


RESEARCH ARTICLE

Multi-population mortality modelling: a Bayesian hierarchical approach

Jianjie Shi^{1†}, Yanlin Shi², Pengjie Wang¹ and Dan Zhu¹ 

¹Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia and ²Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, Australia

Corresponding author: Dan Zhu; Email: dan.zhu@monash.edu

Received: 5 November 2022; **Revised:** 11 June 2023; **Accepted:** 26 July 2023; **First published online:** 25 August 2023

Keywords Lee–Carter model; multi-population approach; Markov Chain Monte Carlo; vector error correction model; structural analysis

Abstract

Modelling mortality co-movements for multiple populations has significant implications for mortality/longevity risk management. This paper assumes that multiple populations are heterogeneous sub-populations randomly drawn from a hypothetical super-population. Those heterogeneous sub-populations may exhibit various patterns of mortality dynamics across different age groups. We propose a hierarchical structure of these age patterns to ensure the model stability and use a Vector Error Correction Model (VECM) to fit the co-movements over time. Especially, a structural analysis based on the VECM is implemented to investigate potential interdependence among mortality dynamics of the examined populations. An efficient Bayesian Markov Chain Monte-Carlo method is also developed to estimate the unknown parameters to address the computational complexity. Our empirical application to the mortality data collected for the Group of Seven nations demonstrates the efficacy of our approach.

1. Introduction

Mortality modelling is an important topic in actuarial science and insurance practice, dating back to the deterministic and one-dimensional Gompertz model. Over the past decades, stochastic approaches are rapidly developed, represented by the seminal (Lee and Carter, 1992), or LC model. The LC model is based on a combination of temporal trends and age patterns in logged central mortality rates, which has become a standard for model and project dynamics of mortality rates. Its forecasting performance is well documented in existing literature (see, for example, Lee and Miller, 2001, among others). Intrinsically, the LC model assumes that a single time-varying index drives the temporal dynamics of mortality rates. The associated forecast relies on the extrapolation of this index with an appropriate statistical linear time-series model. The popularity of LC stems from its simple construction and straightforward interpretations (Lee, 2000; Denuit *et al.*, 2007).

This paper focuses on modelling and forecasting multi-population mortality rates, which are more comprehensive than a single-population framework and are receiving attention from recent literature. For example, a multi-population framework is used to study a group of countries with similar socioeconomic situations or males and females in the same population (Li and Li, 2017; Boonen and Li, 2017). These models are motivated to assess the demographic basis risk involved in an index-based longevity hedge by comparing and projecting the reference and target populations' mortality experience (Li and Lee, 2005). For instance, when projecting mortality for a smaller community with a thin volume of mortality data, the forecaster may aim to improve the credibility by modelling the smaller population jointly

[†]This author thanks the support from Australian Government Research Training Program (RTP) Scholarship.

with a larger population. Another useful approach is to simultaneously forecast mortality rates of both males and females of the same population, thereby ensuring consistency of the sex differentials.

Regarding the multi-population mortality modelling, Li and Lee (2005) are among the first to extend the LC model and propose a multiple-population counterpart, or the LL model. The LL framework uses a common factor to describe the long-term temporal trend shared by all countries within the investigated group to model the mortality rates. A country-specific factor is also adopted to describe the short-term country-specific patterns, effectively avoiding the undesirable divergence of mortality forecasts among populations in the long run.

More recently, other extensions of the LC model to the multi-population framework have flourished, and examples include Yang and Wang (2013), Zhou *et al.* (2014), Li *et al.* (2015) and Danesi *et al.* (2015). For instance, Kleinow (2015) proposed a common-age-effect (CAE) model to allow more than two populations, which extends the LL model. The age effect of CAE model, however, is assumed identical for all populations. This is motivated by the observation that obtained age effects are close to each other when estimated among different countries of similar socioeconomic structures. Thus, the number of parameters (i.e., age effects) can be reduced when simultaneously modelling their mortality experiences.

As pointed out by Chen *et al.* (2015), the CAE-type models might only be justified by the long-term mortality co-integration, yet it seems too strong to model the short-term mortality dependence. Alternatively, an ARMA-GARCH process with heavy-tailed innovations was proposed to filter the mortality dynamics of each population. The residual risk could then be fitted via a one-factor copula model. However, the increased complexity in its modelling structure is the drawback when compared with the LC-type models.

Extended from the classic LC model, this paper proposes a new multi-population approach based on a hierarchical structure to model all examined populations. Our approach effectively balances the dichotomy of short-term predictive power and long-term coherence. Specifically, unlike the existing CAE framework, in our model, population-wise age effects are allowed to improve the short-term forecasting accuracy. Further, similar to the models of Yang and Wang (2013) and Zhou *et al.* (2014), the long-term coherence is considered, since all population-wise age effects are random vectors generated from the *same* parametric distribution. More importantly, our model manages to achieve the improved flexibility while retaining a relatively parsimonious model specification. Given the limited availability of mortality data, such a hierarchical structure effectively utilises cross-sectional information in the estimation and forecasting. As for the temporal dimensionality, a Vector Error Correction Model (VECM) is employed to model the co-movements among sub-populations. Relevant parameter restrictions are discussed for identification purpose, which also provides asymptotically coherent forecasts of mortality rates. The employed VECM also enables subsequent structural analyses, which examine the interdependence of mortality dynamics of sub-populations.

In terms of the parameter estimation of our model, the traditional singular value decomposition (SVD) for estimating the LC model is no longer feasible. Although the integrated likelihood function exists analytically, its exact computation involves large dimensional Kronecker products, resulting in difficulty to implement the standard maximum likelihood estimation algorithms. Due to the high dimensionality and the limited sample size of mortality data, other popular alternative such as the Expectation-Maximization technique is not pursued in this paper. To combat against the complexity, we implement the Bayesian estimation approach.

Over the past decades, Bayesian inference has become a popular statistical approach for its ability to capture complicated dynamics and its inherent parameter variability assumption (Czado *et al.*, 2005; Pedroza, 2006; Kogure and Kurachi, 2010; Li *et al.*, 2015; Wong *et al.*, 2018; Lin and Tsai, 2022). The first Bayesian attempt to implement the LC model via state-space model is made by Pedroza (2006) for a single-population framework, followed by a two-population age-period-cohort model developed in Cairns *et al.* (2011). The benefits of using Bayesian methods in mortality modelling are fourfold:

- It is particularly suitable for constructing hierarchical models, where the inference is obtained iteratively through the conditional likelihoods.
- It allows for the imputation of the prior knowledge that one can blend in the belief of long-term coherence among the population groups (Hyndman *et al.*, 2013) via prior hyper-parameters.
- Unlike the frequentist framework, the model parameters are random variables, which are convenient for superannuation and life insurance fund with heterogeneity among the policyholders (Cairns, 2000).
- Bayesian methods can straightforwardly deal with missing data (Li *et al.*, 2019).

Despite the sophisticated framework, our proposed hierarchical model can be efficiently estimated via the Bayesian Markov Chain Monte-Carlo (MCMC) algorithm. Nevertheless, a naive implementation of the MCMC is undesirable, due to the resulting slow mixing of the chain and poor estimation results. To address this, we develop an efficient MCMC algorithm for sampling posterior distributions and predictive densities. The model is first rewritten as a state-space model under a matrix-variate Gaussian formulation (Gupta and Varga, 1992). An algorithm is then proposed to sample the age random effects for all the populations in one block. Finally, instead of using the standard Kalman filter (Koopman and Durbin, 2003) for sampling the latent factors, we employ an efficient precision sampler of Chan and Jeliuzkov (2009) that substantially reduces the computational cost.

To demonstrate its usefulness, we apply our multi-population LC model to empirical mortality data of the Group of Seven (G7) countries, consisting of Canada, France, Germany, Italy, Japan, the United Kingdom and the United States. The sample ranges from 1956 to 2019. Compared with other competing models, our baseline approach achieves favourable in-sample and out-of-sample forecasting results. A structural analysis of mortality dynamics is further implemented.

The contributions of our research are fourfold. First, we provide a parsimonious hierarchical framework and propose a multi-population LC model. This Bayesian hierarchical model extends Pedroza (2006)'s state-space representation by introducing random age effects to account for multi-population modelling. Compared with existing approaches, such as the CAE and those proposed in Yang and Wang (2013) and Zhou *et al.* (2014), the more flexible population-wise random effects are allowed with minimal additional complexity. Lin and Tsai (2022) also consider introducing Bayesian hierarchical method to multi-population modelling, but their underlying model of log mortality rate is a random walk with a drift process. In contrast, our paper develops a Bayesian hierarchical method based on the classical Lee–Carter model and its multi-factor extension. Second, via the VECM approach with an appropriate shrinkage prior, the proposed model has co-integrated temporal factors and thus achieves the desirable long-term coherence across populations. Thus, same as the seminal LL model, the long-run divergence in forecast mortality rates among populations is prevented, which cannot be ensured if independent LC models were adopted (Tuljapurkar *et al.*, 2000). Compared to the LL specification and other coherent LC extensions (Yang and Wang, 2013; Zhou *et al.*, 2014), our model is more flexible, such that temporal interactions across populations could be investigated. In particular, our VECM approach enables a structural analysis to examine the interdependence of mortality dynamics of G7 populations. To the best of our knowledge, this paper is among the first to consider such analyses. We demonstrate that structural shock to US mortality rate can permanently lead to mortality declines of other countries. In the long run, mortality dynamics of Japan and Germany have non-negligible contributions to US mortality changes. Third, an associated Bayesian algorithm to implement the estimation is developed. Different from a naive MCMC, our approach significantly reduces the computational cost and improves the reliability of estimation. Fourth, the empirical results, demonstrate the outperformance of our model, when both in- and out-of-sample forecasts are considered. Also, our empirical evidence shows significant variation in the age effects and significant temporal interactions among examined populations. This supports the importance of our hierarchical structure to more effectively utilise the cross-sectional information.

The remainder of the paper is organised as follows. Section 2 introduces our multi-population LC model and provides appropriate parameter restrictions for identification and asymptotic coherence. The model is then calibrated in the Bayesian framework in Section 3, and an efficient estimation algorithm

is provided. The in-sample and out-of-sample empirical performances of our proposed models are then presented in Section 4, along with an empirical structural analysis. Finally, we conclude the paper in Section 5.

2. A multi-population Lee–Carter framework

2.1. The classic Lee–Carter specification

Lee and Carter (1992) developed an approach to study mortality data of the United States, which has been widely applied in actuarial and demographic research. The LC model to forecast mortality rates is essentially via extrapolating historical trends and predicting probability distributions of age-specific death rates (ASDR) using standard time-series procedures. The basic LC specification is displayed below:

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

for ages $x = x_0, \dots, \omega$ and years $t = 1, \dots, T$. $\mathbf{y}_t = [\log(m_{x_0,t}), \dots, \log(m_{\omega,t})]'$ is the vector of logged ASDR. We also have that $\boldsymbol{\alpha} = [\alpha_{x_0}, \dots, \alpha_{\omega}]'$, and $n = \omega - x_0 + 1$ denotes the number of age groups. The popularity of LC model stems from its simple interpretation, that α_x is the long term average of $\log(m_{x,t})$, κ_t is the common temporal trend of mortality change and assumed a latent factor, and β_x is the relative sensitivity of the ASDR with respects to the time change.

To estimate this single population Lee–Carter model, the original approach is to use sample average for α_x and apply a singular value decomposition on $[\mathbf{y}_1, \dots, \mathbf{y}_T] - \boldsymbol{\alpha} \mathbf{1}'_T$ to extract $\boldsymbol{\kappa} = [\kappa_1, \dots, \kappa_T]'$. Then, $\boldsymbol{\kappa}$ is adjusted to fit the reported life expectancy at each time. This second stage makes the model fit historical life expectancy exactly. The adjusted $\boldsymbol{\kappa}$ is then modelled using standard time-series methods, typically a random walk with a drift, as suggested in the original work of Lee and Carter (1992), to produce mortality forecasts.

2.2. A multi-population extension of Lee–Cartered model

Despite its popularity, independently fitting single-population LC models are insufficient to comprehensively study the mortality dynamics. To see this, in Figure 1, point forecasts of logged central death rates for age 65 (spanning 2020–2119) are depicted for all the G7 countries, where the LC model is independently fitted to each country’s mortality rate over 1956–2019, with male and female data combined. Obviously, those point forecasts in Figure 1 are non-coherent, that is, the predicted mortality rates of those populations diverge over time. This lack of long-term coherence motivated the class of coherent multi-population models (Li and Lee, 2005; Li *et al.*, 2015).

Studying mortality experience in the multiple population context is a more complicated problem. Early explorations by Tuljapurkar *et al.* (2000) treated the mortality movements of each population independently and modelled the age-specific mortality via

$$\log(m_{x,t}^i) = \alpha_x^i + \beta_x^i \kappa_t^i + \epsilon_{x,t}^i$$

for $i = 1, \dots, I$, where I denoting the number of populations. Various extensions were studied in the literature to allow dependence across populations, such as adopting population-dependent κ_t^i 's in Yang and Wang (2013) and Zhou *et al.* (2014), and common age effects in Kleinow (2015).

To address the aforementioned empirical issues, this paper introduces a framework that jointly models all populations’ temporal effects and allows heterogeneous age effects. In particular, to ensure the coherent forecasts, a common distribution is shared among these age effects. The concept of coherence is formally defined and discussed in Section 2.3. Listed below is the specification of our baseline model.

