Astin Bulletin (2024), **54**, pp. 327–359
doi:10.1017/asb.2023.39

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

Expressive mortality models through Gaussian process kernels

Jimmy Risk[1] and Mike Ludkovski[2]

[1]Mathematics & Statistics, Cal Poly Pomona, Pomona, CA 91676, USA and [2]Statistics & Applied Probability, University of California, Santa Barbara, CA 93106-3110, USA
Corresponding author: Jimmy Risk; Email: jrisk@cpp.edu

Received: 21 June 2023; **Revised:** 22 December 2023; **Accepted:** 22 December 2023;
First published online: 15 February 2024

Keywords: Gaussian process kernel engineering; genetic algorithms; mortality surfaces; Human Mortality Database

Abstract

We develop a flexible Gaussian process (GP) framework for learning the covariance structure of Age- and Year-specific mortality surfaces. Utilizing the additive and multiplicative structure of GP kernels, we design a genetic programming algorithm to search for the most expressive kernel for a given population. Our compositional search builds off the Age–Period–Cohort (APC) paradigm to construct a covariance prior best matching the spatio-temporal dynamics of a mortality dataset. We apply the resulting genetic algorithm (GA) on synthetic case studies to validate the ability of the GA to recover APC structure and on real-life national-level datasets from the Human Mortality Database. Our machine learning-based analysis provides novel insight into the presence/absence of Cohort effects in different populations and into the relative smoothness of mortality surfaces along the Age and Year dimensions. Our modeling work is done with the PyTorch libraries in Python and provides an in-depth investigation of employing GA to aid in compositional kernel search for GP surrogates.

1. Introduction

Gaussian process (GP) models (Ludkovski *et al.*, 2018; Huynh and Ludkovski, 2021b,a) provide a nonparametric spatio-temporal paradigm for longevity analysis within Age–Period–Cohort (APC) modeling. This approach runs parallel to the existing APC models and the newer deep learning-driven approaches (Nigri *et al.,* 2019; Perla *et al.,* 2021; Richman and Wüthrich, 2021). The underlying prediction belongs to the class of *spatial smoothers* and is similar to smoothing splines (Hastie and Tibshirani, 1990). Among the main strengths of GPs are their flexibility, uncertainty quantification, and capabilities for multi-population analysis. Moreover, through their *covariance kernel*, GPs offer a direct view into the inter-dependence of Age–Year-specific mortality rates, which enables the modeler to focus on capturing the respective covariance structure. The covariance kernel of a GP determines the properties of its distribution, including its posterior mean function, smoothness, and more. This offers valuable insight into the underlying dynamics of the process of interest, which is not possible with black-box methods like neural networks.

Matching the APC decomposition of the two-dimensional Age–Year mortality service into three univariate directions, one may consider kernels that reflect the Age structure of mortality, its evolution in time, and its cohort effects. In the existing GP mortality literature, this is straightforwardly translated into a *separable* GP kernel – a product of a univariate kernel in Age, a univariate kernel in Year, and if desired, a univariate kernel in Birth Cohort. While offering a satisfactory performance, this choice is quite restrictive and handicaps the ability of GPs to discover data-driven dependence. With this motivation in mind, we explore GP kernel composition and discovery for mortality models. Our first goal in this article is thus to unleash an automated process for finding the covariance structures most appropriate

for mortality analysis. One motivation is that different (sub-)populations have *different* APC structures, and hence a one-size-fits-all approach is inadequate.

To accomplish this, we propose a new variant of a genetic programming algorithm that iteratively explores the kernel space to discover the most suitable kernels. Our approach tailors previous proposals for kernel discovery using genetic algorithms (GAs) to longevity analysis. We represent APC models through a tree structure comprising addition and multiplication of Age, Year, and Cohort terms and utilize specialized mutation operations to explore such compositions. We assign probabilistic weights to each discovered mortality structure based on the Bayesian information criterion (BIC), indicating its plausibility for a given population. In turn, relative likelihood of two APC structures can be compared based on their Bayes' factors. Through considering several synthetic mortality surfaces, we validate our GA's ability to recover known APC structures. In particular, the GA successfully identifies the presence of additive versus multiplicative effects, the presence of specific terms (such as cohort effect or a nonstationary effect), and the overall complexity of the mortality dependence structure.

Our second motivation is to link ideas in mortality modeling literature to the structures of different GP kernel families. We investigate a variety of kernels, vastly expanding upon the limited number of kernels (such as Squared Exponential and Matérn) that have been considered for mortality so far. By introducing and testing new GP kernel families, we remove the limitation of directly postulating the kernel family to be used, which leads to hidden restrictions and assumptions on the data. Furthermore, additional kernels specifically represent richer structures including random walk, periodicity, and more general ARIMA processes. We observe that existing APC covariance structures, including the well-known Lee–Carter (Lee, 2000) and Cairns–Blake–Dowd (CBD) (Cairns *et al.,* 2011) families, can be exactly matched through additive GP kernels. By testing various kernels, one can find better fits for a mortality surface and answer-related questions, for example, involving the strength or structure of a cohort effect.

Our core approach to above is compositional kernel search. Kernel composition utilizes the fact that kernels are closed under addition and multiplication. On the one hand, compositional kernels offer a rich and descriptive structure of underlying mortality dynamics. On the other hand, they naturally fit the "general procedure" (Hunt and Blake, 2014), already used in the mortality literature, that adds and multiplies APC components. Such compositions and modifications are already performed in the aforementioned Lee–Carter and CBD models (see Cairns *et al.*, 2011 for thorough discussion).

The workhorse of our analysis is a GA that uses the concept of generations to gradually discover better-and-better kernels through a mutation–selection mechanism. Given a mortality dataset, the GA described below generates vast quantities (in the thousands) of potential kernels. These kernels are sequentially fitted to the dataset and ranked according to a statistical fitness function. The GA then probabilistically promotes exploration of the most fit kernels and discards less fit ones. This procedure allows to automate the exploration of the best-performing GP models for mortality modeling. Early proposals for compositional kernels with GPs involved forward search minimizing the BIC to construct tree-based representations of kernels (Duvenaud *et al.,* 2013; Duvenaud, 2014). Jin (2020) and Roman *et al.* (2021) build upon the idea using a GA; the former analyzes performance on several multi-dimensional synthetic test functions, and the latter on univariate time series. We extend and tailor these strategies for mortality modeling, contributing to the GP and GA methodological literatures.

Armed with the outputs of the GA, we address the following fundamental questions about mortality surfaces, which are of intrinsic interest:

- The presence, or lack thereof, of a *Cohort effect*. Our method offers a rigorous Bayesian non-parametric evidence on whether including Birth Cohort effect is beneficial. Since cohort effects are known to be population-specific, this is an important *model selection* question.
- The relative smoothness between the Age- and Year-covariance structures. Classic APC models assume a random-walk structure in calendar time, and (implicitly) a smooth (infinitely differentiable) structure in Age. In contrast, existing GP models have postulated a fixed smoothness (e.g., twice differentiable) in both coordinates. Our method sheds light on whether those

assumptions impact predictions and how much smoothness is most consistent with mortality data.
- Additive versus multiplicative structure in mortality covariance. There have been many proposals and comparative analyses of APC models that variously combine Age and Year terms. We provide an analogous analysis for GP models. In particular, our approach is able to quantify the *complexity* of the best-fitting kernels, giving new insights about how many different terms are necessary.

The rest of the article is organized as follows. In Section 2, we review GP models for mortality surfaces, emphasizing the primordial task of kernel selection and illustrating its impact on model predictions. In Section 3, we develop the GA tool for compositional kernel search. Validity of the GA methodology is asserted in Section 4 through a recovery of known kernels on synthetic mortality surfaces. Section 5 analyzes the output of GA for the initial case study of JPN Females. Section 5.4 then provides a cross-sectional analysis across multiple national-level datasets to address the two questions of presence/absence of Cohort effects and the relative smoothness in Age and Calendar Year. Section 6 concludes.

2. GP models for mortality

A GP is a collection of random variables $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$, such that for any $\ell \in \mathbb{N}$ and $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\} \subseteq \mathbb{R}^d$, the vector $\left[f(\mathbf{x}_1), f(\mathbf{x}_2), ..f(\mathbf{x}_\ell)\right]^\top$ has a multivariate normal distribution (Williams and Rasmussen, 2006) (denoted MVN). For mortality surfaces over APC, $d = 3$. A GP is uniquely defined by its mean function $m : \mathbb{R}^d \to \mathbb{R}$ and covariance kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (Adler, 2010). The kernel $k(\cdot, \cdot)$ must be a symmetric positive-definite function. In this case, for $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$,

$$\mathbb{E}[f(\mathbf{x}_i)] = m(\mathbf{x}_i), \tag{2.1}$$

$$\operatorname{cov}\big(f(\mathbf{x}_i), f(\mathbf{x}_j)\big) = k\big(\mathbf{x}_i, \mathbf{x}_j\big), \tag{2.2}$$

and we write $f \sim \mathcal{GP}(m, k)$. The GP regression model assumes

$$y := y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x}), \tag{2.3}$$

where f is a GP with prior mean $m(\cdot)$ and covariance $k(\cdot, \cdot)$, ϵ is a noise term, and y is a noisy observation. By assuming $\epsilon(\cdot)$ is independent Gaussian white noise with variance $\sigma^2(\cdot)$, properties of multivariate normal random variables imply that $\{y(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ is a GP with mean and covariance functions

$$\mathbb{E}[y(\mathbf{x})] = m(\mathbf{x}), \qquad k_y\big(\mathbf{x}_i, \mathbf{x}_j'\big) = k\big(\mathbf{x}_i, \mathbf{x}_j\big) + \sigma^2(\mathbf{x}_i)\delta_{i=j}, \tag{2.4}$$

where δ is the Dirac delta. It is important to distinguish that $\sigma^2(\mathbf{x}_i)\delta_{i=j}$ is nonzero when the *indices* i and j are equal: it is possible to have two observations at the same location $\mathbf{x}$ but coming from different samples, thus not sharing noise.

2.1. GP regression

Given a dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, the GP assumption and observation likelihood (2.3) imply $[\mathbf{f}, \mathbf{y}]^\top \sim \mathcal{MVN}$, where $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$ and $\mathbf{y} = [y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n)]^\top$, so that the posterior $\mathbf{f}|\mathbf{y} \sim \mathcal{MVN}$ as well. More generally, for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, $[f(\mathbf{x}), f(\mathbf{x}'), \mathbf{y}]^\top \sim \mathcal{MVN}$, so that the posterior finite dimensional distribution is fully known:

$$[f_*(\mathbf{x}), f_*(\mathbf{x}')]^\top := \left([f(\mathbf{x}), f(\mathbf{x}')]^\top | \mathbf{y}\right) \sim \mathcal{MVN}\left([m_*(\mathbf{x}), m_*(\mathbf{x}')]^\top, \begin{bmatrix} k_*(\mathbf{x}, \mathbf{x}) & k_*(\mathbf{x}, \mathbf{x}') \\ k_*(\mathbf{x}', \mathbf{x}) & k_*(\mathbf{x}', \mathbf{x}') \end{bmatrix}\right), \tag{2.5}$$

where, for arbitrary $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, the posterior covariance kernel is defined as $k_*(\mathbf{x}, \mathbf{x}') := \operatorname{cov}(f(\mathbf{x}), f(\mathbf{x}')|\mathbf{y})$. The Kolmogorov extension theorem ensures that $\{f_*(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ defines a GP. Furthermore, (setting $m(\mathbf{x}) \equiv 0$ temporarily) the *posterior mean* and *variance* are explicitly given by

$$m_*(\mathbf{x}) = \mathbf{K}(\mathbf{x}, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \Delta(\mathbf{X}, \mathbf{X})\right]^{-1} \mathbf{y}, \tag{2.6}$$

$$\mathbf{K}_*(\mathbf{x}, \mathbf{x}') = \mathbf{K}\left([\mathbf{x}, \mathbf{x}']^\top, \mathbf{X}\right) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \Delta(\mathbf{X}, \mathbf{X})\right]^{-1} \mathbf{K}\left(\mathbf{X}, [\mathbf{x}, \mathbf{x}']^\top\right), \tag{2.7}$$

where $\mathbf{X}$ denotes the $n \times d$ matrix with rows $\mathbf{x}_i$, $i = 1, \ldots, n$, and for $\mathbf{U}$, $\mathbf{V}$ being $\ell \times d$ and $m \times d$, respectively, $\mathbf{K}(\mathbf{U}, \mathbf{V}) = \left[k(\mathbf{u}_i, \mathbf{v}_j)\right]_{1\le i\le \ell, 1\le j\le m}$ denotes the $\ell \times m$ matrix of pairwise covariances. $\Delta(\mathbf{U}, \mathbf{V})$ has entries $\sigma^2(\mathbf{u}_i)\delta_{i=j}\delta_{\mathbf{u}_i=\mathbf{v}_j}$; in our case $\Delta(\mathbf{X}, \mathbf{X})$ is a $n \times n$ diagonal matrix with entries $\sigma^2(\mathbf{x}_i)$. In the case of constant noise variance $\sigma^2 := \sigma^2(\mathbf{x}_i)$, $\Delta(\mathbf{X}, \mathbf{X}) = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix.

2.2. GP kernels

GP regression encodes the idea that similar inputs (according to the kernel) yield similar outputs. This can be seen through the posterior mean being a weighted average of observed data, since $m_*(\mathbf{x}) = \mathbf{w}^\top \mathbf{y}$ holds for $\mathbf{w}^\top = \mathbf{K}(\mathbf{x}, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \Delta(\mathbf{X}, \mathbf{X})\right]^{-1}$. Various types of kernels exist to encode similarity according to domain knowledge. Akin to covariance matrices, the only requirement for a function $k(\cdot, \cdot)$ to be a covariance kernel is that it is symmetric and *positive-definite*, that is for all $n = 1, 2, \ldots$, and $x_1, \ldots, x_n \in \mathbb{R}^d$, we must have the *Gram matrix* $\mathbf{K}(\mathbf{X}, \mathbf{X})$ be positive semi-definite.

A *stationary* kernel $k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ is one that can be written as a function of $\mathbf{x}_i - \mathbf{x}_j$, that is, $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = k_S(\mathbf{x}_i - \mathbf{x}_j)$ and is thus invariant to translations in the input space. A kernel is further called *isotropic* if it is only a function of $\mathbf{r} = \|\mathbf{x}_i - \mathbf{x}_j\|$, where $\|\cdot\|$ is the ℓ^2 Euclidean distance, so that we can write $k_I(\mathbf{r}) = k\left(\mathbf{x}_i, \mathbf{x}_j\right)$. Yaglom (1957) uses Bochner's theorem to derive a similar Fourier transform specific to isotropic kernels, providing a way to derive kernels from spectral densities. Stationary kernels are usually assumed to be normalized, since $k_S(\mathbf{x} - \mathbf{x}')/k_S(\mathbf{0}) = 1$ whenever $\mathbf{x} = \mathbf{x}'$; this allows for stationary kernels to be nicely interpreted as correlation functions. Let σ_f^2 denote the *process variance* $\operatorname{var}(f(\mathbf{x})) = \sigma_f^2$, we think of the GP f as $f(\mathbf{x}) = \sigma_f g(\mathbf{x})$ when g is a GP with normalized stationary kernel k.

Lastly, a *separable* kernel over $\mathbb{R}^d$ is one that can be written as a product: $k(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^d k_j\left(x^{(j)}, x'^{(j)}\right)$, where $x^{(j)}$ is the jth coordinate of $\mathbf{x}$. Thus, the global kernel is separated as a product over its dimensions, each having its own kernel.

The algebraic properties of positive-definite functions make it straightforward to compose new kernels from existing ones (Genton, 2001; Schölkopf *et al.*, 2002; Shawe-Taylor and Cristianini, 2004; Berlinet and Thomas-Agnan, 2011). The main tool is that kernels are preserved under addition and multiplication, that is, can be combined by sums and products. Hence, if k_1 and k_2 are two kernels and c_1, c_2 are two positive real numbers, then so is $k(\mathbf{x}, \mathbf{x}') = c_1 k_1(\mathbf{x}, \mathbf{x}') + c_2 k_2(\mathbf{x}, \mathbf{x}')$. This is consistent with the properties of GPs: if $f_1 \sim \mathcal{GP}(0, k_1)$ and $f_2 \sim \mathcal{GP}(0, k_2)$ are two independent GPs, then for $c_1, c_2 > 0$ we have $c_1 f_1 + c_2 f_2 \sim \mathcal{GP}(0, c_1 k_1 + c_2 k_2)$. This offers a connection to the framework of generalized additive models. Using this, a GP whose kernel function is of the form $\sigma_{f_1}^2 k_1 + \sigma_{f_2}^2 k_2$ where k_1 and k_2 are normalized stationary kernels can be interpreted as $f(\mathbf{x}) = \sigma_{f_1} f_1(\mathbf{x}) + \sigma_{f_2} f_2(\mathbf{x})$ where f_1 and f_2 are independent GPs with respective kernels k_1 and k_2. Consequently, $\operatorname{var}(f(\mathbf{x})) = \sigma_{f_1}^2 + \sigma_{f_2}^2$. These properties extend inductively to the case of finitely many terms.

