arXiv:2305.01728v1 [stat.ML] 2 May 2023

# 表达 通过高斯过程核构建死亡率模型

Jimmy Risk Mike Ludkovski 数学与统计学 统计学与应用概率 加州州立理工大学波莫纳分校 加利福尼亚大学 波莫纳 CA 91676 SantaBarbara, CA 93106-3110
jrisk@cpp.edu ludkovski@pstat.ucsb.edu

2023年5月4日

**摘要**

我们构建了一个灵活的高斯过程（GP）框架，用于学习年龄特定和年份特定死亡率曲面的协方差结构。利用GP核的加法和乘性结构，我们设计了一种遗传编程算法，为给定人口群体搜索最具表达力的核。我们的组合搜索基于年龄-时期-队列（APC）范式，构建出最符合某个死亡率数据集时空动态的协方差先验。我们将所得遗传算法（GA）应用于合成案例研究，以验证GA恢复APC结构的能力；同时，还将其应用于人类死亡率数据库中的真实国家级数据集。基于机器学习的分析为不同人口群体中队列效应的存在或缺失，以及死亡率曲面沿年龄和年份维度的相对平滑性，提供了新的见解。我们的建模工作使用Python中的PyTorch库完成，并深入研究了如何利用GA辅助GP代理模型的组合核搜索。

关键词：高斯过程核工程、遗传算法、死亡率曲面、人类死亡率数据库，

## 1 引言

一个多世纪以来，将人口死亡率建模为年龄和日历时间的函数一直是活跃的研究领域。1992 年，Lee-Carter模型 [26] 引入了随机的 时期效应 来描述死亡率随时间的演变，由此推动了针对这一主题的统计和概率方法大量涌现。年龄-时期结构被证明不足以满足许多应用的需求，进而促成了基于队列的扩展 [36] 的发展；在这些扩展中，死亡率还根据出生年份作进一步区分。这些年龄-时期-队列（APC）模型 [17, 6] 已成为分析死亡率趋势的重要工具，同时能够分离各自三个主要预测变量的影响。

高斯过程（GP）模型 [27, 19, 18]为 APC 建模中的长寿分析提供了一种非参数时空范式。这种方法与现有的 APC 模型以及较新的深度学习驱动方法 [31, 34, 37]并行。其底层预测属于空间平滑器 类别，与平滑样条 [15]相似。GP 的主要优势包括灵活性、不确定性量化以及进行多人口群体分析的能力。此外，通过其协方差核，GP 能够直接呈现年龄-年份特定死亡率之间的相互依赖关系，从而使建模者能够专注于捕捉相应的

协方差结构。高斯过程的协方差核决定了其分布的性质，包括后验均值函数、平滑性等。这为理解目标过程的底层动态提供了有价值的洞见，而神经网络等黑盒方法无法实现这一点。

与将二维年龄-年份死亡率服务的年龄-时期-队列分解为三个单变量方向相对应，我们可以考虑能够反映死亡率年龄结构、其随时间的演变以及队列效应的核。在现有的高斯过程死亡率文献中，这通常直接转化为一个可分离的高斯过程核——由年龄上的单变量核、年份上的单变量核以及（如有需要）出生队列上的单变量核相乘而成。虽然这种选择的表现令人满意，但限制颇多，也阻碍了高斯过程发现数据驱动依赖关系的能力。基于这一动机，我们探索用于死亡率模型的高斯过程核组合与发现。因此，本文的首要目标是实现一个自动化流程，以找到最适合死亡率分析的协方差结构。其一项动机在于，不同的（亚）群体具有不同的APC结构，因此一刀切的方法并不合适。为此，我们设计了一种遗传编程算法，在核空间中迭代搜索，以找到拟合度最高的核。我们的方法在先前通过遗传算法（GA）进行核发现的方案基础上，针对长寿分析进行了调整，并引入了专门的变异操作。在此过程中，我们通过考察合成死亡率曲面，评估遗传算法恢复已知APC结构的能力，包括识别队列效应以及加法效应与乘性效应的能力。

我们的第二个动机，是将死亡率建模文献中的思想与不同高斯过程核族的结构联系起来。我们研究了多种核，大幅扩展了迄今用于死亡率建模的有限核种类（如平方指数核和 Mat´ern）。通过引入并检验新的高斯过程核族，我们消除了必须直接预先指定所用核族的限制，因为这种做法会对数据施加隐含的约束和假设。此外，我们还引入了其他核，以专门表示更丰富的结构，包括随机游走、周期性以及更一般的 ARIMA过程。我们发现，现有的年龄-时期-队列协方差结构，包括知名的 Lee-Carter模型 [26] 和 Cairns-Blake-Dowd模型 [6] ，都可以通过加性高斯过程核精确匹配。通过检验各种核，可以为死亡率曲面找到更好的拟合，并回答相关问题，例如队列效应的强度或结构。

我们实现上述目标的核心方法是组合核搜索。核组合利用了核在加法和乘法运算下保持封闭这一性质。因此，组合核能够丰富且细致地描述潜在的死亡率动态。我们的分析主要依靠遗传算法（GA），该算法运用代的概念，通过变异-选择机制逐步发现越来越好的核。给定一个死亡率数据集，下文介绍的遗传算法会生成大量（数千个）候选核。随后，这些核会依次拟合该数据集，并根据统计适应度函数进行排名。接着，遗传算法以概率方式推动对适应度最高的核进行探索，并淘汰适应度较低的核。这一过程能够自动探索表现最佳的死亡率高斯过程模型。早期关于高斯过程组合核的研究采用前向搜索，通过最小化贝叶斯信息准则（BIC）来构建基于树的核表示 [11, 10]。Jin [22] 和 Roman等人 [38] 在此基础上采用遗传算法；前者分析了算法在多个多维合成测试函数上的表现，后者则研究了单变量时间序列。我们针对死亡率建模扩展并调整了这些策略。

组合式高斯过程搜索框架与死亡率建模天然契合。核可以视为随机过程的协方差函数；在年龄-时期-队列应用中，搜索通过对年龄、时期和队列效应进行加法和乘法来实现。此类组合

及其修改已在前述 Lee-Carter模型和 Cairns-Blake-Dowd模型（CBD）中实现（详见 [6]）。借助遗传算法的输出，我们探讨以下关于死亡率曲面的基本问题，这些问题本身就具有研究价值：

- 是否存在队列效应。我们的方法提供严谨的贝叶斯非参数证据，用于判断纳入出生队列效应是否有益。由于队列效应具有群体特异性，这是一项重要的模型选择问题。

- 年龄与年份协方差结构之间的相对平滑性。经典年龄-时期-队列模型假设日历时间具有随机游走结构，并（隐含地）假设年龄具有平滑（无限可微）结构。相比之下，现有的高斯过程模型假设两个坐标方向都具有固定的平滑性（例如二阶可微）。我们的方法揭示了这些假设是否会影响预测，以及何种平滑性最符合死亡率数据。

- 死亡率协方差中的加法结构与乘性结构。已有许多关于年龄-时期-队列模型的提案和比较分析，以不同方式组合年龄项与年份项。我们为高斯过程模型提供了类似的分析。具体而言，我们的方法能够量化拟合效果最佳的核的复杂度，从而为需要多少个不同项带来新的见解。

本文其余部分安排如下。第 2 节回顾用于死亡率曲面的高斯过程模型，重点介绍核选择这一首要任务，并说明其对模型预测的影响。第 3 节介绍用于组合核搜索的遗传算法工具。第 4 节通过在合成死亡率曲面上恢复已知核，验证遗传算法方法的有效性。第 5 节分析遗传算法在日本女性初始案例研究中的输出。第 5.4 节随后对多个国家层面的数据集进行横截面分析，以探讨队列效应是否存在，以及年龄与日历年的相对平滑性。第 6 节总结全文。

## 2 用于死亡率的高斯过程模型

高斯过程 (GP) 是一组随机变量 $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$，对于任意 $\ell \in \mathbb{N}$ 和 $\{\mathbf{x}_1,\ldots,\mathbf{x}_\ell\} \subseteq \mathbb{R}^d$，向量 $[f(\mathbf{x}_1)\ f(\mathbf{x}_2)\ \ldots f(\mathbf{x}_\ell)]^\top$ 服从多元正态分布[44] (记为 MVN)。对于死亡率曲面，$d = 3$。高斯过程由其均值函数$m : \mathbb{R}^d \to \mathbb{R}$ 和协方差核 $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ [1]唯一确定。核 $k(\cdot,\cdot)$ 必须是对称正定函数。在这种情况下，对于 $\mathbf{x}_i,\mathbf{x}_j \in \mathbb{R}^d$,

$$\mathbb{E}[f(\mathbf{x}_i)] = m(\mathbf{x}_i), \tag{1}$$
$$\operatorname{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i,\mathbf{x}_j), \tag{2}$$

并记为 $f \sim \mathcal{GP}(m,k)$。高斯过程回归模型假设

$$y := y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x}), \tag{3}$$

其中，$f$ 是一个高斯过程，其先验均值为 $m(\cdot)$，协方差为 $k(\cdot,\ \cdot)$；$\epsilon$ 是噪声项，$y$是含噪观测。假设 $\epsilon(\cdot)$是方差为 $\sigma^2(\cdot)$的独立高斯白噪声，根据多元正态随机变量的性质可知，$\{y(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ 是一个高斯过程，其均值函数和协方差函数为

$$\mathbb{E}[y(\mathbf{x})] = m(\mathbf{x}), \qquad k_y(\mathbf{x}_i,\mathbf{x}'_j) = k(\mathbf{x}_i,\mathbf{x}_j) + \sigma^2(\mathbf{x}_i)\delta_{i=j}, \tag{4}$$

其中，$\delta$ 是狄拉克 delta 函数。需要注意的是，当索引 $i$ 和 $j$ 相等时，$\sigma^2(\mathbf{x}_i)\delta_{i=j}$ 非零：同一位置 $\mathbf{x}$ 可能有来自不同样本的两次观测，因此它们不共享噪声。

### 2.1 高斯过程回归

给定数据集 $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$，高斯过程假设和观测似然 (3) 推出 $[\mathbf{f}, \mathbf{y}]^\top \sim \mathcal{MVN}$，其中 $\mathbf{f} = [f(\mathbf{x}_1) \ldots, f(\mathbf{x}_n)]^\top$，$\mathbf{y} = [y(\mathbf{x}_1) \ldots, y(\mathbf{x}_n)]^\top$，因此后验 $\mathbf{f}|\mathbf{y} \sim \mathcal{MVN}$ 也同样成立。更一般地，对于 $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, $[f(\mathbf{x}) \ f(\mathbf{x}'), \mathbf{y}]^\top \sim \mathcal{MVN}$，后验有限维分布完全已知:

$$[f_*(\mathbf{x}), f_*(\mathbf{x}')]^\top := \left([f(\mathbf{x}), f(\mathbf{x}')]^\top | \mathbf{y}\right) \sim \mathcal{MVN}\left([m_*(\mathbf{x}), m_*(\mathbf{x}')]^\top, \begin{bmatrix} k_*(\mathbf{x},\mathbf{x}) & k_*(\mathbf{x},\mathbf{x}') \\ k_*(\mathbf{x}',\mathbf{x}) & k_*(\mathbf{x}',\mathbf{x}') \end{bmatrix}\right), \quad (5)$$

其中，对于任意 $\mathbf{x},\mathbf{x}' \in \mathbb{R}^d$，后验协方差核定义为 $k_*(\mathbf{x},\mathbf{x}') := \mathrm{cov}(f(\mathbf{x}) \ f(\mathbf{x}')|\mathbf{y})$。Kolmogorov 扩张定理保证 $\{f_*(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ 定义了一个高斯过程。此外，后验均值 和 方差 可明确表示为

$$m_*(\mathbf{x}) = \mathbf{K}(\mathbf{x}, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \Delta(\mathbf{X}, \mathbf{X})\right]^{-1} \mathbf{y}, \quad (6)$$

$$\mathbf{K}_*(\mathbf{x}, \mathbf{x}') = \mathbf{K}([\mathbf{x}, \mathbf{x}']^\top, \mathbf{X}) \left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \mathbf{\Delta}(\mathbf{X}, \mathbf{X})\right]^{-1} \mathbf{K}(\mathbf{X}, [\mathbf{x}, \mathbf{x}']^\top), \quad (7)$$

其中 $\mathbf{X}$ 表示行向量为 $\mathbf{x}_i$、$i = 1$、.....、$n$ 的 $n \times d$ 矩阵；对于 $\mathbf{U}$、$\mathbf{V}$ 分别为 $\ell \times d$ 和 $m \times d$ 的情况，$\mathbf{K}(\mathbf{U}, \mathbf{V}) = [k(\mathbf{u}_i, \mathbf{v}_j)]_{1\le i\le \ell,\ 1\le j\le m}$ 表示成对协方差的 $\ell \times m$ 矩阵。$\mathbf{\Delta}(\mathbf{U}, \mathbf{V})$ 的元素为 $\sigma^2(\mathbf{u}_i)\delta_{i=j}\delta_{\mathbf{u}_i=\mathbf{v}_j}$；在我们的情况下，$\mathbf{\Delta}(\mathbf{X}, \mathbf{X})$ 是一个 $n \times n$ 对角矩阵，其元素为 $\sigma^2(\mathbf{x}_i)$。在噪声方差为常数 $\sigma^2$ 的情况下：$= \sigma^2(\mathbf{x}_i)$，$\mathbf{\Delta}(\mathbf{X}, \mathbf{X}) = \sigma^2 I_n$，其中 $I_n$ 是 $n \times n$单位矩阵。

### 2.2 高斯过程核

高斯过程回归将这样一种思想编码其中：相似的输入（依据核来判断）会产生相似的输出。后验均值是观测数据的加权平均，因此可以看出这一点，因为 $m_*(\mathbf{x}) = \mathbf{w}^\top \mathbf{y}$ 成立，其中$\mathbf{w}^\top = \mathbf{K}(\mathbf{x},\mathbf{X}) \ [\mathbf{K}(\mathbf{X},\mathbf{X}) + \Delta(\mathbf{X},\mathbf{X})]^{-1}$。存在多种核，可依据领域知识对相似性进行编码。与协方差矩阵类似，函数 $k(\cdot,\cdot)$ 成为协方差核的唯一要求是它具有对称性和正定性，即对于所有 $n = 1, 2, \ldots,$ 以及 $x_1, \ldots, x_n \in \mathbb{R}^d$，我们必须确保格拉姆矩阵 $\mathbf{K}(\mathbf{X}, \mathbf{X})$ 为半正定。

平稳核 $k(\mathbf{x}_i, \mathbf{x}_j)$ 是指可以写成 $\mathbf{x}_i - \mathbf{x}_j$ 的函数的核，即 $k(\mathbf{x}_i, \mathbf{x}_j) = k_S(\mathbf{x}_i - \mathbf{x}_j)$，因此对于输入空间中的平移具有不变性。如果一个核仅是 $\mathbf{r} = \|\mathbf{x}_i - \mathbf{x}_j\|$ 的函数，则进一步称其为各向同性核，其中 $\|\cdot\|$ 是 $\ell^2$ 欧几里得距离，因此可以写作 $k_I(\mathbf{r}) = k(\mathbf{x}_i, \mathbf{x}_j)$。Yaglom [45] 利用博赫纳定理推导出一种适用于各向同性核的类似傅里叶变换的方法，从而可以根据谱密度推导核函数。通常假设平稳核经过归一化，因为当 $\mathbf{x} = \mathbf{x}'$ 时，$k_S(\mathbf{x} - \mathbf{x}')/k_S(\mathbf{0}) = 1$；这样便可以将平稳核自然地解释为相关函数。最后，关于 $\mathbb{R}^d$ 的可分离核是指可以写成乘积的核：$k(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^d k_j(x^{(j)}, x'^{(j)})$，其中 $x^{(j)}$ 是 $\mathbf{x}$ 的第 $j$ 个坐标。因此，全局核可以分解为各维核函数的乘积，每个维度都有自己的核函数。

正定函数的代数性质使得从已有核构造新核变得十分简单 [39, 40, 3, 14]。主要工具是核在加法下保持不变。

以及乘法运算，即可以通过求和与求积进行组合。因此，如果 $k_1$ 和 $k_2$ 是两个核，且 $c_1$、$c_2$ 是两个正实数，那么 $k(\mathbf{x}, \mathbf{x}') = c_1 k_1(\mathbf{x}, \mathbf{x}') + c_2 k_2(\mathbf{x}, \mathbf{x}')$ 也是如此。这与高斯过程的性质一致：如果 $f_1 \sim \mathcal{GP}(0, k_1)$ 和 $f_2 \sim \mathcal{GP}(0, k_2)$ 是两个相互独立的高斯过程，那么对于 $c_1$、$c_2 > 0$，我们有 $c_1 f_1 + c_2 f_2 \sim \mathcal{GP}(0, c_1 k_1 + c_2 k_2)$。这为其与广义加性模型框架之间建立了联系。尽管对于核的乘积不存在类似的结果（高斯过程的乘积不再是高斯过程，因为多元高斯分布不具有保持性），但当一个核是平稳的，并且随着 $|\mathbf{x} - \mathbf{x}'| \to \infty$ 的增大而单调衰减时，核相乘通常具有一种常见的解释。事实上，如果 $k_1$ 是这样一个核，那么乘积 $k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$ 就表示 $k_2$ 的影响会随着 $k_1(\mathbf{x}, \mathbf{x}')$ 的变化而衰减，其中 $|\mathbf{x} - \mathbf{x}'|$ 逐渐增大。这与可分离性相辅相成。例如，$k(\mathbf{x}, \mathbf{x}') = k_1(|x^{(1)} - x'(1)|)k_2(x^{(2)}, x'(2))$ 表示在第二个坐标上存在一定的相似性，而这种相似性可能随着 $|x^{(1)} - x'(1)|$ 的增大而衰减。加法和乘性性质通常与常数核 $k(\mathbf{x}, \mathbf{x}') = c, c > 0$ 结合使用，从而产生缩放效应（乘法）或衰减效应（加法）。

**备注 1**。还有其他核设计策略，例如，若 $g : \mathbb{R}^d \to \mathbb{R}$，则 $k(\mathbf{x},\mathbf{x}') = g(\mathbf{x})g(\mathbf{x}')$ 定义了一个核；这一性质结合乘法，是处理异方差噪声的一种方法，因为 $k(\mathbf{x}_i,\mathbf{x}'_j) = \delta_{i=j}$ 定义了一个核。更多性质和说明可参见 *Genton* *[14]* 或 *Noack* 和 *Sethian* *[32]* 。