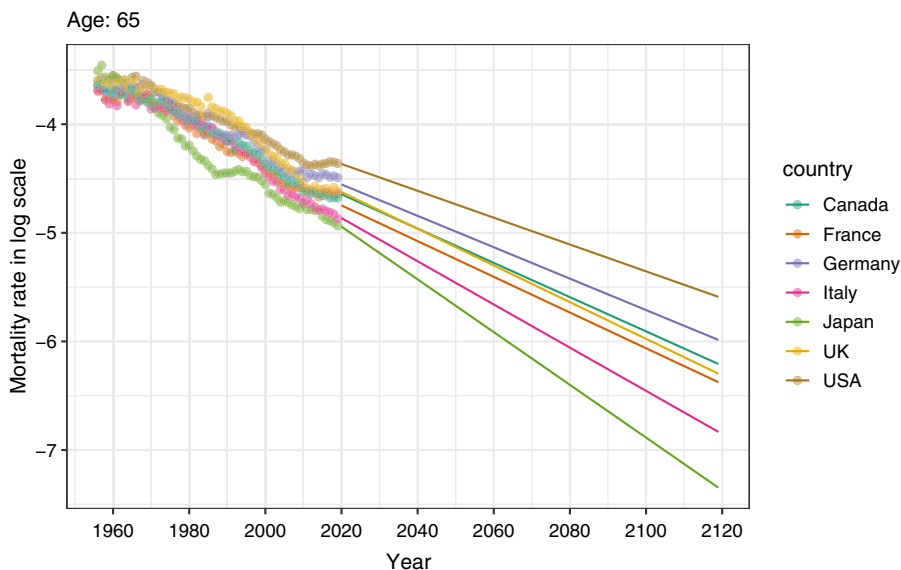


Figure 1. Point forecasts of log mortality rates for Age 65 derived from independent LC models.

Model 1 (The Multi-population LC model). Let $y_t^i = [\log(m_{x_0,t}^i), \dots, \log(m_{\omega,t}^i)]'$ for $i = 1, 2, \dots, I$, we assume that

$$\begin{aligned}
 y_t^i &= \alpha^i + \beta^i \kappa_t^i + \epsilon_t^i, \quad \epsilon_t^i \sim F_\epsilon(\cdot; \Omega) \\
 \alpha^i &= [\alpha_{x_0}^i, \dots, \alpha_\omega^i]' \sim F_a(\cdot | \mu_a, \Sigma_a) \\
 \beta^i &= [\beta_{x_0}^i, \dots, \beta_\omega^i]' \sim F_b(\cdot | \mu_b, \Sigma_b)
 \end{aligned} \tag{2.1}$$

with F_a, F_b and F_ϵ denote some multivariate parametric distributions and

$$\begin{aligned}
 cov(\epsilon_t^i) &= \Omega, \quad \Omega = diag\{g_{x_0}, \dots, g_\omega\} \\
 \mathbb{E}[\alpha^i] &= \mu_a, \quad cov(\alpha^i) = \Sigma_a \\
 \mathbb{E}[\beta^i] &= \mu_b, \quad cov(\beta^i) = \Sigma_b
 \end{aligned}$$

for all $i = 1, 2, \dots, I$. For simplicity, we assume the covariance-variance matrix Ω to be a diagonal matrix with g_k being the error variance for the k th age.

For the temporal movements of latent factors, we consider a Vector Error Correction Model (VECM) for $\kappa_t = [\kappa_t^1, \dots, \kappa_t^I]'$ such that

$$\Delta \kappa_t = b + \Pi \kappa_{t-1} + \xi_t, \quad \xi_t \sim N(0, \Sigma_\kappa), \tag{2.2}$$

where N denotes the multivariate Gaussian distribution. The matrix of long-run multipliers, Π , can be written as $\Pi = \mathbf{c}\mathbf{d}'$, where \mathbf{c} and \mathbf{d} are both full rank $I \times r$ matrices and where $0 \leq r \leq I$ is the number of co-integrating relationships. The matrix \mathbf{d} is called a cointegration matrix and \mathbf{c} is sometimes called a loading matrix. (From the Proposition 1 discussed in Section 2.4, to ensure the long-term coherent mortality forecasts, we require that κ_t^i is an $I(1)$ process for any $i \in \{1, 2, \dots, I\}$. This is a common assumption in mortality literature, for example, Li and Lee (2005), Yang and Wang (2013) and Zhou et al. (2014) among others. In that case, both $\Delta \kappa_t$ and $\mathbf{d}'\kappa_t$ are stationary, and hence, each element of $\mathbf{d}'\kappa_t$ represents a long-run equilibrium relation.) Especially, if $r = I$ then all the elements of κ_t are stationary, while if $r = 0$ then all the series are $I(1)$ processes without any existing co-integration.

Hence, the model parameter is

$$\theta = \{b, \Pi, \Sigma_\kappa, \mu_a, \mu_b, \Sigma_a, \Sigma_b, \Omega\}.$$

In Model 1, the co-movements of population-wise mortality rates are assumed to follow a VECM model without lagged differences, which is equivalent to a restricted VAR(1) model. The VECM is a theoretical-based approach which is useful for combining both long-run co-integration relationship and short-run corrections/adjustments of co-integrated variables towards the long-run equilibrium. Also, the VECM could be used to analyse interdependence of response variables via a structural analysis, which is popular in empirical fields, such as macroeconomics (Kunst and Neusser, 1990; Granger, 2004) and finance (Mukherjee and Naka, 1995). More recently, VECM/co-integration techniques have been employed to produce coherent forecasts in mortality modelling (see, for example, Yang and Wang, 2013; Zhou *et al.*, 2014; Hunt and Blake, 2015; Li and Lu, 2017; Hunt and Blake, 2018; Li *et al.*, 2021; Li and Shi, 2021a,b, among others). This is for the property of VECM such that the distribution $\kappa_t | \kappa_{t-1}$ is known, which then allows coherent mortality forecasts. Please see Jarner and Jallbjørn (2020) for a systematic review on benefits and drawbacks of co-integration based mortality models.

The second unique feature of Model 1 is the allowed heterogeneous age effects, that is, α_x^i 's and β_x^i 's are different across populations. Yet, they are modelled as random effects drawn from a common distribution. The hierarchical structure actually assumes that multiple populations are heterogeneous sub-populations randomly drawn from a hypothetical super-population (represented by the common distribution). The advantages of this hierarchical structure are summarised below. First, this is a parsimonious specification that greatly reduces the dimension of model parameters. Second, the heterogeneity in the age effects is retained, which often provides a better short-term predictive power (Chen *et al.*, 2015). Last, as will be shown below, the common distribution and restrictions on κ_t coefficients will enable asymptotic coherence in mortality forecasting (by Proposition 1, we also need β^i to be deterministic with $|\mu_b| < \infty$). Although it is beyond the scope of this paper, the hierarchical structure could also be extended to other LC-typed models, such as the LL. Compared to existing competitors, it is expected that more effectively utilising the cross-sectional information, as in Model 1, can generally improve forecasting accuracy for mortality data.

The multi-population LC model, as a special case of the linear dynamic factor models, is subject to identification issues (Li and Lu, 2017). The dynamic factor model is a flexible but parsimonious approach to study unobserved heterogeneity, and relevant parameter restrictions for identification purposes have been well discussed in the literature (Bai and Wang, 2015). To identify dynamic factors (i.e., mortality indices κ_t) in the Model 1, we consider two identification assumptions stated below:

Assumption 1. For all $i = 1, \dots, I$,

- (a) $\alpha^i, \beta^i, \kappa_t^i$ and ϵ_t^i are mutually independent
- (b) $\mathbb{E}[\epsilon_t^i] = 0$.

Assumption 2. We assume $\kappa_{-1}^i = [\kappa_2^i, \dots, \kappa_T^i]'$ with $\kappa_1^i = 0$, and

$$\beta^i \sim N_S(\mu_b, \Sigma_b) \text{ where } S = \{\beta_{x_0}^i = 1\},$$

where N_S denotes a multivariate normal distribution truncated on a hyper-plane defined by S . Furthermore, we assume

$$\alpha^i \sim N(\mu_a, \Sigma_a)$$

and

$$\epsilon_t^i \sim N(0, \Omega),$$

where Ω is a diagonal matrix while Σ_a and Σ_b are two general positive-semidefinite matrices.

As a result of Assumption 1, Model 1 reduces to the independent LC model given the random effects. The identification of our model then boils down to the identification of each single-population Lee-Carter model. Following Lee and Carter (1992), in Assumption 2, we let $\beta_{x_0}^i = 1$ and $\kappa_1^i = 0$ for $i \in I$. The merits of the above constraints can be illustrated for the ease of computation and interpretation. (The advantage of this set constraint is that the parameters could still be interpreted as in the classical LC model, except that the interpretation of α_x^i changes slightly. In our case, α_x^i represents the mortality level at the base year. That is, $\mathbb{E}[\log(m_{x,1}^i)] = (\alpha_x^i) + (\beta_x^i)(\kappa_1^i) = \alpha_x^i$.) And the normality assumption is for simplicity purpose in empirical analysis. By introducing the Gaussian assumption (and conditionally conjugate priors in Section 3.1), one can easily obtain analytical full conditional posterior distributions for model parameters (and dynamic factors). This enables the application of a standard Gibbs sampler to approximate joint posterior distribution.

Note that our model offers a generalised framework and nests a range of specifications. For instance, if both Π and Σ_κ are diagonal matrices, it essentially becomes independent LC models for each population group (especially when Π is a zero matrix, all the κ_t^i s will follow independent random walks with drifts). It could also be shown that CAE-type models (Kleinow, 2015) are nested, when associated parameter restrictions are implemented in Model 1. Specifically, the hierarchical random effects can easily reduce to the homogeneous models when Σ_a or Σ_b are zeros, which suggests that the age effects are the same among all populations. In our subsequent study, Bayesian shrinkage priors will be employed to impose the belief of certain nested model when performing the MCMC estimation for Model 1. The implementation of those priors will not enforce a draconian parameter restriction to make Model 1 a reduced structure. In other words, this effectively balances between the prior belief of the long-term coherence as for the CAE model and inherent features of the empirical data.

In comparison, we also consider two restricted cases, namely Model 2 and Model 3. Specifically, Model 2 assumes a homogeneous age effect such that all the β^i 's are the same across different populations. As will be discussed in Section 2.3, this condition is necessary to ensure the coherent forecasts. In that sense, Model 2 is a special case of Model 1 where Σ_b is a zero matrix. Instead of a random effect, $F_b(\cdot | \mu_b, \Sigma_b)$ in Model 2 just serves as a hierarchical prior for the age effect β .

Model 2 (The multi-population LC model with homogeneous age effect β).

$$\begin{aligned} y_t^i &= \alpha^i + \beta \kappa_t^i + \epsilon_t^i, \quad \epsilon_t^i \sim F_\epsilon(\cdot; \Omega) \\ \alpha^i &= [\alpha_{x_0}^i, \dots, \alpha_\omega^i]' \sim F_a(\cdot | \mu_a, \Sigma_a) \\ \beta &= [\beta_{x_0}, \dots, \beta_\omega]' \sim F_b(\cdot | \mu_b, \Sigma_b) \end{aligned} \tag{2.3}$$

with a VAR(1) in the VECM representation for $\kappa_t = [\kappa_t^1, \dots, \kappa_t^I]'$ such that

$$\Delta \kappa_t = b + \Pi \kappa_{t-1} + \xi_t, \quad \xi_t \sim N(0, \Sigma_\kappa). \tag{2.4}$$

Hence, the model parameter is $\theta = \{b, \Pi, \Sigma_\kappa, \mu_a, \mu_b, \Sigma_a, \Sigma_b, \Omega\}$.

In Model 3, we retain hierarchical structures of α^i and β^i but assume that they are population-invariant. That is, in contrast to Model 1, Model 3 is a multi-population LC model with homogeneous age effects α and β . This is also a special case of Model 1, where both Σ_a and Σ_b in Model 1 are zeros. Similar to Model 2, $F_a(\cdot | \mu_a, \Sigma_a)$ and $F_b(\cdot | \mu_b, \Sigma_b)$ are essentially hierarchical priors for the age effects α and β , respectively, rather than random effects.

Model 3 (The Multi-population LC model with homogeneous age effects α and β).

$$\begin{aligned} y_t^i &= \alpha + \beta \kappa_t^i + \epsilon_t^i, \quad \epsilon_t^i \sim F_\epsilon(\cdot; \Omega) \\ \alpha &= [\alpha_{x_0}, \dots, \alpha_\omega]' \sim F_a(\cdot | \mu_a, \Sigma_a) \\ \beta &= [\beta_{x_0}, \dots, \beta_\omega]' \sim F_b(\cdot | \mu_b, \Sigma_b) \end{aligned} \tag{2.5}$$

with a VAR(1) in the VECM representation for $\kappa_t = [\kappa_t^1, \dots, \kappa_t^I]'$ such that

$$\Delta\kappa_t = b + \Pi\kappa_{t-1} + \xi_t, \quad \xi_t \sim N(0, \Sigma_\kappa). \tag{2.6}$$

Hence, the model parameter is $\theta = \{b, \Pi, \Sigma_\kappa, \mu_a, \mu_b, \Sigma_a, \Sigma_b, \Omega\}$.

2.3. Multi-population coherence

In this section, we discuss the concept of long-term coherence in the multi-population mortality modelling framework. Generally speaking, it means that death rates in two modelled populations do not diverge in the long run. Following Li and Lee (2005), the formal definition is stated below.

Definition 1. *The forecasts of a multi-population mortality model are asymptotically coherent if*

$$\lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{m_{x,t}^i}{m_{x,t}^j} \middle| \theta \right]^k < \infty \tag{2.7}$$

for $i, j \in \{1, 2, \dots, I\}$ and $k = 1, 2$.