Although there is no analogous result for a product of kernels (a product of GPs is no longer a GP, since the multivariate Gaussian distribution is not preserved), a common interpretation of multiplying kernels occurs when one kernel is stationary and monotonically decaying as $|\mathbf{x} - \mathbf{x}'| \to \infty$. Indeed, if k_1 is such a kernel, then the product $k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$ offers k_2's effect with a decay according to $k_1(\mathbf{x}, \mathbf{x}')$ as $|\mathbf{x} - \mathbf{x}'|$ increases. This synergizes with separability, for example, $k(\mathbf{x}, \mathbf{x}') = k_1\left(|x^{(1)} - x'^{(1)}|\right) k_2\left(x^{(2)}, x'^{(2)}\right)$ offers some similarity across the second coordinate that could decay as $|x^{(1)} - x'^{(1)}|$ increases. Additive

Table 1. *List of kernel families used in compositional search. C^p indicates that the GP sample paths $x \mapsto f(x)$ have p continuous derivatives; C^0 is continuous but not differentiable. Column $\mathcal{K}_r$ denotes whether the kernel family is in the restricted search set. The linear kernel is used for its year component only.*

Kernel name	Abbv.	Formula $k(x, x'; \theta)$	Properties	$\mathcal{K}_r$
Matérn-1/2	M12	$\exp\left(-\frac{\lvert x-x'\rvert}{\ell_{\text{len}}}\right), \quad \ell_{\text{len}} > 0$	C^0	✓
Matérn-3/2	M32	$\left(1+\frac{\sqrt{3}}{\ell_{\text{len}}}\lvert x-x'\rvert\right)\exp\left(-\frac{\sqrt{3}}{\ell_{\text{len}}}\lvert x-x'\rvert\right), \quad \ell_{\text{len}} > 0$	C^1	
Matérn-5/2	M52	$\left(1+\frac{\sqrt{5}}{\ell_{\text{len}}}\lvert x-x'\rvert+\frac{5}{3\ell_{\text{len}}^2}\lvert x-x'\rvert^2\right)\exp\left(-\frac{\sqrt{5}}{\ell_{\text{len}}}\lvert x-x'\rvert\right)$	C^2	✓
Cauchy	Chy	$\frac{1}{1+\lvert x-x'\rvert^2/\ell_{\text{len}}^2}, \quad \ell_{\text{len}} > 0$	C^∞	
Radial Basis	RBF	$\exp\left(-\frac{(x-x')^2}{2\ell_{\text{len}}^2}\right), \quad \ell_{\text{len}} > 0$	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	$\sigma_0^2 + x \cdot x', \quad \sigma_0 > 0$	Nonstationary	*
Minimum	Min	$t_0^2 + x \wedge x', \quad t_0 > 0$	Nonstat, C^0	✓
Mehler	Meh	$\exp\left(-\frac{\rho^2(x^2+x'^2)-2\rho xx'}{2(1-\rho^2)}\right), \quad -1 \le \rho \le 1$	Nonstationary	

and multiplicative properties are often used in conjunction with the constant kernel $k(\mathbf{x}, \mathbf{x}') = c$, $c > 0$, resulting in a scaling effect (multiplication) or dampening effect (addition).

Remark 1. Several other kernel design strategies exist, for example, if $g : \mathbb{R}^d \to \mathbb{R}$, then $k(\mathbf{x}, \mathbf{x}') = g(\mathbf{x})g(\mathbf{x}')$ defines a kernel; this property along with multiplication is one way to account for heteroskedastic noise, since $k(\mathbf{x}_i, \mathbf{x}'_j) = \delta_{i=j}$ defines a kernel. See, for example, Genton (2001) or Noack and Sethian (2021) for additional properties and explanation.

2.3. Kernel families

Table 1 lists the kernel families we consider. For simplicity, we assume one dimensional base kernels with $x \in \mathbb{R}$ where the full structure for $\mathbf{x} \in \mathbb{R}^d$ is expressed as a separable kernel. Among the nine kernel families, we have kernels that give smooth (C^2 and higher) fits, kernels with rough (non-differentiable) sample paths, and several nonstationary kernels. Here, smoothness refers to the property of the sample paths $\mathbf{x} \mapsto f(\mathbf{x})$ being, say, k-times differentiable, that is, $f(\cdot) \in C^k$. The posterior mean $\mathbf{x} \mapsto m_*(\mathbf{x})$ inherits its similar (but generally less strict) differentiability properties, see Kanagawa *et al.* (2018) for details. The $\mathcal{K}_r$ column indicates whether a kernel is included in the *restricted set of kernels* (used in later sections), in contrast to the *full set of kernels* $\mathcal{K}_f$. This subset $\mathcal{K}_r$ comprises a more compact collection of the most commonly used kernels in the literature. One reason for $\mathcal{K}_r$ is to minimize overlap in terms of kernel properties; as we report below, some kernel families in $\mathcal{K}_f$ apparently yield very similar fits and act as "substitutes" for each other.

All of the kernels listed have hyperparameters, which help to understand their relationship with the data. The quantity ℓ_{len} appearing in many stationary kernels is referred to as the *characteristic*

lengthscale, which is a distance-scaling factor. With the radial basis function (RBF) kernel for example, this loosely describes how far x' needs to move from x in the input space for the function values to become uncorrelated (Williams and Rasmussen, 2006). Thus, the lengthscale of a calibrated GP can be interpreted as the strength of the correlation decay in the training dataset. Out of the stationary kernels, a popular class is the Matérn class. In continuous input space, the value ν in the Matérn-ν corresponds to smoothness: a GP with a Matérn-ν kernel is $\lceil \nu \rceil - 1$ times differentiable in the mean-square sense (Williams and Rasmussen, 2006). The RBF kernel is the limiting case as $\nu \to \infty$, resulting in an infinitely differentiable process. The $\nu = 1/2$ case recovers the well-known Ornstein–Uhlenbeck process, which is mean reverting and non-differentiable. Also non-differentiable but on the nonstationary side, the minimum kernel corresponds to a Brownian motion process when $x \in \mathbb{R}_+$, where $t_0^2 = \text{var}(f(0))$ is the initial variance. For discrete x, Min kernel yields the random walk process, and M12 yields the AR(1) process.

Less commonly studied are the (continuous-time) AR2 (see Parzen, 1961), Cauchy, and Mehler kernels. The continuous-time AR2 kernel acts identically to a discrete-time autoregressive (AR) process of order 2 with complex characteristic polynomial roots when x is restricted to an integer. The heavy-tailed Cauchy probability density function motivates the Cauchy kernel, with the goal of modeling long-range dependence and is a special case of the rational quadratic kernel (see Appendix A). Lastly, the Mehler kernel has a form similar to that of a joint-normal density and acts as a RBF kernel with a nonstationary modification (this can be seen from a "complete the squares" argument). Although Mehler is nonstationary, it remarkably yields a stationary *correlation function* $\text{corr}(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')/\sqrt{k(\mathbf{x}, \mathbf{x})k(\mathbf{x}', \mathbf{x}')}$. See Appendix A for a more thorough discussion of the aforementioned kernels and their properties.

2.4. Varying prior means

For brevity of exposition, the preceding analysis considered zero prior mean $m(\mathbf{x}) = 0$. When the prior mean function $m(\mathbf{x})$ is known and deterministic, the posterior covariance function $k_*(\mathbf{x}, \mathbf{x}')$ remains unchanged compared to the zero-mean case, and the posterior mean $m_*(\mathbf{x})$ is adjusted to:

$$m_*(\mathbf{x}) = m(\mathbf{x}) + \mathbf{K}(\mathbf{x}, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \Delta(\mathbf{X}, \mathbf{X})\right]^{-1} (\mathbf{y} - m(\mathbf{X})), \tag{2.8}$$

where $m(\mathbf{X}) = [m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n)]^\top$. More common is the case of a *parametric mean function* $m(\mathbf{x}) = \sum_{j=1}^d \beta_j h_j(\mathbf{x})$, where the basis functions $h_j(\mathbf{x})$ are fixed and known (e.g., $h_j(\mathbf{x})$ is a jth degree polynomial), and the coefficients $\boldsymbol{\beta}$ are estimated simultaneously with the covariance hyperparameters through maximizing the marginal likelihood. The generalized least squares estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^\top \mathbf{K}_\mathbf{y}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^\top \mathbf{K}_\mathbf{y}^{-1} \mathbf{y}, \qquad \text{with} \quad \mathbf{K}_\mathbf{y} = \mathbf{K}(\mathbf{X}, \mathbf{X}) + \Delta(\mathbf{X}, \mathbf{X}), \tag{2.9}$$

where $\mathbf{H}$ is the $n \times d$ matrix with i, j entry $h_j(\mathbf{x}_i)$. The resulting posterior mean $m_*(\mathbf{x})$ is based on the plug-in estimated trend $\mathbf{H}\hat{\boldsymbol{\beta}}$ and the posterior variance has an additional term to account for the parameter uncertainty in $\boldsymbol{\beta}$:

$$\begin{aligned} m_*(\mathbf{x}; \hat{\beta}) &= \mathbf{h}(\mathbf{x})^\top \hat{\boldsymbol{\beta}} + \mathbf{K}(\mathbf{x}, \mathbf{X}) \mathbf{K}_\mathbf{y}^{-1} \left(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}\right), \qquad \mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_d(\mathbf{x})]^\top; \\ k_*(\mathbf{x}, \mathbf{x}; \hat{\beta}) &= k_*(\mathbf{x}, \mathbf{x}) + \left(\mathbf{h}(\mathbf{x}) - \mathbf{K}(\mathbf{x}, \mathbf{X})^\top \mathbf{K}_\mathbf{y}^{-1} \mathbf{H}\right)^\top \left(\mathbf{H}^\top \mathbf{K}_\mathbf{y} \mathbf{H}\right)^{-1} \left(\mathbf{h}(\mathbf{x}) - \mathbf{K}(\mathbf{x}, \mathbf{X})^\top \mathbf{K}_\mathbf{y}^{-1} \mathbf{H}\right). \end{aligned}$$

For more details, see Roustant *et al.* (2012). It is important to recall that $\mathbf{x}$ is an input to the function, while $\mathbf{X}$ is fixed (from the observed data). Intuitively, this reflects a transition toward prior reliance: in the case of a decaying kernel (e.g., Matérn family), when the predictive location $\mathbf{x}$ distances itself from the rows of $\mathbf{X}$, that is, $\|\mathbf{x} - \mathbf{x}_i\| \to 0$ for all training locations $\mathbf{x}_i$, $i = 1, \ldots, n$, the vector $\mathbf{K}(\mathbf{x}, \mathbf{X}) = [k(\mathbf{x}, \mathbf{x}_i)]_{i=1}^n$ goes to $\mathbf{0}$, causing a reversion to the prior mean and covariance.

2.5. Connections to mortality modeling

For mortality modeling, our core input space is composed of Age and Year coordinates: $x_{ag}, x_{yr} \in \mathbb{R}^2_+$. As the GP can model non-linear relationships, we include Birth Cohort $x_c := x_{ag} - x_{yr}$ as a third coordinate of $\mathbf{x}$, so that $\mathbf{x} = (x_{ag}, x_{yr}, x_c)$. For a given $\mathbf{x}$, denote $D_\mathbf{x}$ and $E_\mathbf{x}$ as the respective observed deaths and exposures (i.e., individuals alive at the beginning of the period) over the corresponding (x_{ag}, x_{yr}) pair. Denote $y(\mathbf{x}) = \log(D_\mathbf{x}/E_\mathbf{x})$ as the *log mortality rate*. The full dataset is denoted $\mathcal{D} = \{\mathbf{x}_i, y_i, D_i, E_i\}_{i=1}^n$. Each mortality observation $y_i := y_i(\mathbf{x}_i)$ is the regression quantity modeled in Equation (2.3), so that $\mathbb{E}[y(\mathbf{x})|f(\mathbf{x})] = f(\mathbf{x})$. The interpretation is that the true mortality rate $f(\mathbf{x})$ is observed in accordance with mean-zero (i.e., unbiased) uncorrelated noise yielding the measured mortality experience y.

Relating to the log-normal distribution, algebra shows that

$$\mathbb{E}[D_\mathbf{x}|f(\mathbf{x})] = E_\mathbf{x} \exp\left(f(\mathbf{x}) + \frac{\sigma^2(\mathbf{x})}{2}\right), \text{ and } \quad \mathbb{E}[D_\mathbf{x}|f(\mathbf{x}), \epsilon(\mathbf{x})] = E_\mathbf{x} \exp(f(\mathbf{x}) + \epsilon(\mathbf{x})),$$

akin to an overdispersed Poisson model in existing mortality modeling literature (see, e.g., Azman and Pathmanathan, 2022). Note that the seminal work of Brouhns *et al.* (2002) for modeling log-mortality rates suggests that homoskedastic noise is unrealistic, since the absolute number of deaths at older ages is much smaller compared to younger ages. As a result, we work with heteroskedastic noise

$$\text{var}(\epsilon|D_\mathbf{x}) = \sigma^2(\mathbf{x}) := \frac{\sigma^2}{D_\mathbf{x}}, \quad \text{where } \sigma^2 \in \mathbb{R}^+. \tag{2.10}$$

Thus, we make observation variance inversely proportional to observed death counts, with the constant σ^2 to be learned as part of the fitting procedure. From a modeling perspective, this works since $D_\mathbf{x}$ is known whenever $E_\mathbf{x}$ and $y(\mathbf{x})$ are. Indeed, Equation (2.5) shows no requirement to know $D_\mathbf{x}$ for out-of-sample forecasting. In the case where full distributional forecasts are desired, one could instead model a noise surface $\sigma^2(\cdot)$ simultaneously with $f(\cdot)$, see, for example, Cole *et al.* (2022).

Remark 2. Our framework makes two different Gaussian assumptions. First, we capture the latent mortality surface $f(\cdot)$ as a GP. This assumption is generally very mild and can be understood as applying standard kernel ridge regression to obtain $m_*(\mathbf{x})$ (with a highly customized and adaptively fitted kernel) within a probabilistic framework (see Kanagawa et al., 2018); it does not say anything specific about mortality itself. The second Gaussianity assumption is about the observation noise $\epsilon(\cdot)$. This assumption is stronger and takes a specific stance on mortality rate observations; its purpose is to maintain Gaussian conjugacy which simplifies the likelihood and facilitates the MLE process. Given the complex observation structure of mortality, we argue that what really matters is not the distribution of ϵ's, but their heteroskedasticity, that is, removing the "identically distributed" assumption imposed by simpler models. This is exactly what we do in (2.10). It is possible to go further and adjust the Gaussian noise likelihood to be heavy tailed (as in Wang et al., 2011; Ahmadi and Gaillardetz, 2014) or to hierarchically model, for example, $D_\mathbf{x} \sim \text{Pois}(E_\mathbf{x} \exp(f(\mathbf{x})))$ where f is a GP. We do not pursue this due to added modeling and computational burdens.

A discussion of GP covariances connects naturally to APC models. A stationary covariance means that the dependence between different age groups or calendar years is only a function of the respective ages/year distances and is not subject to additional structural shifts. Additive and separable covariance structures play an important role in the existing mortality modeling literature, specifically in the APC framework. For example, Lee–Carter and CBD began with Age–Period modeling, which subsequently evolved to add cohort or additional Period effects. Rather than postulating the precise APC terms, the latest (Dowd *et al.*, 2020) CBDX framework adds up to 3 period effects as needed. Similarly, Hunt and Blake (2014) develop a general recipe for constructing mortality models, where core demographic features are represented with a particular parametric form, and combined into a global structure. This can be reproduced with GPs where kernels encode expert judgment to such demographic features. This type of encoding into covariances is already being done in the literature, possibly unknowingly. Take, for example, the basic CBD model whose stochastic part has the form $f(\mathbf{x}) = \kappa^{(1)}_{x_{yr}} + \left(x_{ag} - \overline{\mathbf{x}_{ag}}\right) \kappa^{(2)}_{x_{yr}}$, where

$\overline{\mathbf{x}_{ag}}$ is the average age in $\mathcal{D}$. Under the common assumption that $\left(\kappa^{(1)}_{x_{yr}}, \kappa^{(2)}_{x_{yr}}\right)$ is a multivariate random walk with drift, a routine calculation shows that

$$\mathbb{E}[f(\mathbf{x})] = \kappa_0^{(1)} + \mu^{(1)} x_{yr} + \left(x_{ag} - \overline{\mathbf{x}_{ag}}\right)\left(\kappa_0^{(2)} + \mu^{(2)} x_{yr}\right)$$

and $k_{CBD}(\mathbf{x}, \mathbf{x}') = \text{cov}(f(\mathbf{x}), f(\mathbf{x}')) =$

$$= \left[\sigma_1^2 + \rho\sigma_1\sigma_2\left(x_{ag} + x'_{ag} - 2\overline{\mathbf{x}_{ag}}\right) + \left(x_{ag} - \overline{\mathbf{x}_{ag}}\right)\left(x'_{ag} - \overline{\mathbf{x}_{ag}}\right)\sigma_2^2\right]\left(x_{yr} \wedge x'_{yr}\right). \tag{2.11}$$

We can re-interpret the above as the kernel decomposition $k_{CBD}(x, x') = k_1\left(x_{ag}, x'_{ag}\right) k_2\left(x_{yr}, x'_{yr}\right)$, where $k_2\left(x_{ag}, x'_{ag}\right)$ is a kernel depending only on Age (that could be decomposed into additive components), and k_2 is the minimum kernel depending only on Year. Furthermore, if the multivariate random walk is assumed to be Gaussian, then $f(\mathbf{x})$ actually forms a (discrete-time) GP, and hence the methods detailed in Section 2.1 apply verbatim.

Period and Cohort effects are commonly modeled using time series models (Villegas *et al*., 2015). In particular, Gaussian ARIMA models are popular in existing APC literature, see, for instance, Cairns *et al.* (2011) who single out the usefulness of AR(1), ARIMA(1,1,0), and ARIMA(0,2,1) for Cohort effect, and ARIMA(1,1,0) or ARIMA(2,1,0) for Period effect. This provides another link to (discrete) GP covariance analogues: a Matérn-1/2 covariance corresponds to an AR(1) process, a (continuous-time) AR2 covariance to an AR(2) process with complex unit roots, and a Minimum covariance to a discrete-time random walk.

