arXiv:2305.04625v1 [math.PR] 8 May 2023

签名核

Darrick Lee 和 Harald Oberhauser

2023年5月9日

摘要

签名核是一种适用于序列数据的正定核。它继承了随机分析的理论保证，拥有高效的计算算法，并展现出优异的实证性能。在这篇为即将出版的施普林格手册撰写的简短综述中，我们将对签名核作入门介绍，并着重阐述其理论与计算特性。

1 引言

令 $\mathcal{X}$ 为任意集合，并考虑

$$\mathcal{X}_{\text{seq}} \coloneqq \{\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_L) : L \geq 0, \mathbf{x}_i \in \mathcal{X}\}, \tag{1}$$

集合 L 中任意长度的序列集合。给定 $\mathcal{X}$上的核函数[1] ，

$$\mathrm{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \tag{2}$$

签名核将 k 变换为 $\mathcal{X}_{\text{seq}}$上的核函数，

$$\mathrm{k}_{\mathsf{Sig}} : \mathcal{X}_{\text{seq}} \times \mathcal{X}_{\text{seq}} \to \mathbb{R}. \tag{3}$$

为非欧几里得域（如 $\mathcal{X}_{\text{seq}}$ ）构造核函数通常颇具挑战；尤其是，即使 $\mathcal{X} = \mathbb{R}^d$ 是线性的，序列空间 $\mathcal{X}_{\text{seq}}$ 也是非线性的，因为不同长度的序列之间不存在自然的加法运算。上文在离散时间序列 $\mathcal{X}_{\text{seq}}$ 的定义域上定义了 $\mathbf{k}^{\mathsf{Sig}}$ ，因为在实际应用中，当只能获得离散时间数据时，我们通常采用这种方式；但 $\mathbf{k}^{\mathsf{Sig}}$ 的定义域自然延伸至连续时间路径。更一般地说，签名核之所以受到关注，主要有三个原因。

（i）**它继承了随机分析和经典签名所具备的理论保证**。这些保证体现为普适性（逼近非线性函数的能力）、特征性（刻画概率测度的能力）、对时间重参数化的不变性，以及当序列（离散时间）逼近路径（连续时间）时在缩放极限下的收敛性。

(ii) 它克服了签名特征的瓶颈，无论是在计算复杂度还是表达能力方面都是如此，因为它允许我们使用路径提升至无限维状态空间后的签名。也就是说，它可以间接计算并使用描述底层序列的无限维空间中的张量。

[1]在全文中，我们使用术语 *kernel* 来指代正定函数。

(iii) 它利用了核学习中的模块化工具。给定一个具有良好理论性质的核 k，这些性质通常会传递到相应的签名核 $\mathsf{k}^{\mathsf{Sig}}$，并且核学习中的通用工具也适用。例如，$\mathsf{k}$ 的所谓普适性和特征性[65]会在相应的签名核 $\mathsf{k}^{\mathsf{Sig}}$中得到保留。这还意味着，$\mathsf{k}^{\mathsf{Sig}}$在随机过程的分布之间诱导出一种可计算的度量，即所谓的最大均值差异度量。此外，通过使用 $\mathsf{k}^{\mathsf{Sig}}$ 作为协方差函数，可以在真正的非欧几里得定义域$\mathcal{X}_{\mathrm{seq}}$ 上定义高斯过程，还可以使用通用的核求积构造等。

本综述文章的核心思想是：构成签名的迭代积分，以及由此得到的签名核，对于序列或路径所起的作用，与单项式对于 $\mathbb{R}^d$ 中向量所起的作用相同，见表 1。核心信息是，这既可能是福也是祸。从代数角度看，多项式具有诱人的性质，但在高维空间中很快就会变得计算代价过高；从分析角度看，非紧集上会出现问题，而且往往比起简单的单项式，更灵活的一类函数更适合作为基底。核化使我们能够规避这些缺点。

2 单项式、多项式和张量

经典的多元单项式和矩是便捷的工具，可用于近似 $\mathbb{R}^d$ 上的函数，并通过矩序列表示向量值随机变量的分布。用张量表示向量$\mathbf{x} = (\mathbf{x}^1, \dots, \mathbf{x}^d) \in \mathbb{R}^d$的次数为 m 的单项式，是一种简洁的表达方式。

$$(\mathbb{R}^d)^{\otimes m} \ni \mathbf{x}^{\otimes m} \simeq ((\mathbf{x}^1)^{m_1} \cdots (\mathbf{x}^d)^{m_d})_{m_1+\cdots m_d=m,\, m_i \in \mathbb{N}_0}. \tag{4}$$

我们将 $(\mathbb{R}^d)^{\otimes m}$ 中的元素称为 d 维空间中次数为 m 的张量。

一个玩具示例：$\mathbb{R}^d$ 中的单项式。令 $\mathcal{K} \subset \mathbb{R}^d$ 为紧集，并定义

$$\Phi_{\mathsf{Mon}}(\mathbf{x}) \coloneqq \left(1, \mathbf{x}, \frac{\mathbf{x}^{\otimes 2}}{2!}, \frac{\mathbf{x}^{\otimes 3}}{3!}, \dots\right) \in \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m} \quad \text{and} \quad \Phi_{\mathsf{Mon},m}(\mathbf{x}) \coloneqq \frac{\mathbf{x}^{\otimes m}}{m!}.$$

映射 Φ_{Mon} 以向量 $\mathbf{x} = (\mathbf{x}^i)_{i \in \{1,\dots,d\}} \in \mathbb{R}^d$ 作为输入，并输出一个次数递增的张量序列。该序列由一个标量 $\mathbf{x}^{\otimes 0} \coloneqq 1$、一个向量 $\mathbf{x}^{\otimes 1} = \mathbf{x}$、一个矩阵 $\mathbf{x}^{\otimes 2} = (\mathbf{x}^{i_1}\mathbf{x}^{i_2})_{i_1,i_2 \in \{1,\dots,d\}}$、一个张量 $\mathbf{x}^{\otimes 3} = (\mathbf{x}^{i_1}\mathbf{x}^{i_2}\mathbf{x}^{i_3})_{i_1,i_2,i_3 \in \{1,\dots,d\}}$等组成。 Φ_{Mon}的陪域是一个分级（按 $m = 0, 1, 2, \dots$）线性空间[2]，因此我们可以对 Φ_{Mon}应用线性泛函 w。Stone–Weierstrass定理保证，对于每个 $f \in C(\mathcal{K}, \mathbb{R})$ 和 $\epsilon > 0$，都存在一个 $M \geq 1$及系数 $\mathsf{w}^0, \mathsf{w}^1, \dots, \mathsf{w}^d, \mathsf{w}^{1,1}, \dots, \mathsf{w}^{d,d}, \mathsf{w}^{1,1,1}, \dots, \mathsf{w}^{d,\dots,d} \in \mathbb{R}$，等价地可表示为[3] $\mathsf{w} \in \bigoplus (\mathbb{R}^d)^{\otimes m}$，使得

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| f(\mathbf{x}) - \underbrace{\sum_{m=0}^{M} \mathbf{w}^{i_1,\dots,i_m} \frac{\mathbf{x}^{i_1} \cdots \mathbf{x}^{i_m}}{m!}}_{=: \langle \mathbf{w}, \Phi_{\mathsf{Mon}}(\mathbf{x}) \rangle} \right| < \epsilon.$$

[2]加法按次数定义 $(\mathsf{s}_0, \mathsf{s}_1, \mathsf{s}2, \dots) + (\mathsf{t}_0, \mathsf{t}_1, \mathsf{t}2, \dots) \coloneqq (\mathsf{s}_0 + \mathsf{s}_0, \mathsf{s}_1 + \mathsf{t}_1, \mathsf{s}_2 + \mathsf{t}2, \dots)$。也就是说，标量与标量相加，向量与向量相加，依此类推。[3]回想一下，$\mathsf{w} \in \oplus (\mathbb{R}^d)^{\otimes m}$ 表示一个序列 $\mathsf{w} = (\mathsf{w}^I)$，其中 w^I 仅对有限多个多重指标 I非零；而 $\mathsf{w}' \in \prod (\mathbb{R}^d)^{\otimes m}$ 表示一个序列 $\mathsf{w}' = (\mathsf{w}'^I)$，其中 w'^I 可能对无限多个 I非零。

因此，任意选取 $\mathbf{w}$ 都会导出一个从 $\mathcal{K}$ 到 $\mathbb{R}$的函数，$\left(\mathbf{x} \mapsto \langle \mathbf{w}\ \Phi_{\mathsf{Mon}}(\mathbf{x})\rangle\right) \in \mathbb{R}^{\mathcal{K}}$，这是一个关于 Φ_{Mon}的线性函数。

关于单项式的另一个著名结果是，它们可以刻画随机变量的分布；事实上，矩序列的收敛等价于经典弱收敛。我们将这些结果总结为下面的命题。

命题 1. 设 $\mathcal{K} \subset \mathcal{X} = \mathbb{R}^d$ 为紧集。那么对每个 $f \in C_b(\mathcal{K}, \mathbb{R})$ 和 $\epsilon > 0$，都存在某个 $\mathbf{w} \in \bigoplus (\mathbb{R}^d)^{\otimes m}$，使得

$$\sup_{\mathbf{x}\in\mathcal{K}} |f(\mathbf{x}) - \langle \mathbf{w}, \Phi_{\mathsf{Mon}}(\mathbf{x})\rangle| < \epsilon. \tag{5}$$

设 $\mathbf{X}$、$\mathbf{X}_1, \mathbf{X}_2$、.....是一个取值于 $\mathcal{K}$的随机变量序列。那么

$$\mathrm{Law}(\mathbf{X}) \mapsto \mathbb{E}[\Phi_{\mathsf{Mon}}(\mathbf{X})] \coloneqq \left(1, \mathbb{E}[\mathbf{X}], \frac{\mathbb{E}[\mathbf{X}^{\otimes 2}]}{2!}, \dots\right)$$

是单射，且下列条件等价：

1. $\forall f \in C_b(\mathcal{X}, \mathbb{R}),\quad \lim_{n\to\infty} \mathbb{E}[f(\mathbf{X}_n)] = \mathbb{E}[f(\mathbf{X})]$;
2. $\forall m \in \mathbb{N}$, 极限$_{n\to\infty}\, \mathbb{E}[\mathbf{X}_n^{\otimes m}] = \mathbb{E}[\mathbf{X}^{\otimes m}]$.

证明。 (1) 蕴含 (2) 可立即得出，因为多项式在紧集上有界，因此它们构成 $C_b(\mathcal{K}, \mathbb{R})$的一个子集。要证明反向蕴含，注意到紧致性假设意味着 $(\mathbf{X}_n)$ 的分布具有紧性，因此我们可以使用 Prokhorov定理。于是，证明归结为说明：若 $(\mathbf{X}_n)$ 沿某个子序列弱收敛到随机变量 $\mathbf{Y}$，则 $\mathsf{Law}(\mathbf{Y}) = \mathsf{Law}(\mathbf{X})$。但若 $\mathbf{X}_n$ 弱收敛到 $\mathbf{Y}$，则对任意多项式 p，由于多项式在紧集上有界，有 $\lim_n \mathbb{E}[p(\mathbf{X}_n)] = \mathbb{E}[p(\mathbf{X})]$。因此，对任意多项式 p，必有$\mathbb{E}[p(\mathbf{X})] = \mathbb{E}[p(\mathbf{Y})]$ 。最后，由于多项式在 $C_b(\mathcal{K}, \mathbb{R})$ 中稠密（依据我们上文提到的 Stone–Weierstrass定理），结论得证。 □

上述内容表明，单项式的线性泛函，即 Φ_{Mon} 的坐标的线性组合，可以任意逼近连续函数。此外，Φ_{Mon} 的坐标的期望值刻画了随机变量的分布，而这些坐标的收敛在概率测度空间上诱导出经典弱收敛拓扑。

备注 1（普适性与特征性）。在统计学习文献中，这些性质被称为 Φ_{Mon} 相对于有界连续函数类的**普适性**，以及 Φ_{Mon} 相对于 $\mathcal{K}$ 上的概率测度的**特征性**。一般定义称，一个从集合 $\mathcal{K}$ 映射到希尔伯特空间 H 的映射 $\Phi : \mathcal{K} \to H$，如果线性泛函空间 $\mathcal{F}_0 \coloneqq \{\langle \mathbf{w}\ \Phi(\cdot)\rangle : \mathcal{K} \to \mathbb{R} : \mathbf{w} \in H\}$ 在 $\mathcal{F}$ 中稠密，则称其（i）相对于函数类 $\mathcal{F} \subset \mathbb{R}^{\mathcal{K}}$ 是**通用的**；如果将概率测度映射到 H 的期望映射 $\mu \mapsto \mathbb{E}_{X\sim\mu}[\Phi(X)]$ 是单射，则称其（ii）相对于概率测度 $\mathcal{P}(\mathcal{K})$ 是**特征的**。事实上，如果采用更一般的特征性定义，将分布（即 $\mathcal{F}'$ 的元素）纳入考量，可以证明特征性与普适性是等价的。这使许多证明变得更加容易，参见 [16, 定义 6] 和 [65]。

命题 1 的证明大量运用了紧致性假设，但这并不意味着证明有缺陷。下面的例子表明，在非紧空间中，情况确实会变得很糟。

警告：紧致性和有界性至关重要。 我们不考虑紧集子集$\mathcal{K}$，而是考虑整个空间 $\mathcal{X} = \mathbb{R}^d$。在这种情况下，对于定义在 $\mathcal{X}$ 上且具有无界支撑的随机变量，矩可能不存在；但情况还要更糟：即使所有矩都存在，它们也无法刻画该随机变量的分布。

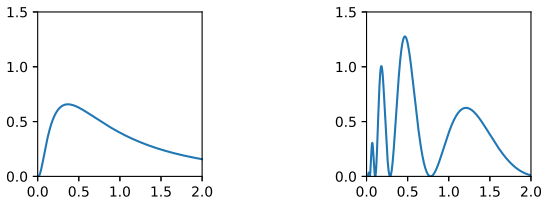

图 1：$\mathbf{X}$ （左）和$\mathbf{Y}$ （右）的概率密度函数。

如下经典反例所示，这种情况在维度为 $d = 1$ 时就已经出现。设$\mathbf{X}$ 服从标准对数正态分布，即 $\exp(\mathbf{X}) \sim N(0, 1)$。存在另一个实值随机变量$\mathbf{Y}$ ，它不服从对数正态分布[4]，，且 $\mathrm{Law}(\mathbf{X}) \neq \mathrm{Law}(\mathbf{Y})$，但对于所有 $m \in \mathbb{N}$，

$$\mathbb{E}[\mathbf{X}^m] = \mathbb{E}[\mathbf{Y}^m] = \exp\left(\frac{m^2}{2}\right).$$

因此，在非紧集上，“特征性”会失效。类似地，通用性（即对方程 (5) 中的函数 $f(\mathbf{x})$ 进行一致逼近）也无望实现，因为当自变量$\mathbf{x}$ 变得无界时，多项式会趋于 $\pm\infty$，其中 $\|\mathbf{x}\| \to \infty$。

多项式在紧集上具有良好的代数和解析性质，但根本问题在于：尽管它们的代数性质可以推广到非紧集上（它们仍然构成一个环），但在分析性质和概率性质方面却会遇到问题，因为它们增长得太快。我们强调，“紧性”论证在这里并不太有用；很容易找到一个紧集，使任意 $\epsilon > 0$ 的概率测度质量 $(1 - \epsilon)$ 集中于该集合；问题在于，我们关心的函数——单项式——在空间中仅承载 ϵ 概率质量的定义域上会发散。处理这一问题的一种经典方法，是对随机变量的矩增长施加条件，但这对我们的目的而言并不令人满意，因为这类假设很快就会变得抽象，而且在机器学习任务中常见的非参数推断场景下，通常无法验证这些假设。毕竟，一个简单的对数正态分布就已经出了问题。最后，一个密切相关的问题是，矩并非稳健统计量：也就是说，如果我们必须根据来自 $\mathcal{X}$ 的独立同分布样本估计期望，那么这些样本中的微小误差就可能导致较大的估计误差。

我们之后会回到所有这些问题，但现在先忽略非紧集这一问题，提出一个更基础的问题：单项式从 $\mathcal{X} = \mathbb{R}^d$推广到空间 $\mathcal{X}_{\mathrm{seq}}$ 中的 $\mathbb{R}^d$ 序列，自然的推广是什么？

3 非交换单项式的迭代积分

现在我们将超越经典设定，考虑序列数据的单项式。为简单起见，我们在 $\mathcal{X} = \mathbb{R}^d$ 中讨论路径和序列，但在第 6 节中，我们会讨论无限维状态空间。尽管实际应用中我们使用离散序列 $\mathcal{X}_{\mathrm{seq}}$ ，但通常.....