As outlined above, forecasting of the original LC model is the extrapolation of historical temporal trends. However, the price to pay for this simplicity is the incapability to ensure coherence. As demonstrated in Figure 1, a clear long-term divergence of the forecast mortality is demonstrated when LC models are independently fitted.

To realise the coherent forecasting, a simple strategy is to assume that $\beta^i = \beta^j$ and that the spread $\kappa_t^i - \kappa_t^j$ is mean-reverting. It can be shown that this condition is sufficient for a standard LC formulation. To see this, if all populations have the same β_x and long term κ_x^i , then the ratios of the mean ASDRs among populations would be constant over time at each age in the forecasts. Otherwise, its projections of some ASDR would differ from those of others over time. In the Proposition described below, the conditions for coherent modelling are derived.

Proposition 1. *Suppose that*

1. *Assumptions 1 and 2 are satisfied,*
2. *β^i are deterministic vectors with $\|\mu_b\| < \infty$ for all $i = 1, 2, \dots, I$ (i.e., Σ_b is a zero matrix); and*
3. *κ_t^i is an I(1) process for any $i \in \{1, 2, \dots, I\}$ and $\kappa_t^i - \kappa_t^j$ is a weakly stationary process for $i, j \in \{1, 2, \dots, I\}$ and $i \neq j$, the forecast ASDR produced by Model 1 are asymptotically coherent.*

Proof. We need to justify that under the parameter constraints in Proposition 1, the mortality forecasts are not divergent for any $i, j \in \{1, 2, \dots, I\}$ and $i \neq j$. Since

$$\begin{aligned} \mathbb{E} \left[\frac{m_{x,t}^i}{m_{x,t}^j} \middle| \theta \right] &= \mathbb{E} \left\{ \exp[\log(m_{x,t}^i) - \log(m_{x,t}^j)] \middle| \theta \right\} \\ &= \mathbb{E} \left\{ \exp[(\alpha_x^i - \alpha_x^j) + (\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j) + (\epsilon_{x,t}^i - \epsilon_{x,t}^j)] \middle| \theta \right\} \\ &= \mathbb{E} \left[\exp(\alpha_x^i - \alpha_x^j) \middle| \theta \right] \cdot \mathbb{E} \left[\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j) \middle| \theta \right] \cdot \mathbb{E} \left[\exp(\epsilon_{x,t}^i - \epsilon_{x,t}^j) \middle| \theta \right], \end{aligned}$$

it is sufficient to prove mortality forecasts to be coherent (i.e., $\lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{m_{x,t}^i}{m_{x,t}^j} \middle| \theta \right] < \infty$) by demonstrating that all the limits of $\mathbb{E} \left[\exp(\alpha_x^i - \alpha_x^j) \middle| \theta \right]$, $\mathbb{E} \left[\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j) \middle| \theta \right]$ and $\mathbb{E} \left[\exp(\epsilon_{x,t}^i - \epsilon_{x,t}^j) \middle| \theta \right]$ are finite when $t \rightarrow \infty$ under the assumptions in Proposition 1. To do so, it is then sufficient to show

that $\mathbb{E}[\exp(\alpha_x^i - \alpha_x^j)|\theta]$, $\mathbb{E}[\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j)|\theta]$ and $\mathbb{E}[\exp(\epsilon_{x,t}^i - \epsilon_{x,t}^j)|\theta]$ are all upper-bounded by their respective time-invariant constants.

We first prove that for any t , both $\mathbb{E}[\exp(\alpha_x^i - \alpha_x^j)|\theta]$ and $\mathbb{E}[\exp(\epsilon_{x,t}^i - \epsilon_{x,t}^j)|\theta]$ are upper-bounded. Based on the prior setting, $\alpha_x^i|\theta \sim N([\mu_a]_x, [\Sigma_a]_x)$ for age group x in population i , where $[\mu_a]_x$ denotes the x -th element in the mean vector μ_a and $[\Sigma_a]_x$ is the x -th diagonal element of the variance-covariance matrix Σ_a . Therefore, $\exp(\alpha_x^i)|\theta \sim \text{log-N}([\mu_a]_x, [\Sigma_a]_x)$; here, log-N is the log-normal distribution. We can further deduce that $\exp(\alpha_x^i - \alpha_x^j)|\theta \sim \text{log-N}(0, 2[\Sigma_a]_x)$ since $\exp(\alpha_x^i)$ and $\exp(\alpha_x^j)$ are identically, independently distributed. Hence, we can see

$$\mathbb{E}[\exp(\alpha_x^i - \alpha_x^j)|\theta] = \exp([\Sigma_a]_x) < \infty,$$

which is a bounded value dependent on the age only. Similarly, for any t ,

$$\mathbb{E}[\exp(\epsilon_{x,t}^i - \epsilon_{x,t}^j)|\theta] = \exp([\Omega]_x) < \infty,$$

where $[\Omega]_x$ represents the x -th value on the diagonal line of Ω .

The remaining task is to prove that $\lim_{t \rightarrow \infty} \mathbb{E}[\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j)|\theta]$ is finite. Based on the prior setting of β_x^i , $\beta_x^i|\theta \sim N([\mu_b]_x, [\Sigma_b]_x)$, and it is easy to show that $\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j)|\theta, \kappa_t^i, \kappa_t^j \sim \text{log-N}([\mu_b]_x, (\kappa_t^i - \kappa_t^j), [\Sigma_b]_x[(\kappa_t^i)^2 + (\kappa_t^j)^2])$. Then according to the assumptions in Proposition 1, that is, $[\Sigma_b]_x = 0$ and $\kappa_t^i - \kappa_t^j$ being a weakly stationary process, for any t , we will have

$$\begin{aligned} \mathbb{E}[\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j)|\theta] &= \mathbb{E}\left\{\mathbb{E}[\exp(\beta_x^i \kappa_t^i - \beta_x^j \kappa_t^j)|\theta, \kappa_t^i, \kappa_t^j]\right\} \\ &= \mathbb{E}\left\{\exp\left([\mu_b]_x(\kappa_t^i - \kappa_t^j) + \frac{1}{2}[\Sigma_b]_x[(\kappa_t^i)^2 + (\kappa_t^j)^2]\right)|\theta\right\} \\ &= \mathbb{E}\left\{\exp([\mu_b]_x(\kappa_t^i - \kappa_t^j))|\theta\right\} \\ &= \exp\left(k_{ij}[\mu_b]_x + \frac{1}{2}K_{ij}[\mu_b]_x^2\right) < \infty \end{aligned}$$

where k_{ij} and K_{ij} represent the stationary mean and variance of $\kappa_t^i - \kappa_t^j$, respectively. The last equality holds because $\kappa_t^i - \kappa_t^j$ follows a stationary Gaussian process. To prove this, it is enough to show that κ_t follows a Gaussian process (stationarity of $\kappa_t^i - \kappa_t^j$ holds by assumption). Starting with an initial state $\kappa_1 = \mathbf{0}$, the VECM form in Equation (2.2) gives us:

$$\begin{aligned} \kappa_t &= \sum_{i=0}^{t-2} \Pi_i^* b + \Pi_{t-1}^* \kappa_1 + \sum_{i=0}^{t-2} \Pi_i^* \xi_{t-i} \\ &= \sum_{i=0}^{t-2} \Pi_i^* b + \sum_{i=0}^{t-2} \Pi_i^* \xi_{t-i}, \quad \xi_{t-i} \stackrel{iid}{\sim} N(0, \Sigma_\kappa). \end{aligned}$$

where $\Pi_i^* = (\mathbb{I} + \Pi)^i$ for $i = 0, \dots, t - 1$. Hence, κ_t follows a Gaussian distribution since it is a linear combination of several i.i.d Gaussian error terms.

In conclusion, under the assumptions of Proposition 1, we could prove that

$$\lim_{t \rightarrow \infty} \mathbb{E}\left[\frac{m_{x,t}^i}{m_{x,t}^j}|\theta\right] < \infty$$

Similarly, we can further prove that

$$\mathbb{E}\left[\left(\frac{m_{x,t}^i}{m_{x,t}^j}\right)^2|\theta\right] = \mathbb{E}\left\{\exp[2(\log(m_{x,t}^i) - \log(m_{x,t}^j))]| \theta\right\} < \infty,$$

which deduces that

$$\text{var} \left[\frac{m_{x,t}^i}{m_{x,t}^j} \middle| \theta \right] < \infty. \quad \square$$

Note that the above requirements for long-term coherence restricts our model towards one with constant β 's. Furthermore, unlike a standard LL approach, where the underlying co-movement is modelled via a multivariate random walk with drift, we assume that the κ_t^i 's are co-integrated I(1) processes. Those then boil Model 1 down to Model 2. However, Model 1 is much more flexible and might be able to provide more accurate short/medium-term forecasts than Models 2 and 3. In the next section, we discuss the use of Bayesian prior techniques in Model 1 to balance both the desirable long-term coherence and short-term data-specific dynamics.

2.4. Structural analysis

Structural analysis is commonly employed to investigate the interdependence of modelled response variables, especially in the macroeconomic literature (see, for example, Forni and Gambetti, 2010; Barigozzi *et al.*, 2021). Such an analysis is also applicable in our multi-population LC model to study the interdependence of mortality dynamics across sub-populations.

To implement a structural analysis, uncorrelated structural shocks need to be constructed first. In a VECM, a usual way is to consider the forecast errors (i.e., ξ_t in Equation (2.2)) as linear combinations of the structural shocks:

$$\xi_t = \Theta_0 u_t, \tag{2.8}$$

where u_t are usually assumed to be orthonormal white noises, that is, $u_t \sim N(0, \mathbb{I}_I)$ with \mathbb{I}_I being an identity matrix of size I . This normalization assumption implies that

$$\Sigma_\kappa = \Theta_0 \Theta_0'. \tag{2.9}$$

The structural shocks could then be explained as random, unexpected events which can influence the mortality rates but exogenous to the currently employed mortality model.

It is widely known that the structural shocks described above are not unique. A common practice is to derive a unique Θ_0 via a Cholesky decomposition of the variance-covariance matrix Σ_κ . This implies that the resulting structural model has a recursive structure. The recursive method identifies structural shocks by imposing short-run restrictions. (There also exists some alternative identification schemes for VAR and VECM via, e.g., long-run restrictions or sign restrictions. Please refer to Lütkepohl, 2005 and Kilian, 2013 for more details about structural VAR and structural VECM.) Specifically, structural shocks of one response variable can only contemporaneously affect variables that ranked after that response. Consequently, a meaningful (non-sample) information is usually needed for identifying the recursive order of the structural shocks.

2.4.1. Impulse Response Function

Impulse response function (IRF) is a popular type of structural analysis. Specifically, this measures the response of one mortality index to an impulse (i.e., an exogenous structural shock) of another. Based on the normalization condition (2.9), $\frac{\partial}{\partial u_t^j} \kappa_{t+h}^i$ is defined as the the h -step IRF of the response of i th population's mortality index κ_{t+h}^i to a one-standard deviation exogenous change in j th structural shock u_t^j .

To derive an analytical form of IRF, we can rewrite the VECM described in Equation (2.2) as

$$\kappa_{t+h} = \sum_{i=0}^{t+h-2} \Pi_i^* b + \sum_{i=0}^{t+h-2} (\Pi_i^* \Theta_0) u_{t+h-i},$$

where $\Pi_i^* = (\mathbb{I} + \Pi)^i$ for $i = 0, \dots, t + h - 1$. Hence, the aforementioned h -step IRF $\frac{\partial}{\partial u_t^i} \kappa_{t+h}^i$ is given by

$$\frac{\partial}{\partial u_t^i} \kappa_{t+h}^i = e_i'(\Pi_h^* B) e_j = e_i'[(\mathbb{I} + \Pi)^h \Theta_0] e_j,$$

where e_i denotes the i^{th} column of the identity matrix \mathbb{I}_t .

2.4.2. Forecast error variance decomposition

Another important component of structural analysis is the forecast error variance decomposition (FEVD). This metric decomposes the variance of forecast error into the contributions from specific exogenous structural shocks. Essentially, FEVD provides information on how much a structural shock contributes to variations of a particular response variable, and the dynamics of those contributions. Specifically, the proportion p_{ij}^h of the h -step forecast error variance of i th population’s mortality index explained by the j th structural shock u_t^j , is given by:

$$p_{ij}^h = \frac{\sum_{k=0}^{h-1} \{e_i'[(\mathbb{I} + \Pi)^k \Theta_0] e_j\}^2}{\sum_{j=1}^I \sum_{k=0}^{h-1} \{e_i'[(\mathbb{I} + \Pi)^k \Theta_0] e_j\}^2},$$

where the denominator $\sum_{j=1}^I \sum_{k=0}^{h-1} \{e_i'[(\mathbb{I} + \Pi)^k \Theta_0] e_j\}^2$ is the h -step forecast error variance of κ_{t+h}^i . For more details about FEVD in the VAR or VECM, please refer to Lütkepohl (2005).

2.5. A multi-population and multi-factor Lee–Carter model

In the previous sections, we mainly focus on extending the classical single-factor LC model to a multi-population specification (please refer to Model 1). It is possible to further incorporate multiple factors, which is discussed in this section.

