Beyond Linearity: A Bayesian Additive Tree Approach to Mortality Forecasting and Nowcasting

Jianjie Shi, Yunyun Wang

Longevity 20, 2025

Contents

Lee-Carter Model

Lee & Carter (1992):

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

- α_x: the average over time of $\log(m_{x,t})$ at age x,
- β_x: the impact of κ_t on $\log(m_{x,t})$ at age x,
- κ_t: the overall temporal trend of $\log(m_{x,t})$.

Moreover, Lee & Carter (1992) found that κ_t could be modelled as a random walk with drift γ, i.e.

$$\kappa_t = \gamma + \kappa_{t-1} + \omega_t.$$

Multi-factor Model

Let

$$\mathbf{y}_t = [\log(m_{x_1,t}), ..., \log(m_{x_N,t})]',$$

Multi-factor Model (FM):

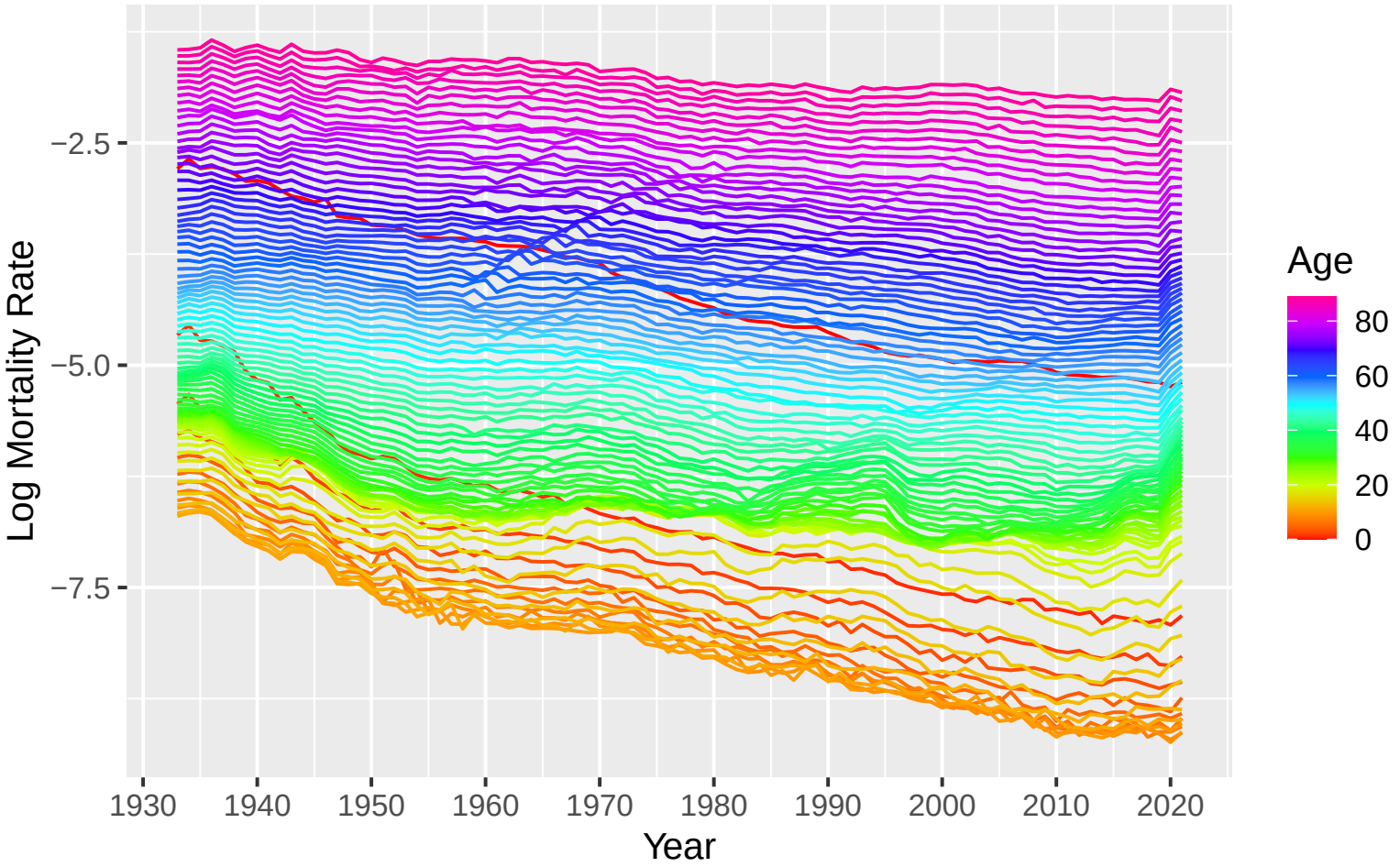
$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\kappa}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \Sigma) \\ \boldsymbol{\kappa}_t &= b + \boldsymbol{\kappa}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \Sigma_\omega) \end{aligned}$$