### 2.3 核族

表1列出了我们考虑的核族。为简单起见，我们假设一维基核，其中 $x \in \mathbb{R}$，完整的 $\mathbf{x} \in \mathbb{R}^d$ 结构表示为可分离核。在这 9 个核族中，有些核可产生平滑（$C^2$ 及更高阶）拟合，有些核具有粗糙（不可微）的样本路径，还有若干非平稳核。这里，平滑性指样本路径 $\mathbf{x} \mapsto f(\mathbf{x})$ 具有 $k$ 阶可微性等性质，即 $f(\cdot) \in C^k$。后验均值 $\mathbf{x} \mapsto m_*(\mathbf{x})$ 具有类似（但通常要求较宽松）的可微性；详情请见[23]。$\mathcal{K}_r$ 列指出某个核是否包含在后续章节使用的受限核集合中。该集合是文献中最常用核的精简集合。采用 $\mathcal{K}_r$ 的一个原因是尽量减少核性质方面的重叠；正如下文所述，$\mathcal{K}_f$ 中的某些核族显然会产生非常相似的拟合，并可相互充当“替代品”。

所列出的所有核都具有超参数，有助于理解它们与数据之间的关系。许多平稳核中出现的量 $\ell_{\text{len}}$ 称为特征长度尺度，它是一个距离缩放因子。以径向基函数核为例，它大致描述了在输入空间中，$x'$ 需要离 $x$ 多远，函数值才会变得不相关 [44]。因此，经过校准的高斯过程的长度尺度可以解释为训练数据集中相关性衰减的强度。在平稳核中，Mat´ern 类是一类常见的核。在连续输入空间中，Mat´ern-$\nu$ 中的值 $\nu$ 对应于平滑性：采用 Mat´ern-$\nu$ 核的高斯过程在均方意义下可微 $\lceil \nu \rceil - 1$ 次 [44]。当 $\nu \to \infty$时，径向基函数（RBF）核是其极限情形，此时得到的过程处处可微。$\nu = 1/2$ 情形对应于广为人知的 Ornstein-Uhlenbeck过程，该过程具有均值回复特性且不可微。同样不可微但属于非平稳一侧的最小核，在 $x \in \mathbb{R}_+$时对应于布朗运动过程，其中 $t_0^2 = \text{var}(f(0))$ 是初始方差。对于离散的 $x$，Min 核得到随机游走过程，而 M12 得到一阶自回归过程。

较少被研究的是（连续时间）AR2（见 Parzen [33]）、柯西核和迈勒核。连续时间AR2核的作用与当 [33] 限制为整数时具有复特征多项式根的二阶离散时间自回归（AR）

| 核名称 | 缩写 | 公式 $k(x,\ x';\theta)$ | 性质 | $\mathcal{K}_r$ |
|---|---|---|---|---|
| Mat´ern-1/2 | M12 | $\exp\left(-\frac{\lvert x-x'\rvert}{\ell_{\text{len}}}\right),\quad \ell_{\text{len}}>0$ | $C^0$ | ✓ |
| Mat´ern-3/2 | M32 | $\left(1+\frac{\sqrt{3}}{\ell_{\text{len}}}\lvert x-x'\rvert\right)\exp\left(-\frac{\sqrt{3}}{\ell_{\text{len}}}\lvert x-x'\rvert\right),\quad \ell_{\text{len}}>0$ | $C^1$ | |
| Mat´ern-5/2 | M52 | $\left(1+\frac{\sqrt{5}}{\ell_{\text{len}}}\lvert x-x'\rvert+\frac{5}{3\ell_{\text{len}}^2}\lvert x-x'\rvert^2\right)\exp\left(-\frac{\sqrt{5}}{\ell_{\text{len}}}\lvert x-x'\rvert\right)$ | $C^2$ | ✓ |
| 柯西 | Chy | $\frac{1}{1+\lvert x-x'\rvert^2/\ell_{\text{len}}^2},\quad \ell_{\text{len}}>0$ | $C^\infty$ | |
| 径向基 | RBF | $\exp\left(-\frac{\lvert x-x'\rvert^2}{2\ell_{\text{len}}^2}\right),\quad \ell_{\text{len}}>0$ | $C^\infty$ | ✓ |
| AR2 | AR2 | $\exp(-\alpha\lvert x-x'\rvert)\left\{\cos(\omega\lvert x-x'\rvert)+\frac{\alpha}{\omega}\sin(\omega\lvert x-x'\rvert)\right\}$ | 周期，$C^1$ | |
| 线性 | Lin | $\sigma_0^2+x\cdot x',\quad \sigma_0>0$ | 非平稳 | * |
| 最小值 | Min | $t_0^2+x\wedge x',\quad t_0\geq 0$ | 非平稳，$C^0$ | ✓ |
| Mehler | Meh | $\exp\left(-\frac{\rho^2(x^2+x'^2)}{2(1-\rho^2)}+\frac{\rho xx'}{1-\rho^2}\right),\quad -1\leq\rho\leq 1$ | 非平稳 | |

表1：组合搜索中使用的核族列表。 $C^p$ 表示高斯过程样本路径 $x\mapsto f(x)$具有 $p$ 连续导数； $C^0$ 连续但不可微。列 $\mathcal{K}_r$表示该核族是否属于受限搜索集合。线性核仅用于其年份分量。

过程完全相同。重尾的柯西概率密度函数促成了柯西核，其目标是对长程依赖进行建模；它也是有理二次核的一个特例（见附录 A）。最后，迈勒核的形式类似于联合正态密度，并且其作用类似于经过非平稳修正的径向基函数核（这可以通过“配方”论证看出）。尽管 Mehler 是非平稳的，但它出人意料地产生了一个平稳的相关函数 $\text{corr}(f(\mathbf{x}),\ f(\mathbf{x}'))\ =k(\mathbf{x},\mathbf{x}')/\sqrt{k(\mathbf{x},\mathbf{x})k(\mathbf{x}',\mathbf{x}')}$。有关上述核及其性质的更深入讨论，请参见附录 A。

## 2.4 与死亡率建模的联系

对于死亡率建模，我们的核心输入空间由年龄和年份坐标组成： $x_a, x_y\in\mathbb{R}_+^2$。由于高斯过程能够建模非线性关系，我们将出生队列 $x_c:=x_a-x_y$ 作为$\mathbf{x}$的第三个坐标，因此$\mathbf{x}=(x_a,x_y,x_c)$。对于给定的$\mathbf{x}$，分别用 $D_{\mathbf{x}}$ 和 $E_{\mathbf{x}}$ 表示对应$(x_a,x_y)$对中观测到的死亡人数和暴露量（即期初存活的个体）。将 $y(\mathbf{x})\ =\log(D_{\mathbf{x}}/E_{\mathbf{x}})$表示为对数死亡率。完整数据集表示为$\mathcal{D}=\{\mathbf{x}_i,y_i,D_i,E_i\}_{i=1}^n$。每个死亡率观测 $y_i:=y_i(\mathbf{x}_i)$都是方程（3）中建模的回归量，因此 $\mathbb{E}[y(\mathbf{x})|f(\mathbf{x})]=f(\mathbf{x})$。其解释是，真实死亡率 $f(\mathbf{x})$是在均值为零（即无偏）的不相关噪声作用下观测到的，从而得到测得的死亡率经验 $y_i$。

关于对数正态分布，通过代数推导可得

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

类似于现有死亡率建模文献中的过度离散泊松模型（例如见 [2]）。请注意，Brouhns 等人 [4]关于对数死亡率建模的开创性研究指出，假设噪声具有同方差并不现实，因为高龄人群的死亡人数绝对值要小得多。

与较年轻年龄相比。因此，我们采用异方差噪声。

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

因此，我们令观测方差与观测到的死亡人数成反比，其中常数 $\sigma^2$ 将在拟合过程中学习。从建模角度看，这种做法是可行的，因为只要已知 $E_{\mathbf{x}}$ 和 $y$（$\mathbf{x}$），$D_{\mathbf{x}}$ 也就已知。事实上，方程（5）表明，样本外预测无需知道$D_{\mathbf{x}}$ 。如果需要完整的分布预测，也可以将噪声曲面 $\sigma^2(\cdot)$ 与 $f(\cdot)$ 同时建模；例如参见 [7]。

对高斯过程协方差的讨论自然会联系到年龄-时期-队列模型。平稳协方差意味着，不同年龄组或日历年之间的依赖关系仅由相应的年龄或年份间隔决定，不受其他结构性变化的影响。加法和可分协方差结构在现有死亡率建模文献中发挥着重要作用，尤其是在年龄-时期-队列框架中。例如，Lee-Carter模型和Cairns-Blake-Dowd（CBD）模型最初是年龄-时期模型，后来逐渐加入队列效应或额外的时期效应。Dowd 等人 [9] 提出的最新 CBDX 框架不再预先设定确切的年龄-时期-队列项，而是按需加入时期效应。类似地，Hunt 和 Blake [16] 提出了构建死亡率模型的一般方法：用特定的参数形式表示核心人口统计特征，并将其组合成一个总体结构。高斯过程也可以做到这一点，其中的核将专家对这些人口统计特征的判断编码其中。这种将信息编码进协方差的做法在文献中早已存在，只是研究者可能并未意识到这一点。以基本 CBD 模型为例，其随机部分的形式为 $f$（$\mathbf{x}$） $= \kappa_{x_{yr}}^{(1)} + (x_{ag} - \mathbf{x}^{ag})\kappa_{x_{yr}}^{(2)}$，其中 $\mathbf{x}^{ag}$是 $\mathcal{D}$中的平均年龄。在通常假设 $(\kappa_{x_{yr}}^{(1)}, \kappa_{x_{yr}}^{(2)})$ 是带漂移的多元随机游走的情况下，经过常规计算可得

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

以及 $k_{CBD}(\mathbf{x}, \mathbf{x}')$ =协方差$(f(\mathbf{x}) \; f(\mathbf{x}'))$ =

$$= \left[\sigma_1^2 + \rho\sigma_1\sigma_2(x_{ag} + x'_{ag} - 2\overline{\mathbf{x}_{ag}}) + (x_{ag} - \overline{\mathbf{x}_{ag}})(x'_{ag} - \overline{\mathbf{x}_{ag}})\sigma_2^2\right] \left(x_{yr} \wedge x'_{yr}\right). \tag{8}$$

我们可以将上述内容重新解释为核分解 $k_{CBD}(x, x') = k_1(x_{ag}, x'_{ag})k_2(x_{yr}, x'_{yr})$，其中 $k_2(x_{ag}, x'_{ag})$ 是一个仅依赖于年龄的核（可以进一步分解为加法分量），而 $k_2$ 是仅依赖于年份的最小核。此外，如果假设多变量随机游走服从高斯分布，那么 $f(\mathbf{x})$ 实际上构成一个（离散时间）高斯过程，因此第 2.1 节中详细介绍的方法可以直接应用。

时期效应和队列效应通常使用时间序列模型 [42] 进行建模。特别是，高斯 ARIMA 模型在现有的年龄-时期-队列文献中十分常见；例如， [6] 指出，一阶自回归过程、ARIMA(1,1,0) 和 ARIMA(0,2,1) 对于队列效应很有用，而 ARIMA(1,1,0) 或 ARIMA(2,1,0) 对于时期效应很有用。这也提供了与（离散）高斯过程协方差类比之间的另一条联系：Mat´ern-1/2 协方差对应于一阶自回归过程，（连续时间）AR2 协方差对应于具有复单位根的 AR(2) 过程，而最小协方差对应于离散时间随机游走。

## 3 高斯过程的遗传编程

从表1中的核构件出发，通过加法和乘法可以构造出无限多种组合核。遗传算法的思路是

通过进化过程自适应地探索核空间。在遗传算法的每一步中，适应度分数较高的核更有可能进化并被传递下去，而适应度较低的核则会被淘汰。进化通过若干种可能的操作实现，每次操作都会随机选取。在遗传算法的第一个子步骤中，会确定下一代核的祖先。这通过“锦标赛”来完成：随机选取第$g$代核，同时优先选择适应度较高的核。同一个核可以在多场锦标赛中入选，也就是说，它可以产生多个子代。在第二个子步骤中，每个祖先都会经过交叉（与另一个祖先混合核分量）或变异（修改唯一祖先的一个分量），从而生成第$(g+1)$代核。

### 3.1 GP 的贝叶斯信息准则和贝叶斯因子

为了评估给定集合 $k \in \mathcal{K}$ 中某个核的适用性，一个具有吸引力的准则是给定数据 $p(k|y) = p(y|k)p(y)/p(k)$ 时该核的后验似然，其中，在均匀先验假设 $p(k) = 1/|\mathcal{K}|$ 下，我们看到 $p(k|y) \propto p(y|k)$。然而，对超参数 $p(y|k) = \int_\theta p(y,\theta|k)d\theta$ 的积分通常难以处理，因此我们使用 贝叶斯信息准则 （BIC）进行近似，其中 $\text{BIC}(k) \approx \log p(y|k)$ 定义为：

$$\text{BIC}(k) = -l_k(\hat{\theta};\mathbf{y}) + \frac{|\hat{\theta}|\log(n)}{2}, \tag{9}$$

其中，$l_k(\theta|\mathbf{y}) = \log p(y|k,\theta)$ 是在给定核 $\hat{k}$ 下对 $y$ 在 $\theta$ 处求得的对数边际似然，$\theta$ 是 $l_k(\theta|\mathbf{y})$ 的最大化者（最大边际似然估计），而 $|\theta|$ 是 $\hat{\theta} = \{\hat{\beta}_0, \hat{\beta}_{ag}, \hat{\theta}_k, \hat{\sigma}^2\}$ 中估计的超参数总数，其中 $\theta_k$ 是所有特定于核的超参数组成的向量。注意，$p(y|k,\theta)$ 是一个多元密度，其均值和协方差由公式（4）决定。BIC 指标已用于 GP 组合核搜索的类似应用中，例如 [11, 10]，，并且通常用于死亡率建模 [5]。在下文中，我们使用 BIC（BIC 越低越好）作为 GA 的适应度指标。

还可以通过再次假设 $\mathcal{K}$ 上的均匀先验并计算贝叶斯因子 （BF），进一步评估 $k_1, k_2 \in \mathcal{K}$ 的相对似然。

$$\text{BF}(k_1,k_2) = \frac{p(k_1|\mathbf{y})}{p(k_2|\mathbf{y})} \approx \exp\Big(\text{BIC}(k_2) - \text{BIC}(k_1)\Big). \tag{10}$$

Gelman 等人 [13] 指出，贝叶斯因子在离散模型选择的情况下效果良好。Jeffreys [21] 的开创性著作给出了一个证据类别表，用于根据贝叶斯因子得出结论；参见附录D中的表15，该表至今仍被频繁使用 [25, 8]。

请注意，根据公式（9）中，贝叶斯信息准则$(k_2)$−贝叶斯信息准则$(k_1)$之间的惩罚差异仅仅是额外核超参数的数量乘以$\log(n)/2$，因为 $\beta_0$，$\beta_{ag}$，$\sigma^2$ 无论核函数选择如何，始终都会进行估计。因此，在高斯过程核选择的应用中，贝叶斯因子会适当地惩罚核复杂度。在表1中，大多数核只有一个 $\theta_k$ （长度尺度），但有些核（如AR2）具有两个超参数，因此会受到相对惩罚。

### 3.2 遗传算法核表示

为了在核空间中进行操作，我们将通过树状结构表示核，参见图 1。内部节点对应于算子（加法或乘法），用于将两个不同的核组合在一起，而叶节点则是作为构件使用的单变量核。我们通过相应的下标指示核所作用的坐标，，，例如 M。

反过来，这类树会被转写为括号表达式，例如 $\kappa$ = add(Exp c, mul(RBF a, add(Mat y, RBF c)))，其对应于年龄-时期-队列核（$k_{M52}(x_{yr}) + k_{RBF}(x_c)$）· $k_{RBF}(x_{ag}) + k_{Exp}(x_c)$。- $\kappa$ 的长度记为 $|\kappa|$，即其节点数量。上述核树的长度为 $|\kappa| = 7$，也就是由 4 个基核和 3 个算子组合而成。请注意，树结构并不唯一，即某些复杂核可以进行置换，并通过不同的树来表示。在下文中，我们将忽略这种非唯一性。

将上述表示转换为其所蕴含的依赖关系结构需要一定的专业知识。一种可视化方法是利用平稳性，将矩阵 $k_S(\mathbf{x} - \mathbf{x}')$ 的热图绘制为其年龄、年份坐标的函数。

### 3.3 遗传算法操作和超参数

整体遗传算法由一组可能的变异和一系列超参数概括。这些内容之所以重要，原因有很多：（i）确保充分探索，使算法不会陷入某种特定的核配置；（ii）提高效率，即以找到最佳核函数所需的代数和每代规模来衡量；（iii）控制膨胀，也就是确保返回的核不会过于复杂，并保留可解释性。可解释性较强的核通常长度较短（低于 10），并避免重复模式。膨胀控制，即避免出现过于复杂或过长的核，是遗传算法面临的一项挑战。Luke 和 Panait [28] 以及 Poli 等人 [35]提出了几种应对方法。

我们的具体高层 GA 超参数列于表 2。我们主要遵循 Sipper 等人 [41] 的指南，该指南对遗传算法的参数空间进行了全面研究。祖先通过锦标赛机制选出：从上一代中独立且均匀地抽取 $T = 7$ 个体，适应度最高（BIC 最低）的个体成为单场锦标赛的胜者。较小的 $T$ 能为后续各代提供多样性，而较大的 $T$则能降低舍弃适应度良好个体的可能性。为抑制膨胀，我们采用 [28]中描述的双重锦标赛流程：将单次锦标赛的所有实例替换为依次进行的两场锦标赛，并以 $D/2$ 的概率选择长度较短的祖先（$D \in [1, 2]$；值越大越倾向简约性而非适应度）。我们采用限制较少的 $D = 1.2$ ，而非建议的 $D = 1.4$，部分原因是 BIC 内置了对复杂模型的惩罚，因此能够抑制膨胀。为进一步提倡简约性，我们采用 Poli 等人 [35] 建议的提升变异，设置为 $p = 0.1$。使用上述参数设置时，我们在实验中很少观察到长度超过 15 的核。

表3详细介绍了交叉和变异操作及其超参数，图1对此进行了直观说明。其中大多数操作都是文献中的标准操作。根据我们的领域知识，我们提出了另一种点变异算子，称为保留特征点变异。该算子会保留正在发生变异的核的坐标，因此，对年龄进行操作的核会被替换为另一个对年龄进行操作的核，以此类推。这一操作保留了已发现的APC结构，并对选定的坐标进行微调。