Model 4 (The multi-population multi-factor LC model). Let $y_t^i = [\log(m_{x_0,t}^i), \dots, \log(m_{\omega,t}^i)]'$ for $i = 1, 2, \dots, I$, we assume that

$$\begin{aligned} y_t^i &= \alpha^i + \sum_{k=1}^p \beta_k^i \kappa_{kt}^i + \epsilon_t^i, \quad \epsilon_t^i \sim F_\epsilon(\cdot; \Omega) \\ \alpha^i &= [\alpha_{x_0}^i, \dots, \alpha_{\omega}^i]' \sim F_a(\cdot | \mu_a, \Sigma_a) \\ \beta_k^i &= [\beta_{k,x_0}^i, \dots, \beta_{k,\omega}^i]' \sim F_{k,b}(\cdot | \mu_{k,b}, \Sigma_{k,b}) \end{aligned} \tag{2.10}$$

with $F_a, F_{k,b}$ and F_ϵ denote some multivariate parametric distributions and

$$\begin{aligned} \text{cov}(\epsilon_t^i) &= \Omega, \quad \Omega = \text{diag}\{g_{x_0}, \dots, g_{\omega}\} \\ \mathbb{E}[\alpha^i] &= \mu_a, \quad \text{cov}(\alpha^i) = \Sigma_a \\ \mathbb{E}[\beta_k^i] &= \mu_{k,b}, \quad \text{cov}(\beta_k^i) = \Sigma_{k,b} \end{aligned}$$

for all $i = 1, 2, \dots, I$ and $k = 1, 2, \dots, p$. For simplicity, we assume the covariance-variance matrix Ω to be a diagonal matrix with g_k being the error variance for the k th age.

For the temporal movements of latent factors, if $\kappa_{kt} = [\kappa_{kt}^1, \dots, \kappa_{kt}^I]'$ is non-stationary, as investigated above, we consider a VECM such that

$$\Delta \kappa_{kt} = b_k + \Pi_k \kappa_{k,t-1} + \xi_{kt}, \quad \xi_{kt} \sim N(0, \Sigma_k^k), \tag{2.11}$$

where N denotes the multivariate Gaussian distribution. While if $\kappa_{kt} = [\kappa_{kt}^1, \dots, \kappa_{kt}^I]'$ is stationary, we consider a VAR model such that

$$\kappa_{kt} = b_k + B_k \kappa_{k,t-1} + \xi_{kt}, \quad \xi_{kt} \sim N(0, \Sigma_k^k), \tag{2.12}$$

Hence, the model parameter is

$$\theta = \{\{b_k\}_{k=1}^p, \{\Pi_k \text{ or } B_k\}_{k=1}^p, \{\Sigma_k^k\}_{k=1}^p, \mu_a, \{\mu_{k,b}\}_{k=1}^p, \Sigma_a, \{\Sigma_{k,b}\}_{k=1}^p, \Omega\}.$$

Similar to Model 1, as a special case of the linear dynamic factor model, we need to impose some additional constraints to identify dynamic factors. Following the discussion in Bai and Wang (2015), two identification assumptions are employed and stated below:

Assumption 3. For all $i = 1, \dots, I$ and $k = 1, \dots, p$,

- (a) $\alpha^i, \beta_k^i, \kappa_{kt}^i$ and ϵ_t^i are mutually independent
- (b) $\mathbb{E}[\epsilon_t^i] = 0$.

Assumption 4. We assume $\kappa_{k-1}^i = [\kappa_{k2}^i, \dots, \kappa_{kI}^i]'$ with $\kappa_{k1}^i = 0$, and

$$\beta_k^i \sim N_S(\mu_{k,b}, \Sigma_{k,b}) \text{ where } S = \{\beta_{k,x_0}^i = \dots = \beta_{k,x_{k-2}}^i = 0, \beta_{k,x_{k-1}}^i = 1\},$$

where N_S denotes a multivariate normal distribution truncated on a hyper-plane defined by S . That is, we constrain the first $k - 1$ elements of β_k^i to be zeros and the k th element to be one. Furthermore, we assume

$$\alpha^i \sim N(\mu_a, \Sigma_a)$$

and

$$\epsilon_t^i \sim N(0, \Omega),$$

where Ω is a diagonal matrix, and Σ_a and $\Sigma_{k,b}$ are general positive-semidefinite matrices.

Finally, we discuss the relevant conditions for coherent modelling of Model 4. The proof of these conditions follows straightforwardly from the the process delineated in the Section 2.3. As such, only the sufficient conditions to achieve coherence are presented below.

Proposition 2. Suppose that

1. Assumptions 3 and 4 are satisfied;
2. For $k = 1, \dots, p$, at least one of κ_{kt}^i is an $I(1)$ process for any $i \in \{1, 2, \dots, I\}$. And for such a κ_{kt}^i , $\kappa_{kt}^i - \kappa_{kt}^j$ should be weakly stationary for $i, j \in \{1, 2, \dots, I\}$ and $i \neq j$, and β_k^i are deterministic vectors with $\|\mu_{k,b}\| < \infty$ for all $i = 1, 2, \dots, I$ (i.e., $\Sigma_{k,b}$ is a zero matrix); and
3. The remaining κ_{kt}^i s are stationary, that is, $I(0)$ processes of any $i \in \{1, 2, \dots, I\}$.

the forecast ASDR produced by Model 4 are asymptotically coherent.

In particular, for a two-factor Model 4, it is plausible to assume that κ_{1t}^i is an $I(1)$ process characterized by mean-reverting $\kappa_{1t}^i - \kappa_{1t}^j$, while κ_{2t}^i is an $I(0)$ process. Thus, κ_{1t}^i can be described by a VECM, whereas κ_{2t}^i conforms to a VAR model.

3. Efficient Bayesian estimation

As discussed above, the co-integration relationship justifies the long-term coherence assumption. However, focusing on this long-run property only may be too strong to model short/medium-term mortality dependence. In this section, we provide a Bayesian method for Model 1 to allow for more flexibility in the short/medium term. The extension of the subsequent Bayesian method to a multi-factor model, specifically Model 4, is convenient. Consequently, a detailed repetition of this process will be forgone.

3.1. Imposing shrinkage priors

Prior specification of Bayesian inference has two components: the parametric class and the prior hyper-parameters given in the parametric family. For the prior parametric distribution, we adopt conditionally conjugate priors structure that implies known conditional density, which is common in the literature for dynamic factor models (see Chan and Jeliazkov, 2009; Bańbura *et al.*, 2010; Njenga and Sherris, 2020, for details). In particular, we set the distribution of the random effect parameters

$$\mu_a \sim N(m_1, P_1^{-1}), \mu_b \sim N(m_2, P_2^{-1})$$

and

$$\Sigma_a \sim IW(\phi_3 \mathbb{I}_N, \nu_3), \Sigma_b \sim IW(\phi_4 \mathbb{I}_N, \nu_4)$$

where IW denote the inverse wishart distribution. For the error variance, we set

$$g_x \sim IG\left(\frac{\phi_5^x}{2}, \frac{\nu_5^x}{2}\right)$$

for $x = x_0, \dots, \omega$. For the VECM coefficients, we consider

$$b \sim N(m_6, P_6^{-1}), \text{vec}(\Pi) \sim N(m_7, P_7^{-1})$$

and

$$\Sigma_\kappa \sim IW(\phi_8 \mathbb{I}_I, \nu_8).$$

Under this construction, all conditional distributions are tractable, and this avoids the use of non-smooth samplers such as the Metropolis–Hasting algorithm. One can employ the standard Gibbs sampling algorithm to estimate the posterior distribution, leading to the algorithm’s geometric convergence.

In Bayesian data analysis, the prior distribution usually demonstrates one’s prior beliefs, which could be independent of empirical data. For example, since mortality rates of all ages are expected to continue to decline in the future, we could set the mean value m_6 of mortality index’s drift term b as negative and the mean value m_2 of age effect μ_b to be positive. In other words, this prior structure effectively balances long-term belief (i.e., coherent mortality forecasts) and short-term empirical dynamics.

Bayesian shrinkage priors have been widely employed in the literature (Litterman, 1986) and are adopted in our estimation. To fulfil the first requirement in Proposition 1, that is, β^i should be a constant, one can set a small value of ϕ_4 for the prior of Σ_b . The second condition is equivalent to the fact that coefficient matrix Π in the VECM model of κ_t should be of a reduced rank. This is to shrink the movements of κ_t towards a co-integrated I(1) process. In particular, the rank of the coefficient matrix Π should be $I - 1$, or equivalently, Π should have a zero eigenvalue.

To specify the prior distribution of Π , we follow a similar procedure developed in Litterman (1986) and employ the Minnesota prior. The basic idea is to “center” the distribution of coefficients in $\mathbb{I}_I + \Pi$ so that the behaviour of each element in κ_t approximates a random walk with drift. Similarly, our prior belief that κ_t ’s are co-integrated over time could be formulated by setting the following moments for the prior distributions of the entries in Π :

$$\mathbb{E}[(\Pi_{ij})] = \begin{cases} -\lambda_1, & \text{if } i=j \\ \lambda_1, & \text{if } i=j-1 \ \& \ j > 1 \\ \lambda_1, & \text{if } i=J \ \& \ j=1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and } \mathbb{V}[(\Pi_{ij})] = \lambda_2^2 \quad (3.1)$$

where λ_1 and λ_2 are two hyper-parameters of the prior distribution of Π . The $\text{vec}(\Pi)$ is assumed to be normally distributed with a diagonal variance-covariance matrix. (As for a multi-factor Model 4, if κ_{kt} is a I(0) process, we can just assign $\mathbb{E}[(B_k)]$ as a null matrix.)

Roughly speaking, this prior specification assumes that κ_t^i is a weighted average of κ_{t-1}^i and κ_{t-1}^{i+1} with the weight λ_1 . In order to avoid the existence of explosive roots, the range of λ_1 should be between 0

Table 1. Hyperparameters used in the empirical analysis in Section 4.

μ_a	$m_1 = (-5)\mathbf{1}_N$	$P_1 = (0.1)^2 \mathbb{I}_N$
μ_b	$m_2 = (0.5)\mathbf{1}_{N-1}$	$P_2 = (0.1)^2 \mathbb{I}_{N-1}$
Σ_a	$\phi_3 = (0.01)^2$	$\nu_3 = N + 3$
Σ_b	$\phi_4 = (0.01)^2$	$\nu_3 = (N - 1) + 3$
g_x	$\phi_5 = (0.01)^2$	$\nu_5 = 3$
b	$m_6 = (-0.1)\mathbf{1}_I$	$P_6 = (0.01)^2 \mathbb{I}_I$
Π	m_7 : given by Equation (3.1)	P_7 : given by Equation (3.1)
Σ_κ	$\phi_8 = (0.1)^2$	$\nu_8 = I + 3$

and 1. In addition, the hyper-parameter λ_2 controls the overall tightness of the prior distribution and represents the relative importance of prior beliefs compared with data-specific information. When λ_2 increases, the prior beliefs become less informative, and the sample information will be more dominant.

For example, when $J = 3$, the prior belief is that

$$\begin{aligned}
 \begin{bmatrix} \Delta \kappa_t^1 \\ \Delta \kappa_t^2 \\ \Delta \kappa_t^3 \end{bmatrix} &= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 \\ 0 & -\lambda_1 & \lambda_1 \\ \lambda_1 & 0 & -\lambda_1 \end{bmatrix} \begin{bmatrix} \kappa_{t-1}^1 \\ \kappa_{t-1}^2 \\ \kappa_{t-1}^3 \end{bmatrix} + \begin{bmatrix} \omega_t^1 \\ \omega_t^2 \\ \omega_t^3 \end{bmatrix} \\
 &= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \lambda_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_{t-1}^1 \\ \kappa_{t-1}^2 \\ \kappa_{t-1}^3 \end{bmatrix} + \begin{bmatrix} \omega_t^1 \\ \omega_t^2 \\ \omega_t^3 \end{bmatrix} \\
 &= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \lambda_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} \kappa_{t-1}^2 - \kappa_{t-1}^1 \\ \kappa_{t-1}^3 - \kappa_{t-1}^2 \end{bmatrix} + \begin{bmatrix} \omega_t^1 \\ \omega_t^2 \\ \omega_t^3 \end{bmatrix}
 \end{aligned} \tag{3.2}$$

which is equivalent to expressing κ_t in the VAR form of

$$\begin{bmatrix} \kappa_t^1 \\ \kappa_t^2 \\ \kappa_t^3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} 1 - \lambda_1 & \lambda_1 & 0 \\ 0 & 1 - \lambda_1 & \lambda_1 \\ \lambda_1 & 0 & 1 - \lambda_1 \end{bmatrix} \begin{bmatrix} \kappa_{t-1}^1 \\ \kappa_{t-1}^2 \\ \kappa_{t-1}^3 \end{bmatrix} + \begin{bmatrix} \omega_t^1 \\ \omega_t^2 \\ \omega_t^3 \end{bmatrix}$$

Equation (3.2) is exactly a VECM form of κ_t with the co-integrating vectors being $[-1 \ 1 \ 0]$ and $[0 \ -1 \ 1]$. In other words, this prior specification supposes that $\kappa_t^i - \kappa_t^j$ is mean-reverting for any choice of i and j . Table 1 summarise all the hyperparameters that will be used in our empirical analysis.

Alternatively, instead of using shrinkage prior, it is also possible to restrict parameters of Model 1 to satisfy the long-term coherence assumptions. For instance, we can make use of Model 2 in which age effect $\beta_i = \beta$ are assumed to be the same across different populations. Additionally, in the VECM form of κ_t , the coefficient Π can be written as a matrix product \mathbf{cd}' . Then we can apply the co-integration relations (i.e., $\kappa_t^i - \kappa_t^j$ is weakly stationary) by setting the co-integration matrix \mathbf{d}' as a $(I - 1) \times I$ matrix

$$\begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

It can be shown that such specifications produce the long-term coherent forecasts. However, as will be discussed in Section 4, this would be too restrictive for short/medium-term forecasting performance. Instead, imposing Bayesian prior to Model 1 is therefore a more flexible method.

