Forecasting high-dimensional functional time series: Application to sub-national age-specific mortality

Cristian Felipe Jimenez Varon[1], Ying Sun[1], Han Lin Shang[2]

1. Department of Statistics. King Abdullah University of Science and Technology
2. Department of Actuarial Studies and Business Analytics. Macquarie University

June 21, 2023

Seminar at Macquarie University. Sydney, Australia.

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Motivation and Introduction

- Most countries worldwide have seen continuous drops in mortality rates, which are also associated with aging populations.
- Policymakers from insurance firms and government departments demand more precise mortality forecasts.
- For planning, several statistical methods have been presented for **forecasting** age-specific central mortality rates, life-table death counts, or survival function.

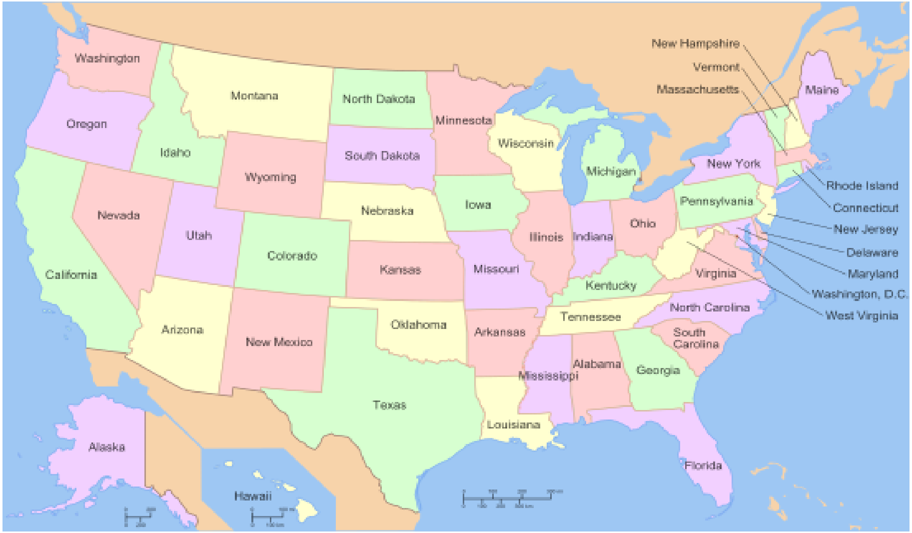
Motivation and Introduction

- Lee and Carter (1992) uses a principal component (PC) method to derive a single time-varying index of the level of mortality rates, from which **forecasts are obtained using a random walk with drift**.
- The model structure is given by $\log(m_{x,t}) = a_t + b_x k_t + \epsilon_{x,t}$
  - a_x is the age pattern averaged across years.
  - b_x is the first PC reflecting the relative change at each age.
  - k_t: is the first set of PC scores by year t.
  - $\epsilon_{x,t}$ is the residual at age x and year t.

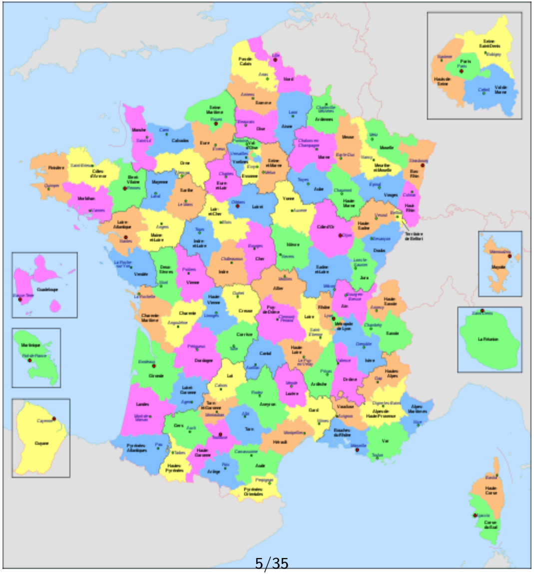
Functional time series (FTS)

- Several approaches have modified and extended the Lee-Carter method.
    - For instance, Hyndman and Ullah (2007) proposed a functional data (FDA) approach along with nonparametric smoothing and high-order principal components for mortality forecasting.
        - In the FDA approach, the functional data are generated from a stochastic process $\{\mathcal{X}_t(u), t \in \mathcal{Z}, u \in \mathcal{I} \subset \mathcal{I}\}$
        - It is assumed that the mortality rate in each year follows an underlying smooth function of age u.
- When mortality rates are collected over time, we refer to the data as functional time series (FTS).
- One major **drawback** of the Lee-Carter method and other contributions is that they **mainly** focus on forecasting mortality for a **single population**.
- **Each population** can be further categorized based on gender, state, ethnic group, socioeconomic position, and other factors.