3. Genetic programming for GPs

Starting with the building blocks of the kernels in Table 1, infinitely many compositional kernels can be constructed through addition and multiplication. The idea of a GA is to adaptively explore the space of kernels via an evolutionary procedure. At each step of the GA, kernels that have a higher "fitness score" are more likely to evolve and be propagated, while lower-fitness kernels get discarded. The evolution is achieved through several potential operations that are selected randomly in each instance. In the first sub-step of the GA, *ancestors* of next-generation kernels are identified. This is done via "tournaments" that aim to randomly pick generation-g kernels, while preferring those with higher fitness. A given kernel can be selected in multiple tournaments, that is, generate more than one child. In the second sub-step, each ancestor undergoes crossover (mixing kernel components with another ancestor) or mutation (modifying a component of the sole ancestor) to generate a generation-$(g+1)$ kernel.

3.1. BIC and Bayes factors for GPs

To evaluate the appropriateness of a kernel within a given set $k \in \mathcal{K}$, an attractive criterion is the posterior likelihood of the kernel given the data $p(k|\mathbf{y}) = p(\mathbf{y}|k)p(\mathbf{y})/p(k)$, where, under a uniform prior assumption $p(k) = 1/|\mathcal{K}|$, we see that $p(k|\mathbf{y}) \propto p(\mathbf{y}|k)$. However, the integral over hyperparameters $p(\mathbf{y}|k) = \int_\theta p(\mathbf{y}, \theta|k)d\theta$ is generally intractable, so we use the BIC as an approximation, where $\text{BIC}(k) \approx \log p(\mathbf{y}|k)$ is defined as:

$$\text{BIC}(k) = -l_k\left(\hat{\theta}; \mathbf{y}\right) + \frac{|\hat{\theta}|\log(n)}{2}, \tag{3.1}$$

where $l_k(\theta|\mathbf{y}) = \log p(\mathbf{y}|k, \theta)$ is the *log marginal likelihood of y* evaluated at θ under a given kernel k, $\hat{\theta}$ is the maximizer (maximum marginal likelihood estimate) of $l_k(\theta|\mathbf{y})$, and $|\hat{\theta}|$ is the total number of estimated hyperparameters in $\hat{\theta} = \left[\hat{\beta}_0, \hat{\beta}_{ag}, \hat{\theta}_k, \hat{\sigma}^2\right]^\top$, where θ_k is a vector of all kernel specific hyperparameters. Note that $p(\mathbf{y}|k, \theta)$ is a multivariate density, with mean and covariances governed by Equation (2.4). The BIC metric has seen use in similar applications of GP compositional kernel search, see, for example, Duvenaud *et al*. (2013), Duvenaud (2014), and is commonly used in mortality modeling (Cairns *et al*., 2009). We employ BIC (lower BIC being better) for our GA fitness metric below.

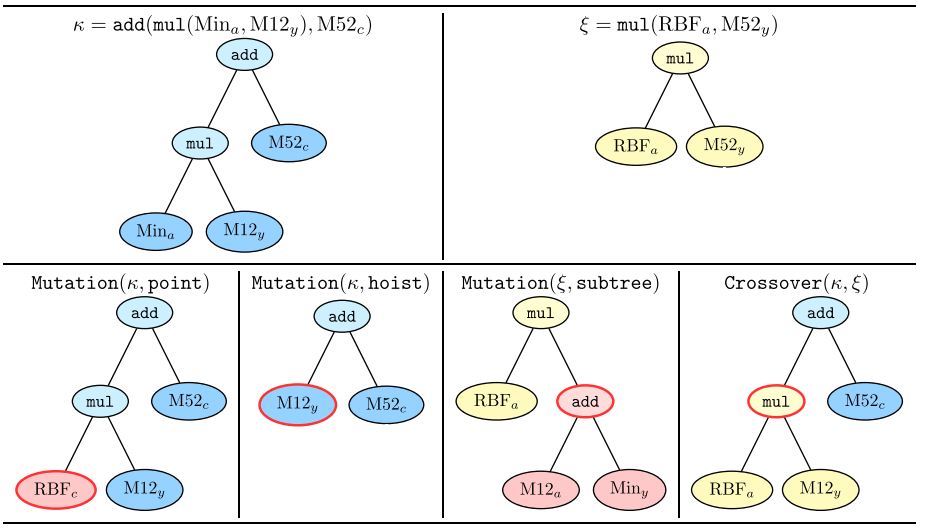

Figure 1. *Representative compositional kernels and GA operations. Bolded red ellipses indicate the node of κ (or ξ) that was chosen for mutation or crossover.*

One can further assess the relative likelihood of $k_1, k_2 \in \mathcal{K}$ by again assuming a uniform prior over $\mathcal{K}$, and computing the *Bayes factor* (BF)

$$\text{BF}(k_1, k_2) = \frac{p(k_1|\mathbf{y})}{p(k_2|\mathbf{y})} \approx \exp\Big(\text{BIC}(k_2) - \text{BIC}(k_1)\Big). \tag{3.2}$$

Obtaining an approximation for $p(k|\mathbf{y})$ further allows the possibility of *Bayesian model averaging* by conditioning over $k \in \mathcal{K}$:

$$p(\mathbf{f}|\mathbf{y}) = \sum_{k \in \mathcal{K}} p(\mathbf{f}|\mathbf{y}, k) p(k|\mathbf{y}). \tag{3.3}$$

Since $p(\mathbf{f}|\mathbf{y}, k)$ is available in the closed form, this provides the full distribution of future forecasts, if desired.

Gelman *et al.* (1995) states that BFs work well in the case of a discrete model selection. The seminal work of Jeffreys (1961) gives a table of *evidence categories* to determine a conclusion from BFs, see Table 15 in Appendix E, which is still frequently used today (Lee and Wagenmakers, 2014; Dittrich *et al.*, 2019).

Note that in accordance with the penalty term $|\hat{\theta}| \log(n)/2$ in Equation (3.1), the difference in penalties in $\text{BIC}(k_2) - \text{BIC}(k_1)$ is simply the number of additional kernel hyperparameters, scaled by $\log(n)/2$, since $\beta_0, \beta_{ag}, \sigma^2$ are always estimated regardless of kernel choice. Thus in the application of GP kernel selection, the BF properly penalizes kernel complexity. In Table 1, most kernels have one θ_k (the lengthscale), but some, like AR2, have two hyperparameters and so incur a larger penalty.

3.2. GA kernel representation

In order to operate in the space of kernels, we shall represent kernels via a tree-like structure, cf. Figure 1. Internal nodes correspond to *operators* (`add` or `mul`) that combine two different kernels together, while leafs are the univariate kernels used as building blocks. We indicate the coordinate operated on by the kernel through the respective subscripts a, y, c, such as M52_a. In turn, such trees are transcribed into bracketed expressions, such as $\kappa = \texttt{add(Exp\_c, mul(RBF\_a, add(Mat\_y, RBF\_c)))}$ corresponding to the Age–Period–Cohort kernel $(k_{M52}(x_{yr}) + k_{RBF}(x_c)) \cdot k_{RBF}(x_{ag}) + k_{Exp}(x_c)$. The *length* of κ, denoted

Table 2. *High level Genetic algorithm parameters and description. Note that* $G \cdot n_g = 4000$ *is the total number of trained GP models in a single run of the GA.*

Parameter	Value	Notes
Population size	$n_g = 200$	Number of individuals per generation
Generations	$G = 20$	Number of generations
Tournament size	$T = 7$	Run a double tournament and select smaller winner with probability p_{DT}
	$p_{DT} = 0.6$	Smaller is selected with probability 0.60

$|\kappa|$, is its number of nodes. The above kernel tree has length $|\kappa| = 7$, namely four base kernels combined with three operators. Observe that the tree structure is not unique, that is, some complex kernels can be permuted and expressed through different trees. In what follows, we will ignore this non-uniqueness.

A certain expertise is needed to convert from an above representation to the dependence structure it implies. One way to visualize is to take advantage of the stationarity and plot the heatmap of the matrix $k_S(\mathbf{x} - \mathbf{x}')$ as a function of its Age, Year coordinates.

3.3. GA operations and algorithm parameters

The overall GA is summarized by the set of possible mutations and a collection of algorithmic parameters. These are important for many reasons: (i) sufficient exploration so that the algorithm does not get trapped in a particular kernel configuration; (ii) efficiency in terms of number of generation and generation size needed to find the best kernels; and (iii) bloat control, that is, ensuring that returned kernels are not overly complex and retain interpretability. Interpretable kernels would tend to have low length (below 10) and avoid repetitive patterns. Bloat control, that is, avoiding the appearance of overly complex/long kernels, is a concern with GAs. Luke and Panait (2006) and Poli *et al.* (2008) suggest several ways to combat it.

Our specific high-level GA parameters are listed in Table 2. We largely follow guidelines from Sipper *et al.* (2018), which offers a thorough investigation of the parameter space of GA algorithms. Ancestors are chosen via a tournament setup, where $T = 7$ individuals are independently and uniformly sampled from the previous generation and a single tournament winner is the fittest (lowest BIC) individual. A smaller T provides diversity in future generations, whereas a larger T reduces chances of leaving behind fit individuals. To combat bloat, we follow the *double tournament* procedure described in Luke and Panait (2006): all instances of a single tournament are replaced by two tournaments run one after the other, with the lower length ancestor chosen with probability $p_{DT} \in [0.5, 1]$; larger values prefer parsimony over fitness. We use the less restrictive $p_{DT} = 0.6$ instead of the suggested $p_{DT} = 0.7$, partially since BIC has a built-in penalization for unwieldy models thereby mitigating bloat. As a further proponent of parsimony, we use hoist mutation as suggested in Poli *et al.* (2008) with $p_h = 0.1$. Using the above parameterization, we rarely observe kernels of length over 15 in our experiments.

Table 3 fully details the crossover and mutation operations and associated algorithmic parameters, with Figure 1 providing a visual illustration. Most of these are standard in the literature. Our domain knowledge suggests an additional point mutation operator which we call *respectful point mutation*. This operator maintains the coordinate of the kernel being mutated, so that a kernel operating on age is replaced by one in age, and so forth. This respects a discovered APC structure and fine tunes the chosen coordinate.

The *arity* (number of arguments) of a function needs to be preserved in crossover and mutation operations. In our setup, the only non-trivial arity functions are `add` and `mul` (both with arity 2), so if these nodes are chosen for a mutation, the point mutation (and respectful point mutation) automatically replaces them with the other operation: `add` with `mul` and vice-versa.

In some GA applications, authors propose to have several hundred (or even thousand) generations. Because times to fit a GP model is non-trivial, running so many generations is computationally

Table 3. *Operator specific GA parameters with description. Note that* $p_c + p_s + p_h + p_p + p_r + p_0 = 1$.

Probability	GA operation	Notes
$p_c = 0.45$	Crossover	
$p_s = 0.2$	Subtree mutation	
$p_h = 0.1$	Hoist mutation	
$p_p = 0.05$	Point mutation	Each node is mutated with another node of same arity with prob. q_p
$p_r = 0.15$	Respectful point mutation	Each node is mutated with another node of same arity and same (age, year, cohort) with prob. q_r
$p_0 = 0.05$	Copy	
$q_p = 0.25$	Point replace	
$q_r = 0.35$	Respectful replace	
$q_a = 0.5$		Probability that `add`/`mul` is included when initializing trees

Algorithm 1: Genetic Algorithm for compositional kernel selection

input: Kernel search set $\mathcal{K}$; genetic algorithm parameters $\mathcal{G}_1 = \{n_g, G, T\}$ and $\mathcal{G}_2$ in Tables 2 and 3 respectively; Dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

output: Kernel expressions and BIC scores $\{\{\kappa_i^g, b_i^g\}_{i=1}^{n_g}\}_{g=0}^{G-1}$

for $i = 1$ **to** n_g **do**
  $\kappa_i^0 \leftarrow$ `InitializeKernel()`;
  $b_i^0 \leftarrow$ `BIC(`κ_i^0`)`
for $g = 1$ to $G - 1$ do
  for $i = 1$ to n_g do
    DetermineAncestor
      $\{\kappa_{(1)}^{g-1}, \ldots, \kappa_{(T)}^{g-1}\} \leftarrow$ `SampleUniform(`$\{\kappa_1^{g-1}, \ldots, \kappa_{n_{g-1}}^{g-1}\}, T$`)`;
      $j_i \leftarrow$ `argmin`$\{b_{(1)}^{g-1}, \ldots, b_{(T)}^{g-1}\}$ // `b's are the respective BIC's of these individuals;`
      $A_i^{g-1} \leftarrow \kappa_{j_i}^{g-1}$;
    `Operation` $\leftarrow$ `Sample(Crossover, (Mutation, Type)) // according to` $\mathcal{G}_2$;
    if `Operation == Crossover` **then**
      $A_i'^{g-1} \leftarrow \kappa_{j_i'}^{g-1}$ // 2nd ancestor (according to DetermineAncestor);
      $\kappa_i^g \leftarrow$ `Crossover(`$A_i^{g-1}, A_i'^{g-1}$`)`;
    else
      $\kappa_i^g \leftarrow$ `Mutation(`A_i^{g-1}`; Type)`;
    $b_i^g \leftarrow$ `BIC(`κ_i^g`)`

prohibitive. Below we use a fixed number of $G = 20$ generations. As shown in Figure 5, kernel exploration seems to stabilize after a dozen or so generations, so there appears to be limited gain in increasing G. In contrast with a typical GA application, we are more interested in interpreting all prototypical well-performing kernels, rather than in optimizing an objective function to an absolute minimum; that is to say, our analysis is sufficient as long as a representative ballpark of optimal models has been discovered.

Denote by κ_i^g the $i = 1, \ldots, n_g$th *individual* in generation $g = 1, \ldots, G-1$, with a corresponding BIC of b_i^g. Algorithm 1 outlines the full GA. `BIC` computes the BIC as in Equation (3.1), which implicitly fits

a GP and optimizes hyperparameters. `SampleUniform` determines ancestors according to a tournament of size T. A crossover or mutation is determined according to the probabilities provided in $\mathcal{G}_2$ (according to Table 3), where `Type` denotes the type of mutation chosen. In the event of a crossover, another ancestor is determined through the same process as the first. The entire algorithm is ran for G generations, where $g = 0$ initializes, and the remaining $G - 1$ steps involve choosing ancestors and offspring. For $g = 0$, `InitializeKernel()` constructs a randomly initialized kernel, where the length is uniformly chosen from $\{3, 5, 7, 9\}$ (respectively 2, 3, 4, 5 base kernels), where the base kernels are uniformly sampled from $\mathcal{K}$, and add/multiply operations are equally likely. Experiments where we initialized with a more diverse set (kernel length from 1 to 15) slowed GA convergence with a negligible effect on end results. Our initialization was chosen with the goal of providing as unbiased of a sampling procedure as possible. Alternative initializations could be used, for example, imposing a diversification constraint on the APC terms in a given kernel, or infusing the initial generation with known-to-be-adequate kernels. It is also worth mentioning that although we fixed $n_g = 200$ for all $g = 0, \ldots, G-1$, this value could vary over g so that the total number of results is $\sum_{g=0}^{G-1} n_g$.

Remark 3. As mentioned, we utilized a double tournament method to combat bloat. Specifically in Algorithm 1, whenever `DetermineAncestor` is run, it is in fact run twice to obtain $A_{i,(1)}^{g-1}$ and $A_{i,(2)}^{g-1}$, and the smaller length winner $A_i^{g-1} = A_{i,(j_0)}^{g-1}$ is chosen with probability $p_{DT} = 0.6$, where $j_0 = \arg\min_{j=1,2} |A_{i,(j)}^{g-1}|$. See Luke and Panait (2006) for details. This is left out of Algorithm 1 for brevity.

Since top-performing kernels are preferred as tournament winners, they become ancestors for future generators and are likely to re-appear as duplicates (either due to a `Copy` operation or a couple of mutations canceling each other). As a result, we observe many duplicates when aggregating all κ_i^g's across generations.

Remark 4. Some kernel compositions result in over-parameterization, for example, `mul(RBF_a, RBF_a)` which is statistically identical to RBF_a. This is handled through the BIC penalty, which prefers the reduced one-term version.

3.4. GP hyperparameter optimization

As mentioned, the hyperparameters of a given kernel $k(\cdot, \cdot; \theta)$ are estimated by maximizing $l_k(\theta|\mathbf{y})$, the marginal log likelihood of the observed data. The optimization landscape of kernel hyperparameters in a GP is non-convex with many local minima, so we take care in this optimization. Since $\mathbf{y}$ is on the log scale, we leave these values untransformed. For given $\mathbf{x} = [x_{ag}, x_{yr}, x_c]^\top \in \mathbb{R}^3$, we perform dimension-wise scaling to the unit interval, for example, if $x_{ag} = [x_{1,ag}, \ldots, x_{n,ag}]^\top$ then $x_{i,ag} \mapsto \frac{x_{i,ag} - \min(x_{ag})}{\max(x_{ag}) - \min(x_{ag})}$ and similarly for x_{yr} and x_c. Lengthscales for stationary kernels can be interpreted on the original scale through the inverse transformation $\ell_{\text{len}} = (\max(x) - \min(x))\, \tilde{\ell}_{\text{len}}$, where $\tilde{\ell}_{\text{len}}$ is on the transformed scale, with similar transformations for the mean function parameters. For interpretability purposes, we utilize this to report values on the original scale in Sections 4 and 5 whenever possible. Nonstationary kernel (i.e., Min, Meh, and Lin) hyperparameters are left transformed, as they generally cannot be interpreted on the original scale.