where

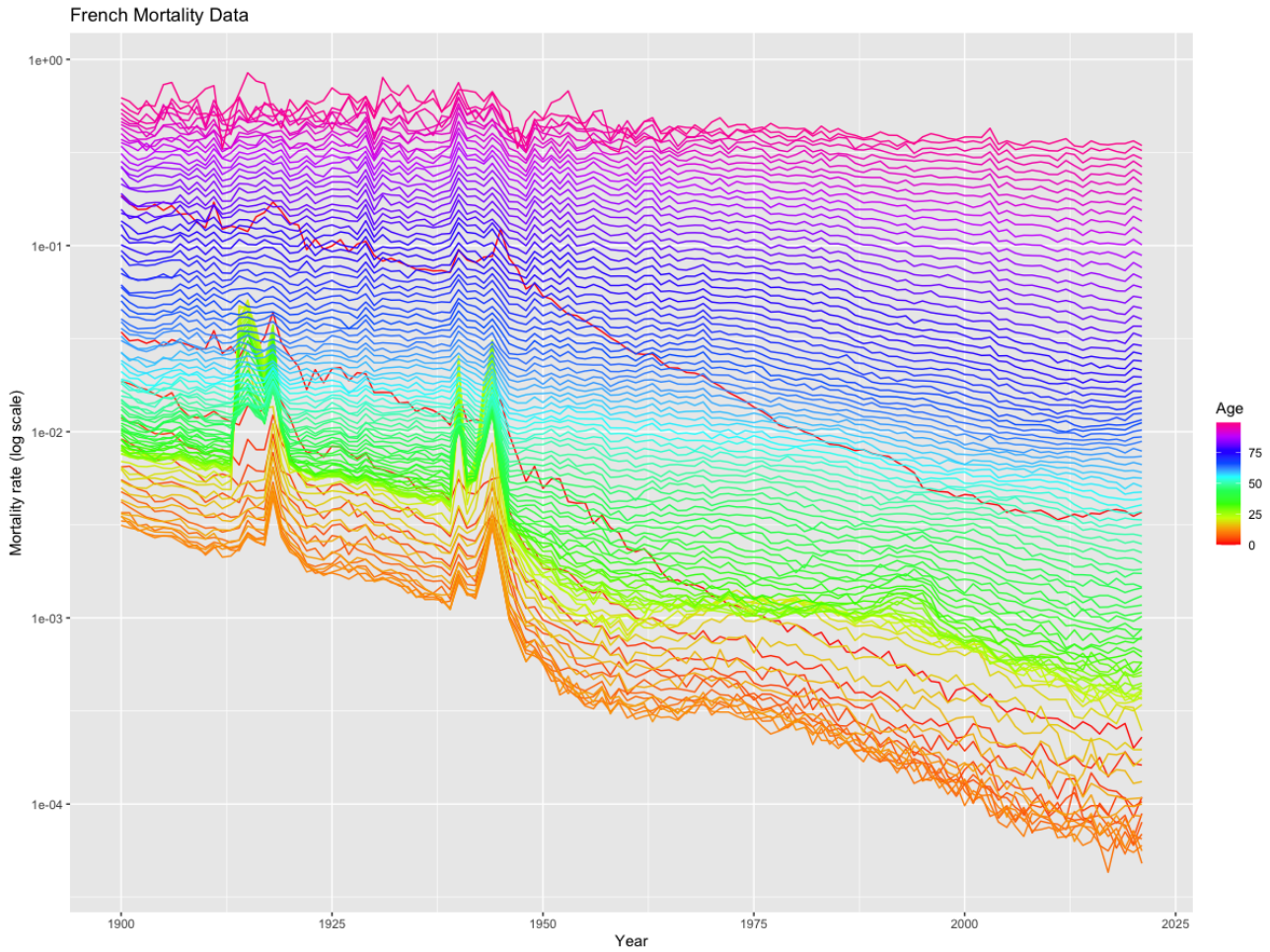
- $\boldsymbol{\kappa}_t = (\kappa_{1t}, \cdots, \kappa_{Jt})'$ is a $J \times 1$ vector of the latent factors
- $\boldsymbol{\Lambda}$ represent the $N \times J$ matrix of factor loadings
- $\boldsymbol{\kappa}_t$ is typically estimated using Singular Value Decomposition