在交叉和变异操作中，需要保留函数的元数（参数数量）。在我们的设置中，唯一具有非平凡元数的函数是add和mul（二者的元数均为2），因此，如果在变异中选中了这些节点，点变异（以及保留特征点变异）会自动将其替换为另一种操作：将add替换为mul，反之亦然。

在一些遗传算法应用中，作者提出使用数百代（甚至数千代）。由于拟合高斯过程模型所需的时间并不短，运行这么多代在计算上难以承受。下文中，我们使用固定数量的 $G = 20$ 代。如图5所示，核的探索似乎在十几代后趋于稳定，因此继续增加

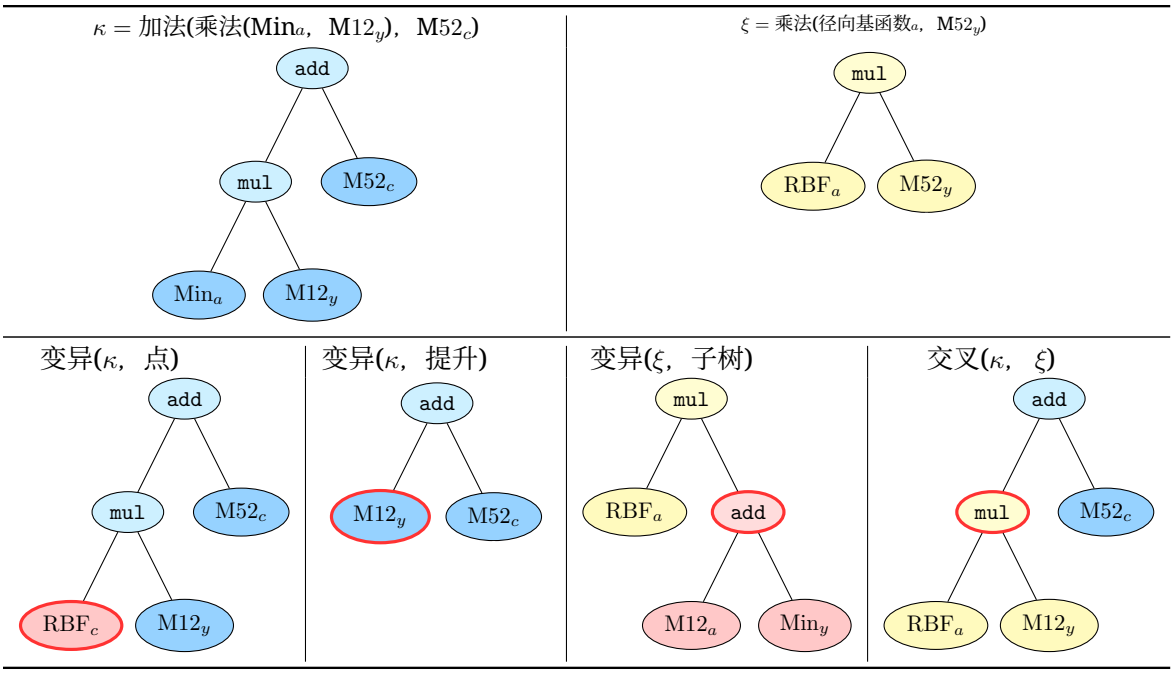



图1：代表性的组合核和遗传算法操作。加粗的红色椭圆表示选定用于变异或交叉的 $\kappa$ （或 $\xi$）节点。

| 参数 | 值 | 注释 |
|---|---|---|
| 人口群体规模 | $n_g = 200$ | 每代的个体数量 |
| 代 | $G = 20$ | 代数 |
| 锦标赛规模 | $T = 7$ | 进行双重锦标赛，并选择较小的获胜者概率为 $D/2$ |
| | $D = 1.2$ | 较小者以 0.60 的概率获胜 |

表 2：遗传算法的高级超参数及其说明。请注意， $G \cdot n_g = 4000$是单次遗传算法运行中训练的高斯过程模型总数。

$G$的收益似乎有限。与典型的遗传算法应用不同，我们更关注对所有典型的高性能核进行解读，而不是将目标函数优化至绝对最小值；也就是说，只要发现一批具有代表性的最优模型，大致涵盖合理范围，我们的分析就足够了。

将第 $i = 1, \ldots,$ $n_g$代中的第 $i = 1, \ldots,$ $n_g$个个体 记为 $\kappa_i^g$，其对应的贝叶斯信息准则为 $b_i^g$ 。算法 1 概述了完整的遗传算法。BIC 按照公式（9）计算贝叶斯信息准则，该过程会隐式地拟合高斯过程并优化超参数。均匀采样根据规模为 $T$的锦标赛确定祖先。根据 $\mathcal{G}_2$ 中给出的概率（参见表 3），确定进行交叉还是变异，其中 Type 表示所选变异的类型。如果发生交叉，则通过与确定第一个祖先相同的过程确定另一个祖先。整个算法运行 $G$ 代，其中 $g = 0$ 次用于初始化，其余 $G - 1$ 步则用于选择祖先和后代。对于 $g = 0$，初始化核函数会构造一个随机初始化的核，其长度从 $\{3, 5, 7, 9\}$中均匀选取（分别为 2,3,4,5 个基核）；基核从 $\mathcal{K}$中均匀采样，加法和乘法运算的概率相同。我们还进行了使用更多样化的初始化方式的实验

| 概率 | 遗传算法操作 | 注释 |
|---|---|---|
| $p_c = 0.45$ | 交叉 | |
| $p_s = 0.2$ | 子树变异 | |
| $p_h = 0.1$ | 提升变异 | |
| $p_p = 0.05$ | 点变异 | 每个节点以相同元数的另一个节点进行变异概率为 $q_p$ |
| $p_r = 0.15$ | 保留式点变异 | 每个节点以相同元数的另一个节点进行变异且相同的（年龄，年份，队列），概率为 $q_r$ |
| $p_o = 0.05$ | Copy | |
| $q_p = 0.25$ | 点替换 | |
| $q_r = 0.35$ | 尊重式替换 | |
| $q_a = 0.5$ | | 初始化时包含加法/乘法的概率 |
| | | 树 |

表 3：各算子特定的遗传算法超参数及说明。注意，$p_c + p_s + p_h + p_p + p_r + p_0 = 1$。

**算法 1：** 用于组合核选择的遗传算法

**输入：** 核搜索集 $\mathcal{K}$；遗传算法超参数 $\mathcal{G}_1 = \{n_g, G, T\}$ 和 $\mathcal{G}_2$ 分别见表 2 和表 3；数据集 $\mathcal{D} = \{(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)\}$

**输出：** 核表达式和 BIC 分数 $\{\{\kappa_i^g, b_i^g\}_{i=1}^{n_g}\}_{g=0}^{G-1}$

**对于** $i = 1$ **到** $m_0$，**执行**
  $\kappa_i^0 \leftarrow$初始化核函数();
  $b_i^0 \leftarrow$贝叶斯信息准则$(\kappa_i^0)$
**对于** $g = 1$ **到** $G - 1$ **执行**
  **对于** $i = 1$ **到** $\eta_g$ **do**
    **DetermineAncestor**
      $\{\kappa^{g-1}_{(1)}, \ldots, \kappa^{g-1}_{(T)}\} \leftarrow$ 均匀采样$(\{\kappa_1^{g-1}, \ldots, \kappa_{n_{g-1}}^{g-1}\}, T)$;
      $j_i \leftarrow \text{argmin}\{b^{g-1}_{(1)} \ldots, b^{g-1}_{(T)}\}$;
      $A_i^{g-1} \leftarrow \kappa_{j_i}^{g-1}$ ;
    操作 $\leftarrow$ 样本（交叉, (变异, 类型)） // 根据 $\mathcal{G}_2$;
    **如果** 操作 == 交叉**则**
      $A'^{g-1}_i \leftarrow \kappa^{g-1}_{j'_i}$ // 第二祖先（根据 DetermineAncestor）；
      $\kappa_i^g \leftarrow$ 交叉$(A_i^{g-1}, A'^{g-1}_i)$;
    **else**
      $\kappa_i^g \leftarrow$变异$(A_i^{g-1}$ ; 类型);
    $b_i^g \leftarrow$贝叶斯信息准则$(\kappa_i^g)$

（核长度从 1 到 15）减缓了遗传算法的收敛，但对最终结果的影响微乎其微。我们的初始化旨在尽可能提供无偏的抽样过程。也可以采用其他初始化方式，例如对给定核中的年龄、时期和队列项施加多样化约束，或在初始代中加入已知效果良好的核。还值得一提的是，尽管我们对所有 $n_g = 200$ ，… , $G-1$都固定了该值，但此值可以随 $g$变化，因此结果总数为 $\sum_{g=0}^{G-1} n_g$。

**注释2.** 如前所述，我们采用双重锦标赛方法来抑制膨胀。具体而言，在算法 *1* 中，每当运行 *DetermineAncestor* 时，实际上都会运行两次，以得到$A_{i\ 1\ ,()}^{g-1}$ 和$A_{i\ 2\ ,()}^{g-1}$，然后以概率 $D/2 = 0.6$选取长度较短的获胜者$A_i^{g-1} = A_{i\ \ ,(j_0)}^{g-1}$ ，其中$j_0 = \mathrm{argmin}_{j=1,2} |A_{i\ \ ,(j)}^{g-1}|$。详情请参见 *[28]*。为简洁起见，算法 *1* 中省略了这一点。

由于锦标赛胜者优先选择表现最佳的核，它们会成为后续生成器的祖先，并且很可能以重复项的形式再次出现（可能是由于复制操作，也可能是几次变异相互抵消所致）。因此，汇总各代所有$\kappa_i^g$ 时，我们观察到许多重复项。

### 3.4 高斯过程超参数优化

如前所述，给定核 $k(\cdot,\ \cdot;\theta)$ 的超参数通过最大化 $l_k(\theta|\mathbf{y})$ 来估计，即最大化观测数据的边际对数似然。高斯过程中的核超参数优化景观是非凸的，且存在许多局部极小值，因此我们对此优化格外谨慎。由于 $\mathbf{y}$ 处于对数尺度，我们不对这些值进行变换。对于给定的 $\mathbf{x} = [x_a, x_y, x_c]^\top \in \mathbb{R}^3$，我们按维度缩放至单位区间。例如，如果 $x_a = [x_{1,a}, \ldots, x_{n,a}]^\top$，那么$x_{i,\ a} \mapsto \frac{x_{i\ ,\ a} - \min(x_a)}{\max(x_a) - \min(x_a)}$，$x_y$ 和 $x_c$也采用类似处理。平稳核的长度尺度可以通过逆变换 $\ell_{\text{len}} = (\max(x) - \min(x))\,\ell_{\text{len}}$ 在原始尺度上进行解释，其中 $\ell_{\text{len}}$ 处于变换后的尺度；均值函数参数也采用类似的变换。为便于解释，只要条件允许，我们就利用这一方法在第4节和第5节中报告原始尺度上的数值。非平稳核（即 Min，Meh 和 Lin）的超参数保留在变换后的尺度上，因为它们通常无法在原始尺度上进行解释。

我们使用带有 GPyTorch 库的 Python [12]来高效处理数据和矩阵运算。由于我们的结果高度依赖准确的似然值（通过贝叶斯信息准则），因此我们关闭了 GPyTorch 的默认矩阵近似方法。使用 Adam [24]最大化 $l_k(\theta|\mathbf{y})$。这是一个计算开销较大的过程，因为对 $l_k(\theta|\mathbf{y})$进行朴素求值需要通过求逆 $\mathbf{K}(\mathbf{X},\mathbf{X})$来进行 $\mathcal{O}(n^3)$。我们的遗传算法运行使用收敛容差 $\varepsilon = 10^{-4}$ ，最大迭代次数为 $\eta_{\max} = 150$。我们发现，大多数简单核都能快速收敛（$\eta \leq 100$），即使达到 $\varepsilon = 10^{-6}$也是如此，见附录中的表13。由于我们需要拟合 $n_g \cdot G \gg 10^3$ 个模型，因此将 $\eta_{\max}$ 保持在相对较小的水平。完成后，排名靠前的几十个核会使用 $\varepsilon = 10^{-6}$ 和 $\eta_{\max} = 1000$重新拟合。请注意，由于 Adam 是一种随机算法，在此次重新拟合过程中，拟合得到的核超参数会发生变化（在我们的实证工作中变化幅度很小）。这体现了这样的理念：遗传算法是一种初步的“鸟瞰式搜索”，用于寻找合理的死亡率结构，之后再对超参数估计进行细化。除少数情况外，最终贝叶斯信息准则值的变化都很小，不过排名靠前的核的排名偶尔会有所调整。

## 4 合成死亡率核恢复

遗传算法的基本原理是进行广泛搜索，从而为给定数据集 $\mathcal{D}$正确识别合适的协方差结构。遗传算法可视为一种初步搜索，用于得出

| 实验 | 真实值核 | $\sigma^2(\mathbf{x})$ | $\beta_0$ | $\beta_{ag}$ |
|---|---|---|---|---|
| SYA | $0.04 \cdot$ 径向基函数$_a$(0.4) $\cdot$ 径向基函数$_y$(0.3) | 0.001 | -5.0 | 3.4 |
| SYB | $0.08 \cdot$径向基函数$_a$(0.586)$\cdot M12_y$(13.33)$+0.02 \cdot M52_c$(0.079) | 0.0004 | -5.568 | 2.974 |
| SYC | $0.0134 \cdot M52_a$(1.132) $\cdot$Min$_y$(0.877) $\cdot M12_c$(96.234)$\cdot$ $\cdot$Meh$_c$(0。8483) | $1.0783/D_{\mathbf{x}}$ | -3.165 | 3.380 |

表 4：合成数据集说明。数据根据式 (4) 生成，作为多元正态实现，其中参数化均值函数为 $m(\mathbf{x}) = \beta_0 + \beta_{ag} x_{ag}$。SYA 和 SYB 为同方差数据。生成 SYC 的异方差噪声时，$D_{\mathbf{x}}$ 取自日本女性数据；有关该数据集的详情，请参见第 5 节。

数千个（$n_g \cdot G$）候选项，随后便可找出拟合效果最好的少数几个核，用以最佳表示 $\mathcal{D}$的协方差结构。为评估这一核发现流程的质量，我们首先尝试三个合成数据集，其中真实的数据生成过程，即 APC 协方差结构，已知，因此可以直接将遗传算法的输出与已知真值进行比较。这样可以检验遗传算法能否恢复真实协方差，并通过贝叶斯因子分析哪些核表达了相似的信息、可以相互替代。此外，还可以评估噪声对核恢复的影响。这些实验也为理解高斯过程核比较中的贝叶斯因子证据类别提供了校准依据（见表15）。较简单的实验聚焦于 2.3 节所述的受限核集 $\mathcal{K}_r$，其中 Lin$_y$ 是唯一出现的线性核（用于模拟日历年中的线性死亡率改善/下降）。其他实验则采用完整核集 $\mathcal{K}_f$，其中包含表1中涵盖所有坐标维度的全部核。

完整的实验设置如下。首先，我们在年龄、时期、队列空间中固定一个高斯过程核，并从该先验中精确采样，生成相应的对数死亡率曲面。由此得到一个合成死亡率曲面。下文的曲面覆盖年龄50–84岁和年份1990–2019年，并包含年龄上的线性参数趋势 $\beta_0 + \beta_{ag} x_{ag}$。因此，我们的训练集包含 $35 \cdot 30 = 1050$个输入，其规模与第5节所用的HMD数据集相同。

每种合成死亡率结构都用一个三字母代码标识。前两个实验（SYA、SYB）采用同分布且相互独立的观测噪声 $\epsilon \sim \mathcal{N}(0,\ \sigma^2)$；第三个实验（SYC）采用异方差噪声 $\epsilon(\mathbf{x}) \sim \mathcal{N}(0,\ \sigma^2/D_{\mathbf{x}})$，其中 $D_{\mathbf{x}}$ 取自真实世界的HMD日本女性数据集，以体现年龄上的现实异质性。具体设置及超参数见表4，所得合成死亡率曲面可在github.com/jimmyrisk/gpga-synthetic-surfaces公开重用。所有合成核及其超参数都符合现实世界的合理情况。具体而言，SYA、SYB和SYC均来自对HMD数据集进行的典型遗传算法运行，因此与其结构相符。

第一个案例研究（SYA）从一个极其简单的结构开始（长度为 3 的核，即由 2 个基核组成），以检验遗传算法在探索核空间时是否能保持简约性。换言之，我们使用 SYA 来验证，在遗传算法的演化过程中，长度更长的核是否会被正确地视为过拟合，并且与真实核相比获得更低的适应度分数。其次，SYA 也用于评估核平滑性的恢复情况，因为 RBF$_a \cdot$RBF$_y$ 在年龄和年份分量上都具有平滑性。由于我们针对给定的训练集最小化贝叶斯信息准则，因此对于某个给定的实现，与生成数据所用的核不同的核实际上可能取得（略微）更低的贝叶斯信息准则。因此，我们进行了两次实验（分别生成 SYA-1 和 SYA-2），以评估抽样变异性。

第二个合成示例（SYB）采用了一个更复杂的长度为 7 的核（包含 3 个基核），并且结合了乘性和加性结构。其灵感来源于

| SYA-1 | | | SYA-2 | | |
|---|---|---|---|---|---|
| BIC | $\widehat{\text{贝叶斯因子}}(k,\ K_0)$ | 核 | BIC | $\widehat{\text{贝叶斯因子}}(k,\ K_0)$ | 核 |
| **-2034.23** | **1.0000*** ** | $\mathbf{RBF_a RBF_y}$ | -2066.93 | 1.1907*** | $\mathrm{M52}_a\,\mathrm{RBF}_y$ |
| -2034.04 | 0.8264*** | $\mathrm{M52}_a\,\mathrm{RBF}_y$ | **-2066.76** | **1.0000*** ** | $\mathbf{RBF_y RBF_a}$ |
| -2031.82 | 0.0902* | $\mathrm{M52}_a\mathrm{M52}_y$ | -2064.63 | 0.1216** | $\mathrm{M52}_a\mathrm{M52}_a\,\mathrm{RBF}_y$ |
| -2031.29 | 0.0526* | $\mathrm{M52}_a\,\mathrm{RBF}_a\,\mathrm{RBF}_y$ | -2064.24 | 0.0801* | $\mathrm{M52}_a\,\mathrm{RBF}_a\,\mathrm{RBF}_y$ |
| -2031.09 | 0.0433* | $\mathrm{M52}_a\mathrm{M52}_a\,\mathrm{RBF}_y$ | -2063.88 | 0.0561* | $\mathrm{M52}_a\mathrm{M52}_a\,\mathrm{RBF}_y$ |

