arXiv:2602.08427v1 [cs.LG] 9 Feb 2026

克里金法与大型神经网络之间的联系

Marius Marinescu富恩拉夫
拉达工程学院 胡安·卡洛斯国王
大学 马德里，西班牙
marius.marinescu@urjc.es

2026年2月10日
（更新）

摘要

人工智能已影响许多学科，如今无处不在。尤其是，空间统计学正处于关键时期，并将日益与人工智能交织融合。在此背景下，一个值得关注的问题是：空间统计学模型与机器学习（ML）模型之间是否存在关联？具体而言，本文探讨克里金法与神经网络之间的联系。乍看之下，两者似乎毫不相关。克里金法及其机器学习领域的对应方法——高斯过程回归——以概率论和随机过程为基础，而许多机器学习模型则被广泛视为黑箱模型。尽管如此，两者之间仍有着密切联系。我们研究了它们之间的联系，并重新审视了相关文献。理解二者的关系并结合两种视角，或许能够提升机器学习技术，使其更具可解释性、可靠性，并能更好地考虑空间因素。

关键词。克里金法、高斯过程、神经网络、机器学习、回归

1 引言

近几十年来，人工智能（AI）和机器学习（ML）革新了众多科学与工程领域，影响范围从计算机视觉延伸到自然语言处理。空间统计学是地球科学、环境建模和资源估算的基石，在这股变革浪潮中也不例外。我们正处于一个关键节点：人工智能技术与传统空间统计方法的结合，有望带来全新的洞见与能力。这种融合引出了一个根本性问题：经典空间统计模型与现代机器学习架构之间究竟存在怎样的内在联系（如果确实存在的话）？具体而言，克里金法——植根于地统计学和概率框架的历史悠久的方法——与看似晦涩难懂的神经网络“黑箱”模型有何关联?

乍看之下，克里金法与神经网络似乎毫无关联。克里金法最初为采矿应用而开发，用于插值空间数据，同时考虑自相关和不确定性；它深深植根于随机过程和线性预测理论。其机器学习对应方法——高斯过程回归（GPR）——将这一基础拓展为完整的概率范式。相比之下，神经网络，尤其是多层感知机（MLP）及其更深层的变体，通常被视为灵活但难以解释的函数逼近器，由数据和优化算法驱动。然而，在这些表面差异之下，隐藏着一种深刻的数学联系，将这两个领域紧密相连。

本文深入探讨这些联系，强调克里金预测器不仅等价于最大后验（MAP）高斯过程回归估计，而且也是大型神经网络的一种极限情形。据作者所知，克里金法从未从这一机器学习视角加以阐述。

通过探索这些联系，我们旨在形成一种统一视角，以同时发挥两个领域的优势：空间统计学的可解释性与不确定性量化，以及机器学习的可扩展性与通用性。这样的理解可以为机器学习模型注入空间感知，同时以神经网络架构的自适应能力丰富空间统计学。

本文其余部分组织如下。首先，第1节介绍克里金法与高斯过程回归的历史背景。然后，第2节阐明克里金法与高斯过程回归之间的数学联系。接下来，第3节论证多层感知器在无限宽度极限下收敛于高斯过程回归，并给出示例与数值说明。最后，我们讨论更广泛的意义，包括向深度网络或机器学习训练的扩展，并作出结论。

1.1 克里金法与高斯过程回归的历史背景

In t本节简要介绍　　克里金法及其机器学习　　对应方法——高斯过程回归。

1.1.1 克里金法

Daniel Gerhardus Krige 曾在南非威特沃特斯兰德盆地的金矿工作（Minnitt & Assibey- Bonsu 2015）。这些矿山面临的主要挑战之一，是准确估算矿体的黄金含量。

从技术经济角度来看，只有当开采和加工成本不超过可从矿山中提取的金属价值时，才值得开采。矿区块段的真实品位，即单位岩石中有价值物质的含量，在开采前并不清楚，因此需要通过样本进行估算。20世纪50年代初，估算方法很简单，就是取属于该矿区块段或位于其边界处的数据的平均品位。D. Krige 注意到，金矿床具有空间连续性，观测值并非相互独立，而是彼此相关；他还惊讶地发现，平均而言，低品位矿区块段的估值偏低，而高品位矿区块段的估值偏高。

他发现，矿区块段品位观测值接近对数正态概率律（例如，参见 Krige（1951，图 2）），并运用经典回归理论，在给定样本的情况下改进矿区块段品位估计，使其无偏（例如，参见 Krige（1962，图 24 和 25））。通过采用他的方法，从恰当的统计学视角解决了使用样本均值所带来的系统误差问题，并大幅提高了矿区块段品位估计的准确性，参见 Krige（1951，第 6 节——基于统计理论的改进估计）。