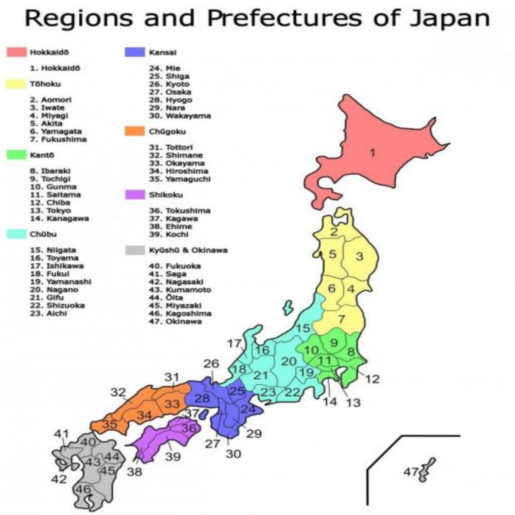
Example of high-dimensional time series: USA



Example of high-dimensional time series: France



Example of high-dimensional time series: Japan

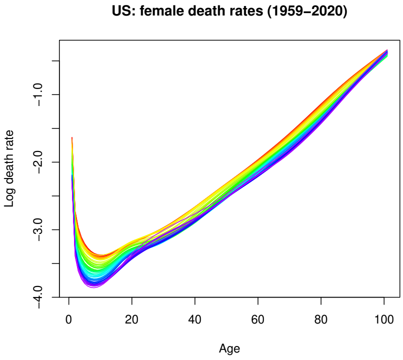


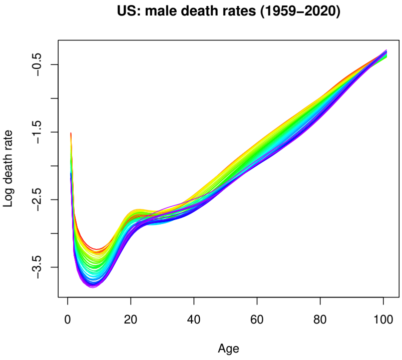
Features of high-dimensional functional time series

- We consider modeling and forecasting high-dimensional functional time series (HDFTS), which can be **cross-sectionally correlated** and **temporally dependent**.
- Two-way functional median polish decomposition, which is robust against outliers. Two-way functional ANOVA.
- The two-way functional ANOVA and median polish decompose HDFTS into deterministic and time-varying components.
- Dynamic functional principal component analysis, is implemented to produce **forecasts** for the time-varying components.
- Forecast curves are obtained by **combining** the forecasts of the **time-varying components** with the **deterministic components**.

US mortality

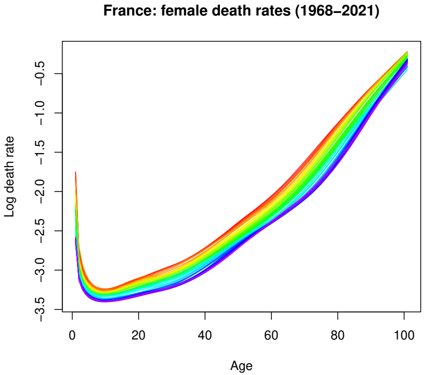
- US mortality database has a complete set of state-level life tables for studying geographic variation in mortality across the US.
- Data cover 50 states and the District of Columbia for each year between 1959 and 2020 with mortality data up to age 110.
- Ages from 0 to 100 in single years of age (u), last age group including all ages above 100.

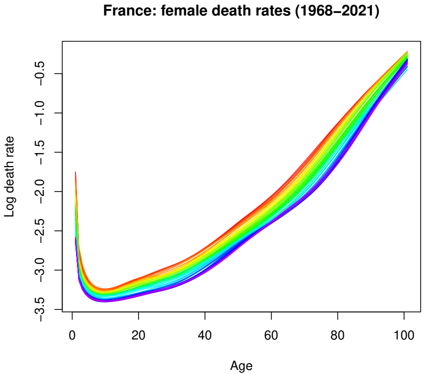




French mortality

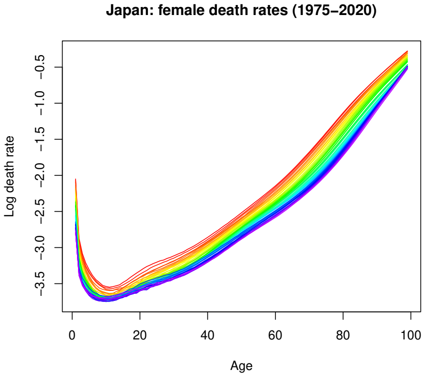
- French Human Mortality Database has mortality by departments.
- France has 97 departments, of which two (Seine and Seine et Oise) do not have any data from 1968 to 2021.

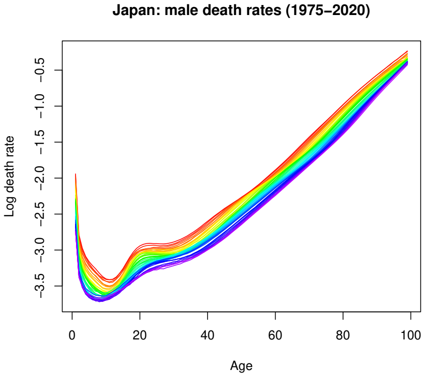




Japanese Mortality

- Japanese Mortality Database has mortality by prefecture.
- Ages from 0 to 98 in single years of age, last age group including all ages at and above 99.