表5：第一个合成案例研究 SYA 中拟合度最高的五个非重复核。加粗的是$K_0 = \mathrm{RBF}_y\mathrm{RBF}_a$，即数据生成中使用的真实核。SYA-1 和 -2 表示所使用的训练实现。

Lee-Carter模型，并具有乘性的年龄-时期分量，以及较不显著的加法队列效应（系数为0.02，而非0.08）。其目的是：(i) 检验遗传算法能否识别这种高层次结构（包括在正确项上进行加法和乘法），(ii) 将长度较短的核识别为欠拟合，以及 (iii) 区分年龄、年份和队列这三类项。

第三种（SYC）是纯乘性的，具有更复杂的核时期-队列结构，包含4个分量，代表真实实验中的结果。Min$_y$ 核加入了一个非平稳的时期效应，而队列效应则相当复杂，由一个平滑的非平稳Mehler项以及粗糙的M12$_c$组成。请注意，较大的长度尺度 $\ell_{\text{len}}$ =6062.76对应于一个近乎非平稳的过程。通过在三维输入空间中施加由四个核构成的可分离结构，本实验考察遗传算法在可能存在过度参数化和/或共线性的情况下，能否在多个维度上恢复非平稳性和粗糙性。此外，由于 Meh $\notin \mathcal{K}_r$，我们使用SYC作为实验，以评估高斯过程能否通过更简单的核近似Meh协方差结构，具体做法是分别在 $\mathcal{K}_r$ 和 $\mathcal{K}_f$条件下运行遗传算法$\mathcal{K}_r$ 。

SYA 和 SYB 均仅针对 $\mathcal{K}_r$进行搜索，其中 SYA 测试基本的年龄-时期-队列设置，SYB 则侧重于恢复加性结构和线性系数。使用 $\mathcal{K}_f$ 可能会使分析复杂化，因此留待 SYC 处理。每次遗传算法运行的初始代随机数生成器种子各不相同；所有遗传算法运行的超参数均相同，具体如第 2 节和第 3 节所述。

**注3**。 公式$(10)$ 中的理论贝叶斯因子假定所有超参数都已积分消去 $p(y|k) = \int_\theta p(y|k,\theta)d\theta$，或者在使用贝叶斯信息准则时以最大似然估计替代。因此，通过贝叶斯因子检验的假设纯粹关乎核函数的选择。

## 4.1 合成结果

上一节提出的问题的答案见表 5（SYA）、表7（SYB）和表 8（SYC）。在所有表格中， $K_0$ 表示生成合成ˆ数据的已知核。为实现核恢复（忽略超参数估计），根据生成的数据为 $K_0$ 估计 $\theta$。

### SYA

表5显示了 SYA-1 和 SYA-2 中排名前5的拟合度最高的核。估计的贝叶斯$\widehat{\ }$因子（使用贝叶斯信息准则）列于 BF($k, K_0$) $=$ exp(贝叶斯信息准则($K_0$) $-$ 贝叶斯信息准则($k$)) 这一列。对于 SYA-1，真实核$K_0$ $=$ 径向基函数$_y$径向基函数$_a$ 被发现，且其贝叶斯信息准则最低。其次最佳的（具有较大的贝叶斯人片在线观看

| | SYA-1 | | | | SYA-2 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mathrm{M12}_a$ | $\mathrm{M32}_a$ | $\mathrm{M52}_a$ | $\mathrm{RBF}_a$ | $\mathrm{M12}_a$ | $\mathrm{M32}_a$ | $\mathrm{M52}_a$ | $\mathrm{RBF}_a$ |
| $\mathrm{M12}_y$ | -63.98 | -38.96 | -37.49 | -39.41 | -96.99 | -68.85 | -66.12 | -65.25 |
| $\mathrm{M32}_y$ | -35.83 | -8.13 | -5.98 | -7.12 | -53.24 | -21.19 | -17.67 | -16.06 |
| $\mathrm{M52}_y$ | -32.29 | -4.70 | -2.41* | -3.33* | -42.75 | -11.36 | -8.09 | -6.92 |
| $\mathrm{RBF}_y$ | -30.24 | -3.14* | -0.19*** | 0.00*** | -32.67 | -2.31* | 0.17*** | 0.00*** |

表 6：对数贝叶斯因子 $\log\ BF(k, K_0) = \mathrm{BIC}(K_0) - \mathrm{BIC}(k)$，针对 $K_0 = \mathrm{RBF}_a\,\mathrm{RBF}_y$ 和 $k = k_a k_y$，其中 $k_a$ 和 $k_y$ 分别对应行标签和列标签；$\mathrm{BIC}(K_0)$是在 SYA-1（左侧面板）和 SYA-2（右侧面板）上计算的。

因子为0.826）的模型是$\mathrm{M52}_a\mathrm{RBF}_y$，除了年龄使用了平滑程度稍低的Mat´ern-$5/2$ 核而非RBF核之外，其结构完全相同。需要注意的是，根据附录中的表15，只有当贝叶斯因子低于 $1/3 \approx 0.3333$时，才值得提及这两种核之间存在差异，这表明这两个模型实际上没有区别。我们的解释是，抽样变异使得 $M52_a$ 成为 $RBF_a$的一个相似备选项。然而，值得注意的是，两张表中均未出现任何阶数更低的Mat´ern核。其余三行具有相似的核结构，只是额外加入了多余的年龄核或时期核。贝叶斯因子均低于0.1，表明有充分证据反对将这些核视为生成数据集的合理备选项。这一实验证实了遗传算法恢复整体结构的能力，同时也表明，在区分 $M52_a$ 与$\mathrm{RBF}_a$时存在可以理解的困难。表5右侧面板中的SYA-2考虑了相同人口分布下另一种实际实现的死亡率曲面，以评估遗传算法在不同运行之间的稳定性。结果呈现出类似的情况，不过有趣的是，现在BIC最低的核变成了$\mathrm{M52}_a\mathrm{RBF}_y$，$\widehat{\ }$其 $BF\left(k, K_0\right) = 1.1907$。这意味着它的BIC低于真实核，展示了对数据过拟合的可能性。与此同时，由于贝叶斯因子非常接近1，这种差异在统计上仍不值得一提，因此完全可以归因于抽样变异。除此之外，我们再次观察到仅有两个真正合理的备选项，以及一些合理性明显较低的备选项（贝叶斯因子介于0.05和0.13之间）。

为进一步评估平滑性检测，将乘性年龄-时期核 $k = k_a k_y$ 拟合到 SA1 和 SA2 两个数据集，其中 $k_a$ 和 $k_y$ 依次取 M12,M32, M52 和径向基函数（按平滑性递增排序），每个训练曲面共得到 16 种组合。所得 BIC 差异列于表 6。请注意，真实值核 $K_0 = \mathrm{RBF}_a\mathrm{RBF}_y$生成的死亡率曲面在年龄和时期两个维度上都具有无穷阶可微性。在两种情况下，决定性证据始终表明，任一曲面都不包含 M12 分量（$\mathrm{BIC}(K_0)$ –$\mathrm{BIC}(k)$ 的范围为 $-96.99$ 至 $-30.24$；请记住，低于 -4.61 的值表示有利于 $K_0$的决定性证据）。如上所述，对于 SYA-1 和 SYA-2，$\mathrm{M52}_a$ 都是 $\mathrm{RBF}_a$ 的合理替代。对于 SYA-1，只有强有力的证据反对 $\mathrm{M52}_y$ （而 SYA-2 则是决定性证据）。在这两种情况下，都有决定性证据反对 $\mathrm{M32}_y$；使用 $\mathrm{RBF}_y$ 时，只有强有力的证据反对年龄分量 $\mathrm{M32}_a$ （否则证据是决定性的）。这种差异可能是因为 $\mathrm{M32}_a$ 具有额外的灵活性，能够捕捉一些波动，而这些波动通常由年龄维度上的参数化均值函数来拟合。

**SYB**

接下来，我们讨论 SYB，其中训练曲面由 $K_0 = 0.08 \cdot \mathrm{RBF}_a\,\mathrm{M12}_y + 0.02 \cdot \mathrm{M52}_c$ 生成。表7显示，共发现四种合理的核（其贝叶斯因子均高于

| SYB | | | | | |
|---|---|---|---|---|---|
| BIC | $\widehat{\text{贝叶斯因子}}(k,\ K_0)$ | 核 | $\hat{\sigma}^2$ | $\hat{\beta}_0$ | $\hat{\beta}_{ag}$ |
| -2468.0 | 1.033*** | $0.014 \cdot$ 径向基函数$_a$(0.574) Min$_y$(4.4) $+ 0.018 \cdot$ M52$_c$(0.083) | $3.52 \cdot 10^{-4}$ | $-4.249$ | 3.264 |
| **-2467.8** | **1.000*** ** | $\mathbf{0.063 \cdot}$ **径向基函数$_\mathbf{a}$(0.585)M12$_\mathbf{y}$(8.937)** $\mathbf{+ 0.0180 \cdot}$ **M52$_\mathbf{c}$(0.084)** | $\mathbf{3.448 \cdot 10^{-4}}$ | $\mathbf{-4.145}$ | **3.332** |
| -2465.9 | 0.161** | $0.060 \cdot$ 径向基函数$_a$(0.574)M12$_y$(8.267) $+ 0.016 \cdot$ 径向基函数$_c$(0.051) | $3.60 \cdot 10^{-4}$ | $-4.249$ | 3.264 |
| -2465.7 | 0.127** | $0.014 \cdot$ 径向基函数$_a$(0.574) Min$_y$(4.3) $+ 0.016 \cdot$ 径向基函数$_c$(0.051) | $3.60 \cdot 10^{-4}$ | $-12.01$ | 3.264 |
| -2464.7 | 0.047* | $0.022 \cdot$ 径向基函数$_a$(0.591) Lin$_y$(2.8)M12$_y$(10.480) $+ 0.016 \cdot$ M52$_c$(0.083) | $3.45 \cdot 10^{-4}$ | $-4.312$ | 3.298 |

表7：第二个合成案例研究 SYB 中拟合度最高的五个非重复项核。粗体表示数据生成中使用的真实核 $K_0 = 0.08 \cdot$径向基函数$_a$(0.586) $\cdot M12_y$(13.33) $+ 0.02 \cdot M52_c$(0.079)，其中 $\sigma^2 = 4 \cdot 10^{-4}, \beta_0 = -5.568,$ $\beta_1 = 2.974$。

0.1）。GA 以略有不同的超参数发现了真实值（表7中以粗体标出），以及三个密切相关的替代项。具体而言，这四个拟合度最高的核都识别出了正确数量的年龄-时期-队列项，即一个乘性的年龄-时期项加上一个加法的队列项。四者都找回了 RBF$_a$ 项，且相应的年龄长度尺度与真实的 $\ell^a_{\text{len}} = 0.586$非常接近。队列效应通过真实值 M52$_c$ （排名前两位的候选项）或 RBF$_c$捕捉到，这与我们在上文对 SYA 观察到的替代现象类似。相应的长度尺度估计准确，与真实的 $\ell^c_{\text{len}} = 0.079$ 相差不超过 0.3。时期效应通过真实值 M12$_y$ 或 Min$_y$捕捉到。生成数据的 $K_0$具有很大的 $\ell^y$ 长度 $= 13.33$，对于 M12$_y$ 项而言，这在去标准化后再变换时对应于一个一阶自回归过程，其中 $\phi_0 = 0.9975$ 。在 30 年的训练范围内，这与采用 Min 核的随机游走几乎无法区分。因此，以 Min$_y$ 替代并不令人意外，估计的 $\ell_{\text{len}}$范围较宽也在意料之中。例如，第二行中的 $\hat{\ell}^y$ 长度 $= 8.933$ 对应的 AR(1) 持久性参数为 $\phi = 0.9960$，与真实的 $\phi_0$非常接近。此外，我们还观察到所有非核超参数（$\sigma^2$、$\beta_0$、$\beta_{ag}$）都能稳定地恢复，估计值接近其真实值。

最后，GA 在学习年龄-时期分量（系数 0.08）支配队列分量（系数 0.02）方面也非常成功，并保留了两个分量的相对振幅。请注意，Min 和 Lin 核包含一个偏移量，从数学上看，这会产生与真实值 0.08 和 0.02 相似的线性系数。例如，第一行中的第一项可简化为 $0.0141 \cdot \text{RBF}_a \cdot (4.42 + x_{yr} \wedge x'_{yr}) = 0.0623 \cdot \text{RBF}_a + 0.0141 \cdot \text{RBF}_a \cdot (x_{yr} \wedge x'_{yr})$。因此，表7中的全部五种核都发现，第一个分量比第二个分量大 3.5–4 倍。

## SYC

表 8 展示了 SYC 的结果。左侧面板（在 $\mathcal{K}_r$上的结果）展示了在 Meh$_c \notin \mathcal{K}_r$ 的情况下，高斯过程如何通过其他基核近似$K_0 =$ M52$_a$ Min$_y$(M12$_c$Meh$_c$)。多个特征都得到了完美恢复：年龄成分中，M52$_a$ 是唯一出现的年龄项；乘性结构中，核的数量正确（四个）；以及队列效应中，存在 M12$_c$ 。Meh$_c$ 项被 RBF$_c$ 或 M52$_c$取代。请注意，在 $\mathcal{K}_r$ 中找到的最佳 $\widehat{\ }$核的贝叶斯因子仅为 $BF(k,\ K_0) = 0.1930$，这表明这种替代并不完美，而且 Meh 相较于 RBF 所引入的非平稳性类型在统计上显著。

观察右侧面板，M52$_a$M12$_y$M12$_c$M52$_c$ 明显胜出，它仅将 Min$_y$ 误识别为其他项，而非 M12$_y$，这与 SYB 中观察到的类似替换相同。该核的贝叶斯因子高于 0.83；根据贝叶斯范式，对于训练数据而言，它比第二行中的次选项（将 M12$_c$ 换成 Min$_c$，以及 Meh$_c$）更有可能，概率高出 $3.75 = 0.8321/0.2218$ 倍。

| SYC | | | | | |
|---|---|---|---|---|---|
| 受限搜索集 $\mathcal{K}_r$ | | | 完整搜索集 $\mathcal{K}_f$ | | |
| BIC | $\widehat{\text{贝叶斯因子}}(k,\ K_0)$ | 核 | BIC | $\widehat{\text{贝叶斯因子}}(k,\ K_0)$ | 核 |
| -2785.35 | 0.1930** | $\mathrm{M52}_a\,\mathrm{Min}_y\,\mathrm{M12}_c\,\mathrm{RBF}_c$ | -2786.82 | 0.8321*** | $\mathbf{M52}_a\mathrm{M12}_y\mathbf{M12}_c\mathbf{Meh}_c$ |
| -2785.06 | 0.1444** | $\mathrm{M52}_a\,\mathrm{Min}_y\,\mathrm{M12}_c\mathrm{M52}_c$ | -2785.49 | 0.2218** | $\mathrm{M52}_a\mathrm{M12}_y\,\mathrm{Min}_c\,\mathrm{RBF}_c$ |
| -2784.99 | 0.1339** | $\mathrm{M52}_a\mathrm{M12}_y\mathrm{M12}_c$径向基函数$_c$ | -2785.41 | 0.2046** | $\mathrm{Chy}_a\mathrm{Min}_y\,\mathrm{M12}_c\,\mathrm{Meh}_c$ |
| -2784.73 | 0.1032** | $\mathrm{M52}_a\mathrm{M12}_y\mathrm{M12}_c\mathrm{M52}_c$ | -2785.21 | 0.1671** | $\mathrm{Chy}_a\mathrm{M12}_y\mathrm{M12}_c\,\mathrm{Meh}_c$ |
| -2784.21 | 0.0615* | $\mathrm{M52}_a\mathrm{Min}_y$ 径向基函数$_y\,\mathrm{M12}_c$ | -2784.87 | 0.1185** | $\mathrm{Chy}_a\mathrm{M12}_y\,\mathrm{Meh}_y\,\mathrm{M12}_c$ |

表 8：在两次独立运行中针对 SYC 找到的最佳非重复核，一次针对 $\mathcal{K}_r$，另一次针对 $\mathcal{K}_f$。真实核为 $K_0 = \mathrm{M52}_a\,\mathrm{Min}_y(\mathrm{M12}_c\,\mathrm{Meh}_c)$。注意，$\mathrm{Meh}_c \notin \mathcal{K}_r$。

到 $\mathrm{RBF}_c$）。与左侧面板一样，这有助于确立 Meh 的非平稳性的重要性。除此之外，结果与 $\mathcal{K}_r$ 相似，只是加入了 $\mathrm{Chy}_a$ 来替代 $\mathrm{M52}_a$。有趣的是，这次 GA 运行避开了真实值 $K_0$，尽管通过对其他表现最佳的核进行一次交叉或非子树变异，就能得到它。这表明，GA 未能在 20 代内充分探索核空间，这可能受到包含 Min 时超参数估计收敛缓慢的影响。更多讨论请参见附录 B。

SYC 合成曲面还展示了 GA 识别总项数$\widehat{\quad}$ $\widehat{\quad}$在 $K_0$ 中的能力，以及识别加性结构的能力。使用简写 $\mathrm{BF} = \mathrm{BF}(k, K_0)$，GA 找到：

- 具有三个分量（而非四个）的最佳核是 $\mathrm{Chy}_a\mathrm{M12}_y\mathrm{M12}_c$ (贝叶斯因子 $= 0.0206$)$\widehat{\quad}$，位于 $\widehat{\mathcal{K}_f}$；而 $\mathrm{M52}_a\,\mathrm{M12}_y\mathrm{M12}_c$ (贝叶斯因子 $= 0.0182$)位于 $\mathcal{K}_r$。因此，BIC 准则使遗传算法正确地拒绝了这些核，认为它们对于 SYC 数据集而言过短；
- 包含五个分量的最佳核是 $\mathrm{M52}_a(\mathrm{M12}_y\mathrm{M12}_y)(\mathrm{M12}_c\mathrm{Meh}_c)$（贝叶斯因子 $= 0.0108$），$\widehat{\quad}$位于 $\mathcal{K}_f\widehat{\quad}$；而在 $\mathcal{K}_r$ 中，最佳核是 $\mathrm{M52}_a\mathrm{M12}_y\mathrm{Lin}_y\mathrm{M12}_c\mathrm{M52}_c$（贝叶斯因子 $= 0.0108$）。同样，遗传算法及其 BIC 准则会正确惩罚过长的核。
- 带有加性成分的最佳核是 $\mathrm{M52}_a(\mathrm{M12}_y + \mathrm{M12}_c\mathrm{RBF}_c)$（贝叶斯因子 $= 0.0070$）$\widehat{\quad}$以及 $\widehat{\mathrm{Chy}_a}\mathrm{Meh}_c(\mathrm{M12}_c + \mathrm{M12}_y)$（贝叶斯因子 $= 0.0067$），位于 $\mathcal{K}_r$。这有力地证明，贝叶斯信息准则能够正确识别单个加法项的存在。

