theta, \tilde{\sigma}_\theta^2), \quad \eta \sim \mathcal{N}(\tilde{\mu}_\eta, \tilde{\sigma}_\eta^2) \tag{A6a}$$

$$\lambda \sim \mathcal{N}_{[-1,1]}(\tilde{\mu}_\lambda, \tilde{\sigma}_\lambda^2), \quad \sigma_\varepsilon^2 \sim \mathrm{IG}(\tilde{a}_\varepsilon, \tilde{b}_\varepsilon), \quad \sigma_\kappa^2 \sim \mathrm{IG}(\tilde{a}_\kappa, \tilde{b}_\kappa), \quad \sigma_\gamma^2 \sim \mathrm{IG}(\tilde{a}_\gamma, \tilde{b}_\gamma) \tag{A6b}$$

where $\mathrm{N}_{[-1,1]}$ denotes a truncated Gaussian with support $[-1, 1]$ and $\mathrm{IG}(\tilde{a}, \tilde{b})$ denotes an inverse-gamma distribution with mean $\tilde{b}/(\tilde{a} - 1)$ and variance $\tilde{b}^2/((\tilde{a} - 1)^2(\tilde{a} - 2))$ for $\tilde{a} > 2$. The posteriors of the static parameters are then obtained as follows:[1]

$$\alpha_x|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\alpha_x} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\alpha^2 \sum_{t=1}^{T}(y_{x,t} - \beta_x \kappa_t - \gamma_t^x) + \tilde{\mu}_\alpha \sigma_\varepsilon^2}{\tilde{\sigma}_\alpha^2 T + \sigma_\varepsilon^2}, \frac{\tilde{\sigma}_\alpha^2 \sigma_\varepsilon^2}{\tilde{\sigma}_\alpha^2 T + \sigma_\varepsilon^2} \right), \tag{A7}$$

$$\beta_x|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\beta_x} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\beta^2 \sum_{t=1}^{T}(y_{x,t} - (\alpha_x + \gamma_t^x))\kappa_t + \tilde{\mu}_\beta \sigma_\varepsilon^2}{\tilde{\sigma}_\beta^2 \sum_{t=1}^{T} \kappa_t^2 + \sigma_\varepsilon^2}, \frac{\tilde{\sigma}_\beta^2 \sigma_\varepsilon^2}{\tilde{\sigma}_\beta^2 \sum_{t=1}^{T} \kappa_t^2 + \sigma_\varepsilon^2} \right), \tag{A8}$$

$$\theta|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\theta} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\theta^2 \sum_{t=1}^{T}(\kappa_t - \kappa_{t-1}) + \tilde{\mu}_\theta \sigma_\kappa^2}{\tilde{\sigma}_\theta^2 T + \sigma_\kappa^2}, \frac{\tilde{\sigma}_\theta^2 \sigma_\kappa^2}{\tilde{\sigma}_\theta^2 T + \sigma_\kappa^2} \right), \tag{A9}$$

$$\eta|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\theta} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\eta^2 \sum_{t=1}^{T}(\gamma_t - \lambda\gamma_{t-1}) + \tilde{\mu}_\eta \sigma_\gamma^2}{\tilde{\sigma}_\eta^2 T + \sigma_\gamma^2}, \frac{\tilde{\sigma}_\eta^2 \sigma_\gamma^2}{\tilde{\sigma}_\eta^2 T + \sigma_\gamma^2} \right), \tag{A10}$$

$$\lambda|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\lambda} \sim \mathcal{N}_{[-1,1]}\left(\frac{\tilde{\sigma}_\lambda^2 \sum_{t=1}^{T}((\gamma_t^{x_1} - \eta)\gamma_{t-1}^{x_1}) + \tilde{\mu}_\lambda \sigma_\gamma^2}{\tilde{\sigma}_\lambda^2 \sum_{t=1}^{T}(\gamma_{t-1}^{x_1})^2 + \sigma_\gamma^2}, \frac{\tilde{\sigma}_\lambda^2 \sigma_\gamma^2}{\tilde{\sigma}_\lambda^2 \sum_{t=1}^{T}(\gamma_{t-1}^{x_1})^2 + \sigma_\gamma^2} \right), \tag{A11}$$

