Expressive Mortality Models through Gaussian Process Compositional Kernels

Insurance Data Science
Bayes Business School, June 2023

Mike Ludkovski

Dept of Statistics & Applied Probability UC Santa Barbara

Overview

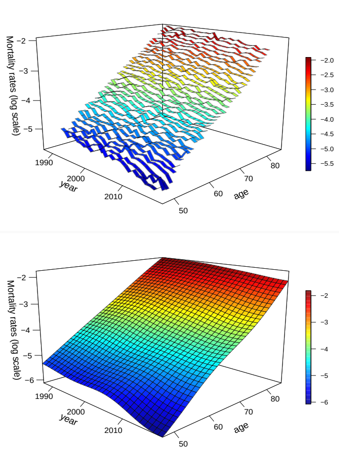
- Expressive longevity modeling w/Gaussian Process models
- Compositional kernel search via Genetic Algorithms
- Proof of concept: synthetic datasets
- Results w/HMD datasets
- Take-aways about mortality surface structures

Joint with Jimmy Risk (Cal Poly Pomona)
Preprint: arxiv:2305.01728

Mortality modeling

- A 2-D table indexed by Age and Year: $x = (x^n_{ag}, x^n_{yr})$
- Raw observed **log-rates** $Y(x^n) = f(x^n) + \epsilon^n$
- Learn $f(\cdot)$ the latent log-mortality surface:
    - Smooth observed mortality experience (remove $\epsilon(x)$)
    - Uncover patterns in mortality evolution and mortality improvement factors
    - Quantify uncertainty (intrinsic; model-driven)
    - Focus on interpretation rather than forecasting

Denmark Male mortality



What are the factors driving mortality?

- Age-Period M1: $f(\boldsymbol{x}) = \alpha(x_{ag}) + \beta(x_{ag})\kappa(x_{yr})$ – Lee & Carter (1992)
- Then add a Cohort term (M3). Then add more terms...
- Dowd-Cairns-Blake (2020) CBDX: $f(\boldsymbol{x}) = \alpha(x_{ag}) + \sum_{i=1}^{I} \beta_i(x_{ag})\kappa_i(x_{yr}) + \gamma(x_{co})$ - adaptive sum $I \in \{1,2,3\}$ of Age-Period + "residual" Cohort term, κ is RW w/drift
- Hunt & Blake (2014): "general procedure" to pick an APC structure
- Gaussian Process Age-Period: $f = \mathcal{GP}(m, k)$ where k is multiplicative in x_{yr}, x_{ag} – L-Risk-Zail (2018)
- Huynh-L (2021): Age-Period-Cohort + multi-population;
- Neural network APC: Perla et al (2021); Richman & Wüthrich (2021)
- **How to flexibly express f?**

Statistical Framework for Gaussian Process Mortality Surfaces

- Input x, true response surface $f(x)$, observations $y(x)$: training dataset $\mathcal{D} = (x^{1:n}, y^{1:n})$
- Specify prior distribution and then compute conditional distribution given the data $p(f|\mathcal{D}) \propto p(\mathbf{y}|f)p(f) = \{likelihood\} \cdot \{prior\}$
- Response surface is a Gaussian random field w/prior $f \sim \mathcal{GP}(m, k)$
- **Covariance** kernel $k(x^i, x^j) = \mathbb{E}[(f(x^i) - m(x^i))\,(f(x^j) - m(x^j))]$
- Observation likelihood $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{\Delta})$, w/$\mathbf{\Delta} = diag(\sigma^2(x^i))$, $\varepsilon(x^i) \sim \mathcal{N}(0, \sigma^2(x^i))$
- Gaussian prior + Gaussian likelihood $\Rightarrow$ Gaussian posterior $f|\mathcal{D} \sim \mathcal{GP}(m_*, k_*)$
- Posterior based on multivariate Gaussian conditioning $f(x)|\mathcal{D} \sim \mathcal{N}(m_*(x), s_*^2(x))$

$$\text{mean: } m_*(x) = \mathbf{k}(x)^T \underbrace{(\mathbf{K} + \mathbf{\Delta})^{-1}\mathbf{y}}_{=:c}, \qquad K_{ij} = k(x^i, x^j), k_i = K(x, x^i)$$

$$\text{cov: } s_*(x, x') = K(x, x') - \mathbf{k}(x)^T(\mathbf{K} + \mathbf{\Delta})^{-1}\mathbf{k}(x')$$