Two-way functional median polish (FMP)

- Let $\mathcal{Y}_{t,s}^{g}(u)$ be $\log_{10}$ mortality for age u, state s, gender g at year t.
- $\mathcal{Y}_{t,s}^{g}(u)$ can be decomposed as

$$\mathcal{Y}_{t,s}^{g}(u) = \mu(u) + \alpha_s(u) + \beta^g(u) + \mathcal{X}_{t,s}^{g}(u), \qquad u \in \mathcal{I}$$

  - u is a continuous variable, but observed at $(u_1, \dots, u_p)$ grid points.
  - $\mu(u)$: functional grand effect
  - $\alpha_s(u)$: functional row effect; $\text{median}_s\{\alpha_s(u)\} = 0$
  - $\beta^g(u)$: functional column effect; $\text{median}_g\{\beta^g(u)\} = 0$
  - $\mathcal{X}_s^g(u) = [\mathcal{X}_{1,s}^{g}(u), \dots, \mathcal{X}_{T,s}^{g}(u)]$: functional residual; $\text{median}_s\{\mathcal{X}_{t,s}^{g}\} = \text{median}_g\{\mathcal{X}_{t,s}^{g}\} = 0$
- Deterministic components (states and genders) + time-varying components (functional residuals).

Long-run covariance estimation

- For a stationary residual process $\mathcal{X}^{g}_{t,s}(u)$, long-run covariance function

$$C(u,v) = \sum_{l=-\infty}^{\infty} \gamma_l(u,v) = \sum_{l=-\infty}^{\infty} \text{cov}\left[\mathcal{X}^{g}_{0,s}(u), \mathcal{X}^{g}_{l,s}(v)\right]$$

  where $u, v \in \mathcal{I}$ and l denote a time-series lag variable.

- For a finite sample, a natural estimator of $C(u,v)$ is

$$\widehat{C}_T(u,v) = \frac{1}{T}\sum_{|l|=0}^{|l|\le T}(T-|l|)\widehat{\gamma}_l(u,v) \tag{1}$$

  where

$$\widehat{\gamma}_l(u,v) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T-l}\left[\mathcal{X}^{g}_{t,s}(u) - \overline{\mathcal{X}}^{g}_{s}(u)\right]\left[\mathcal{X}^{g}_{t+l,s}(v) - \overline{\mathcal{X}}^{g}_{s}(v)\right] & \text{if } l \ge 0 \\ \frac{1}{T}\sum_{t=1-l}^{T}\left[\mathcal{X}^{g}_{t,s}(u) - \overline{\mathcal{X}}^{g}_{s}(v)\right]\left[\mathcal{X}^{g}_{t+l,s}(v) - \overline{\mathcal{X}}^{g}_{s}(v)\right] & \text{if } l < 0 \end{cases}$$

Kernel estimator of the long-run covariance

- The long-run covariance function can be seen as a sum of autocovariance functions with decreasing weights.
- It is common in practice to determine the optimal lag value of l to balance the trade-off between squared bias and variance.
- Some approaches use the kernel sandwich estimator

$$\widehat{\widehat{C}}_{T,b}(u,v) = \sum_{l=-\infty}^{\infty} W_q\left(\frac{l}{b}\right)\widehat{\gamma}_l(u,v)$$

  - b: bandwidth
  - $W_q(\cdot)$:symmetric weight function with bounded support of order q.
  - Rice and Shang (2017) propose a plug-in algorithm for obtaining the optimal bandwidth parameter to minimize the asymptotic mean-squared normed error between the estimated and actual long-run covariance functions.

Dynamic functional principal components

- Via the Mercer's lemma, the estimated long-run covariance function $\widehat{\widehat{C}}_{T,b}(u,v)$ can be approximated by

$$\widehat{\widehat{C}}_{T,b}(u,v) = \sum_{k=1}^{\infty} \theta_k \phi_k(u) \phi_k(v)$$

  - $\theta_1 > \theta_2 > \ldots > 0$: eigenvalues of $\widehat{\widehat{C}}_{T,b}(u,v)$
  - $[\phi_1(u), \phi_2(u), \ldots]$ orthonormal functional principal components.
- Via Karhunen-Loève expansion of the realization of a stochastic process,

$$\mathcal{X}^g_{t,s}(u) = \overline{\mathcal{X}}^g_s(u) + \sum_{k=1}^{\infty} \gamma^g_{k,t,s} \phi^g_{k,s}(u)$$

where $\gamma^g_{k,t,s} = \left\langle \mathcal{X}^g_{t,s}(u) - \overline{\mathcal{X}}^g_s(u), \phi^g_{k,s}(u) \right\rangle$, denotes the k^{th} set of principal component scores for time t.

Selection of the K functional principal components

We select K as the minimum of leading principal components reaching 95% of the total variance explained, such that

$$K = \underset{K:K\geq 1}{\operatorname{argmin}} \left\{ \sum_{k=1}^{K} \widehat{\theta}_k \bigg/ \sum_{k=1}^{T} \widehat{\theta}_k \mathbb{1}_{\{\widehat{\theta}_k > 0\}} \geq 0.95 \right\}$$

where $\mathbb{1}\{\cdot\}$ represents the binary indicator function.