促使 D. Krige 开展这项工作的原因之一，是当时采矿领域对统计学的运用还很初级。以下引用他在 Krige（1951）中的一段话：“目前，这些方法几乎完全是运用简单的算术和经验公式，并以实践经验为指导，却忽视了通过审慎的统计分析所能获得的诸多益处……”关于 D. Krige 的担忧，更多详情请参阅前述论文。

另一方面，由于未考虑观测值之间的（空间）自相关结构，估计量的效率并不高。这正是 G. Matheron 介入的地方。法国学界开始常规使用术语“Le Krieage”，而 G. Matheron 在其出版物中以 D. Krige 的工作为敬，将“Kriging”一词创造出来（Cressie 1990）。他呼吁所有关注空间插值的科学家采用这一术语，此后该术语也在盎格鲁-撒克逊采矿术语中普及开来。克里金法所做的事情与 Matheron 创造的“Kriging”一词所指的内容并不完全相同。给定一个随机场，Matheron 将克里金法定义为一种利用现有观测值来预测未观测值或区块平均值的方法。具体而言，他在空间统计学框架下推导出了最佳线性无偏预测量（BLUP），但没有使用这一术语（Matheron 1971）。另一方面，D. Krige 最初的想法是，用矿区全部数据的样本均值与该矿区块段的样本均值的凸和来估计该矿区块段的真实品位。因此，他使用的并非所有观测值（如 Matheron 所用）的加权和，而是全局样本均值与局部样本均值的加权和。更多详情见 Marinescu（2024）。

Matheron 被认为是地统计学的创始人（Chilès & Desassis 2018年）。在 Agterberg（2004）中，他被誉为二十世纪最杰出的数理统计学家之一，其地位堪比 Ronald Fisher 或 John Tukey。

1.1.2 高斯过程回归

我们今天所熟知的高斯过程回归，其形式化及广泛应用源自不同的发展脉络，始于20世纪90年代，主要由机器学习领域的进步所推动。

高斯过程回归的概念基于高斯过程的概念。Andrey Kolmogorov 是高斯过程理论基础最早的奠基者之一，他为随机过程理论奠定了基础（Kolmogorov 1938）。与发展历程较为分散的克里金法不同，Kolmogorov 在概率论和随机过程方面的研究提供了严谨的数学框架，后来被证明是高斯过程回归发展的重要基础。由 Kolmogorov 的昔日学生撰写、探讨高斯过程数学基础的开创性著作是 Ibragimov & Rozanov（1978）。

[1]在采矿领域，矿区块段指矿山中正在开采或已为开采做好准备的特定区域或分区。

将高斯过程用于估计，源于先驱统计学家和数学家们的研究，他们致力于理解并建模随时间和空间变化的随机过程。20世纪下半叶，研究人员认识到高斯过程在对不同类型的数据进行建模时具有很强的适用性。这一时期见证了重要理论进展和实际应用的发展。20世纪90年代末至21世纪初，在机器学习兴起和计算资源日益充足的推动下，高斯过程回归迎来了一个重要的转折点。

Christopher Williams 和 Carl Rasmussen 是推动机器学习转型的关键人物。他们编写了颇具影响力的教科书《机器学习中的高斯过程》（Williams & Rasmussen 2006），汇总了此前的理论进展，并将其应用于包括回归和分类在内的广泛问题。他们展示了高斯过程回归的实用价值：它不仅能提供解，还能通过后验分布衡量不确定性。此外，他们还讨论了高斯过程回归与其他数学模型和机器学习模型之间的联系（Williams & Rasmussen 2006，第 6 章和第 7 章）。

高斯过程回归具有若干吸引人的特性和优势，因此广受欢迎：

- 基于随机过程，具有清晰的数学基础。
- 能够为预测提供不确定性度量。
- 模型具有灵活性。通过选择不同的核函数，高斯过程回归可以对广泛的目标函数进行建模，并捕捉不同的数据模式。
- 其非参数特性。与参数模型不同，高斯过程回归不假设底层（先验）函数具有固定形式，因此能够高度适应复杂数据集。
- 高斯过程回归的可解释性。其概率特性以及协方差函数的显式形式，使高斯过程回归具有可解释性，与黑箱模型形成对比。

