Supplementary Material to *Multi-population Mortality Modelling: A Bayesian Hierarchical Approach*

Jianjie Shi*[1], Yanlin Shi[2], Pengjie Wang[1], and Dan Zhu†[1]

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

July 27, 2023

A The Precision Sampler: Computational Complexity

Another significant contribution of our paper is introducing the precision sampling for the Lee-Carter type model when estimated via Bayesian MCMC. Assuming the model specification of linearity and Gaussianality, an alternative approach to sampling latent states κ_t is to use Kalman Filter. Yet, due to the precision matrix's sparsity, as shown by Figure 1, one can greatly reduce the computational complexity in modelling mortality rates via state-space models.

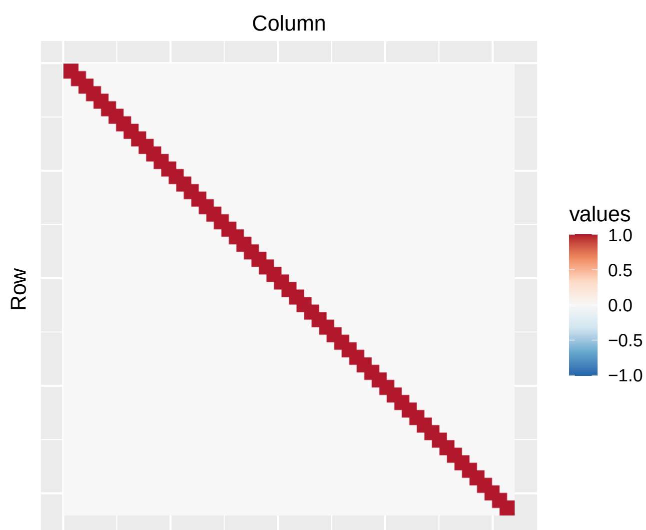


Figure 1: A simplified κ_t's precision matrix: non-zero elements are drawn in red whereas zeroes are just blanks

*This author thanks the support from Australian Government Research Training Program (RTP) Scholarship.
†Email address: dan.zhu@monash.edu

We compare computational complexities under several scenarios with different dimensions in both ages and time. We could observe that precision sampler is much faster than the commonly-used Kalman Filter, especially when the dimension of age is large.

Table 1: Computational time (in seconds) for 10000 iterations under different scenarios.

	Kalman Filter		Precision Sampler	
	$1986 \sim 2016$	$1956 \sim 2016$	$1986 \sim 2016$	$1956 \sim 2016$
$0 \sim 30$	3954	7046	2211	8250
$0 \sim 60$	20410	35578	4623	13684
$0 \sim 100$	74399	134695	13885	32431

In this section, we also compare the mixing performances of Kalman Filter and precision sampler. To illustrate it, we apply both methods to the mortality data with ages from 0 to 60 and years from 1986 to 2016. The effective sample size for the simulated parameters is reported. The effective sample size is the sample size required to give the same numerical variance as the MCMC sample if that sample were a simple random sample. It can be seen, from the comparison, that the proposed precision sampler improves the MCMC mixing than the Kalman Filter.

Table 2: Effective sample size of some selected parameters for 10000 iterations (after burn-in period)

	Kalman Filter	Precision Sampler
κ_{2016}^{7}	74	150
μ_a^{60}	5284	6758
μ_b^{60}	1530	1374
$\Sigma_a^{60,60}$	7837	7117
$\Sigma_b^{60,60}$	948	512
Ω^{60}	2526	2624
b^7	2822	2815
$\Pi^{7,7}$	8886	9110
$\Sigma_k^{7,7}$	3240	5229

In addition, we also present trace plots for these selected parameters below.