Multivariate functional principal component analysis

- By stacking female and male populations,

$$\boldsymbol{\mathcal{X}}_{t,s}(u) = \boldsymbol{\Phi}_s(u)\boldsymbol{\Gamma}_{t,s}$$

- $\boldsymbol{\mathcal{X}}_{t,s}(u) = [\mathcal{X}^{\text{F}}_{t,s}(u), \mathcal{X}^{\text{M}}_{t,s}(u)]^\top$
- Combined functional principal scores

$$\boldsymbol{\Gamma}_{t,s} = \left[\gamma^{\text{F}}_{1,t,s}, \dots, \gamma^{\text{F}}_{K,t,s}, \gamma^{\text{M}}_{1,t,s}, \dots, \gamma^{\text{M}}_{K,t,s}\right]^\top$$

$\boldsymbol{\Gamma}_{t,s}$ is a $((2 \times K) \times 1)$ vector

- Combined principal components

$$\boldsymbol{\Phi}_s(u) = \begin{pmatrix} \phi^{\text{F}}_{1,1}(u) & \dots & \phi^{\text{F}}_{K,1}(u) & 0 & \dots & 0 \\ 0 & \dots & 0 & \phi^{\text{M}}_{1,2}(u) & \dots & \phi^{\text{M}}_{K,2}(u) \end{pmatrix}$$

$\boldsymbol{\Phi}_s(u)$ is a $2 \times (2 \times K)$ matrix

h-step-ahead point forecasts

- By conditioning on $\mathbf{\Phi}_s(u)$, obtain h-step-ahead point forecasts

$$\begin{aligned}\widehat{\boldsymbol{\mathcal{X}}}_{T+h|T,s}(u) &= E\Big[\boldsymbol{\mathcal{X}}_{T+h,s}(u)\Big|\boldsymbol{\mathcal{X}}_{1,s}(u),\dots,\boldsymbol{\mathcal{X}}_{T,s}(u);\mathbf{\Phi}_s(u)\Big] \\ &= \overline{\boldsymbol{\mathcal{X}}}_s(u) + \mathbf{\Phi}_s(u)\widehat{\mathbf{\Gamma}}_{T+h|T,s}\end{aligned}$$

  where the empirical mean function $\overline{\boldsymbol{\mathcal{X}}}_s(u) = [\overline{\mathcal{X}}_s^{\text{F}}(u), \overline{\mathcal{X}}_s^{\text{M}}(u)]$
- Use univariate time series forecasting method to obtain forecast principal component score $\widehat{\mathbf{\Gamma}}_{T+h|T,s}$.
- With the forecasted functional residuals, add back the deterministic component.

$$\widehat{\mathcal{Y}}^{g}_{T+h|T,s}(u) = \mu(u) + \alpha_s(u) + \beta^{g}(u) + \widehat{\mathcal{X}}^{g}_{T+h|T,s}(u)$$

Sieve bootstrap

1) Center the observed functional time series by calculating $\mathcal{Z}^g_{t,s}(u) = \mathcal{X}^g_{t,s}(u) - \overline{\mathcal{X}}^g_s(u)$
2) Apply FPCA to $\boldsymbol{\mathcal{Z}}^g_s(u) = [\mathcal{Z}^g_{1,s}(u), \dots, \mathcal{Z}^g_{T,s}(u)]$ to obtain estimated functional principal components and their scores.
3) Fit a VAR(p), process to the "forward" series of the estimated scores

$$\gamma^g_{m,s} = \sum_{j=1}^{p} A_{j,p} \gamma^g_{m-j,s} + \epsilon^g_{m,s}, \qquad m = p+1, \dots, T$$

where $\epsilon^g_{m,s}$ being residuals, $A_{j,p}$: forward VAR(p) coefficient.

Sieve bootstrap

4) Generate

$$\gamma^{g,*}_{T+h,s} = \sum_{j=1}^{p} A_{j,p}\gamma^{g,*}_{T+h-j,s} + \epsilon^{g,*}_{T+h,s}$$

where we set $\gamma^{g,*}_{T+h-j} = \gamma_{T+h-j}$ if $T+h-j \leq T$ and $\epsilon^{g,*}_{T+h,s}$ is iid resampled from the set of centered residuals $(\epsilon^{g}_{m,s} - \overline{\epsilon}^{g}_{s})$, $\overline{\epsilon}^{g}_{s} = (T-p)^{-1}\sum_{m=p+1}^{T}\epsilon^{g}_{t,s}$

5) Compute

$$\mathcal{X}^{g,*}_{T+h,s}(u) = \overline{\mathcal{X}}^{g}_{s}(u) + \sum_{k=1}^{K}\gamma^{g,*}_{k,T+h,s}\phi^{g}_{k,s}(u) + U^{g,*}_{T+h,s}(u)$$

where $U^{g,*}_{T+h,s}(u)$ is iid resampled from the set $\{U^{g}_{t,s}(u) - \overline{U}^{g}_{s}(u), t = 1, 2, \dots, T\}$, $\overline{U}^{g}_{s}(u) = T^{-1}\sum_{t=1}^{T}U^{g}_{t,s}(u)$ and $U^{g}_{t,s}(u) = \mathcal{X}^{g}_{t,s}(u) - \sum_{k=1}^{K}\gamma^{g}_{k,t,s}\phi^{g}_{k,s}(u)$

Sieve bootstrap

6) Fit a VAR(p) process to the "backward" series of the estimated scores;

$$\gamma^{g}_{\nu,s} = \sum_{j=1}^{p} B_{j,p}\gamma^{g}_{\nu+j,s} + \xi^{g}_{\nu,s}, \qquad \nu = 1, 2, \ldots, T-p$$

where $B_{j,p}$ denotes the backward VAR(p) coefficient.