高斯过程回归的发展历程展现了现代统计学与机器学习方法的跨学科特性。从20世纪初期的理论研究起源，到实际应用以及发展成为现代机器学习中的基础工具，高斯过程回归不断适应并拓展其应用范围。如今，它已成为一种功能强大且用途广泛的方法，可用于对复杂数据建模。

2 克里金法与高斯过程回归

本节介绍克里金法与高斯过程回归之间的数学关系。高斯过程回归的形式体系将使我们能够把克里金法与大型神经网络联系起来。给定一个带有观测值的随机场，克里金法的目标是预测一个未观测值或区块平均值。克里金法基于以下方程：

$$Y(\boldsymbol{x}_i) = Z(\boldsymbol{x}_i) + \varepsilon_i,\ i = 1, \ldots, n$$

其中 $Y(\boldsymbol{x}_i)$ 为观测值， $Z(\boldsymbol{x}_i)$ 是一个二阶随机场（不一定是高斯的）， ε_i 为若干独立同分布噪声。令 m 表示 Z 的均值函数，k 表示自协方差函数。为进行预测，克里金法寻求统计学中所称的最佳线性无偏预测[2]。著名的简单克里金法（简单克里金）解为（Chilès & Desassis（2018年））：

$$\hat{Z}(\boldsymbol{x}_*) = m_* + \boldsymbol{k}_*^\top (\Sigma + \sigma^2 I)^{-1} (Y - \boldsymbol{m})$$

其中 $m_* = m(\boldsymbol{x}_*)$、 $\boldsymbol{m} = \big(m(\boldsymbol{x}_1)\ m(\boldsymbol{x}_2)\ldots\ldots,\ m(\boldsymbol{x}_n)\big)^\top$， $\boldsymbol{k}_* = \big(k(\boldsymbol{x}_1,\ \boldsymbol{x}_*)\ k(\boldsymbol{x}_2,\ \boldsymbol{x}_*)\ldots\ldots,\ k(\boldsymbol{x}_n,\ \boldsymbol{x}_*)\big)^\top$， $\Sigma = (k(\boldsymbol{x}_i,\ \boldsymbol{x}_j))_{i,\ j=1,\ \ldots\ldots,\ n}$ 以及 $\sigma^2 = \mathbb{V}[\varepsilon]$。

可以看出，我们考虑的是均值（可能）非零但已知的简单克里金描述。现在我们介绍高斯过程回归的设定。高斯过程回归的形式体系将 $Y(\boldsymbol{x})$ 视为一个高斯随机场。因此，

$$\big(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n)\big) \sim \mathcal{N}(m(X),\ \Sigma + \sigma^2 I)$$

新位置 $Z_* = Z(\boldsymbol{x}_*)$ 处的估计值通过计算条件分布 $Z_* \mid Y = \boldsymbol{y}$ 得出：

$$\begin{bmatrix} Y \\ Z_* \end{bmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} \boldsymbol{m} \\ m_* \end{bmatrix}, \begin{bmatrix} \Sigma + \sigma^2 I & \boldsymbol{k}_* \\ \boldsymbol{k}_*^\top & k_{**} \end{bmatrix}\Big) \qquad \Rightarrow \qquad Z_* \mid \boldsymbol{y} \sim \mathcal{N}\left(m_* + \boldsymbol{k}_*^\top (\Sigma + \sigma^2 I)^{-1} (\mathbf{y} - \boldsymbol{m}),\right.$$
$$\left. k_{**} - \boldsymbol{k}_*^\top \Sigma^{-1} \boldsymbol{k}_*\right).$$

[2]估计确定性量（例如经典回归中的参数）时，通常使用“最佳线性无偏估计”这一术语；估计随机量时，则使用“最佳线性无偏预测”这一术语。

其中 $k_{**} = k(\boldsymbol{x}_*,\ \boldsymbol{x}_*)$。该条件分布的最大后验（**MAP**）给出简单克里金预测器，而且其方差相同（Marinescu (2024)）。
因此，克里金预测与高斯过程回归最大后验预测器相同。从这个意义上说，这两种技术可以视为等价的。

我们使用简单克里金来描述克里金法与高斯过程回归之间的联系。其他克里金法变体，例如普通克里金和泛克里金，也与高斯过程回归直接相关。普通克里金和泛克里金方法使用所谓的克里金方程组，这是一种巧妙的设置，将未知均值函数参数隐藏在拉格朗日乘子中，从而实现预测。可以证明，该过程等价于对均值参数应用最佳线性无偏估计量方法，并对预测应用最佳线性无偏预测方法。有关简单克里金、普通克里金和泛克里金与高斯过程回归之间联系的完整讨论，请参见 Marinescu（2024）。
从现在起，我们将交替使用克里金法和高斯过程回归这两个术语，因为按照前文所述的含义，它们在数学上是等价的。