[4]随机变量 $\mathbf{Y}$ 由密度 $\mathbb{P}(\mathbf{Y} \in dy) = f(y)(1 + \sin(2\pi\log(y))$ 显式指定，其中 f 表示对数正态随机变量 $\mathbf{X}$ 的密度。事实上， [38] 表明，存在不可数多个与 $\mathbf{X}$ 同分布的随机变量。

便于将我们的对象表述为连续路径。然而，通过分段线性插值，我们可以自然地将离散序列视为连续路径。事实上，假设我们有一个序列 $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_L) \in \mathcal{X}_{\text{seq}}$。滥用记号，我们定义相应的连续路径 $\mathbf{x} : [0, L] \to \mathbb{R}^d$ 通过

$$\mathbf{x}(t) = (1 - t)\mathbf{x}_{i-1} + t\mathbf{x}_i, \qquad t \in [i - 1, i]. \tag{6}$$

在连续情形下，我们将考虑有界变差路径

$$\mathcal{X}_{\text{1-var}} \coloneqq \{\mathbf{x} \in C([0, T], \mathcal{X}) : T > 0,\ \|\mathbf{x}\|_1 < \infty, \mathbf{x}(0) = 0\}$$

其中

$$\|\mathbf{x}\|_1 \coloneqq \sup_{\substack{0=t_0<\cdots<t_n=T \\ n\in\mathbb{N}}} \sum_{i=1}^{N} |\mathbf{x}(t_i) - \mathbf{x}(t_{i-1})|$$

表示路径通常的 1-变差。由于本文所讨论的方法（特别是路径签名）具有平移不变性，我们将路径限制为从原点开始。此外，为便于阐述，我们取 $\mathcal{X} = \mathbb{R}^d$ 为有限维。为 $\mathcal{X}_{\text{1-var}}$ 赋予拓扑有多种方式，我们采用与 [16] 中相同的方式。最后，我们注意到，方程 6 定义了一个嵌入

$$\mathcal{X}_{\text{seq}} \hookrightarrow \mathcal{X}_{\text{1-var}}.$$

前一篇文章 [13]介绍了路径签名，

$$\Phi_{\mathsf{Sig}}(\mathbf{x}) \coloneqq \left(1, \int d\mathbf{x}, \int d\mathbf{x}^{\otimes 2}, \dots\right) \in \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}$$

它将有界变差路径映射为张量序列。次数为 m 的张量表示为一个 m 重迭代积分，

$$\Phi_{\mathsf{Sig},m}(\mathbf{x}) \coloneqq \int d\mathbf{x}^{\otimes m} \coloneqq \int_{0<t_1<\cdots<t_m<T} d\mathbf{x}(t_1) \otimes \cdots \otimes d\mathbf{x}(t_m) \in (\mathbb{R}^d)^{\otimes m},$$

并且按照惯例，$(\mathbb{R}^d)^{\otimes 0} = \mathbb{R}$。或者，在坐标表示下，我们可以将其识别为

$$(\mathbb{R}^d)^{\otimes m} \ni \int d\mathbf{x}^{\otimes m} \simeq \left(\int_{0<t_1<\cdots<t_m<T} d\mathbf{x}^{i_1}(t_1) \cdots d\mathbf{x}^{i_m}(t_m)\right)_{i_1,\dots,i_m\in\{1,\dots,d\}}$$

其中

$$\int_{0<t_1<\cdots<t_m<T} d\mathbf{x}^{i_1}(t_1) \cdots d\mathbf{x}^{i_m}(t_m) = \int_{0<t_1<\cdots<t_m<T} \dot{\mathbf{x}}^{i_1}(t_1) \cdots \dot{\mathbf{x}}^{i_m}(t_m) dt_1 \cdots dt_m$$

由于有界变差路径几乎处处可微，因此该积分可定义为黎曼–斯蒂尔切斯积分。对于规则性较低的路径，需要谨慎处理构成签名的迭代积分的定义，但相同的原则仍然适用（见第9节）；对于取值于无限维状态空间中的路径，同样的构造也适用，我们将在第 6 节讨论希尔伯特空间的情形。这类迭代积分可以视为单项式的推广，下面我们将着重强调二者的相似之处。

一种分级描述。 理解签名映射的一种方式

$$\Phi_{\mathsf{Sig}} : \mathcal{X}_{\text{1-var}} \to \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m} \quad \mathbf{x} \mapsto \left(1, \int d\mathbf{x}, \int d\mathbf{x}^{\otimes 2}, \dots\right)$$

也就是说，它是单项式映射的自然推广

$$\Phi_{\mathsf{Mon}} : \mathcal{X} \to \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}, \quad \mathbf{x} \mapsto \left(1, \mathbf{x}, \frac{\mathbf{x}^{\otimes 2}}{2!}, \dots\right)$$

从定义域 $\mathcal{X}$ 到 $\mathcal{X}_{\text{1-var}}$。二者都将底层对象（向量或路径）描述为次数递增的一系列张量。

签名推广单项式。 签名 Φ_{Sig} 是 Φ_{Mon}的严格推广。对于直线， $[0,1] \ni t \mapsto t \cdot \mathbf{v}$ 且 $\mathbf{v} \in \mathbb{R}^d$，我们有

$$\Phi_{\mathsf{Sig}}(t \mapsto t \cdot \mathbf{v}) = \left(1, \mathbf{v}, \frac{\mathbf{v}^{\otimes 2}}{2!}, \frac{\mathbf{v}^{\otimes 3}}{3!}, \dots\right) = \Phi_{\mathsf{Mon}}(\mathbf{v}) \tag{7}$$

这直接由 Φ_{Sig} 的定义得出，因为

$$\int d(t \cdot \mathbf{v})^{\otimes m} = \int_{0<t_1<\cdots<t_m<1} \mathbf{v}^{\otimes m} dt_1 \cdots dt_m = \frac{\mathbf{v}^{\otimes m}}{m!}.$$

尽管它们的定义相似，但也有一些需要强调的重要差异。

真正的非线性定义域。 不同于 $\mathbb{R}^d$， Φ_{Sig} 的定义域 $\mathcal{X}_{\text{1-var}}$ 不是线性的： $\mathcal{X}_{\text{1-var}}$中的元素没有自然的加法。特别是，对于 $T \neq T'$，不存在将路径 $\mathbf{x} : [0,T] \to \mathbb{R}^d$ 与 $\mathbf{y} : [0,T'] \to \mathbb{R}^d$ 相加的自然方式。

非对称张量。 映射 Φ_{Sig} 和 Φ_{Mon} 的陪域之间存在一个重要区别。后者的取值位于 $\prod(\mathbb{R}^d)^{\otimes m}$ 的对称张量子集中。例如，如果将 Φ_{Mon} 中的次数为 $m = 2$ 的张量视为 $d \times d$ 矩阵，那么 $\mathbf{x}^{\otimes 2}$ 是一个对称矩阵，因为 $\mathbf{x}$ 的 (i,j) 坐标 $^{\otimes 2}$ 正是 $\mathbf{x}$ 的第 i 个坐标与第 j 个坐标的通常乘积，而 $\mathbf{x}^i \cdot \mathbf{x}^j = \mathbf{x}^j \cdot \mathbf{x}^i$。

然而，一般而言， Φ_{Sig} 中的张量并不对称。例如，再次考察 $\int d\mathbf{x}^{\otimes 2}$ 的 (i,j) 和 (j,i) 坐标，我们一般有

$$\int_{0<t_1<t_2<T} d\mathbf{x}^i(t_1) d\mathbf{x}^j(t_2) \neq \int_{0<t_1<t_2<T} d\mathbf{x}^j(t_1) d\mathbf{x}^i(t_2).$$

非交换单项式。 由于 Φ_{Sig} 产生的是非对称张量，这些单项式背后的代数结构是非交换的。给定两个张量 $\mathbf{v} \in (\mathbb{R}^d)^{\otimes m}$ 和 $\mathbf{w} \in (\mathbb{R}^d)^{\otimes k}$，张量积 $\mathbf{v} \otimes \mathbf{w} \in (\mathbb{R}^d)^{\otimes (m+k)}$ 按坐标定义为

$$(\mathbf{v} \otimes \mathbf{w})^{i_1,\dots,i_{m+k}} := \mathbf{v}^{i_1,\dots,i_m} \cdot \mathbf{w}^{i_{m+1},\dots,i_{m+k}}.$$

一般而言，张量积是非交换的，$\mathbf{v} \otimes \mathbf{w} \neq \mathbf{w} \otimes \mathbf{v}$。然而，当我们局限于对称张量（例如 Φ_{Mon}）时，张量积是可交换的

$$\mathbf{x}^{\otimes m} \otimes \mathbf{x}^{\otimes k} = \mathbf{x}^{i_1} \cdots \mathbf{x}^{i_{m+k}} = \mathbf{x}^{\otimes k} \otimes \mathbf{x}^{\otimes m}.$$

备注 2. 我们将 Φ_{Sig} 称为签名（特征）映射；不过，将路径展开为迭代积分级数是一种经典构造，许多学术领域都对此进行过研究，只是采用了不同的名称，例如 Chen-Fliess级数、Magnus展开、时间序指数、路径指数、时间序演算等。其现代起源可追溯至 Chen 在 [10, 11, 12] 中的工作，这可以视为经典 Volterra级数 [61]的扩展。签名丰富的李代数性质在自由李代数[55] 的研究中受到广泛关注；到了 20 世纪 80 年代，控制理论家广泛使用 Φ_{Sig} 来线性化受控微分方程 [29, 6]。随后，这些成果大多被推广到所谓的粗糙路径理论[50] 中，以处理“粗糙”轨迹，例如随机过程的样本路径，因为对这类路径无法应用标准积分。详细的历史不在本文讨论范围之内，但上述参考文献展现了丰富的文献，来自不同领域的研究者都曾使用（各种变体的）我们所称的 Φ_{Sig} 来开展各自的研究。

4 签名的普适性与特征性

为了进一步确立单项式 $\Phi_{\mathsf{Mon}}:\mathbb{R}^d \to \prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$与签名 $\Phi_{\mathsf{Sig}}:\mathcal{X}_{\text{1-var}} \to \prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$之间的关系，我们现在探讨如何利用签名研究 $\mathcal{X}_{\text{1-var}}$上的函数和测度。特别地，给定某个 $\mathbf{w} \in \bigoplus(\mathbb{R}^d)^{\otimes m}$，我们像之前一样定义函数 $\langle \mathbf{w}\ \Phi_{\mathsf{Sig}}(\cdot)\rangle : \mathcal{X}_{\text{1-var}} \to \mathbb{R}$；在坐标中，

$$\langle \mathbf{w}, \Phi_{\mathsf{Sig}}(\mathbf{x})\rangle \coloneqq \sum_{m=0}^{M} \sum_{i_1,\ldots,i_m \in \{1,\ldots,d\}} \mathbf{w}^{i_1,\ldots,i_m} \int_{0<t_1<\cdots<t_m<T} d\mathbf{x}^{i_1}(t_1)\cdots\mathbf{x}^{i_m}(t_m).$$

直接计算可知，两个这样的线性泛函的乘积仍可表示为关于 Φ_{Sig} 的一个线性泛函。更正式地，对于 $\mathbf{w}_1,\mathbf{w}_2 \in \bigoplus(\mathbb{R}^d)^{\otimes m}$，存在某个$\mathbf{w}_{1,2} \in \bigoplus(\mathbb{R}^d)^{\otimes m}$，使得

$$\langle \mathbf{w}_1, \Phi_{\mathsf{Sig}}(\mathbf{x})\rangle\langle \mathbf{w}_2, \Phi_{\mathsf{Sig}}(\mathbf{x})\rangle = \langle \mathbf{w}_{1,2}, \Phi_{\mathsf{Sig}}(\mathbf{x})\rangle \tag{8}$$

对于所有 $\mathbf{x} \in \mathcal{X}_{\text{1-var}}$。换言之，（非交换）多项式的乘积仍然是（非交换）多项式。尽管这一结论的证明只需直接应用分部积分，但 $\mathbf{w}^{1,2}$ 的代数结构（即所谓的 shuffle 结构）引出了许多有趣的代数问题，不过这些问题超出了本简短导论的讨论范围。关键在于，(8) 表明，路径的线性泛函构成一个代数。如果我们还具有点分离性质，即其中

$$\mathbf{x} \mapsto \Phi_{\mathsf{Sig}}(\mathbf{x})$$

是单射，则我们可以立即将命题 1 从 $\mathbb{R}^d$ 中的紧子集推广到 $\mathcal{X}_{\text{1-var}}$ 中的紧子集。然而，一般而言，$\mathbf{x} \mapsto \Phi_{\mathsf{Sig}}(\mathbf{x})$ 仅在所谓的树状等价意义下才是单射。有关树状等价的详细信息，请参阅 [35] 。

我们可以通过引入时间坐标来规避下一个命题中的这种树状等价：我们不使用 $t \mapsto \mathbf{x}(t) \in \mathbb{R}^d$，而是考虑

$$t \mapsto \bar{\mathbf{x}}(t) \coloneqq (t, \mathbf{x}(t)) \in \mathbb{R}^{d+1}. \tag{9}$$

当我们考虑此类时间参数化路径时，签名 Φ_{Sig} 是单射的，由此我们得到命题 1 的以下推广。

	单项式映射	路径签名
定义域	$\mathcal{X} = \mathbb{R}^d$	$\mathcal{X}_{\text{1-var}}$
陪域	$\prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$	$\prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$
Map	$\Phi_{\mathsf{Mon}}(\mathbf{x}) = \left(\frac{\mathbf{x}^{\otimes m}}{m!}\right)_{m\geq 0}$	$\Phi_{\mathsf{Sig}}(\mathbf{x}) = \left(\int d\mathbf{x}^{\otimes m}\right)_{m\geq 0}$
张量类型	对称	非对称
连续性	关于欧几里得度量	关于 1-变差度量
阶乘衰减	$\|\Phi_{\mathsf{Mon},\,m}(\mathbf{x})\| \leq \frac{\|\mathbf{x}\|^m}{m!}$	$\|\Phi_{\mathsf{Sig},\,m}(\mathbf{x})\| \leq \frac{\|\mathbf{x}\|_1^m}{m!}$
缩放	$\Phi_{\mathsf{Mon},m}(\lambda\mathbf{x}) = \lambda^m\Phi_{\mathsf{Mon},m}(\mathbf{x})$	$\Phi_{\mathsf{Sig},\,m}(\lambda\mathbf{x}) = \lambda^m\Phi_{\mathsf{Sig},\,m}(\mathbf{x})$
普适性	紧集 $\mathcal{K}\subset\mathcal{X}$ 关于 $C(\bar{\mathcal{K}},\mathbb{R})$	紧集 $\mathcal{K}\subset\mathcal{X}_{\text{1-var}}$ 关于 $C(\mathcal{K},\mathbb{R})$
特征性	关于 $\mathcal{P}(\bar{\mathcal{K}})$	关于 $\mathcal{P}(\mathcal{K})$

表 1：经典单项式与签名的比较。

命题 2. 设 $\mathcal{K}\subset\mathcal{X}_{1\text{-}var}$ 为一个紧集，并用 $\bar{\mathcal{K}} = \{\bar{\mathbf{x}} : \mathbf{x}\in\mathcal{K}\}$ 表示由$\bar{}$中的路径构成的集合，这些路径根据公式 *9* 增加了时间坐标。则对于每个 $f\in C_b(\mathcal{K},\mathbb{R})$ 以及$\epsilon>0$，存在一个 $\mathbf{w}\in\bigoplus(\mathbb{R}^{d+1})^{\otimes m}$，使得

$$\sup_{\mathbf{x}\in\mathcal{K}}|f(\bar{\mathbf{x}}) - \langle\mathbf{w},\Phi_{\mathsf{Sig}}(\bar{\mathbf{x}})\rangle| < \epsilon.$$

设$\mathbf{X}$、$\mathbf{X}_1,\mathbf{X}_2$、.....是一个取值于 $\mathcal{K}$ 的随机变量序列，也就是说，每个 $\mathbf{X}_i = (\mathbf{X}_i(t))_{t\in[0,T_i]}$都是一个随机过程。那么期望签名映射

$$\text{Law}(\bar{\mathbf{X}})\mapsto\mathbb{E}[\Phi_{\mathsf{Sig}}(\bar{\mathbf{X}})]$$

是单射的，并且

1. $\forall f\in C_b(\mathcal{K},\mathbb{R})\ \lim_{n\to\infty}\mathbb{E}[f(\mathbf{X}_n)] = \mathbb{E}[f(\mathbf{X})]$,
2. $\forall m\in\mathbb{N}$, 极限$_{n\to\infty}\mathbb{E}[\int d\bar{\mathbf{X}}_n^{\otimes m}] = \mathbb{E}[\int d\bar{\mathbf{X}}^{\otimes m}]$.