We use Python with the GPyTorch library (Gardner *et al.*, 2018) to efficiently handle data and matrix operations. Since our results rely heavily on accurate likelihood values (through BIC), we turn off GPyTorch's default matrix approximation methods. Maximizing $l_k(\theta|\mathbf{y})$ is done using Adam (Kingma and Ba, 2014). This is an expensive procedure, as a naive evaluation of $l_k(\theta|\mathbf{y})$ is $\mathcal{O}(n^3)$ from inverting $\mathbf{K}(\mathbf{X}, \mathbf{X})$. Our GA runs use a convergence tolerance of $\varepsilon = 10^{-4}$ and maximum iterations of $\eta_{\max} = 300$. We found most simple kernels to converge quickly ($\eta \leq 100$) even up to $\varepsilon = 10^{-6}$, see Table B.1 in the Appendix. We keep $\eta_{\max}$ relatively small since we need to fit $n_g \cdot G \gg 10^3$ models. Upon completion, the top few dozen kernels are refit with $\varepsilon = 10^{-6}$ and $\eta_{\max} = 1000$. Note that since Adam is a stochastic algorithm, the fitted kernel hyperparameters vary (slightly in our empirical work) during this refitting.

Table 4. *Description of synthetic datasets. Data are generated as multivariate normal realizations according to the Equations in (2.4), with parametric mean function $m(\boldsymbol{x}) = \beta_0 + \beta_{ag} x_{ag}$. SYA and SYB are homoskedastic. In generating SYC's heteroskedastic noise, D_x comes from the JPN Female data, see Section 5 for details regarding this dataset.*

Exprmnt	Ground truth kernel	$\sigma^2(\mathbf{x})$	β_0	β_{ag}
SYA	$0.04 \cdot \text{RBF}_a(13.6) \cdot \text{RBF}_y(8.7)$	0.001	–10.0	0.1
SYB	$0.08 \cdot \text{RBF}_a(19.92) \cdot M12_y(386.6) + 0.02 \cdot M52_c(4.98)$	0.0004	–9.94	0.087
SYC	$0.4638 \cdot \text{M52}_a(37.7) \cdot \text{Chy}_y(56.6) \cdot \text{M12}_y(1810) \cdot \cdot \text{M12}_c(7378)$	$1.0781/D_{\mathbf{x}}$	–11.8835	0.1134

This reflects the idea that the GA is a preliminary "bird's-eye search" to find plausible mortality structures, thereafter refining hyperparameter estimates. Except in a few cases, changes in final BIC values are minimal, though the ranking of top kernels can occasionally get adjusted.

4. Synthetic mortality kernel recovery

The premise of the GA is to carry out an extensive search that can correctly identify appropriate APC covariance structure(s) for a given dataset $\mathcal{D}$. The GA can be thought of as an initial search to yield a few thousand $(n_g \cdot G)$ candidates, after which one can identify the top few best-fitting kernels as the ones to best represent the covariance structure of $\mathcal{D}$. To assess the quality of this kernel discovery process, we first try three synthetic datasets where the true data generating process, that is, the APC covariance structure, is given, and therefore we can directly compare the outputs of the GA to a known truth. This checks whether the GA can recover the true covariance, and by using BFs, analyzes which kernels express similar information by substituting others. Furthermore, it allows to assess the effect of noise on kernel recovery. These experiments also provide a calibration to understand evidence categories for BFs (see Table 15) in the context of GP kernel comparison. Simpler experiments focus on the restricted set of kernels $\mathcal{K}_r$ as described in Section 2.3, noting that Lin_y is the only appearance of linear (to model linear mortality improvement/decline in calendar year). Additional experiments utilize the *full set* $\mathcal{K}_f$ of kernels, which includes all kernels in Table 1 over all coordinate dimensions.

The full experimental setup is as follows. First, we fix a GP kernel in the APC space and generate a respective log-mortality surface by sampling exactly from that prior. This creates a synthetic mortality surface. Below our surface spans Ages 50–84 and Years 1990–2019 and includes a linear parametric trend $m(\mathbf{x}) = \beta_0 + \beta_{ag} x_{ag}$ in Age. Thus, our training sets consist of $35 \cdot 30 = 1050$ inputs, identical in size to the HMD datasets used in Section 5.

A three letter code is used to identify each synthetic mortality structure. The first two experiments (SYA, SYB) use identically distributed, independent observation noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and the third (SYC) takes heteroskedastic noise $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2/D_{\mathbf{x}})$, with $D_{\mathbf{x}}$ coming from the real-world HMD Japan Female dataset to capture realistic heterogeneity in age. The precise setups along with the hyperparameters are described in Table 4, and the resulting synthetic mortality surfaces are available for public re-use at github.com/jimmyrisk/gpga-synthetic-surfaces. All of the synthetic kernels and their hyperparameters are real-world-plausible. Namely, SYA, SYB, and SYC all came from prototypical GA runs on HMD datasets and hence match their structure.

The first case study (SYA) starts with an exceptionally simple structure (kernel of length 3, i.e., 2 base kernels) as a check on whether the GA preserves parsimony when exploring the kernel space. In other words, we use SYA to validate that longer-length kernels would correctly be perceived as over-fitting during the GA evolution and receive lower fitness scores compared to the true kernel. As a secondary effect, SYA addresses recovery of kernel smoothness, as $\text{RBF}_a \cdot \text{RBF}_y$ is smooth in both Age and Year components. Since we minimize BIC for the given training set, it is plausible that a different kernel

Table 5. *Top five fittest non-duplicate kernels for the first synthetic case study SYA. Bolded is $K_0 = RBF_yRBF_a$, the true kernel used in data generation. SYA-1 and -2 denote the realization trained on.*

SYA-1			SYA-2		
BIC	$\widehat{\text{BF}}(k, K_0)$	Kernel	BIC	$\widehat{\text{BF}}(k, K_0)$	Kernel
–2034.23	**1.0000*****	$\mathbf{RBF}_a\mathbf{RBF}_y$	–2066.93	1.1907***	$M52_aRBF_y$
–2034.04	0.8264***	$M52_aRBF_y$	**–2066.76**	**1.0000*****	$\mathbf{RBF}_y\mathbf{RBF}_a$
–2031.82	0.0902*	$M52_aM52_y$	–2064.63	0.1216**	$M52_aM52_aRBF_y$
–2031.29	0.0526*	$M52_aRBF_aRBF_y$	–2064.24	0.0801*	$M52_aRBF_aRBF_y$
–2031.09	0.0433*	$M52_aM52_aRBF_y$	–2063.88	0.0561*	$M52_aM52_aRBF_y$

from the generating one might actually achieve a (slightly) lower BIC for a given realization, so this experiment is performed twice (generating SYA-1 and SYA-2) to assess sampling variability.

The second synthetic example (SYB) features a more sophisticated kernel of length 7 (with 3 base kernels) and moreover combines both multiplicative and additive structure. It is motivated by Lee–Carter models and has a multiplicative age-period component with a less prominent additive cohort effect (coefficient of 0.02 versus 0.08). Its purpose is to (i) check whether the GA is able to identify such high-level structure (including addition and multiplication, on the correct terms), (ii) identify low-length kernels as under-fitting, and (iii) distinguish between all three of Age, Year, and Cohort terms.

The third experiment (SYC) employs a multiplicative four-component kernel, primarily to test the GA's handling of complex structures and potential over-parameterization due to dual period effects (Chy_y and M12_y). Except for Chy_y, all kernels are in $\mathcal{K}_r$. The large lengthscales of M12_y and $M12_c$ both represent near nonstationarity. The analysis spans both $\mathcal{K}_r$ and $\mathcal{K}_f$, with one focus on how the GA approximates Chy_y using $\mathcal{K}_r$-kernels.

Both SYA and SYB are done purely searching over $\mathcal{K}_r$, with SYA testing a basic APC setup, and SYB emphasizing recovering additive structure and linear coefficients. Using $\mathcal{K}_f$ could muddle the analysis and is therefore left for SYC. Random number generator seeding for the initial generation is unique to each GA run; all GA runs have identical algorithmic parameters as described in Tables 2 and 3.

Remark 5. The theoretical BF in Equation (3.2) assumes all hyperparameters have been integrated out $p(y|k) = \int_\theta p(y|k, \theta)d\theta$, or replaced with MLE's when using BIC. Thus, the hypothesis being tested through the BFs is purely about the kernel choice.

4.1. Synthetic results

Answers to questions presented in the previous section are found in Table 5 for SYA, Table 7 for SYB, and Table 8 for SYC. In all tables, K_0 denotes the known kernel that generated the synthetic data. With the goal in mind to establish kernel recovery (ignoring hyperparameter estimation), $\hat{\theta}$ is estimated for K_0 from the generated data.

SYA

Table 5 shows the results of the top 5 fittest kernels for SYA-1 and SYA-2. The estimated BF (using BIC) is the column $\widehat{\text{BF}}(k, K_0) = \exp(\text{BIC}(K_0) - \text{BIC}(k))$. For SYA-1, the true kernel $K_0 = \text{RBF}_y\text{RBF}_a$ is discovered and appears with lowest BIC. Next best (with a large BF of 0.826) is $M52_aRBF_y$, which has an identical structure aside from using the slightly less smooth Matérn-5/2 kernel for age instead of RBF. Note that the BF would need to be below $1/3 \approx 0.3333$ to even be worth mentioning a difference between these kernels according to Table 15 in the Appendix, suggesting an indifference of the two models. Our interpretation is that sampling variability makes $M52_a$ a similar alternative to RBF_a. Note,

Table 6. *Logarithmic Bayes Factor* $\log \widehat{BF}(k, K_0) = BIC(K_0) - BIC(k)$ *for* $K_0 = RBF_a RBF_y$ *and* $k = k_a k_y$, *where* k_a *and* k_y *are in the respective row and column labels and* $BIC(K_0)$ *is evaluated over SYA-1 (left panel) and SYA-2 (right panel).*

	SYA-1				SYA-2			
	$M12_a$	$M32_a$	$M52_a$	RBF_a	$M12_a$	$M32_a$	$M52_a$	RBF_a
$M12_y$	–63.98	–38.96	–37.49	–39.41	–96.99	–68.85	–66.12	–65.25
$M32_y$	–35.83	–8.13	–5.98	–7.12	–53.24	–21.19	–17.67	–16.06
$M52_y$	–32.29	–4.70	–2.41*	–3.33*	–42.75	–11.36	–8.09	–6.92
RBF_y	–30.24	–3.14*	–0.19***	0.00***	–32.67	–2.31*	0.17***	0.00***

however, that any lower Matérn order does not appear in either table. The remaining three rows have a similar kernel structure with superfluous Age or Period kernels added. The BFs are below 0.1, suggesting strong evidence against these kernels as being plausible alternatives for the generated dataset. This experiment confirms the GA's ability to recover the overall structure, with an understandable difficulty in distinguishing between $M52_a$ and RBF_a. SYA-2 in the right panel of Table 5 considers a different realized mortality surface under the same population distribution to assess stability across GA runs. It tells a similar story, though interestingly the lowest-BIC kernel is now $M52_a RBF_y$, with $\widehat{BF}(k, K_0) = 1.1907$. This means that it achieves a lower BIC than the true kernel, showcasing possibility of overfitting to data. At the same time, since the BF is so close to 1, this is still not statistically worth mentioning and hence can be fully chalked up to sampling variability. Otherwise, we again observe only two truly plausible alternatives, and very similar less-plausible (BF between 0.05 and 0.13) alternates.

To further assess smoothness detection, multiplicative Age–Period kernels $k = k_a k_y$ are fit to both SYA-1 and SYA-2 datasets, where k_a and k_y range over M12, M32, M52 and RBF (in order of increasing smoothness), resulting in a total of 16 combinations per training surface. The resulting differences in BIC are provided in Table 6. Note that the ground truth kernel $K_0 = RBF_a RBF_y$ generates a mortality surface that is infinitely differentiable in both Age and Period. In both cases, decisive evidence is always against either surface having a M12 component (with $BIC(K_0) - BIC(k)$ ranging from -96.99 to -30.24 – recall that anything below –4.61 is decisive evidence for K_0). As found above, $M52_a$ is a reasonable surrogate for RBF_a for both SYA-1 and SYA-2. SYA-1 shows only strong evidence against $M52_y$ (as opposed to decisive for SYA-2). In both cases, there is decisive evidence against $M32_y$, with only strong evidence on the age component $M32_a$ when using RBF_y (otherwise, it is decisive). This difference is likely explained by the additional flexibility of $M32_a$ to pick up some fluctuations which would normally be reserved for the parametric mean function in age.

SYB

Next, we discuss SYB where the training surface is generated from $K_0 = 0.08 \cdot RBF_a M12_y + 0.02 \cdot M52_c$. Table 7 shows that there a total of four plausible kernels found (with BFs above 0.1). The GA discovers the ground truth (bolded in Table 7) with slightly different hyperparameters, and three closely related alternates. In particular, these four fittest kernels all identify the correct number of APC terms with a multiplicative Age–Period term plus an additive Cohort term. The RBF_a term is recovered in all four, and the corresponding Age lengthscales are very close to the true $\ell^a_{\text{len}} = 0.586$. The Cohort effect is captured either via the ground truth $M52_c$ (top-2 candidates) or RBF_c, a substitution phenomenon similar to what we observe for SYA above. The respective lengthscale is correctly estimated to be within 0.3 of the true $\ell^c_{\text{len}} = 0.079$. The Period effect is captured either by the ground truth $M12_y$ or by Min_y. The data-generating K_0 has a very large $\ell^y_{\text{len}} = 13.33$ for the $M12_y$ term, which corresponds to an AR(1) process with $\phi_0 = \exp(-1/13.33) = 0.9975$ after unstandardizing then transforming. Over the training range of 30 years, this is almost indistinguishable from a random walk with a Min kernel. As a result, the

Table 7. *Top five fittest non-duplicate kernels for the second synthetic case study SYB. Bolded is the true kernel used in data generation* $K_0 = 0.08 \cdot RBF_a(19.3) \cdot M12_y(386.6) + 0.02 \cdot M52_c(4.98)$*, with* $\sigma^2 = 4 \times 10^{-4}$, $\beta_0 = -9.942$, $\beta_1 = 0.0875$.

SYB					
BIC	$\widehat{\text{BF}}(k, K_0)$	Kernel	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_{ag}$
–2468.0	1.033^{***}	$0.014 \cdot \text{RBF}_a(19.5)\text{Min}_y(4.42) + 0.018 \cdot \text{M52}_c(5.23)$	3.52×10^{-4}	−9.049	0.096
–2467.8	$\mathbf{1.000}^{***}$	$\mathbf{0.063 \cdot \text{RBF}_a(19.89)\text{M12}_y}(259.2) + \mathbf{0.0180 \cdot \text{M52}_c}(5.29)$	$\mathbf{3.448} \times 10^{-4}$	**−9.045**	**0.098**
–2465.9	0.161^{**}	$0.060 \cdot \text{RBF}_a(19.5)\text{M12}_y(239.7) + 0.016 \cdot \text{RBF}_c(3.21)$	3.60×10^{-4}	−9.049	0.096
–2465.7	0.127^{**}	$0.014 \cdot \text{RBF}_a(19.5)\text{Min}_y(4.3) + 0.016 \cdot \text{RBF}_c(3.21)$	3.60×10^{-4}	−4.249	3.298
–2464.7	0.047^{*}	$0.022 \cdot \text{RBF}_a(20.1)\text{Lin}_y(2.8)\text{M12}_y(303.9) + 0.016 \cdot \text{M52}_c(5.23)$	3.45×10^{-4}	−9.162	0.097

Table 8. *Fittest non-duplicate kernels for SYC in two separate runs, one over* $\mathcal{K}_r$ *and the other over* $\mathcal{K}_f$*. The true kernel is* $K_0 = M52_a\left(Chy_y M12_y\right) M12_c$*. Note that* $Chy_y \notin \mathcal{K}_r$*. All* $\widehat{BF}$ *values are in the* *** *evidence category.*

SYC					
Restricted search set $\mathcal{K}_r$			Full search set $\mathcal{K}_f$		
BIC	$\widehat{\text{BF}}(k, K_0)$	Kernel	BIC	$\widehat{\text{BF}}(k, K_0)$	Kernel
–2723.57	1.9874	$\text{M52}_a(\text{RBF}_y\text{M12}_y)\text{M12}_c$	–2723.57	1.9874	$\text{M52}_a(\text{RBF}_y\text{M12}_y)\text{M12}_c$
–2723.53	1.9055	$\text{M52}_a(\text{RBF}_y\text{Min}_y)\text{M12}_c$	–2722.89	1.0000	$\text{M52}_a(\text{Meh}_y\text{M12}_y)\text{M12}_c$
–2723.45	1.7642	$\text{M52}_a(\text{RBF}_y\text{M12}_y)\text{Min}_c$	–2722.89	1.0000	$\mathbf{M52}_a(\mathbf{Chy}_y\mathbf{M12}_y)\mathbf{M12}_c$
–2722.81	0.9258	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{M12}_c$	–2722.85	0.9646	$\text{M52}_a(\text{Chy}_y\text{Min}_y)\text{M12}_c$
–2722.71	0.8369	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{Min}_c$	–2722.81	0.9258	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{M12}_c$

substitution with Min_y is unsurprising, as is the wide range of estimated ℓ_{len}. For example, $\hat{\ell}^y_{\text{len}} = 259.057$ in the second row corresponds to AR(1) persistence parameter of $\phi = 0.9960$, very close to the true ϕ_0. We furthermore observe stable recovery of all non-kernel hyperparameters $\left(\sigma^2, \beta_0, \beta_{ag}\right)$, with estimates close to their true values.

Finally, the GA is also very successful in learning that the Age–Period component (coefficient 0.08) dominates the Cohort component (coefficient 0.02), conserving the relative amplitude of the two components. Note that the Min and Lin kernels include an offset, which mathematically result in similar linear coefficients to the true 0.08 and 0.02. For example, the first term in the first row simplifies to $0.0141 \cdot \text{RBF}_a \cdot \left(4.42 + x_{yr} \wedge x'_{yr}\right) = 0.0623 \cdot \text{RBF}_a + 0.0141 \cdot \text{RBF}_a \cdot \left(x_{yr} \wedge x'_{yr}\right)$. This results in all five kernels in Table 7 discovering the first component to be larger than the second by a factor of 3.5–4.