3 克里金法作为多层感知机

本节将展示克里金法与具有一个隐藏层的多层感知机之间的联系。考虑一个多层感知机，其输入为 $\boldsymbol{x}$ （表示空间变量），具有一个包含 L 个神经元的隐藏层，并采用传递函数 h（假设该函数有界），如图1所示。那么，其输出（表示预测）可写为:

$$y(\boldsymbol{x}) = b_0 + \sum_{j=1}^{L} b_j h(\boldsymbol{x}; \boldsymbol{a}_j)$$

输出是隐藏单元的线性组合，并带有偏置 b_0。这种架构具有一项非常重要的性质。研究表明，随着隐藏神经元数量趋于无穷，只要激活函数局部黎曼可积且非多项式，它就能在紧集上均匀逼近任意目标连续函数（通用逼近器）（Hornik（1993））。高斯过程回归模型也有类似的性质（Micchelli 等人（2006），Williams & Rasmussen（2006））。

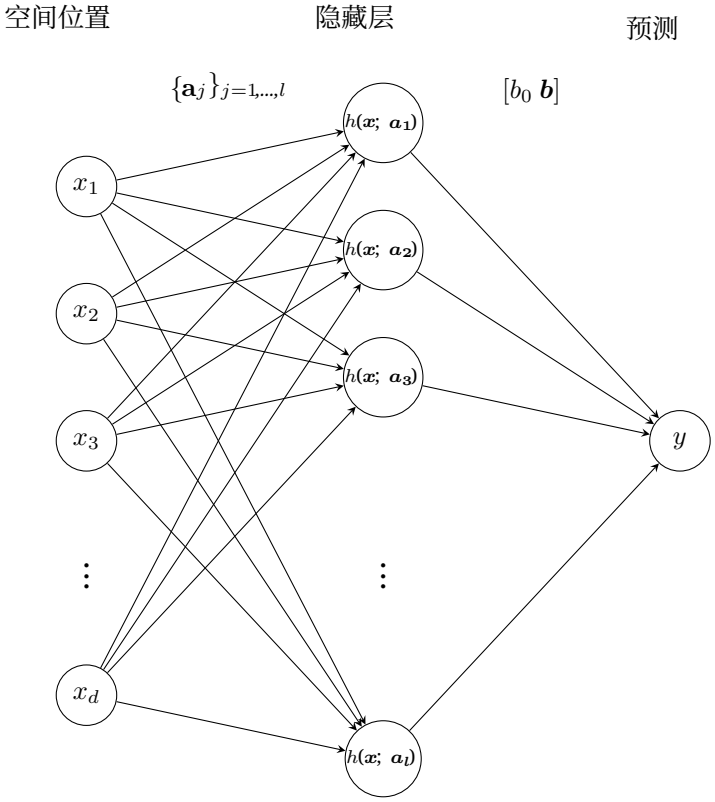


图 1：一个包含一个隐藏层和传递函数 h 的多层感知机架构。

核名称	传递函数	核	带 $\boldsymbol{a}$ 非适当的核
线性	$h(\boldsymbol{x};\boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{x}$ $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0},\ \Sigma)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = \boldsymbol{x}^\top \Sigma \boldsymbol{x'}$	-
神经网络	$h(\boldsymbol{x};\boldsymbol{a}) =$ 误差函数$(\boldsymbol{a}^\top \boldsymbol{x})$ $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0},\ \Sigma)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = \frac{2}{\pi}$ 反正弦 $\frac{2\boldsymbol{x}^\top \Sigma \boldsymbol{x'}}{(1+2\boldsymbol{x}^\top \Sigma \boldsymbol{x})(1+2\boldsymbol{x'}^\top \Sigma \boldsymbol{x'})}$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = \frac{2}{\pi}$ 反正弦 $\frac{\boldsymbol{x}^\top \boldsymbol{x'}}{\|\|\boldsymbol{x}\|\|^2 \|\|\boldsymbol{x'}\|\|^2}$
平方指数	$h(\boldsymbol{x};\ \boldsymbol{a}) = \sqrt{2}$ 余弦 $(\boldsymbol{a}^\top \boldsymbol{x} + \varphi_0)$ $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\sigma^2} I) \quad \varphi_0 \sim \mathcal{U}(0, 2\pi)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = \exp\left(-\frac{\|\|\mathbf{x}-\mathbf{x'}\|\|^2}{2\sigma^2}\right)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = \begin{cases} 1 & \text{如果 } \boldsymbol{x} = \boldsymbol{x'} \\ 0 & \text{如果 } \boldsymbol{x} \neq \boldsymbol{x'} \end{cases}$
非平稳 平方指数	$h(\boldsymbol{x};\ \boldsymbol{a}) = \exp\left(-\|\|\boldsymbol{x}-\boldsymbol{a}\|\|^2/2\sigma_g^2\right)$ $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 I)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) \propto$ 指数$\left(-\frac{\mathbf{x}^\top}{c_1}\right)$ 指数$\left(-\frac{\|\|\mathbf{x}-\mathbf{x'}\|\|^2}{c_2}\right)$ 指数$\left(-\frac{\mathbf{x'}^\top}{c_1}\right)$ 对于某个常数 $\mathbf{c}_\circ \quad c_1, c_2{}_\circ$	$k(\boldsymbol{x},\ \boldsymbol{x'}) \propto$ 指数$\left(-\frac{\|\|\mathbf{x}-\mathbf{x'}\|\|^2}{4\sigma_g^2}\right)$ ———