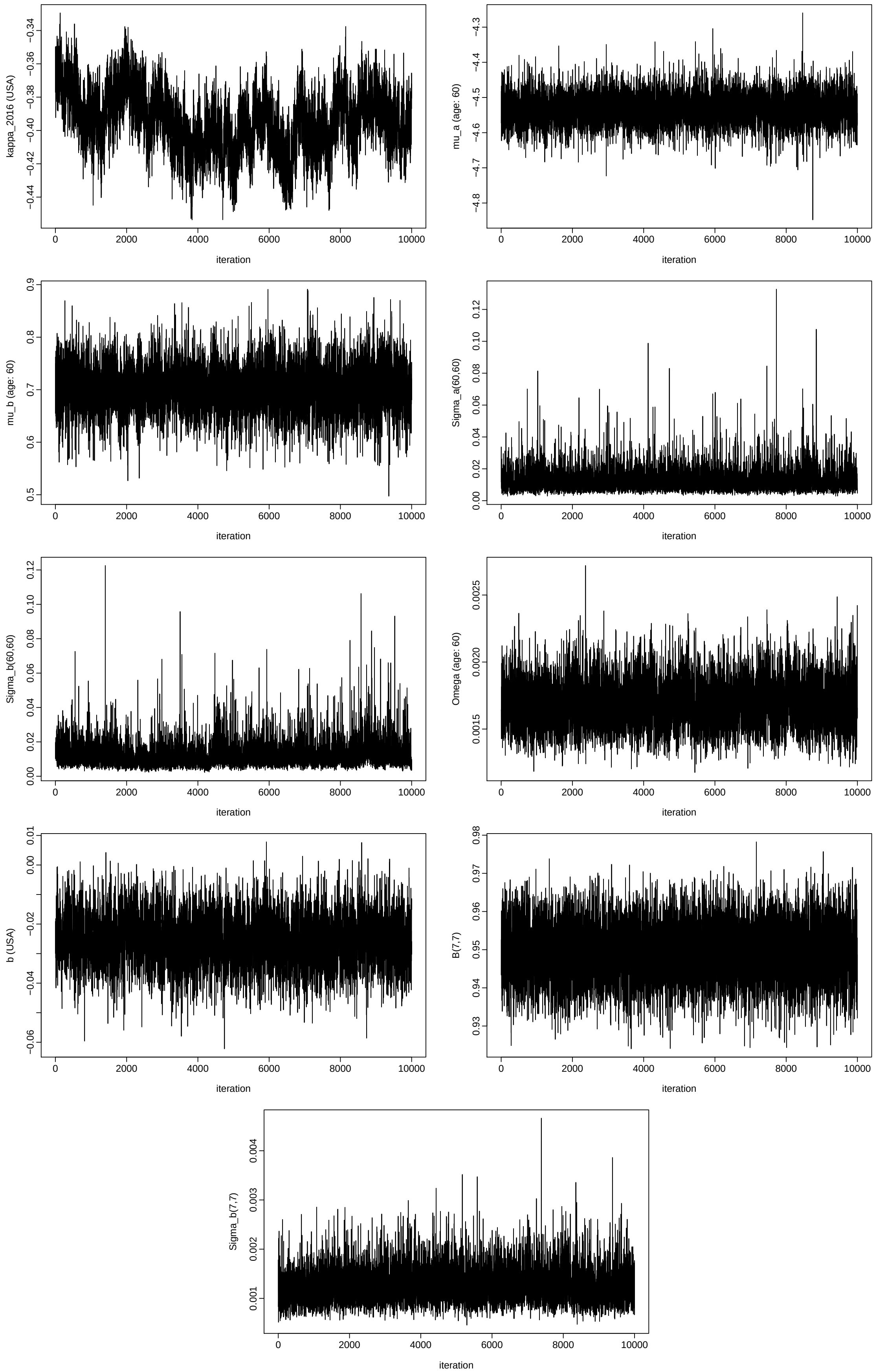


Figure 2: Trace plots (after burn-in period) of selected parameters (Precision Sampler)

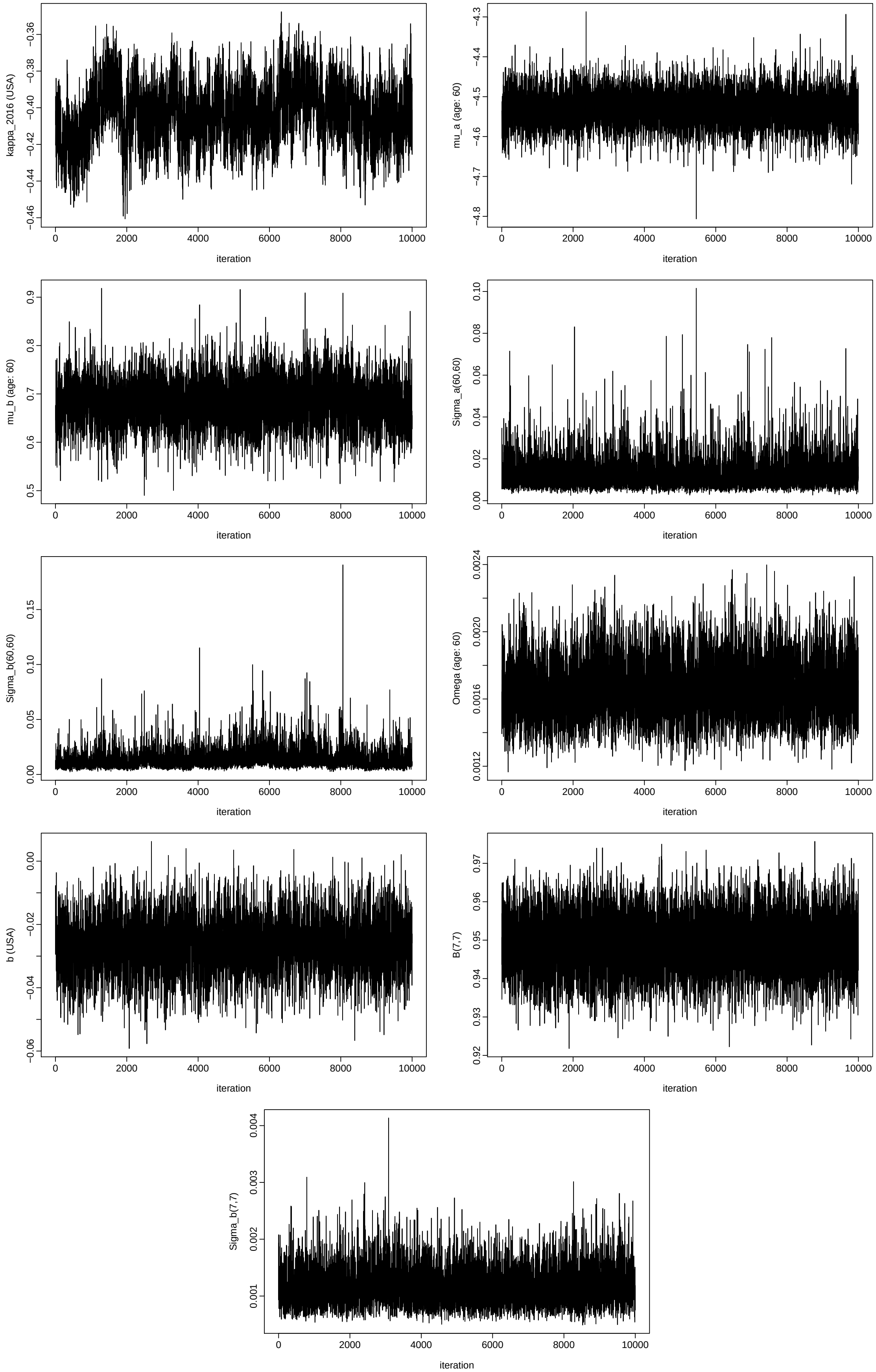


Figure 3: Trace plots (after burn-in period)) of selected parameters (Kalman Filter)

B State-Space Representations of Lee-Carter Model and Li & Lee Model

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).