7) Generate a pseudo-time series of the scores $\{\gamma^{g,*}_{1,s}, \ldots, \gamma^{g,*}_{T,s}\}$ by setting $\gamma^{g,*}_{t,s} = \gamma^{g}_{t,s}$ for $t = T, T-1, \ldots, T-w+1$
8) By using for $t = T-w, T-w-1, \ldots, 1$, the backward VAR representation $\gamma^{g,*}_{\nu,s} = \sum_{j=1}^{p} B_{j,p}\gamma^{g,*}_{\nu+j,s} + \xi^{g,*}_{\nu,s}$
9) Generate a pseudo-functional time series $\{\mathcal{X}^{g,*}_{1,s}, \ldots, \mathcal{X}^{g,*}_{T,s}\}$

Sieve bootstrap

10) For each bootstrapped $\mathcal{X}_{t,s}^{g,*}(u)$, we apply a functional time-series forecasting method to obtain its h-step-ahead forecast, denoted by $\widehat{\mathcal{X}}_{T+h|T,s}^{g,*}(u)$
11) Model calibration error, $\omega_{T+h,s}^{g,*}(u) = \mathcal{X}_{T+h,s}^{g,*}(u) - \widehat{\mathcal{X}}_{T+h|T,s}^{g,*}(u)$, is the difference between the VAR extrapolated forecasts and the model-based forecasts.
12) Search for an optimal tuning parameter δ, where the symmetric prediction interval $(-\delta \times \text{sd}[\omega_{T+h,s}^{g,1}, \dots, \omega_{T+h,s}^{g,B}], \delta \times \text{sd}[\omega_{T+h,s}^{g,1}, \dots, \omega_{T+h,s}^{g,B}])$ achieves the smallest coverage probability difference between the empirical and nominal coverage probabilities based on the in-sample data.

Sieve bootstrap

13) Using the same functional time-series forecasting method, we apply it to the original functional time series to obtain the h-step-ahead forecast, denoted by $\widehat{\mathcal{X}}^{g}_{T+h|T,s}(u)$.
14) We add the deterministic component. The prediction interval of mortality curves is

$$\widehat{\mathcal{Y}}^{g,\ell}_{T+h|T,s}(u) = \mu(u) + \alpha_s(u) + \beta^{g}(u) + \widehat{\mathcal{X}}^{g,\ell}_{T+h|T,s}(u)$$

where ℓ symbolizes either the lower or upper bound.

Point forecast evaluation

- Rolling window scheme: with a training set of size T, produce $(T+h)-$step-ahead forecast.
- Iterates over $h = 1, \dots, H = 10$, the training set rolls one-step-ahead each time until $T + H$.
- We use the root mean squared prediction error (RMSPE) and the mean absolute prediction error (MAPE) to evaluate the point forecast accuracy.

Point forecast errors

- For each of the states and gender as

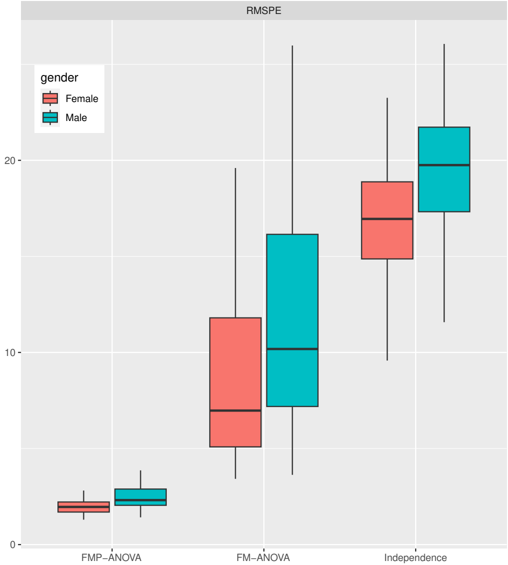
$$\text{RMSPE}_s^g(h) = \sqrt{\frac{1}{Hp}\sum_{\zeta=h}^{H}\sum_{i=1}^{p}\left[\frac{\mathcal{Y}_{T+\zeta,s}^{g}(u_i) - \widehat{\mathcal{Y}}_{T+\zeta,s}^{g}(u_i)}{\mathcal{Y}_{T+\zeta,s}^{g}(u_i)}\right]^2} \times 100$$