- Fitting: learn the hyperparameters controlling the covariance structure

Expressive GP Kernels

Kernel Families: Lots of Choices

- Kernel k determines all structural properties: (non)stationarity, smoothness of the GP mean and sample paths
- Default choice is a multiplicative + separable. Ex: RBF Age-Period kernel (LRZ 2018)

$$k(x,x') = \eta^2 \exp\left(-\frac{(x_{ag}-x'_{ag})^2}{2\ell_{ag}^2}\right) \cdot \exp\left(-\frac{(x_{yr}-x'_{yr})^2}{2\ell_{yr}^2}\right) = k_{\mathrm{RBF}}(x_{ag},x'_{ag}) \cdot k_{\mathrm{RBF}}(x_{yr},x'_{yr})$$

Kernel Name	Abbv.	Formula $k(x,x';\theta)$	Properties	$\mathcal{K}_r$
Matérn-1/2	M12	$\exp\left(-\frac{\lvert x-x'\rvert}{\ell_{\mathrm{len}}}\right), \quad \ell_{\mathrm{len}} > 0$	C^0	✓
Matérn-3/2	M32	$\left(1+\frac{\sqrt{3}}{\ell_{\mathrm{len}}}\lvert x-x'\rvert\right)\exp\left(-\frac{\sqrt{3}}{\ell_{\mathrm{len}}}\lvert x-x'\rvert\right), \quad \ell_{\mathrm{len}} > 0$	C^1	
Matérn-5/2	M52	$\left(1+\frac{\sqrt{5}}{\ell_{\mathrm{len}}}\lvert x-x'\rvert+\frac{5}{3\ell_{\mathrm{len}}^2}\lvert x-x'\rvert^2\right)\exp\left(-\frac{\sqrt{5}}{\ell_{\mathrm{len}}}\lvert x-x'\rvert\right)$	C^2	✓
Cauchy	Chy	$\frac{1}{1+\lvert x-x'\rvert^2/\ell_{\mathrm{len}}^2}, \quad \ell_{\mathrm{len}} > 0$	C^∞	
Radial Basis	RBF	$\exp\left(-\frac{(x-x')^2}{2\ell_{\mathrm{len}}^2}\right), \quad \ell_{\mathrm{len}} > 0$	C^∞	✓
AR2	AR2	$\exp(-\alpha\lvert x-x'\rvert)\left\{\cos(\omega\lvert x-x'\rvert)+\frac{\alpha}{\omega}\sin(\omega\lvert x-x'\rvert)\right\}$	Periodic, C^1	
Linear	Lin	$\sigma_0^2 + x\cdot x', \quad \sigma_0 > 0$	Non-stationary	*
Minimum	Min	$t_0^2 + x \wedge x', \quad t_0 > 0$	Non-stat, C^0	✓
Mehler	Meh	$\exp\left(-\frac{\rho^2(x^2+x'^2)-2\rho xx'}{2(1-\rho^2)}\right), \quad -1 \le \rho \le 1$	Non-stationary	

Compositional Kernel Search

- Interested in recovering mortality dependence structure from data
- Cast a broad net to seek the "best" kernel
- Idea of "Automatic Model Construction with Gaussian Processes" (Duvenaud, 2015): look at thousands of potential kernels
- Extract ~ 100 best-fitting kernels for a given population and analyze this aggregate collection:
  - Smoothness of mortality experience across Age and across Year
  - Presence/absence of a Cohort effect
  - Additive structures (linking to multi-scale) vs classical multiplicative APC
  - Relative structures across populations (how does discovered structure vary; which countries have more "complex" mortality patterns)
- Analogue of the "general procedure" in APC frameworks

Searching Through Kernels

- Space of kernels has nice **algebraic properties**
- Kernels are stable under addition $(k_1 + k_2)$ and multiplication $(k_1 \cdot k_2)$
- Index kernels by Age k_a; Period/Year k_y and birth Cohort k_c
- Consider about a dozen of common GP families, compose them through `add` & `mult`
- e.g $\kappa =$ `add(Exp_c, mul(RBF_a, add(Mat_y, RBF_c)))` corresponds to

$$(k_{M52}(x_{yr}) + k_{RBF}(x_c)) \cdot k_{RBF}(x_{ag}) + k_{Exp}(x_c)$$

- Kernel **length**: number of terms $|\kappa| = 7$ above: 4 base kernels + 3 operators
- Compare kernels via BIC (log marginal likelihood of data + complexity penalty)

$$BIC(k) = -\ell_k(\hat{\theta}; \mathrm{y}) + \frac{|\hat{\theta}| \log(\mathrm{n})}{2}$$