Specifically, the single-factor LC model, applicable to each distinct population denoted as i, is predicated on the following assumptions:

$$y_t^i = \alpha^i + \beta^i \kappa_t^i + \epsilon_t^i, \quad \epsilon_t^i \sim N(0, \Omega_i)$$
$$\kappa_t^i = b^i + \kappa_{t-1}^i + \xi_t^i, \quad \xi_t^i \sim N(0, \sigma_{\kappa,i}^2)$$

where $\Omega_i = diag\{g_{x_0,i}, \cdots, g_{\omega,i}\}$. While the two-factor LC model assumes that

$$y_t^i = \alpha^i + \beta_1^i \kappa_{1t}^i + \beta_2^i \kappa_{2t}^i + \epsilon_t^i, \quad \epsilon_t^i \sim N(0, \Omega_i)$$
$$\kappa_{1t}^i = b_1^i + \kappa_{1,t-1}^i + \xi_{1t}^i, \quad \xi_{1t}^i \sim N(0, \sigma_{1,\kappa,i}^2)$$
$$\kappa_{2t}^i = b_2^i + B_2^i \kappa_{2,t-1}^i + \xi_{2t}^i, \quad \xi_{2t}^i \sim N(0, \sigma_{2,\kappa,i}^2)$$

Similarly, the single-factor Li & Lee model, as posited by Li and Lee (2005), advocates for the use of a common age effect, denoted as β, along with a common factor, κ_t, to model the log mortality rates. Specifically, the model is as follows:

$$y_t^i = \alpha^i + \beta \kappa_t + \epsilon_t^i, \quad \epsilon_t^i \sim N(0, \Omega)$$
$$\kappa_t = b + \kappa_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\kappa^2)$$

This model is referred to as the common factor model in their original publication. In contrast, the two-factor Li & Lee model, which is also referred to as the augmented common factor model, is formulated as:

$$y_t^i = \alpha^i + \beta_1 \kappa_{1t} + \beta_2^i \kappa_{2t}^i + \epsilon_t^i, \quad \epsilon_t^i \sim N(0, \Omega)$$
$$\kappa_{1t} = b_1 + \kappa_{1,t-1} + \xi_{1t}, \quad \xi_{1t} \sim N(0, \sigma_{1,\kappa}^2)$$
$$\boldsymbol{\kappa}_{2t} = b_2 + B_2 \boldsymbol{\kappa}_{2,t-1} + \xi_{2t}, \quad \xi_{2t} \sim N(0, \Sigma_{2,\kappa})$$

where $\boldsymbol{\kappa}_{2t} = (\kappa_{2t}^1, \cdots, \kappa_{2t}^I)'$ adheres to a VAR(1) model. The Li & Lee model can be regarded as a particular instance of the multi-level dynamic factor model, and Bai and Wang (2015) comprehensively discusses the constraints for its identification.

C Marginal Likelihood

Suppose we have two different models M_i $(i = 1, 2)$ to explain data y, and M_i depends on parameters θ^i. For each model M_i, we can derive the marginal likelihood as $\pi(y|M_i) = \int \pi(y|\theta^i, M_i)\pi(\theta^i|M_i)d\theta^i$ where $\pi(y|\theta^i, M_i)$ and $\pi(\theta^i|M_i)$ are the corresponding likelihood and prior distribution. Then to compare M_1 and M_2, it is common to use the posterior odds ratio, which is simply the ratio of posterior model probabilities:

$$\frac{\pi(M_1|y)}{\pi(M_2|y)} = \frac{\pi(y|M_1)}{\pi(y|M_2)} \frac{\pi(M_1)}{\pi(M_2)}$$

where $\pi(M_i)$ is referred to as the prior model probabilities for M_i. Especially, when $\pi(M_1) = \pi(M_2)$, the posterior odds ratio becomes simply the ratio of marginal likelihood and is called Bayes factor. In this case, $\pi(y|M_1) > \pi(y|M_2)$ is equivalent to $\pi(M_1|y) > \pi(M_2|y)$, providing the evidence in favour of model M_1 over M_2. For a more detailed discussion of the marginal likelihood and Bayesian model comparison, please see Koop (2003).

However, evaluating the marginal likelihood is usually a computationally challenging task. The most commonly-used Bayesian information criteria (or BIC) is just used to approximate 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, i.e.,

$$\frac{1}{\pi(y)} = \int \frac{\pi(\theta|y)}{\pi(y|\theta)}\, d\theta = \mathbb{E}\left(\frac{1}{\pi(y|\theta)}\middle| y\right).$$

where $\pi(\theta|y)$ is the posterior distribution of parameter θ given the observed data y, $\pi(y|\theta)$ is the likelihood function and $\pi(y)$ is the marginal likelihood. This suggests that $\pi(y)$ can be approximated by the sample harmonic mean of the likelihood:

$$\frac{1}{\pi(y)} = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{1}{\pi(y|\theta^i)}\middle| y\right)$$

based on R draws $\{\theta^i\}$ from the posterior distribution $\pi(\theta|y)$.

D Simulating from the Posterior Predictive Distribution

Assume that historical data and future log mortality rates distribute independently, given all the latent random states and parameters, the posterior predictive distribution of $\{\mathbf{y}_{T+s}\}_{s=1}^{h}$ is

$$\begin{aligned} p(\{\mathbf{y}_{T+s}\}_{s=1}^{h}|\mathcal{F}_T) = \int & p(\{\mathbf{y}_{T+s}\}_{s=1}^{h}|\{\kappa_{T+s}\}_{s=1}^{h}, \{\alpha^i\}_{i=1}^{I}, \{\beta^i\}_{i=1}^{I}, \theta) \cdot p(\{\kappa_{T+s}\}_{s=1}^{h}|\{\kappa_t\}_{t=1}^{T}, \theta)\cdot \\ & p(\{\kappa_t\}_{t=1}^{T}, \{\alpha^i\}_{i=1}^{I}, \{\beta^i\}_{i=1}^{I}, \theta|\mathcal{F}_T) d\kappa_t d\alpha^i d\beta^i d\theta, \end{aligned} \tag{1}$$

where $\mathcal{F}_T$ represents the information set up to time T.