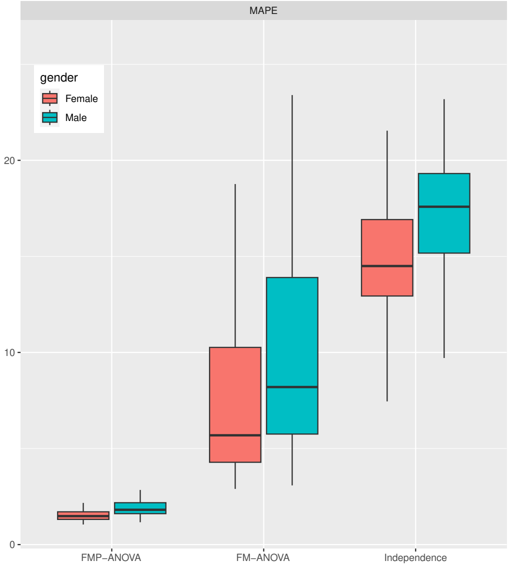
$$\text{MAPE}_s^g(h) = \frac{1}{Hp}\sum_{\zeta=h}^{H}\sum_{i=1}^{p}\left|\frac{\mathcal{Y}_{T+\zeta,s}^{g}(u_i) - \widehat{\mathcal{Y}}_{T+\zeta,s}^{g}(u_i)}{\mathcal{Y}_{T+\zeta,s}^{g}(u_i)}\right| \times 100$$

  - $\mathcal{Y}_{T+\zeta,s}^{g}(u_i)$ represents the holdout sample for state s and gender g.
  - $\widehat{\mathcal{Y}}_{T+\zeta,s}^{g}(u_i)$ represents the corresponding point forecasts.
- Average over H different number of forecast horizons

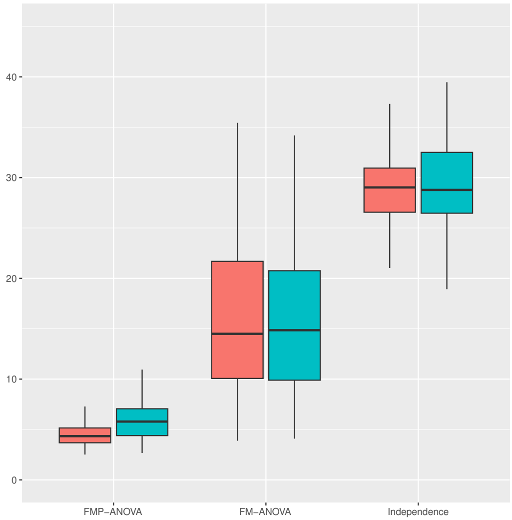
$$\overline{\text{RMSPE}}_s^g = \frac{1}{H}\sum_{h=1}^{H}\text{RMSPE}_s^g(h) \qquad \overline{\text{MAPE}}_s^g = \frac{1}{H}\sum_{h=1}^{H}\text{MAPE}_s^g(h)$$