## 5 人类死亡率数据库数据的结果

在使用合成数据验证了我们的生成式核探索之后，我们转向现实的实证分析。我们的讨论侧重于回顾性分析，即根据不同核在拟合给定训练集时的表现来评估它们。因此，我们不追求样本外指标，例如用于衡量预测准确度的（概率）分数，而是专注于结合定性比较来考察贝叶斯信息准则（BIC）分数。回顾性评估与 APC 方法的核心思路相呼应，即在第二步引入样本外动态之前，先通过奇异值分解对数据矩阵进行分解。选择这种方法的另一个原因是，预测准确度面临诸多挑战（例如如何处理近期由 COVID 导致的 2020年或2021年死亡率等特殊数据），而且目前尚无公认的标准评估方法。在

相比之下，BIC 为 GP 提供了一个单一且“简洁”的统计拟合度量，而且还与贝叶斯因子（BF）解释相联系，即相对证据优势。

作为我们的第一个案例研究，我们考虑日本女性人口群体。我们使用涵盖50—84岁和1990年至2018年的HMD数据集。遗传算法的顶层超参数和交叉/变异超参数与上一节相同（分别见表2和表3）。

表9提供了 $\mathcal{K}_f$ 中排名靠前且 BIC 分数最低的核的汇总统计量。为了给出具有代表性的横截面汇总，我们考察排名前10、前50，以及第51-100、第101-150和第151-200名最佳核的统计量。请注意，GA 一共提出了 $20 \cdot 200 = 4000$ 个核，因此排名前200的核对应于表现最佳的5%组合。我们发现，拟合效果最佳的是纯乘性核函数（单个加性成分），包含3个或4个项。这包括年龄、时期和队列坐标各一个项，以及可能的第4个项，通常位于队列或年份坐标中。在受限类别 $\mathcal{K}_r$ 中，排名前10的核均属于这种形式；在 $\mathcal{K}_f$ 中找到的排名前10的核也有9个属于这种形式。

| 范围 | BIC max | BIC min | len | 加法 非数值 组成统计 | | num age | num year | num coh | 粗糙 age | 粗糙 year | 粗糙 coh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 日本女性 | | | | | | | | | | | |
| 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% |
| 日本女性重新运行 | | | | | | | | | | | |
| 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% |
| 在 $\mathcal{K}$ 中搜索日本女性 $_r$ | | | | | | | | | | | |
| 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% |
| 基于 $\mathcal{D}_{rob}$训练的日本女性 | | | | | | | | | | | |
| 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% |

表9：针对日本女性在 $\mathcal{K}_r$ 和 $\mathcal{K}_f$上的重新运行与稳健性检验

上述APC结构还在年份和队列中包含一个粗糙（不可微或仅一次可微）成分。这与时间序列模型关于死亡率随时间演变的逻辑相符。请注意，在我们的设定中，这可以解读为不同年龄间观测噪声的强相关性，换言之，存在环境扰动（流行病、热浪及其他共病因素），导致死亡率逐年受到特异性影响。另一方面，年龄的最佳拟合是平滑的，最常见的是通过M52核实现。这与死亡率的年龄结构是平滑函数这一直觉相符。表9记录了强烈且明确无疑的队列效应：在所有排名靠前的核中均有体现，占比100%。这与Willets [43]的观点一致。

他指出，日本女性人口的强烈队列效应可以外推至更高年龄。

存在多个因子（长度大于 3）通常表明以下情况之一或两者兼有：(i) 多尺度依赖结构；(ii) 模型设定错误。一方面，由于我们只考虑少数几类核族，如果真实的相关结构无法由其中任何一类匹配，算法就会用现有核的组合来替代。例如，在年份维度同时使用粗糙核和平滑核，表明 M12 和径向基函数都无法单独实现良好拟合。另一方面，存在加性结构，或一般而言需要许多项（尤其是超过 5 项），说明数据的相关结构具有许多特征，因此无法用任何简单的方式描述。

在日本女性数据中，GA 对简约性的偏好得到了证实：表现最佳的核长度最短。我们观察到一个普遍规律：长度随排名上升而增加。具体而言，排名逐渐靠后时，我们开始看到包含两个加性成分的核。由此可以得出，第二个加性成分只带来了轻微的拟合改进，而这一改进不足以抵消复杂度惩罚，因此基于简约性原则，该成分会被舍弃。

整体贝叶斯信息准则最低的组合核是

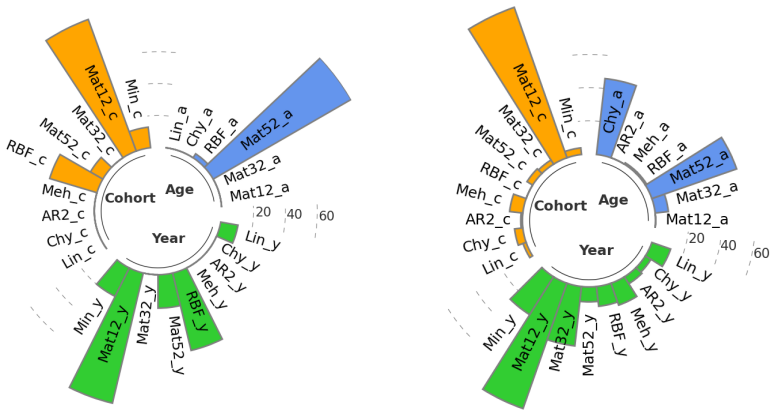
$$k^*_{JPN-FEM} = 0.4638 \cdot M52_a(1.11) \cdot \mathrm{Chy}_y(1.95) \cdot M12_y(62.42) \cdot M12_c(117.11).$$

请注意， $k^*_{JPN-FEM}$及其两个时期项具有纯乘性结构，既捕捉了局部的粗糙特性，也捕捉了较长范围的依赖关系。表10列出了次优的备选方案，分别来自 $\mathcal{K}_f$ 和 $\mathcal{K}_r$。我们发现，将范围限制在较小的 $\mathcal{K}_r$内，损失极小：在 $\mathcal{K}_f$ 中找到的五个排名靠前的核里，有三个实际上属于 $\mathcal{K}_r$。因此，扩大搜索范围并不会改善贝叶斯信息准则，这表明新增至 $\mathcal{K}_f$ 的大多数核选项，要么与 $\mathcal{K}_r$ 中的基础核十分相似，要么并不能针对性地改善人类死亡率数据库数据。事实上，根据表9，贝叶斯因子改进倍数仅为 exp(0.19)。此外，我们还发现，不同排名靠前的核之间具有很强的超参数稳定性。例如，年龄的长度尺度（在表10中始终通过 M52 核捕捉）始终处于范围 $[1.09, 1.13]$ 内。同样，队列中 M12 核的长度尺度也处于较窄的范围 $[95, 120]$ 内。

| 1990—2018年50—84岁日本女性人类死亡率数据库数据集 | | | | | |
|---|---|---|---|---|---|
| $\mathcal{K}_r$ | | | $\mathcal{K}_f$ | | |
| BIC | $\widehat{\mathrm{BF}}$ | 核 | BIC | $\widehat{\mathrm{BF}}$ | 核 |
| -2725.288 | 0.995 | $\mathrm{M52}_a$(径向基函数$_y$ $\mathrm{M12}_y$)$\mathrm{M12}_c$ | -2725.293 | 1 | $\mathrm{M52}_a$(Chy$_y$ $\mathrm{M12}_y$)$\mathrm{M12}_c$ |
| -2725.270 | 0.977 | $\mathrm{M52}_a$($\mathrm{M52}_y\mathrm{M12}_y$)$\mathrm{M12}_c$ | -2725.270 | 0.977$^\dagger$ | $\mathrm{M52}_a$($\mathrm{M52}_y\mathrm{M12}_y$)$\mathrm{M12}_c$ |
| -2725.233 | 0.941 | $\mathrm{M52}_a$(径向基函数$_y$ Min$_y$)$\mathrm{M12}_c$ | -2725.221 | 0.931$^\dagger$ | $\mathrm{M52}_a$($\mathrm{M52}_y$ Min$_y$)$\mathrm{M12}_c$ |
| -2725.221 | 0.931 | $\mathrm{M52}_a$($\mathrm{M52}_y$ Min$_y$)$\mathrm{M12}_c$ | -2724.623 | 0.512$^\dagger$ | $\mathrm{M52}_a$($\mathrm{M52}_y\mathrm{M12}_y$) Min$_c$ |
| -2724.640 | 0.520 | $\mathrm{M52}_a$($\mathrm{M52}_y\mathrm{M12}_y$) Min$_c$ | -2724.510 | 0.457 | $\mathrm{M52}_a$($\mathrm{M32}_y\mathrm{M12}_c$)$\mathrm{M12}_c$ |

表10：两次独立运行中针对日本女性拟合最好的非重复项核，一次针对 $\mathcal{K}_f$ 和$\widehat{\ }$，另一次针对 $\mathcal{K}_r$。贝叶斯因子 BF 是相对于找到的最佳核 $k^*_{JPN-FEM}$而言的，且显著性均不明显。 $\mathcal{K}_f$ 列下带匕首标记的核也属于 $\mathcal{K}_r$

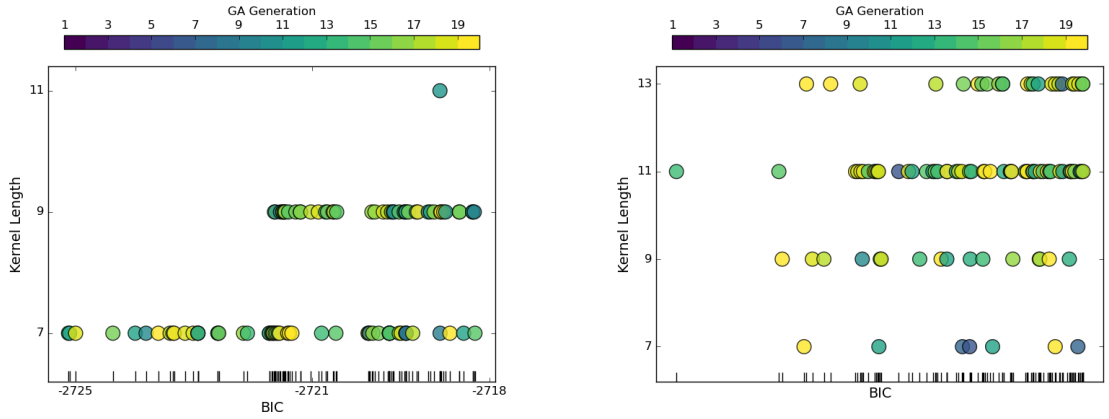
图2展示了不同核出现的频率。我们考察遗传算法返回的前100个唯一核，并显示每个所示核在返回的组合核中出现的次数。请注意，有时同一个核会出现多次。左侧面板考察遗传算法在 $\mathcal{K}_r$ 中搜索的情况，因此未考虑许多核族；右侧面板则考察完整的 $\mathcal{K}_f$。




$\mathcal{K}_r$ $\mathcal{K}_f$

图2：不同核在日本女性模型中的出现频率。

图2中的条形图表明，在考察前100个提案时，所选核存在显著差异。不过，我们仍能观察到典型的分解形式：若干“主要”因素，例如日本女性的$\mathrm{M12}_c$和$\mathrm{M12}_y$，以及额外的残差核。后者会产生二阶效应，且不易识别，因此有多种核会出现在5%至15%的提案中。例如，时期中的几乎每个核族都可用于构造良好的组合核。所选核的这种异质性表明，对于给定数据集，谈论“唯一的”高斯过程模型并不恰当，因为有多个差异颇大的拟合都能取得良好效果。




日本女性 美国男性

图3：遗传算法找到的排名前100个核的性质。

图3展示了若干拟议核的汇总统计量及其贝叶斯信息准则分数。我们展示排名前100的唯一核，并按总长度（y轴）和发现代数（颜色）排列。首先可以看到，沿着BIC排序（x轴）向下，核的聚集程度逐渐增加。换言之，通常有少数（1–5个）表现最佳的核，随后随着BIC分数降低，出现越来越多同样优秀的备选项。这与贝叶斯因子的解释相符：将表现最佳的核视为“真相”，我们会发现几个合理备选项、二十来个有一定合理性的备选项，以及几十个

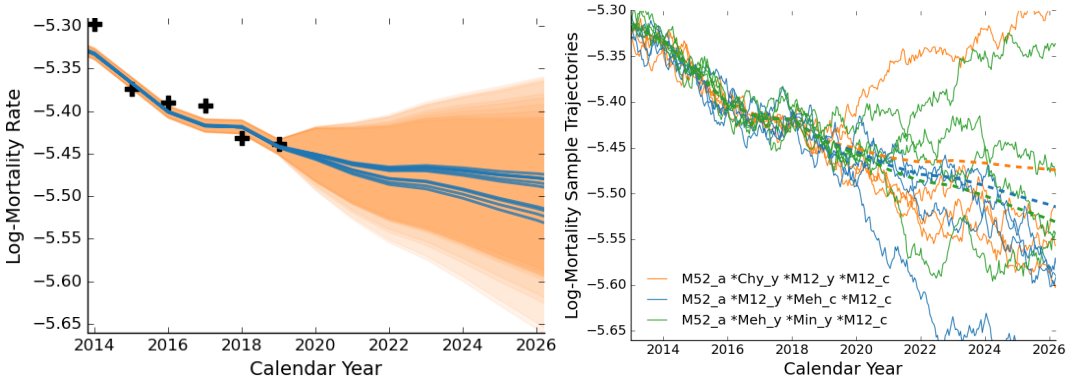



图4： $\mathcal{K}_f$ 中针对日本女性，65岁的排名前10位核函数的预测。左图：排名前10位核函数的预测均值和90%后验区间。作为比较，我们还展示了2014–2019年间观测到的5个对数死亡率（黑色加号）。右图：来自3个代表性核函数的4条样本路径。

合理性较弱的备选项。各群体的贝叶斯因子分布各不相同；有些情况下只有 $\sim 50-60$ 个合理备选项，另一些情况下则远超一百个。

其次，我们观察到，大多数最佳核函数是在经历 $10+$ 代后找到的，这与遗传算法的探索逻辑相符：逐步缩小搜索范围，找到适应度最高的核族。不过，这种模式并不十分明显，偶尔也会较早发现最佳核函数。

第三，我们观察到核复杂度与适应度之间存在一种模式，这与上述逻辑相符：表现最佳的核函数往往长度相同（彼此也非常相似，通常只相差一次变异）；但在考察（相对而言）合理的备选项时，我们既能找到更简约的核函数，也能找到更复杂的核函数。这体现了简约性权衡：较短的核函数似然较低，但复杂度惩罚较小；较长的核函数对数似然较高，但受到的惩罚也更大。

在图 4 中，我们展示了日本女性，65岁的死亡率对数的样本内拟合和未来预测。左侧面板使用排名前 10 的核，并给出其后验均值和预测区间。样本内拟合受到严格约束，而随着我们逐渐远离训练样本，样本外预测变得更加异质。具体而言，出现了将核分为不同组的双峰预测：一些核预测未来死亡率将改善，另一些则预测下降趋势会有所减缓。叠加的 90% 后验预测区间表明，未来死亡率轨迹存在一个共同区域；随着日历年递增，区间向两侧对称扩展，而在更远的未来年份，略微偏向较低的死亡率。这种预测方法可作为贝叶斯模型平均的基础，并以贝叶斯因子作为权重。此外，我们还观察到方差呈平方根式扩展，这在随机游走死亡率模型中很常见。

图的右侧面板通过使用三个具有代表性的核模拟路径，研究高斯过程（GP）的随机性。样本内路径紧密聚集在其后验均值附近，难以明显偏离观测数据。轨迹中观察到的粗糙性是由于在日历年或队列中加入了 M12 或 Min 分量所致。在考察样本外轨迹时，单个核的影响变得更加明显，尤其是在绿色和蓝色路径中。这些路径中缺乏均值回归的现象更为显著，这可能归因于日历年中的非平稳 Min 核（绿色）以及队列中的 Mehler 核（蓝色）。

### 5.1 稳健性检验

为验证上述遗传算法（GA）的结果，我们进行了两项检验：（i）从头重新运行该算法，以验证不同遗传算法运行结果的稳定性；（ii）在扩展数据集 $\mathcal{D}_{\text{rob}}$ 上运行遗传算法，即增加 2 个日历年（起始年份由 1990 年提前至 1988 年），以及 2 个额外的年龄组（范围由 50–84 岁扩大至 48–86 岁）。

在这两种情况下，我们预期结果与上文讨论的“主”运行非常相似。虽然 GA 会进行随机排列，且初始化是随机的，但我们预计，每代使用 200 个核、运行 20 代，足以让 GA 充分探索，因此最终表现最佳的核在不同 GA 运行中应保持不变。这是我们接受 GA 输出作为“真正”最佳拟合核的首要依据。同样，尽管 BIC 指标取决于具体数据集，但我们预计，即使数据集略有改动，该指标也足够稳定，因此可以将最佳拟合核视为适用于相关人口群体的正确核，而非仅适用于所选取的特定数据子集。