Specifically, the empirical posterior predictive distribution of $\boldsymbol{y}_{T+s}$ is obtained via the following steps:

1. Sample a realisation of $\{\kappa_t\}_{t=1}^{T}, \{\alpha^i\}_{i=1}^{I}, \{\beta^i\}_{i=1}^{I}$ and θ from their empirical posterior distribution;
2. Given those simulated values at Step 1, sample a realisation of $\{\kappa_{T+s}\}_{s=1}^{h}$ from its conditional predictive distribution $p(\{\kappa_{T+s}\}_{s=1}^{h}|\{\kappa_t\}_{t=1}^{T}, \theta)$;
3. Given the simulated values at Steps 1 and 2, sample a realisation of $\{\mathbf{y}_{T+s}\}_{s=1}^{h}$ from its conditional predictive distribution
$p(\{\mathbf{y}_{T+s}\}_{s=1}^{h}|\{\kappa_{T+s}\}_{s=1}^{h}, \{\alpha^i\}_{i=1}^{I}, \{\beta^i\}_{i=1}^{I}, \theta)$; and
4. Repeat steps 1-3, until the required number of simulations is fulfilled.

E In-sample Results for the Two-factor Model 4

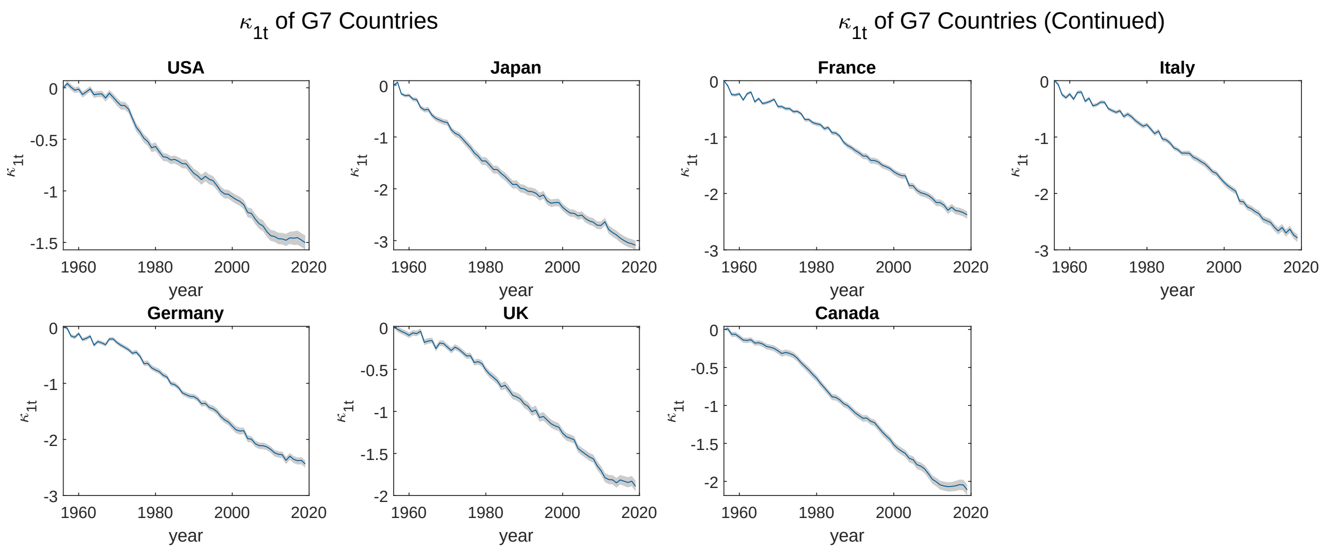


Figure 4: Temporal plots of estimated first latent factor κ^i_{1t} for all the G7 countries (solid line: posterior mean of; grey area: 99% credible interval)

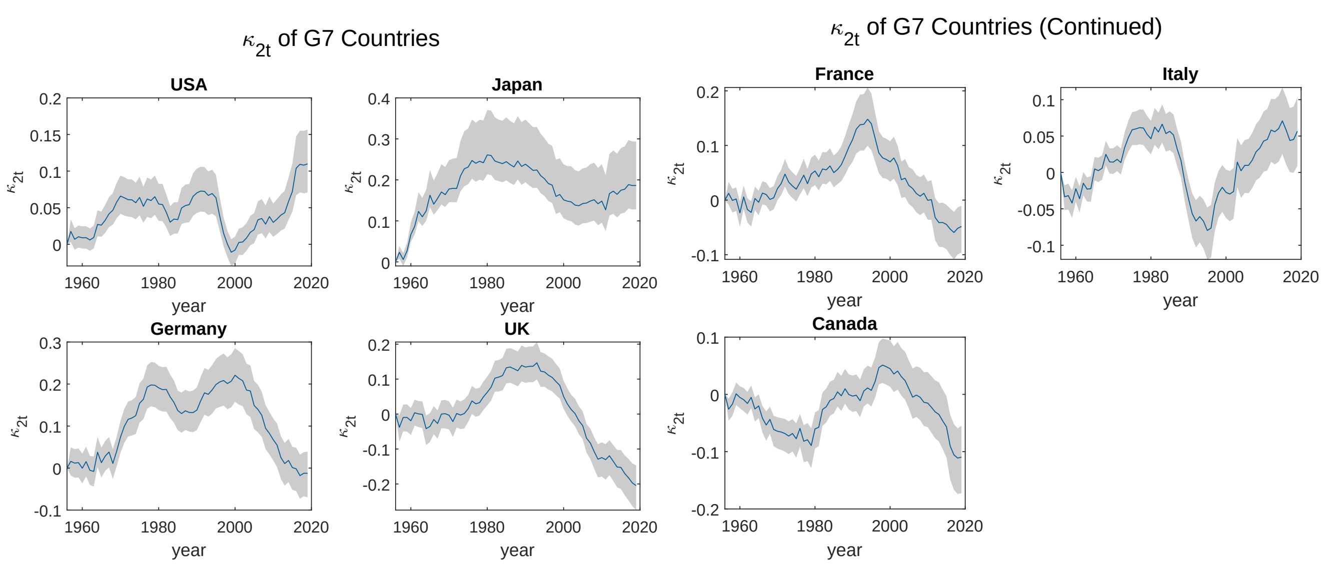


Figure 5: Temporal plots of estimated first latent factor κ^i_{2t} for all the G7 countries (solid line: posterior mean of; grey area: 99% credible interval)