这个命题是命题 1 的直接推广；事实上，可以用相同的方法证明。唯一的区别是，需要证明$\mathbf{x}\mapsto\Phi_{\mathsf{Sig}}(\bar{\mathbf{x}})$是单射。不过，这并不太难，例如参见 [48]。如果不进行时间坐标增广，那么将 $\mathcal{K}$ 替换为路径等价类的集合（所谓的树状等价）[5]时，同样的结论也成立。考虑到这些性质，我们可以看到，签名映射的表现与经典单项式非常相似，我们在表 1 中总结了这些结果。

5 好的、坏的与丑的

虽然经典单项式具有许多吸引人的性质，但单项式也有一些缺点，而这些缺点也会遗传到“非交换单项式”中，也就是构成 Φ_{Sig} 的迭代积分。

[5]路径和无参数化路径有多种拓扑可供选择；关于无参数化路径的详细研究，参见 -para参见 [9]。

很好。单项式是构造更复杂函数的基本构件，并且具有良好理解的代数结构。例如，单项式的线性组合（即多项式）在乘法下封闭，并构成一个环；任何具有紧支撑的连续函数都可以由多项式一致逼近。

糟糕。• 单项式以及多项式在非紧集定义域上表现极差；它们很擅长提供局部近似，但在无界定义域上会发散。

• 在无界定义域上，矩序列可能不存在。但即使它存在，也就是说所有矩都存在，矩序列通常也不能唯一刻画随机变量的分布。

• 计算复杂度：张量 $\mathbf{x}^{\otimes m} \in (\mathbb{R}^d)^{\otimes m}$ 有 $\binom{d+m-1}{m} = O(d^m)$ 个坐标，因此在中等规模的 d 下，计算成本会迅速变得难以承受，这使得回归 $f(\mathbf{x}) \approx \langle \mathbf{w}\ \Phi_{\mathsf{Mon}}(\mathbf{x}) \rangle$ 以及处理矩 $\mathbb{E}[\mathbf{X}^{\otimes m}]$ 都过于昂贵。

• 非稳健性：底层概率测度的微小变化可能导致矩序列发生巨大变化。这会给估计带来问题；例如，样本中的一个离群值就可能彻底破坏矩估计。

丑陋之处。虽然单项式理论上可以逼近函数并捕捉概率分布，但在实践中可能并非理想选择。在数据科学中，可以说最成功的方法是从数据本身学习近似函数。在实践中，这意味着提供一个规模庞大的参数化函数族，并在将其应用于数据时对参数进行优化；而标准单项式很少是最优选择。

由于 Φ_{Sig} 具有泛化能力 Φ_{Mon}，它也存在相同的问题。不过，这些问题早已为人所知，现代统计学和机器学习也已开发出解决这些缺陷的工具。本文余下部分将重点介绍以下两种工具，并针对 Φ_{Sig}展开阐述。

• **核化**让我们能够同时考虑丰富的非线性函数，同时避免计算单项式时出现组合爆炸。其方法是绕过对各个单项式的计算，通过开发高效算法直接计算单项式之间的内积。

• **稳健统计学**于 20 世纪 80 年代发展起来 [41]，旨在提供一套形式体系，用于量化给定统计量的稳健性，并提供工具，使给定统计量在这一精确的数学意义上具备稳健性。我们借此同时解决 Φ_{Mon} 和 Φ_{Sig}在非紧集上的问题，并应对它们普遍缺乏稳健性的问题。

将核化和稳健性方法扩展到路径签名 Φ_{Sig}所面临的问题涉及不同方面，但二者可以结合。我们将在第 6 节和第 7 节讨论核化；第 8 节将讨论稳健化。

6 核学习

我们简要回顾核化背后的主要思想，并请参阅 [63, 19]，以获取关于核学习的详细介绍。

核化简述。给定一个集合 $\mathcal{K}$，我们希望找到一个特征映射 $\varphi: \mathcal{K} \to H$，将 $\mathcal{K}$ 中的元素映射到希尔伯特空间 H，使得 φ 提供足够丰富的非线性函数，从而首先，φ上的线性泛函 $\mathbf{x} \mapsto \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle$能够充分表达丰富的函数集合并对其进行逼近；其次，$\mathbb{E}[\varphi(\mathbf{X})]$能够捕捉取值于 $\mathcal{K}$的随机变量$\mathbf{X}$的分布。我们遇到的例子包括以 $\varphi = \Phi_{\mathsf{Mon}}$ 为特征映射的 $\mathcal{K} \subset \mathcal{X} = \mathbb{R}^d$，以及以 $\varphi = \Phi_{\mathsf{Sig}}$ 为特征映射的$\mathcal{K} \subset \mathcal{X}_{1\text{-var}}$。即使对于 $\mathcal{K} = \mathbb{R}$，φ和 H 也有很多选择，因此让我们退一步，单纯思考什么可以作为特征空间 H的良好候选。

我们希望 H 具有线性结构、足够大，并且相对于底层数据空间 $\mathcal{K}$具有规范性。对于任意集合 $\mathcal{K}$，都可以与之关联的一个规范线性空间是定义在 $\mathcal{K}$ 上的实值函数 $\mathbb{R}^{\mathcal{K}}$ 空间。因此，我们取 H 为 $\mathbb{R}^{\mathcal{K}}$ 的子集，即 H 将是定义在 $\mathcal{K}$ 上的实值函数空间。由此，特征映射 φ 将是函数值的，

$$\mathcal{X} \ni \mathbf{x} \mapsto \varphi(\mathbf{x}) := \mathrm{k}(\mathbf{x}, \cdot) \in H \subset \mathbb{R}^{\mathcal{K}}. \tag{10}$$

虽然这种方法有望通过将 $\mathcal{K}$ 嵌入 H（通常为无限维）空间中，提供丰富的“非线性函数”，但其缺点在于，学习算法需要计算 $\varphi(\mathbf{x}) = \mathrm{k}(\mathbf{x}, \cdot)$，这使得直接使用（10）式变得不可行。核学习基于三个观察结果。

1. 如果我们假设 k 是正定的，就可以将 H 识别为由 k 诱导的再生核希尔伯特空间（再生核希尔伯特空间），也就是说，H 是.....的闭包。

$$\left\{ \mathbf{x} \mapsto \sum_{i=1}^{n} \mathrm{k}(\mathbf{x}_i, \mathbf{x}) \,:\, \mathbf{x}_i \in \mathcal{K},\, n \in \mathbb{N} \right\} \subset \mathbb{R}^{\mathcal{K}}.$$

这引出了再生性质，其中 $\langle f, \mathrm{k}(\mathbf{x}, \cdot) \rangle_H = f(\mathbf{x})$ 对于 $f \in H$ 成立。作为再生性质的一个推论，我们有 $\langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle_H = \mathrm{k}(\mathbf{x}, \mathbf{y})$。因此，我们可以将 k 视为一种“非线性内积”。

2. 即使 $\mathrm{k}(\mathbf{x}, \mathbf{y})$ 是 $\varphi(\mathbf{x})$ 和 $\varphi(\mathbf{y})$ 在（通常为无限维的）H 中的内积，其计算通常仍然十分高效。例如，或许最流行的核是径向基函数（RBF）核，$\mathrm{k}(\mathbf{x}, \mathbf{y}) = \exp(-\sigma |\mathbf{x} - \mathbf{y}|^2)$，它是无限维映射 $\varphi(\mathbf{x})$ 的内积。特征映射 φ 的内积计算起来很高效，这通常被称为“核技巧”。

3. 许多学习算法都可以重新表述，使其无需计算 $\varphi(\mathbf{x})$，只需计算内积 $\mathrm{k}(\mathbf{x}, \mathbf{y}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle$。

总而言之，对于给定的数据集 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{X}$，核化使我们能够处理取值于无限维空间 H 中的高度非线性特征映射 $\varphi(\mathbf{x})$，而只需计算格拉姆矩阵 $(\mathrm{k}(\mathbf{x}_i, \mathbf{x}_j))_{i,j \in N}$。

注释 3. 核化利用对偶性实现复杂度权衡：虽然在计算核函数时，我们不必担心 H 的维度，但必须计算整个格拉姆矩阵，其计算量会随样本数量 N 呈二次增长。这是核化背后对偶性的直接结果。

为了获得一些直观认识，我们重新审视并核化第 2 节中的经典单项式特征映射，

$$\Phi_{\mathsf{Mon}} : \mathbb{R}^d \to \prod_{m \geq 0} (\mathbb{R}^d)^{\otimes m}, \mathbf{x} \mapsto \left(1, \mathbf{x}, \frac{\mathbf{x}^{\otimes 2}}{2}, \ldots \right).$$

此外，我们用以下符号表示

$$\Phi_{\mathsf{Mon},:M} = \mathbb{R}^d \to \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}, \mathbf{x} \mapsto \left(1, \mathbf{x}, \frac{\mathbf{x}^{\otimes 2}}{2}, \dots, \frac{\mathbf{x}^{\otimes M}}{M!}, 0, 0, \dots\right)$$

同一个映射，但截断在第 M 级；这里 $0 \in (\mathbb{R}^d)^{\otimes M}$ 表示次数为 M 的张量，其所有元素均为零。Φ_{Mon} 和 $\Phi_{\mathsf{Mon},:M}$ 的陪域相同，但后者只有有限个非零元素。我们称 $\Phi_{\mathsf{Mon},:M}$ 截断于 M。

张量的内积。通过设定，在 $\mathbb{R}^d$ 上定义一个自然的内积。

$$\langle \mathbf{s}_m, \mathbf{t}_m \rangle_m \coloneqq \sum_{i_1,\dots,i_m=1}^{d} \mathbf{s}_m^{i_1,\dots,i_m} \cdot \mathbf{t}_m^{i_1,\dots,i_m},$$

其中 $\mathbf{s}_m, \mathbf{t}_m \in (\mathbb{R}^d)^{\otimes m}$。该内积可以扩展到 $\prod(\mathbb{R}^d)^{\otimes m}$ 的子集上

$$\langle \mathbf{s}, \mathbf{t} \rangle \coloneqq \sum_{m=0}^{\infty} \langle \mathbf{s}_m, \mathbf{t}_m \rangle_m,$$

使上述求和收敛的条件；其中 $\mathbf{s} = (\mathbf{s}_0, \mathbf{s}_1, \dots), \mathbf{t} = (\mathbf{t}_0, \mathbf{t}_1, \dots) \in \prod(\mathbb{R}^d)^{\otimes m}$。因此，若记 $\mathbf{e}_i \in \mathbb{R}^d$ 为除第 i 个分量为 1 外其余分量全为零的向量，则 $\prod(\mathbb{R}^d)^{\otimes m}$ 的一组正交基为

$$\{\mathbf{e}_{i_1} \otimes \cdots \otimes \mathbf{e}_{i_m} : i_1, \dots, i_m \in \{1, \dots, d\}^m, m \in \mathbb{N}\}.$$

下面是我们在全文中一直使用的一条基础但关键的性质。

$$\langle \mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_m, \mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_m \rangle = \langle \mathbf{x}_1, \mathbf{y}_1 \rangle \cdots \langle \mathbf{x}_m, \mathbf{y}_m \rangle \text{ for all } \mathbf{x}_1, \dots, \mathbf{x}_m, \mathbf{y}_1, \dots, \mathbf{y}_m \in \mathbb{R}^d. \tag{11}$$

这已经暗示了我们如何降低计算复杂度。朴素求值左侧需要我们存储张量 $\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_m$ 和 $\mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_m$ 的 d^m 个坐标，然后计算内积，因此需要 $O(d^m)$ 内存。相比之下，计算右侧只需要 $O(dm)$ 内存。

玩具示例 1：避免计算复杂度。令 $\mathcal{X} \subset \mathbb{R}^d$ 并定义

$$\mathrm{k}^0_{\mathsf{Mon},:M} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad \mathrm{k}^0_{\mathsf{Mon}}(\mathbf{x}, \mathbf{y}) = \langle \Phi_{\mathsf{Mon},:M}(\mathbf{x}), \Phi_{\mathsf{Mon},:M}(\mathbf{y}) \rangle$$

我们用 0 表示这是第“0”版，因为我们会在下一个例子中对它进行改进。但现在请注意，利用 (11) 我们得到的是

$$\mathrm{k}^0_{\mathsf{Mon},:M}(\mathbf{x}, \mathbf{y}) = \sum_{m=0}^{M} \frac{\langle \mathbf{x}^{\otimes m}, \mathbf{y}^{\otimes m} \rangle}{(m!)^2} = \sum_{m=0}^{M} \frac{\langle \mathbf{x}, \mathbf{y} \rangle^m}{(m!)^2}. \tag{12}$$

这已经表明，计算成本关于 d 呈线性。回想一下霍纳法，我们还可以进一步提高效率：公式 (12) 是关于 $\langle \mathbf{x}、\mathbf{y} \rangle$ 的多项式，可以高效地计算为

$$k^0_{\mathsf{Mon},:M}(\mathbf{x}, \mathbf{y}) = 1 + \langle \mathbf{x}, \mathbf{y} \rangle \left(1 + \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{2^2} \left(1 + \dots \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{(M-1)^2} \left(1 + \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{M^2}\right)\right)\right), \tag{13}$$

这种“霍纳型”求值有三个优点：首先，相较于式 (12) 的朴素求值，它只需要 M 次加法和 M 次乘法；其次，它在数值上稳定，这也与式 (12) 形成对比——式 (12) 会将尺度不同的标量的幂相加，使浮点运算变得困难；第三，它让我们能够用核 $\mathrm{k}(\mathbf{x},\mathbf{y})$ 替换内积 $\langle \mathbf{x},\mathbf{y} \rangle$，我们将在下面的玩具示例中对此进行说明。