SYC

Table 8 presents the SYC results. Multiple models exhibit BFs greater than 1, or very close to 1, indicating that it is impossible to tell which is the "best" compound kernel. At the same time, all kernels in the top-5 list for $\mathcal{K}_f$ are small variations of the true kernel (which shows up in the third spot), differing by only

one kernel component. Moreover, $\mathrm{M52}_a$ and $\mathrm{M12}_c$ appear consistently, underscoring their importance in the true model. We yet again observe BIC-equivalence between Chy and RBF (see Appendix A) as well as M12 and Min. Same pattern holds for searching in $\mathcal{K}_r$ that substitutes RBF_y or $\mathrm{M52}_y$ for the true Chy_y term.

Against the above non-uniqueness of the best expressive kernel, the GA does provide strong evidence on the overall structure, namely the number of total terms in K_0 and the presence of an additive structure. The best kernel with three components (rather than four) is $\mathrm{M52}_a\mathrm{Min}_y\mathrm{M12}_c$ with $\widehat{\mathrm{BF}} < 1 \times 10^{-6}$; the best kernel with five components is $\mathrm{M52}_a(\mathrm{Lin}_y\mathrm{M12}_y\mathrm{RBF}_y)\mathrm{M12}_c$ ($\widehat{\mathrm{BF}} = 0.0419$). Thus, the BIC criterion leads the GA to correctly reject kernels that are too short or too long for the SYC dataset. Finally, the best kernel with an additive component is $\mathrm{M52}_a(\mathrm{M12}_y + \mathrm{RBF}_y)\mathrm{M12}_c$ ($\widehat{\mathrm{BF}} = 0.0595$), providing decisive evidence that the BIC also properly learns the lack of any further additive terms.

5. Results on Human Mortality Database data

After validating our generative kernel exploration with synthetic data, we move to realistic empirical analysis. Unless otherwise mentioned, we use the same prior mean $m(\mathbf{x}) = \beta_0 + \beta_{\mathrm{ag}} x_{\mathrm{ag}}$ as in the previous section, which is a reasonable prior trend for the age ranges we consider. Our discussion focuses on retrospective analysis, namely the performance of different kernels assessed in terms of fitting a given training set. Thus, we do not pursue out-of-sample metrics, such as (probabilistic) scores for predictive accuracy and concentrate on looking at the BIC scores augmented with a qualitative comparison. Retrospective assessment parallels the core of APC methods that seek to decompose the data matrix via singular value decomposition, prior to introducing the out-of-sample dynamics in the second step. A further reason for this choice is that predictive accuracy is fraught with challenges (such as handling idiosyncratic data like the recent 2020 or 2021 mortality driven by COVID), and there is no canonical way to assess it. In contrast, BIC offers a single "clean" measure of statistical fit for a GP and moreover connects to the BF interpretation of relative preponderance of evidence.

Below we consider four representative national populations from HMD covering three countries and both genders. As our first case study, we consider the Japanese Female population. We utilize the HMD dataset covering Ages 50–84 and years 1990–2018. The same top-level and crossover/mutation algorithmic parameters of the GA are used as in the last section (Tables 2 and 3, respectively).

Table 9 provides summary statistics regarding the top kernels in $\mathcal{K}_f$ that achieve the lowest BIC scores. In order to provide a representative cross-sectional summary, we consider statistics for the top-10, top-50, and then 51–100, 101–150, and 151–200th best kernels. Recall that there are a total of $20 \cdot 200 = 4000$ kernels proposed by the GA, so top-200 correspond to the best 5% of compositions. We find that the best fit is provided by purely multiplicative kernels (single additive component) with 3 or 4 terms. This includes one term for each of APC coordinates, plus a possible 4th term, usually in Cohort or Year. In the restricted class $\mathcal{K}_r$, all of the top-10 kernels are of this form, as are 9 out of top-10 kernels found in $\mathcal{K}_f$.

The above APC structure moreover includes a rough (non-differentiable or only once-differentiable) component in Year and in Cohort. This matches the logic of time-series models for evolution of mortality over time. Note that in our setting, it can be interpreted as a strong correlation of observed noise across Ages, in other words the presence of environmental disturbances (epidemics, heat waves, other co-morbidity factors) that yield year-over-year idiosyncratic impacts on mortality. On the other hand, in Age best fits are smooth, most commonly via the M52 kernel. This matches the intuition that the Age-structure of mortality is a smooth function. Table 9 documents a strong and unequivocal cohort effect: present 100% in all top kernels. This is consistent with Willets (2004) who states that a strong cohort effect for the Japanese Female population can be projected into older ages.

The presence of multiple factors (length above 3) generally indicates one or both of the following: (i) multi-scale dependence structure and (ii) model mis-specification. On the one hand, since we are considering only a few kernel families, if the true correlation structure is not matched by any of them,

Table 9. *Summary statistics of the top kernels for the re-run and robust checks for JPN Females across both $\mathcal{K}_r$ and $\mathcal{K}_f$. `addtv comps` refers to the average number of additive components (the frequency of appearance of the "+" operator plus one); `num` refers to the total average number of kernel terms in the respective coordinate; `non-stat` reports the percentage of returned compositional kernels that include any of Min, Meh families; `rough` reports the fraction that include any of the M12, Min, AR2 families. For each row we average all metrics among the respective kernels: top-10, top-50, and those ranked 51–100.*

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%
101–150	–2717.03	–2718.17	5.02	1.14	4%	1.30	2.18	1.54	6%	98%	100%
151–200	–2715.77	–2717.01	5.10	1.48	12%	1.28	2.36	1.46	6%	100%	100%
JPN Female Re-run											
1–10	–2723.05	–2725.29	4.00	1.00	0%	1.00	1.60	1.40	0%	100%	100%
1–50	–2719.82	–2725.29	4.14	1.08	10%	1.08	1.58	1.48	0%	98%	100%
51–100	–2718.30	–2719.82	4.46	1.26	8%	1.14	1.62	1.70	6%	100%	100%
JPN Female Search in $\mathcal{K}_r$											
1–10	–2724.11	–2725.27	4.00	1.00	0%	1.00	1.70	1.30	0%	100%	100%
1–50	–2721.19	–2725.27	4.48	1.10	8%	1.14	1.96	1.38	0%	100%	100%
51–100	–2718.06	–2721.19	4.72	1.50	18%	1.16	1.96	1.60	0%	100%	100%
JPN Female trained on $\mathcal{D}_{\text{rob},1}$											
1–10	–2724.11	–2725.29	4.00	1.40	40%	1.00	1.50	1.50	0%	100%	100%
1–50	–2716.84	–2725.29	4.42	1.12	18%	1.14	1.64	1.64	0%	100%	100%
51–100	–2714.96	–2716.58	4.70	1.16	12%	1.18	1.68	1.84	0%	100%	100%
JPN Female trained on $\mathcal{D}_{\text{rob},2}$											
1–10	–2724.33	–2725.23	4.00	1.00	0%	1.00	1.10	1.90	0%	100%	100%
1–50	–2719.09	–2725.23	4.22	1.00	4%	1.04	1.38	1.80	0%	100%	100%
51–100	–2718.16	–2719.10	4.62	1.18	14%	1.32	1.40	1.90	0%	100%	100%

the algorithm is going to substitute with a combination of the available kernels. Thus, for example, using both a rough and a smooth kernel in Year indicates that neither of the M12 or RBF fit well on their own. On the other hand, the presence of additive structure, or in general the need for many terms (especially over 5) suggests that there are many features in the correlation structure of the data, and hence it does not admit any simple description.

In JPN Females, the GA's preference for parsimony is confirmed by the fact that the best-performing kernels are the shortest. We observe a general pattern that Length is increasing in Rank. In particular, going down the rankings, we start to see kernels with two additive components. We may conclude that the second additive component provides a minor improvement in fit, which is outweighed by the complexity penalty and hence rejects on the grounds of parsimony.

Table 10. *Fittest non-duplicate kernels for Japanese Females in two separate runs, one over $\mathcal{K}_f$ and the other over $\mathcal{K}_r$. Bayes Factors $\widehat{BF}$ are relative to the best found kernel $k^*_{JPN-FEM}$ and all have insubstantial significance. †Daggered kernels under $\mathcal{K}_f$ column are those that also belong to $\mathcal{K}_r$.*

Japan Female HMD Dataset for 1990–2018 and Ages 50–84					
$\mathcal{K}_r$			$\mathcal{K}_f$		
BIC	$\widehat{\text{BF}}$	Kernel	BIC	$\widehat{\text{BF}}$	Kernel
–2725.288	0.995	$\text{M52}_a(\text{RBF}_y\text{M12}_y)\text{M12}_c$	–2725.293	1	$\text{M52}_a(\text{Chy}_y\text{M12}_y)\text{M12}_c$
–2725.270	0.977	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{M12}_c$	–2725.270	0.977†	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{M12}_c$
–2725.233	0.941	$\text{M52}_a(\text{RBF}_y\text{Min}_y)\text{M12}_c$	–2725.221	0.931†	$\text{M52}_a(\text{M52}_y\text{Min}_y)\text{M12}_c$
–2725.221	0.931	$\text{M52}_a(\text{M52}_y\text{Min}_y)\text{M12}_c$	–2724.623	0.512†	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{Min}_c$
–2724.640	0.520	$\text{M52}_a(\text{M52}_y\text{M12}_y)\text{Min}_c$	–2724.510	0.457	$\text{M52}_a(\text{M12}_y\text{M32}_y)\text{M12}_c$

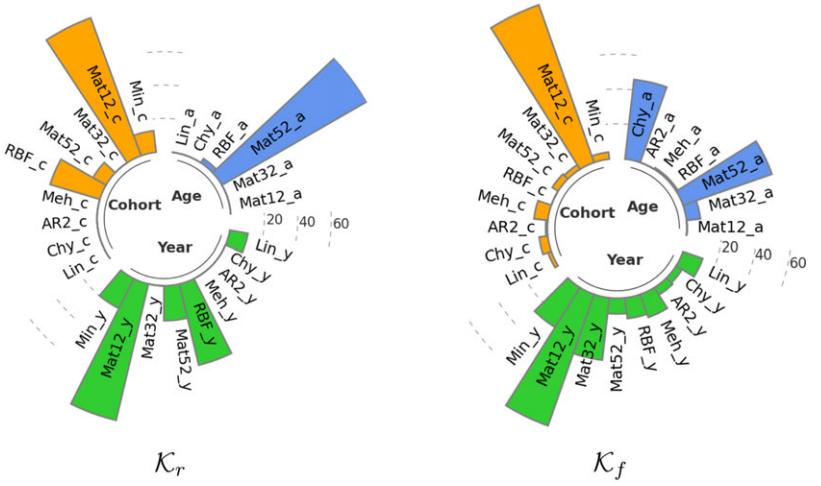

Figure 2. *Frequency of appearance of different kernel families in JPN Female models.*

The compositional kernel that achieves the lowest overall BIC is

$$k^*_{JPN-FEM} = 0.4638 \cdot M52_a(37.7) \cdot \text{Chy}_y(56.6) \cdot M12_y(1810) \cdot M12_c(7378).$$

Note the purely multiplicative structure of $k^*_{JPN-FEM}$ and its two Period terms, capturing both the local rough nature and the longer-range dependence. Table 10 lists the next-best alternatives, both within $\mathcal{K}_f$ and within $\mathcal{K}_r$. We see minimal loss from restricting to the smaller $\mathcal{K}_r$, as three of five top kernels found in $\mathcal{K}_f$ actually belong to $\mathcal{K}_r$. Thus, casting a "wider net" does not improve BIC, suggesting that most of the kernel options added to $\mathcal{K}_f$ are either close substitutes to the base ones in $\mathcal{K}_r$ or do not specifically help with HMD data. Indeed, the BF improvement factor is just exp (0.19) from Table 9. Moreover, we also find that there is strong hyperparameter stability across different top kernels. For example, we find that the lengthscale in Age (which is always captured via a M52 kernel in Table 10) is consistently in the range [37, 38.5]. Similarly, the lengthscales for the M12 kernel in Cohort are large (> 6000).

Figure 2 visualizes the frequency of the appearance of different kernels. We consider the top 100 unique kernels returned by the GA and show the number of times each displayed kernel is part of the composite kernel returned. Note that sometimes the same kernel can show more than once. In the left panel, we consider GA searching in $\mathcal{K}_r$, hence many of the families are not considered; the right panel looks at the full $\mathcal{K}_f$.

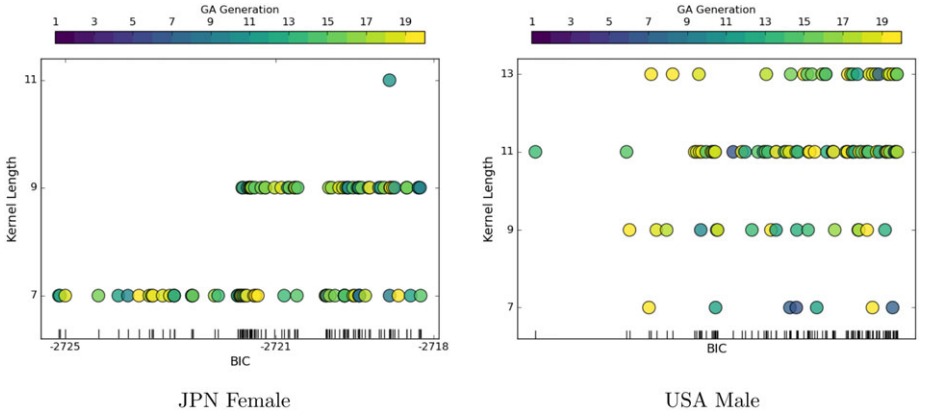

Figure 3. *Properties of the top 100 kernels found by GA.*

The barplots in Figure 2 indicate that there is a substantial variability in selected kernels when considering the top 100 proposals. Nevertheless, we observe the typical decomposition into "principal" factors, such as $M12_c$ and $M12_y$ for JPN Females, plus additional residual kernels. The latter generate second-order effects and are not easily identifiable, leading to a variety of kernels showing up for 5–15% of the proposals. For example, nearly every kernel family in Period can be used to construct a good compositional kernel. This heterogeneity of kernels picked indicates that it is not appropriate to talk about "the" GP model for a given dataset, as there are several, quite diverse fits that work well.

Figure 3 shows several summary statistics of proposed kernels against their BIC scores. We display the top 100 unique kernels, arranged according to their total length (y-axis) and generation found (color). First, we observe that there is an increasing clustering of kernels as we march down the BIC order (x-axis). In other words, there is typically a handful (1–5) of best-performing kernels, and then more and more equally good alternatives as the BIC decreases. This matches the interpretation of BFs: accepting the best-performing kernel as the "truth", we find several plausible alternatives, a couple dozen of somewhat plausible ones, and many dozens of weakly plausible ones. The spread of the respective BF factors varies by population; in some cases there are only $\sim 50 - 60$ plausible alternatives, in others there are well over a hundred.

Second, we observe that most of the best kernels are found after 10+ generations, matching the logic of the GA exploring and gradually zooming into the most fit kernel families. However, that pattern is not very strong, and occasionally the best kernel is discovered quite early on.

Third, we observe a pattern in terms of kernel complexity vis-a-vis its fitness, matching the above logic: the best-performing kernels tend to be of same length (and are very similar to each other, often just 1 mutation away), but as we consider (weakly) plausible alternatives, we can find both more parsimonious and more complex kernels. This captures the parsimony trade-off: shorter kernels have lower likelihood but smaller complexity penalty; longer kernels have higher log-likelihood but are penalized more.

In Figure 4, we present in-sample and future forecasts of log-mortality for JPN Female Age 65. The left panel uses the top-10 kernels, providing their posterior mean and prediction intervals. The in-sample fit is tightly constrained, while the out-of-sample prediction becomes more heterogeneous as we move away from the training sample. In particular, there is a bimodal prediction that groups kernels, with some projecting future mortality improvement and others moderating the downward trend. Examining the GA output, we find that there are two "clusters" of kernels among the best-performing ones. Some of them contain a Mehler kernel, either in Year or in Cohort, and others do not. The ones that do form the "bottom" cluster in Figure 4, that is, they predict relatively large improvements in future Japanese mortality. The ones that do not utilize Mehler (all top-5 belong to this category) predict more moderate MI. The overlayed 90% posterior prediction intervals indicate a common region for future mortality

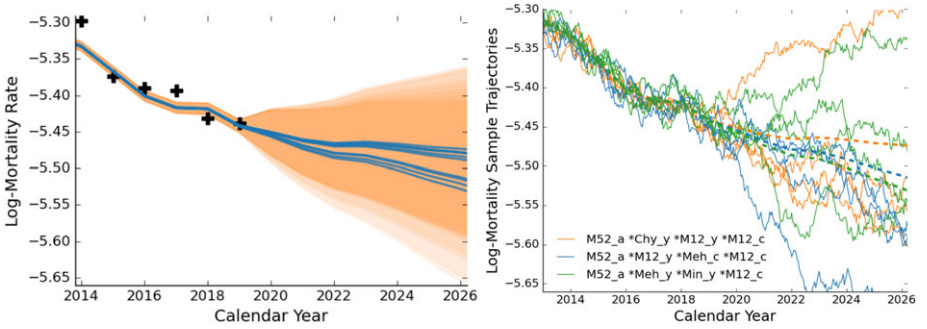

Figure 4. *Predictions from the top 10 kernels in $\mathcal{K}_f$ for JPN Females Age 65. Left: predictive mean and 90% posterior interval from the top-10 kernels. For comparison, we also display (black plusses) the six observed log-mortality rates during 2014–2019. Right: four sample paths from each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

trajectories, with symmetric fanning as calendar year increases and a slight skew toward lower mortality rates deeper in future years. This forecasting approach can serve as a basis for Bayesian model averaging, utilizing the BFs as weights. Furthermore, we observe a square-root type fanning of variance, which is common in random-walk mortality models.