$$\sigma_\varepsilon^2|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\sigma_\varepsilon^2} \sim \mathrm{IG}\left(\tilde{a}_\varepsilon + \frac{pT}{2}, \tilde{b}_\varepsilon + \frac{1}{2} \sum_{x=x_1}^{x_p} \sum_{t=1}^{T}(y_{x,t} - (\alpha_x + \beta_x\kappa_t + \gamma_t^x))^2 \right), \tag{A12}$$

$$\sigma_\kappa^2|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\sigma_\kappa^2} \sim \mathrm{IG}\left(\tilde{a}_\kappa + \frac{T}{2}, \tilde{b}_\kappa + \frac{1}{2} \sum_{t=1}^{T}(\kappa_t - (\kappa_{t-1} + \theta))^2 \right), \tag{A13}$$

$$\sigma_\gamma^2|\boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\sigma_\gamma^2} \sim \mathrm{IG}\left(\tilde{a}_\gamma + \frac{T}{2}, \tilde{b}_\gamma + \frac{1}{2} \sum_{t=1}^{T}(\gamma_t^{x_1} - \lambda\gamma_{t-1}^{x_1})^2 \right). \tag{A14}$$

---

[1] For simplicity, we denote $\boldsymbol{y} = \boldsymbol{y}_{1:T}$, $\boldsymbol{\varphi} = \boldsymbol{\varphi}_{0:T}$ and $\boldsymbol{\psi}_{-h} = (\psi_1, \ldots, \psi_{h-1}, \psi_{h+1}, \ldots, \psi_{2p-2})$.

Appendix A.4. Posteriors for Static Parameters in the Extended Models

In order to develope sampling algorithm for models incorporating European countries population information, we add the conjugate priori assumptions to Equation (A6) related to static parameters of Equation (7)

$$[\mathbf{\Omega}]_{i,j} \sim \mathcal{N}(\tilde{\mu}_{\Omega}, \tilde{\sigma}_{\Omega}^2), \quad \Psi_j \sim \mathcal{N}(\tilde{\mu}_{\Psi}, \tilde{\sigma}_{\Psi}^2), \quad \sigma_{\varrho}^2 \sim \mathrm{IG}(\tilde{a}_{\varrho}, \tilde{b}_{\varrho}) \tag{A15}$$

The posteriors of static parameters from Equation (7) are essential for Gibbs backward sampling regardless of the considered case for the extended model. We refer by small letters i, j to ages x_i, $x_j \in \{x_1, \ldots, x_p\}$, for $i, j \in \{1, \ldots, p\}$, and letters $m, l \in \{1, \ldots, k\}$ to the components of the matrix $\mathbf{F}_t$. An element labelled as im corresponds to $[\mathbf{F}_t]_{i,m}$ or a latent variable ϱ_t^{im}. Then

$$\Psi_{im} | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-\Psi_{im}} \sim \mathrm{N} \left(\frac{\tilde{\sigma}_{\Psi}^2 \sum_{t=1}^{T} (\varrho_t^{im} - \boldsymbol{\varrho}_{t-1}^{T} [\mathbf{\Omega}]_{im,\cdot}) + \tilde{\mu}_{\psi} \sigma_{\varrho}^2}{\tilde{\sigma}_{\Psi}^2 T + \sigma_{\varrho}^2}, \frac{\tilde{\sigma}_{\Psi}^2 \sigma_{\varrho}^2}{\tilde{\sigma}_{\Psi}^2 T + \sigma_{\varrho}^2} \right),$$