反余弦核 I	$h(\boldsymbol{x};\boldsymbol{a}) = \delta_{\text{Heaviside}}(\boldsymbol{a}^\top \boldsymbol{x})$ $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, I)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = 1 - \frac{\theta}{\pi}$, 余弦 $\theta = \frac{\boldsymbol{x}^\top \boldsymbol{x'}}{\|\|\boldsymbol{x}\|\| \cdot \|\|\boldsymbol{x'}\|\|}$	-
反余弦核 II	$h(\boldsymbol{x};\boldsymbol{a}) = \max(0,\ \boldsymbol{a}^\top \boldsymbol{x}) \ -$ ReLU $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, I)$	$k(\boldsymbol{x},\ \boldsymbol{x'}) = \frac{\|\|\boldsymbol{x}\|\| \cdot \|\|\boldsymbol{x'}\|\|}{\pi}$(正弦 $\theta + (\pi - \theta)$ 余弦 θ), 余弦 $\theta = \frac{\boldsymbol{x}^\top \boldsymbol{x'}}{\|\|\boldsymbol{x}\|\| \cdot \|\|\boldsymbol{x'}\|\|}$	-

表 1：核函数与传递函数列表（非详尽）。神经网络和非平稳平方指数核可见于 Williams（1998年）。反余弦 I 和 II 可见于 Cho 和 Saul（2009年）。文献中未找到的线性核和平方指数核的计算见附录 A 和 B。

为了使神经网络（NN）具有随机性，我们引入一些随机因素。令 b 服从方差为 σ_b^2 的独立零均值分布，并令每个 $\boldsymbol{a}_j$ 均为独立同分布。于是， $E[Y(\boldsymbol{x})] = 0$ 按构造成立，且协方差为

$$E[Y(\boldsymbol{x})Y(\boldsymbol{x'})] = \sigma_b^2 + \sum_{j=1}^{L} \sigma_b^2 E[h(\boldsymbol{x};\boldsymbol{a}_j)h(\boldsymbol{x'};\boldsymbol{a}_j)] = \\ = \sigma_b^2 + L \cdot \sigma_b^2 E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x'};\boldsymbol{a})]$$

其中，在第一个等式中我们使用了独立性，最后一个等式成立，是因为所有 e
隐藏单元都服从相同的分布 。 Y大写是为了强调现在是随机的。

由此可知，在式 3 中令尺度参数 $L \cdot \sigma_b^2 \to$ cte.，协方差有界，因为假设 h 有界。有趣的是，式 3 所描述的随机过程是一个方差有界的独立同分布（i.d.d.）和（式 3）；根据中心极限定理（CLT），当隐藏单元数量趋于无穷时，它收敛到高斯过程。通过计算$E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x'};\boldsymbol{a})]$,，可以得到刻画每个神经网络的协方差（以及变差函数）的显式形式。具体而言，

$$\int_{\mathbb{R}^L} h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x'};\boldsymbol{a}) f_{\boldsymbol{a}}(\boldsymbol{u})\, d\boldsymbol{u} \ \overset{\text{NN Cov}}{\propto} \ k(\boldsymbol{x},\boldsymbol{x'}) \overset{\text{Kriging Kernel}}{}$$