Other specifications for $\boldsymbol{\kappa}_t$: VAR, ARIMA, etc.

US Mortality Data



French Mortality Data



Contents

FM-BART

Factor Model with Bayesian Additive Regression Tree (BART):

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\kappa}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \Sigma) \\ \boldsymbol{\kappa}_t &= F(\boldsymbol{X}) + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, \Sigma_\omega) \end{aligned} \tag{1}$$

- A nonparametric specification for F using BART:

$$F(\boldsymbol{X}) = (f_1(\boldsymbol{X}), \cdots, f_N(\boldsymbol{X})), \quad f_j(\boldsymbol{X}) = \sum_{k=1}^{m} g_{jk}(\boldsymbol{X} \mid T_{jk}, M_{jk})$$

- $\boldsymbol{X}$: lags of $\boldsymbol{\kappa}_t$, Macro/Financial variables
- Here, Σ_ω is assumed to be a diagonal matrix. Please refer to Huber & Rossini (2021) for further discussions.

FM-BART

The FM-BART can handle:

- Nonlinear main effects and multi-way interaction effects
- Outliers caused by extreme events, e.g., wars, pandemics
- Allowing for different smoothness or functional forms for factors
- Incorporation of expert knowledge via prior distributions

BART: Univariate Case

Bayesian Additive Regression Trees (BART) (Chipman et al, 2010) is a Bayesian "sum-of-trees" model, i.e.,

$$y = f(\boldsymbol{x}) + \epsilon = \sum_{j=1}^{m} g(\boldsymbol{x}|T_j, M_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

BART: Univariate Case

Bayesian Additive Regression Trees (BART) (Chipman et al, 2010) is a Bayesian "sum-of-trees" model, i.e.,

$$y = f(\boldsymbol{x}) + \epsilon = \sum_{j=1}^{m} g(\boldsymbol{x}|T_j, M_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- T_j is a binary regression tree
- $M_j = \{\mu_{1,j}, \cdots, \mu_{b_j,j}\}$ is a vector of parameters associated with the b_j terminal nodes of T_j
- The decision rules are binary splits of the predictor space, typically of the form $\{x_i \leq c\}$ or $\{x_i > c\}$
- $g(\boldsymbol{x}|T_j, M_j) = \sum_{i=1}^{b_j} \mu_{i,j} \mathbb{I}\{x_{(k)} \in \mathcal{S}_i\}$

Binary Regression Trees (Huber & Rossini, 2021)