The right panel of the figure investigates the stochasticity of the GP by simulating paths using three representative kernels. In-sample paths cluster closely around their posterior means, with observed difficulties in deviating far from the observed data. The observed roughness in the trajectories is a consequence of including a M12 or Min component in Calendar Year or Cohort. When examining the trajectories out-of-sample, the impact of individual kernels becomes more apparent, particularly in the green and blue paths. Lack of mean reversion is more evident in these paths, which could be attributed to the presence of the nonstationary Min kernel in calendar year (green) and Mehler kernel in cohort (blue).

5.1. Robustness check

To validate the above results of the GA, we perform two checks: (i) re-run the algorithm from scratch, to validate stability across GA runs; (ii) run the GA on two modified training datasets: $\mathcal{D}_{\text{rob},1}$, $\mathcal{D}_{\text{rob},2}$. For $\mathcal{D}_{\text{rob},1,}$ we augment with two extra calendar years (beginning at 1988 instead of 1990), and four extra Age groups (48–86 instead of 50–84). For $\mathcal{D}_{\text{rob},2,}$ we shift the dataset by 4 years in time, namely to 1986–2015, considering same Age range 50–84.

In all above cases, we expect results to be very similar to the "main" run discussed above. While the GA undertakes random permutations and has a random initialization, we expect that with 200 kernels per generation and 20 generations, the GA explores sufficiently well that the ultimate best-performing kernels are invariant across GA runs. This is the first justification to accept GA outputs as the "true" best-fitting kernels. Similarly, while the BIC metric is determined by the precise dataset, it ought to be sufficiently stable when the dataset undergoes a small modification, so that the top kernels can be interpreted as being the right ones for the population in question, and not just for the particular data subset picked.

The above robustness checks are summarized below and in Table 9. These confirm that the GA is stable both across its own runs (see the "Re-run" listings) and when subjected to slightly modifying the training dataset or "rolling" it in time. We observe that we recover essentially the same kernels (both in $\mathcal{K}_r$ and $\mathcal{K}_f$), and moreover the best kernels/hyperparameters are highly stable as we enlarge the dataset (see the "Robust" listings). This pertains both to the very top kernels, explicitly listed below, but also to the larger set of top-100 kernels, see the summary statistics in Table 9. Re-assuringly, the kernel

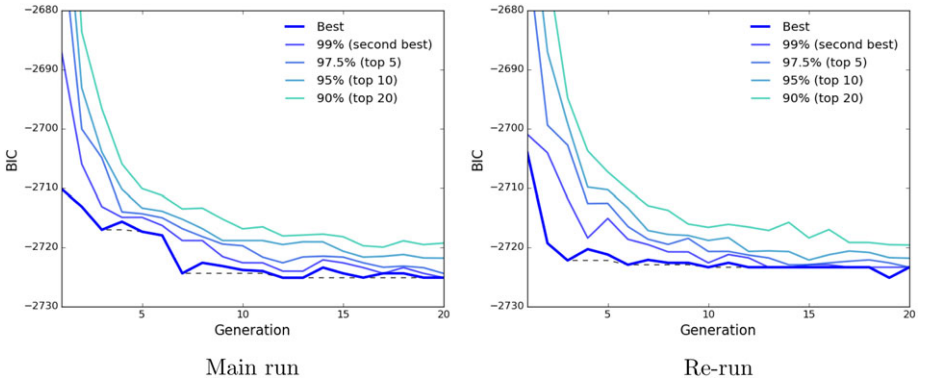

Figure 5. *Summary statistics of best kernels proposed by GA as a function of generation g.*

lengthscales listed below change minimally when run on a larger dataset, confirming the stability of the MLE GP sub-routines. In particular, the same kernel is identified as the best one during the re-run in $\mathcal{K}_r$, and it shows up yet again as the best for $\mathcal{D}_{\text{rob,1}}$, with just slightly modified parameters:

$$\text{original } \mathcal{D}\text{: } 0.4651 \cdot \text{M52}_a(37.7) \cdot \text{M52}_y(52.2) \cdot \text{M12}_y(1821) \cdot \text{M12}_c(7412);$$
$$\text{enlarged } \mathcal{D}_{\text{rob,1}}\text{: } 0.4646 \cdot \text{M52}_a(37.7) \cdot \text{M52}_y(52.2) \cdot \text{M12}_y(1819) \cdot \text{M12}_c(7403).$$

Four out of the five best kernels repeat when working with $\mathcal{D}_{\text{rob,1}}$. This stability can be contrasted with Cairns *et al.* (2011) who comment on sensitivity of SVD-based fitting to date range. The other alternatives continue to follow familiar substitution patterns. Of note, with the run over $\mathcal{D}_{\text{rob,1}}$ there is the appearance of Chy_a, but no appearance of Min_y, M32_y, or Meh_c. As can be seen in Figure 2, Chy_a is actually quite common.

Furthermore, all runs (original, re-run, enlarged dataset) always select M52 or Cauchy kernel for the Age effect, Matérn-1/2 (or sometimes Min) in Period, typically augmented with a smoother kernel like M32_y, Meh_y, Chy_y, and M12_c in Cohort. We furthermore record very similar frequency of different kernels among top-100 proposals, and similar BIC scores for the re-run.

As another validation of GA convergence, Figure 5 shows the evolution of fitness scores over generations. We display the BIC of the best kernel in generation g, as well as the second best (99% quantile across 200 kernels), 5th best (97.5%), 10th best (95%), and 20th best (90%) across the main run of the GA and a "re-run". The experiments in each panel differ only through a different initial seed for the first generation. In both settings, only minimal performance increases (according to the minimum BIC) are found beyond generation 12 or so. Since in each new generation there is inherent randomness in newly proposed kernels, there is only distributional convergence of the BIC scores as new kernels are continuously tried out. This churn is indicated by the flat curves of the respective within-generation BIC quantiles. In sum, the GA converges to its "equilibrium" after about a dozen generations, validating our use of $G = 20$ for analysis.

5.1.1. Robustness to prior mean

To examine the stability of top-ranking kernel compositions under different prior mean functions, we compare our primary choice of the linear prior mean $m_{lin}(\mathbf{x}) = \beta_0 + \beta_{ag} x_{ag}$ to the following two alternatives:

(i) A constant mean function $m_c(\mathbf{x}) = \beta_0$;
(ii) A calendar year-averaged mean $m_{ya}(\mathbf{x}) = n_{\text{yr}}^{-1} \sum_{x_{\text{yr}} \in \mathcal{D}} y(x_{\text{ag}}, x_{\text{yr}})$.

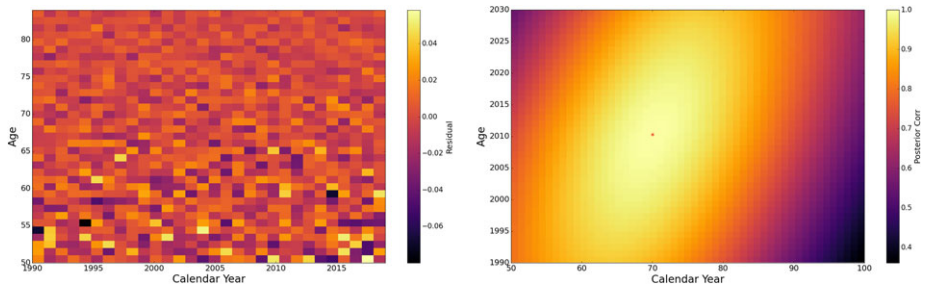

Figure 6. *Left: residuals from the best kernel in $\mathcal{K}_f$ for JPN Females. Right: implied prior correlation $k(\boldsymbol{x}_0, \boldsymbol{x}')$ of the best kernel as function of $\boldsymbol{x}'$ relative to the cell $\boldsymbol{x}_0 = (70, 2010)$ shown as the red dot.*

The constant mean m_c is the simplest possible prior and drops the second coefficient $\beta_{ag} = 0$, while still estimating β_0 alongside covariance hyperparameters during MLE. It is expected to yield a worse goodness-of-fit and also lead to longer lengthscales, as the GP must match the linear trend directly. The year-averaged mean m_{ya} is an example of a pre-determined (rather than fitted) mean function that provides an age-specific data-driven trend; this would generally be a more accurate de-trending compared to the linear function in $m_{lin}(\cdot)$. Year-averaging de-trending is used in Cairns *et al.* (2009) (comparing the Lee–Carter with and without cohort models), and also appears in the Renshaw–Haberman model (Renshaw and Haberman, 2006). Below our goal is not to compare performance across different mean functions but to assess robustness of covariance structure to choice of mean function.

Re-running the GA with these alternatives, we find the following top kernels:

$$\begin{aligned} m_{ya}\text{:}\quad & 0.1567 \cdot \mathrm{Chy}_a(45.4) \cdot (\mathrm{RBF}_y(24.7)\mathrm{M12}_y(579.9)) \cdot \mathrm{M12}_c(2562); \\ m_c\text{:}\quad & 2.1117 \cdot \mathrm{M52}_a(40.9) \cdot \mathrm{M12}_y(7873) \cdot (\mathrm{M12}_c(31494)\mathrm{Meh}_c(0.473)). \end{aligned}$$

In all, these results are remarkably similar to the primary run and demonstrate that the choice of the mean function mostly affects kernel *lengthscales* but not the selected kernel types. Throughout the three choices of $m(\cdot)$, we consistently get a purely multiplicative kernel with four terms, including a smooth Age effect (captured either with Chy or M52 kernels that appear fully substitutable), a quasi-nonstationary Year effect with a $\mathrm{M12}_y$ term, a similar cohort effect with a $\mathrm{M12}_c$ term, and a fourth smooth term, mostly in Cohort using one of Meh, Chy, M52 kernels. We note that with a constant mean there is less of a mean-reversion effect, that is, the GP is more focused on extrapolating the Year pattern, see the very large lengthscales in Year and Cohort. Also, as expected, we find a reduced GP process variance $\sigma^2_{f,ya} \approx 0.15 \ll \sigma^2_{f,lin} \approx 0.45$ for the year-averaged case (where the de-trended residuals modeled by the GP are smaller in magnitude) and a much larger $\sigma^2_{f,c} \approx 2$ for the constant prior mean.

5.2. Analysis of model residuals

In Figure 6, the left panel displays the residuals that compare the realized log-mortality rates of JPN Females with the GP prediction from the best-performing kernel. The absence of any identifiable structure, especially along the SW-NE diagonals that correspond to Birth Cohorts, indicates a statistically sound fit, consistent with the expected uncorrelated and identically distributed residuals. Additionally, we observe distinct heteroskedasticity, where residuals for smaller Ages exhibit higher variance. This is due to the smaller number of deaths at those Ages, resulting in a more uncertain inferred mortality rate, despite the larger number of exposures. Generally, the observation variance is lowest around Age 80.

The right panel of Figure 6 shows the implied prior correlation relative to the cell (70, 2010). The strong diagonal shape indicates the importance of the cohort effect. Moreover, we observe that the correlation decays about the same in Period (vertical) as in Age (horizontal), with the inferred lengthscales imposing a dependence of about ± 12 years in each direction.

Table 11. *Results from GA runs on JPN Male, US Male, and SWE Female. Throughout we search within the full set $\mathcal{K}_f$. See Table 10 for the full definition of all the columns.*

Range	BIC max	BIC min	len	addtv comps	non-stat.	num age	num year	num coh	rough age	rough year	rough coh
					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%

Table 12. *Best-performing kernel in $\mathcal{K}_r$ and $\mathcal{K}_f$ for each of the four 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.*

Pop'n/Search Set	N_{pl}	Top Kernel
JPN Female $\mathcal{K}_r$	90	$0.464 \cdot \mathrm{M52}_a(37.4) \cdot \mathrm{RBF}_y(38.6)\mathrm{M12}_y(1812) \cdot \mathrm{M12}_c(7438)$
JPN Female $\mathcal{K}_f$	95	$0.4638 \cdot \mathrm{M52}_a(37.7) \cdot \mathrm{Chy}_y(56.6)\mathrm{M12}_y(1810) \cdot \mathrm{M12}_c(7378)$
JPN Male $\mathcal{K}_r$	89	$0.1491 \cdot \mathrm{M52}_a(32.3) \cdot \mathrm{RBF}_y(33.4)\mathrm{M12}_y(761.0) \cdot \mathrm{M12}_c(1569)$
JPN Male $\mathcal{K}_f$	112	$0.2130 \cdot \mathrm{M52}_a(37.1) \cdot \mathrm{M12}_y(1311.6) \cdot M32_c(54.2)\mathrm{M12}_c(2566)$
US Male $\mathcal{K}_r$	57	$0.017 \cdot \mathrm{M12}_a(171.4) \cdot \mathrm{M52}_y(14.5)M12_y(299.6) \cdot \mathrm{M52}_c(22.7)\mathrm{M12}_c(315)$
US Male $\mathcal{K}_f$	35	$0.010 \cdot \mathrm{AR2}_a(38.1, 63.9) \cdot \mathrm{M12}_y(701.2) \cdot \mathrm{M32}_c(45.4) \cdot [4.6211 \cdot \mathrm{M12}_c(849.9) + 0.011 \cdot \mathrm{M32}_a(0.68) \cdot \mathrm{M52}_c(6.30)]$
SWE Female $\mathcal{K}_r$	200+	$0.2527 \cdot \mathrm{RBF}_a(17.68) \cdot \mathrm{M12}_y(2138) \cdot \mathrm{RBF}_c(39.1)$
SWE Female $\mathcal{K}_f$	200+	$0.2094 \cdot \mathrm{Chy}_a(35.7) \cdot \mathrm{M12}_y(1951) \cdot \mathrm{Meh}_c(0.857)$

5.3. Male versus female populations

We proceed to apply the GA to the JPN Male population. The rationale behind this comparison is the assumption of similar correlation structures between the genders, which enables us to both highlight the similarities and pinpoint the observed differences.

As expected, the JPN Male results (see Tables 11 and 12) strongly resemble those of JPN Females. Once again, we detect a strong indication of a single multiplicative term, characterized by APC structure with smooth Age and rough Period and Cohort effects. Just like for Females, a (rough) Cohort term is selected in all (100%) of the top-performing kernels for JPN Males. The best-fitting individual kernels, as shown in Table 12, are also similar, with M52 in Age, M52 or RBF in Year, and M12 in Cohort being identified as the optimal choices.

Some differences, such as more Cohort-linked kernels for JPN Males versus Females, are also observed. The Cohort lengthscale is much larger for females (7600 vs. 2500), and so is the rough $\mathrm{M12}_y$

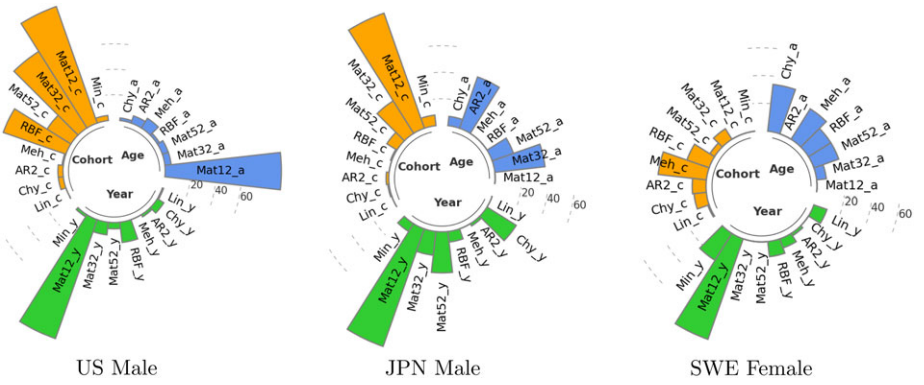

Figure 7. *Frequency of appearance of different kernels from $\mathcal{K}_f$ in US, SWE, and JPN Male models.*

lengthscale (1800 vs. 1100), while the Age lengthscales are almost identical. One interpretation is that there is more idiosyncratic noise in Male mortality, leading to faster correlation decay.

5.4. Analysis across countries

To offer a broader cross-section of the global mortality experience, we next also consider US males and Sweden females. In total, we thus analyze four datasets: US males, Japan females and males, and Sweden females. We note that Sweden is much smaller (10M population compared to 130M in Japan and 330M in USA) than the other two countries and therefore has much noisier data.

USA Males: The US male data lead to kernels of much higher length compared to all other datasets. The GA returns kernels with 5–6 base kernels and frequently includes two or even three additive terms. Moreover, the APC pattern is somewhat disrupted, possibly due to collinearity between the multiple Period and Cohort terms.

To demonstrate some of the observed characteristics, let us examine the top kernel in $\mathcal{K}_f$, as presented in Table 12. This kernel comprises 11 terms, including 2 additive terms. However, we note that the second term has a significantly smaller coefficient, indicating that it serves as a "correction" term that is introduced to account for a less prominent and identifiable feature relative to the primary terms. Additionally, we observe that this kernel incorporates both rough Cohort term $\mathrm{M12}_c$ and smoother ones, namely $\mathrm{M32}_c$ and $\mathrm{M52}_c$. This points toward a multi-scale Cohort effect, where a few exceptional years (such as birth years during the Spanish Flu outbreak in 1918–1919) are combined with generational patterns (e.g., Baby Boomers vs. the Silent Generation). Unlike other datasets, the US Male data even include a RBF_c term. Finally, the Age effect is described by the AR2 kernel, which is also commonly observed in the JPN Male population.

Moving down the list, there are also shorter kernels with a single component (no "+"), for example, $0.0129 \cdot \mathrm{M12}_a(117.0) \cdot (\mathrm{RBF}_y(18.3)\mathrm{M12}_y(228.8)) \cdot (\mathrm{M32}_c(27.1)\mathrm{M12}_c(244.4))$ which is fourth-best, and the length-7 $0.0113 \cdot \mathrm{M12}_a(88.7) \cdot (\mathrm{M32}_y(20.3)\mathrm{M12}_y(189.4)) \cdot \mathrm{M12}_c(183.3)$ which is seventh-best. In all, for US males we can find a plausible kernel of length 7, 9, 11, 13 when the best-performing one has length 11. This wide distribution of plausible kernel lengths (and a wide range of proposed kernel families) is illustrated in the right panel of Figure 3 and the middle column of Figure 7.