- Bayes Factor: $\mathrm{BF}(k_1, k_2) = \frac{p(k_1|\mathrm{y})}{p(k_2|\mathrm{y})} \approx \exp\Big(\mathrm{BIC}(k_2) - \mathrm{BIC}(k_1)\Big)$ to assess significance

How to Search: Genetic Algorithm

- Represent kernels via a binary tree
- Mutation-selection to propagate the "fittest" kernel-trees across generations
- Generation **0**: Randomly select n_g kernels
- Generation $\boldsymbol{g}$:
  - Sample fit parents from the $g-1$ generation (based on **BIC**)
  - Evolve them (mutate, crossover, replace operations) into a new offspring
  - Add offspring to generation g
- Repeat for $g = 1, 2, \ldots, G$

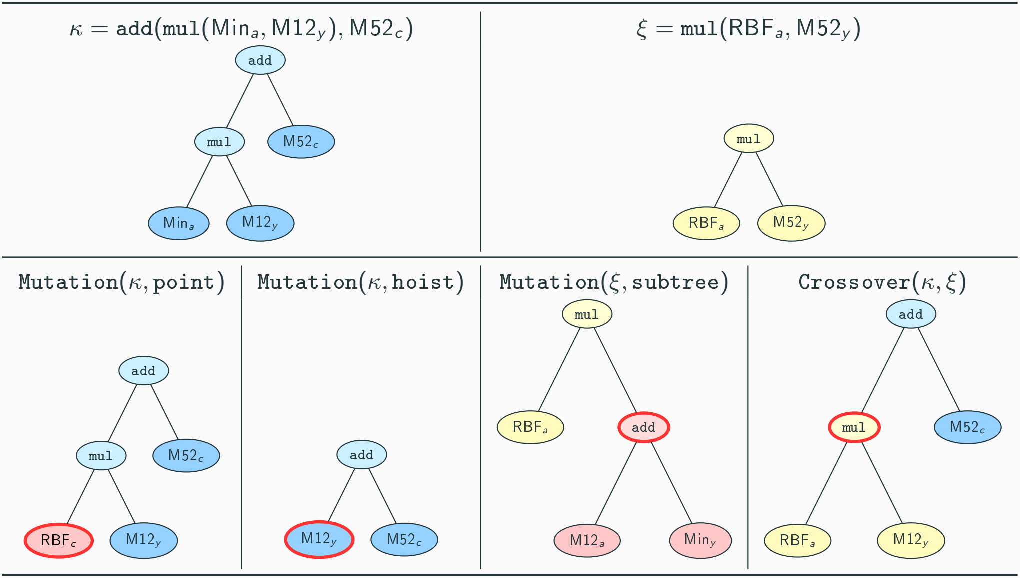


Figure 1: Representative compositional kernels and GA operations. Bolded red ellipses indicate the node of κ (or ξ) that was chosen for mutation or crossover.

GA Implementation

- Fit GPs using the `GPyTorch` library in Python
- Maximize $\ell_k(\theta|y)$ via Adam SGD
- Standardize inputs into $[0, 1]^2$
- Use $n_g = 200$ kernels per generation and $G = 20$ generations (a total of 4000 candidates)
- Tends to converge after 10-12 generations
- Double tournament of size $T = 7$ to select ancestors
- Some customization regarding the relative probability of mutation operations and how to initialize the zeroth generation
- Big potential challenge of GA: bloat (want kernel length ≤ 15 or so)
- Largely follow Luke & Panait (2006); Poli et al (2008); Sipper et al (2018)

Results

Synthetic Experiments

- Can the GA recover the true structure?
- Can the GA detect additivity?
- Is the GA stable?

Three synthetic datasets (35 ages x 28 years) generated with a specified GP K_0

Exprmnt	Ground Truth Kernel	$\sigma^2(x)$	β_0	β_{ag}
SYA	$0.04 \cdot \text{RBF}_a(0.4) \cdot \text{RBF}_y(0.3)$	0.001	-5.0	3.4
SYB	$0.08 \cdot \text{RBF}_a(0.586) \cdot M12_y(13.33) + 0.02 \cdot M52_c(0.079)$	0.0004	-5.568	2.974
SYC	$0.0134 \cdot M52_a(1.132) \cdot \text{Min}_y(0.877) \cdot M12_c(96.234) \cdot$ $\cdot \text{Meh}_c(0.8483)$	$1.0783/D_x$	-3.165	3.380

Table 1: Description of synthetic data sets. Data is generated with prior mean $m(x) = \beta_0 + \beta_{ag} x_{ag}$. SYA and SYB are homoskedastic. In generating SYC's heteroskedastic noise, D_x comes from the JPN Female data.