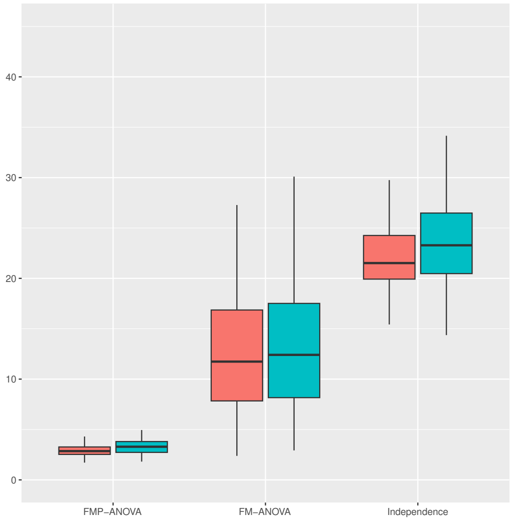
US data results



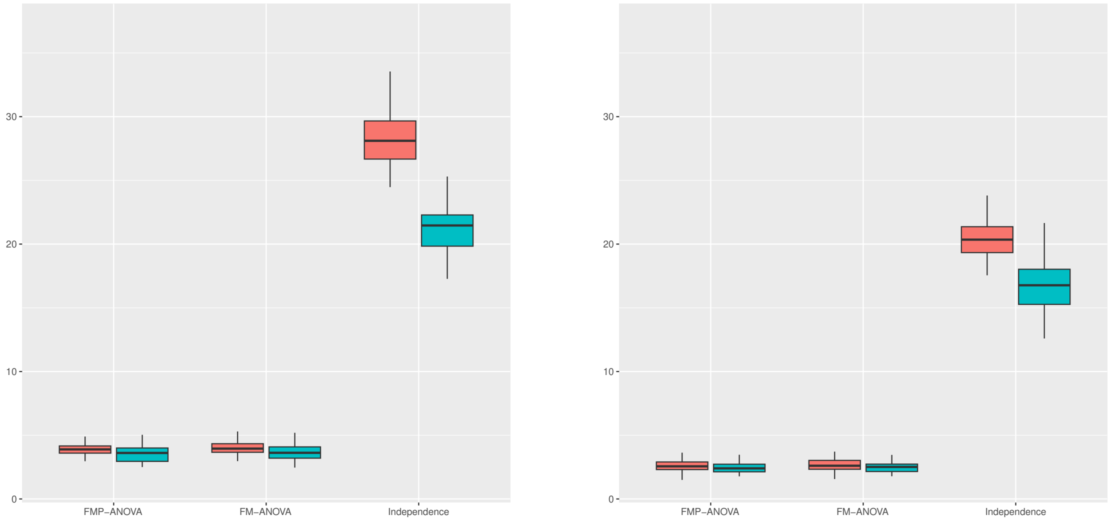


French data results





Japan data results



Interval forecast evaluation

- Empirical coverage probability is defined as follows

$$\text{Empirical coverage}_s^g = 1 - \frac{1}{Hp}\sum_{\zeta=h}^{H}\sum_{i=1}^{p}\Big[\mathbb{1}\Big\{\mathcal{Y}_{T+\zeta|T,s}^{g}(u_i) > \widehat{\mathcal{Y}}_{T+\zeta|T,s}^{g,\text{ub}}(u_i)\Big\} + \mathbb{1}\Big\{\mathcal{Y}_{T+\zeta|T,s}^{g}(u_i) < \widehat{\mathcal{Y}}_{T+\zeta|T,s}^{g,\text{lb}}(u_i)\Big\}\Big]$$

  - H denotes the number of curves in the forecasting period.
  - p denotes the number of discretized points for the age.
  - $\widehat{\mathcal{Y}}_{T+\zeta|T,s}^{g,\text{ub}}$ and $\widehat{\mathcal{Y}}_{T+\zeta|T,s}^{g,\text{lb}}$ denote the upper and lower bounds.
- Pointwise CPD is defined as

$$\text{CPD}_s^g = \Big|\text{Empirical coverage}_s^g - \text{Nominal coverage}\Big|$$

The lower the CPD_s^g value, the better the forecasting method's performance.

Interval score

- Scoring rule for the interval forecast at discretized point u_i is

$$
\begin{aligned}
S^{g}_{\alpha,\zeta,s}\left[\widehat{\mathcal{Y}}^{g,\text{lb}}_{T+\zeta|T,s}(u_i), \widehat{\mathcal{Y}}^{g,\text{ub}}_{T+\zeta|T,s}(u_i), \mathcal{Y}^{g}_{T+\zeta|T,s}(u_i)\right] &= \left[\widehat{\mathcal{Y}}^{g,\text{ub}}_{T+\zeta|T,s}(u_i) - \widehat{\mathcal{Y}}^{g,\text{lb}}_{T+\zeta|T,s}(u_i)\right] \\
&+ \frac{2}{\alpha}\left[\widehat{\mathcal{Y}}^{g,\text{lb}}_{T+\zeta|T,s}(u_i) - \mathcal{Y}^{g}_{T+\zeta|T,s}(u_i)\right]\mathbb{1}\left\{\mathcal{Y}^{g}_{T+\zeta|T,s}(u_i) < \widehat{\mathcal{Y}}^{g,\text{lb}}_{T+\zeta|T,s}(u_i)\right\} \\
&+ \frac{2}{\alpha}\left[\mathcal{Y}^{g}_{T+\zeta|T,s}(u_i) - \widehat{\mathcal{Y}}^{g,\text{ub}}_{T+\zeta|T,s}(u_i)\right]\mathbb{1}\left\{\widehat{\mathcal{Y}}^{g,\text{ub}}_{T+\zeta|T,s}(u_i) > \mathcal{Y}^{g}_{T+\zeta|T,s}(u_i)\right\}
\end{aligned}
$$

  where α: denotes a level of significance.

- Mean interval score for the total of T series as

$$
\overline{S}^{g}_{\alpha,s} = \frac{1}{Hp}\sum_{\zeta=h}^{H}\sum_{i=1}^{p} S^{g}_{\alpha,\zeta,s}\left[\widehat{\mathcal{Y}}^{g,\text{lb}}_{T+\zeta|T,s}(u_i), \widehat{\mathcal{Y}}^{g,\text{ub}}_{T+\zeta|T,s}(u_i), \mathcal{Y}^{g}_{T+\zeta|T,s}(u_i)\right]
$$

- The optimal interval score is achieved when $\mathcal{Y}^{g}_{T+\zeta|T,s}(u_i)$ lies between $\widehat{\mathcal{Y}}^{g,\text{lb}}_{T+\zeta|T,s}(u_i)$ and $\widehat{\mathcal{Y}}^{g,\text{ub}}_{T+\zeta|T,s}(u_i)$, with the distance between the upper bound and the lower bound being minimal.