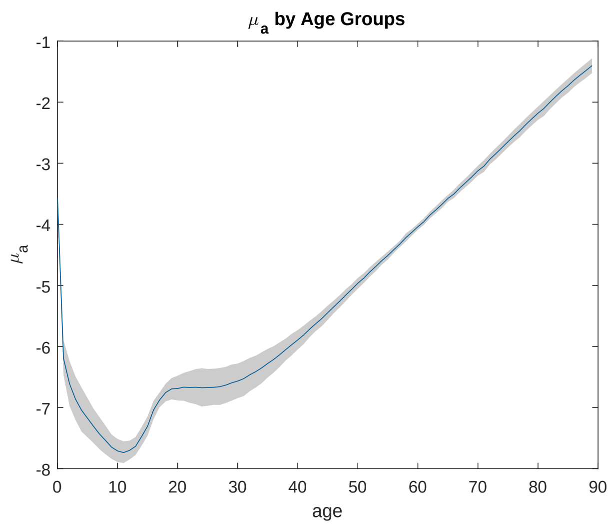


Figure 6: Estimated age effects μ_a (solid line: posterior mean; grey area: 99% credible interval)

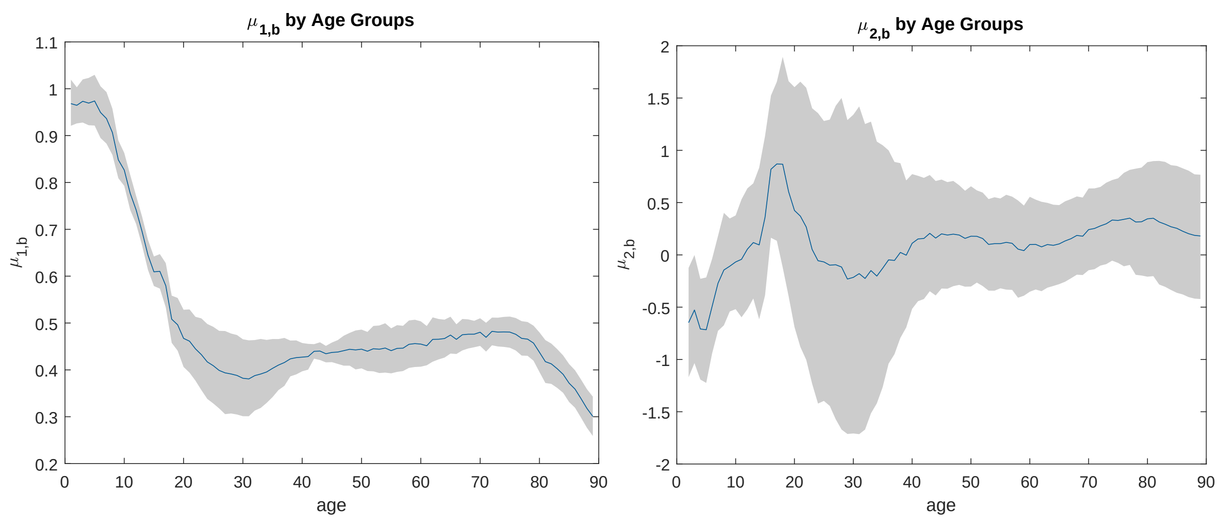


Figure 7: Estimated age effects $\mu_{1,b}$ and $\mu_{2,b}$ (solid line: posterior mean; grey area: 99% credible interval)

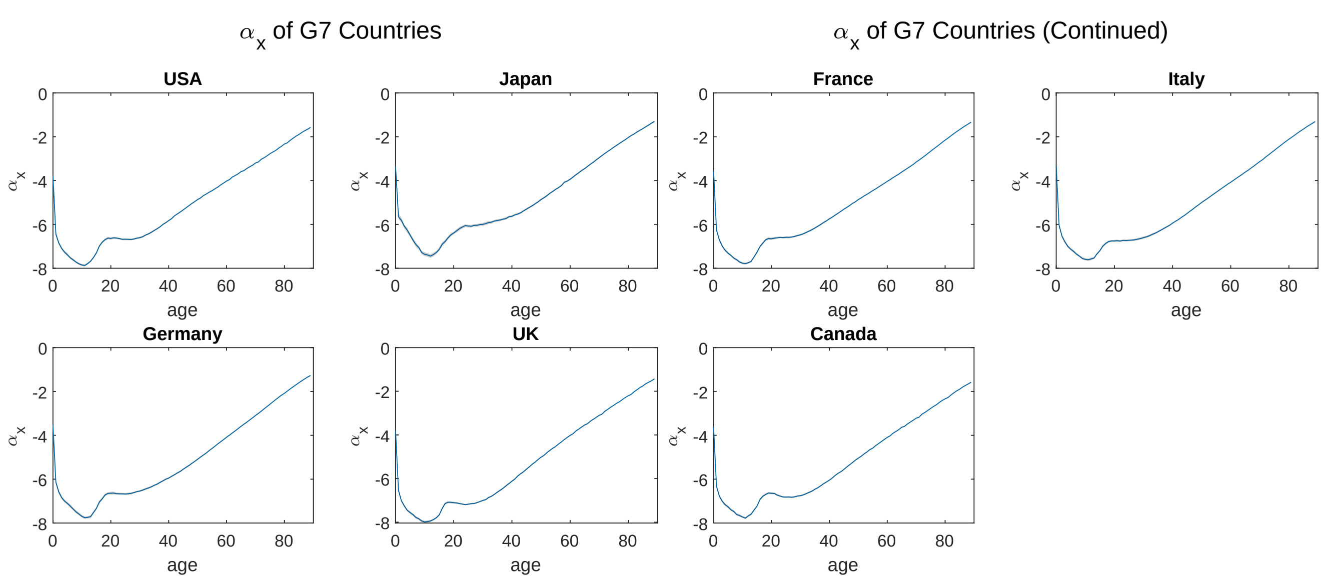


Figure 8: Estimated age effects α^i's (solid line: posterior mean; grey area: 99% credible interval)