3.2. The proposed MCMC algorithm

In this section, we discuss a block-based Gibbs sampler to estimate the proposed multi-population LC models. Note that high-dimensional latent random states are to be sampled (i.e., the α^i 's, β^i 's and κ_t 's), in addition to the model parameters. Given the model parameters, random effects α^i and β^i can be sampled iteratively, and the VECM/VAR can be obtained via the Kalman Filter. However, due to the high dimensionality of mortality data, a naive implementation of such a sampler will lead to extensive computational costs. In this section, we present an efficient precision sampler.

3.2.1. Jointly sampling the random age effects

As the discussion in Chan and Jeliakzov (2009), the efficiency of MCMC can be greatly improved, if the number of blocks could be reduced. Therefore, the following joint conditional posterior of all the random age effects is developed, which allows us to sample them as a whole block.

First, rewrite the general model as

$$\mathbf{Y}_t = A + \boldsymbol{\beta} \text{diag}(\boldsymbol{\kappa}_t) + \boldsymbol{\epsilon}_t$$

where $\mathbf{Y}_t = [y_t^1, \dots, y_t^I]$, $A = [\alpha^1, \dots, \alpha^I]$, $\boldsymbol{\beta} = [\beta^1, \dots, \beta^I]$, $\boldsymbol{\epsilon}_t = [\epsilon_t^1, \dots, \epsilon_t^I]$. Since $\beta_{x_0}^i = 1$ for $i \in I$, $\boldsymbol{\beta}$ can be expressed as $[\mathbf{1}_I, \boldsymbol{\beta}'_{(-1)}]'$. Based on this, we can sample

$$\text{vec}(A) | (\boldsymbol{\beta}, \theta, \mathbf{Y}, \boldsymbol{\kappa}) \sim N(\tilde{\boldsymbol{\mu}}_a, K_a^{-1})$$

with

$$K_a = T\mathbb{I}_I \otimes \Omega^{-1} + \mathbb{I}_I \otimes \Sigma_a^{-1}$$

$$\tilde{\boldsymbol{\mu}}_a = K_a^{-1} \left[\text{vec} \left(\Omega^{-1} \sum_{t=1}^T (\mathbf{Y}_t - \boldsymbol{\beta} \text{diag}(\boldsymbol{\kappa}_t)) \right) + \mathbf{1}_I \otimes (\Sigma_a^{-1} \boldsymbol{\mu}_a) \right]$$

and

$$\text{vec}(\boldsymbol{\beta}_{(-1)}) | (A, \theta, \mathbf{Y}, \boldsymbol{\kappa}) \sim N(\tilde{\boldsymbol{\mu}}_b, K_b^{-1})$$

where

$$K_b = \sum_{t=1}^T \text{diag}(\boldsymbol{\kappa}_t^2) \otimes \Omega_{(-1)}^{-1} + \mathbb{I}_I \otimes \Sigma_b^{-1}$$

$$\tilde{\boldsymbol{\mu}}_b = K_b^{-1} \left[\text{vec} \left(\Omega_{(-1)}^{-1} \sum_{t=1}^T (\mathbf{Y}_t^{(-1)} - A_{(-1)}) \text{diag}(\boldsymbol{\kappa}_t) \right) + \mathbf{1}_I \otimes (\Sigma_b^{-1} \boldsymbol{\mu}_b) \right].$$

In this formulation, $\Omega_{(-1)}$ means Ω without the first row and column. $\boldsymbol{\beta}_{(-1)}$, $\mathbf{Y}_t^{(-1)}$ and $A_{(-1)}$, respectively, represent $\boldsymbol{\beta}$, \mathbf{Y}_t and A with the first rows removed. This is equivalent to sampling them independently for each population, yet avoiding using a loop.

3.2.2. Precision sampler for all the κ 's

The sampling of all the κ_t 's usually involves the Kalman filter. In this paper, we develop the precision sampler that allows us to sample all the latent time-series jointly. This improves the transparency in the integrated likelihood and, consequently, allows for the associated model comparison. Recall that for the

identification purpose, we have set $\kappa_1 = 0$. Hence, sampling is only required for $\kappa = [\kappa'_2, \dots, \kappa'_T]$. For the proposed model, we can rewrite

$$\mathbf{Y} = \mathbf{1}_{T-1} \otimes \text{vec}(A) + \mathcal{B}\kappa + \epsilon$$

where $\mathbf{Y} = [\text{vec}(\mathbf{Y}_2)', \dots, \text{vec}(\mathbf{Y}_T)']'$, $\epsilon = [\text{vec}(\epsilon_2)', \dots, \text{vec}(\epsilon_T)']'$, and

$$\mathcal{B} = \mathbb{I}_{T-1} \otimes ((\mathbb{I}_I \otimes \boldsymbol{\beta})M_I)$$

where M_I is a matrix of dimension $I^2 \times I$ that first converts the column vector into a diagonal matrix and then vectorizes it, that is,

$$M_I \kappa_i = \text{vec}(\text{diag}(\kappa_i)).$$

In a similar fashion, we can write the VECM of latent factors as

$$H\kappa = \mathbf{1}_{T-1} \otimes b + \boldsymbol{\xi}$$

where $\boldsymbol{\xi} = [\xi'_2, \dots, \xi'_T]'$ and

$$H = \begin{bmatrix} \mathbb{I}_I & \mathbf{0} & \dots & \dots & \dots \\ -(\mathbb{I}_I + \Pi) & \mathbb{I}_I & \mathbf{0} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & -(\mathbb{I}_I + \Pi) & \mathbb{I}_I \end{bmatrix}$$

Hence, we derive the precision sampler

$$\kappa | (\theta, \mathbf{Y}, \boldsymbol{\beta}, A) \sim N(\mu_k, K_k^{-1})$$

with

$$K_k = H'(\mathbb{I}_{T-1} \otimes \Sigma_k^{-1})H + \mathcal{B}'(\mathbb{I}_{(T-1)I} \otimes \Omega^{-1})\mathcal{B}$$

$$\mu_k = K_k^{-1}[\mathcal{B}'(\mathbb{I}_{(T-1)I} \otimes \Omega^{-1})(\mathbf{Y} - \mathbf{1}_{T-1} \otimes \text{vec}(A)) + H'(\mathbf{1}_{T-1} \otimes (\Sigma_k^{-1}b))].$$

The most significant advantage of our precision sampler is that the precision matrix K_k is a symmetric block-banded matrix with very few non-zero elements. In Section A of Supplementary Material, we provide a simplified example of κ_i 's precision matrix with red points representing non-zero elements in the matrix. Our sampling algorithm's exact computational advantages are also demonstrated via simulation explorations in Section A of Supplementary Material. The presented precision matrix there is an intermediate product during the simulation, when the empirical analysis detailed in Section 4 is conducted. It can be seen that only a small number of non-zero entries falling in a narrow band around the precision matrix's diagonal. From a computational point of view, this implies that we could reduce storage and computational costs by exploiting efficient algorithms designed for the sparse matrix. More details regarding the precision sampler can be found in Chan and Jeliazkov (2009).

4. Empirical application to the modelling of G7 mortality data

In this section, we examine the mortality data of the G7 countries, which are sourced from Human Mortality Database (2019). In a related study, the empirical results of Tuljapurkar *et al.* (2000) with data over 1950–1994 suggest a universal decline in mortality across all age groups in the G7 populations. They alluded that this trend places a constraint on any theory of society-driven mortality decline and provides a basis for stochastic mortality forecasting via the LC-type model. Instead of independently fitting LC models, this section presents a comprehensive analysis of the G7 mortality data based on our proposed multi-population LC models.

Table 2. *Logarithm of marginal likelihoods for different mortality models.*
 (a) *Marginal log-likelihood for the single-factor mortality models.*

λ_2	Model 1	Model 2	Model 3	Lee-Carter	Li & Lee
0.01	51963	45786	26425		
0.1	51980	45787	26429	54890	30830
0.00001	51961	45778	26423		

(b) *Marginal log-likelihood for the two-factor mortality models.*

λ_2	Model 4	LC	Li & Lee
0.01	62949		
0.1	62944	65254	55676
0.00001	62946		

4.1. Empirical data set

We use the crude (un-smoothed) annual data for the period 1956–2019 for all the G7 countries. For each country, age-sex-specific death rates are available annually, from age 0 to age 110. Since mortality measures at very old ages are unreliable (Lee and Carter, 1992), we constrain the maximum age as 89, such that $\omega = 89$ as in the model. Also, male and female mortality rates are combined for the following analyses.

4.2. Model comparison: preliminary results

To demonstrate the usefulness of our model, we first undertake a preliminary comparative analysis of a range of popular single-factor and two-factor models. For the single-factor case, our consideration encompasses Model 1, Model 2, Model 3, the LC model (Lee and Carter, 1992), and the single-factor Li & Lee model (Li and Lee, 2005). With respect to the two-factor models, we compare the performance of Model 4, the two-factor LC model, and the two-factor Li & Lee model (Li and Lee, 2005).

To obtain comparable results, both the LC models and the Li & Lee models have been reformulated according to the state-space representations suggested by Pedroza (2006) (please refer to Section B of Supplementary Material for details). In particular, we consider the marginal likelihood as the basis for model comparison, which is a standard technique in Bayesian analysis (Koop, 2003). It is worth mentioning that evaluating the marginal likelihood is usually a computationally challenging task. In practice, the most commonly used Bayesian information criteria (or BIC) approximates twice the log of the marginal likelihood (Schwarz, 1978). To address the computational issue, Newton and Raftery (1994) proposed a simple way to calculate marginal likelihood by using the posterior harmonic mean of the likelihood. Please refer to the Section C of Supplementary Material for more details. To more comprehensively compare the prediction accuracy of each model, we present the out-of-sample forecasting results in Section 4.4.

In Table 2, we present the marginal log-likelihoods for Model 1–4, the LC model, and the Li & Lee model. For Model 1–3, we consider three distinct values of the hyper-parameter λ_2 : 0.01 (moderate), 0.1 (weak), and 0.00001 (strong). These same values of λ_2 are also considered for the first factor in Model 4.

In general, two-factor models significantly outperform single-factor mortality models in terms of the marginal likelihood. Geweke and Amisano (2011) demonstrated that the in-sample marginal likelihood is intimately connected to the one-step-ahead predictive likelihood, thereby making it a good measure of short-term forecasting accuracy. Thus, the two-factor models showcase superior short-term forecasting capabilities compared to their single-factor counterparts.

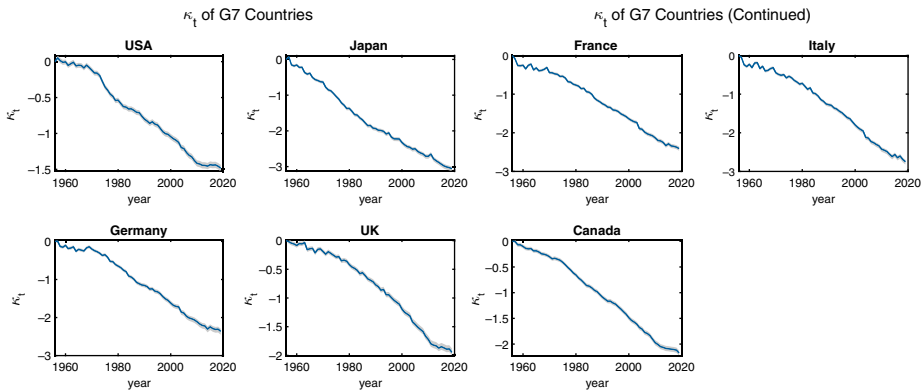


Figure 2. Temporal plots of estimated mortality index κ_t^i for all the G7 countries (solid line: posterior mean of; grey area: 99% credible interval).

Among all the single-factor models, the LC model exhibits the greatest marginal likelihood. In contrast, the Li & Lee model only outperforms Model 3, ranking it as the second-worst model. Among Model 1–3, Model 1, which incorporates heterogeneous random age effects, has the highest marginal likelihood. Conversely, Model 3, characterized by homogeneous age effects, has the lowest marginal likelihood. The marginal likelihood of Model 2 considerably surpasses Model 3, although it remains inferior to Model 1. Therefore, even though Model 2 satisfies long-term coherence conditions, Model 1 emerges as the preferred model when we focus on its superior short-term predictive performance. Moreover, Model 1 exhibits greater flexibility than Model 2, considering that Model 2 is actually a special case of Model 1 when Σ_b converges to a zero matrix. Essentially, Model 1, employing a Bayesian shrinkage prior, more appropriately balances the trade-off between short to medium-term predictive accuracy and long-term coherence. The superior short- to medium-term predictive accuracy is attributable to the heterogeneous random age effects, while the long-term coherence is approximately attained by imposing a Bayesian shrinkage prior. Furthermore, in the selection of hyper-parameters, the optimal choice is determined to be $\lambda_2 = 0.1$ among three distinct λ_2 values.

Similarly, when considering two-factor mortality models, the two-factor LC model has the highest marginal likelihood. Conversely, the two-factor Li & Lee model results in the lowest marginal likelihood, which suggests comparatively less accurate short-term forecasting performance. Regarding the selection of the hyper-parameter λ_2 for Model 4, it is discerned that $\lambda_2 = 0.01$ is the optimal choice.