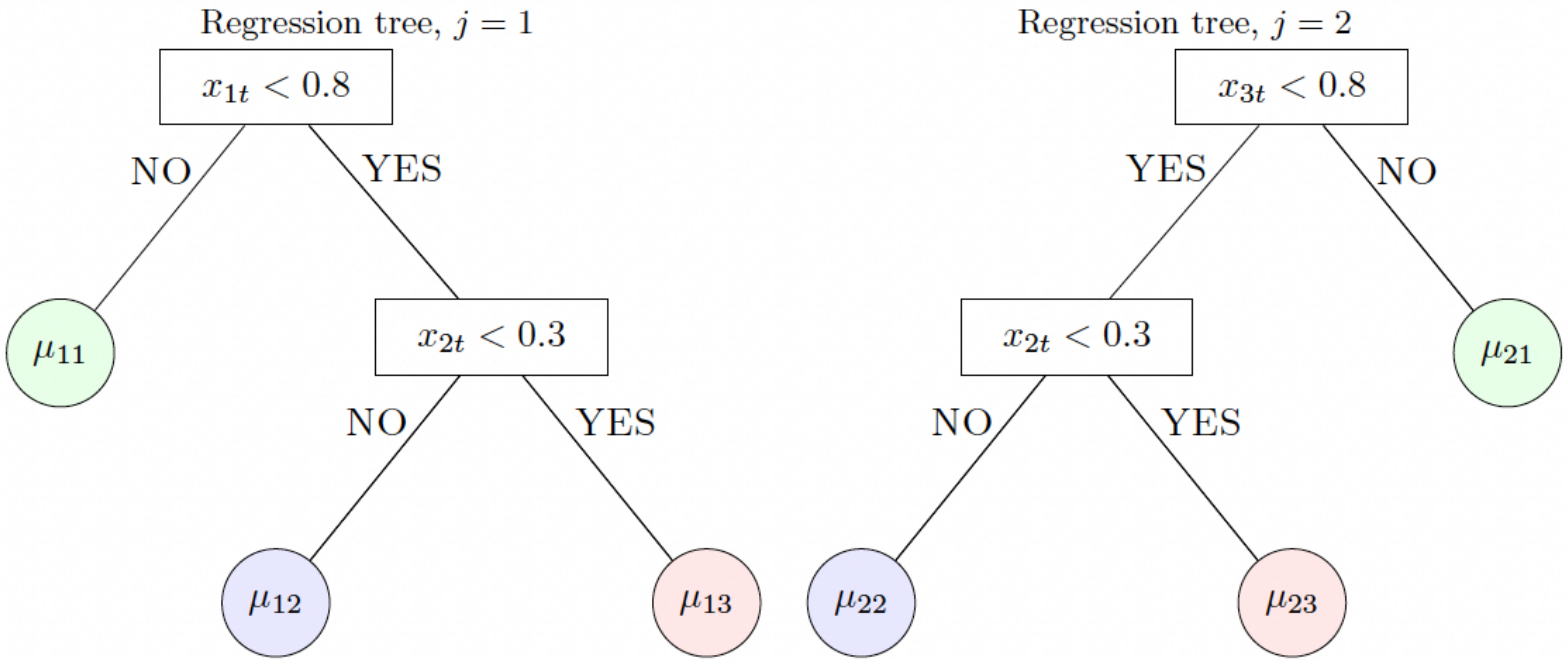


Figure 2: Example of a sum of regression tree, $g(\boldsymbol{x}|\mathcal{T}_j, \boldsymbol{m}_j)$, with internal nodes labeled by their splitting rules and leaf nodes labeled with the corresponding parameters μ_{jk}, where $j = 1, 2$ and $k = 1, 2, 3$.

Bayesian Additive Regression Trees

Characteristics of BART:

- With a large number of trees, a sum-of-trees model gains increased representation flexibility
- Each individual tree is "weakened" by a strongly influential regularization prior
- Fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm

Contents

Bayesian Inference

- Equation by equation estimation: $f_j(\boldsymbol{X}) \approx \sum_{k=1}^{m} g_{jk}(\boldsymbol{X} \mid T_{jk}, M_{jk})$
- **Independent and Symmetric Prior**:

$$p\left(\left(T_{j1}, M_{j1}\right), \cdots, \left(T_{jm}, M_{jm}\right), \sigma_j^2\right)$$
$$= p(\sigma_j^2) \times \prod_{k=1}^{m} \prod_{q=1}^{b_{jq}} p\left(\mu_{jk,q} \mid \mathcal{T}_{jk}\right) \times p\left(\mathcal{T}_{jk}\right)$$