玩具示例 2：组合非线性函数。现在我们有了一种高效计算 $\mathrm{k}^0_{\mathsf{Mon}}$ 的方法，但众所周知，这种核在实践中表现不佳 [62]。原因正如前面所提到的：尽管单项式具有良好的理论性质，但它们的表达能力不足；在实践中，其他核函数往往表现更好，例如基于其他非线性函数内积的 $\mathbb{R}^d$，如 RBF 核。

一种自然的处理方式是首先应用一个非线性映射

$$\varphi : \mathbb{R}^d \to H$$

并且只有在此之后才计算 H 中的单项式，也就是说，我们考虑的是

$$\Phi_{\mathsf{Mon},:M}\left(\varphi(\mathbf{x})\right) \equiv \left(1, \varphi(\mathbf{x}), \frac{\varphi(\mathbf{x})^{\otimes 2}}{2!}, \ldots, \frac{\varphi(\mathbf{x})^{\otimes M}}{M!}, 0, 0, \ldots\right). \tag{14}$$

这保留了代数结构，同时也支持表达能力强得多的特征映射，因为 φ 通过 φ添加了（无穷多个）额外的非线性函数。如果我们将这一底层映射取为恒等映射 $\varphi = \mathrm{id}$，那么 (14) 与 (4) 一致；但一般来说，我们可以使用任意非线性函数 φ。遗憾的是，实际应用中无法直接使用映射 $\Phi_{\mathsf{Mon}} \circ \varphi$，因为当 H 是无限维时，即使对于 $M = 2$，我们也无法直接计算 φ。这里的关键观察是，如果我们将 $\varphi(\mathbf{x}) = \mathrm{k}(\mathbf{x}\ \cdot)$ 取为核 $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 的特征映射，且该核的再生核希尔伯特空间为H，那么内积

$$\mathrm{k}_{\mathsf{Mon},:M}(\mathbf{x},\mathbf{y}) := \langle \Phi_{\mathsf{Mon},:M} \circ \varphi(\mathbf{x})\, \Phi_{\mathsf{Mon},:M} \circ \varphi(\mathbf{y}) \rangle = \langle \Phi_{\mathsf{Mon},:M} \circ \mathrm{k}(\mathbf{x}\ \cdot)\, \Phi_{\mathsf{Mon},:M} \circ \mathrm{k}(\mathbf{y}\ \cdot) \rangle \tag{15}$$

可以精确计算！事实上，这只需将欧几里得内积 $\langle \mathbf{x},\mathbf{y} \rangle$替换为公式 (12) 和 (13) 中的 $\mathrm{k}(\mathbf{x},\mathbf{y})$。注意，核 $\mathrm{k}_{\mathsf{Mon},:M}$ 取决于 M 和 φ；对这些量的最佳选择非常重要，我们将在第 7.3 节中进一步讨论这一点。

虽然计算本身很简单，但令人惊讶的是，尽管 $\Phi_{\mathsf{Mon},:M} \circ \varphi$ 的陪域实际上是无限维的，我们仍然能够完成这一计算；这种以失去对特征映射的直接访问能力为代价，换取计算内积能力的权衡，通常称为核技巧。因此，核 $\mathrm{k}_{\mathsf{Mon},:M}$ 不仅继承了单项式良好的代数性质，还继承了通过 k 访问无限维 H 的检验效能。

7 签名核

现在，我们将第 6 节中的相同核化方法应用于签名特征映射Φ_{Sig}，以获得一个可计算的核，并赋予它更多非线性函数。我们先考虑一个核

$$\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \tag{16}$$

在我们路径的状态空间上；此后我们将 k 称为静态核。我们不再将 $\mathbb{R}^d$ 中的元素提升到 k 的再生核希尔伯特空间 H 中，而是用它将路径 $\mathbf{x}$ 映射为，

$$t \mapsto \mathbf{x}(t) \in \mathbb{R}^d,$$

一条路径 $\mathbf{k_x}$,

$$t \mapsto \mathrm{k}_{\mathbf{x}}(t) \coloneqq \mathrm{k}(\mathbf{x}(t), \cdot) \in H,$$

它现在在 H 中演化。对于一般的静态核 $\mathbf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$，其再生核希尔伯特空间 H 将是无限维的，因此路径 $\mathbf{k_x} : t \mapsto \mathbf{k_x}(t)$ 会在无限维空间中演化。我们的目标是利用路径 $\mathbf{k_x}$ 的签名，

$$\Phi_{\mathsf{Sig}}(\mathrm{k}_{\mathbf{x}}) = \left(1, \int d\,\mathrm{k}_{\mathbf{x}}, \int d\,\mathrm{k}_{\mathbf{x}}^{\otimes 2}, \dots\right), \tag{17}$$

以及它的截断版本

$$\Phi_{\mathsf{Sig},:M}(\mathrm{k}_{\mathbf{x}}) \coloneqq \left(1, \int d\,\mathrm{k}_{\mathbf{x}}, \int d\,\mathrm{k}_{\mathbf{x}}^{\otimes 2}, \dots, \int d\,\mathrm{k}_{\mathbf{x}}^{\otimes M}, 0, 0, \dots\right) \tag{18}$$

来表示路径 $\mathbf{x}$。再次强调，直观上

$$\Phi_{\mathsf{Sig}}(\mathrm{k}_{\mathbf{x}}) \in \prod_{m \geq 0} H^{\otimes m}$$

比 $\Phi_{\mathsf{Sig}}(\mathbf{x})$ 的表达能力强得多，因为它会在应用核 $\mathbf{k}$ 的非线性函数后计算路径的迭代积分。因此，当 $\mathbf{k}$ 是非线性的且其再生核希尔伯特空间很大时，我们就会有丰富得多的（迭代积分）坐标可供选择。

当然，坏消息仍然是，即使对于 $m = 2$，当 H 是无限维时，精确存储 $\int d\mathbf{k}_{\mathbf{x}}^{\otimes 2} \in H^{\otimes 2}$ 通常也是不可行的，这正是我们感兴趣的情形。[43] 中的主要洞见是，也许会令人惊讶的是，内积

$$\mathrm{k}_{\mathsf{Sig},:M}(\mathbf{x}, \mathbf{y}) \coloneqq \langle \Phi_{\mathsf{Sig},:M}(\mathrm{k}_{\mathbf{x}}), \Phi_{\mathsf{Sig},:M}(\mathrm{k}_{\mathbf{y}}) \rangle \tag{19}$$

当 $M < \infty$ 为分段线性路径 $\mathbf{x}$,$\mathbf{y}$ 时，可以精确计算；或许不那么令人意外的是，它继承了签名特征的许多优良性质。

定理 1. 设$\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 是一个核，并设 $M \in \mathbb{N}$，并设

$$\mathrm{k}_{\mathsf{Sig},:M} : \mathcal{X}_{1\text{-}var} \times \mathcal{X}_{1\text{-}var} \to \mathbb{R} \tag{20}$$

如 (19) 中所定义。随后，对于分段线性路径$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_{L_\mathbf{x}})$ 和 $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_{L_\mathbf{y}})$，可以在 $O(M^3 L^2)$ 个计算步骤内精确计算 $\mathrm{k}_{\mathsf{Sig},:M}(\mathbf{x}, \mathbf{y})$，其中 $L = \max(L_\mathbf{x}, L_\mathbf{y})$。

即使在 $M = \infty$的情况下，当 $\mathbf{k}^{\mathsf{Sig}}(\mathbf{x}, \mathbf{y}) \equiv \langle \Phi_{\mathsf{Sig},:\infty}(\mathbf{x}) \Phi_{\mathsf{Sig},:\infty}(\mathbf{y}) \rangle \equiv \langle \Phi_{\mathsf{Sig}}(\mathbf{x}), \Phi_{\mathsf{Sig}}(\mathbf{y}) \rangle$时，对于分段线性路径，$\mathbf{k}^{\mathsf{Sig}}(\mathbf{x}, \mathbf{y})$也存在良好的近似，我们将在下一节进一步讨论。

7.1 近似

上述定理 1 提供了一种强大的方法来计算签名核，并支持多种应用。然而，除了精确计算 $\mathbf{k}_{\mathsf{Sig},:M}$ 之外，近似也因各种原因而十分重要。

格拉姆矩阵近似。在机器学习语境中，计算瓶颈不仅在于计算单个 $\mathbf{k}_{\mathsf{Sig},:M}(\mathbf{x}、\mathbf{y})$，还在于计算格拉姆矩阵

$$G = (\mathrm{k}_{\mathsf{Sig},:M}(\mathbf{x}, \mathbf{y}))_{\mathbf{x},\mathbf{y} \in \mathcal{D}}$$

其中， $\mathcal{D} \subset \mathcal{X}_{\text{seq}}$ 是我们获得的有限数据集。然而，由于数据集可能非常庞大，通常无法直接计算该矩阵的 $|\mathcal{D}|^2$ 个元素。可以应用诸如 Nystr¨om 方法之类的通用核学习方法，以低秩矩阵近似替代原始格拉姆矩阵，其复杂度可关于 $|\mathcal{D}|$ 呈线性增长。类似地，在高斯过程（GP）中，不仅需要近似 G，还需要在贝叶斯更新规则中计算其逆矩阵。这一过程的复杂度甚至超过数据点数量的平方级；同样，也已有一些通用的 GP 方法来处理这一问题。

未截断签名核的近似。 我们将截断级别 $M \in \mathbb{N}_\infty$视为一个应根据数据进行调优的超参数。虽然定理 1 给出了有限 M 时的精确算法（在实践中，定理 1 的算法适用于 $M < 10$），但我们希望了解当 M 很大甚至为 $M = \infty$时，签名核的近似效果如何。

从离散时间到连续时间的近似。 有时，数据本身确实是一个离散时间序列，但有时，底层对象是一条连续时间路径，而我们仅获得了它的离散时间样本。签名核对连续时间路径有明确定义，因此可用于量化时间离散化误差。这些极限路径通常不具有有界变差，例如，布朗运动仅具有有限的p-变差（对于 $p > 2$），因此需要格外谨慎。

上述三点分别涉及不同的问题，下面将对此进行更详细的讨论。

格拉姆矩阵近似。核学习中的一个普遍问题是，格拉姆矩阵的计算量会随数据点数量 $N = |\mathcal{D}|$ 呈二次方增长 $\mathcal{D}$。具体而言，每个数据点$\mathbf{x} \in \mathcal{D}$ 都对应一个序列，因此我们需要计算签名核 N^2 次。即使单次计算 $\mathsf{ks}_{\mathsf{ig},:M}$ 的开销很小，这种二次复杂度也会很快变得难以承受。适用于任意核 k 的通用方法是 Nyström 近似；其最简单的形式是对一般核的格拉姆矩阵进行近似

$$G = (\mathrm{k}(\mathbf{x}_i, \mathbf{y}_j))_{i,j=1,\dots,N}$$

通过从 $\{1, \dots, n\}$ 中均匀随机抽取一个大小为 r 的子集 J，并定义

$$\tilde{G} = CW^{-1}C^\top$$

˜其中 $C = (\mathsf{k}(\mathbf{x}_i, \mathbf{x}_j))_{i=1,\dots,N,\, j\in J}$，且 $W = (\mathsf{k}(\mathbf{x}_i, \mathbf{x}_j))_{i,j\in J}$。注意， G是一个秩为 r的随机矩阵，因此计算成本按 r^2 增长；当 $r \ll N$时，计算量可大幅降低。除了对列进行均匀子采样之外，还有更复杂的 Nyström 近似方案 [23]。所有这些通用方法都可应用于签名核：选择 J 相当于从数据集中选取一部分序列。不过，这只能解决序列数量 N 带来的复杂度问题，却无法解决另一个计算瓶颈，即序列长度带来的二次复杂度 L。解决后者的一个简单方法是对每个序列进行子采样；这虽然可行，但仍是蛮力方法。更精细的方案是在时间点层面也应用 Nyström 近似； [43][算法 5] 表明，这两种 Nyström 近似可以同时进行：简单来说，不仅要重新加权各个序列的重要性，也要重新加权这些序列中各时间点的重要性。

矩阵计算的成本很快变得难以承受的另一种情况是高斯过程。我们将在第9节讨论使用签名核的高斯过程，并】【。

同样，通用的高斯过程近似方法也适用。不过，我们还可以利用签名核的特定结构，进一步降低算法的复杂度。这可以通过选择特殊的张量 [68] 或使用对角化论证 [46]来实现。

截断近似。 签名张量呈阶乘级衰减（见表 1），这使我们能够对截断签名与未截断签名之间的差异进行如下量化。

命题 3. 设 $\Phi_{\mathsf{Sig},:\infty} = \Phi_{\mathsf{Sig}}$ 为未截断的签名映射，并设 $M > 0$。那么，对于任意 $\mathbf{x} \in \mathcal{X}_{1\text{-}var}$，我们有

$$\|\Phi_{\mathsf{Sig},:\infty}(\mathrm{k}_{\mathbf{x}}) - \Phi_{\mathsf{Sig},:M}(\mathrm{k}_{\mathbf{x}})\| \leq \frac{e^{\|\,\mathrm{k}_{\mathbf{x}}\,\|_1}\,\|\,\mathrm{k}_{\mathbf{x}}\,\|_1^{M+1}}{(M+1)!}. \tag{21}$$

证明见 [16, 引理 46]。尽管阶乘衰减很理想，我们仍要强调，序列的变差范数会出现在其中。此外，通过应用柯西-施瓦茨不等式，我们可以得到核的截断误差的类似界。

命题 4. 令 $\mathrm{k}_{\mathsf{Sig},:\infty}$ 为未截断的签名核，令 $\mathrm{k}_{\mathsf{Sig},:M}$ 为截断至 $M > 0$ 的签名核。则对于任意 $\mathbf{x},\mathbf{y} \in \mathcal{X}_{1\text{-}var}$，我们有

$$|\,\mathrm{k}_{\mathsf{Sig},:\infty}(\mathbf{x},\mathbf{y}) - \mathrm{k}_{\mathsf{Sig},:M}(\mathbf{x},\mathbf{y})| \leq \frac{e^C C^{M+1}}{(M+1)!},$$

其中 $C = \text{最大值}\{\|\,\mathrm{k}_{\mathbf{x}}\,\|_1,\ \|\,\mathrm{k}_{\mathbf{y}}\,\|_1\}$。

以上定量说明，通过选择足够高的截断级别 M，我们可以很好地近似 $\mathbf{k}_{\mathsf{Sig},\,\infty}$。在实践中，$M = 10$ 是可行的计算，而且由于 $\mathbf{10!} \approx 10^6$，求和中尾项的贡献可以忽略不计。不过，我们要指出，序列的 1-变差以常数的形式出现。因此，阶乘衰减开始起作用的级别 M 由路径长度决定。这与计算矩阵指数的情况类似：矩阵范数决定了这一截断级别，参见 [51]中的图 1。

除了对有限 M 使用精确算法，并将 $\mathbf{k}_{\mathsf{Sig},\,:M}$ 视为 $\mathbf{k}_{\mathsf{Sig},\,:\infty}$的近似之外，也可以通过数值求解偏微分方程 [57]，直接近似 $\mathbf{k}_{\mathsf{Sig},\,:\infty}$。在这种情况下，$\mathbf{k}_{\mathsf{Sig},\,:\infty}$ 的近似误差源于偏微分方程求解器的离散化。具体而言，误差以步长来量化；更多详情请参见 [57]。

离散时间近似。 虽然我们是用连续路径来表述核的，但通过公式 6 中的分段线性插值，便可立即得到适用于离散序列的核。给定一个离散序列 $\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_L) \in \mathcal{X}_{\text{seq}}$ 以及静态核，我们考虑由此诱导出的连续路径 $\mathbf{k}_{\mathbf{x}} : [0, L] \to H$，其定义为

$$\mathrm{k}_{\mathbf{x}}(t) = (1-t)\,\mathrm{k}_{\mathbf{x}_{i-1}} + t\,\mathrm{k}_{\mathbf{x}_i}, \qquad t \in [i-1, 1]. \tag{22}$$