4.3. Fitted in-sample results and inferences

4.3.1. The temporal trend of mortality rates

In this subsection concerning in-sample results, our primary focus is on the single-factor model. Specifically, we utilize Model 1, setting $\lambda_2 = 0.1$ for the in-sample analyses within this subsection. This selection is premised on its superior short-term forecasting performance relative to other competitive specifications. (Please refer to the Section E of Supplementary Material for the in-sample analyses for Model 4.) Figure 2 exhibits the temporal plot of fitted κ_t^i separately for each country. The solid line represents the posterior mean of κ_t^i , and the grey area depicts the corresponding 99% credible band. It appears that κ_t^i has a persistent declining trend and thus indicates a non-stationary process. Those posterior means are consistent with the downward trends of historical mortality data of the G7 countries. The narrow widths of the credible band imply that the estimation of κ_t is reliable.

To compare the differences in mortality declines over years, we plot all the estimated posteriors means of κ_t^i 's in Figure 3. Despite some slowed mortality improvements over recent periods, the overall patterns

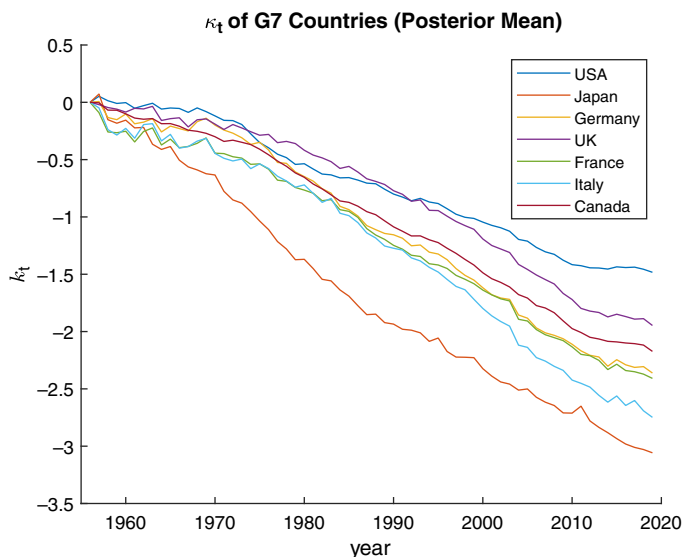


Figure 3. Comparison of estimated mortality index κ_t^i for all the G7 countries.

might be roughly fitted by linear trends, especially for those post-1970s. According to the behaviours of κ_t^i , G7 countries can be divided into three distinct groups:

1. Japan has the lowest mortality index across all seven countries, and the Japanese κ_t^i declines at almost a constant rate (linear pattern) over the examined six decades.
2. Canada, France, (West) Germany, Italy and the UK: Although their κ_t^i 's decline rates are substantially lower than the Japanese counterpart before 1980s, all those countries' κ_t^i 's tend to exhibit similar speeds of decline after 1990. Figure 3 shows that the estimated κ_t^i 's of the six countries are almost parallel to each other over 1990–2019.
3. USA has the highest mortality rates among G7 countries over the sample period. Unlike the other G7 countries, the marginal decline rate of the USA's κ_t^i decelerates, especially over most recent years. A flat curve is displayed since 2010. This somewhat deviates from the rest G7 countries' overall temporal trend.

4.3.2. The age effects of mortality rates

In Figure 4, we plot the estimated age effects, that is, μ_a and μ_b , respectively. Recall that the heterogeneous age effects in Model 1 are drawn randomly from a common distribution characterised by μ_a and μ_b , to ensure the coherence. For age x , μ_a^x represents the (common) first-year (i.e., 1956) level of log mortality rate, and μ_b indicates the (common) age-specific loading of κ_t^i . From the widths of 99% credible bands shown in Figures 4 and 5, the estimation uncertainty is relatively lower for μ_a than for μ_b . The age pattern of μ_a has a classic 'tick' shape, which declines from age 0 to reach a minimum around age 12. The pattern then almost uniformly increases, except for a famous 'accidental-hump' over ages 15–25. The estimates of μ_b^x are also consistent with empirical findings, such that mortality declines are faster at young ages than at very old ages.

In addition to μ_a and μ_b , Figures 5 and 6 present estimated age effects α^i and β^i for each country, respectively. Recall that those are the unique feature of Model 1 to allow flexible population-wise age effects. Despite some minor differences, all the α^i 's demonstrate rather similar patterns. More differences can be observed for the country-dependent β^i 's, suggesting various age-specific decline speeds across

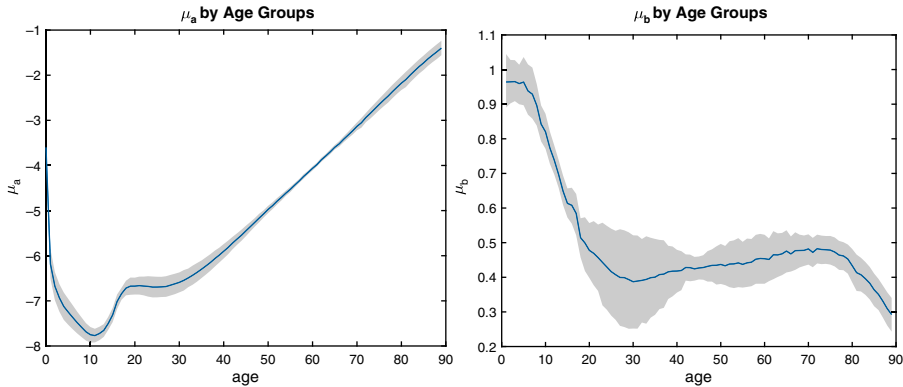


Figure 4. Estimated age effects μ_a and μ_b (solid line: posterior mean; grey area: 99% credible interval).

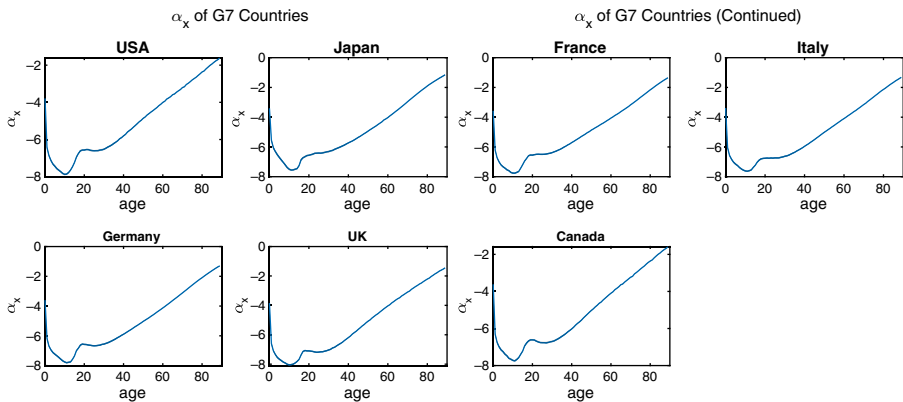


Figure 5. Estimated age effects α^i 's (solid line: posterior mean; grey area: 99% credible interval).

the G7 countries. In Figure 7, we plot the posterior means of α^i 's and β^i 's for all the G7 countries together to facilitate comparison. The population-specific variations are demonstrated by differences among the corresponding curves. This validates the effectiveness of our specification to enable heterogeneous age effects in the Model 1. Again, we observe that α^i 's of all the countries are relatively close to each other, whereas the β^i 's are more heterogeneous, especially for ages over 20–40 and 50–80.

4.3.3. Cross-sectional dynamic structure

We now investigate relationships among κ_i in the VECM, characterised by the coefficient matrix Π . In Table 3, posterior means of parameter Π are presented, together with their corresponding standard errors. Those estimates could provide more information about interdependence of temporal factors in each country. We firstly verify the impact of prior belief of long-term coherence on the coefficient matrix Π . Specifically, recall that our prior belief suggests that $\kappa_i^i - \kappa_i^j$ is weakly stationary for any $i \neq j$. It can be seen that most of the diagonal elements of Π are shrunk to $-\lambda_1$ (-0.1), and most of the first super-diagonal elements are also close to λ_1 (0.1). Most of the remaining cross-sectional relationships are insignificant, since a small λ_2 (0.1) is adopted to be consistent with our prior belief of the long-term coherence.

In Figure 8, we report posterior distributions of eigenvalues' modulus computed from the simulated VECM coefficient Π . Recall that to meet the coherence conditions, the Minnesota-type prior is adopted

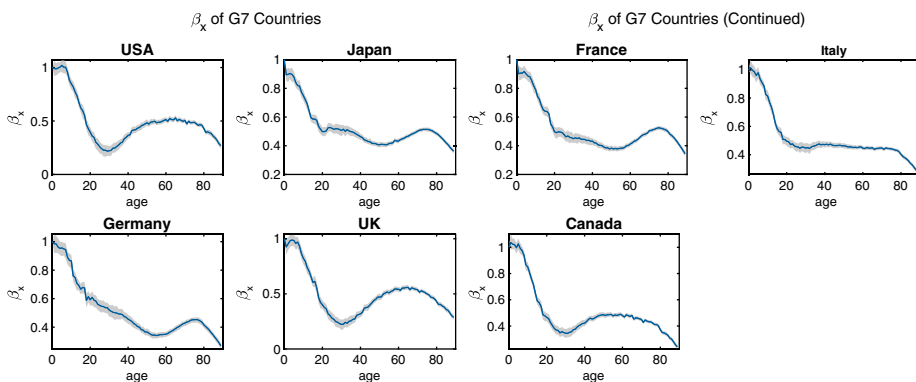


Figure 6. Plots of estimated age effects β^i 's (solid line: posterior mean; grey area: 99% credible interval).

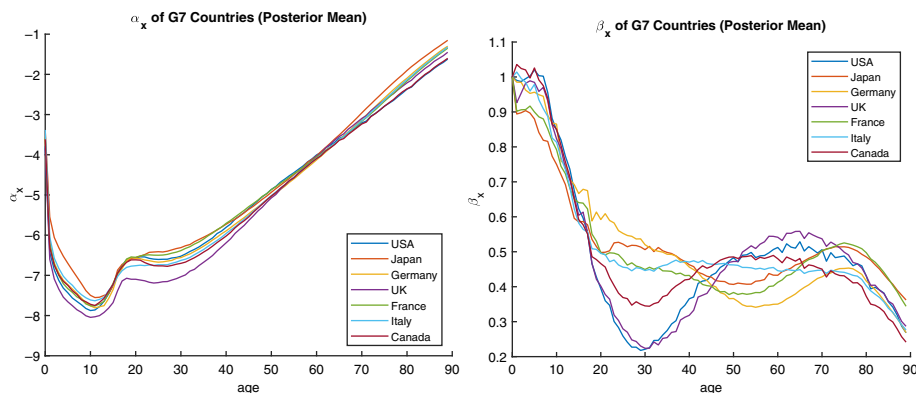


Figure 7. Comparison of estimated age effects α^i 's and β^i 's for all G7 countries.

such that the coefficient Π has a reduced rank. Figure 8 shows that the smallest eigenvalue (measured by modulus) of Π is close to 0. Hence, the VECM coefficient Π is indeed of rank $I - 1$. This supports the desirable co-integration of temporal factors, and thus all the countries' mortality rates will not diverge in the long run (Li and Lu, 2017).

4.3.4. Structural analysis

As stated in Section 2.4, a meaningful recursive relationship is needed to identify unique structural shocks. In this section, since GDP is believed an important factor on the mortality improvements (Boonen and Li, 2017), we choose an order according to G7 countries' total GDP as of 2019, which are retrieved from World Bank Open Data (<https://data.worldbank.org>). Specifically, the response variables are ranked as USA, Japan, Germany, UK, France, Italy and Canada. Thus, this recursive structure implies that the US structural shock could contemporaneously affect mortality indices of others, while the shock of Canadian population cannot affect any of other contemporaneous mortality indices.

In Figure 9, we plot the responses of all the mortality indices to the US mortality shock of one standard deviation (around 3%). The solid blue lines represent posterior means of IRFs, and the grey areas and dashed red lines correspond to 68% and 95% credible intervals, respectively (It is common to use 68% and 95% credible intervals in the macroeconometrics literature, for example, see baumeister2019structural). Since mortality rates are consistently improving (reducing), our results suggest that

Table 3. Estimated coefficient matrix Π in the VECM of κ_t (with standard errors displayed in parentheses).