Bayesian Inference

Regularization Priors:

- $\mathcal{T}_{jk}$: Tree structure
  (1) splitting terminal node: the probability that a node at depth d is non-terminal is $\alpha(1+d)^{-\beta}$, $\alpha \in (0,1), \beta \in [0,\infty)$
  (2) splitting rule: predictors and threshold are chosen uniformly at random
- $\mu_{jk,q} \mid \mathcal{T}_{jk}$: $\mu_{jk,q} \sim N(0, \sigma^2_{\mu_j})$, $\sigma_{\mu_j} \propto \frac{1}{\sqrt{m}} \rightarrow$ shrinkage
- σ_j^2: $\sigma_j^2 \sim \nu_j \xi_j / \chi^2_{\nu_j}$;
- The choice of m: fully nonparametric/cross-validation/fixed ($m = 200$)

Bayesian Variable Selection

Splitting rule for the tree: Let $s = (s_1, \cdots, s_L)$ be such that s_i is the probability that predictor x_i is used to construct a given split.

- Typically it is assumed that $s_j = L^{-1}$ so that predictors are chosen uniformly at random.

Bayesian Variable Selection

Splitting rule for the tree: Let $s = (s_1, \cdots, s_L)$ be such that s_i is the probability that predictor x_i is used to construct a given split.

- Typically it is assumed that $s_j = L^{-1}$ so that predictors are chosen uniformly at random.
- High-Dimensional (Linero, 2018): setting s to follow a sparsity-inducing Dirichlet distribution,

$$(s_1, \cdots, s_L) \sim \mathcal{D}\left(\frac{\alpha}{P}, \cdots, \frac{\alpha}{P}\right),$$

where α plays a central role in determining the degree of sparsity.

Backfitting MCMC Algorithm

Define a partial residual as

$$\boldsymbol{R}_{ji} = \boldsymbol{\kappa}_{\cdot \boldsymbol{j}} - \sum_{k \neq i} g_{jk}(\boldsymbol{X} | T_{jk}, M_{jk})$$

A Gibbs sampler entails m successive draws is

1. For i from 1 to m, draw (T_{ji}, M_{ji}) using the Metropolis-Hasting

$$T_{ji} | \boldsymbol{R}_{ji}, \sigma_j^2, \qquad M_{ji} | T_{ji}, \boldsymbol{R}_{ji}, \sigma_j^2$$

2. Draw σ_j^2 from full conditional

$$\sigma_j^2 \mid (T_{j1}, M_{j1}), \cdots, (T_{jm}, M_{jm}), \boldsymbol{\kappa}$$

Bayesian Inference

- One-step-ahead predictive distribution is given by:

$$p\left(\boldsymbol{\kappa}_{T+1} \mid \boldsymbol{\kappa}_{1:T}\right) = \int p\left(\boldsymbol{\kappa}_{T+1} \mid \boldsymbol{\kappa}_{1:T}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{\kappa}_{1:T}\right) d\boldsymbol{\theta},$$

$$\boldsymbol{\kappa}_{T+1} \mid \boldsymbol{\kappa}_{1:T}, \boldsymbol{\theta} \sim N\left(F\left(\boldsymbol{X}_{T+1}\right), \boldsymbol{\Sigma}\right)$$

Bayesian Inference

- One-step-ahead predictive distribution is given by:

$$p\left(\boldsymbol{\kappa}_{T+1} \mid \boldsymbol{\kappa}_{1:T}\right) = \int p\left(\boldsymbol{\kappa}_{T+1} \mid \boldsymbol{\kappa}_{1:T}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{\kappa}_{1:T}\right) d\boldsymbol{\theta},$$

$$\boldsymbol{\kappa}_{T+1} \mid \boldsymbol{\kappa}_{1:T}, \boldsymbol{\theta} \sim N\left(F\left(\boldsymbol{X}_{T+1}\right), \boldsymbol{\Sigma}\right)$$