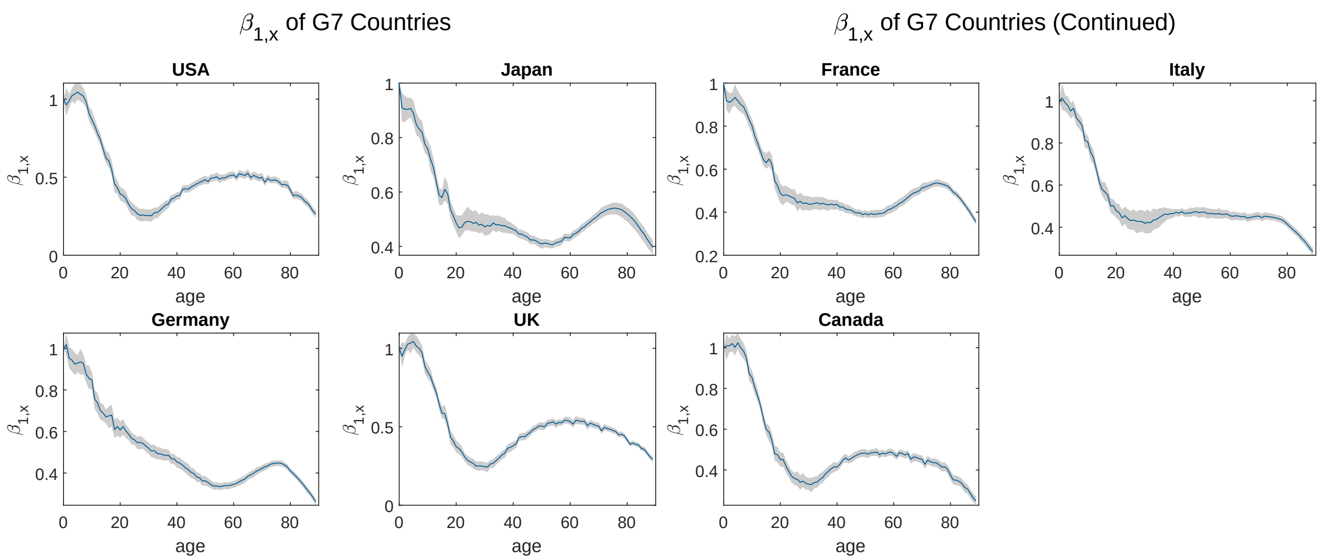


Figure 9: Plots of estimated age effects β_1^i's (solid line: posterior mean; grey area: 99% credible interval)

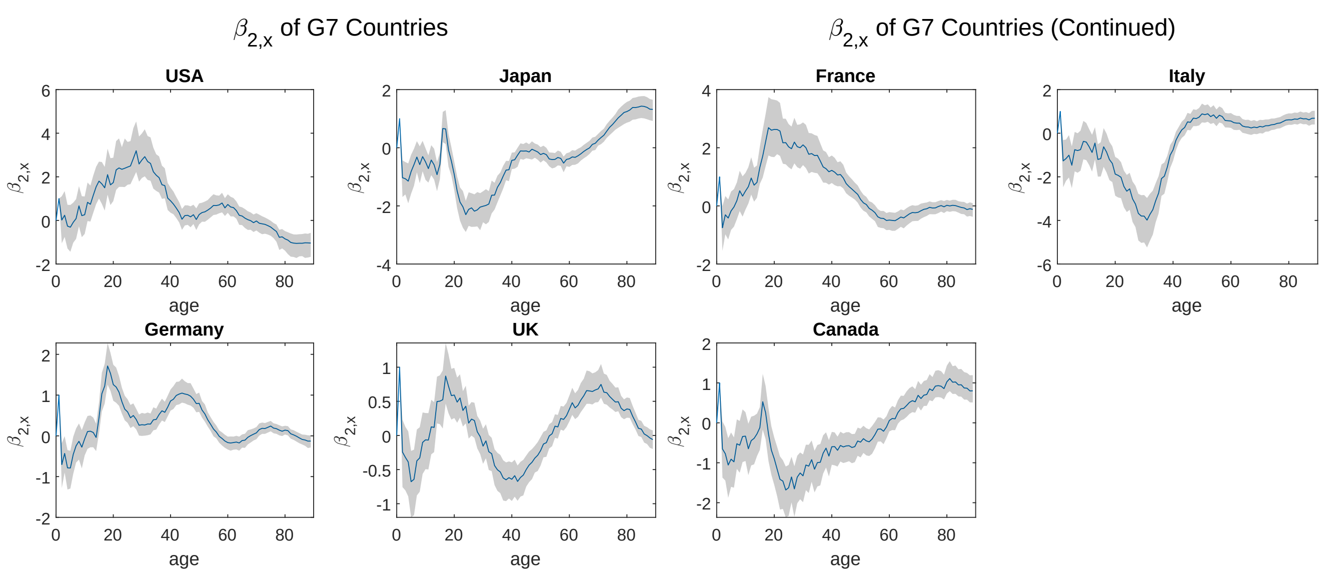


Figure 10: Plots of estimated age effects β_2^i's (solid line: posterior mean; grey area: 99% credible interval)