$$[\mathbf{\Omega}]_{im,jl} | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-[\mathbf{\Omega}]_{im,jl}} \sim$$
$$\mathcal{N} \left(\frac{\tilde{\sigma}_{\Omega}^2 \sum_{t=1}^{T} \left[\left(\varrho_t^{im} - \Psi_{im} - \sum_{xh \neq jm} \varrho_{t-1}^{xh} [\mathbf{\Omega}]_{im,xh} \right) \varrho_{t-1}^{jl} \right] + \tilde{\mu}_{\Omega} \sigma_{\varrho}^2}{\tilde{\sigma}_{\Omega}^2 \sum_{t=1}^{T} (\varrho_{t-1}^{jl})^2 + \sigma_{\varrho}^2}, \frac{\tilde{\sigma}_{\Omega}^2 \sigma_{\varrho}^2}{\tilde{\sigma}_{\Omega}^2 \sum_{t=1}^{T} (\varrho_{t-1}^{jl})^2 + \sigma_{\varrho}^2} \right), \tag{A16}$$

$$\sigma_{\varrho}^2 | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-\sigma_{\varrho}^2} \sim \mathrm{IG} \left(\tilde{a}_{\varrho} + \frac{pkT}{2}, \tilde{b}_{\varrho} + \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{p} \sum_{m=1}^{k} \left(\varrho_t^{im} - \Psi_{im} - \boldsymbol{\varrho}_{t-1}^{T} [\mathbf{\Omega}]_{im,\cdot} \right)^2 \right),$$

where $\tilde{\boldsymbol{\varphi}}$ is a vector of latent variables from Equation (8b) and $\boldsymbol{\psi}$ is a vector of static parameters updated to Equation (7). Depends on the cases f extended model, we have the following replacement of the posterioris from Appendix A.3

Case 1 Global factors $\mathbf{F}_t$ in the observation equation

$$\alpha_x | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-\alpha_x} \sim \mathcal{N} \left(\frac{\tilde{\sigma}_{\alpha}^2 \sum_{t=1}^{T} \left(y_{x,t} - [\tilde{\mathbf{B}}_t]_{x,\cdot}^{T} \tilde{\boldsymbol{\varphi}}_t \right) + \tilde{\mu}_{\alpha} \sigma_{\varepsilon}^2}{\tilde{\sigma}_{\alpha}^2 T + \sigma_{\varepsilon}^2}, \frac{\tilde{\sigma}_{\alpha}^2 \sigma_{\varepsilon}^2}{\tilde{\sigma}_{\alpha}^2 T + \sigma_{\varepsilon}^2} \right), \tag{A17}$$

$$\beta_x | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-\beta_x} \sim \mathcal{N} \left(\frac{\tilde{\sigma}_{\beta}^2 \sum_{t=1}^{T} \left[y_{x,t} - \left(\alpha_x + \gamma_t^x + [\tilde{\mathbf{F}}_t]_{x,\cdot}^{T} \boldsymbol{\varrho}_t \right) \right] \kappa_t + \tilde{\mu}_{\beta} \sigma_{\varepsilon}^2}{\tilde{\sigma}_{\beta}^2 \sum_{t=1}^{T} \kappa_t^2 + \sigma_{\varepsilon}^2}, \frac{\tilde{\sigma}_{\beta}^2 \sigma_{\varepsilon}^2}{\tilde{\sigma}_{\beta}^2 \sum_{t=1}^{T} \kappa_t^2 + \sigma_{\varepsilon}^2} \right) \tag{A18}$$

$$\sigma_{\varepsilon}^2 | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-\sigma_{\varepsilon}^2} \sim \mathrm{IG} \left(\tilde{a}_{\varepsilon} + \frac{pT}{2}, \tilde{b}_{\varepsilon} + \frac{1}{2} \sum_{x=x_1}^{x_p} \sum_{t=1}^{T} \left[y_{x,t} - \left(\alpha_x + [\tilde{\mathbf{B}}_t]_{x,\cdot}^{T} \tilde{\boldsymbol{\varphi}}_t \right) \right)^2 \right] \tag{A19}$$