- Iterative Prediction. Let $\tilde{\boldsymbol{\kappa}}_{T+1}$ denote the draws of $\boldsymbol{\kappa}_{T+1}$. We can draw $\tilde{\boldsymbol{\kappa}}_{T+2}$ from:

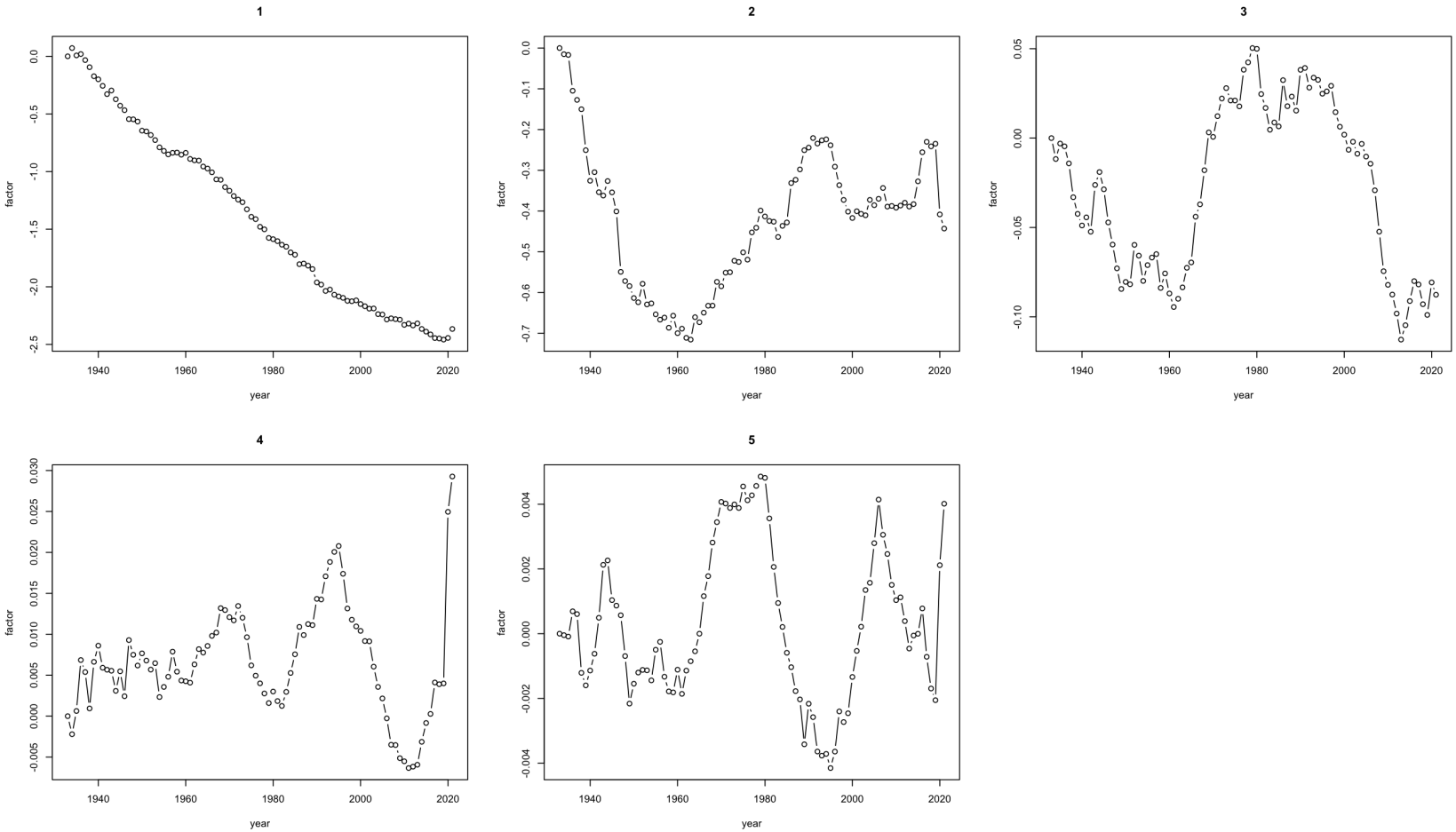
$$\tilde{\boldsymbol{\kappa}}_{T+2} \sim N\left(F\left(\tilde{\boldsymbol{X}}_{T+2}\right), \boldsymbol{\Sigma}\right)$$