Another sign that the US data have inherent complexity is the wide gap in BF of the best kernel in $\mathcal{K}_f$ compared to that in $\mathcal{K}_r$, by far the biggest among all populations. Thus, restricting to $\mathcal{K}_r$ materially worsens the fit. In fact, we observe that all the top kernels in $\mathcal{K}_r$ are purely multiplicative (such as $\mathrm{M12}_a \cdot \mathrm{M52}_y\mathrm{M12}_y \cdot \mathrm{M52}_c\mathrm{M12}_c$), which is unlikely to be the correct structure for these data and moreover hints at difficulty in capturing the correlation in each coordinate, leading to multiple Period and Cohort terms. Within $\mathcal{K}_f$ only 35 plausible alternatives are found, 3–5 times fewer than in other datasets.

SWE Female: The Swedish Females dataset turns out to have two distinguishing features. First, it yields the simplest and shortest kernels that directly match the APC structure of three multiplicative terms. The average kernel length reported in Table 11 is the smallest for SWE Females, and additive terms appear very rarely. When kernels with more than three terms are proposed, these are usually still all-multiplicative and add either a second Period term (e.g. $\text{Meh}_a \cdot \text{M12}_y \cdot \text{Meh}_y \cdot \text{Meh}_c$) or a second Age term, though both are smooth: $(\text{Chy}_a\text{RBF}_a) \cdot \text{M12}_y \cdot \text{Meh}_c$).

Second, and unlike all other populations above, the Cohort effect is ambiguous in Sweden. About 15% of the top performing kernels (and 30% in $\mathcal{K}_r$) have no Cohort terms at all, instead proposing two terms for the Period effect. Those that do include a Cohort effect, use either a RBF_c or a Meh_c term, indicating no short-term cohort features, but only generational ones. Nine of the top-10 kernels in $\mathcal{K}_f$ and only 7 out of 10 in $\mathcal{K}_r$ have Cohort terms. In contrast, in JPN and USA rough cohort terms are present in every single top-100 kernel. Once again, our results are consistent with the literature, see Murphy (2010) who discusses the lack of clarity on cohort effect for the Swedish female population.

A third observation is that SWE Female shows a compression of BIC values, that is, a lot of different kernels are proposed with very similar BICs. The GA returns over two hundred kernels with a BF within a ratio of 30 to the top one. This indicates little evidence to distinguish many different choices from each other and could be driven by the lower complexity of the Swedish mortality data.

5.5. Discussion

Best kernel families: The barplots in Figure 7 suggest that there is no clear-cut covariance structure that fits mortality patterns. Consequently, the models often propose a combination (usually a product, sometimes a sum) of various kernels. Moreover, there is no one-size-fits-all solution as far as different populations are concerned. For instance, the Age effect is typically modeled via a M12 kernel for US Males, a M32 kernel for JPN Males, and a Chy (or RBF or M52) kernel for SWE Females. Additionally, Chy may be identified as a possible Period term for US and JPN Males but never for SWE Females. As such, it is recommended to select different kernels for different case studies. This represents one of the significant differences compared to the classical APC framework, where the SVD decomposition is invariant across datasets, and researchers must manually test numerous combinations, as illustrated in Cairns *et al.* (2011).

Necessity of Cohort Effect: To assess the impact of including a Birth Cohort term, we re-run the GA while excluding all cohort-specific kernels. This is a straightforward adjustment to the implementation and can be used to test whether cohort effects could be adequately explained through a well-chosen Age–Period kernel combination.

We first evaluate our models by comparing the Bayesian information criterion (BIC) of the top kernel that excludes Cohort terms with that of the full $\mathcal{K}_f$. Additionally, we examine the residuals heatmap to detect any discernible diagonal patterns. Our findings indicate that the Cohort effect is overwhelmingly needed for US Males, JPN Females, and JPN Males. The absence of a Cohort term results in a significant increase in BIC, with a difference of 235 for JPN Male, 198 for US Male, and 111 for JPN Female. To put things in perspective, a BIC difference is considered significant only when it surpasses 6.802. The results are confirmed by Figure 8, which illustrates pronounced diagonals in the bottom row (the no-cohort models), contradicting the assumption of independent residuals.

For Sweden, the difference is only 1.24, which corresponds to a BF of 0.2894, indicating no significance. Moreover, the corresponding no-cohort residuals still appear satisfactory, and the associated kernel $0.1125 \cdot \text{Chy}_a(28.9) \cdot \text{Meh}_y(0.653) \cdot \text{M12}_y(1094)$ is ranked ninth-best in the original $\mathcal{K}_f$. Once again, our results are consistent with the literature, see Murphy (2010) who discusses the lack of clarity on cohort effect for Sweden.

Cairns *et al.* (2011) suggested that cohort effects might be partially or completely explained by well-chosen age and period effects. We find a partial confirmation of this finding in that in many populations there are more than two kernels used to explain the Period and Cohort dependence, and there is a clear

substitution between them. However, this is a second-order effect; the primary necessity of including Cohort is unambiguous except for Sweden. For Sweden females, the need for Birth Cohort dependence is quite weak.

Kernel Substitution: Through its mutation operations, the GA naturally highlights substitution effects among different kernel families. Substitution of one kernel with another is intrinsic to the evolution of the GA, and by ranking the kernels in terms of their BIC, we observe the presence of many compositions that achieve nearly same performance and differ just by one term. The above effect is especially noticeable in purely multiplicative kernels that are prevalent in all populations except the USA. In this case, we can frequently observe that one of the terms can be represented by two or three kernel families, with the rest of the terms staying fixed.

We observed that certain kernels are commonly used as substitutes for each other, such as the RBF and M52 kernels, as well as the M12 and Min kernels. Although the latter pair differs by stationarity, the sample paths generated by Min are visually indistinguishable from those generated by M12 when the lengthscale parameter ℓ is large (usually $\ell_{\text{Len}} > 1000$). These substitution patterns are in agreement with our synthetic results, as discussed in Section 4. As an example, when training on the Japan Female dataset and searching in $\mathcal{K}_r$, the top kernel is of the form $\text{M52}_a \cdot (\text{RBF}_y \cdot \text{M12}_y) \cdot \text{M12}_c$, while the second-best according to BIC is $\text{M52}_a \cdot (\text{M52}_y \cdot \text{M12}_y) \cdot \text{M12}_c$. This preserves the same macro-structure while replacing one of the two Period terms, cf. Table 10. Additionally, the next two ranked kernels are very similar, but substitute M12 with Min: the third-best is $\text{M52}_a \cdot (\text{RBF}_y \cdot \text{Min}_y) \cdot \text{M12}_c$, and the fourth-best is $\text{M52}_a \cdot (\text{M52}_y \cdot \text{Min}_y) \cdot \text{M12}_c$. It is important to note that during substitution, the lengthscales (and sometimes process variances) change, as these have a different meaning for different kernel families. For example, the RBF lengthscales tend to be about 50% smaller than those for its substitute, M52.

The Cauchy kernel and M52 are also interchangeable, although Chy and RBF are less so. This effect of multiple substitutes for smooth kernels is nicely illustrated for SWE Females, where the four top kernels fix the Period and Cohort effects according to $\text{M12}_y \cdot \text{Meh}_c$ and then propose any of $\text{Chy}_a, \text{RBF}_a, \text{Meh}_a, \text{M52}_a$ for the Age effect. Substitutions are more common within $\mathcal{K}_f$, since the availability of more kernel families contributes to "collinearity" and hence more opportunities for substitution. At the same time, we occasionally observe the ability to find a more suitable kernel family in $\mathcal{K}_f$. For example, often we observe both a rough and a smooth kernel in Year, indicating that neither M12 nor RBF fit well on their own; in that case M32 sometimes appears to be a better single substitute. Similarly, Meh_c is the most common choice for SWE Females and is replaced with RBF_c or $\text{RBF}_c \cdot \text{M12}_c$ in $\mathcal{K}_r$. Consequently, proposed kernels from $\mathcal{K}_f$ tend to be a bit shorter on average than those from $\mathcal{K}_r$.

A further substitution effect happens between Period and Cohort terms. In JPN Females and SWE Females, we tend to observe a total of three terms, and there is a substitution between using two Period and one Cohort or one Period and two Cohort terms. For instance, in JPN female among top-10 kernels we find both $\text{M52}_a \cdot \text{M12}_y \cdot (\text{Meh}_c\text{M12}_c)$ and $\text{M52}_a \cdot (\text{Meh}_y\text{M12}_y) \cdot \text{M12}_c$.

Additivity and Nonstationarity: Returning to the topic of additive components, our results generally suggest that the additive structure is generally weak. Specifically, introducing an additional additive component often provides only a minor improvement in goodness of fit, which is offset by the complexity penalty in BIC, resulting in lower BICs for additive kernels. Hence, additive kernels tend to be rejected by the GA on the grounds of parsimony. As a result, most kernels with four terms (length 7) and the majority with five terms are purely multiplicative.

Summarizing nonstationarity is a challenging task due to various factors, so any table providing the percentage of nonstationary data should be viewed with reservation. A more comprehensive analysis of nonstationarity can be seen through the frequency diagrams shown in Figures 2 and 7. It is important to note that M12 lengthscales tend to be large in fitting. When M12 has a large lengthscale, the resulting processes are visually indistinguishable from those generated by the nonstationary Min kernel. Therefore, in our data, the presence of M12 indicates potential nonstationarity. From this perspective, our findings suggest that all populations exhibit a (potentially) nonstationary period effect. The other nonstationary kernels (Lin, Meh) are rare but do occur, mostly in the SWE female population.

6. Conclusion

Our work analyzes the use of a GA to discover kernels (i.e., covariance structures) for GP surrogates of mortality surfaces. The GA performed excellently in our synthetic experiments, indicating its promising role as a tool for model selection and validation when using the GP framework for realistic data analysis. In particular, it successfully detected the smoothness of the data generating process, demonstrated robustness across samples (SYA), distinguished additive versus multiplicative APC structures, identified relatively small cohort effects (SYB), and found the correct number of base kernels and identified multiple nonstationarities over Period and Cohort (SYC) coordinates. Additionally, all experiments illustrate the "substitution" effect, where one kernel approximates the impact of another. For instance, M12 kernel with large ℓ_{len} can substitute for Min and vice versa.

When applied to the HMD datasets, our results strongly suggest that best fits to mortality data are provided by GP models that include a rough (non-differentiable or only once-differentiable) component in Year and Cohort, and smooth terms in Age. This matches the classical assumption that the Age-structure of mortality is a smooth function, while the temporal dynamics are random-walk-like. The only exception is the US data, where Age structure is proposed to be non-differentiable, while the Cohort term is smooth.

Among non-standard kernels, we find that Cauchy kernels are often picked, with Chy_a showing up in 34% of SWE Female and 28% of JPN Female top-100, and Chy_y in 26% of JPN Male. Mehler kernels also appear, though infrequently except for Meh_c in SWE females (35 out of top 100 kernels).

Historical data analysis and the SYC experiment both revealed a lack of clarity on determining one single covariance structure. This is unsurprising, given the presence of surrogate kernels that mimic one another (e.g. Chy_a instead of M52_a, or Meh_c instead of RBF_c) and the complexity of the search problem when using the larger search set $\mathcal{K}_f$ (twice as many kernels). Although this benefits BIC optimization by better approximating the truth, finding a precise and expressive covariance structure requires a smaller $\mathcal{K}$ that includes key families that express the modeler's prior beliefs about the underlying data process. This is a challenging knowledge to have as it requires expertise in the application area (e.g. mortality modeling) and properties of GP kernels.

In this article, we emphasized qualitative examination of the kernels of the best-fitting models. In parallel, the GA output also supports model averaging analysis (see Section 3.1). Model averaging not only provides insights into model risk but also robustifies predictions, making it particularly advantageous for actuarial applications such as capital requirements for annuities and pension funds where distributional analysis like tail value-at-risk is essential.

Additional avenues for future work include transferring our approach to smaller populations (under 10–20 million). In contrast to our data where the signal to noise ratio was high and hence the GA had a favorable setting to infer the best latent structure, this is likely to require a significant adjustment since with higher noise the GA will have a harder time differentiating candidate kernels. Similarly, exploring subpopulations, such as insured individuals, can provide insights into the differences in dynamics compared to the general population. This analysis can help uncover distinct patterns and factors that influence mortality within these specific groups. Furthermore, breaking down heterogeneous populations, like the US Males, into subpopulations might be beneficial. With this approach, one aims to identify more concise and interpretable kernels that capture the unique characteristics and dynamics of each subpopulation. By considering variations within the population, we can gain a deeper understanding of localized mortality trends. A suitable framework could be to jointly model multiple populations using multi-output GPs (Huynh and Ludkovski, 2021a,b).

Instead of employing the BIC criterion, one could consider other criteria, in particular those geared toward forecasting performance. GPs support exact predictive mean and probability density formulas for "leave-out-one" (LOO) analysis (Williams and Rasmussen, 2006, Section 5.3). Thus, one could investigate a GA based on ranking kernels by their LOO cross-validation score. Our entire approach was geared to in-sample goodness-of-fit analysis. It is left to future research to investigate predictive performance of the proposed GP compositional kernels. Compared to retrospective assessment, it is

possible that simpler models outperform more complex ones; in other words the selected BIC-driven complexity penalty might not carry to discriminating for best predictive accuracy. The related question of overfitting to the training set to the detriment of out-of-sample performance would require revisiting the impact of $\mathcal{D}$ on the kernels picked by the GA. Proper assessment of out-of-sample projections would require the use of sliding training windows, as well as consideration of probabilistic metrics to test not just GP posterior mean but also its posterior (co)variance, for example, proper coverage of its posterior credible bands.

There are also several extensions to our kernel search. Although `add` and `mul` encompass a variety of possibilities with interpretable results, several alternatives exist. A starting point is the fact that for any real-valued function $g(\cdot)$, we have that $k(\mathbf{x}, \mathbf{x}') = g(\mathbf{x})g(\mathbf{x}')$ defines a kernel. This can be combined with multiplication to provide nonstationary modifications, that is, $g(\mathbf{x})k_S(\mathbf{x} - \mathbf{x}')g(\mathbf{x}')$. Another possibility is *warping*: $k(\mathbf{x}, \mathbf{x}') = k_0(\psi(\mathbf{x}), \psi(\mathbf{x}'))$ being a valid kernel for any function ψ and kernel k_0. See Genton (2001) for a thorough discussion of potential transformations.

Additionally, further kernels can be considered. One example is that the Cauchy kernel is a special case of the rational quadratic kernel indexed by α (see Appendix A). Given the popularity of Chy in our results, it may be worthwhile to explore α values beyond $\alpha = 1$, or consider a direct search over α. It is also possible to incorporate a Cohort effect into non-separable kernels over Age and Year dimensions, where $x_{yr} - x_{ag}$ can naturally appear in expressions such as $\exp(-[x_{ag}, x_{yr}]^\top A[x_{ag}, x_{yr}])$ to define a kernel when A is positive definite.

Changepoint detection can be naturally implemented with GPs, utilizing as kernel $\sigma(x)k_1(x, x')\sigma(x') + \overline{\sigma}(x)k_2(x, x')\overline{\sigma}(x')$, where $\overline{\sigma}(x) = 1 - \sigma(x)$ and $\sigma(\cdot)$ is an activation function, like the sigmoid $\sigma(x) = 1/(1 + \exp(-x))$. This is useful when there is a nonstationary shift from one mortality structure to another, that is, the transitioning from younger to older ages, or in the presence of a temporal mortality shift. Rather than prescribing a hard cutoff, one could design kernels that automatically explore that possibility.

Lastly, more work is warranted to understand the limitations of the GA as currently built. Additional analysis of synthetic experiments can help clarify the impact of the signal-to-noise ratio and the size of the dataset on the ability of the GA to appropriately explore and find the best-fitting kernels. Similarly, we leave to a future study the analysis of whether it is always better to "throw in the kitchen sink", as far as including as many diverse kernels as feasible, or whether a pre-selection could be beneficial.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2023.39

References

Adler, R.J. (2010) *The Geometry of Random Fields*. SIAM.

Ahmadi, S.S. and Gaillardetz, P. (2014) Two factor stochastic mortality modeling with generalized hyperbolic distribution. *Journal of Data Science*, **12**, 1–18.

Azman, S. and Pathmanathan, D. (2022) The GLM framework of the Lee–Carter model: A multi-country study. *Journal of Applied Statistics*, **49**(3), 752–763.

Berlinet, A. and Thomas-Agnan, C. (2011) *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York: Springer Science & Business Media.

Brouhns, N., Denuit, M. and Vermunt, J.K. (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**(3), 373–393.

Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A. and Balevich, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**(1), 1–35.

Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D. and Khalaf-Allah, M. (2011) Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, **48**(3), 355–367.

Cole, D.A., Gramacy, R.B. and Ludkovski, M. (2022) Large-scale local surrogate modeling of stochastic simulation experiments. *Computational Statistics & Data Analysis*, 107537.

Dittrich, D., Leenders, R.T.A. and Mulder, J. (2019) Network autocorrelation modeling: A Bayes factor approach for testing (multiple) precise and interval hypotheses. *Sociological Methods & Research*, **48**(3), 642–676.