Case 2 Global factors $\mathbf{F}_t$ in the state equation of κ_t

$$\theta | \boldsymbol{y}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{\psi}_{-\theta} \sim \mathcal{N} \left(\frac{\tilde{\sigma}_{\theta}^2 \sum_{t=1}^{T} (\kappa_t - \kappa_{t-1} - \tilde{\mathbf{f}}_t^{T} \boldsymbol{\varrho}_t) + \tilde{\mu}_{\theta} \sigma_{\kappa}^2}{\tilde{\sigma}_{\theta}^2 T + \sigma_{\kappa}^2}, \frac{\tilde{\sigma}_{\theta}^2 \sigma_{\kappa}^2}{\tilde{\sigma}_{\theta}^2 T + \sigma_{\kappa}^2} \right), \tag{A20}$$

$$\sigma_{\kappa}^2 | \boldsymbol{y}, \boldsymbol{\varphi}, \boldsymbol{\psi}_{-\sigma_{\kappa}^2} \sim \mathrm{IG} \left(\tilde{a}_{\kappa} + \frac{T}{2}, \tilde{b}_{\kappa} + \frac{1}{2} \sum_{t=1}^{T} \left[\kappa_t - \left(\kappa_{t-1} + \theta + \tilde{\mathbf{f}}_t^{T} \boldsymbol{\varrho}_t \right) \right]^2 \right) \tag{A21}$$

Case 3 Global factors $\mathbf{F}_t$ in the state equation of γ_t

$$\eta|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\theta} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\eta^2 \sum_{t=1}^T \left(\gamma_t^{x_1} - \lambda\gamma_{t-1}^{x_1} - \boldsymbol{\varrho}_t^T[\tilde{\mathbf{F}}_t]_{x_1,\cdot}\right) + \tilde{\mu}_\eta \sigma_\gamma^2}{\tilde{\sigma}_\eta^2 T + \sigma_\gamma^2}, \frac{\tilde{\sigma}_\eta^2 \sigma_\gamma^2}{\tilde{\sigma}_\eta^2 T + \sigma_\gamma^2}\right), \tag{A22}$$

$$\lambda|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\lambda} \sim \mathcal{N}_{[-1,1]}\left(\frac{\tilde{\sigma}_\lambda^2 \sum_{t=1}^T ((\gamma_t^{x_1} - \eta - \boldsymbol{\varrho}_t^T[\tilde{\mathbf{F}}_t]_{x_1,\cdot})\gamma_{t-1}^{x_1}) + \tilde{\mu}_\lambda \sigma_\gamma^2}{\tilde{\sigma}_\lambda^2 \sum_{t=1}^T (\gamma_{t-1}^{x_1})^2 + \sigma_\gamma^2}, \frac{\tilde{\sigma}_\lambda^2 \sigma_\gamma^2}{\tilde{\sigma}_\lambda^2 \sum_{t=1}^T (\gamma_{t-1}^{x_1})^2 + \sigma_\gamma^2}\right), \tag{A23}$$

$$\sigma_\gamma^2|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\sigma_\gamma^2} \sim \mathrm{IG}\left(\tilde{a}_\gamma + \frac{T}{2}, \tilde{b}_\gamma + \frac{1}{2}\sum_{t=1}^T \left(\gamma_t^{x_1} - \lambda\gamma_{t-1}^{x_1}\right)^2\right). \tag{A24}$$

Appendix A.5. Application of the Constraints

The constraints are applied for every iteration of the sampler. Let the current iteration be denoted by i, then the following procedures are performed:

1. The constraints of the vector $\boldsymbol{\beta}^{(i)}$ are applied after sampling the arbitrary vector of static parameters $\boldsymbol{\beta}^{(i)}$, the vector is mapped into a vector of transformed parameters, $\tilde{\boldsymbol{\beta}}^{(i)}$ by the following rescaling $\tilde{\boldsymbol{\beta}}^{(i)} = \frac{\boldsymbol{\beta}^{(i)}}{\sum_x \beta_x^{(i)}}$. Then we replace $\boldsymbol{\beta}^{(i)}$ with $\tilde{\boldsymbol{\beta}}^{(i)}$ and proceed to the next steps of the sampler.