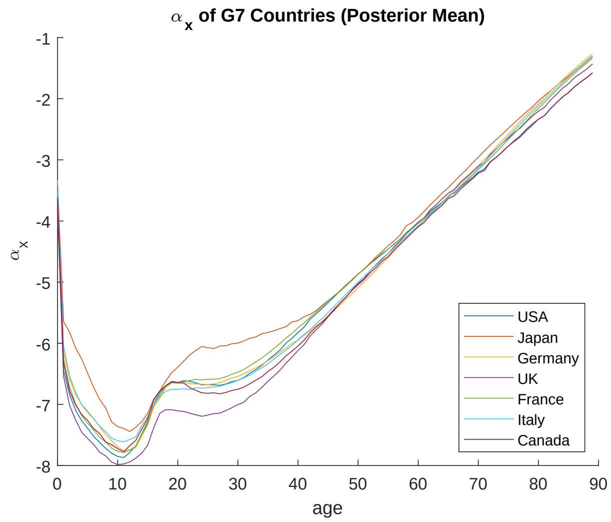


Figure 11: Comparison of estimated age effects α^i's for all G7 countries

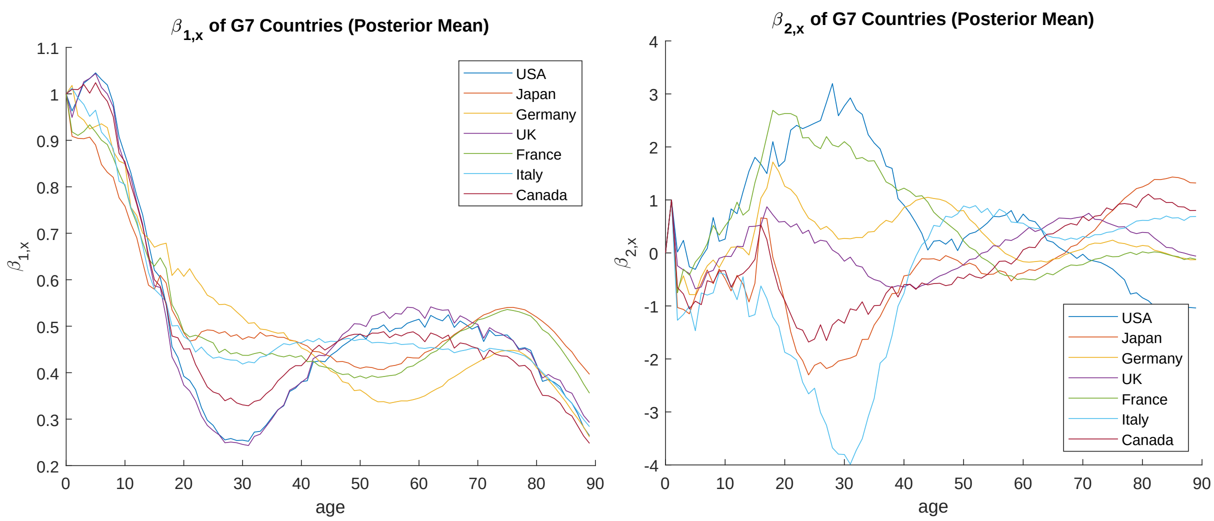


Figure 12: Comparison of estimated age effects β_1^i's and β_2^i's for all G7 countries

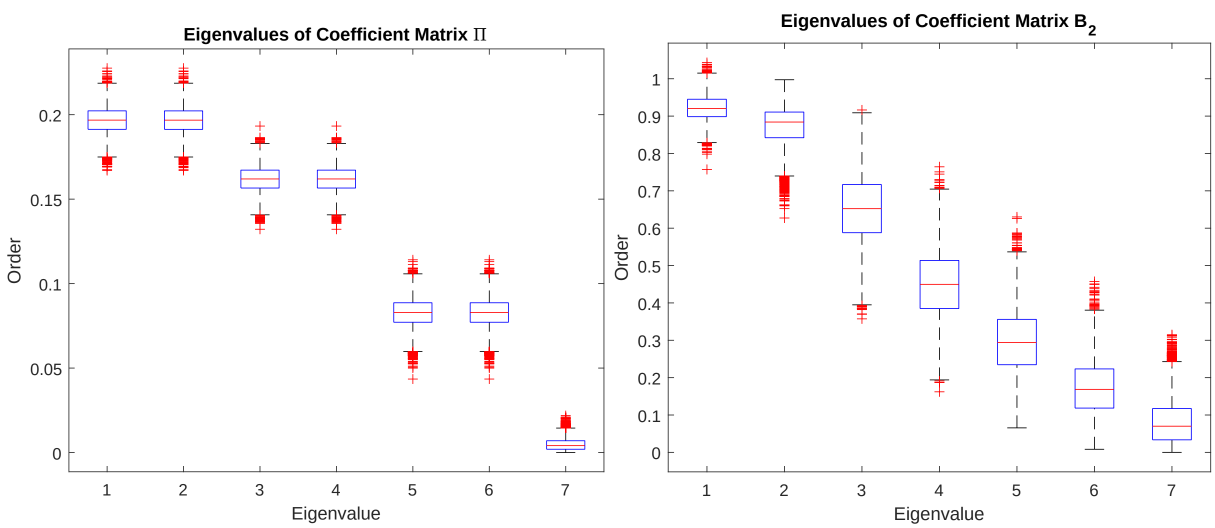


Figure 13: Posteriors of eigenvalues' modulus of the simulated coefficient matrices Π and B_2 (ordered by the size of modulus)

F Average Lengths of Prediction Intervals

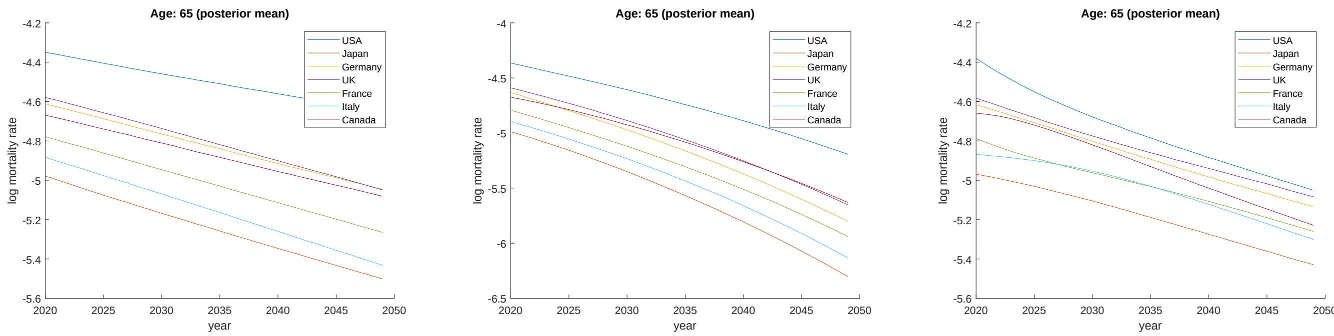
Table 3: Average Lengths of the 95% prediction intervals produced by the single-factor mortality models