上述稳健性检验汇总如下，并列于表9。这些检验证实，GA 无论在自身的多次运行之间（见“重新运行”条目），还是在训练数据集略有改动时，都具有稳定性。我们发现，得到的核基本相同（在$\mathcal{K}_r$ 和 $\mathcal{K}_f$中均如此），此外，随着数据集扩大，最佳核/超参数也高度稳定（见“稳健”条目）。这既体现在下文明确列出的顶级核上，也体现在规模更大的前100个核集合中，详见表9中的汇总统计量。令人欣慰的是，在更大的数据集上运行时，下文列出的核长度尺度变化极小，这证实了最大似然估计 GP 子程序的稳定性。具体而言，在 $\mathcal{K}_r$的重新运行中，同一个核被确定为最佳核；在 $\mathcal{D}_{\text{rob}}$中它再次成为最佳核，只是参数略有调整：

$$\begin{aligned}
\text{原始 } \mathcal{D}\text{:}\quad & 0.4651 \cdot \mathrm{M52}_a(1.11) \cdot \mathrm{M52}_y(1.80) \cdot \mathrm{M12}_y(62.79) \cdot \mathrm{M12}_c(117.65);\\
\text{扩大后 } \mathcal{D}_{\text{rob}}\text{:}\quad & 0.4646 \cdot \mathrm{M52}_a(1.11) \cdot \mathrm{M52}_y(1.80) \cdot \mathrm{M12}_y(62.72) \cdot \mathrm{M12}_c(117.50).
\end{aligned}$$

在使用 $\mathcal{D}_{rob}$ 时，五个最佳核函数中有四个会重复出现。这种稳定性与 Cairns等人 [6] 的评论形成对比，他们指出基于 SVD 的拟合对日期范围较为敏感。其他替代方案仍遵循熟悉的替换模式。值得注意的是，在对 $\mathcal{D}_{\text{rob}}$ 进行的运行中出现了 $\mathrm{Chy}_a$，但没有出现 $\mathrm{Min}_y$,$\mathrm{M32}_y$ 或 $\mathrm{Meh}_c$。如图2所示，$\mathrm{Chy}_a$ 实际上相当常见。

此外，所有运行（原始运行、重新运行、扩大数据集）在年龄效应上始终选择 M52 或柯西核，在时期上选择 Mat´ern-1/2（有时为 Min），在队列上通常再搭配更平滑的核，如 $\mathrm{M32}_y$, $\mathrm{Meh}_y$,$\mathrm{Chy}_y$ 和 $\mathrm{M12}_c$。我们还记录到，前 100 个候选方案中不同核的出现频率非常相似，重新运行的贝叶斯信息准则分数也相近。

作为对遗传算法收敛性的另一项验证，图5展示了适应度分数随代数的演变。我们展示了第 $g$ 代中最佳核的贝叶斯信息准则，以及在遗传算法主要运行和一次“重新运行”过程中，次优核（200个核中的99%分位数）、第5优核（97.5%）、第10优核（95%）和第20优核（90%）的贝叶斯信息准则。每个面板中的实验仅因第一代使用了不同的初始种子而有所不同。在这两种设置下，根据最小贝叶斯信息准则，在大约第12代之后只能发现极小的性能提升。由于每一新代中提出的新核具有固有的随机性，随着新核被持续尝试，贝叶斯信息准则分数仅实现分布上的收敛。这种变化由相应代内贝叶斯信息准则分位数的平坦曲线所表明。总之，遗传算法在大约十几代后收敛至其“均衡”状态，从而验证了我们使用 $G = 20$ 进行分析的做法。

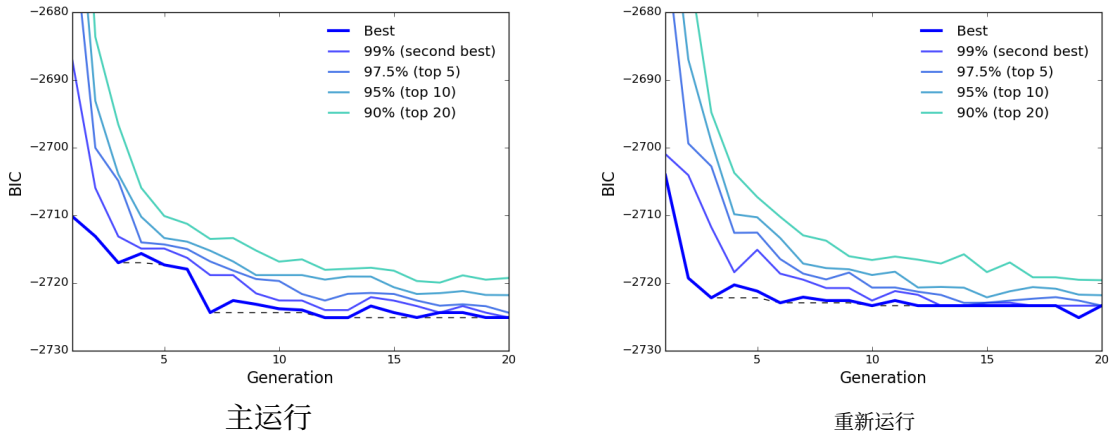



图5：遗传算法提出的最佳核函数汇总统计量随代数的变化 $g$。

## 5.2 模型残差分析

在图6中，左侧面板展示了日本女性实际对数死亡率与表现最佳的核所对应的高斯过程（GP）预测值之间的残差。残差中不存在任何可识别的结构，尤其是在对应出生队列的西南—东北方向对角线上，这表明拟合在统计学上是可靠的，与残差预期具有不相关性且服从相同分布的特征一致。此外，我们观察到明显的异方差性，即年龄较小时的残差表现出更高的方差。这是由于这些年龄段的死亡人数较少，导致推断出的死亡率具有更大的不确定性，尽管暴露量较大。总体而言，观测方差在80岁左右达到最低。

图6右侧面板显示了相对于单元格 (70, 2010年) 的先验相关性。明显的对角线形状表明队列效应十分重要。此外，我们观察到，相关性在时期（纵向）和年龄（横向）上的衰减程度大致相同，推断出的长度尺度使两个方向上的依赖范围均约为 $\pm 12$ 年。

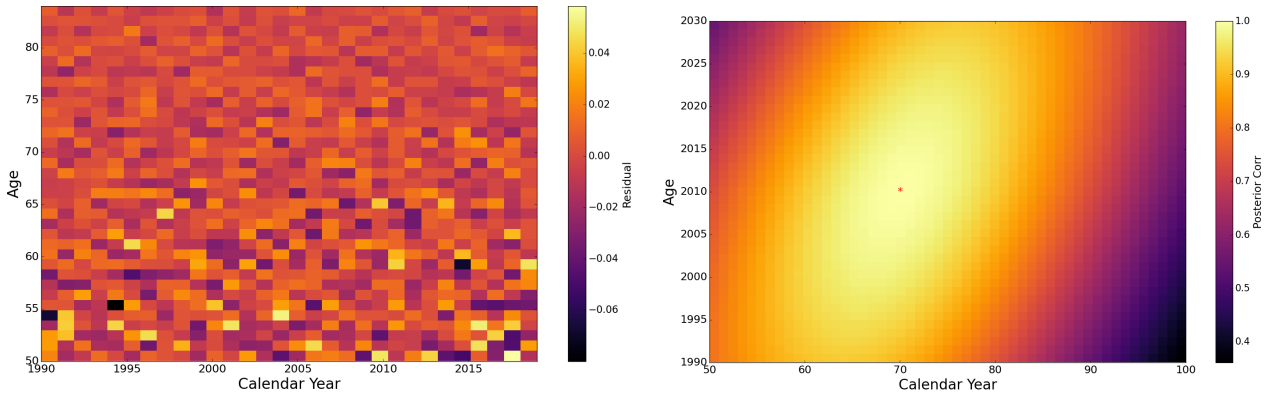



图6：右：最佳核函数隐含的先验相关性。左：$\mathcal{K}_f$ 中日本女性的最佳核函数残差。

### 5.3 男性与女性群体

接下来，我们将遗传算法应用于日本男性群体。进行这一比较的依据是假设两种性别之间具有相似的相关结构，这使我们能够突出相似之处并指出观察到的差异。

正如预期，日本男性的结果（见表11和表12）与日本女性的结果非常相似。我们再次发现，有力证据表明存在一个单一乘法项，其特征是具有年龄-时期-队列结构，年龄平滑，而时期和队列效应粗糙。与女性群体的情况一样，所有（100%）表现最佳的日本男性核中都选择了一个（粗糙的）队列项。表12所示的最佳单个核也很相似：年龄维度的最优选择为M52，年份维度为M52或径向基函数，队列维度则为M12。

也观察到了一些差异，例如日本男性比女性采用了更多队列相关核。女性的队列长度尺度大得多（120对40），粗糙M12 y长度尺度也更大（62对39），而年龄长度尺度几乎相同。一种解释是，男性死亡率中的特异性噪声更多，导致相关性衰减更快。

### 5.4 跨国家分析

为更全面地了解全球死亡率状况，接下来我们还将考察美国男性和瑞典女性。总计，我们分析 4 个数据集：美国男性、日本女性和男性、瑞典女性。需要指出的是，瑞典人口规模远小于另外两个国家（人口为 1000 万，而日本为 1.3 亿，美国为 3.3 亿），因此其数据噪声也大得多。

| 范围 | BIC max | BIC min | len | 加法非数值组成统计 | | age | num year | num coh | 粗糙 age | 粗糙 year | 粗糙 coh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 日本男性 | | | | | | | | | | | |
| 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% |
| 美国男性 | | | | | | | | | | | |
| 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% |
| 瑞典女性 | | | | | | | | | | | |
| 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% |

表11：遗传算法在日本男性、美国男性和瑞典女性数据上的运行结果。整个过程中，我们在完整集合 $\mathcal{K}_f$ 中进行搜索。

**美国男性：** 与所有其他数据集相比，美国男性数据得到的核长度要大得多。遗传算法返回的核包含 5-6 个基核，并且经常包含两个甚至三个加性项。此外，年龄-时期-队列模式受到了一定干扰，这可能是多个时期项和队列项之间存在共线性所致。

为了展示一些观测到的特征，我们来考察 $\mathcal{K}_f$ 中排名第一的核，如

| Pop'n/搜索集 | $N_{pl}$ | 最佳核 |
|---|---|---|
| 日本女性 $\mathcal{K}_r$ | 90 | $0.464 \cdot \mathrm{M52}_a(1.1) \cdot$ 径向基函数$_y(1.33)\mathrm{M12}_y(62.51) \cdot \mathrm{M12}_c(118.06)$ |
| 日本女性 $\mathcal{K}_f$ | 95 | $0.4638 \cdot \mathrm{M52}_a(1.11) \cdot \mathrm{Chy}_y(1.95)\mathrm{M12}_y(62.42) \cdot \mathrm{M12}_c(117.11)$ |
| 日本男性 $\mathcal{K}_r$ | 89 | $0.1491 \cdot \mathrm{M52}_a(0.95) \cdot$ 径向基函数$_y(1.15)\mathrm{M12}_y(26.24) \cdot \mathrm{M12}_c(24.90)$ |
| 日本男性 $\mathcal{K}_f$ | 112 | $0.2130 \cdot \mathrm{M52}_a(1.09) \cdot \mathrm{M12}_y(39.09) \cdot M32_c(0.86)\mathrm{M12}_c(40.73)$ |
| 美国男性 $\mathcal{K}_r$ | 57 | $0.017 \cdot \mathrm{M12}_a(5.04) \cdot \mathrm{M52}_y(0.50)M12_y(10.33) \cdot \mathrm{M52}_c(0.36)\mathrm{M12}_c(5.00)$ |
| 美国男性 $\mathcal{K}_f$ | 35 | $0.01 \cdot \mathrm{AR2}_a(1.12,1.88) \cdot \mathrm{M12}_y(24。18) \cdot \mathrm{M32}_c(0。72) \cdot [4.6211 \cdot \mathrm{M12}_c(13。49) + 0.01 \cdot \mathrm{M32}_a(0。02) \cdot \mathrm{M52}_c(0。1)]$ |
| 瑞典女性 $\mathcal{K}_r$ | 200+ | $0.2527 \cdot$ 径向基函数$_a(0.52) \cdot \mathrm{M12}_y(73.74) \cdot$ 径向基函数$_c(0.62)$ |
| 瑞典女性 $\mathcal{K}_f$ | 200+ | $0.2094 \cdot \mathrm{Chy}_a(1.05) \cdot \mathrm{M12}_y(67.27) \cdot \mathrm{Meh}_c(0.60)$ |

表12：在 $\mathcal{K}_r$ 和 $\mathcal{K}_f$ 中，针对所考虑的4个人口群体表现最佳的核。 $N_{pl}$ 表示BIC与最佳核相差不超过6.802的备选核数量，因此根据贝叶斯因子准则将其判定为"合理"。

表12所示。该核由11个项组成，其中包括2个加性项。不过，我们注意到第二项的系数显著较小，这表明它起到了"校正"项的作用，可能是为了捕捉一个不太突出、难以识别，且通常被主要项掩盖的特征。此外，我们还观察到，该核同时包含粗糙的队列项 $\mathrm{M12}_c$ 和较平滑的队列项 $\mathrm{M32}_c$ 及 $\mathrm{M52}_c$。这表明存在多尺度队列效应：少数特殊年份（例如1918—1919年西班牙流感暴发期间的出生年份）与代际模式（如婴儿潮一代与沉默的一代）相结合。与其他数据集不同，美国男性数据甚至包含一个 $\mathrm{RBF}_c$ 项。最后，年龄效应由 $AR2$ 核描述，这种情况在日本男性群体中也很常见。

继续往下看，还能找到一些长度较短且只有一个加性成分的核，例如 $0.0129 \cdot \mathrm{M12}_a(3.44) \cdot$(径向基函数$_y(0.63)\mathrm{M12}_y(7.89)) \cdot (\mathrm{M32}_c(0.43)\mathrm{M12}_c(3.88)$，其排名第四；以及长度为7 $0.0113 \cdot$ 的M $12_a(2.61) \cdot (\mathrm{M32}_y(0.7)\mathrm{M12}_y(6.53)) \cdot \mathrm{M12}_c(2.91)$，其排名第七。总的来说，对于美国男性，即使表现最佳的核长度为11，我们仍能找到长度为7、9、11和13的合理核。图3右侧面板和图7中间列展示了这种合理核长度的广泛分布（以及多种拟议核族）。

美国数据具有内在复杂性的另一个迹象是， $\mathcal{K}_f$ 中最佳核的贝叶斯因子与 $\mathcal{K}_r$中的相比差距很大，在所有人口群体中远高于其他数据集。因此，限定为 $\mathcal{K}_r$会显著降低拟合效果。事实上，我们观察到， $\mathcal{K}_r$ 中所有排名靠前的核都是纯乘性的（例如 $\mathrm{M12}_a \cdot \mathrm{M52}_y\mathrm{M12}_y \cdot \mathrm{M52}_c\mathrm{M12}_c$），这不太可能是该数据的正确结构，而且也表明难以捕捉各坐标上的相关性，导致出现多个时期项和队列项。在 $\mathcal{K}_f$ 中只找到 35 个合理备选项，比其他数据集少 3 到 5 倍。

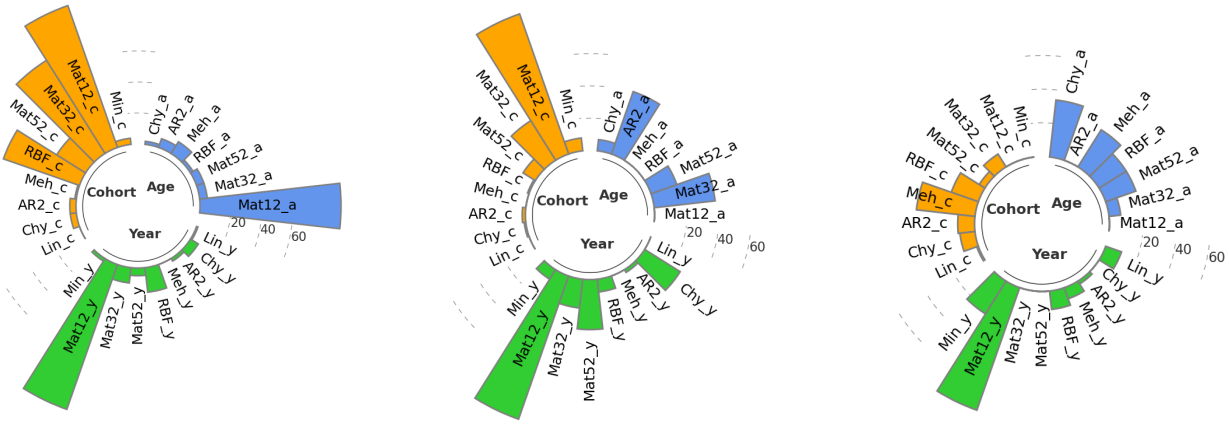
**瑞典女性：**瑞典女性数据集有两个显著特征。首先，它得到的核最简单、最短，且与由 3 个乘性项构成的 APC 结构直接吻合。表11所报告的平均核长度在瑞典女性数据集中最小，而且加性项极少出现。当提出包含 3 个以上项的核时，这些核通常仍然全是乘性的，并会增加第二个时期项（例如 $\mathrm{Meh}_a \cdot \mathrm{M12}_y \cdot \mathrm{Meh}_y \cdot \mathrm{Meh}_c$）或第二个年龄项，不过两者都是平滑的：（$\mathrm{Chy}_a\mathrm{RBF}_a$）· $\mathrm{M12}_y \cdot \mathrm{Meh}_c$）。

第二，与上述所有其他人口群体不同，瑞典的队列效应存在歧义。表现最佳的核中约有15%（在 $\mathcal{K}_r$中为30%）完全不包含队列项，而是

提出了两个时期效应项。包含队列效应的核采用的要么是径向基函数$_c$，要么是Meh$_c$ 项，这表明其中没有短期队列特征，只有代际特征。在 $\mathcal{K}_f$ 的前10个核中有9个包含队列项，而在 $\mathcal{K}_r$ 中仅有7个包含。相比之下，在日本和美国，每一个前100个核中都包含粗糙的队列项。我们的结果再次与文献一致；请参见Murphy [30]，其讨论了瑞典女性人口的队列效应缺乏明确性。

第三个观察结果是，瑞典女性的 BIC 值较为集中，即许多不同的核都被提出，而它们的 BIC 值非常相近。遗传算法返回了两百多个核，其贝叶斯因子与最佳核的比值在 30 以内。这表明，几乎没有证据能够区分许多不同的选择，而这可能是由瑞典死亡率数据较低的复杂度所致。

## 5.5 讨论




美国男性　　日本男性　　瑞典女性

图7：频率　　$\mathcal{K}_f$ 在美国、瑞典和日本男性模型中不同核的出现概率　　.