Contents

Mortality Data

- Source of Dataset: **Human Mortality Database**
- **Populations**: USA
- **Period**: 1933 $\sim$ 2021
- **Range of ages**: 0 $\sim$ 89

Plots of Factors



In-sample Performances

No. of factors	BART_lag1	BART_lag2	BART_lag4	ARIMA
1	0.00817	0.00833	0.00831	0.00861
2	0.00295	0.00315	0.00303	0.00362
3	0.00187	0.00188	0.00189	0.00539
4	0.00168	0.00161	0.00119	0.00223
5	0.00138	0.00102	0.00086	0.00196

Forecasting Performance Evaluation

- We separate the data into an in-sample period covering 1933 to year $t = 2011, \cdots, 2020$
- In each iteration, obtain the posterior predictive density of the h-step ahead forecast of $\boldsymbol{y}_{t+h}$
- The root mean squared forecast error (RMSFE):

$$\mathbb{E}[\text{RMSFE(h)}] = \frac{1}{n}\sum_{i=1}^{n} \text{RMSFE}^{(\text{i})}(\text{h})$$

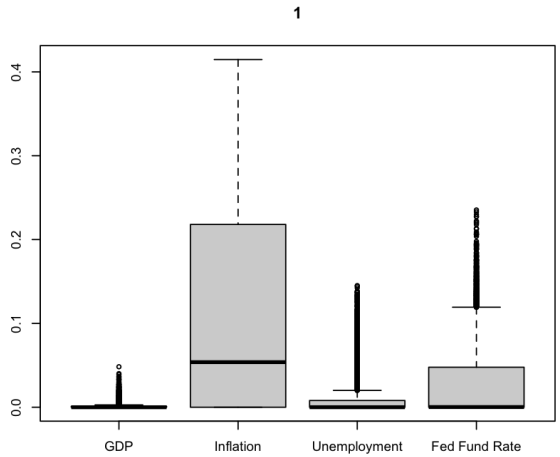
$$= \frac{1}{n}\sum_{i=1}^{n} \sqrt{\frac{\sum_{x=0}^{N}\sum_{t=t_0}^{T-h}[y^{o}_{x,t+h} - \hat{y}^{(i)}_{x,t+h}|\mathcal{F}_t]^2}{(N+1)(T-h-t_0+1)}}$$

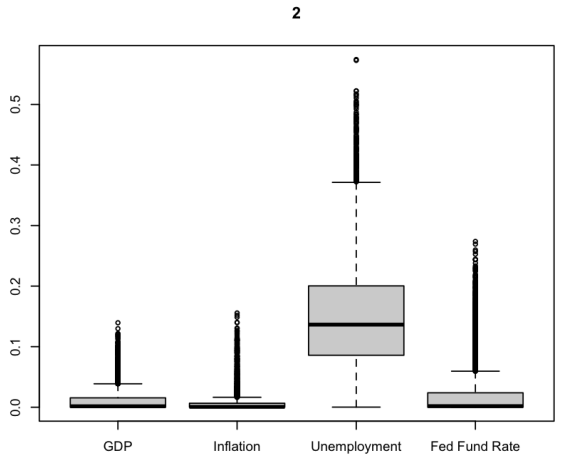
Out-of-sample Performances

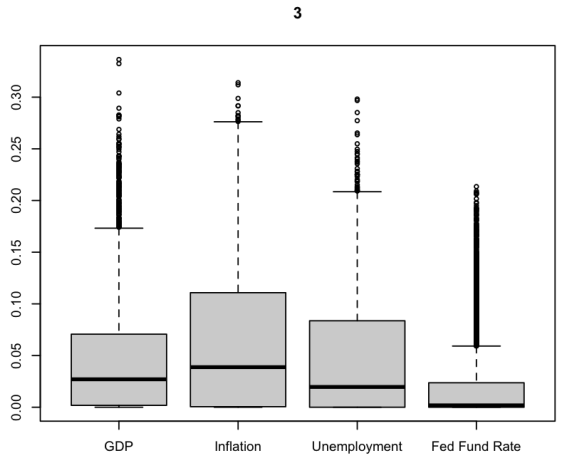
Table: Comparison of Out-of-sample Performances