然后，我们可以使用来自 [43] 的算法，结合定理 1 中的复杂度界限，高效地为离散数据计算签名核函数。

我们还可以将序列视为连续路径的离散化。给定一条连续路径 $\mathbf{x} : [0, T] \to \mathbb{R}^d$，我们考虑相对于区间 $[0, T]$ 的划分$\pi = (t_i)_{i=0}^L$的离散化 $\mathbf{x}^\pi \in \mathcal{X}_{\text{seq}}$，其中 $0 = t_0 < \ldots < t_L = T$。将上述插值应用于 $\mathbf{x}^\pi$，我们得到一条由 $\mathbf{x}^\pi$诱导的近似连续路径 $\mathbf{k}_{\mathbf{x}}^\pi$。因此，我们得到一个近似 $\Phi_{\mathsf{Sig},:M}(\mathbf{k}_{\mathbf{x}}^\pi)$，当 $M \to \infty$ 且划分 π 的网格 $\sup_{t_i \in \pi} |t_i - t_{i-1}|$ 趋于零时，它收敛于签名 $\Phi_{\mathsf{Sig}}(\mathbf{k}_{\mathbf{x}})$。收敛速率由以下结果给出；更多估计可参见 [43]。

命题 5 ([43])。 设 $\mathbf{x} \in \mathcal{X}_{1\text{-}var}$ 的定义域为 $[0,T]$，并设 $\pi = (t_i)_{i=0}^{L}$ 是 $[0,T]$ 的一个划分。那么，

$$\|\Phi_{\mathsf{Sig},:\infty}(\mathrm{k}_{\mathbf{x}}) - \Phi_{\mathsf{Sig},:M}(\mathrm{k}_{\mathbf{x}}^{\pi})\| \leq e^{\|\,\mathrm{k}_{\mathbf{x}}\,\|_1} \|\,\mathrm{k}_{\mathbf{x}}\,\|_1 \cdot \max_{i=1,\dots,L} \|\,\mathrm{k}_{\mathbf{x}}\,|_{[t_{i-1},t_i]}\|_1. \tag{23}$$

7.2 算法

只有当底层静态核 $\mathrm{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 的再生核希尔伯特空间 H是低维的情况下，才能通过内积直接计算签名核 $\mathrm{k}^{\mathsf{Sig}} : \mathcal{X}_{\mathrm{seq}} \times \mathcal{X}_{\mathrm{seq}} \to \mathbb{R}$。如果 $\mathcal{X} = \mathbb{R}^d$且 k 是欧几里得内积，那么 $\mathrm{k}^{\mathsf{Sig}}$ 就只是普通签名的内积，而即使如此，计算成本也会迅速变得很高。然而，直觉上，通过首先将路径提升到一个在高维或无限维再生核希尔伯特空间 H中演化的路径，我们能够获得更强的表达能力。事实上，经验数据也表明了这一点，参见例如 [68],，即使用径向基函数、Matérn 核或拉普拉斯核等常见选择作为静态核，会得到一个签名核 $\mathrm{k}^{\mathsf{Sig}}$，其表现优于由 $\mathbb{R}^d$上的内积核所诱导的「平凡」签名核。对于非平凡核，无法直接进行计算，因为一般核 k 的再生核希尔伯特空间 H 是无限维的，而且我们无法计算在一般再生核希尔伯特空间 H中演化的路径的签名。不过，至少存在两种不同的算法。

动态规划。 其思路是仿照第 6 节中玩具示例 1 和玩具示例 2 对单项式核 $\mathrm{k}_{\mathsf{Mon}}$in $\mathbb{R}^d$ 的论证。具体而言，给定 $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_L) \in \mathcal{X}_{\mathrm{seq}}$，我们考虑 $\mathrm{k}_{\mathbf{x}}$ 的签名，也就是将$\mathbf{x}$ 提升为式 22 中定义的 H 序列。首先，一条线性线段的签名（见式 7）为

$$\Phi_{\mathsf{Sig}}(\mathrm{k}_{\mathbf{x}}\,|_{[i-1,i]}) = \Phi_{\mathsf{Mon}}(\delta_i\,\mathrm{k}_{\mathbf{x}}) = \left(1, \delta_i\,\mathrm{k}_{\mathbf{x}}, \frac{(\delta_i\,\mathrm{k}_{\mathbf{x}})^{\otimes 2}}{2!}, \dots\right),$$

其中 $\delta_i\,\mathrm{k}_{\mathbf{x}} \coloneqq \mathrm{k}^{\mathbf{x}_{i+1}} - \mathrm{k}(\mathbf{x}_i, \cdot) \in H$。其次，利用所谓的陈恒等式（该恒等式将序列拼接表示为张量乘法），我们可以将整个分段线性路径的签名表示为

$$\Phi_{\mathsf{Sig}}(\mathrm{k}_{\mathbf{x}}) = \Phi_{\mathsf{Mon}}(\delta_1\,\mathrm{k}_{\mathbf{x}}) \otimes \dots \otimes \Phi_{\mathsf{Mon}}(\delta_L\,\mathrm{k}_{\mathbf{x}}).$$

次数 m 分量可以显式表示为

$$\Phi_{\mathsf{Sig},m}(\mathrm{k}_{\mathbf{x}}) = \sum_{\mathbf{i}:1\leq i_1\leq\dots\leq i_m\leq L} \frac{1}{\mathbf{i}!}\delta_{i_1}\,\mathrm{k}_{\mathbf{x}} \otimes \dots \otimes \delta_{i_m}\,\mathrm{k}_{\mathbf{x}} \in H^{\otimes m},$$

其中，对于 $\mathbf{i} = (i_1, \dots, i_m)$，我们定义 $\mathbf{i}! \coloneqq n_1! \cdots n_k!$，若$\mathbf{i}$ 包含 $k = |\{i_1, \dots, i_m\}|$ 个不同元素，且 $n_1, \dots, n_k$ 表示它们在 $\mathbf{i}$ 中出现的次数。因此，截断签名核可以写为

$$\mathrm{k}_{\mathsf{Sig},:M}(\mathbf{x}, \mathbf{y}) = \sum_{m=0}^{M} \sum_{\substack{\mathbf{i}:1\leq i_1\leq\dots\leq i_m\leq L \\ \mathbf{j}:1\leq j_1\leq\dots\leq j_m\leq L}} \frac{1}{\mathbf{i}!\cdot\mathbf{j}!} \prod_{r=1}^{m} \nabla^{\mathrm{k}}_{i_r,j_r}(\mathbf{x}, \mathbf{y}),$$

其中

$$\nabla^{\mathrm{k}}_{i,j}(\mathbf{x}, \mathbf{y}) \coloneqq \langle \delta_i\,\mathrm{k}_{\mathbf{x}}, \delta_j\,\mathrm{k}_{\mathbf{y}} \rangle_H = \mathrm{k}(\mathbf{x}_{i-1}, \mathbf{y}_{j-1}) - \mathrm{k}(\mathbf{x}_i, \mathbf{y}_{i-1}) - \mathrm{k}(\mathbf{x}_{i-1}, \mathbf{y}_j) + \mathrm{k}(\mathbf{x}_i, \mathbf{y}_j).$$

最后，我们注意到，可以应用霍纳法来得到

$$\mathrm{k}_{\mathsf{Sig},:M}(\mathbf{x},\mathbf{y}) = 1 + \sum_{\substack{i_1\geq 1\\ j_1\geq 1}} \nabla^{\mathrm{k}}_{i_1,j_1}(\mathbf{x},\mathbf{y}) \cdot \left(1 + \sum_{\substack{i_2\geq i_1\\ j_2\geq j_1}} \frac{\nabla^{\mathrm{k}}_{i_2,j_2}(\mathbf{x},\mathbf{y})}{C_2(\mathbf{i}_2,\mathbf{j}_2)} \cdot \left(1 + \ldots \sum_{\substack{i_M\geq i_{M-1}\\ j_M\geq j_{M-1}}} \frac{\nabla^{\mathrm{k}}_{i_M,j_M}(\mathbf{x},\mathbf{y})}{C_M(\mathbf{i}_M,\mathbf{j}_M)} \right) \right), \tag{24}$$

其中 $\mathbf{i}_m = (i_1,\ldots,i_m)$、$\mathbf{j}_m = (j_1,\ldots,j_m)$，以及 $C_m(\mathbf{i}_m,\mathbf{j}_m) = n_{\mathbf{i}_m}\cdot n_{\mathbf{j}_m}$，其中 $n_{\mathbf{i}_m}\geq 1$是满足以下条件的最大数：对于所有 $k = m - n_{\mathbf{i}_m},\ldots,m$，$i_k = i_m$。常数$C_m(\mathbf{i}_m,\mathbf{j}_m)$用于考虑 $\frac{1}{\mathbf{i}!\cdot\mathbf{j}!}$ 项中重复计数的情况。采用递归算法实现（24）十分直接，而且兼具经典霍纳法的优点。

偏微分方程求解器。另一种方法见于 [57]，它聚焦于 $M=\infty$的情形。其思路是通过签名核定义函数 $K:[0,L]^2\to\mathbb{R}^d$

$$K(s,t) = \mathrm{k}_{\mathsf{Sig}}(\mathbf{x}|_{[0,s]},\mathbf{y}|_{[0,t]})$$

在将 $\mathbf{x}$ 限制于 $[0,s]$、将$\mathbf{y}$ 限制于 $[0,t]$后计算得到。于是，该函数满足以下偏微分方程

$$\frac{\partial^2 K}{\partial s\partial t}(s,t) = \nabla^{\mathrm{k}}_{\lceil s\rceil,\lceil t\rceil}(\mathbf{x},\mathbf{y})\cdot K(s,t),\quad K(s,0)=K(0,t)=1.$$

这类偏微分方程称为所谓的古尔萨偏微分方程。因此，计算归结为应用偏微分方程求解器；[57] 提出了这样一种方案，并分析了误差率。先验而言，偏微分方程方法聚焦于 $M=\infty$，但对 M 进行交叉验证通常能提升性能，而 [8]则将偏微分方程方法推广为对任意 $M\geq 1$近似 $\mathbf{k}_{\mathsf{Sig}}$。

两种算法都效果良好，并且有可用的软件包[6]，可将底层算法改造为在图形处理器上运行。这两种算法截然不同，即使只计算一次 $\mathbf{k}_{\mathsf{Sig},:M}(\mathbf{x},\mathbf{y})$ 也是如此。对于有限的 M，动态规划方法是精确的；对于 $M=\infty$，近似误差则源于阶乘衰减。相比之下，偏微分方程方法所提供的近似由偏微分方程求解器所用数值格式的性质决定，例如步长。类似地，当需要计算整个格拉姆矩阵时，这两种算法也适用于不同的计算策略。这些选择之间的权衡可能因数据集而异，因此我们建议两种算法都试一试。

7.3 超参数调优

对于给定的参数选择，若已有快速计算格拉姆矩阵的算法，那么绝大部分计算预算都应投入到超参数调优中。一般而言，这对于任何核都成立，但对于签名核尤其如此，因为参数选择会对性能产生巨大影响。签名核具有以下超参数。

[6] `https://github.com/tgcsaba/KSig`和 `https://github.com/crispitagorico/sigkernel`。

1. **静态核参数** Θ_{k}。签名核以定义在 $\mathcal{X}$ 上的核 $\mathbf{k}$ 为输入，并将其转换为定义在 $\mathcal{X}_{\mathrm{seq}}$ 上的核 $\mathbf{k}^{\mathsf{Sig}}$ 。通常，静态核 $\mathbf{k}$ 是参数化的；例如，经典的 RBF 核 $\mathbf{k}(\mathbf{x},\mathbf{y}) = \exp(\theta\|\mathbf{x} - \mathbf{y}\|^2)$ 具有一个超参数 θ。一般而言，超参数并不局限于标量，我们用 Θ_{k} 表示所有可能超参数的集合。因此，签名核 $\mathbf{k}^{\mathsf{Sig}}$ 会继承静态核 $\mathbf{k}$ 的所有这些超参数。

2. **截断参数** $\mathbb{N}_\infty$。在实践中，截断级别 $M \in \mathbb{N}_\infty$ 的最优选择因数据集而异，甚至可能低至 $M = 2$。粗略来说，人们预期静态核 $\mathbf{k}$ 的非线性与最优截断级别 M 之间存在权衡；但即使不进行核化，最优 M 的选择也很微妙， [27]。事实上，一个极端情况是采用平凡的内积核，实验表明此时最优 M 较大；相比之下，对于径向基函数及其他核函数，通常较小的 M 才是最优选择。不过，这些只是经验规律，并且会因数据集而异。

3. **预处理参数** $\Theta_{preprocess}$。在时间序列建模中，预处理原始时间序列是常见做法。例如，人们不使用 $\mathbb{R}^d$ 中的原始序列，而是考虑 $\mathbb{R}^{2d}$ 中的序列，其中额外的坐标通过对原始时间序列进行滞后处理（滑动窗口）得到。类似地，对于多变量时间序列，不同坐标的变化尺度可能不同，因此对它们进行归一化是合理的；例如，参见 [68] ，其中介绍了一些适用于签名核的标准时间序列预处理方法。一般而言，预处理步骤种类繁多，我们用 $\Theta_{preprocess}$ 表示这些步骤。

4. **代数参数** $\Theta_{algebra}$。签名可以通过多种方式推广，也可以用其他对象替代迭代积分。这些对象的具体构造方式超出了本文的讨论范围，但这通常能提升实证性能。我们在此不再进一步讨论，而是请读者参阅 [43, 67] 中的“非几何”签名，它要求我们超越对签名的常规（Hopf）代数处理，见 [22]。由此引入了另一组参数 $\Theta_{algebra}$。

5. **归一化参数** $\Theta_{normalization}$。正如我们将在下一节讨论的那样，归一化能使签名核成为更稳健的统计量；特别是，这解决了上述关于非紧性的问题。归一化有多种方式， [16] 将进一步讨论这一点。此外，我们可以独立地重新缩放不同次数张量的内积，从而得到不同的签名核 [8]。我们将这些归一化参数和重缩放参数记为 $\Theta_{normalization}$。

因此，由参数集 Θ_{k} 的静态核 $\mathbf{k}$ 构建的签名核 $\mathbf{k}^{\mathsf{Sig}}$ 的超参数集为

$$\Theta_{\mathbf{k}^{\mathsf{Sig}}} := \Theta_{\mathrm{k}} \times \mathbb{N}_\infty \times \Theta_{preprocess} \times \Theta_{algebra} \times \Theta_{normalization}.$$

我们强调，与大多数核函数一样，超参数选择对于获得出色性能至关重要。根据问题的不同，需要采用不同的方法；一个基本的区分是有监督问题与无监督问题。

有监督学习。给定一个目标函数，原则上可以通过直接进行交叉验证，利用网格搜索优化 $\theta \in \Theta_{\mathbf{k}^{\mathsf{Sig}}}$。然而，由于 $\Theta_{\mathbf{k}^{\mathsf{Sig}}}$ 通常是高维的，因此贝叶斯优化等更复杂的超参数调优算法会很有用，而且大多数机器学习工具箱都已实现这些算法。无论采用哪种方式，都可以使用标准的超参数调优方法，而简单的网格搜索通常效果不错。

无监督学习。与有监督问题相比，情况不那么简单，因为优化准则往往并非显而易见。更糟的是，对于签名核，所谓的中位数启发式等标准启发式方法 [30] 并无理论依据，实际上通常会导致较差的性能。例如，我们会在第9节的假设检验部分讨论这一点。在那里，中位数启发式用于某些核函数，但没有理由认为这种启发式方法也适用于签名核；事实上，这会导致检验效能较差。一般而言，对于无监督问题，必须考虑具体应用。