**最佳核族：** 图7中的条形图表明，没有明确的协方差结构能够拟合死亡率模式。因此，模型通常会提出各种核的组合（通常是乘积，有时是求和）。此外，对于不同人口群体而言，并不存在一种适用于所有情况的解决方案。例如，对于美国男性，年龄效应通常使用 M12 核建模；对于日本男性，通常使用 M32 核；对于瑞典女性，则通常使用 Chy（或径向基函数或 M52）核。此外，对于美国男性和日本男性，Chy 可能会被识别为一个可能的时期项，但对于瑞典女性则从未如此。因此，建议针对不同的案例研究选择不同的核。这与经典的年龄-时期-队列框架相比，是一个显著差异：在经典框架中，奇异值分解在不同数据集之间保持不变，研究人员必须像 Cairns等人 [6] 所示那样，手动测试大量组合。

**队列效应的必要性：**
为评估纳入出生队列项的影响，我们在排除所有队列专属核的情况下重新运行遗传算法。这只需对实现做简单调整，便可用于检验通过精心选择的年龄-时期核组合，是否足以解释队列效应。

首先，我们通过比较不包含队列项的最佳核与完整的 $\mathcal{K}_f$ 的贝叶斯信息准则（BIC）来评估我们的模型。此外，我们还检查残差热图，以检测是否存在明显的对角线模式。我们的研究结果表明，队列效应

对于美国男性、日本女性和日本男性而言必不可少。缺少队列项会导致 BIC 显著增加：日本男性增加 235，美国男性增加 198，日本女性增加 111。作为参考，只有当 BIC 差异超过 6.802 时，才认为其具有显著性。图8进一步证实了这些结果：底行（无队列模型）呈现出明显的对角线，这与残差相互独立的假设相矛盾。

对于瑞典，差异仅为 1.24，对应的贝叶斯因子为 0.2894，表明不具有显著性。此外，相应的无队列残差仍然令人满意，相关核 $0.1125 \cdot \text{Chy}_a(0.85) \cdot \text{Meh}_y(1.16) \cdot \text{M12}_y(37.74)$ 在原始结果中排名第九。再次说明，我们的结果与文献一致；参见 [30]，其讨论了瑞典队列效应缺乏明确结论的问题。

Cairns等人 [6] 提出，队列效应或许可以通过精心选择年龄和时期效应得到部分或完全解释。我们对此发现得到部分证实：在许多人口群体中，解释时期和队列依赖关系所用的核超过两个，且它们之间存在明显的替代关系。不过，这只是次要效应；除瑞典外，纳入队列的必要性是明确无疑的。对于瑞典女性，出生队列依赖关系的必要性相当弱。

**核替换：**

遗传算法通过其变异操作，自然地凸显了不同核族之间的替换效应。用一个核替换另一个核是遗传算法演化过程的内在特征。按照贝叶斯信息准则对核进行排名后，我们发现，许多组合的性能几乎相同，且仅相差一个项。上述效应在纯乘性核中尤为明显，这类核在除美国以外的所有人口群体中都很常见。在这种情况下，我们经常发现，其中一个项可以由两个或三个核族表示，而其余项保持不变。

我们发现，某些核经常互相替代，例如径向基函数核与 M52 核，以及 M12 核与 Min 核。尽管后一对核在平稳性方面有所不同，但当长度尺度参数 $\ell$ 较大时（通常为 $\ell_{\text{Min}} > 25$），Min 生成的样本路径在视觉上与 M12 生成的样本路径无法区分。这些替换模式与我们的合成结果一致，详见第 4 节。例如，在日本女性数据集上训练并在 $\mathcal{K}_r$中进行搜索时，排名第一的核形式为 $\text{M52}_a \cdot (\text{RBF}_y \cdot \text{M12}_y) \cdot \text{M12}_c$，而按贝叶斯信息准则排名第二的核为 $\text{M52}_a \cdot (\text{M52}_y \cdot M12_y) \cdot M12_c$。它保留了相同的宏观结构，同时替换了两个时期项中的一个；参见表10。此外，排名接下来的两个核非常相似，但它们用 Min 替换了 M12：排名第三的是 $\text{M52}_a \cdot (\text{RBF}_y \cdot \text{Min}_y) \cdot \text{M12}_c$，排名第四的是 $\text{M52}_a \cdot (\text{M52}_y \cdot \text{Min}_y) \cdot \text{M12}_c$。需要注意的是，替换时长度尺度（有时还有过程方差）会发生变化，因为它们在不同核族中的含义不同。例如，径向基函数核的长度尺度往往比其替代核 M52 的长度尺度小约 50%。

柯西核和M52也可以相互替代，尽管Chy和径向基函数的可互换性较低。瑞典女性的例子很好地展示了平滑核有多个替代项这一现象：排名靠前的四个核根据$\text{M12}_y \cdot \text{Meh}_c$确定时期效应和队列效应，然后为年龄效应提出$\text{Chy}_a$、 径向基函数$_a$、 $\text{Meh}_a$、 $\text{M52}_a$中的任意一种。在 $\mathcal{K}_f$中，替代现象更为常见，因为可用核族更多，从而加剧了“共线性”，也带来了更多替代机会。与此同时，我们偶尔也会发现，在 $\mathcal{K}_f$中可以找到更合适的核族。例如，我们经常在年份中同时观察到一个粗糙核和一个平滑核，这表明M12和径向基函数单独拟合时效果都不理想；在这种情况下，M32有时会成为更好的单一替代项。同样，$\text{Meh}_c$ 是瑞典女性最常见的选择，在 $\mathcal{K}_r$中则由径向基函数$_c$ 或径向基函数$_c \cdot \text{M12}_c$ 取代。因此， $\mathcal{K}_f$ 提出的核往往更短。

平均而言比来自 $\mathcal{K}_r$ 的那些。

时期项和队列项之间还存在进一步的替代效应。在日本女性和瑞典女性中，我们通常观察到总共三个项，并且在采用两个时期项和一个队列项，还是一个时期项和两个队列项之间存在替代关系。例如，在日本女性的前 10 个核中，我们同时发现了 $\mathrm{M52}_a \cdot \mathrm{M12}_y \cdot (\mathrm{Meh}_c \cdot \mathrm{M12}_c)$ 和 $\mathrm{M52}_a \cdot \mathrm{Meh}_y\, \mathrm{M12}_y \cdot \mathrm{M12}_c$。

**加法性与非平稳性：**

回到加性分量这一主题，我们的结果总体表明，加性结构通常较弱。具体而言，引入额外的加性分量往往只能略微改善拟合优度，而这一改善会被贝叶斯信息准则中的复杂度惩罚抵消，因此加性核的贝叶斯信息准则值更低。因此，遗传算法往往基于简约性而拒绝加性核。由此，大多数包含 4 个项（长度为 7）的核以及大多数包含 5 个项的核都是纯乘性的。

由于多种因素的影响，对非平稳性进行概括颇具挑战，因此，对于任何提供非平稳数据占比的表格，都应持审慎态度。图 2 和图 7 所示的频率图能够更全面地展示非平稳性。需要注意的是，拟合时 M12 的长度尺度往往较大。当 M12 的长度尺度较大时，生成的过程在视觉上与非平稳 Min 核生成的过程无法区分。因此，在我们的数据中，M12 的出现表明可能存在非平稳性。从这一角度来看，我们的研究结果表明，所有人口群体都表现出（可能的）非平稳时期效应。其他非平稳核（Lin、Meh）较为少见，但确实存在，且主要出现在瑞典女性人口群体中。

# 6 结论

我们的研究分析了使用遗传算法为死亡率曲面的高斯过程代理模型发现核（即协方差结构）的做法。遗传算法在我们的合成实验中表现出色，表明在使用高斯过程框架进行真实数据分析时，它有望成为模型选择与验证的工具。具体而言，它成功检测出了数据生成过程的平滑性，在不同样本（SYA）中表现稳健，区分了加法与乘性 APC 结构，识别出相对较小的队列效应（SYB），并找到了正确的基核数量，同时识别出时期和队列坐标上的多重非平稳性（SYC）。此外，所有实验都体现了“替代”效应，即一个核能够近似另一个核的影响。例如，长度尺度较大的 M12 核可以替代 Min 核，反之亦然。

应用于 HMD 数据集时，我们的结果强烈表明，最适合死亡率数据的是这样的高斯过程模型：年份和队列中包含粗糙（不可微或仅一次可微）的成分，而年龄项则是平滑的。这与经典假设相符：死亡率的年龄结构是平滑函数，而时间动态则类似随机游走。唯一的例外是美国数据，其中年龄结构被认为不可微，而队列项则是平滑的。

在非标准核中，我们发现柯西核经常入选：$\mathrm{Chy}_a$ 出现在瑞典女性排名前 100 的核中，占 34%；出现在日本女性排名前 100 的核中，占 28%；$\mathrm{Chy}_y$ 则出现在日本男性排名前 100 的核中，占 26%。迈勒核也有出现，不过除 $\mathrm{Meh}_c$ 在瑞典女性中较常见（排名前 100 的核中有 35 个）外，其他情况并不频繁。

历史数据分析和 SYC 实验都表明，难以明确确定单一的协方差结构。考虑到存在彼此模仿的替代核（例如，$\mathrm{Chy}_a$ 代替 $\mathrm{M52}_a$，或 $\mathrm{Meh}_c$ 代替 $\mathrm{RBF}_c$），以及

使用较大的搜索集 $\mathcal{K}_f$（核的数量翻倍）时所面临的搜索问题的复杂性，这并不令人意外。虽然更大的搜索集能更好地逼近真实情况，从而有利于贝叶斯信息准则优化，但要找到精确且富有表达力的协方差结构，就需要一个更小的 $\mathcal{K}$，其中包含能够表达建模者对底层数据生成过程先验信念的关键类别。要掌握这类知识颇具挑战，因为这需要具备应用领域（例如死亡率建模）以及高斯过程核性质方面的专业知识。

本文的重点之一是考察拟合效果最佳的模型所采用的核。与此同时，遗传算法的输出也支持模型平均分析，即同时使用多个模型。模型平均有助于深入了解模型风险，并增强预测对任意核函数选择中特有差异的稳健性。因此，它为贝叶斯高斯过程提供了一种高效的替代方案，如 [19] 中所采用的方法。由于篇幅所限，这部分内容留待另行研究。

未来研究还可将我们的方法应用于规模较小的人口群体（人口不足 1000 万至 2000 万）。与我们的数据不同，我们的数据中信噪比较高，因此遗传算法能够在有利条件下推断最佳潜在结构；对于较小群体，这可能需要进行重大调整，因为噪声较高时，遗传算法将更难区分候选核。一个解决方案是构建一个能够联合建模多个群体的版本，采用多输出高斯过程 [18, 19]。

此外，还可以考虑其他核。例如，柯西核是有理二次核的一个特例，其索引为 $\alpha$（见附录 A）。鉴于 Chy 在我们的结果中颇受欢迎，值得探索 $\alpha$ 超出 $\alpha = 1$ 的取值，或直接对 $\alpha$ 进行搜索。还可以在年龄和年份维度上的不可分核中纳入队列效应；当 $A$ 为正定时，$x_{yr} - x_{ag}$ 可以自然地出现在 $\exp(-[x_{ag}, x_{yr}]^\top A[x_{ag}, x_{yr}])$ 等表达式中，以定义核。

变点检测可以通过高斯过程自然实现，将 $\sigma(x)k_1(x,\ x')\sigma(x')+\sigma(x)k_2(x,\ x')\sigma(x')$用作核，其中 $\sigma(x) = 1 - \sigma(x)$，而 $\sigma(\cdot)$是一个激活函数，例如 sigmoid$\sigma(x) = 1/(1 + \exp(-x))$。当死亡率结构从一种状态非平稳地转变为另一种状态时，例如从较年轻年龄段转向较年长年龄段，或出现时间上的死亡率转变时，这种方法很有用。与其设定一个硬性阈值，不如设计能够自动探索这种可能性的核。

此外，精心选择的超参数约束有助于找到更好的结构。在加法情形下，额外项总是会通过引入一个额外的尺度系数来增加贝叶斯信息准则（BIC）的惩罚。可以考虑添加一个使用相同尺度系数的项（即因子分解），或者将现有尺度因子按固定幅度（例如 0.5 或 0.25）缩减。通过重复使用现有超参数，这样就不会增加贝叶斯信息准则（BIC）的惩罚。此外，还可以通过约束时期（Period）与队列（Cohort）的长度尺度之间的关系，进一步提高可解释性。

最后，仍需进一步研究目前构建的遗传算法的局限性。对合成实验进行更多分析，有助于阐明信噪比和数据集大小对遗传算法能否充分探索、从而找到拟合效果最佳的核的影响。同样，关于“把所有东西都一股脑加进来”是否总是更好的问题，我们留待未来研究探讨：也就是说，在可行范围内尽可能纳入多种不同的核，还是预先筛选会更有益。

## 参考文献


[1] R. J. Adler。 随机场的几何。SIAM，2010年。

[2] S. Azman 和 D. Pathmanathan。《李-卡特模型的广义线性模型框架：一项多国家研究》。《应用统计学杂志》，49(3):752–763，2022年。

[3] A. Berlinet 和 C. Thomas-Agnan。概率论与统计学中的再生核希尔伯特空间。Springer Science & Business Media，2011年。[4] N. Brouhns、M. Denuit 和 J. K. Vermunt。采用泊松对数双线性回归构建投影生命表。*Insurance: Mathematics and Economics*，31(3):373–393，2002年。[5] A. J. Cairns、D. Blake、K. Dowd、G. D. Coughlan、D. Epstein、A. Ong 和 I. Balevich。利用英格兰和威尔士及美国的数据对随机死亡率模型进行定量比较。*North American Actuarial Journal*，13(1):1–35，2009年。[6] A. J. Cairns、D. Blake、K. Dowd、G. D. Coughlan、D. Epstein 和 M. Khalaf-Allah。死亡率密度预测：对六种随机死亡率模型的分析。*Insurance: Mathematics andEconomics*，48(3):355–367，2011年。[7] D. A. Cole、R. B. Gramacy 和 M. Ludkovski。大规模随机模拟实验的局部代理建模。*Computational Statistics & Data Analysis*，第107537页，2022年。[8] D. Dittrich、R. T. A. Leenders 和 J. Mulder。网络自相关建模：用于检验（多个）精确假设和区间假设的贝叶斯因子方法。*Sociological Methods & Research*，48(3):642–676，2019年。[9] K. Dowd、A. J. Cairns 和 D. Blake。CBDX：Cairns–Blake–Dowd 系列中的主力死亡率模型。*Annals of Actuarial Science*，14(2):445–460，2020年。[10] D. Duvenaud。利用高斯过程进行自动模型构建。博士论文，剑桥大学，2014年。[11] D. Duvenaud、J. Lloyd、R. Grosse、J. Tenenbaum 和 G. Zoubin。通过组合核搜索发现非参数回归中的结构。载于国际机器学习会议，第1166–1174页。PMLR，2013年。[12] J. Gardner、G. Pleiss、K. Q. Weinberger、D. Bindel 和 A. G. Wilson。GPyTorch：采用 GPU 加速的黑箱矩阵－矩阵高斯过程推断。*Advances in Neural Information Processing Systems*，31，2018年。[13] A. Gelman、J. B. Carlin、H. S. Stern 和 D. B. Rubin。贝叶斯数据分析。Chapman and Hall/CRC，1995年。[14] M. G. Genton。机器学习中的核类别：统计学视角。*Journal of Machine Learning Research*，2(12月):299–312，2001年。[15] T. Hastie 和 R. Tibshirani。广义加性模型。统计学与应用概率专著。Chapman and Hall/CRC，1990年。[16] A. Hunt 和 D. Blake。构建死亡率模型的一般程序。*North AmericanActuarial Journal*，18(1):116–138，2014年。[17] A. Hunt 和 D. Blake。年龄／时期／队列死亡率模型中的可识别性。*Annals of Actuarial Science*，14(2):500–536，2020年。[18] N. Huynh 和 M. Ludkovski。多个群体中的死因死亡率联合模型。*arXiv* 预印本 *arXiv:2111.06631*，2021年。

[19] N. Huynh 和 M. Ludkovski。用于多个人口群体长寿建模的多输出高斯过程。*Annals of Actuarial Science*，15(2)：318–345，2021年。[20] P. J¨ahnichen、F. Wenzel、M. Kloft 和 S. Mandt。可扩展广义动态主题模型。载于国际人工智能与统计会议，第1427–1435页。PMLR，2018年。[21] H. Jeffreys。概率论。牛津大学出版社，1961年。[22] S.-S. Jin。使用基于树的遗传编程进行组合核学习以实现高斯过程回归。《结构与多学科优化》，62(3)：1313–1351，2020年。[23] M. Kanagawa、P. Hennig、D. Sejdinovic 和 B. K. Sriperumbudur。高斯过程与核方法：关于联系与等价性的综述。*arXiv*预印本 *arXiv:1807.02582*，2018年。[24] D. P. Kingma 和 J. Ba。Adam：一种用于随机优化的方法。*arXiv*预印本 *arXiv:1412.6980*，2014年。[25] M. D. Lee 和 E.-J. Wagenmakers。贝叶斯认知建模：实用课程。剑桥大学出版社，2014年。[26] R. Lee。用于死亡率预测的李-卡特方法及其各种扩展与应用。*North American Actuarial Journal*，4(1)：80–91，2000年。[27] M. Ludkovski、J. Risk 和 H. Zail。用于死亡率和改善因子的高斯过程模型。《*ASTIN*公报》：国际精算师协会期刊，48(3)：1307–1347，2018年。[28] S. Luke 和 L. Panait。遗传编程膨胀控制方法的比较。《进化计算》，14(3)：309–344，2006年。[29]F. G. Mehler。关于将任意多个变量的函数按拉普拉斯高阶函数 h¨oherer 展开的研究。1866年。[30] M. Murphy。重新审视出生队列效应对死亡率的主导作用。《人口与发展评论》，36(2)：365–390，2010年。[31] A. Nigri、S. Levantesi、M. Marino、S. Scognamiglio 和 F. Perla。融合深度学习的李–卡特模型。《*Risks*》，7(1)：33，2019年。[32] M. M. Noack 和 J. A. Sethian。用于领域感知高斯过程的先进平稳核和非平稳核设计。*arXiv*预印本 *arXiv:2102.03432*，2021年。[33] E. Parzen。一种时间序列分析方法。《数理统计年刊》，32(4)：951–989，1961年。[34] F. Perla、R. Richman、S. Scognamiglio 和 M. V. W¨uthrich。使用深度学习进行死亡率时间序列预测。《斯堪的纳维亚精算杂志》，2021(7)：572–598，2021年。[35] R. Poli、W. B. Langdon、N. F. McPhee 和 J. R. Koza。遗传编程指南。施普林格，2008年。[36] A. E. Renshaw 和 S. Haberman。用于死亡率降低因子的李–卡特模型的基于队列的扩展。*Insurance: Mathematics and Economics*，38(3)：556–570，2006年。