Synthetic Results

SYA-1			SYA-2		
BIC	$\widehat{\mathrm{BF}}(k, K_0)$	Kernel	BIC	$\widehat{\mathrm{BF}}(k, K_0)$	Kernel
-2034.23	**1.0000***	$\mathbf{RBF_a\,RBF_y}$	-2066.93	1.1907***	$M52_a\,RBF_y$
-2034.04	0.8264***	$M52_a\,RBF_y$	**-2066.76**	**1.0000***	$\mathbf{RBF_y\,RBF_a}$
-2031.82	0.0902*	$M52_a M52_y$	-2064.63	0.1216**	$M52_a M52_a\,RBF_y$
-2031.29	0.0526*	$M52_a\,RBF_a\,RBF_y$	-2064.24	0.0801*	$M52_a\,RBF_a\,RBF_y$
-2031.09	0.0433*	$M52_a M52_a\,RBF_y$	-2063.88	0.0561*	$M52_a M52_a\,RBF_y$

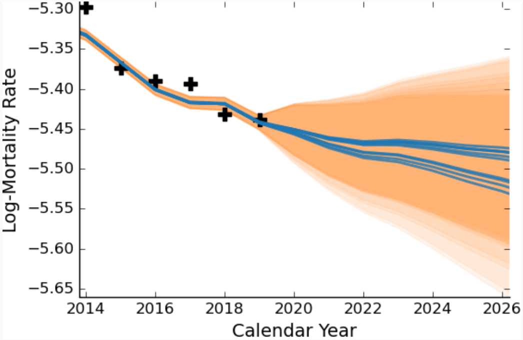
Table 2: Top five fittest non-duplicate kernels for the first synthetic case study SYA. Bolded is $K_0 = RBF_y\,RBF_a$, the true kernel used in data generation. SYA-1 and -2 denote the realization trained on.

- GA finds the true optimum for SYA (+2 plausible alternatives)
- Correctly identifies the # of terms and the additive age × year + cohort structure for SYB
- Correctly identifies the # of terms and the multiplicative structure for SYC
- Closely recovers the ground truth GP hyperparameters
- Can fully distinguish relative smoothness in Age and Year
- Stable results across re-runs
- Validates GA convergence

Application to HMD National Datasets

Human Mortality Database:

- **Four** representative datasets:
    - different pop'n size;
    - different demographics;
    - both genders
- JPN Females and Males
- US Males; SWE Females
- Years 1990–2018 and ages 50–84



Predictions from the top 10 kernels in $\mathcal{K}_f$ for JPN Females Age 65. We show the predictive mean and 90% posterior interval from the top-10 kernels, as well as the observed log-mortality rates (+) during 2014–2019.

Illustration: Japan Females

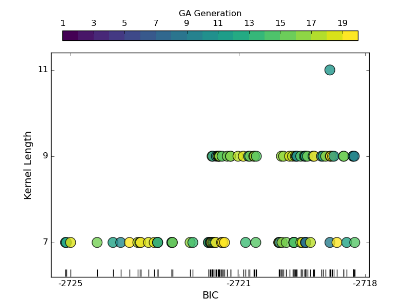
Lowest BIC: $k^*_{JPN-FEM} = 0.4638 \cdot M52_a(1.11) \cdot Chy_y(1.95) \cdot M12_y(62.42) \cdot M12_c(117.11)$.