Dowd, K., Cairns, A.J. and Blake, D. (2020) CBDX: A workhorse mortality model from the Cairns–Blake–Dowd family. *Annals of Actuarial Science*, **14**(2), 445–460.
Duvenaud, D. (2014) *Automatic model construction with Gaussian processes*. Ph.D. Thesis, University of Cambridge.
Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J. and Zoubin, G. (2013) Structure discovery in nonparametric regression through compositional kernel search. *International Conference on Machine Learning*, pp. 1166–1174. PMLR.
Gardner, J., Pleiss, G., Weinberger, K.Q., Bindel, D. and Wilson, A.G. (2018) GpyTorch: Blackbox matrix-matrix Gaussian Process inference with GPU acceleration. *Advances in Neural Information Processing Systems, 31*.
Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.
Genton, M.G. (2001) Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, **2**, 299–312.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Monographs on Statistics & Applied Probability. Boca Raton, FL: Chapman and Hall/CRC.
Hunt, A. and Blake, D. (2014) A general procedure for constructing mortality models. *North American Actuarial Journal*, **18**(1), 116–138.
Huynh, N. and Ludkovski, M. (2021a) Joint models for cause-of-death mortality in multiple populations. arXiv preprint arXiv:2111.06631.
Huynh, N. and Ludkovski, M. (2021b) Multi-output Gaussian processes for multi-population longevity modelling. *Annals of Actuarial Science*, **15**(2), 318–345.
Jähnichen, P., Wenzel, F., Kloft, M. and Mandt, S. (2018) Scalable generalized dynamic topic models. *International Conference on Artificial Intelligence and Statistics*, pp. 1427–1435. PMLR.
Jeffreys, H. (1961) *The Theory of Probability*. Oxford: Oxford University Press.
Jin, S.-S. (2020) Compositional kernel learning using tree-based genetic programming for Gaussian process regression. *Structural and Multidisciplinary Optimization*, **62**(3), 1313–1351.
Kanagawa, M., Hennig, P., Sejdinovic, D. and Sriperumbudur, B.K. (2018) Gaussian processes and kernel methods: A review on connections and equivalences. arXiv preprint arXiv:1807.02582.
Kingma, D.P. and Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Lee, M.D. and Wagenmakers, E.-J. (2014) *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.
Lee, R. (2000) The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, **4**(1), 80–91.
Ludkovski, M., Risk, J. and Zail, H. (2018) Gaussian process models for mortality rates and improvement factors. *ASTIN Bulletin: The Journal of the IAA*, **48**(3), 1307–1347.
Luke, S. and Panait, L. (2006) A comparison of bloat control methods for genetic programming. *Evolutionary Computation*, **14**(3), 309–344.
Mehler, F.G. (1866) Ueber die entwicklung einer function von beliebig vielen variablen nach laplaceschen functionen höherer ordnung.
Murphy, M. (2010) Reexamining the dominance of birth cohort effects on mortality. *Population and Development Review*, **36**(2), 365–390.
Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S. and Perla, F. (2019) A deep learning integrated Lee–Carter model. *Risks*, **7**(1), 33.
Noack, M.M. and Sethian, J.A. (2021) Advanced stationary and non-stationary kernel designs for domain-aware Gaussian processes. arXiv preprint arXiv:2102.03432.
Parzen, E. (1961) An approach to time series analysis. *The Annals of Mathematical Statistics*, **32**(4), 951–989.
Perla, F., Richman, R., Scognamiglio, S. and Wüthrich, M.V. (2021) Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, **2021**(7), 572–598.
Poli, R., Langdon, W.B., McPhee, N.F. and Koza, J.R. (2008) *A Field Guide to Genetic Programming*. Springer.
Renshaw, A.E. and Haberman, S. (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**(3), 556–570.
Richman, R. and Wüthrich, M.V. (2021) A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, **15**(2), 346–366.
Roman, I., Santana, R., Mendiburu, A. and Lozano, J.A. (2021) Evolving Gaussian process kernels from elementary mathematical expressions for time series extrapolation. *Neurocomputing*, **462**, 426–439.
Roustant, O., Ginsbourger, D. and Deville, Y. (2012) Dicekriging, diceoptim:n Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, **51**, 1–55.
Schölkopf, B., Smola, A.J. and Bach, F. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.
Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.
Sipper, M., Fu, W., Ahuja, K. and Moore, J.H. (2018) Investigating the parameter space of evolutionary algorithms. *BioData Mining*, **11**(1), 1–14.
Villegas, A., Kaishev, V.K. and Millossovich, P. (2015) StMoMo: An R package for stochastic mortality modelling. *7th Australasian Actuarial Education and Research Symposium*.
Wang, C.-W., Huang, H.-C. and Liu, I.-C. (2011) A quantitative comparison of the Lee-Carter model under different types of non-Gaussian innovations. *The Geneva Papers on Risk and Insurance-Issues and Practice*, **36**, 675–696.
Willets, R.C. (2004) The cohort effect: Insights and explanations. *British Actuarial Journal*, **10**(4), 833–877.

Williams, C.K. and Rasmussen, C.E. (2006) *Gaussian Processes for Machine Learning*, Vol. 2. Cambridge, MA: MIT Press.
Yaglom, A.M. (1957) Some classes of random fields in n-dimensional space, related to stationary random processes. *Theory of Probability & Its Applications*, **2**(3), 273–320.

Appendix A. Notes on Kernels

Unless otherwise stated, the following assumes $x, x' \in \mathbb{R}$. In this section, we use the fact that for a GP f, its derivative f' exists (in the mean-square sense) if and only if $\frac{\partial^2 k}{\partial x \partial x'}(x, x')$ exists (Adler 2010).

Stationary

Matérn-1/2: is the covariance of an Ornstein–Uhlenbeck (OU) process (Berlinet and Thomas-Agnan, 2011). The OU process follows a linear mean-reverting stochastic differential equation; it has continuous, nowhere differentiable paths. The mean-reversion localizes dependence and has been advocated (Jähnichen *et al*., 2018) for capturing small-scale effects. The lengthscale ℓ_{len} controls the rate of mean-reversion (lower values revert more quickly).
The re-parametrization

$$k(x, x') = \exp\left(-\frac{-|x - x'|}{\ell_{\text{len}}}\right) = \phi^{|x-x'|}, \qquad \phi = \exp\left(-1/\ell_{\text{len}}\right), \tag{A.1}$$

shows that when x is discrete, M12 is equivalent to an AR(1) process with persistence parameter ϕ. In particular, ℓ_{len} large (i.e. $\phi \simeq 1$) mimics the nonstationary random walk process and its Min kernel, allowing sample paths to deviate far from their mean and weakening stationarity.

AR2: The covariance kernel associated with a (continuous x) second-order autoregressive (AR(2)) process is Parzen (1961)

$$k(x, x'; \alpha, \gamma) = \frac{\exp\left(-\alpha|x - x'|\right)}{4\alpha\gamma^2}\left\{\cos\left(\omega|x - x'|\right) + \frac{\alpha}{\omega}\sin\left(\omega|x - x'|\right)\right\},$$

where $\omega^2 = \gamma^2 - \alpha^2 > 0$. Notably, this kernel has two parameters and, thus, a higher BIC penalty during the GA optimization. One can show that $\frac{\partial^2 k}{\partial x \partial x'}(x, x')$ exists for all $x, x' \in \mathbb{R}$, but $\frac{\partial^4 k}{\partial^2 x \partial^2 x'}(x, x')$ does not. Thus, a GP with AR2 kernel is (mean-square) once but not twice differentiable, that is, $f \in C^1$.
For consistency with the Matérn family of kernels, we reparameterize according to $\alpha = 1/\ell_{\text{len}}$, $\omega = \pi/p$ and normalize for $k(x, x; \ell_{\text{len}}, p) = 1$, so that

$$k\big(x, x'; \ell_{\text{len}}, p\big) = \exp\left(-\frac{|x - x'|}{\ell_{\text{len}}}\right)\left\{\cos\left(\frac{\pi}{p}|x - x'|\right) + \frac{p}{\pi\ell_{\text{len}}}\sin\left(\frac{\pi}{p}|x - x'|\right)\right\}, \tag{A.2}$$

where, under the re-parametrization, $\gamma^2 = \omega^2 + \alpha^2 = \frac{1}{\ell_{\text{len}}^2} + \frac{\pi^2}{p^2}$. Through trigonometric identities, one can see that this is the same covariance function as a stationary discrete-time AR(2) process (written as $f(x) = \phi_1 f(x-1) + \phi_2 f(x-2) + \epsilon(x)$) in the case of complex characteristic roots, that is, $\phi_1^2 + 4\phi_2 < 0$, with parameters related by

$$\phi_2 = -\exp\left(-2/\ell_{\text{len}}\right), \qquad \phi_1 = 2\cos\left(\pi/p\right)\sqrt{-\phi_2}. \tag{A.3}$$

Cauchy: This kernel is fat-tailed and has long-range memory, which means that correlations decay not exponentially but polynomially, leading to a long-range influence between inputs (Jähnichen *et al*., 2018). The Chy kernel function is given by

$$k(x, x'; \ell_{\text{len}}) = \frac{1}{1 + \frac{(x-x')^2}{\ell_{\text{len}}^2}} \tag{A.4}$$

which is a special case, $\alpha = 1$, of the rational quadratic kernel $k(x, x'; \alpha, \ell_{\text{len}})$:$n=\left(1+\frac{(x-x')^2}{\alpha \ell_{\text{len}}^2}\right)^{-\alpha}$.

Since $k(x, x')$ in (A.4) is infinitely differentiable in both arguments, the associated GP is also in C^∞ like RBF. One way to interpret Chy is as a marginalized version of the RBF kernel with an exponential prior on $1/\ell_{\text{RBF}}^2$ (with rate $\ell_{\text{Chy}}^2/2$):

$$\int_0^\infty \exp\left(-u \cdot \frac{(x-x')^2}{2}\right) \cdot \ell_{\text{Chy}}^2 \exp\left(-\ell_{\text{Chy}}^2 \cdot u\right) du = \frac{1}{1+\frac{(x-x')^2}{\ell_{\text{Chy}}^2}}$$

Nonstationary

Linear: the linear kernel connects Gaussian processes to Bayesian linear regression. In particular, if $f(x) = \beta_0 + \beta_1 x$ where $x \in \mathbb{R}$ and there are priors $\beta_0 \sim N\left(0, \sigma_0^2\right)$, $\beta_1 \sim N(0, 1)$, then $f \sim \mathcal{GP}(0, k_{\text{Lin}})$, that is, $k(x, x') = \sigma_0^2 + x \cdot x'$. Note that k_{Lin} can be scaled to yield a prior variance on β_1.

Mehler: The nonstationary (Mehler, 1866) kernel is

$$\begin{aligned} k(x, x'; \rho) &= \exp\left(-\frac{\rho^2(x^2+x'^2) - 2\rho x x'}{2(1-\rho^2)}\right), \qquad -1 \le \rho \le 1 \\ &= k_{RBF}(x, x'; \ell_{\text{len}}) \cdot \exp\left(\frac{\rho}{\rho+1} x x'\right) \end{aligned} \tag{A.5}$$

where $\ell_{\text{len}}^2 = \frac{1-\rho^2}{\rho^2}$. Reported hyperparameters in the text are ρ (not ℓ_{len}). By the above decomposition, we can interpret the Mehler kernel as another C^∞ kernel that provides a nonstationary scaling to RBF. Initial experiments always found $\rho > 0$, which causes an increase in covariance for larger values of x and x'. One way to see the effect of the nonstationary component is through variance and correlation. If $f \sim \mathcal{GP}(m, k_{\text{Meh}})$, then $\text{var}(f(x)) = k_{\text{Meh}}(x, x) = \exp\left(\frac{\rho}{\rho+1} \cdot x^2\right)$ which illustrates an increase in process variance as x increases. Remarkably, this results in a stationary *correlation function* $\text{corr}(f(x), f(x')) = \exp\left(-\frac{\rho}{2(1-\rho^2)}(x-y)^2\right)$. Thus, Mehler is appropriate when one desires RBF dynamics combined with increasing process variance.

Remark 6. The Mehler kernel is a valid kernel function for $\rho > 0$ in the sense that it is positive definite, as it admits the basis expansion $k(x, x') = \frac{1}{\sqrt{1-\rho^2}} \sum_{k=0}^\infty \frac{\rho^k}{k!} H_{e_k}(x) H_{e_k}(x')$ and hence for all $n \in \mathbb{N}$, $\mathbf{a} \in \mathbb{R}^n$ and $x_1, \ldots, x_n \in \mathbb{R}$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x, x') = \frac{1}{\sqrt{1-\rho^2}} \sum_{k=0}^\infty \rho^k \left(\sum_{i=1}^n a_i h_k(x_i)\right)^2 \ge 0,$$

where $H_{e_k} = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{x^2/2}$ is the kth probabilist's Hermite polynomial.

Appendix B. More on GP Hyperparameter Convergence

The mutation operations of our genetic algorithm depend crucially on the relative comparison of the log-likelihoods of the given dataset across different kernels k's given its direct role in computing BIC. Hence, an accurate value of the maximal likelihood for a given kernel is a pre-requisite to identify which kernels are fitter than others and hence explore accordingly. Computing the likelihood is equivalent to inferring the MLE for the kernel hyperparameters and is known to be a challenging optimization task. In our implementation, this optimization is done via stochastic gradient descent (SGD) through Adam (Kingma and Ba, 2014), up to a given number $\eta_{\max}$ of iterations or until a pre-set tolerance threshold is reached.

Table B.1 *Number of training steps η_ε needed for the likelihood $l_{K_0}^{(\eta)}\left(\hat{\theta}|\mathbf{y}\right)$ to be within ε of l_{min} when using respective K_0, across the synthetic case studies. $BIC\left(\hat{K}_0^{(\eta_\varepsilon)}\right)$ corresponds to $l_{K_0}^{(\eta)}\left(\hat{\theta}|\mathbf{y}\right)$.*

	ε	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
SYA-1	η_ε	6	62	89	90	90
	$\mathrm{BIC}\left(\hat{K}_0^{(\eta_\varepsilon)}\right)$	–2066.14	–2066.67	–2066.76	–2066.77	–2066.77
SYA-2	η_ε	39	57	143	179	3241
	$\mathrm{BIC}\left(\hat{K}_0^{(\eta_\varepsilon)}\right)$	–2033.23	–2034.17	–2034.22	–2034.23	–2034.23
SYB	η_ε	165	239	603	835	997
	$\mathrm{BIC}\left(\hat{K}_0^{(\eta_\varepsilon)}\right)$	–2467.19	–2467.96	–2468.06	–2468.07	–2468.07
SYC	η_ε	128	206	283	353	415
	$\mathrm{BIC}\left(\hat{K}_0^{(\eta_\varepsilon)}\right)$	–2721.86	–2722.78	–2722.88	–2722.89	–2722.89

In this section, we present additional evidence on how fast this convergence occurs in our synthetic experiments. Namely, we evaluate GP hyperparameter convergence during training by fitting each of the true kernels K_0 from initialization for SYA, SYB, and SYC, indexing the intermediate log marginal likelihoods after η steps as $l_k^{(\eta)}\left(\hat{\theta}|\mathbf{y}\right)$. Given a gold-standard $l_{\max} = \max_{1\leq\eta} l_k^{(\eta)}\left(\hat{\theta}|\mathbf{y}\right)$, we record the number of training steps needed to achieve a log marginal likelihood within $\varepsilon \in \{10^{-3}, 10^{-4}, \ldots, 10^{-7}\}$ of $l_{\min}$:

$$\eta_\varepsilon = \min\left\{\eta : |l_k^{(\eta)}\left(\hat{\theta}|\mathbf{y}\right) - l_{\max}| \leq \varepsilon\right\}. \tag{A.6}$$

We present the results in Table B.1. SYA-1 and -2 show inconsistency for $\varepsilon = 10^{-7}$ probably because the SGD optimization hyperparameters were calibrated to the $\varepsilon = 10^{-6}$ case. SYB takes four times as many training steps as SYA for $\varepsilon = 10^{-4}$, and ten times as many for $\varepsilon = 10^{-6}$. SYC converges quickly, matching SYB up to $\varepsilon = 10^{-4}$ and maintaining rapid convergence rates for $\varepsilon = 10^{-5}$, 10^{-6}, and 10^{-7}. Despite incorporating four base kernels, including two Period components and heteroskedastic noise, SYC converges at a comparable rate to SYA and more swiftly than SYB for lower ε values, underscoring its efficiency in reaching $l_{\max}$. Consequently, for computational tractability, our experiments have restricted $\eta_{\max} = 300$ (running time is linear in $\eta_{\max}$), with the result that we achieve a tolerance of better than 10^{-4}.

Appendix C. Runtime and Computational Complexity Analysis

Computational efficiency of our methods is driven separately by the Gaussian process (GP) and the genetic algorithm (GA) components. The GP's most computationally intensive operation is the inversion of the covariance matrix $\mathbf{K}$ during likelihood evaluation, which is $\mathcal{O}(N^3)$. Each GP fitting process iterates MLE optimization steps η times, up to the maximum $\eta_{\max} = 300$ specified by the user (see Section B).

For the GA, the runtimes are linear in n_g and G, resulting in an overall runtime of $n_g \times G \times \mathcal{O}(N^3)$. However, the GA is inherently parallelizable, allowing for a theoretical time complexity reduction to $G \times \mathcal{O}(N^3)$ with n_g paralleled environments. Empirically, a single GA run with the default parameters takes approximately 8 h for a full kernel set search ($\mathcal{K}_f$) on a home PC with AMD Ryzen 1950X 16-Core 3.40GhZ, 32GB RAM, and NVIDIA GeForce GTX 1080 Ti GPU, somewhat affected by the complexity of the dataset: mortality surfaces with more terms, like the US Males, take longer to fit.

Speed-up options for GP fitting, such as variational and sparse GPs, could reduce the GP runtime complexity to $\mathcal{O}(N \log N)$ or $\mathcal{O}(N)$. These methods however cannot use BIC for model evaluation and inherently carry a loss of accuracy due to information compression (although, empirically this is often minor). In theory, these could be employed during the GA selection phase, followed by exact methods for final model evaluation in the last generation, achieving speed-up factors of 1–2 orders of magnitude.