2. The constraints for the latent processes $\kappa_t^{(i)}$ and γ_{t-x} are applied after the finalisation of Forward Backward. The arbitrary filtered estimates of the processes are transformed to $\tilde{\kappa}_t^{(i)} = \kappa_t^{(i)} - \bar{\kappa}^{(i)}$ and $\tilde{\gamma}_{t-x}^{(i)} = \gamma_{t-x}^{(i)} - \bar{\gamma}^{(i)}$ for $\bar{\kappa}^{(i)} = \frac{1}{N}\sum_{k=1}^N \kappa_{t_k}^{(i)}$ and $\bar{\gamma}^{(i)} = \frac{1}{N+p-1}\sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c^{(i)}$. Then we replace $\kappa_t^{(i)}$ with $\tilde{\kappa}_t^{(i)}$ and $\gamma_t^{(i)}$ with $\tilde{\gamma}_t^{(i)}$ and and proceed to the next steps of the sampler.

If one models the full cohort model the constraints of the vector $\boldsymbol{\beta}^\gamma$ are applied applied in the same fashion as for the vector $\boldsymbol{\beta}$. For a simplified cohort model, the vector of parameters is set to the vector of ones and is not sampled.

Appendix B. Description of Stochastic Mortality Models Utilizing Factor Extraction from European Demographic Data

Recalling the notation from the Equation (8), all models which utilize features extracted from Demographic data have the following the state-space representation

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \tilde{\mathbf{B}}_t\tilde{\boldsymbol{\varphi}}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0,\sigma_\varepsilon^2 \mathbb{I}_{21}), \tag{A25a}$$

$$\tilde{\boldsymbol{\varphi}}_t = \tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\varphi}}_{t-1} + \tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\kappa}}_t \overset{iid}{\sim} \mathcal{N}(0,\tilde{\Upsilon}). \tag{A25b}$$

As we examine only **Case 1** defined in Section 3, the corresponding transition matrices of the observation and state equations are equal to

$$\tilde{\mathbf{B}}_{t\ 21\times(22+m)} = \left(\ \mathbf{B}_{21\times22} \ \middle| \ \tilde{\mathbf{F}}_t\ \right)$$

$$\tilde{\boldsymbol{\Lambda}}_{(22+m)\times(22+m)} = \left(\begin{array}{c|c} \boldsymbol{\Lambda}_{22\times22} & \mathbf{0}_{22\times m} \\ \hline \mathbf{0}_{m\times22} & \boldsymbol{\Omega}_{m\times m} \end{array}\right) \tag{A26}$$

where m is the dimensionality of the latent process vector $\boldsymbol{\varrho}_t$ which corresponds to the factor matrix $\tilde{\mathbf{F}}_t$. The structure of the matrix depends on the models what is further discussed in the next two subsections.

Appendix B.1. DFM-PC-B Model

The model is constructed as follow

Step 1: Take the first eigenvector of robustly standardized Birth counts which is of vector-type, country specific

$$[\mathbf{F}]_{\cdot,1} = \begin{bmatrix} f_{1,1}^{AUT} \\ \vdots \\ f_{21,1}^{UKR} \end{bmatrix} ;$$

Step 2: Take the mean across the countries (components of the first eigenvector vector) which is a scalar $\hat{f}$ and use a one per age group latent process to model additional supplementary information per age group;

Step 3: Incorporate $\hat{f}$ as elements of the diagonal matrix $\tilde{\mathbf{F}}$

$$\tilde{\mathbf{F}} = \begin{bmatrix} \hat{f} & 0 & 0 & \cdots & 0 \\ 0 & \hat{f} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \hat{f} \end{bmatrix}_{21 \times 21}$$

The corresponding ϱ_t is age group specific vector, that is

$$\varrho_t = \left(\varrho_t^0, \ldots, \varrho_t^{95} \right)_{1 \times 21}$$

Appendix B.2. The Models of the Class DFM-PC-D and DFM-PC-Mx

The models are constructed as follow

Step 1: Take the first eigenvector of robustly standardized corresponding data set which is of matrix-type, age and country specific

$$\mathbf{F} = \begin{bmatrix} f_{1,1}^{AUT} & \cdots & f_{1,28}^{UKR} \\ \vdots & \ddots & \vdots \\ f_{21,1}^{AUT} & \cdots & f_{21,28}^{UKR} \end{bmatrix}_{21 \times 28}$$

Step 2: Notice that the matrix $\tilde{\mathbf{F}}$ is equal to $\mathbf{F}$

Step 3: Use a one per country latent process to model the impact of country specific vector;

The corresponding ϱ_t is country specific vector, that is

$$\varrho_t = \left(\varrho_t^{AUT}, \ldots, \varrho_t^{UKR} \right)_{1 \times 28}$$

References

(Basilevsky 1994) Basilevsky, Alexander T. 1994. *Statistical Factor Analysis and Related Methods*. Hoboken: John Wiley & Sons, Inc.