	USA	Japan	Germany	UK	France	Italy	Canada
USA	-0.1789* (0.009)	0.0997* (0.0023)	-0.0000 (0.0005)	-0.0000 (0.0004)	0.0000 (0.0003)	0.0000 (0.0002)	-0.0000 (0.0002)
Japan	-0.1183 (0.0966)	-0.0049 (0.0445)	0.0881* (0.0196)	-0.0065 (0.0159)	-0.0036 (0.0133)	-0.0039 (0.011)	-0.0010 (0.0099)
Germany	0.1701 (0.1188)	0.0179 (0.0433)	-0.1507* (0.0647)	0.0023 (0.059)	0.0123 (0.0579)	0.0048 (0.0466)	0.0055 (0.047)
UK	0.0478 (0.1062)	-0.0121 (0.0385)	0.0143 (0.0713)	-0.1981* (0.0674)	0.0780 (0.0676)	0.0414 (0.0552)	0.0248 (0.0566)
France	0.0915 (0.1296)	0.0360 (0.0476)	0.0388 (0.0817)	0.0776 (0.0784)	-0.2859* (0.0796)	0.0510* (0.065)	0.0128 (0.0673)
Italy	-0.0948 (0.1442)	0.0511 (0.0538)	0.1021 (0.0951)	0.1794* (0.0921)	-0.0555 (0.0951)	-0.2808* (0.0782)	0.1308* (0.0798)
Canada	0.1289* (0.0785)	0.0238 (0.032)	0.0340 (0.0678)	-0.0023 (0.0793)	0.0070 (0.0862)	0.0203 (0.0737)	-0.1876* (0.0802)

Note: * 0 is outside the 90% credible interval;

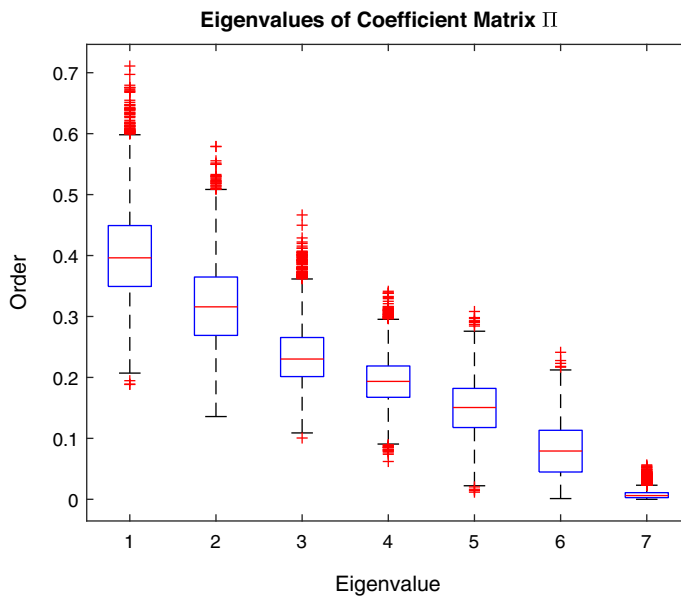


Figure 8. Posteriors of eigenvalues' modulus of the simulated coefficient matrix Π (ordered by the size of modulus).

a reduction in the US mortality rates may significantly lead to permanent decline of mortality rates in other G7 countries. For the contemporaneous impact, the point estimates of IRF range from 1 to 3%, whereas the influence at the 50th step varies within roughly the same range.

In Figure 10, the dynamics of FEVD of the US mortality index contributed by structural shocks of G7 populations are plotted. First, we observe that contributions of shocks of US mortality rates constantly reduce over time and achieves a minimum just below 30% at the 50th step. As for the contributions of other G7 countries, shocks of Japanese mortality rates is the largest (around 60% at the 50th step),

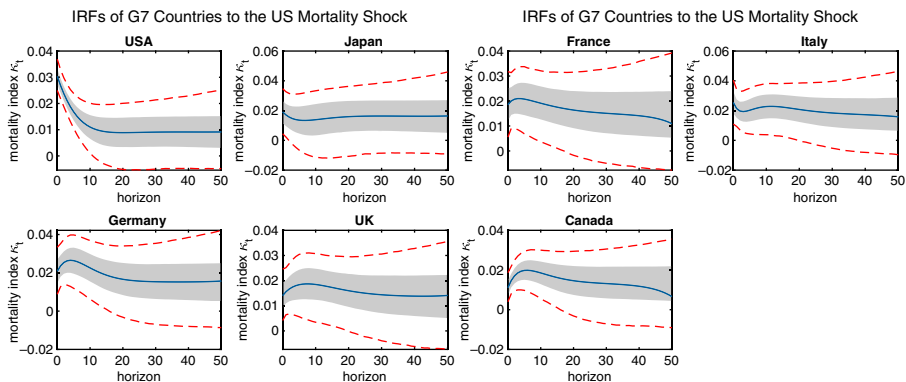


Figure 9. IRFs of G7 countries' mortality indices to a one standard deviation US mortality shock. (blue solid line: posterior mean; grey area: 68% credible interval; red dashed line: 95% credible interval).

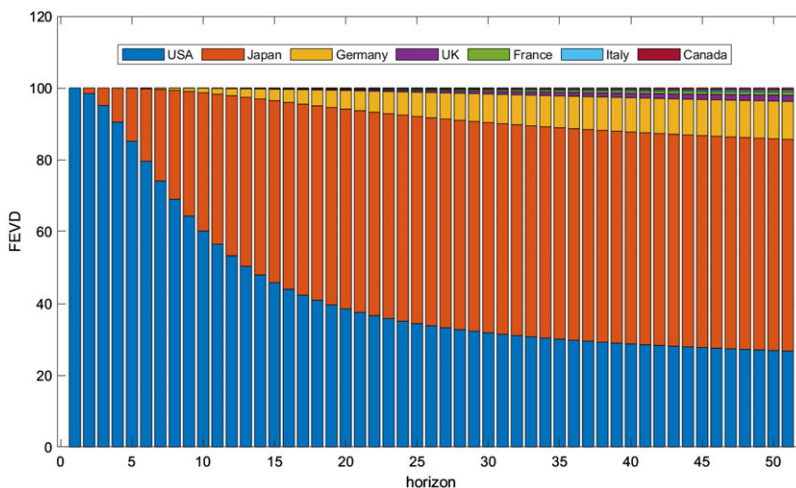


Figure 10. FEVD of the US mortality index with relative contributions of all the population-specific structural shocks (posterior means).

followed by those of Germany population (around 10% at the 50th step). This is consistent with the assumed recursive order that Japan and German are the second and third largest economies, respectively, among all G7 countries.

Interpreting the impacts of mortality rates among different countries is typically challenging due to the multitude of influencing factors. Despite this, changes of Japan's mortality rates could potentially influence US mortality rates through the following channels, given the tight social and economic connections between the two countries.

First, a significant shift in Japan's mortality rates could indicate a new health trend, disease emergence or healthcare breakthrough. If these factors are globally pervasive or if the associated research is disseminated worldwide, they could potentially impact US mortality rates. For instance, if a medical breakthrough originates in Japan and later adopted by the US, it could potentially reduce the US mortality rates. Given that Japan has the highest life expectancy worldwide, it is plausible that Japanese mortality rates serve as an upper limit for other countries.

Second, as a significant global economy and a key trading partner of the US, shocks to Japan’s mortality rates could affect the size and productivity of its workforce, which could have a long-term economic impacts. Those impacts could then lead to non-negligible impacts to the US economy, which eventually affects factors such as the healthcare conditions and thus influence the US mortality rates. Specifically, a significant rise in mortality among Japan’s working-age population could have negative impacts on labour economics and thus reduce its economic outputs. Consequently, this might introduce economic downturns in the US, which then lead to increased mortality rates due to factors such as increased stress level and reduced healthcare spending.

4.4. Out-of-sample forecasting results

The out-of-sample forecasts of our proposed models can be obtained using the posterior predictive distributions of log mortality rates y_{T+s} , where $s = 1, 2, \dots, h$, with h the largest forecasting step. This will be based on the simulations of latent random states and model parameters. Please see the Section D of Supplementary Material for more details on the simulation of posterior predictive distributions.

4.4.1. Out-of-sample forecasting performance comparison

In this section, we evaluate and compare forecasting performances of popular mortality models as considered in Section 4.2, including Models 1–4, the LC model and Li & Lee model. Our complete sample period is from 1956 to 2019, with the training sample spanning 1956–2009 and the test sample covering 2010–2019. At each step, we use only the data up to time t , denoted as \mathcal{F}_t , to obtain the posterior predictive density of the h -step ahead forecast y_{t+h} .

Consistent with existing studies, such as Li and Lu (2017) and Li and Shi (2021a), the popular metric root mean squared forecast error (RMSFE) is employed to evaluate forecasting accuracy. Denote $\mathbb{E}(y_{x,T+h}|\mathcal{F}_T)$ as the h -step-ahead point forecast (posterior expectation) for age x at year T , the RMSFE for the next h years is defined as

$$\text{RMSFE}(h) = \text{RMSFE}^{(i)}(h) = \frac{1}{n} \sum_{i=1}^n \sqrt{\frac{\sum_{x=0}^N \sum_{t=t_0}^{T-h} [y_{x,t+h}^o - \mathbb{E}(y_{x,t+h}|\mathcal{F}_t)]^2}{(N+1)(T-h-t_0+1)}}$$

where $y_{x,T+t}^o$ is the observed values of log mortality rate $\log(m_{x,T+t})$ for age x at year $T + t$.

Table 4 provides a comprehensive comparison of the RMSFEs for all fitted single-factor mortality models. The RMSFEs are calculated at different forecast horizons, ranging from 1 to 10, and three various shrinkage hyperparameters ($\lambda_2 = 0.1, 0.01, 0.00001$) are used for Models 1–3. Our results suggest that the forecasting performance varies with the forecast horizon and the shrinkage hyperparameter. Overall, the single-factor LC model demonstrates the best forecasting performance, while the common factor Li & Lee model substantially underperforms others. Despite this, it is worth noting that the LC model achieves a higher forecasting accuracy at the cost of producing incoherent forecasts in a longer term, whereas the single-factor Li & Lee model achieves the long-run coherence with the least accurate forecasts. Our proposed models, namely Model 1–3, manage to balance the trade-off between short-term forecasting performance and long-term coherence. Across different shrinkage hyperparameters and among the three proposed specifications, Model 1 is the best performing candidate in terms of forecasting accuracy.

Table 4 also indicates that the RMSFEs of the three proposed models are influenced by the shrinkage hyperparameters. Regardless of the model used, both weak ($\lambda_2 = 0.1$) and strong ($\lambda_2 = 0.00001$) priors outperform the moderate prior ($\lambda_2 = 0.01$). This may be attributed to the flexibility of the weak prior and the strong prior’s ability to capture the long-term trend. Conversely, the moderate prior lacks these two advantages, resulting in poor forecasting performance. In particular, it is worth mentioning that Model 1 with a weak prior performs as the second best model across all competitors (including LC and Lee & Li models).

Table 4. RMSFEs of single-factor mortality models with different shrinkage hyperparameters and forecast horizons.

Horizons	Model 1			Model 2			Model 3			LC	Li & Lee
	0.1	0.01	0.00001	0.1	0.01	0.00001	0.1	0.01	0.00001		
λ_2										/	/
$h = 1$	0.1219	0.1252	0.1214	0.1364	0.1398	0.1384	0.1785	0.1777	0.1769	0.1179	0.1732
$h = 2$	0.1312	0.1409	0.1312	0.1436	0.1538	0.1482	0.1833	0.1840	0.1820	0.1266	0.1808
$h = 3$	0.1383	0.1567	0.1395	0.1486	0.1682	0.1565	0.1849	0.1880	0.1854	0.1332	0.1890
$h = 4$	0.1493	0.1799	0.1514	0.1579	0.1911	0.1686	0.1895	0.1977	0.1921	0.1430	0.1978
$h = 5$	0.1595	0.2064	0.1622	0.1669	0.2184	0.1799	0.1945	0.2103	0.1993	0.1512	0.2059
$h = 6$	0.1688	0.2343	0.1715	0.1754	0.2477	0.1895	0.1989	0.2237	0.2056	0.1587	0.2124
$h = 7$	0.1811	0.2660	0.1838	0.1859	0.2817	0.2016	0.2041	0.2394	0.2136	0.1686	0.2211
$h = 8$	0.1881	0.2959	0.1906	0.1926	0.3161	0.2085	0.2060	0.2532	0.2174	0.1733	0.2251
$h = 9$	0.1952	0.3326	0.2013	0.1995	0.3574	0.2199	0.2077	0.2761	0.2235	0.1798	0.2330
$h = 10$	0.1977	0.3610	0.2104	0.2034	0.3908	0.2280	0.2112	0.2949	0.2297	0.1831	0.2378

Table 5. RMSFEs of two-factor mortality models with different shrinkage hyperparameters and forecast horizons.

Horizons	Model 4			Lee–Carter	Li & Lee
	0.1	0.01	0.00001		
λ_2				/	/
$h = 1$	0.1082	0.1077	0.1071	0.1018	0.1173
$h = 2$	0.1212	0.1205	0.1201	0.1120	0.1271
$h = 3$	0.1325	0.1325	0.1318	0.1205	0.1369
$h = 4$	0.1473	0.1484	0.1462	0.1339	0.1488
$h = 5$	0.1612	0.1636	0.1587	0.1494	0.1631
$h = 6$	0.1736	0.1795	0.1689	0.1713	0.2030
$h = 7$	0.1885	0.1983	0.1816	0.2021	0.2987
$h = 8$	0.1981	0.2137	0.1889	0.2430	0.7311
$h = 9$	0.2097	0.2118	0.1999	0.3208	1.5741
$h = 10$	0.2134	0.2060	0.2085	0.4188	5.5744

In Table 5, we further compare the forecasting performances of three two-factor mortality models: Model 4, the two-factor LC model and the augmented common factor Li & Lee model. Among them, Model 4 is the best performing model. When compared to single-factor counterparts, the improvements achieved by the two-factor models are primarily evident in short-term forecasting, particularly when the forecast horizon h is small. When the forecast horizon expands, Model 4 exhibits forecasting accuracy comparable to that of its single-factor counterpart. However, for the LC model and the Li & Lee model, the forecasting performances significantly deteriorate when $h > 5$. This may be attributed to the hierarchical structure employed by Model 4, which results in fewer parameters than the LC and Li & Lee models and consequently reduces the possibility of overfitting. Finally, unlike the single-factor models, we find that the choice of hyperparameters does not have substantially affect the forecasting performance of Model 4.

Apart from the point forecasts, prediction intervals with high coverage are usually more important for mortality models to be used in actuarial practices. In Tables 6 and 7, we present the coverage ratios of the 95% prediction intervals of single-factor models and two-factor models, respectively. The average widths of the prediction intervals (to demonstrate the efficiency) are displayed in the Section F

Table 6. Coverage ratios of the 95% prediction intervals produced by the single-factor mortality models.