其中， f 是 $\boldsymbol{a}$的密度函数。

因此，克里金法与多层感知机模型密切相关。具体而言，在上述假设下，随着隐藏单元数量趋于无穷，多层感知机模型趋近于高斯过程。此外，可以直接根据多层感知机模型的设定推导出该高斯过程的均值函数和协方差（核）。由于这种收敛基于中心极限定理，通常只需 30 个隐藏单元，就足以很好地近似高斯过程的极限分布。第3节的数值演示中展示了相关例子。如需具体的（理论）界，可使用 Berry-Esseen 界（Vershynin（2018年））。

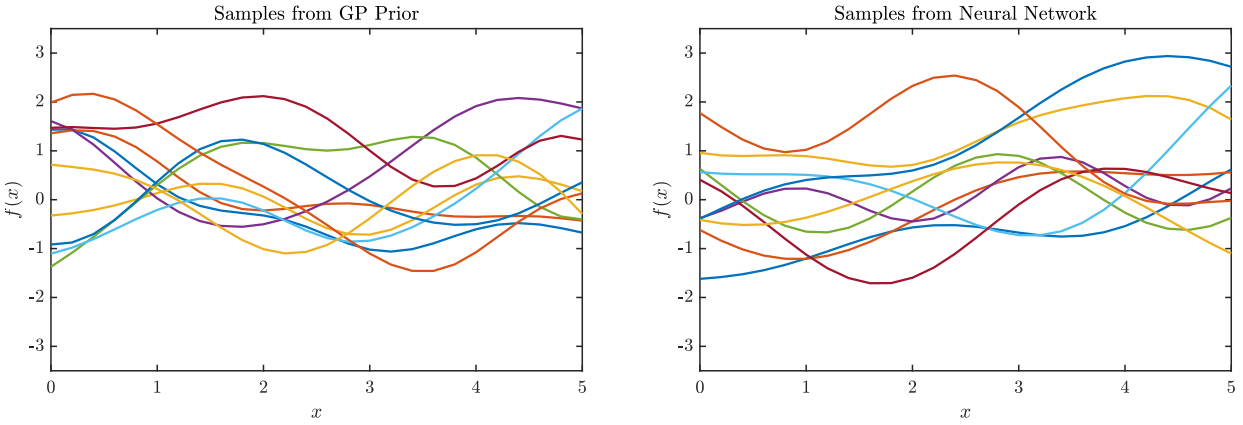
得益于 [5-6], 等人的研究，我们介绍了若干传递函数及其对应的协方差函数示例，列于表 1 中。可以看到，线性核、平方指数核和神经网络核等知名核，都有相应的多层感知机传递函数。通过为 $\boldsymbol{a}$ 选择不同的概率分布，也可以构造新的平稳核。对于许多传递函数，可能难以用初等函数显式计算式 3 中的积分，因此可以采用数值近似。总体而言，大多数多层感知机传递函数都会产生非平稳核（表 1 中只有平方指数核是平稳核）。这或许可以解释神经网络架构为何具有如此广泛的适用性，并能很好地适应许多不同的应用问题。

请注意，文献中所谓的“sigmoid核”从来都不是正定的，因此不是有效的协方差函数（Williams & Rasmussen（2006年）），所以将其省略。

3.1 一些数值证明

在本节中，我们针对特定情况，对第3节中描述的理论关系进行数值检验。具体而言，我们分别从多层感知机和高斯过程中进行模拟，并对随机样本执行统计检验，以评估它们是否可以被视为服从相同分布。

图2展示了一组包含10条样本路径的图。左图由采用平方指数核的高斯过程生成，右图则由采用余弦激活函数的多层感知机模型生成。



(a) 采用平方指数核的高斯过程先验。

(b) 采用余弦激活函数、包含20个隐藏单元的多层感知机。

图2：来自高斯过程和神经网络的样本。长度尺度 σ 设为1。

为了检验它们是否来自同一分布，我们使用了基于函数型数据深度概念的随机函数秩检验（López-Pintado 与 Romo 2009，第 8 节）。该检验利用由样本中两个或多个函数的图形界定的带这一概念。借助这一工具，可以衡量每个随机函数的深度或中心性，从而自然地对它们进行由中心向外的排序。该方法有多种应用，包括使用秩检验[3] 来判断两组曲线是否来自同一总体（更多信息请参见 López-Pintado 与 Romo (2009)）。原假设是它们来自同一总体，而备择假设则是它们并非来自同一总体。

为了应用该检验，我们分别使用多层感知机模型和高斯过程模型，从线性随机过程和平方指数随机过程中生成了 50 个随机实现。对于多层感知机模型，仅使用了 200 个隐藏单元；在两种情况下，$\boldsymbol{a}$ 的协方差矩阵都设为单位矩阵。应用该检验后发现，两种情况下都应接受原假设，对应的 p-值分别为 0.63 和 0.91。我们还测试了其他配置，包括改变样本量、隐藏单元数量和方差，结果相似。因此，这些检验从经验上支持了本节讨论的理论结果。