Horizons	Model 1			Model 2			Model 3			Lee-Carter	Li & Lee
λ_2	0.1	0.01	0.00001	0.1	0.01	0.00001	0.1	0.01	0.00001	/	/
h=1	0.3155	0.3174	0.3174	0.353	0.3546	0.3545	0.553	0.5536	0.5538	0.3057	0.4595
h=2	0.3305	0.3347	0.332	0.3668	0.3702	0.3673	0.562	0.5629	0.5613	0.3227	0.4624
h=3	0.3463	0.3527	0.3447	0.3819	0.387	0.3785	0.5716	0.5723	0.5681	0.3387	0.4654
h=4	0.3633	0.3716	0.356	0.3985	0.4056	0.3886	0.5827	0.5827	0.5743	0.3541	0.4688
h=5	0.3826	0.3924	0.3667	0.4176	0.4262	0.398	0.5953	0.5943	0.5802	0.3687	0.4725
h=6	0.4038	0.4153	0.3765	0.4394	0.4502	0.407	0.6094	0.6078	0.5859	0.3826	0.4763
h=7	0.4277	0.44	0.386	0.4647	0.4778	0.4159	0.6264	0.6225	0.5915	0.3964	0.4806
h=8	0.4534	0.4673	0.3945	0.493	0.5082	0.4247	0.6447	0.6394	0.597	0.41	0.4849
h=9	0.4727	0.4975	0.4033	0.5109	0.5419	0.4331	0.6588	0.6582	0.6023	0.4222	0.4896
h=10	0.4917	0.5323	0.413	0.5332	0.5822	0.443	0.6745	0.6801	0.6077	0.4359	0.4946

Table 4: Average Lengths of the 95% prediction intervals produced by the two-factor mortality models

Horizons	Model 4			Lee-Carter	Li & Lee
λ_2	0.1	0.01	0.00001	/	/
h=1	0.2715	0.2715	0.2735	0.4078	0.4476
h=2	0.2961	0.2943	0.2957	0.5302	0.6913
h=3	0.3199	0.3144	0.3129	0.6555	1.1
h=4	0.3456	0.3335	0.3277	0.8153	1.8788
h=5	0.3759	0.3519	0.3405	1.0232	3.4736
h=6	0.4113	0.3709	0.3526	1.3337	6.8664
h=7	0.453	0.3907	0.364	1.7489	14.2252
h=8	0.5032	0.4104	0.3749	2.2675	30.8982
h=9	0.526	0.4301	0.3846	3.1303	71.7165
h=10	0.5563	0.4503	0.3943	4.1759	195.9341

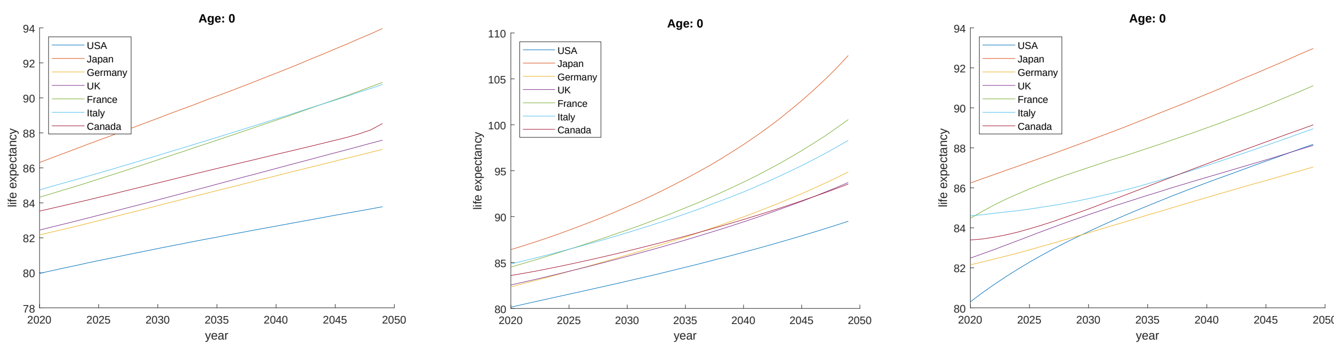
G Long-run Predictions of Mortality Rates Using Model 2



(a) $\lambda_2 = 0.1$ (b) $\lambda_2 = 0.01$ (c) $\lambda_2 = 0.00001$

Figure 14: Point forecasts of log mortality rates at age 65 for all G7 countries

H Future Life Expectancy at Birth



(a) $\lambda_2 = 0.1$ (b) $\lambda_2 = 0.01$ (c) $\lambda_2 = 0.00001$

Figure 15: Point forecasts of log mortality rates at age 65 for all G7 countries

Here we also present the points forecasts of life expectancy at birth (e_0) for all the G7 countries. Since e_0 is determined by mortality rates of all ages, it is a useful statistic to represent the overall forecasts of a population. The results under different shrinkage hyper-parameters are plotted. As expected, by using the strong prior ($\lambda_2 = 0.00001$), all G7 countries demonstrate increasing e_0 in (nearly) parallel fashions in the long run. Although as seen from Figure 15, life expectancies are not strictly parallel to each other, especially for Japan. This is because we choose to impose a strong shrinkage prior rather than using a restricted model specification. Thus, the final result balances the trade-off between the short/medium-term prediction accuracy and the long-term coherence. Although not shown here, Model 2 has similar results of life expectancies as Model 1. It means that, to ensure long-term coherence, the co-integration relationship in the VECM plays a more important role than the common age effect. It can be expected that using the more restricted VECM form displayed in the end of Section 3.1 could obtain more coherent forecasts in the long run.

References

Bai, J. and P. Wang (2015). Identification and bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics 33*(2), 221–240.

Koop, G. (2003). *Bayesian econometrics*. John Wiley & Sons.

Li, N. and R. Lee (2005). Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography 42*(3), 575–594.

Newton, M. A. and A. E. Raftery (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(1), 3–48.

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.