[37] R. Richman 和 M. V. W¨uthrich。李–卡特模型向多个群体的神经网络扩展。*Annals of Actuarial Science*，15(2)：346–366，2021年。[38] I. Roman、R. Santana、A. Mendiburu 和 J. A. Lozano。从初等数学表达式演化高斯过程核，用于时间序列外推。《神经计算》，462：426–439，2021年。[39] B. Sch¨olkopf、A. J. Smola 和 F. Bach。核方法学习：支持向量机、正则化、优化及其他。麻省理工学院出版社，2002年。[40] J. Shawe-Taylor、N. Cristianini 等。用于模式分析的核方法。剑桥大学出版社，2004年。[41] M. Sipper、W. Fu、K. Ahuja 和 J. H. Moore。研究遗传算法的参数空间。《生物数据挖掘》，11(1)：1–14，2018年。[42] A. Villegas、V. K. Kaishev 和 P. Millossovich。StMoMo：用于随机死亡率建模的 R软件包。载于第七届澳大拉西亚精算教育与研究研讨会，2015年。[43] R. C. Willets。队列效应：洞见与解释。《英国精算期刊》，10(4)：833–877，2004年。[44] C. K. Williams 和 C. E. Rasmussen。用于机器学习的高斯过程，第2卷。麻省理工学院出版社，马萨诸塞州剑桥，2006年。[45] A. M. Yaglom。 $n$维空间中与平稳随机过程相关的某些随机场类别。《概率论及其应用》，2(3)：273–320，1957年。

# A 核函数札记

除非另有说明，以下均假设 $x,\ x' \in \mathbb{R}$。本节利用这样一个事实：对于高斯过程 $f$，其导数 $f'$在均方意义下存在，当且仅当 $\frac{\partial^2 k}{\partial x \partial x'}(x,\ x')$ 存在 [1]。 ——

**平稳**

Mat´ern-1/2：是 Ornstein-Uhlenbeck（OU）过程 [3] 的协方差。OU 过程遵循线性均值回复随机微分方程；其路径连续且处处不可微。均值回复使依赖关系局部化，并被建议用于 [20] 捕捉小尺度效应。长度尺度 $\ell_{\text{len}}$ 控制均值回复的速率（数值越低，回复越快）。

重新参数化

$$k(x, x') = \exp\left(\frac{-|x - x'|}{\ell_{\text{len}}}\right) = \phi^{|x-x'|}, \qquad \phi = \exp(-1/\ell_{\text{len}}), \tag{11}$$

这表明，当 $x$ 为离散值时，M12 等价于一个持久性参数为 $\phi$ 的一阶自回归过程。特别是，较大的 $\ell_{\text{len}}$ （即 $\phi \simeq 1$）会模拟非平稳随机游走过程及其 Min 核，使样本路径能够大幅偏离其均值，并削弱平稳性。

AR2：与（连续 $x$）二阶自回归（AR(2)）过程相关的协方差核为 [33]

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

其中 $\omega^2=\gamma^2-\alpha^2>0$。值得注意的是，该核有两个参数，因此在遗传算法优化中会受到更高的贝叶斯信息准则惩罚。可以证明，对于所有 $x,x'\in\mathbb{R}$，$\frac{\partial^2 k}{\partial x\partial x'}(x,x')$ 都存在，但 $\frac{\partial^4 k}{\partial^2 x\partial^2 x'}(x,x')$ 不存在。因此，采用 AR2核的高斯过程具有一次（均方）可微性，但不具有二次可微性，即 $f\in C^1$。

为保持与 Matérn 核族的一致性，我们按 $\alpha=1/\ell_{\text{len}},\omega=\pi/p$ 重新参数化，并针对 $k(x,x;\ell_{\text{len}},p)=1$ 进行归一化，使得

$$k(x,x';\ell_{\text{len}},p)=\exp\left(-\frac{|x-x'|}{\ell_{\text{len}}}\right)\left\{\cos\left(\frac{\pi}{p}|x-x'|\right)+\frac{p}{\pi\ell_{\text{len}}}\sin\left(\frac{\pi}{p}|x-x'|\right)\right\},\tag{12}$$

其中，在重新参数化后，$\gamma^2=\omega^2+\alpha^2=\frac{1}{\ell_{\text{len}}^2}+\frac{\pi^2}{p^2}$. 通过三角恒等式可知，这与平稳离散时间AR(2)过程的协方差函数相同（写作 $f(x)=\phi_1 f(x-1)+\phi_2 f(x-2)+\epsilon(x)$），适用于复特征根的情况，即 $\phi_1^2+4\phi_2<0$，参数之间的关系为

$$\phi_2=-\exp(-2/\ell_{\text{len}}),\qquad \phi_1=2\cos(\pi/p)\sqrt{-\phi_2}.\tag{13}$$

柯西：该核具有厚尾和长程记忆，这意味着相关性不是按指数衰减，而是按多项式衰减，从而导致输入之间产生长程影响 [20]。Chy核函数定义为

$$k(x,x';\ell_{\text{len}})=\frac{1}{1+\frac{(x-x')^2}{\ell_{\text{len}}^2}}\tag{14}$$

这是一种特殊的】【。情况下，$\alpha=1$，有理二次核的 $k(x,\ x';\ \alpha,\ \ell_{\text{len}}):=\left(1+\frac{(x-x')^2}{\alpha\ell_{\text{len}}^2}\right)^{-\alpha}$.

由于式（14）中的 $k(x,\ x')$ 关于两个参数均为无限可微，因此相应的高斯过程也具有类似径向基函数核的性质。理解 Chy 的一种方式是：它是径向基函数核关于 $1/\ell_{\text{RBF}}^2$ 的边缘化版本，其中采用了速率为 $\ell_{\text{Chy}}^2/2$ 的指数先验：

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

**非平稳性**

线性：线性核将高斯过程与贝叶斯线性回归联系起来。具体而言，如果 $f(x)=\beta_0+\beta_1 x$ 其中 $x\in\mathbb{R}$，并且先验为 $\beta_0\sim N(0,\sigma_0^2)$、$\beta_1\sim N(0,1)$，那么$f\sim\mathcal{GP}(0,k_{\text{Lin}})$，也就是说，$k(x,x')=\sigma_0^2+x\cdot x'$。请注意，$k_{\text{Lin}}$ 可以进行缩放，从而得到关于 $\beta_1$的先验方差。

Mehler：非平稳迈勒 [29]核为

$$k(x,x';\rho)=\exp\left(-\frac{\rho^2(x^2+x'^2)-2\rho xx'}{2(1-\rho^2)}\right),\qquad -1\le\rho\le 1$$
$$=k_{RBF}(x,x';\ell_{\text{len}})\cdot\exp\left(\frac{\rho}{\rho+1}xx'\right)\tag{15}$$

其中 $\ell^2_{\text{len}}=\frac{1-\rho^2}{\rho^2}$。因此，我们可以将迈勒核解释为另一种 $C^\infty$ 核，它为径向基函数提供非平稳缩放。初步实验始终发现 $\rho>0$，这会导致 $x$ 和 $x'$取较大值时协方差增加。观察非平稳成分影响的一种方式是考察方差和相关性。如果 $f\sim\mathcal{GP}(m,k_{\text{Meh}})$，那么 $\text{var}(f(x))=k_{\text{Meh}}(x,x)=\exp\left(\frac{\rho}{\rho+1}\cdot x^2\right)$，这说明随着$x$增加，过程方差也会增加。值得注意的是，这会产生一个平稳的相关函数 $\text{corr}(f(x)\ f(x'))=\exp\left(-\frac{\rho}{2(1-\rho^2)}(x-y)^2\right)$。因此，当希望将径向基函数的动态特性与不断增加的过程方差相结合时，Mehler 非常合适。

**注 4.** 迈勒核对于 $\rho>0$ 是一个有效的核函数，因为它是正定的，并且具有基展开式 $k(x,x')=\frac{1}{\sqrt{1-\rho^2}}\sum_{k=0}^{\infty}\frac{\rho^k}{k!}H_{e_k}(x)H_{e_k}(x')$，因此对于所有 $n\in\mathbb{N}$、~~$\mathbf{a}\in\mathbb{R}^n$~~ 和 $x_1,\ldots,x_n\in\mathbb{R}$，

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

其中 $H_{e_k}=(-1)^ke^{x^2/2}\frac{d^k}{dx^k}e^{x^2/2}$ 是第 $k$ 个概率论家 *Hermite* 多项式。

## B 关于高斯过程超参数收敛的更多内容

我们的遗传算法中的变异操作，关键取决于对给定数据集在不同核 $k$ 下对数似然的相对比较，因为这直接关系到贝叶斯信息准则的计算。因此，准确估计 $l(k)$ 是判断哪些核比其他核拟合得更好、进而据此进行探索的先决条件。计算似然等价于推断核超参数的最大似然估计，这被认为是一项具有挑战性的优化任务。在我们的实现中，该优化通过 Adam[24], 进行随机梯度下降（SGD），迭代次数达到给定数值 $\eta_{max}$ 时，或达到预设的容差阈值时停止。

本节将通过合成实验中的补充证据，展示这一收敛过程发生得有多快。具体而言，我们从 SYA、SYB 和 SYC 的初始化出发，分别拟合每个真实核 $K_0$，并在训练过程中评估高斯过程超参数的收敛情况，将经过 $\eta$ 步后的中间对数边际似然记为 $l_k^{(\eta)}(\hat{}\theta|\mathbf{y})$。给定一个黄金标准 $l_{\min}=\min_{1\le\eta}l_k^{(\eta)}(\hat{}\theta|\mathbf{y})$，我们记录达到与 $l_{\min}$ 相差不超过 $\varepsilon\in\{10^{-3},10^{-4},\ldots,10^{-7}\}$ 的对数边际似然所需的训练步数：

$$\eta_\varepsilon=\min\left\{\eta:|l_k^{(\eta)}(\hat{\theta}|\mathbf{y})-l_{\min}|\le\varepsilon\right\}.\tag{16}$$

我们在表13中给出了结果。对于 $\varepsilon=10^{-7}$，SYA-1和-2表现出不一致，这可能是因为随机梯度下降优化超参数是针对 $\varepsilon=10^{-6}$ 案例进行校准的。对于 $\varepsilon=10^{-4}$，SYB所需的训练步数是SYA的四倍；对于 $\varepsilon=10^{-6}$，则是十倍。

| | $\varepsilon$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ | $10^{-7}$ |
|---|---|---|---|---|---|---|
| SYA-1 | $\eta_\varepsilon$ | 6 | 62 | 89 | 90 | 90 |
| | 贝叶斯信息准则($\hat{K}_0^{(\eta_\varepsilon)}$) | -2066.14 | -2066.67 | -2066.76 | -2066.77 | -2066.77 |
| SYA-2 | $\eta_\varepsilon$ | 39 | 57 | 143 | 179 | 3241 |
| | 贝叶斯信息准则($\hat{K}_0^{(\eta_\varepsilon)}$) | -2033.23 | -2034.17 | -2034.22 | -2034.23 | -2034.23 |
| SYB | $\eta_\varepsilon$ | 165.00 | 239.00 | 603.00 | 835.00 | 997.00 |
| | 贝叶斯信息准则($\hat{K}_0^{(\eta_\varepsilon)}$) | -2467.19 | -2467.96 | -2468.06 | -2468.07 | -2468.07 |
| SYC | $\eta_\varepsilon$ | 669.00 | 2356.00 | 4500.00 | 6553.00 | 8318.00 |
| | 贝叶斯信息准则($\hat{K}_0^{(\eta_\varepsilon)}$) | -2794.16 | -2795.11 | -2795.20 | -2795.21 | -2795.21 |

表13：在各合成案例研究中，使用相应的 $K_0$时，使似然 $l_{K_0}^{(\eta)}(\hat{\theta}|\mathbf{y})$与 $l_{\min}$ 的差值在 $\varepsilon$ 以内所需的训练步数 $\eta_\varepsilon$。BIC($\hat{K}_0^{(\eta_\varepsilon)}$)对应于 $l_{K_0}^{(\eta)}(\hat{\theta}|\mathbf{y})$。

对于 $\varepsilon = 10^{-4}$ 和 $10^{-6}$，学习 SYC 的超参数大约比学习 SYB 的超参数慢十倍。总之，SYA 和 SYB 的收敛到 $l_{\min}$ 很快，而 SYC 则非常慢，部分原因是后者拥有更多超参数、非平稳项以及非常数方差。

为保证计算可行性，在实验中我们将范围限制为 $\eta_{max} = 300$（运行时间与 $\eta_{max}$ 呈线性关系），结果是：SYA 的容差优于 $10^{-5}$，SYB 的容差优于 $10^{-4}$，但 SYC 的 $l_k$ 可能会被误估至多达 $\pm 1$。这解释了为何在 SYC 实验中难以恢复 $K_0$：遗传算法难以正确地对排名靠前的核进行排序。

# C 补充表格和图

| 范围 | BIC max | BIC min | len | 加法非-组成统计 | 数值 | age | num year | num coh | 粗糙 age | 粗糙 year | 粗糙 coh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 日本女性 $\mathcal{K}_r$：重新运行 | | | | | | | | | | | |
| 1-10 | -2723.53 | -2725.27 | 4.00 | 1.00 | 0% | 1.00 | 1.50 | 1.50 | 0% | 100% | 100% |
| 1-50 | -2720.11 | -2725.27 | 4.48 | 1.02 | 10% | 1.12 | 1.82 | 1.54 | 0% | 100% | 100% |
| 51-100 | -2717.87 | -2720.11 | 5.16 | 1.24 | 14% | 1.28 | 2.30 | 1.58 | 0% | 100% | 100% |
| 在 $\mathcal{D}_{rob}$ 上训练的日本女性，并在 $\mathcal{K}_r$ 中搜索 | | | | | | | | | | | |
| 1-10 | -2716.89 | -2725.27 | 4.10 | 1.00 | 0% | 1.00 | 1.70 | 1.40 | 0% | 100% | 100% |
| 1-50 | -2713.77 | -2725.27 | 4.64 | 1.20 | 2% | 1.20 | 1.88 | 1.56 | 0% | 100% | 100% |
| 51-100 | -2712.33 | -2714.64 | 5.00 | 1.40 | 6% | 1.22 | 1.88 | 1.90 | 0% | 100% | 100% |

表14：在日本女性人类死亡率数据库数据上进行的其他遗传算法运行

# D 解读贝叶斯因子

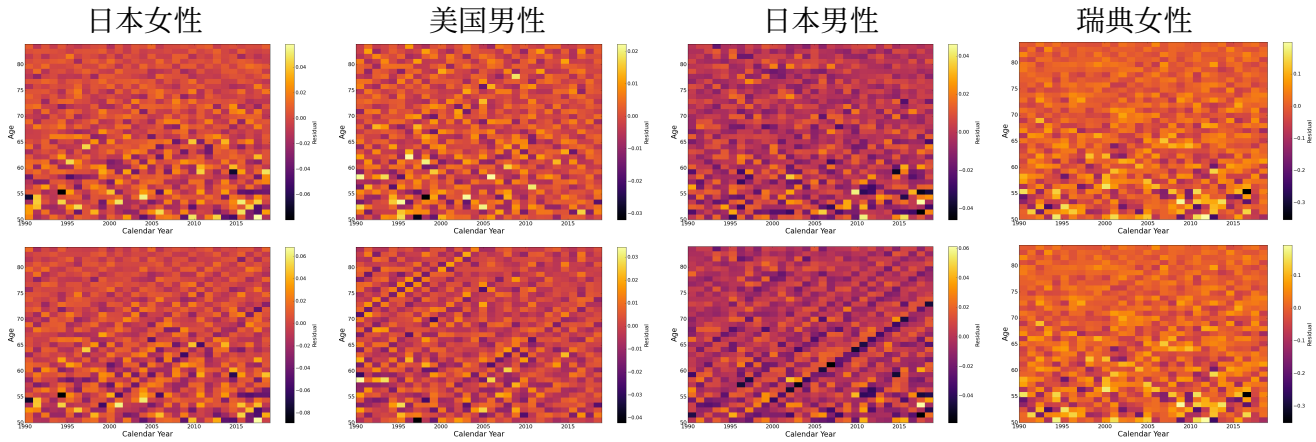



图8： $\mathcal{K}_f$ 中排名最高的核（上排）以及不含队列项的最佳核（下排）的残差图。下排的核分别为：日本女性（$0.0144 \cdot [0.0133 \cdot \text{Meh}_a(0.1) \cdot \text{AR2}_y(1.45, 0.26) + 30.5479 \cdot \text{M32}_y(2.43)] \cdot \text{M52}_a(1.12) \cdot \text{M12}_y(62.46)$；日本男性： $0.082 \cdot M52_a(0.91) \cdot \text{M52}_y(0.9) \cdot [0.0114 \cdot AR2_y(4.53, 0.28) \cdot$ 径向基函数$_a(0.13) + 1.0985 \cdot \text{M12}_y(19.6)]$；美国男性： $0.0194 \cdot \text{Chy}_a(0.78) \cdot \text{M12}_a(22.69) \cdot \text{Chy}_y(0.55) \cdot [1.7718 \cdot \text{M12}_y(25.09) + 0.0111 \cdot \text{M32}_a(0.07) \cdot$ 径向基函数$_y(0.09)]$；瑞典女性： $0.1125 \cdot \text{Chy}_a(0.85) \cdot \text{Meh}_y(1.16) \cdot \text{M12}_y(37.74)$。

| 贝叶斯因子($K_1$， $K_2$) | 对数贝叶斯因子($K_1$， $K_2$) | 代码 | 解释 |
|---|---|---|---|
| $> 100$ | $> 4.61$ | | 支持 $K_1$ 的决定性证据 |
| 30–100 | 3。40 至 4。61 | . | 支持 $K_1$ 的非常强的证据 |
| 10–30 | 2.30 至 3.40 | * | 有力支持 $K_1$ |
| 3–10 | 1.10 至 2.30 | ** | 实质性支持 $K_1$ |
| 1–3 | 0 至 1.10 | *** | 仅值得一提 |

表15：在假设 $l(K_1) < l(K_2)$ 的情况下， [21], 给出的贝叶斯因子的证据类别。请注意，对于第二列，$\text{BIC}(K_2) - \text{BIC}(K_1) \approx \log \text{BF}(K_1, K_2)$。代码列表示显著性水平，其中更多的 $*$ 表明高斯过程分布不同的证据较弱。