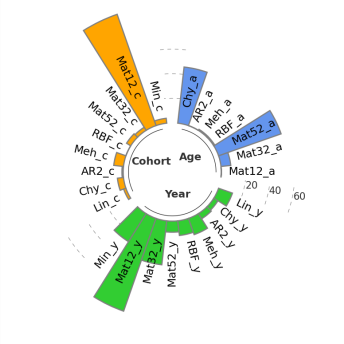
Japan Females during 1990-2018 and Ages 50-84		
BIC	$\widehat{\text{BF}}$	Kernel
-2725.293	1	$M52_a(Chy_y\ M12_y)M12_c$
-2725.270	$0.977^\dagger$	$M52_a(M52_y M12_y)M12_c$
-2725.221	$0.931^\dagger$	$M52_a(M52_y\ Min_y)M12_c$
-2724.623	$0.512^\dagger$	$M52_a(M52_y M12_y)\ Min_c$
-2724.510	0.457	$M52_a(M32_y M12_c)M12_c$

Above: fittest non-duplicate kernels for HMD Japanese Females over $\mathcal{K}_f$. Bayes Factors $\widehat{\text{BF}}$ are relative to the best $k^*_{JPN-FEM}$ and none are significant. Daggered kernels also belong to $\mathcal{K}_r$.

Top Right: Properties of top 100 kernels

Bottom Right: Frequency of different kernels among top 100 candidates





GA Results based on searching within the full set $\mathcal{K}_f$

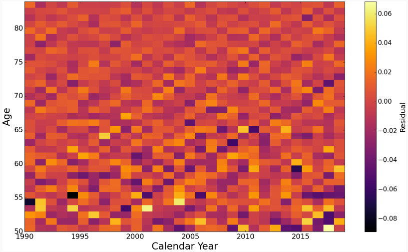
Range	BIC max	BIC min	len	addtv comps	non-stat.	num age	num year	num coh	rough age	rough year	rough coh
JPN Female											
1-10	-2723.68	-2725.29	4.00	1.00	0%	1.00	1.80	1.20	0%	100%	100%
1-50	-2720.64	-2725.29	4.34	1.08	10%	1.12	1.90	1.32	0%	100%	100%
51-100	-2718.24	-2720.62	4.60	1.20	18%	1.12	2.20	1.28	0%	100%	100%
JPN Male											
1-10	-2978.43	-2980.53	4.10	1.00	0%	1.00	1.60	1.50	0%	100%	100%
1-50	-2975.36	-2980.53	4.26	1.10	0%	1.06	1.70	1.50	18%	100%	100%
51-100	-2974.25	-2975.32	4.60	1.00	0%	1.04	2.14	1.42	64%	100%	100%
US Male											
1-10	-3163.54	-3170.29	5.70	2.30	0%	1.50	1.50	2.70	100%	100%	100%
1-50	-3160.32	-3170.29	5.78	2.24	0%	1.40	1.54	2.84	100%	100%	100%
51-100	-3157.93	-3160.24	6.14	2.38	2%	1.46	1.72	2.96	100%	100%	98%
SWE Female											
1-10	-1624.34	-1625.57	3.00	1.00	0%	1.00	1.00	1.00	0%	100%	0%
1-50	-1622.74	-1625.57	3.02	1.00	6%	1.00	1.24	0.78	0%	100%	14%
51-100	-1622.04	-1622.74	3.42	1.04	16%	1.10	1.38	0.94	0%	100%	6%

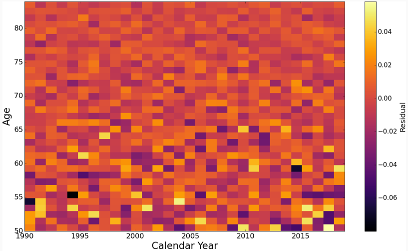
Discussion

- Additive vs Multiplicative Structure
  - Generally, multiplicative APC is sufficient: find evidence for additivity only in US
  - Often, the found kernel has several multiplicative terms in the same coordinate
  - Interpret as (i) multi-scale effects; (ii) insufficient fit with the selected base kernels
  - When kernel is additive, one term tends to dominate. Interpret as primary effect + correction/residual (à la boosted models)
- Kernel smoothness confirms accepted folklore:
  - Rough (non-differentiable) in Period and Cohort
  - Smooth (at least twice-differentiable) in Age
  - Potentially non-stationary (i.e. random-walk like) Period effect
  - Roughness in Period is driven by environmental (vs idiosyncratic that is smoothed) noise
- Substitution effect: often observe multiple plausible (BIC-wise) alternatives:
  - E.g M52/ RBF / Chy are close substitutes
  - Min and M12 also often substituted
  - Alternates yield very similar predictions and log-likelihood
  - Effect amplified as the search space is increased

Cohort Effect

- Overwhelming evidence for cohort effect in Japan and US
- BIC differences of 6+ (Bayes factors of 100+)
- Clear deterioration of residual heatmaps if remove Cohort
- Top panel: Japan Female **w/out Cohort**; bottom: w/Cohort
- Less obvious cohort effect in Sweden (confirming prior discussion)





Conclusion

No one-size-fits-all:

- Mortality experiences are heterogeneous across populations
- Need expressive kernels for a proper fit
- GA + GP is a powerful, interpretable tool to discover structure

Whereto next:

- Multi-population analysis (Huynh & L, 2022, 2023)
- Noise modeling
- Bayesian model averaging

Thank You!