(Cairns et al. 2009) Cairns, Andrew J. G., David Blake, Kewin Dowd, Guy D. Coughlan, David Epstein, Alen Ong, and Igor Balevich. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* 13: 1–35.

(Carter and Kohn 1994) Carter, Chris K., and Robert Kohn. 1994. On Gibbs sampling for state-space models. *Biometrika* 81: 541–53.

(Celeux et al. 2006) Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington. 2006. Deviance information criteria for missing data models. *Bayesian Anal* 1: 651–73.

(Davies 1987) Davies, P. Laurie 1987. Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics* 15: 1269–92.

(Dempster et al. 1977) Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society. Series B (Methodological)* 39: 1–38.

(Diebold and Li 2006) Diebold, Francis X., and Canlin Li. 2006. Forecasting the term structure of government bond yields. *Journal of Econometrics* 130: 337–64.

(Erbas et al. 2010) Erbas, Bircan, Muhammed Akram, Dorota M. Gertig, Dallas English, John L. Hopper, Anne M. Kavanagh, and Rob Hyndman. 2010. Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the United States and England-Wales. *Journal of Epidemiology* 20: 159–65.

(Frahm and Jaekel 2010) Frahm, Gabriel, and Uwe Jaekel. 2010. A generalization of Tyler's M-estimators to the case of incomplete data. *Computational Statistics & Data Analysis* 54: 374–93.

(Friedman and Tukey 1974) Friedman, Jerome H., and John W. Tukey. 1974. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers* C-23: 881–90.

(Fung et al. 2016) Fung, Man Chung, Gareth W. Peters, and Pavel V. Shevchenko. 2016. A unified approach to mortality modelling using state-space framework: Characterisation, identification, estimation and forecasting. *Annals of Actuarial Science* 1–47. doi:10.1017/S1748499517000069.

(Fung et al. 2017) Fung, Man Chung, Gareth William Peters, and Pavel V. Shevchenko. 2017 Cohort Effects in Mortality Modelling: A Bayesian State-Space Analysis. Available online: https://ssrn.com/abstract=2907868 (accessed on 31 January 2017).

(Gaille and Sherris 2015) Gaille, Sererine Arnold, and Michael Sherris. 2015. Causes-of-Death Mortality: What Do We Know on Their Dependence? *North American Actuarial Journal* 19:116–28.

(Gavrilov and Gavrilova 2015) Gavrilov, Leonid A., and Natalia S. Gavrilova. 2015. Predictors of Exceptional Longevity: Effects of Early-Life and Midlife Conditions, and Familial Longevity. *North America Actuarial Journal* 19: 174–86.

(Girosi and King 2008) Girosi, Federico, and Gary King. 2008. *Demographic Forecasting*. Princeton: Princeton University Press.

(Haberman and Renshaw 2011) Haberman, Steven, and Arthur Renshaw. 2011. A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics* 48: 35–55.

(Hanewald 2011) Hanewald, Katja 2011. Explaining Mortality Dynamics: The Role of Macroeconomic Fluctuations and Cause of Death Trends. *North American Actuarial Journal, Series B* 290–314.

(Huber 1964) Huber, Peter J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35: 73–101.

(Huber and Ronchetti 2009) Huber, Peter J., and Elvezio M. Ronchetti. 2009. *Robust Statistics*. Wiley Series in Probability and Statistics. Hoboken: John Wiley & Sons, Inc., p. 380.