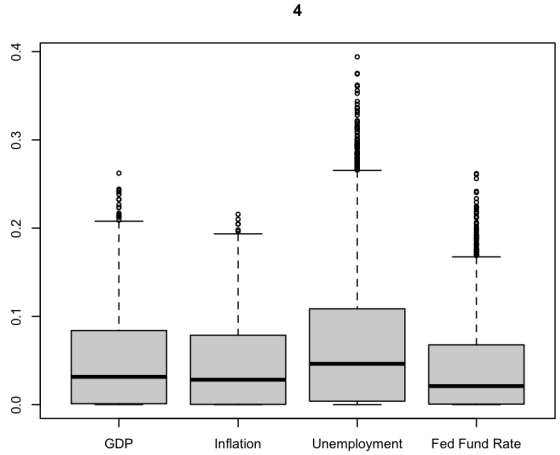
Horizon	FM-BART-1	FM-BART-3	FM-BART-5	FM-RF-1	FM-RF-3	FM-RF-5	FM-ARIMA-1	FM-ARIMA-3	FM-ARIMA-5	Pure RF
1	0.0267	0.0110	0.0091	0.0267	0.0226	0.0632	0.0266	0.0093	**0.0071**	0.0246
2	0.0331	**0.0158**	0.0209	0.0328	0.0225	0.0667	0.0359	0.0183	0.0165	0.0314
3	0.0367	**0.0184**	0.0264	0.0368	0.0327	0.1438	0.0427	0.0234	0.0203	0.0347
4	0.0408	**0.0213**	0.0249	0.0412	0.0354	0.1602	0.0507	0.0274	0.0255	0.0370
5	0.0474	0.0256	**0.0243**	0.0472	0.0343	0.1585	0.0632	0.0353	0.0360	0.0420
6	0.0540	0.0331	**0.0270**	0.0538	0.0317	0.1413	0.0812	0.0511	0.0525	0.0492
7	0.0608	0.0404	0.0405	0.0605	**0.0351**	0.1365	0.1026	0.0663	0.0749	0.0575
8	0.0704	**0.0431**	0.0586	0.0711	0.0507	0.1372	0.1296	0.0848	0.1085	0.0727
9	0.0868	**0.0602**	0.0733	0.0891	0.0655	0.1232	0.1769	0.1316	0.1422	0.0987
10	0.1042	0.0837	**0.0772**	0.1081	0.0840	0.1077	0.2191	0.1710	0.1520	0.1157

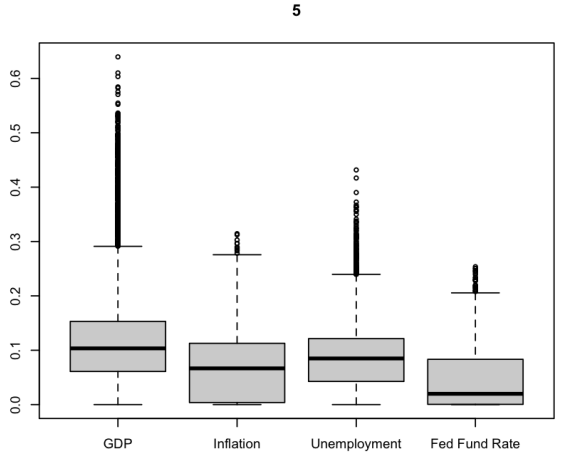
Macro/Finance Effect: Variable Importance











Some Further Considerations

(1) More Empirical Results
  - Explore other countries' data to validate the model.
  - Compare with other Machine Learning methods, eg. neural network

(2) Extend to Multi-population Mortality Modeling
  - Expand the model to encompass different populations for broader applicability.

The End

Thanks!