3.2 进一步的联系

GP 与 MLP 之间的联系可以推广到深度神经网络（DNNs）和卷积神经网络（CNNs）。在无限宽度极限下，即每个隐藏层中的神经元数量趋于无穷大，同时采用经过适当缩放的随机初始化，DNN 和 CNN 会收敛为 GP。这个极限 GP 的核由网络架构决定，被称为神经网络高斯过程（NNGP）核（Lee 等人（2017年），Garriga-Alonso 等人（2018年））。此外，Yang（2019年）还将无限宽度极限下神经网络与 GP 的等价性推广并描述到一大类神经网络架构中。

神经网络模型的常见训练方式也与高斯过程回归估计有关。所谓的神经切线核（NTK）是一种核，用于描述深度人工神经网络在通过梯度下降进行训练时的演化（Jacot 等人（2018年）、Chizat 等人（2019年））。在无限宽度状态下，针对网络参数进行梯度下降会产生由神经切线核支配的线性化训练动力学，而神经切线核本身定义了一种类似于高斯过程中协方差的核。有趣的是，这一性质表明，深度神经网络训练可以被解释为在高维特征空间中执行核回归，从而弥合核方法训练与深度学习之间的概念鸿沟。

[3]该秩检验可以被视为 应用于随机函数的推广 双侧 Wilcoxon 符号秩检验。

高斯过程回归也与支持向量机和相关向量机有关（Williams & Rasmussen，2006年，第6章），但它们并不等同。前者通常能产生更稀疏的解，且其预测分布也有所不同。

总而言之，克里金法和高斯过程回归处于多个关键机器学习范式的交汇处，展现了深刻的概念联系。这表明，它们并非孤立的方法，而是连接概率推断、核方法与深度学习的更广阔领域的一部分。归根结底，基于核的视角能够提升机器学习技术，使其更具可解释性、可靠性，并更充分地考虑空间特性。

参考文献

Agterberg, F. (20 04), ‘Georges matheron：空间统计学的创始人’，地球科学史 **23**(2), 205–334 .

Chilès, J.-P. & Desassis, N. (2018年), ‘克里金法五十年’，数学地球科学手册：国际数学地球科学协会五十年 pp.589–612.

Chizat, L., Oyallon, E. & Bach, F.（2019），《可微编程中的惰性训练》，神经信息处理系统进展 **32**。

Cho, Y. &Saul, L.（2009），《深度学习的核方法》，神经信息处理系统进展 **22**。Cressie, N.（1990），《克里金法的起源》，数学地质学 **22**，239–252。

Garriga-Alonso, A., Rasmussen, C. E. & Aitchison, L.（2018），《深度卷积网络作为浅层高斯过程》，*arXiv* 预印本 *arXiv:1808.05587* 。

Hornik, K.（1993），《神经网络逼近的一些新结果》，神经网络 **6**(8)，1069–1072。

Ibragimov, I. A. & Rozanov, Y. A.（1978），高斯随机过程，第9卷，Springer Science & Business Media。

Jacot, A., Gabriel, F. & Hongler, C.（2018年），《神经切线核：神经网络中的收敛与泛化》，神经信息处理系统进展 **31**。

Kolmogorov, A. N.（1938年），《概率论的解析方法》，*Uspekhi matematicheskikhnauk* (5), 5–41。

Krige, D. G.（1951年），《威特沃特斯兰德矿山估值若干基本问题的统计方法》，南部非洲采矿与冶金学会期刊 **52**(6), 119–139。

Krige, D. G.（1962年），《矿山估值中的统计应用》，南非矿山测量师学会期刊 **12**(2)。

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J. & Sohl-Dickstein, J.（2017年），《深度神经网络作为高斯过程》，*arXiv* 预印本 *arXiv:1711.00165* 。