8 鲁棒矩、签名与核函数

在第 5 节中，我们看到，当考虑无界定义域时，单项式特征 Φ_{Mon} 会失去逼近函数和刻画随机变量分布的能力。签名特征 Φ_{Sig}通常也会面临这些问题。对于 $\mathcal{K} \subset \mathbb{R}^d$ 的情形，假设单项式定义在紧集上或许是合理的；但对于 $\mathcal{K} \subset \mathcal{X}_{\text{1-var}}$ 的情形，对签名作出这样的假设则要强得多[7] 。

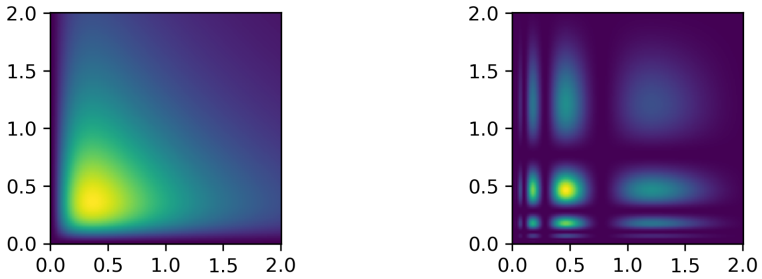

图 2：$\mathbf{V}$ （左）和$\mathbf{W}$ （右）的概率密度函数。

例 1（不具特征性）。 存在两个随机过程$\mathbf{X}$、$\mathbf{Y}$ ，其样本路径均为直线，使得

$$\mathbb{E}[\Phi_{\mathsf{Sig}}(\mathbf{X})] = \mathbb{E}[\Phi_{\mathsf{Sig}}(\mathbf{Y})].$$

例如，可以取

$$\mathbf{X}(t) = t \cdot \mathbf{V} \ \textit{and} \ \mathbf{Y}(t) = t \cdot \mathbf{W}$$

其中$\mathbf{V}$ 是一个取值于 $\mathbb{R}^2$ 、服从如下密度的随机变量

$$\mathbf{V} \sim \rho_1(\mathbf{v}) = \frac{1}{2\pi \cdot \mathbf{v}_1\mathbf{v}_2} \exp\left(\frac{-(\log^2(\mathbf{v}_1) + \log^2(\mathbf{v}_2))}{2}\right)$$

以及$\mathbf{W}$ 取自

$$\rho_2(\mathbf{w}_1, \mathbf{w}_2) = \rho_1(\mathbf{w}_1, \mathbf{w}_2) \prod_{i=1}^{2} \left(1 + \sin(2\pi \log(\mathbf{w}_i))\right).$$

[7]回顾一下，拓扑向量空间局部紧当且仅当它是有限维的。另请参阅注释 5：即使给定一个承载测度几乎全部支撑的紧集，这对于特征性而言仍然不够。

一种方法是对矩的衰减率作出更强的模型假设，从而能够刻画分布；在 $\mathbb{R}^d$ 中，这是经典的，对于路径，我们参考 [14]。然而，在推断的语境下，我们通常不希望作出过强的假设，而且上述例子提醒我们，即使对于平凡的随机过程，也可能出现真正严重的问题。另一个不同但相关的问题是统计学中的稳健性概念。

例2 (不具备稳健性)。 假设$\mathbf{X}$ 是一个服从分布 μ 的实值随机变量。固定$\epsilon \in (0,1)$ 和 $\mathbf{x} \in \mathbb{R}$，并令 $\mathbf{X}_{\mathbf{x},\epsilon}$ 为服从分布 $\mu_{\mathbf{x},\epsilon} = (1-\epsilon)\mu + \epsilon\delta_{\mathbf{x}}$ 的扰动随机变量。当 $\mathbf{x} \gg \mathbb{E}[\mathbf{X}]$，时，很容易看出，离群值狄拉克测度 $\delta_{\mathbf{x}}$ 对均值 $\mathbb{E}[\mathbf{X}_{\mathbf{x},\epsilon}]$ 具有过大的影响，而这种影响在高阶矩 $\mathbb{E}[\mathbf{X}_{\mathbf{x},\epsilon}^{\otimes m}]$ 中会进一步放大。

例2令人担忧的原因是，在实践中，我们只能获得有限个样本来估计统计量，例如 $\mathbb{E}[\mathbf{X}^{\otimes m}]$ 或 $\mathbb{E}[\int d\mathbf{X}^{\otimes m}]$。在实践中，这些有限的样本取自现实世界观测，而数据不干净的情况很常见；也就是说，如上例中的 $\mathbf{x} \gg \mathbb{E}[\mathbf{X}]$ 这类离群值并不少见，原因包括传感器误差、数据记录等[8]。不过，某些统计量仍然可以得到良好的估计，而如何设计此类稳健特征一直是现代统计学的核心课题 [36, 41]。讽刺的是，时间序列的标准机器学习基准测试中，离群值相当少见，但这类离群值在“现实世界”的应用中却经常出现。

8.1 稳健统计学

所谓影响函数 量化了统计量的稳健性。

定义 1 ([36])。 设 $\mathcal{X}$ 是一个拓扑空间， $\mathcal{P}(\mathcal{X})$ 是 $\mathcal{X}$ 上博雷尔概率测度构成的空间。对于映射 $T\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$，定义在 $\mu \in \mathcal{P}(\mathcal{X})$处的影响函数 为

$$\mathsf{IF}(\mathbf{x}; T, \mu) := \lim_{\epsilon \searrow 0} \frac{T(\epsilon\delta_{\mathbf{x}} + (1-\epsilon)\mu) - T(\mu)}{\epsilon}$$

for 对每个使该极限存在的 $x \in \mathcal{X}$。如果 $\sup_{\mathbf{x}\in\mathcal{X}} |\mathsf{IF}(\mathbf{x}; T, \mu)| < \infty$，则称 T 在 μ 处是 B-稳健的 .

例2表明，统计量“ m 阶矩”，

$$T(\mu) = \int \mathbf{x}^{\otimes m}\, \mu(d\mathbf{x})$$

并非 B-稳健；用同样的论证，我们立即可以看出，期望签名并不是用于描述随机过程概率律的稳健统计量。两种情况下的问题都在于，单项式 $\mathbf{x}^{\otimes m}$ （分别为 $\int d\mathbf{x}^{\otimes m}$）会在自变量变大时急剧增大。因此，如果我们在 N 个样本中观察到一个极端值 $\mathbf{x}$，那么最终的经验测度中就会出现一个权重为 $\epsilon = \frac{1}{N}$、位于 $\mathbf{x}$ 处的狄拉克测度，而质量为 $\epsilon\delta_{\mathbf{x}}$ 的 $\mathbf{x}^{\otimes m}$ 加权项可能会主导估计量。

8.2 稳健签名

解决这一问题的一种方法，是在$\mathbf{x}$ 变大时补偿指数增长。对于$\mathbf{x} \in \mathbb{R}^d$的单项式映射，可以通过将$\mathbf{x}^{\otimes m}$ 重新缩放为 $\lambda_m(\mathbf{x})\mathbf{x}^{\otimes m}$来实现，其中 $\lambda_m : \mathbb{R}^d \to \mathbb{R}$ 是一个当 $\|\mathbf{x}\| \to \infty$时快速衰减的函数。惩罚项 λ 的精确选择之所以具有挑战性，是因为它必须同时衰减得足够快，使统计量具有稳健性；但又不能衰减得过快，以保留通用性和

[8]形式上，我们获得的是来自$\mathbf{X}_{\mathbf{x},\ \epsilon}$ 而非来自$\mathbf{X}$的样本

特征性质。具体而言，我们希望得到命题 1 和命题 2 的稳健类比。作为一项技术说明，为了处理非局部紧空间，在接下来的两个定理中，我们为连续有界函数空间赋予严格拓扑。这一拓扑由 [31], 引入，关于其在稳健性方面的应用，可参见 [16]。下文将缩放因子 λ_m 重新表述为单个归一化映射$\lambda : \prod(\mathbb{R}^d)^{\otimes m} \to \prod(\mathbb{R}^d)^{\otimes m}$。

定理 2。 存在一个映射 $\lambda : \prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m} \to \prod_{m\geq 0}(\mathbb{R}^d)^{\otimes m}$，使得

$$\lambda \circ \Phi_{\mathsf{Mon}} : \mathbb{R}^d \to \prod_{m\geq 0} (\mathbb{R}^d)^{\otimes m}, \quad \mathbf{x} \mapsto \lambda(\Phi_{\mathsf{Mon}}(\mathbf{x}))$$

1. 对于配备严格拓扑的 $C_b(\mathbb{R}^d, \mathbb{R})$ 具有通用性，并且

2. 对 $\mathbb{R}^d$ 上的有符号测度空间具有特征性。

此外， $\mu \mapsto \mathbb{E}_{\mathbf{X}\sim\mu}[\lambda \circ \Phi_{\mathsf{Mon}}(\mathbf{X})]$ 是 *B*-稳健的。

特别地，集合 $\{\mathbf{x} \mapsto \langle \mathbf{w}\ \lambda \circ \Phi_{\mathsf{Mon}}(\mathbf{x})\rangle : \mathbf{w} \in \bigoplus(\mathbb{R}^d)^{\otimes m}\}$ 在严格拓扑下于 $C_b(\mathbb{R}^d, \mathbb{R})$ 中稠密，并且归一化矩 $\mathbb{E}[\lambda \circ \Phi_{\mathsf{Mon}}(\mathbf{X})]$ 总能刻画 $\mathbf{X}$ 的分布。函数 λ 的存在并非显而易见；尽管 λ 未以显式形式给出，但可以高效计算，见 [16, 命题 14]。我们发现，同样的论证也适用于签名。

定理 3。 存在一个映射 $\lambda : \prod_{m\geq 0} \bar{H}^{\otimes m} \to \prod_{m\geq 0} \bar{H}^{\otimes m}$，使得

$$\lambda \circ \Phi_{\mathsf{Sig}} : H_{1-var} \to \prod_{m\geq 0} \bar{H}^{\otimes m}, \quad \mathbf{x} \mapsto \lambda(\Phi_{\mathsf{Sig}}(\bar{\mathbf{x}}))$$

1. 对配备严格拓扑的[9] $C_b(\bar{H}_{1-var}, \mathbb{R})$ 具有通用性，并且

2. 对 H_{1-var} 上的有符号测度空间具有特征性。

¯此外， $\mu \mapsto \mathbb{E}_{\mathbf{X}\sim\mu}[\lambda \circ \Phi_{\mathsf{Sig}}(\mathbf{X})]$ 是 *B*-稳健的。

我们在定理 3 中针对希尔伯特空间中演化的路径进行表述，是出于核化的考虑，但对于具有 $\mathsf{k} = \mathsf{id}$ 和 $H = \mathbb{R}^d$ 的普通签名，结论同样成立。此外，若考虑定义在树状路径等价类上的函数或测度，则不添加时间坐标的类似结论也成立，见 [16, 定理 21]。

备注 4 （张量归一化即路径归一化。）。 张量层面的归一化 λ 等价于对底层路径进行归一化。也就是说， $\lambda \circ \Phi_{\mathsf{Sig}}(\mathbf{x})$ 可以改写为 $\Phi_{\mathsf{Sig}}(\mathbf{x} \, n(\mathbf{x}))$，其中函数 $n(\mathbf{x}) \in \mathbb{R}$ 取决于路径 $\mathbf{x}$。这样，我们可以将 λ 视为一种非常具体的数据归一化，见 [16]。

注 5 （近乎紧集的支撑并无帮助。）. 我们强调，对于每个 $\epsilon > 0$，都很容易找到一个承载 $1-\epsilon$ 概率测度质量的紧集。问题在于，我们所关注的函数——单项式或迭代积分——在承载剩余 ϵ 概率测度质量的集合上表现很差。归一化 λ 将这些函数变为有界函数，因此表现良好，同时仍保留必要的代数结构。

[9] 与命题 2 类似， $\bar{H}_{1-var}$ 表示对 H_{1-var} 的路径添加时间坐标后得到的路径。具体而言， $\bar{H}_{1-var}$ 中的路径是在 $H \oplus \mathbb{R}$ 中演化且具有有界 1-变差的路径。

8.3 鲁棒（签名）核

归一化 λ 使我们能够克服特征映射 Φ_{Mon} 和 Φ_{Sig} 在非紧集上的缺陷。给定一个静态核 $\mathrm{k}:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 及其对应的特征映射 $\varphi:\mathcal{X}\to H$，我们可以通过以下方式定义其对应核 $\mathrm{k}_{\mathsf{Mon}}$ 和 $\mathrm{k}^{\mathsf{Sig}}$ 的鲁棒版本：

$$\mathrm{k}_{\mathsf{Mon}}(\mathbf{x},\mathbf{y}) := \langle \lambda\circ\Phi_{\mathsf{Mon}}(\varphi(\mathbf{x})),\ \lambda\circ\Phi_{\mathsf{Mon}}(\varphi(\mathbf{y}))\rangle \text{ 和 } \mathrm{k}^{\mathsf{Sig}}(\mathbf{x},\mathbf{y}) := \langle \lambda\circ\Phi_{\mathsf{Sig}}(\varphi(\mathbf{x})),\ \lambda\circ\Phi_{\mathsf{Sig}}(\varphi(\mathbf{y}))\rangle.$$

稍微滥用一下术语，我们沿用非鲁棒版本的相同记号。上述普适性和特征性结果（定理 2 和定理 3）可立即转化为相应的核化表述。

从计算方面来看，利用 [43, 16], 归一化 λ 的结果本身可以进行核化，e
ker从而得到以下结论。

命题 6. 设 $\mathrm{k}:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 为一个核，$\mathbf{x}$、$\mathbf{y}\in\mathcal{X}_{seq}$ 为两个最大长度为L的序列，M 为截断级别。则鲁棒截断签名核 $\mathrm{k}^{\mathsf{Sig},:M}(\mathbf{x},\mathbf{y})$ 可通过 $O(L^2(c_{\mathrm{k}}+M^3)+q)$ 个计算步骤求得，并使用 $O(M^2L^2+M+q)$ 内存，其中 c_{k} 表示对静态核k 进行单次求值的成本，q 表示求出次数为 $2M$ 的一元多项式的非负根的成本。

命题 6 中的算法其实很简单，只要我们知道如何计算签名核即可，详情请参见 [16] 。

9 应用与进一步发展

拥有一个可高效计算、适用于任意空间中序列数据的核，使我们能够将核方法应用于各种各样的问题。下面我们将讨论其中一些问题。

随机过程的最大均值差异。给定一个集合 $\mathcal{X}$，我们通常关注该空间上概率测度 $\mathcal{P}(\mathcal{X})$ 的空间。我们已经遇到过特征性的概念，它表示我们可以将测度 $\mu\in\mathcal{P}(\mathcal{X})$ 唯一地表示为一个期望，即

$$\mu\mapsto\mathbb{E}_{\mathbf{X}\sim\mu}[\varphi(\mathbf{X})]$$

是单射。如果 φ 的陪域，即特征空间，具有范数，那么这将诱导出一个度量c

$$d(\mu,\nu)=\|\mathbb{E}_{\mathbf{X}\sim\mu}[\varphi(\mathbf{X})]-\mathbb{E}_{\mathbf{Y}\sim\nu}[\varphi(\mathbf{Y})]\|.$$