References

Williams, C. K. and Rasmussen, C. E. 2006.
Gaussian processes for machine learning, the MIT Press.

M. Ludkovski, J. Risk, H. Zail
Gaussian Process Models for Mortality Rates and Improvement Factors
ASTIN Bulletin, 48(3), pp. 1307–1347, 2018
Reproducible R notebook: github.com/jimmyrisk/GPmortalityNotebook

N. Huynh, M. Ludkovski
Multi-Output Gaussian Processes for Multi-Population Longevity Modeling
Annals of Actuarial Science, 15(2), 318-345, 2021 arXiv:2003.02443

N. Huynh, M. Ludkovski, H. Zail
Multipopulation Longevity Analysis: a Spatial Random Field Approach
SOA 2020 Living to 100 Symposium

N. Huynh, M. Ludkovski
Joint Models for Cause-of-Death Mortality in Multiple Populations
Annals of Actuarial Science, to Appear, 2023 arxiv:2111.06631

M. Ludkovski, J. Risk
Expressive Mortality Models through Gaussian Process Kernels
arxiv:2305.01728, 2023

Best Found Kernels

Pop'n/Search Set	N_{pl}	Top Kernel
JPN Female $\mathcal{K}_r$	90	$0.464 \cdot \mathrm{M52}_a(1.1) \cdot \mathrm{RBF}_y(1.33)\mathrm{M12}_y(62.51) \cdot \mathrm{M12}_c(118.06)$
JPN Female $\mathcal{K}_f$	95	$0.4638 \cdot \mathrm{M52}_a(1.11) \cdot \mathrm{Chy}_y(1.95)\mathrm{M12}_y(62.42) \cdot \mathrm{M12}_c(117.11)$
JPN Male $\mathcal{K}_r$	89	$0.1491 \cdot \mathrm{M52}_a(0.95) \cdot \mathrm{RBF}_y(1.15)\mathrm{M12}_y(26.24) \cdot \mathrm{M12}_c(24.90)$
JPN Male $\mathcal{K}_f$	112	$0.2130 \cdot \mathrm{M52}_a(1.09) \cdot \mathrm{M12}_y(39.09) \cdot M32_c(0.86)\mathrm{M12}_c(40.73)$
US Male $\mathcal{K}_r$	57	$0.017 \cdot \mathrm{M12}_a(5.04) \cdot \mathrm{M52}_y(0.50)M12_y(10.33) \cdot \mathrm{M52}_c(0.36)\mathrm{M12}_c(5.00)$
US Male $\mathcal{K}_f$	35	$0.01 \cdot \mathrm{AR2}_a(1.12, 1.88) \cdot \mathrm{M12}_y(24.18) \cdot \mathrm{M32}_c(0.72) \cdot [4.6211 \cdot \mathrm{M12}_c(13.49) + 0.01 \cdot \mathrm{M32}_a(0.02) \cdot \mathrm{M52}_c(0.1)]$
SWE Female $\mathcal{K}_r$	200+	$0.2527 \cdot \mathrm{RBF}_a(0.52) \cdot \mathrm{M12}_y(73.74) \cdot \mathrm{RBF}_c(0.62)$
SWE Female $\mathcal{K}_f$	200+	$0.2094 \cdot \mathrm{Chy}_a(1.05) \cdot \mathrm{M12}_y(67.27) \cdot \mathrm{Meh}_c(0.60)$

Table 3: Best performing kernel in $\mathcal{K}_r$ and $\mathcal{K}_f$ for each of the 4 populations considered. N_{pl} is the number of alternate kernels that have a BIC within 6.802 of the top kernel and hence are judged "plausible" based on the BF criterion.