López-Pintado, S. & Romo, J.（2009年），《关于函数型数据深度概念》，美国统计学会会刊**104**(486), 718–734。

Marinescu, M.（2024），《阐释并联系克里金法与高斯过程回归》，*arXiv* 预印本 *arXiv:2408.02331* 。

Matheron, G.（1971），《区域化变量理论及其应用》，枫丹白露数学形态学中心论文集(5)。
Micchelli, C. A., Xu, Y. & Zhang, H.（2006），《通用核》，机器学习研究期刊 **7**(12)。Minnitt, R. & Assibey-Bonsu, W.（2015），《谨以此文纪念 dg krige frssaf 教授》，南非采矿与冶金学会会刊**115**(1), VIII–XI。Vershynin, R.（2018），《高维概率：数据科学应用导论》，剑桥统计与概率数学丛书，剑桥大学出版社。Williams, C. K.（1998年），《无限神经网络的计算》，神经计算 **10**(5), 1203–1216。
Williams, C. K. & Rasmussen, C. E.（2006年），《机器学习中的高斯过程》，麻省理工学院出版社，马萨诸塞州剑桥。

Yang, G.（2019年），《任意架构的宽前馈神经网络或循环神经网络都是高斯过程》，神经信息处理系统进展**32**。

附录

A 线性核

我们将证明

$$E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})] = \boldsymbol{x}^\top \Sigma \boldsymbol{x}',$$

当 $\boldsymbol{a} \sim \mathcal{N}(0,\ \Sigma)$ 且 h 线性时。

证明

$$E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})] = E[\boldsymbol{a}^\top \boldsymbol{x} \cdot \boldsymbol{a}^\top \boldsymbol{x}'] = E[\boldsymbol{x}^\top \boldsymbol{a} \cdot \boldsymbol{a}^\top \boldsymbol{x}'] =$$
$$= \boldsymbol{x}^\top E[\boldsymbol{a} \cdot \boldsymbol{a}^\top] \boldsymbol{x}' = \boxed{\boldsymbol{x}^\top \Sigma \boldsymbol{x}'}$$

其中我们有　用到了 $\boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{a}$，因为它是标量，并且 $E[\boldsymbol{a} \cdot \boldsymbol{a}^\top]$ 按定义是 $\boldsymbol{a}$ 的协方差　　n.

B 平稳平方指数核

我们将证明

$$E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})] = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right),$$

当 $\boldsymbol{a} \sim \mathcal{N}(0, \Sigma = \sigma^{-2}I)$、 $\varphi_0 \sim \mathcal{U}(0, 2\pi)$且 $h(\boldsymbol{x};\ \boldsymbol{a}) = \sqrt{2}$ 余弦 $(\boldsymbol{a}^\top \boldsymbol{x} + \varphi_0)$。

证明

$$E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})] = E[\sqrt{2}\cos(\boldsymbol{a}^\top \boldsymbol{x} + \varphi_0)\sqrt{2}\cos(\boldsymbol{a}^\top \boldsymbol{x}' + \varphi_0)]$$

现在我们使用三角恒等式:

$$\cos a \cos b = \frac{1}{2}(\cos(a-b) + \cos(a+b))$$

因此得到:

$$E[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})] =$$
$$E[\cos(\boldsymbol{a}^\top(\boldsymbol{x}-\boldsymbol{x}'))] + E[\cos(\boldsymbol{a}^\top(\boldsymbol{x}+\boldsymbol{x}')) + 2\varphi_0] =$$
$$E[\cos(\boldsymbol{a}^\top(\boldsymbol{x}-\boldsymbol{x}'))]$$

第二个期望值为零，因为它是随机相位的平均值: $\int_0^{2\pi}$ 余弦 $(2\varphi_0 + cte.) \cdot \mathcal{U}(0,\ 2\pi)\ d\varphi_0 = 0$。

剩余的期望值仅取决于 $\boldsymbol{a}$。现在利用 $E\big[\text{余弦}\big(\boldsymbol{a}^\top \boldsymbol{t}\big)\big] = \text{Re}\big(E\big[e^{i\boldsymbol{a}^\top \boldsymbol{t}}\big]\big)$，以及 $\boldsymbol{a} \sim \mathcal{N}(0,\ \Sigma)$ 的特征函数为

$$E\left[e^{i\boldsymbol{a}^\top \boldsymbol{t}}\right] = \exp\big(-\tfrac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\big).$$

令 $\boldsymbol{t} = \boldsymbol{x} - \boldsymbol{x}'$，并取实部（该值本身已为实数且为正），得到

$$E_{\boldsymbol{a},\varphi_0}\big[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})\big] = \exp\big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^\top \Sigma(\boldsymbol{x}-\boldsymbol{x}')\big).$$

特别地，对于 $\Sigma = \sigma^{-2}I$

$$\boxed{\mathbb{E}_{\boldsymbol{a},\varphi_0}\big[h(\boldsymbol{x};\boldsymbol{a})h(\boldsymbol{x}';\boldsymbol{a})\big] = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right)}.$$