需要注意的是，这些期望取值于一个无限维空间。然而，对于选择 $\varphi(\mathbf{x})=\mathrm{k}(\mathbf{x},\cdot)$，可以证明上述量能够高效计算；利用 H 是一个再生核希尔伯特空间进行直接计算可得

$$\begin{aligned}\|\mathbb{E}[\varphi(\mathbf{X})]-\mathbb{E}[\varphi(\mathbf{Y})]\|^2 &= \langle\mathbb{E}[\varphi(\mathbf{X})]-\mathbb{E}[\varphi(\mathbf{Y})],\mathbb{E}[\varphi(\mathbf{X})]-\mathbb{E}[\varphi(\mathbf{Y})]\rangle\\ &= \langle\mathbb{E}[\varphi(\mathbf{X})],\mathbb{E}[\varphi(\mathbf{X})]\rangle-2\langle\mathbb{E}[\varphi(\mathbf{X})],\mathbb{E}[\varphi(\mathbf{Y})]\rangle+\langle\mathbb{E}[\varphi(\mathbf{Y})],\mathbb{E}[\varphi(\mathbf{Y})]\rangle\\ &= \mathbb{E}[\mathrm{k}(\mathbf{X},\mathbf{X}')]-2\mathbb{E}[\mathrm{k}(\mathbf{X},\mathbf{Y})]+\mathbb{E}[\mathrm{k}(\mathbf{Y},\mathbf{Y}')],\end{aligned}$$

其中 $\mathbf{X},\mathbf{X}'\sim\mu$ 分别与 $\mathbf{Y},\mathbf{Y}'\sim\nu$ 相互独立。当只给出 μ 和 ν 的有限样本时，右侧很容易估计；特别是，

$$d_{\mathrm{k}}(\mu,\nu)=\sqrt{\mathbb{E}[\mathrm{k}(\mathbf{X},\mathbf{X}')]-2\mathbb{E}[\mathrm{k}(\mathbf{X},\mathbf{Y})]+\mathbb{E}[\mathrm{k}(\mathbf{Y},\mathbf{Y}')]}$$

是一种可以利用有限样本轻松估计的度量。上述度量也称为核最大均值差异（MMD），因为再次通过直接计算可以证明，

$$d_{\mathrm{k}}(\mu,\nu)=\sup_{f\in H:\|f\|\leq 1}\left|\int_{\mathcal{X}}f(\mathbf{x})\mu(d\mathbf{x})-\int_{\mathcal{X}}f(\mathbf{x})\nu(d\mathbf{x})\right|.$$

因此， d_{k} 是我们通过针对 k 的再生核希尔伯特空间（RKHS）单位球中的函数检验测度所能获得的最大差异。最大均值差异（MMD）已在机器学习（ML）中有诸多应用，综述请见 [52] 。有了签名核，我们便可立即得到路径上概率测度的 MMD；特别地，我们得到随机过程概率律的一种度量。签名核的理论保证也由此转化为签名核 MMD 的理论保证 $d_{\mathrm{k_{Sig}}}$。不过，签名 MMD 所诱导的概率测度空间拓扑更为微妙。当 RKHS 是有限维时，它诱导弱收敛；但 RKHS 为无限维时则不然。详情请见 [16]。

假设检验。最受欢迎的假设检验方法之一是双样本检验，而签名核使我们能够将其用于随机过程。特别是，我们可以检验

$$H_0:\mathrm{Law}(\mathbf{X})=\mathrm{Law}(\mathbf{Y})\text{ against }H_1:\mathrm{Law}(\mathbf{X})\neq\mathrm{Law}(\mathbf{Y})$$

其中，$\mathbf{X}=(\mathbf{X}_t)$ 和 $\mathbf{Y}=(\mathbf{Y}_t)$ 是两个随机过程。我们获得的数据是来自 $\mathbf{X}$ 的 m 条样本轨迹，以及来自 $\mathbf{Y}$的 n 条样本轨迹。此类情况并不少见，例如，在一项医学试验中，患者队列被分为两个亚组（例如，一个亚组服用药物，另一个不服用），并在试验期间测量医学标志物。另一个例子见 [2] ，其中将签名 MMD 应用于检验并验证现实世界经济情景中的假设。根据 [33]，若满足以下条件，我们便拒绝 H_0

$$d_{\mathrm{k_{Sig}}}(\hat{\mu}_X,\hat{\mu}_Y)>c(\alpha,m,n). \tag{25}$$

这里，$\hat{}\mu_X=\sum_{i=1}^m\delta_{\mathbf{x}_i}$ 和$\hat{}\mu_Y=\sum_{j=1}^n\delta_{\mathbf{y}_j}$ 是由观测到的序列[10] $\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{y}_1,\ldots,\mathbf{y}_m\in\mathcal{X}_{\mathrm{paths}}$给出的经验测度，而 $c(\alpha,m,n)$是一个常数，使检验（25）的显著性水平为[11] α。在 [33],中，针对$d_{\mathrm{k_{Sig}}}(\hat{}\mu_X,\hat{}\mu_Y)$提出了若干估计量，这些估计量在偏差程度和计算复杂度方面有所不同。函数 c 由 [33], 中的渐近分布确定，但这种选择通常会导致检验标准过于保守，也就是说，很少会拒绝 H_0。在实践中，置换检验通常能在最大均值差异假设检验中取得更好的经验表现，这一点对签名核尤其适用；详情以及与其他假设检验的实验和基准比较，请参阅 [16] 。除了检验分布是否等价之外，核学习还使我们能够检验随机变量 [34] 各坐标之间的独立性。需要注意的是，对于随机过程，检验各坐标之间的独立性颇具挑战，因为不同坐标在不同时间可能存在统计依赖关系。然而，量化并检验随机过程的独立性有重要应用，例如随机过程的独立成分分析[60, 59, 5]。

[10]稍微滥用一下记号，每个$\mathbf{x}_i\in\mathcal{X}_{\mathrm{seq}}$ 表示一个完整的序列，而不是像之前那样表示序列中的单个项。[11]回想一下，检验的第一类错误是错误地拒绝 H_0 的概率，而第二类错误是错误地接受 H_0。如果第一类错误的上界为 α ，那么就称该检验的检验效能为 α。

最后请注意，签名核 $\mathsf{k}_{\mathsf{Sig}}$以及由此得到的最大均值差异 $d_{\mathsf{k}_{\mathsf{Sig}}}$由超参数 $\theta_{\mathsf{k}_{\mathsf{Sig}}} \in \Theta_{\mathsf{k}_{\mathsf{Sig}}}$ 参数化；即使在简单的有监督学习中，超参数的选择对取得良好表现也至关重要。使用标准网格搜索来选择最佳的 θ 不再适用于假设检验，因为使用给定数据会引入相关性，从而使检验失效。这是核最大均值差异在检验问题中的普遍难题，人们已提出多种旨在优化检验效能的解决方案。实践中效果较好的一种做法是遵循 [66] 的建议，用 $d_{\mathsf{k}_{\mathsf{Sig}}}$ 替换为

$$(\mathbf{x}, \mathbf{y}) \mapsto \sup_{\theta \in \Theta} d_{\mathsf{k}_{\mathsf{Sig}}(\theta)}(\mathbf{x}, \mathbf{y})$$

其中，我们用 \(k_{\{v1\}}(\{v2\})\) 表示 \(k_{\{v3\}}\) 对 \(\{v4\}\) 的依赖关系。上述定义仍然构成一种度量，尽管它并非检验效能的最优选择，但能够保证在给定显著性水平下进行样本检验；此外，在实际应用中，它与签名核配合使用时效果良好，详情请参见 [16]。另一个相关问题是静态核 \(k\) 本身的选择：对于此前几乎所有签名核的应用，使用欧几里得内积核作为静态核，其表现都不如 RBF 等非线性静态核。然而，对于假设检验，在某些数据集上，将标准内积作为签名核的静态核，其表现可以与基于非线性静态核构建的签名核相媲美。关于这一问题及其与超参数选择问题之间的关系，我们再次参见 [16] 。

生成模型中的签名 MMD。签名 MMD 是路径空间上任意两个测度之间的度量，并且可以高效计算。特别是，它不对所比较的两个测度作任何参数化假设，因此可用于生成模型，以判断生成器生成的样本是否与给定的目标分布足够接近。这一方法已应用于生成对抗网络（GAN），用于对金融数据建模 [7]，也用于（神经）随机微分方程的生成模型 [42]。同样，从理论角度来看，超参数和静态核的选择颇具挑战性，但这并非签名核所特有，机器学习文献中的通用启发式方法或许也适用。

具有签名协方差的高斯过程。 我们通常不仅需要点预测，还需要估计相关的不确定性。贝叶斯方法提供了一种系统化的方法来更新和量化不确定性；特别是，高斯过程是一种强大且非参数的方法，能够灵活地为数据的函数设置先验；参见 [54]。形式上， $\mathcal{X}$ 上的高斯过程是一种随机过程 $f = (f_{\mathbf{x}})_{\mathbf{x} \in \mathcal{X}}$，由 $\mathbf{x} \in \mathcal{X}$索引，使得 $f \sim N(m, \mathrm{k})$ 由均值和协方差函数确定。

$$m : \mathcal{X} \to \mathbb{R} \text{ and a positive definite } \mathrm{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

其中 $f_{\mathbf{x}}$ 服从高斯分布，且彩票官网

$$\mathbb{E}[f_{\mathbf{x}}] = m(\mathbf{x}) \text{ and } \operatorname{Cov}(f_{\mathbf{x}}, f_{\mathbf{y}}).$$

需要注意的是，在这里我们打破了使用大写字母表示随机变量的惯例，转而写作 f 而非 F，以与高斯过程（GP）文献中的惯例保持一致。不失一般性，可以取 $m(\mathbf{x}) = 0$ ，并在给定核 k 的情况下，考虑任意空间 $\mathcal{X}$ 中的数据。如果核 k 具有良好的性质，这些性质通常也会转化为高斯过程的理想性质。给定签名核，我们可以立即考虑针对序列数据的高斯过程。然而，这会引出两个问题，其中一个更偏理论层面，另一个则更偏应用层面。

- 在非紧空间上，具有连续样本轨迹的高斯过程并非必然存在；简单反例可参见 [1]。确保其存在所需的正则性估计，可利用 Dudley 的经典结果 [24] 推导，并通过界定覆盖数的增长来获得，这需要利用 $\mathbf{k}$ 和 $\mathcal{X}$ 的结构。在 [68], 中，计算了签名核的此类覆盖数估计，由此可得一个以路径为指标的高斯过程存在；此外，还能得到该高斯过程样本路径的连续模$\mathbf{x} \mapsto f_{\mathbf{x}}$。

- 普通高斯过程推断需要对大型格拉姆矩阵进行求逆，其计算复杂度随样本数量 n 呈 $O(n^3)$ 增长。对于大多数数据集而言，这种计算成本过高，因此人们开发了多种近似方法。其中，最受欢迎的当属变分推断和诱导点方法。要将这些方法高效地应用于签名核，需要利用张量的结构。文献 [68] 引入了具有签名协方差的高斯过程，并提出了一种基于稀疏张量的变分方法；而文献 [46] 则提出了一种采用格拉姆矩阵对角近似的变分方法，从而能够快速进行矩阵求逆。

字符串、时间序列与经典序列核。序列的核学习是一个经典课题，其中状态空间为有限或离散的情况——在我们的记号中，$\mathcal{X}$是一个有限或可数的字母集合，而 $\mathcal{X}_{\text{seq}}$ 是由 $\mathcal{X}$ 中的字母构成的字符串集合——受到了广泛关注。Haussler 提出了关系卷积核框架[37], 由此产生了所谓的字符串核；例如，参见 [49, 47] ，了解其在文本学习和生物学中的算法与应用。这些方法通过对齐核 [3, 64, 20] 得到了扩展，使其能够处理一般状态空间，并搜索所有可能的对齐方式（“时间扭曲”），尽管这可能无法得到正定核。然而，经典字符串核和全局对齐核都可以作为签名核构造的特例或微小变体，而其他变体也是可能的，参见 [43, 第 5 节和注释 4。10] 以及 [67, 附录 B。3], 从而为这些核提供了理论保证。

时间参数化（不）变性。签名 $\mathbf{x} \mapsto \Phi_{\mathsf{Sig}}(\mathbf{x})$ 对路径 $\mathbf{x} = (\mathbf{x}(t))_{t\in[0,T]}$ 的起点 $\mathbf{x}(0)$ 的平移以及所谓的“树状”等价 [35]而言是单射。后者意味着路径在重新参数化和回溯意义下等价。具体来说，$\Phi_{\mathsf{Sig}}(\mathbf{x}) = \Phi_{\mathsf{Sig}}(\mathbf{x}^{\rho})$，其中 $\rho : [0, S] \to [0, T]$ 是非递减的，且 $\mathbf{x}^{\rho}(t) \coloneqq \mathbf{x}(\rho(t))$。如第 4 节所示，添加时间坐标 $\bar{\mathbf{x}}(t) \coloneqq (t, \mathbf{x}(t))$ 会消除任何树状等价，即 $\mathbf{x} \mapsto \Phi_{\mathsf{Sig}}(\bar{\mathbf{x}})$ 在 $\mathcal{X}_{\text{1-var}}$ 上是单射，因此能够区分仅相差时间变换的路径。消除这样一大类不变性（时间变换集合是无限维的）是一种强大的工具，在工程学 [56] 领域备受关注；这一方法称为“时间扭曲”。对于上文讨论的字符串离散情形，这称为全局对齐 [44]。第 4 节讨论的相同理论保证，也适用于定义在此类等价类上的函数或概率测度的相应结论。

弱拓扑与适应拓扑。签名 MMD 是定义在 $\mathcal{X}_{\text{seq}}$ 或 $\mathcal{X}_{\text{paths}}$ 上的概率测度空间上的一种度量。可以说，概率测度空间上最重要的拓扑是弱拓扑，我们在上文讨论了签名及签名 MMD 分别在何种条件下能够度量这一拓扑。然而，弱拓扑完全忽略了随机过程的滤过。考虑滤过结构、用于描述随机过程的概率律的更精细拓扑的必要性，早已得到认识 [39], 并且近年来重新引起了广泛关注 [53]。这些更精细的拓扑统称为高秩拓扑（秩越高，所考虑的滤过结构就越多。

考虑在内）。在 [4], 中，引入了考虑滤过的高秩签名，从而能够度量这些高秩拓扑。构造签名核的方法同样可以基于静态核，运用于高秩签名； [58, 40]使用偏微分方程求解器来估计各种机器学习任务中相关的最大均值差异。

粗糙路径。全文中，我们假设路径具有有界变差，也就是说，我们在 $\mathcal{X}_{\text{1-var}}$ 上定义了 kSig 。然而，许多随机过程的样本路径具有无界变差，因此并不属于 $\mathcal{X}_{\text{1-var}}$。借助随机分析和粗糙路径理论中的工具，可以相当直接地解决这一问题。对此有两种理解方式：

(i) 实际上，我们得到的是序列，而不是路径；这些序列总能嵌入为有界变差路径，$\mathcal{X}_{\text{seq}} \hookrightarrow \mathcal{X}_{\text{paths}}$。(ii) 尽管如 (i) 所述，我们可以将任意序列嵌入为有界变差路径，但在高频缩放极限下，这种做法可能会失效，因为迭代积分可能不收敛。更彻底的做法是为所谓的粗糙路径定义签名核。这需要我们指定第一阶迭代积分，因为它们无法定义为黎曼–斯蒂尔切斯积分。 [43, 附录 B] 中，我们为所谓的几何粗糙路径完成了这一步；这类路径的分段线性近似会收敛[12] ，因此可以规范地选取更高阶迭代积分。这涵盖了半鞅以及许多其他常见随机过程。若指定其他的高阶积分选取方式[13] ，同一算法也适用，但这些选取方式通常无法从常见的序列数据集中获得。