(Hunt and Villegas 2015) Hunt, Andrew, and Andres M. Villegas. 2015. Robustness and convergence in the Lee-Carter model with cohort effects. *Insurance: Mathematics and Economics* 64: 186–202.

(Hyndman and Yasmeen 2012) Hyndman, Rob J., and Farah Yasmeen. 2012. Common functional principal component models for mortality forecasting. In *Contributions in Infinite-Dimensional Statistics And Related Topics*. chp. 29, pp. 161–66.

(Jamshidian 1997) Jamshidian, Mortaza 1997. An EM Algorithm for ML Factor Analysis with Missing Data. In *Lecture Notes in Statistics*. New York: Springer, pp. 247–58.

(Jolliffe 2002) Jolliffe, I. T. 2002. *Principal Component Analysis* New York: Springer.

(Kogure and Kurachi 2010) Kogure, Atsuyuki, and Yoshiyuki Kurachi. 2010. A Bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics* 46: 162–72.

(Lee and Carter 1992) Lee, Ronald D., and Lawrence R. Carter. 1992. Modeling and forecasting US mortality. *Journal of the American Statistical Association* 87: 659–75.

(Little and Rubin 2002) Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. Hoboken: John Wiley & Sons, Inc.

(Lopuhaa 1989) Lopuhaa, Hendrik P. 1989. On the Relation between S-Estimators and M-Estimators of Multivariate Location and Covariance. *The Annals of Statistics* 17: 1662–83.

(Maronna 1976) Maronna, Ricardo Antonio 1976. Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics* 4: 51–67.

(Murray and Lopez 1997) Murray, Christopher J. L., and Alan D. Lopez. 1997. Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. *The Lancet* 1498–1504.

(Niu and Melenberg 2014) Niu, Geng, and Bertrand Melenberg. 2014. Trends in Mortality Decrease and Economic Growth. *Demography* 51: 1755–73.

(Pedroza 2006) Pedroza, Claudia 2006. A Bayesian forecasting model: predicting US male mortality. *Biostatistics* 7: 530–50.

(Renshaw and Haberman 2006) Renshaw, Arthur E., and Steven Haberman. 2006. A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* 38: 556–70.

(Rousseeuw and Yohai 1984) Rousseeuw, Peter, and Victor Yohai. 1984. *Robust Regression by Means of S-Estimators*. Robust and Nonlinear Time Series Analysis, New York: Springer, pp. 256–72.

(Roweis 1998) Roweis, Sam T. 1998. EM Algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, pp. 626–32.

(Rubin and Thayer 1982) Rubin, Donald B., and Dorothy T. Thayer. 1982. EM algorithms for ML factor analysis. *Psychometrika* 47: 69–76.

(Shang and Hyndman 2016) Shang, Han Lin, and Rob J. Hyndman. 2016. Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics* 26: 330–43.

(Spiegelhalter et al. 2002) Spiegelhalter, David J., Nicola G. Best, and Bradley P. Carlin. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* 64: 583–639.

(Sun et al. 2016) Sun, Ying, Prabhu Babu, and Daniel P. Palomar. 2016. Robust Estimation of Structured Covariance Matrix for Heavy-Tailed Elliptical Distributions. *IEEE Transactions on Signal Processing* 64: 3576–90. [1506.05215].

(Tipping and Bishop 1999) Tipping, Michael E., and Christopher M. Bishop. 1999. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61: 611–22.

(Tyler 1987) Tyler, David E. 1987. A Distribution-Free M-Estimator of Multivariate Scatter. *The Annals of Statistics* 15: 234–51.

(Tyler 1987) Tyler, David E. 1987. Statistical Analysis for the Angular Central Gaussian Distribution on the Sphere. *Biometrika* 74: 579.

(Willets 2004) Willets, R. C. 2004. The cohort effect: insights and explanations. *British Actuarial Journal* 10: 833–77.

(Wilmoth et al. 2007) Wilmoth, J. R., K. Andreev, and D. Jdanov. 2007. *Methods Protocol for the Human Mortality Database*, Technical Report. Available online: http://www.mortality.org/Public/Docs/MethodsProtocol.pdf