Horizons	Model 1			Model 2			Model 3			Lee-Carter Li & Lee	
	0.1	0.01	0.00001	0.1	0.01	0.00001	0.1	0.01	0.00001	/	/
λ_2											
$h = 1$	0.8737	0.8630	0.8640	0.8708	0.8633	0.8684	0.9181	0.9205	0.9178	0.8629	0.8500
$h = 2$	0.8621	0.8337	0.8407	0.8665	0.8432	0.8617	0.9086	0.9189	0.9044	0.8568	0.8340
$h = 3$	0.8520	0.7937	0.8252	0.8629	0.8139	0.8423	0.8946	0.9163	0.8956	0.8500	0.8139
$h = 4$	0.8354	0.7347	0.8052	0.8528	0.7612	0.8070	0.8782	0.9120	0.8862	0.8361	0.7900
$h = 5$	0.8196	0.6783	0.7804	0.8497	0.6966	0.7786	0.8619	0.9132	0.8757	0.8204	0.7743
$h = 6$	0.8187	0.6149	0.7635	0.8460	0.6260	0.7559	0.8467	0.9143	0.8740	0.8108	0.7648
$h = 7$	0.8099	0.5611	0.7317	0.8433	0.5671	0.7349	0.8325	0.9115	0.8675	0.7952	0.7571
$h = 8$	0.8127	0.5106	0.7212	0.8508	0.5021	0.7296	0.8148	0.9148	0.8667	0.8021	0.7508
$h = 9$	0.8151	0.4532	0.7048	0.8476	0.4325	0.7183	0.8040	0.9222	0.8675	0.7944	0.7341
$h = 10$	0.8127	0.4333	0.6952	0.8556	0.4175	0.7159	0.8079	0.9222	0.8730	0.7873	0.7317

Table 7. Coverage ratios of the 95% prediction intervals produced by the two-factor mortality models.

Horizons	Model 4			Lee-Carter	Li & Lee
	0.1	0.01	0.00001	/	/
λ_2					
$h = 1$	0.8822	0.8837	0.8759	0.9508	0.9392
$h = 2$	0.8624	0.8663	0.8478	0.9499	0.9683
$h = 3$	0.8482	0.8496	0.8242	0.9528	0.9863
$h = 4$	0.8311	0.8195	0.7941	0.9599	0.9943
$h = 5$	0.8130	0.7995	0.7627	0.9672	0.9979
$h = 6$	0.8092	0.7797	0.7438	0.9705	1.0000
$h = 7$	0.8071	0.7667	0.7321	0.9730	1.0000
$h = 8$	0.8085	0.7661	0.7201	0.9799	1.0000
$h = 9$	0.8373	0.7730	0.7071	0.9873	1.0000
$h = 10$	0.8571	0.7905	0.6905	0.9968	1.0000

of Supplementary Material. It is worth noting that, for single-factor models, our proposed models can provide larger coverage ratios with comparable forecasting efficiency. In the case of two-factor models, LC and Li % Lee models results in abnormally wide prediction intervals at almost all forecasting steps. Despite the resulting large coverage ratios, those intervals are undesirable, compared with those of Model 4.

4.4.2. Long-run predictions of mortality rates

We now compare the long-run predictions of mortality rates for all the G7 countries using the Model 1, for its best out-of-sample performance among the four proposed specifications. The forecast horizon h is chosen as 30 years beyond 2019 (up to 2049). To facilitate the comparison, the point forecasts (posterior means) are presented and discussed in this section.

To illustrate the long-term and age-specific forecasts, we present point forecasts of log mortality rates at age 65 for all the G7 countries in Figure 11. The left, middle and right panel displays results corresponding to Model 1 with a weak ($\lambda_2 = 0.1$), moderate ($\lambda_2 = 0.01$) and strong ($\lambda_2 = 0.00001$) shrinkage parameter, respectively. We also display long-run forecasts of log mortality rates at age 65 using

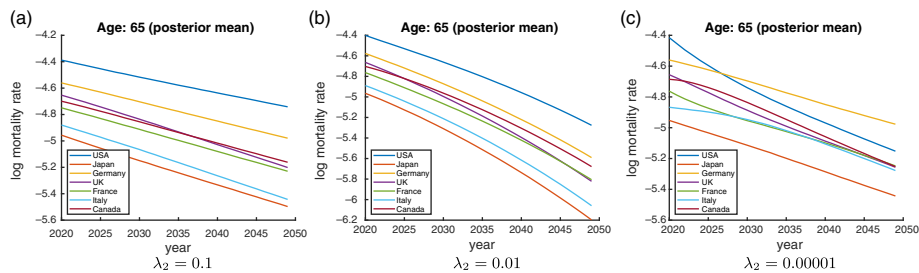


Figure 11. Point forecasts of log mortality rates at age 65 for all G7 countries.

Model 2 in the Section G of Supplementary Material, where the results are similar to those from Model 1. When $\lambda_2 = 0.00001$, despite the presence of some short-term fluctuations, the age-specific mortality rates exhibit parallel declining trends, implying their non-divergence over the long term. In contrast, when $\lambda_2 = 0.1$, the data demonstrate relatively more diversified declining patterns, signifying the long-term divergence of these age-specific mortality rates across the G7 countries. This underscores the effectiveness of our prior parameter (λ_2) in achieving the long-term coherence. Similar observations can be made from the plots of future life expectancy at birth provided in the Section H of Supplementary Material. By employing a strong shrinkage prior, the future life expectancy also appears to be nearly parallel across different countries.

5. Conclusion

In this paper, we present a new multi-population mortality framework based on the seminal Lee–Carter model. In particular, a hierarchical structure is assumed for the age effects, and a (structural) VECM is employed to fit the co-movements of the mortality dynamics. By employing the Bayesian inference with a shrinkage prior, the proposed model is flexible on balancing the short-term empirical patterns and long-term coherence in mortality forecasting. Building on the Bayesian MCMC literature, we construct an efficient precision block sampler that largely reduces the extensive computational cost of Kalman filter and small-blocked sampling. The application to the G7 data set demonstrates the usefulness of our model in understanding the mortality dynamics for actuarial practice.

Supplementary Material. To view supplementary material for this article, please visit <https://doi.org/10.1017/asb.2023.29>.

References

- Bai, J. and Wang, P. (2015) Identification and bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics*, **33**(2), 221–240.
- Barigozzi, M., Lippi, M. and Luciani, M. (2021) Large-dimensional dynamic factor models: Estimation of impulse-response functions with $i(1)$ cointegrated factors. *Journal of Econometrics*, **221**(2), 455–482.
- Bañbura, M., Giannone, D. and Reichlin, L. (2010) Large bayesian vector auto regressions. *Journal of Applied Econometrics*, **25**(1), 71–92.
- Baumeister, C. and Hamilton, J.D. (2019) Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *American Economic Review*, **109**(5), 1873–1910.
- Boonen, T.J. and Li, H. (2017) Modeling and forecasting mortality with economic growth: A multipopulation approach. *Demography*, **54**(5), 1921–1946.
- Cairns, A.J. (2000) A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, **27**(3), 313–330.
- Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D. and Khalaf-Allah, M. (2011) Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, **41**(1), 29–59.
- Chan, J.C. and Jeliaskov, I. (2009) Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, **1**(1-2), 101–120.

- Chen, H., MacMinn, R. and Sun, T. (2015) Multi-population mortality models: A factor copula approach. *Insurance: Mathematics and Economics*, **63**, 135–146.
- Czado, C., Delwarde, A. and Denuit, M. (2005) Bayesian poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, **36**(3), 260–284.
- Danesi, I.L., Haberman, S. and Millossovich, P. (2015) Forecasting mortality in subpopulations using lee–carter type models: A comparison. *Insurance: Mathematics and Economics*, **62**, 151–161.
- Denuit, M., Devolder, P. and Goderniaux, A.-C. (2007) Securitization of longevity risk: Pricing survivor bonds with wang transform in the Lee-Carter framework. *Journal of Risk and Insurance*, **74**(1), 87–113.
- Forni, M. and Gambetti, L. (2010) The dynamic effects of monetary policy: A structural factor model approach. *Journal of Monetary Economics*, **57**(2), 203–216.
- Geweke, J. and Amisano, G. (2011) Hierarchical markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, **26**(1), 1–29.
- Granger, C.W. (2004) Time series analysis, cointegration, and applications. *American Economic Review*, **94**(3), 421–425.
- Gupta, A. and Varga, T. (1992) Characterization of matrix variate normal distributions. *Journal of Multivariate Analysis*, **41**(1), 80–88.
- Human Mortality Database (2019) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).
- Hunt, A. and Blake, D. (2015) Modelling longevity bonds: Analysing the swiss re kortis bond. *Insurance: Mathematics and Economics*, **63**, 12–29. Special Issue: Longevity Nine - the Ninth International Longevity Risk and Capital Markets Solutions Conference.
- Hunt, A. and Blake, D. (2018) Identifiability, cointegration and the gravity model. *Insurance: Mathematics and Economics*, **78**, 360–368. Longevity risk and capital markets: The 2015-16 update.
- Hyndman, R.J., Booth, H. and Yasmeen, F. (2013) Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography*, **50**(1), 261–283.
- Jarner, S.F. and Jallbjørn, S. (2020) Pitfalls and merits of cointegration-based mortality models. *Insurance: Mathematics and Economics*, **90**, 80–93.
- Kilian, L. (2013) Structural vector autoregressions. In *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing.
- Kleinow, T. (2015) A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, **63**, 147–152.
- Kogure, A. and Kurachi, Y. (2010) A bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics*, **46**(1), 162–172.
- Koop, G. (2003) *Bayesian Econometrics*. New York: John Wiley & Sons.
- Koopman, S.J. and Durbin, J. (2003) Filtering and smoothing of state vector for diffuse state-space models. *Journal of Time Series Analysis*, **24**(1), 85–98.
- Kunst, R. and Neusser, K. (1990) Cointegration in a macroeconomic system. *Journal of Applied Econometrics (Chichester, England)*, **5**(4), 351–365.
- Lee, R. (2000) The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, **4**(1), 80–91.
- Lee, R. and Miller, T. (2001) Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, **38**(4), 537–549.
- Lee, R.D. and Carter, L.R. (1992) Modeling and forecasting us mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.
- Li, H., De Waegenaere, A. and Melenberg, B. (2015) The choice of sample size for mortality forecasting: A bayesian learning approach. *Insurance: Mathematics and Economics*, **63**, 153–168.
- Li, H. and Li, J. S.-H. (2017) Optimizing the Lee-Carter approach in the presence of structural changes in time and age patterns of mortality improvements. *Demography*, **54**(3), 1073–1095.
- Li, H. and Lu, Y. (2017) Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin*, **47**(2), 563–600.
- Li, H., Lu, Y. and Lyu, P. (2021) Coherent mortality forecasting for less developed countries. *Risks*, **9**(9), 151.
- Li, H. and Shi, Y. (2021a) Forecasting mortality with international linkages: A global vector-autoregression approach. *Insurance: Mathematics and Economics*, **100**, 59–75.
- Li, H. and Shi, Y. (2021b) Mortality forecasting with an age-coherent sparse var model. *Risks*, **9**(2), 35.
- Li, J.S.-H., Zhou, K.Q., Zhu, X., Chan, W.-S. and Chan, F.W.-H. (2019) A bayesian approach to developing a stochastic mortality model for China. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1523–1560.
- Li, J.S.-H., Zhou, R. and Hardy, M. (2015) A step-by-step guide to building two-population stochastic mortality models. *Insurance: Mathematics and Economics*, **63**, 121–134.
- Li, N. and Lee, R. (2005) Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography*, **42**(3), 575–594.
- Lin, T. and Tsai, C.C.-L. (2022) Hierarchical bayesian modeling of multi-country mortality rates. *Scandinavian Actuarial Journal*, **2022**(5), 375–398.
- Litterman, R.B. (1986) Forecasting with Bayesian vector autoregressions-five years of experience. *Journal of Business & Economic Statistics*, **4**(1), 25–38.

- Lütkepohl, H. (2005) *New Introduction to Multiple Time Series Analysis*. New York: Springer Science & Business Media.
- Mukherjee, T.K. and Naka, A. (1995) Dynamic relations between macroeconomic variables and the Japanese stock market: An application of a vector error correction model. *The Journal of Financial Research*, **18**(2), 223–237.
- Newton, M.A. and Raftery, A.E. (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**(1), 3–48.
- Njenga, C.N. and Sherris, M. (2020) Modeling mortality with a bayesian vector autoregression. *Insurance: Mathematics and Economics*, **94**, 40–57.
- Pedroza, C. (2006) A Bayesian forecasting model: Predicting us male mortality. *Biostatistics* **7**(4), 530–550.
- Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.
- Tuljapurkar, S., Li, N. and Boe, C. (2000) A universal pattern of mortality decline in the g7 countries. *Nature*, **405**(6788), 789–792.
- Wong, J.S., Forster, J.J. and Smith, P.W. (2018) Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics*, **83**, 206–221.
- Yang, S.S. and Wang, C.-W. (2013) Pricing and securitization of multi-country longevity risk with mortality dependence. *Insurance: Mathematics and Economics*, **52**(2), 157–169.
- Zhou, R., Wang, Y., Kaufhold, K., Li, J.S.-H. and Tan, K.S. (2014) Modeling period effects in multi-population mortality models: Applications to solvency ii. *North American Actuarial Journal*, **18**(1), 150–167.