Stability check by re-estimating with a slightly larger dataset (+2 years, +2 age groups):

$$\text{original } \mathcal{D}: \; 0.4651 \cdot \mathrm{M52}_a(1.11) \cdot \mathrm{M52}_y(1.80) \cdot \mathrm{M12}_y(62.79) \cdot \mathrm{M12}_c(117.65);$$

$$\text{enlarged } \mathcal{D}_{rob}: \; 0.4646 \cdot \mathrm{M52}_a(1.11) \cdot \mathrm{M52}_y(1.80) \cdot \mathrm{M12}_y(62.72) \cdot \mathrm{M12}_c(117.50).$$

Comparing Scenarios of Future Mortality

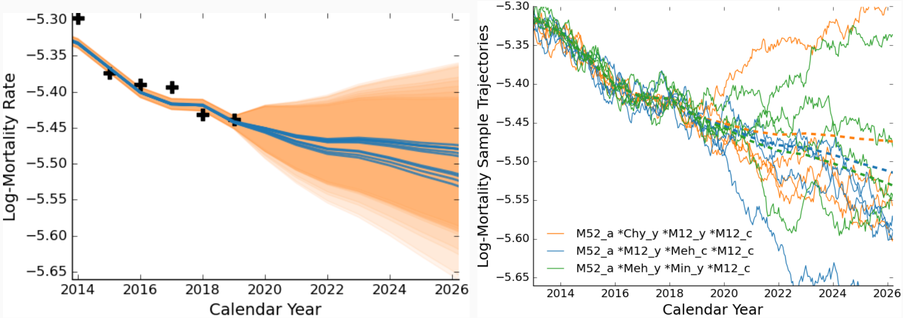
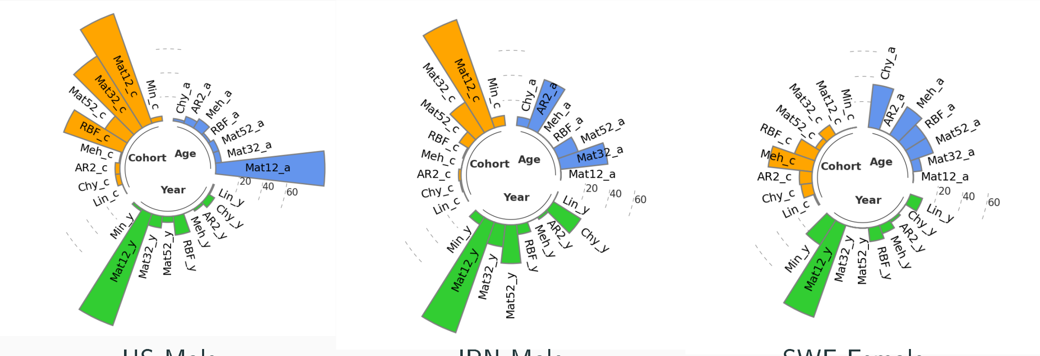
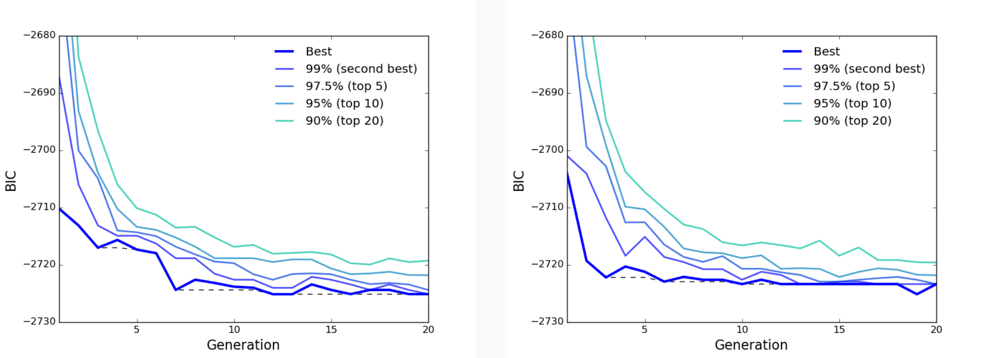


Figure 2: Predictions from the top 10 kernels in $\mathcal{K}_f$ for JPN Females Age 65. *Left:* predictive mean and 90% posterior interval from the top-10 kernels. For comparison we also display (black plusses) the 5 observed log-mortality rates during 2014–2019. *Right:* 4 sample paths from 3 representative kernels.



US Male JPN Male SWE Female

Figure 3: Frequency of appearance of different kernels from $\mathcal{K}_f$ in US, SWE and JPN Male models.



Main run   Re-run

Figure 4: Summary statistics of best kernels proposed by GA as a function of generation g.