超越（几何）粗糙路径。 本文始终聚焦于经典签名，但对于许多应用而言，所谓的非几何粗糙路径或分支粗糙路径提供了另一种将路径描述为张量序列的方式。到目前为止，我们讨论的大部分内容只需稍作修改，就可以推广到分支粗糙路径；也就是说，同样可以定义分支粗糙路径之间的“签名核”[14]。甚至还可以更进一步，认为其核心结构是将一个序列嵌入非交换代数中，并通过代数中的非交换乘法将序列元素“串接”起来。乘法具有非交换性至关重要，因为这能以代数方式捕捉序列的顺序。从这个角度看，我们迄今讨论的签名和核，通过选用所谓的 shuffle Hopf 代数，并以李群指数映射作为嵌入，实现了这一抽象方法；分支粗糙路径则以 Connes–Kreimer 代数取代 shuffle 代数。遵循这一思路，我们可以直接对代数结构进行参数化，并从数据中学习最佳选择。相关理论超出了本文的讨论范围，但请参阅 [67]，其中展示了如何通过参数化底层代数结构并超越几何粗糙路径来提升基准测试表现（往往提升显著），以及如何应对计算方面的挑战。同样，请参阅关于

[12]在 [43, 附录 B], $\mathrm{k}^{+}\ (d,\ m)$ 的记号中，其中 $d = \lfloor p \rfloor$ 表示 p 粗糙路径的 p-变差。在我们当前的记号中， d， m 是参数 $\theta_{algebra} \in \Theta_{algebra}$ 的一部分，而该参数属于 kSig。[13]例如，为了描述 d 维空间中的一条“粗糙”路径，可以考虑状态空间 $\mathcal{X} = G_{m,\ d}$ 中的序列 $\mathbf{x} = (\mathbf{x}_0,\ \ldots\ldots,\ \mathbf{x}_L)$ ，其中的元素是类群元素；非正式地说，$\mathbf{x}_i$ 是底层路径在区间 $[t_i, t_{i+1}]$ 上的前 m 个“迭代积分”。按照第 7.3 节的记号，从原始序列数据生成这类类群元素可视为一个预处理步骤。[14]这一点见 arXiv 版本 v1 中的 [17, 附录 D]。为简洁起见，附录 D 不属于 [16] 的已出版版本[17] 。

非几何签名核[15] 的讨论，见 [43, 注释 4.5]。除了提升表现之外，这类变化还会引出有趣且颇具挑战性的代数问题，而这些问题已无法在 Hopf代数框架内处理。有关底层代数结构的详细研究，请参阅 [22] 。

演化（非欧氏）数据。 签名核将定义域 $\mathcal{X}$ 上的任意静态核转换为相应序列空间 $\mathcal{X}_{\text{seq}}$上的核，使我们能够为一般（非欧氏）空间中演化的对象设计核函数。一个例子来自拓扑数据分析领域，其中一种称为持久同调的方法，可以量化数据的多尺度拓扑结构‘，例如高度非线性对象持久性图中的点云 [21]。在 [15], 中，静态持久性图被转换为路径，而签名核能够捕捉静态点云的拓扑性质。给定一个演化的点云，例如表示一个智能体群，我们可以获得一系列持久性图，用于概括数据的动态拓扑结构，而签名核同样为执行回归等学习任务提供了有效方法 [32]。另一个例子基于这样的观察：签名的定义只需要序列导数上的内积。对于在李群上演化的路径，其导数位于相应的李代数中，因此可以完全以相同的方式构造签名核 [45]。最后，利用签名核颇具吸引力的理论性质，还可以通过签名核研究一种适用于序列的一般方法，以分析经典的深度学习架构；例如， [28] 从签名核的角度研究循环神经网络，而 [18] 则从这一角度研究残差网络的缩放。

无似然推断。 给定一个参数化统计模型，似然函数能够刻画该模型对数据的拟合程度。通常，底层模型较为复杂，导致似然函数难以处理，从而无法实际求得最优参数。所谓无似然推断旨在解决这些挑战， [25] 将签名核及其最大均值差异与所谓的密度比估计相结合，从而开发出一种无似然推断方法；另请参见 [26] ，了解其在所谓近似贝叶斯计算中的应用。

致谢

HO和DL获得了香港创新科技署（InnoHK CIMDA项目）的支持。HO获得了英国工程与自然科学研究委员会 [资助，资助编号为EP/S026347/1]。

参考资料

[1] R.J. Adler 和 J.E. Taylor。《随机场与几何》。数学专著丛书。施普林格纽约，2009年。[2] Herv´e Andr`es、Alexandre Boumezoued 和 Benjamin Jourdain。基于签名的现实世界经济情景验证。*arXiv* 预印本 *arXiv:2208.07251*，2022年。[3] Claus Bahlmann、Bernard Haasdonk 和 Hans Burkhardt。使用支持向量机进行在线手写识别——一种核方法。载于第八届国际手写识别前沿研讨会论文集，第49–54页。IEEE，2002年。

[15]在 [43] 的记法中，记作 $\mathrm{k}^{+}{}_{(d,m)}$，其中 $d \neq m$ 。

[4] Patric Bonnier、Chong Liu 和 Harald Oberhauser。适应拓扑与高阶签名。《应用概率年刊》，33(3):2136–2175，2023年。[5] Patric Bonnier 和 Harald Oberhauser。随机过程的签名累积量、有序划分与独立性。《伯努利》，26(4):2727–2757，2020年。[6] Roger W Brockett。Volterra级数与几何控制理论。《自动化》，12(2):167–176，1976年。[7] Hans Buehler、Blanka Horvath、Terry Lyons、Imanol Perez Arribas 和 Ben Wood。面向小数据环境的数据驱动型市场模拟器。arXiv 预印本 arXiv:2006.14498，2020年。[8] Thomas Cass、Terry Lyons 和 Xingcheng Xu。一般签名核。arXiv:2107.00447，2021年。[9] Thomas Cass 和 William F. Turner。无参数化路径空间上的拓扑。arXiv:2206.*11153 [*math*]*,2022年6月。[10] Kuo-Tsai Chen。迭代积分与指数同态ŧ。《伦敦数学会会刊》，s3-4(1):502–512，1954年。[11] Kuo-Tsai Chen。路径积分、几何不变量与广义Baker-Hausdorff公式。《数学年刊》，65(1):163–178，1957年。[12] Kuo-Tsai Chen。路径积分——通过非交换形式幂级数忠实表示路径。《美国数学会汇刊》，89(2):395–407，1958年。[13] Ilya Chevyrev 和 Andrey Kormilitzin。机器学习中的签名方法入门。arXiv:1603.*03788 [*cs, stat*]*, 2016.[14] Ilya Chevyrev 和 Terry Lyons。几何粗糙路径上测度的特征函数。《概率年刊》，44(6):4049–4082，2016年。[15] Ilya Chevyrev、Vidit Nanda 和 Harald Oberhauser。拓扑数据分析中的持久路径与签名特征。IEEE模式分析与机器智能汇刊，42:192–202，2018年。[16] Ilya Chevyrev 和 Harald Oberhauser。用签名矩刻画随机过程的分布。机器学习研究杂志，23(176):1–42，2022年。[17] Ilya Chevyrev 和 Harald Oberhauser。用签名矩刻画随机过程的分布。arXiv预印本1810.10971.v1，2022年。[18] Nicola Muca Cirone、Maud Lemercier 和 Cristopher Salvi。神经签名核作为受控残差网络的无限宽度与深度极限。arXiv:2303.*17671 [*cs, math*]*, 2023年3月。[19] Felipe Cucker 和 Steve Smale。学习的数学基础。《美国数学学会通报》，39(1):1–49，2002年。[20] Marco Cuturi、Jean-Philippe Vert、Øystein Birkenes 和 Tomoko Matsui。基于全局对齐的时间序列核。2007年IEEE国际声学、语音与信号处理会议——ICASSP ’07，2:II–413–II–416，2006年。[21] Tamal Krishna Dey 和 Yusu Wang。《数据分析计算拓扑学》。剑桥大学出版社，2022年2月。

[22] Joscha Diehl、Kurusch Ebrahimi-Fard 和 Nikolas Tapia。广义迭代和签名。arXiv 预印本 arXiv:2012.04597，2020 年。[23] Petros Drineas、Michael W Mahoney 和 Nello Cristianini。关于用于改进基于核的学习的格拉姆矩阵近似方法 Nyström。机器学习研究杂志，6(12)，2005 年。[24] Richard M Dudley。高斯过程的样本函数。载于 RM Dudley 选集，第 187–224 页。施普林格，2010 年。[25] Joel Dyer、Patrick Cannon 和 Sebastian M. Schmon。利用签名比率估计，为昂贵的时间序列模拟器进行摊销式无似然推断。载于国际人工智能与统计学会议，2022 年。[26] Joel Dyer、John Fitzgerald、Bastian Rieck 和 Sebastian M Schmon。基于签名最大均值差异的面板数据近似贝叶斯计算。载于 NeurIPS 2022 时序图学习研讨会，2022 年。[27] Adeline Fermanian。基于截断签名的函数型线性回归。《多元分析杂志》，192:105031，2020 年。[28] Adeline Fermanian、Pierre Marion、Jean-Philippe Vert 和 G´erard Biau。将循环神经网络构建为核方法：一种神经常微分方程方法。《神经信息处理系统进展》，34:3121–3134，2021 年。[29] Michel Fliess。非线性因果泛函与非交换不定元。《法国数学会通报》，109:3–40，1981 年。[30] Damien Garreau、Wittawat Jitkrittum 和 Motonobu Kanagawa。中位数启发式的大样本分析。arXiv:1707.*07269 [*math，stat*]*,2018 年 10 月。[31] Robin Giles。严格拓扑的一种推广。《美国数学会汇刊》，161:467–474，1971 年。[32] Chad Giusti 和 Darrick Lee。签名、Lipschitz 自由空间与持久性图的路径。arXiv:2108.*02727 [*math，stat*]*,2021 年 8 月。[33] Arthur Gretton、Karsten M. Borgwardt、Malte J. Rasch、Bernhard Sch¨olkopf 和 Alexander Smola。核双样本检验。《机器学习研究杂志》，13:723–773，2012 年 3 月。[34] Arthur Gretton、Kenji Fukumizu、Choon Teo、Le Song、Bernhard Sch¨olkopf 和 Alex Smola。核独立性统计检验。《神经信息处理系统进展》，20，2007 年。[35] Ben Hambly 和 Terry Lyons。有界变差路径的签名唯一性与约化路径群。《数学年刊》，171(1):109–167，2010 年。[36] Frank R. Hampel。《稳健统计学：基于影响函数的方法》。威利，1986 年。[37] David Haussler 等。离散结构上的卷积核。技术报告，Citeseer，1999 年。[38] Chris C Heyde。关于对数正态分布的一种性质。《英国皇家统计学会杂志：B辑（方法论）》，25(2):392–393，1963 年。

[39] Douglas N. Hoover 和 H. Jerome Keisler。适应概率分布。美国数学会汇刊，286:159–201，1984 年。[40] Blanka Horvath、Maud Lemercier、Chong Liu、Terry Lyons 和 Cristopher Salvi。通过分布回归实现最优停止：一种更高秩的签名方法，2023 年。[41] Peter J. Huber 和 Elvezio M. Ronchetti。稳健统计学。John Wiley & Sons，2011 年 9 月。[42] Patrick Kidger、James Foster、Xuechen Li、Harald Oberhauser 和 Terry J Lyons。作为无限维 GAN 的神经随机微分方程。载于国际机器学习会议，第 5453–5463 页。PMLR，2021 年。[43] Franz J. Kiraly 和 Harald Oberhauser。适用于顺序排列数据的核。《机器学习研究杂志》，20(31):1–45，2019 年。[44] Joseph B Kruskal。序列比较概述：时间扭曲、字符串编辑与大分子。*SIAM* 评论，25(2):201–237，1983 年。[45] Darrick Lee 和 Robert Ghrist。李群上的路径签名。*arXiv:2007.06633 [cs,math, stat]*, 2020。[46] Maud Lemercier、Cristopher Salvi、Thomas Cass、Edwin V Bonilla、Theodoros Damoulas 和 Terry J Lyons。Siggpde：扩展序列数据上的稀疏高斯过程。载于国际机器学习会议，第 6233–6242 页。PMLR，2021 年。[47] Christina S. Leslie 和 R. Kuang。利用非精确匹配对蛋白质序列进行快速字符串核计算。《机器学习研究杂志》，5:1435–1455，2004 年。[48] Christian Litterer 和 Harald Oberhauser。关于扩散泛函的 Chen–Fliess 近似。《数学月刊》，175(4):577–593，2014 年 12 月。[49] Huma Lodhi、Craig Saunders、John Shawe-Taylor、Nello Cristianini 和 Christopher J. C. H. Watkins。利用字符串核进行文本分类。《机器学习研究杂志》，2:419–444，2002 年。[50] Terry Lyons 和 Zhongmin Qian。系统控制与粗糙路径。Clarendon，牛津，2007 年。[51] Cleve Moler 和 Charles Van Loan。计算矩阵指数的十九种不可靠方法：二十五年后。*SIAM* 评论，45(1):3–49，2003 年。[52] Krikamol Muandet、Kenji Fukumizu、Bharath Sriperumbudur、Bernhard Sch¨ olkopf 等。概率分布的核均值嵌入：综述与拓展。基础与趋势®机器学习，10(1-2):1–141，2017 年。[53] Gudmund Pammer。关于离散时间适应弱拓扑的一则注记。*arXiv* 预印本 *arXiv:2205.00989*，2022 年。[54] Carl Edward Rasmussen 和 Christopher K. I. Williams。面向机器学习的高斯过程。自适应计算与机器学习系列。麻省理工学院出版社，美国马萨诸塞州剑桥，2005 年 11 月。[55] Christophe Reutenauer。自由李代数。伦敦数学会专著。牛津大学出版社，牛津、纽约，1993 年 6 月。

[56] Hiroaki Sakoe 和 Seibi Chiba。用于口语词识别的动态规划算法优化。*IEEE* 声学、语音与信号处理汇刊，26(1):43– 49，1978。[57] Cristopher Salvi、Thomas Cass、James Foster、Terry Lyons 和 Weixin Yang。签名核是 Goursat偏微分方程的解。*SIAM*数据科学数学杂志，3(3):873–899，2021年1月。[58] Cristopher Salvi、Maud Lemercier、Chong Liu、Blanka Horvath、Theodoros Damoulas 和 Terry Lyons。用于捕获随机过程滤过的高阶核均值嵌入。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第 34 卷，第 16635–16647 页。Curran Associates, Inc.，2021。[59] Alexander Schell。盲源分离的稳健性分析。*arXiv:2303.10104[*数学、统计学*]*,2023年3月。[60] Alexander Schell 和 Harald Oberhauser。连续时间信号的非线性独立成分分析。*arXiv:2102.02876*，2021年6月。[61] Martin Schetzen。非线性系统的 *Volterra* 与 *Wiener* 理论。Krieger 出版社，2006。[62] Bernhard Sch¨olkopf，马克斯·普朗克智能系统研究所所长，德国图宾根机器学习教授 Bernhard Sch¨olkopf、Alexander J. Smola 和 Francis 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。[63] Bernhard Scholkopf 和 Alexander J Smola。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，2018。[64] Hiroshi Shimodaira、Ken-ichi Noma、Mitsuru Nakai 和 Shigeki Sagayama。支持向量机中的动态时间对齐核。神经信息处理系统进展，14，2001。[65]Carl-Johann Simon-Gabriel、Alessandro Barp、Bernhard Sch¨olkopf 和 Lester Mackey。利用最大均值差异度量弱收敛。*arXiv* 预印本 *arXiv:2006.09268*，2020。[66] Bharath K. Sriperumbudur、Kenji Fukumizu、Arthur Gretton、Gert R. G. Lanckriet 和 Bernhard Sch¨olkopf。概率分布的再生核希尔伯特空间嵌入的核选择与可分类性。载于神经信息处理系统大会，2009。[67] Csaba Toth、Patric Bonnier 和 Harald Oberhauser。Seq2Tens：一种通过低秩张量投影高效表示序列的方法。载于国际学习表征会议，2021。[68] Csaba Toth 和 Harald Oberhauser。利用签名协方差高斯过程从序列数据中进行贝叶斯学习。国际机器学习大会论文集，1，2020。