Functional median polish. Empirical coverage probability

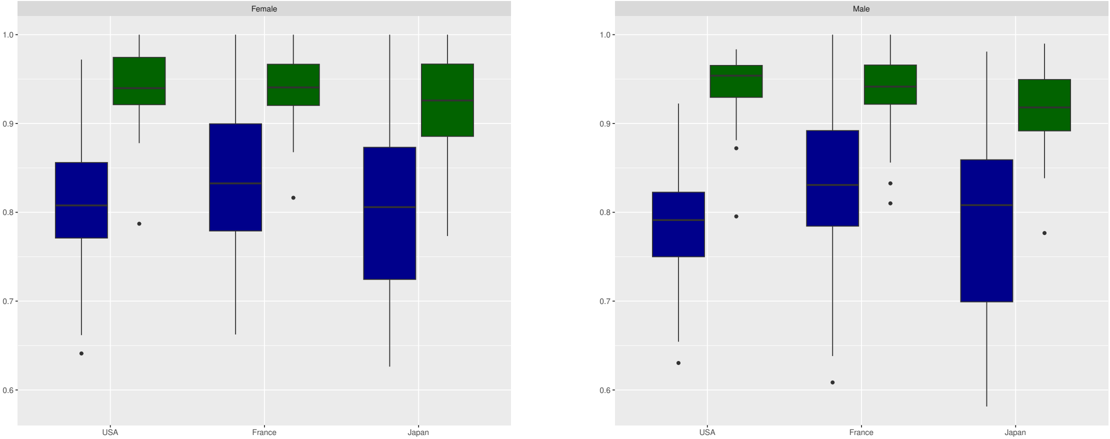
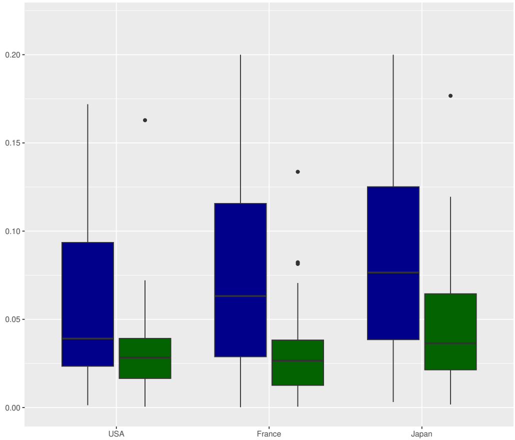
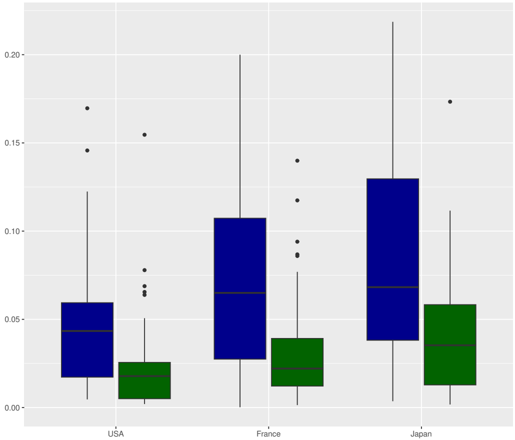


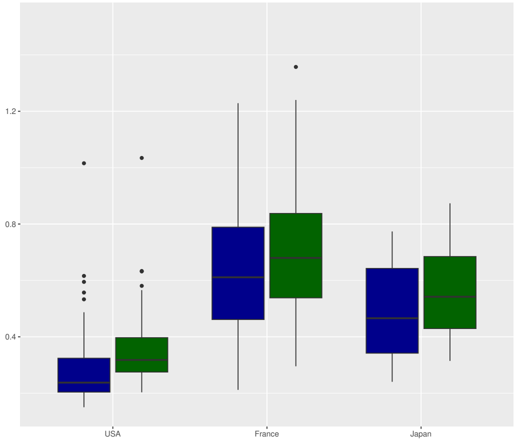
Figure: Consider two nominal coverage probabilities 80% (dark blue) and 95% (dark green). Each plot contains the US (most left), France (center), and Japan (most right).

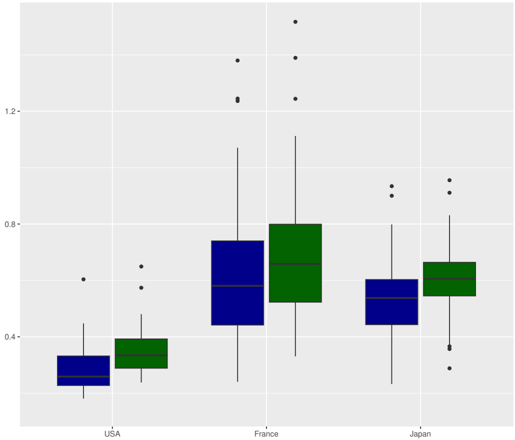
Functional median polish. CPD





Functional median polish. Interval score





Conclusion

- FMP and functional ANOVA produce more accurate forecasts than the ones from the independent FTS forecasting method.
- FMP performs better than functional ANOVA for the US and France, but not for Japan.
- The individual forecast errors for horizons $h = 1, \dots, H$, obtained from both methods for each state, are available in a developed shiny app `https://cristianjv.shinyapps.io/HDFTSForecasting/`.

Paper: Jimenez-Varon, C. F., Y. Sun, and H. L. Shang (2023). Forecasting high-dimensional functional time series: Application to sub-national age-specific mortality.

Thank you

References

Hyndman, R. and M. S. Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis 51*(10), 4942–4956.

Lee, R. D. and L. R. Carter (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association: Applications & Case Studies 87*(419), 659–671.

Rice, G. and H. L. Shang (2017). A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. *Journal of Time Series Analysis 38*(4), 591–609.